Genomic Prediction of Drought Tolerance during Seedling Stage in Maize using Low-Cost Markers

Ao Zhang  
Shenyang Agricultural University

Shan Chen  
Shenyang Agricultural University

Zhenhai Cui  
Northeast Institute of Geography and Agroecology Chinese Academy of Sciences

Yubo Liu  
Shanghai Academy of Agricultural Sciences

Yuan Guan  
Shanghai Academy of Agricultural Sciences

Shuang Yang  
Shenyang Academy of Agricultural Sciences

Jingtao Qu  
Sichuan Agricultural University

Juchao Nie  
Fumeng County modern agricultural development service center

Dongdong Dang  
Shanghai Academy of Agricultural Sciences

Cong Li  
Shenyang Agricultural University

Xiaomei Dong  
Shenyang Agricultural University

Jinjuan Fan  
Shenyang Agricultural University

Yanshu Zhu  
Shenyang Agricultural University

Xuecai Zhang  
CIMMYT: Centro Internacional de Mejoramiento de Maiz y Trigo

Jose Crossa  
CIMMYT: Centro Internacional de Mejoramiento de Maiz y Trigo

Huiying Cao  
Shenyang Agricultural University

Yanye Ruan
Genomic Prediction of Drought Tolerance during Seedling Stage in Maize using Low-Cost Markers

Ao Zhang1,2,3+, Shan Chen2+, Zhenhai Cui4, Yubo Liu1,3, Yuan Guan1, Shuang Yang5, Jingtao Qu6, Juchao Nie7, Dongdong Dang1,2, Cong Li2, Xiaomei Dong2, Jinjuan Fan2, Yanshu Zhu2, Xuecai Zhang1,3, Jose Crossa3, Huiying Cao3, Yanye Ruan2✉, Hongjian Zheng1✉

1 CIMMYT-China Specialty Maize Research Center, Crop Breeding and Cultivation Research Institute, Shanghai Academy of Agricultural Sciences, Shanghai, China.
2 College of Bioscience and Biotechnology, Shenyang Agricultural University, Shenyang, Liaoning, China. Shenyang City Key Laboratory of Maize Genomic Selection Breeding, Shenyang, Liaoning, China.
3 International Maize and Wheat Improvement Center (CIMMYT), El Batan, Texcoco, México.
4 Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China.
5 Shenyang Academy of Agricultural Sciences, Liaoning, China.
6 Maize Research Institute, Sichuan Agricultural University, Wenjiang, Sichuan, China
7 Fumeng County Modern Agricultural Development Service Center, Fuxin, Liaoning Province, China

+These authors contributed equally to this research presented in this article.
Corresponding author:
✉ Yanye Ruan, E-mail address: yanyeruan@syau.edu.cn;
✉ Hongjian Zheng, E-mail address: hjzh6188@163.com.

Abstract

Drought tolerance in maize is a complex and polygenic trait, especially in the seedling stage. In plant breeding, such traits can be improved by genomic selection (GS), which has become a practical and effective tool. In the present study, a natural maize population named Northeast China core population (NCCP) consisting of 379 inbred lines were genotyped with diversity arrays technology (DArT) and genotyping-by-sequencing (GBS) platforms. Target traits of seedling emergence rate (ER), seedling plant height (SPH), and grain yield (GY) were evaluated under two natural drought environments in northeast China. Adequate genetic variants have been found for genomic selection, they are not stable enough between
two years. Similarly, the heritability of the three traits is not stable enough, and the
heritabilities in 2019 (0.88, 0.82, 0.85 for ER, SPH, GY) are higher than that in 2020 (0.65,
0.53, 0.33) and cross-two-year (0.32, 0.26, 0.33). The current research obtained two kinds of
marker sets: the SilicoDArT markers were from DArT-seq, and SNPs were from the GBS and
DArT-seq. In total, a number of 11,865 SilicoDArT, 7,837 DArT's SNPs, and 91,003 GBS
SNPs were used for analysis after quality control. The results of phylogenetic trees showed
that the population was rich in consanguinity. Genomic prediction results showed that the
average prediction accuracies estimated using the DArT SNP dataset under the 2-fold
cross-validation scheme were 0.27, 0.19, and 0.33, for ER, SPH, and GY, respectively. The
result of SilicoDArT is close to the SNPs from DArT-seq, those were 0.26, 0.22, and 0.33.
For SPH, the prediction accuracies using SilicoDArT were more than ones using DArT SNP,
In some cases, alignment to the reference genome results in a loss to the prediction. The trait
with lower heritability can improve the prediction accuracy using filtering of linkage
disequilibrium. For the same trait, the prediction accuracy estimated with two types of DArT
markers was consistently higher than those estimated with the GBS SNPs under the same
genotyping cost. Our results show the prediction accuracy has been improved in some cases
of controlling population structure and marker quality, even when the density of the marker is
reduced. In the initial maize breeding cycle, Silicodart markers can obtain higher prediction
accuracy with a lower cost. However, higher marker density platforms i.e. GBS may play a
role in the following breeding cycle for the long term. The natural drought experimental
station can reduce the difficulty of phenotypic identification in a water-scarce environment.
The accumulation of more yearly data will help to stabilize the heritability and improve
predictive accuracy in maize breeding. The experimental design and model for drought
resistance also need to be further developed.

**Keywords** Maize (*Zea mays* L.); GBS; DArT; GS; Drought Stress

**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>BLUE</td>
<td>Best linear unbiased estimate</td>
</tr>
<tr>
<td>BP</td>
<td>Breeding population</td>
</tr>
<tr>
<td>DArT</td>
<td>Diversity arrays technology</td>
</tr>
<tr>
<td>DL</td>
<td>Deep learning</td>
</tr>
<tr>
<td>ER</td>
<td>Emergence rate</td>
</tr>
<tr>
<td>GBS</td>
<td>Genotyping-by-sequencing</td>
</tr>
<tr>
<td>GEBV</td>
<td>Genomic estimated breeding values</td>
</tr>
<tr>
<td>GS</td>
<td>Genomic selection</td>
</tr>
</tbody>
</table>
1. Introduction

Maize (*Zea mays* L.) is an important source of food, animal feed, and energy because of its high yield potential (Jiang et al. 2018). Consequently, poor maize harvests, usually caused by environmental stresses, will have a tremendously detrimental impact on human life. Seed germination and seedling establishments are the initial stages of maize life and influence seedling emergence rate, uniformity, and robustness, which determine maize yield potential (Tian et al. 2014). However, these stages of maize are frequently subjected to severe drought stress worldwide due to increasingly global warmer weather, which decreases the emergence rate and retards seedling growth. Thus, it is necessary to breed more drought-tolerant maize hybrids.

Selection of elite drought-resistant maize inbred lines is the basis of obtaining drought-tolerant hybrids. However, traditional methods of inbred line selection are not only time-consuming, but also labor-intensive and inefficient. In recent decades, with the development of high-throughput and inexpensive genotyping technologies, genomic selection (GS) has become a practical and effective tool for animal and plant breeding (Michel et al. 2016; Zhang et al. 2017b). The principle of GS is to establish the prediction model based on
the genotypic and phenotypic data from a training population (TP), which is used to derive genomic estimated breeding values (GEBVs) for all the individuals in the breeding population (BP) from their genomic profiles (Meuwissen et al. 2001), guiding breeders to select excellent individuals from the population without phenotypic data. GS has been shown to improve the efficiency of recurrent selection of bi-parental populations (Beyene et al. 2015; Massman et al. 2013; Môro et al. 2019; Vivek et al. 2017) and multi-parent population (Zhang et al. 2017b) to minimize breeding cycle times and maximize the genetic gain per year. At present, GS has been widely used in maize trait selection, including yield (Liu et al. 2018), development (Cui et al. 2020; Zhang et al. 2019), resistance (Beyene et al. 2015; Cao et al. 2017; Vivek et al. 2017) and grain quality (Guo et al. 2020). Compared with the widely used marker-assisted selection (MAS) technology, the GS strategy is advantageous in that it does not need to identify significant markers, genes, or quantitative trait loci (QTL) related to traits in advance, and can be used to predict the performance of populations without genetic structure analysis (Bernardo 2016). Moreover, GS depends on all genes/loci affecting phenotypes, including major genes/loci and minor genes/loci (Crossa et al. 2017; Desta and Ortiz 2014; Xiao et al. 2017), thus it is suitable for the selection of complex quantitative traits.

Genomic prediction can be based on any method that can capture the association between the genotypic data and associated phenotypes (or breeding values) of a training population. Over the past two decades, several different statistical models and machine learning methods have been proposed for GP. Among the parametric models, ridge regression best linear unbiased predictions (RR-BLUP) and genomic-BLUP (GBLUP) are most commonly used in plant breeding that assumes a normal distribution of SNP effects (Heslot et al. 2012). Over the past two decades, several different statistical models and machine learning methods have been proposed for GS. Among the parametric models, ridge regression best linear unbiased predictions (RR-BLUP) and genomic-BLUP (GBLUP) are most commonly used in plant breeding that assumes a normal distribution of SNP effects (Heslot et al. 2012). Bayes A assumes a prior distribution of effects with a higher probability of moderate to large effects, while Bayes B and Bayes C model assume some marker effects to be zero. Reproducing Kernel Hilbert Spaces (RKHS) regression is equivalent to a GBLUP with a non-linear kernel (de los Campos et al. 2009). Deep Learning (DL) is a type of machine learning (ML)
approach that is a subfield of artificial intelligence, which method is a nonparametric models (Montesinos-López et al. 2021). In most traits, there is no significant difference between different models. RR-BLUP is a classical method with lower computational requirements and usually performs well (Kwong et al. 2017; Maulana et al. 2021; Resende et al. 2012).

Prediction accuracy of GS, calculated based on the correlation between the predicted phenotypic value (GEBV) and the observed phenotypic value, is used to evaluate the reliability of GS results (Combs and Bernardo 2013). A variety of factors are known to influence GS prediction accuracy, which include training population size, relatedness between training and test individuals, DNA marker type, quality and density, trait heritability, statistical models, Linkage Disequilibrium (LD) and population structure (Liu et al. 2018; Norman et al. 2018). The quality and density of genomic markers are influenced by genotyping platforms. The ideal sequencing platform should be inexpensive, provide high-throughput and genomic coverage, as well as be replicable and sensitive. Genotyping-by-sequencing (GBS), which can simplify complex genomes, has now been used as a high-throughput and cost-effective tool for screening molecular markers in many crop species (Elshire et al. 2011). The Diversity Arrays Technology (DArT) is a DNA hybridization-based molecular marker technique that can detect variation at numerous genomic loci without simultaneous sequence information. DArT marker has been used in maize molecular assisted selection (Dos Santos et al. 2016). In addition, DArT can be combined with next-generation sequencing platforms known as DArT-seq, which permits simultaneous detection of several thousand DNA polymorphisms by scoring the presence or absence of DNA fragments in genomic representations generated from genomic DNA through a process of complexity reduction (Kilian et al. 2012); however further studies must be performed to confirm its reliability.

The process from maize seed germination to seedling establishment involves a variety of complicated metabolic transformations, physiological activities, and cellular and tissue differentiation, which are regulated by many genomic loci, including major and micro effect loci under a drought environment. An abundant variation of these loci takes place among different maize germplasms. In this study, we used the combined genotyping data of natural maize population obtained by DArT and GBS sequencing platforms, as well as the phenotypic
data of emergence rate (ER), seedling plant height (SPH) for seedling stage, and grain yield (GY) for the mature stage under drought environment, to conduct GS of drought resistance. The main objectives were to: (1) assess the phenotypic variation in the NCCP under the natural drought environment; (2) estimate the genomic prediction accuracy for ER, SPH, and GY traits using DArT and GBS markers; (3) try to improve the prediction accuracies by increasing the marker quality; and (4) control population structure for improve the prediction accuracies.

2. Materials and Methods

2.1 Plant materials and phenotypic data evaluation

A collection of 391 inbred lines for genomic prediction analyses in the current study. All these inbred lines were selected from China, Mexico, America, etc., to adjust to Northeast China, designated as the northeast China core population (NCCP) in the Fuxin Mongolian autonomous county, Liaoning Province, China (42° 06′ N, 122° 55′ E) in 2019 and 2020, respectively, where drought stress occurs frequently during spring season. Some germplasm unsuitable for the local environment has been removed. The trial was planted on 12 May 2019 and harvested on 7 October 2019 and planted again on 11 May 2020 and harvested on 9 October 2020. In the 2019 maize growing season, the average temperature was 21 °C and the average rainfall was 4 mm. In the 2020 maize growing season, the average temperature was 21°C and the average rainfall was 3 mm. During the maize seedling stage, the highest temperature across two years was 35°C, and the precipitation was less than 20 mm, causing serious drought stress in the seedling stage, where three leaves were maintained in the whole growing season (Fig.1). Trials in our research did not take irrigation measures, but the seeds were planted before precipitation according to the weather forecast for germination. A completely randomized block design with three replications was applied in each evaluated environment. The inbred lines were sown using a seed for each hole in a single plot, 3 m long, with plant spacing of 10 cm and row spacing of 60 cm. Target traits of emergence rate (ER), seedling plant height (SPH) and grain yield (GY) were evaluated to represent the drought
tolerance of the tested plant materials. The target traits of ER and SPH traits were measured at
the seedling stage, i.e. 20 days after planting. While GY was measured at the maturity stage,
the moisture content readings were made by LDS-1G Grain Moisture Meter. ER was
measured as the proportion of surviving plant number to number planted. SPH was measured
as the distance from the plant base to the highest of the seedling plant. The average dry weight
of five plants was regarded as the GY score for entries.

2.2 Phenotypic data analysis and heritability estimation

The best linear unbiased estimate (BLUE) values and broad-sense heritability ($H^2$) of ER,
SPH and GY were calculated within and across years using the META-R software version
6.04 (Alvarado et al. 2020) (http://hdl.handle.net/11529/10201). The linear mixed models used in
META-R are implemented in the LME4 R-package, functions of lmer and REML were used
to estimate the variance components.

$$Y_{ijk} = \mu + \text{Gen}_i + \text{Env}_j + \text{Gen}_i \times \text{Env}_j + \text{Rep}_k + \epsilon_{ijk}$$

Where $Y_{ijk}$ is the trait of interest, $\mu$ is the overall mean, $\text{Gen}_i$, $\text{Env}_j$, and $\text{Gen}_i \times \text{Env}_j$
are the effects of the $i$-th genotype, $j$-th year, and $i$-th genotype by $j$-th year interaction,
respectively. $\text{Rep}_k$ is the effect of $k$-th replication. $\epsilon_{ijk}$ is the residual effect of the $i$-th
genotype, $j$-th year, $k$-th replication. Genotype is considered as the fixed effect, whereas all
other terms are declared as the random effects. Years with heritability below 0.05 were
excluded from the across location analysis.

Broad-sense heritability ($H^2$) based on the entry means within trials was estimated as follows:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{n\text{Envs}} + \frac{\sigma^2}{n\text{Envs} \times n\text{reps}}}$$

Where $\sigma_g^2$, $\sigma^2$, and $\sigma_{ge}^2$ are the genotypic variance, error variance, and
genotype-by-environment interaction variance, respectively, and $n\text{reps}$ and $n\text{Envs}$ are the
numbers of replications and environments, respectively (Carena et al. 2010).

2.3 Genotyping and quality control

Leaf samples of three individual plants were collected for each inbred line during the seeding
stage for DNA extraction with a CTAB procedure, and genotyping was carried out by DArT (https://www.diversityarrays.com/) and GBS (Elshire et al. 2011) platform.

Genotyping of all inbred lines with the GBS method was performed at Wuhan medical laboratory, BGI Co., Ltd, followed a protocol widely used by the maize research community (Elshire et al. 2011). The genomic DNA was digested with the ApeKI restriction enzyme, and a DNA library was constructed in 96-plex and sequenced on Illumina-hiseq XTen sequencing system. Each sample obtained an average of about seven hundred and fifty thousand reads. The B73 RefGen_v4 was the reference genome and applied BWA software to match all sequencing data to the reference genome. Details in single nucleotide polymorphism (SNP) calling and imputation have been previously described (Cao et al. 2017). Each line in our population has an average of approximately 67,000 SNPs spread across different chromosomes. In total, 768,558 loci were aggregated to cover all SNPs, and 760,831 of them were mapped to chromosomes 1-10, while 7,727 could not be anchored to any of the chromosomes (Table S1).

Genotyping of all the inbred lines with DArT-seq method was carried out at the SAGA (https://seedsofdiscovery.org/about/genotyping-platform/) sequencing lab, jointly established by DArT company and CIMMYT. Two enzymes of PstI (CTGCAG) and HpaII (CCGG) were used to digest the DNA samples to reduce the complexity of the genome. For each 96 well plate, 16% of the samples were replicated to assess reproducibility (Pereira et al. 2020). After enzyme digestion, DNA from different samples was linked with barcodes of different base combinations and sequenced to construct a DNA simplified sequencing library. Equimolar amounts of amplification products from each sample were pooled by plate and amplified by c-Bot (Illumina) bridge PCR, followed by fragment sequencing on Illumina Hiseq 2500 (www.illumina.com). The short fragment sequencing technology (150bp) and the simplified sequencing DNA library of mixed samples was sequenced on a single lane (Kilian et al. 2012) (https://www.diversityarrays.com/). SNPs were called using the DArTsoft analytical pipeline (http://www.diversityarrays.com/software.html) (Chen et al. 2016). DArT's SNP development is different from GBS, which does not rely on the reference genome information, but mainly depends on the DArT company's sequencing the sample library data. All reads were aligned by sequence analysis based on maize tags of metagenome representation. Two types of data
were generated by DArT-seq: SilicoDArT and SNP. The former is the presence/absence variation of the tag sequences, while the last one, SNPs, have been widely used in genetics and molecular marker-assisted selection breeding fields. The type of SNP data was obtained by B73 RefGen_v4 using the BLAST tool, and some of the fragments in SilicoDArT cannot be aligned to the chromosomes and abandoned. For this reason, the number of SilicoDArT is more than SNP. SilicoDArT of 62,794 were acquired, and 39,659 SNPs were located on maize chromosomes 1-10; however, 547 SNPs were anchored on the zero chromosome (Table S1).

In both DArT and GBS datasets, TASSEL 5 (Bradbury et al. 2007) was used to filter out markers with minor allele frequency (MAF) < 0.05 and a missing rate > 20%. The samples with a missing rate of more than 20% were rejected. The number of lines in DArT and GBS is 379 and 378, respectively. In total, the markers are 11,865 SilicoDArT and 7,837 SNPs in the DArT platform and 91,003 SNPs in the GBS dataset (Table 1). Imputation was performed using the method of Cao et al. (2017).

2.4 SNP distribution across the whole maize genome

The distribution of SNPs in the genome is one of the indicators of evaluation marking availability. A package named Rideogram (Hao et al. 2020), written in R programming (Team 2021), was used to assess SNPs in the DArT-seq and GBS platform. It is worth noting that SilicoDArT does not implement this analysis because the location information on the genome is not available. The base position information of the reference genome, RefGen_v4, was obtained on the Gramene Website (https://ensembl.gramene.org/Zea_mays/Info/Index). The SNP distributed plot was drawn using ideogram function. We provided the integrated script to plot a heatmap of the distribution of the SNP on maize chromosomes based on HMP format at the URL of https://aozhangchina.github.io/R/chromosomeheatmapTool/ChromosomeHeatmap.html.

2.5 Genomic prediction model

The RR-BLUP model assumes homogenous variance of all markers and shrinks all marker
effects equally to zero. RR-BLUP is equivalent to BLUP and uses the realized relationship matrix estimated from the markers. The genomic prediction analysis was performed with the ridge regression BLUP (rrBLUP) package (Endelman, 2011). The mixed model is described as:

\[
y = X\beta + Zu + \epsilon
\]

where \(y\) is the vector \((n \times 1)\) of observations, \(X\) is the vector \((n \times 1)\) of individuals and \(\beta\) is the fixed effects; \(\epsilon\) is the vector \((n \times 1)\) of independently random residuals with assumed distribution \(N(0, I\sigma^2_\epsilon)\), \(Z\) is the design matrix \((n \times m)\) for random effects and \(u\) is the vector of random effects with \(u \sim N(0, K\sigma^2_u)\) and \(K\) being an identity matrix in this case. In addition, \(n\) is the number of individuals, and \(m\) is the number of markers (Endelman 2011; Liu et al. 2018). Variance components are estimated by REML using the spectral decomposition algorithm of Kang et al. (2008). The \texttt{mixed.solve} function, a linear mixed-model equation estimator, estimates marker effects and GEBVs. GEBVs are derived from the realized (additive) relationship matrix of individuals calculated from marker genotypes (Tecle et al. 2014).

\[2.6\] Effect of Training Population Size (TPS) on Genomic Prediction Accuracy Estimation

To evaluate the impact of TPS on the accuracy of genomic prediction, the random cross-validation scheme was used to randomly generate different training and prediction sets and assess the prediction accuracy within each TPS. The inbred lines ranged from 10% to 90%, with an interval of 10%, were extracted from the total lines and were used to establish the training population to predict the remaining inbred lines. The genomic prediction was repeated 100 times for each TPS. For the same trait, the prediction accuracies estimated from the DArT and GBS datasets were compared.

\[2.7\] Effect of Marker Quality on Genomic Prediction Accuracy Estimation

Compared with the normal growth conditions, genomic prediction in drought environments are more complicated. Higher marker quality has been proven to help improve prediction accuracy. We try to enhance the quality of markers to improve the prediction accuracy under
drought environments. Different levels of MAF and missing rates were filtered for the GBS/DArT markers and controlled the marker quality and quantity. A combination of MAF and missing rate was used to filter markers, i.e., the sets were filtered by MAF from 0.10 to 0.40 with an interval of 0.10, and the missing rate from 0 to 60% with an interval of 20%. A 2-fold cross-validation used to compare the prediction accuracy of different quality levels, and the analysis was repeated 100 times for each marker quality.

For evaluating the effect of linkage disequilibrium (LD) among markers on the estimation of the genomic prediction accuracy, non-independent SNPs caused by strong LD were excluded in the GBS/DArT dataset. TASSEL 5 (Bradbury et al. 2007) were used to perform LD analysis, and LD decay plots were developed using the LD decay Plot Tool written by Zhang Ao based on the base functions using the R language. The plotting script is available online on the webpage [https://aozhangchina.github.io/R/LDdecay/LDdecayPlotTool.html](https://aozhangchina.github.io/R/LDdecay/LDdecayPlotTool.html). SNP sets in GBS/DArT dataset were filtered by R language, and the SNPs were removed when $r^2$ was greater than or equal to specific threshold. Genomic prediction accuracy across 100 times repetition was calculated using 2-fold cross-validation.

### 2.8 Effect of Genetic Structure on Genomic Prediction Accuracy Estimation

In our previous research, the relationship between the training and testing population significantly impacts prediction accuracy (Zhang et al. 2017a). NCCP consists of multiple heterotic groups in the current study, such as the Reid, Reid_Chenese, Lancaster, SS, NSS, Tangsipingtou, Lv, Huang, Longdan Jidan, CML, Mixed, etc. Some elite lines of Northeast China were also included in the population, including Shen118, Shen3336, Liao6049, Dan340, Ji818, etc.

A phylogenetic tree is used to demonstrate the genetic relationship among all the inbred lines used in the present study. The genetic distances among all the inbred lines were calculated in Bio-R (Biodiversity Analysis with R) software developed by CIMMYT using the filtered DArT and GBS datasets (https://hdl.handle.net/11529/10820). The formula of the modified Rogers distance (MRD) used in the present study is as follows:
\[ MRD_{xy} = \sqrt{\frac{\sum_{l=1}^{L} \sum_{a=1}^{n_l} (p_{lax} - p_{lay})^2}{2L}} \]

Where \( p_{lax} \) is the estimated frequency of allele \( a \), which is located in locus \( l \), and genotype \( x \); \( L \) is the number of loci, \( n_l \) is the number of alleles in the locus \( l \). Markers from both platforms were used to create the phylogenetic tree. Based on the genetic distance matrix, the Neighbor-Joining (NJ) method of MEGA X and Fig Tree software was used to create the phylogenetic tree's construction (Saitou and Nei 1987).

We performed a principal component analysis (PCA) to assess the prediction accuracy affected by genetic structures. All the inbred lines used in the present study were divided into two subgroups because it is necessary to ensure that the subsets have enough material for genomic prediction. The \texttt{prcomp} function from the stats package and \texttt{plot} function from the base package were executed to draw PCA plots under the R environment. Genomic prediction accuracy across 100 times repetition was estimated within each subgroup using 2-fold cross-validation.

### 2.9 Cross-validation for assessment

In this study, the size of the training population with the highest average prediction accuracy and the lowest standard deviation is about 50%, which is equivalent to 2-fold cross-validation. A 2-fold cross-validation analysis was implemented for all traits, i.e., 50% of individuals randomly selected for the training population and 50% in the testing population. Additionally, different training population sizes were used to evaluate the performance of population size. For prediction, regression models were evaluated by the prediction accuracy, a Pearson’s correlation between the GEBVs and the phenotypic data in the testing population (Thavamanikumar et al. 2015). All predictions are repeated 100 times, and the mean value is taken as the average prediction accuracy.
3. Results

3.1 Phenotypic data and correlation analysis

The phenotypic mean and standard deviation of the three traits in 2019, 2020 and BLUEs across two years are shown in Table 2. Genetic variants sufficient for genomic selection were found under the drought environment. Phenotypic mean of ER ranged from 0.16 to 1.00, with an average value of 0.81 in 2019; and ranged from 0.32 to 0.90, with an average value of 0.70 in 2020; and the BLUE, ranged from 0.60 to 0.84 with an average of 0.76 across two years.

For SPH, the phenotypic mean ranged from 9.14 to 20.64 cm, and had an average value of 14.96 cm in 2019; from 8.61 to 15.06 cm with 11.67 cm in 2020; and for BLUE, from 12.00 cm to 14.48 cm, with an average of 13.36 cm. The mean ranged from 12.22 to 93.96, 7.59 to 23.32, and 18.32 to 38.19 for GY trait in 2019, 2020 and BLUE, respectively. The average was 40, 10.97, 25.04 for three kinds of environments in the GY trait. The coefficient of variation (CV) shows a trend that 2020 is greater than 2019 in three traits (Table 2).

Correspondingly, the environmental variance component is a great proportion, especially for GY and SPH. The low to moderate heritabilities were gained for three traits except in 2019, while there are very high heritabilities in 2019. The $H^2$ are 0.88, 0.82, and 0.85 for ER, SPH, and GY in 2019, respectively; Correspondingly 0.65, 0.53, and 0.33 in 2020; and 0.32, 0.26 and 0.33 across two years (Table 2). The phenotypic data analysis based on BLUEs revealed that ER, SPH, and GY showed normal distribution or skewed normal distribution (Fig.2). All three traits showed significant positive correlations, and the correlation coefficients are 0.36, 0.35, and 0.25 ($P < 0.01$) between ER and SPH, ER and GY, and SPH and GY, respectively.

3.2 Quality control and marker distribution on genotypic data

Two kinds of marker sets were obtained in the current research. The SilicoDArT markers were from DArT-seq, and SNPs were from the GBS and DArT-seq. The plots of MAF and missing rate distribution before filtering for three marker sets are shown in Fig. 3. The SilicoDArT markers had an average MAF of 0.10, and SNPs' average MAF of 0.11 of DArT-seq at pre-filtering. For SNPs from GBS, the minor allele frequency mean before
filtering is 0.11. Before filtering, the highest peak is found at the MAF intervals of 0 to 0.05 for three marker sets. However, we are not sure whether the small MAF values are from the actual situation or sequencing errors, so they can not be used for further analysis. The MAF distribution of other intervals (0.05 to 0.5) is relatively uniform. Average minor allele frequency after filtering are 0.24, 0.24, and 0.23 for SilicoDArT, SNPs from DArT-seq, and SNPs from GBS, respectively (Table 1). Consistent MAFs were found on both genotyping platforms (Fig.3 a,b,c), potentially indicating that sequencing quality is reliable. It is worth noting that higher quality marker sets were used in the present study.

Results of the distribution of missing rates showed a decreasing trend in SilicoDArT and SNPs from DArT-seq markers, and the missing rate is mainly concentrated in the range of 0-10%. On the other hand, the highest peak, accounting for more than 40 percent, was found in the GBS markers for the missing rate, which appears at the interval of 90% to 100%. For each interval of 0 to 10%, 10% to 20%, 20% to 30%, 30% to 40%, and 40% to 50% in GBS markers, the missing ratio accounts for about 10%. There were almost zero accounts in the range of 50% to 90% (Fig.3 d,e,f). The irregular missing rates results of GBS markers were not apparent and thus may not be suitable for alignment to a single reference genome in the SNP calling process due to the diversity of materials.

High-quality marker sets were obtained after filtering had been done, and the average missing rates are 0.070, 0.073, and 0.068 for SilicoDArT, SNPs from DArT-seq, and GBS markers (Table 1). Heterozygosity is usually a critical index for marker filtering, especially for inbred lines. In this research, the proportion heterozygous of the three marker datasets was very low before and after filtering, which means our material is homozygous and implies the sequencing is accurate. Specifically, the averages of heterozygosities were less than 0.01 before filtering for the three marker sets. And after filtering, the heterozygosity rates were increased, but all were still less than 0.03 (Table S1 and Table 1).

When comparing pre-filtering and post-filtering, the number of markers decreased by 81%, 80%, and 88% for SilicoDArT and SNPs from DArT-seq and SNPs from GBS, respectively (Table S1 and Table 1). The missing rate is mainly responsible for the reduction of markers, and those were 0.25, 0.24, and 0.56 for the SilicoDArT, SNPs of DArT, and SNPs of GBS shown in Table S1. The high missing rate indicates that the diversity of the population
is rich, but a large number of markers will be lost when the marker set is summarized. The SNP marker sets from DArT-seq and GBS cover almost the entire maize genome (Fig.4). Generally, the density of markers at both ends of the chromosomes was high and decreased toward the centromere position. The SNP marker sets from DArT-seq and GBS platforms were enriched at both ends of chromosomes 5 and 8. In addition, markers from GBS had a more uniform distribution than ones in DArT, whereas DArT’s SNPs had higher density on the long arms of chromosomes 8 and 9. Uneven distribution may potentially lead to a decrease in prediction accuracy when implementing a genomic prediction.

3.3 The phylogenetic tree among NCCP

The results of the phylogenetic tree are shown in Fig.5 using SNPs of DArT-seq and GBS. SilicoSNP does not have a specific chromosome location information of each marker, so the analysis is not performed. The consanguinity or heterosis group of the NCCP were divided into 13 groups, including the NSS, SS, Huanglvxi, Lvxi, Reid, Mixed, France, PA, Jidan, Longdan, Lvdahonggu, Tangsipingtou and Unknown. The CML series from CIMMYT is a tropical material, and only two can be planted in northeast China, where they were classified into the Mixed. Some small groups and inbred lines from the multiple parent selection also were classified as unknown. DArT SNPs produced a better phylogenetic tree, and the markers more clearly separate the Reid, Lvxi, and Huanglvxi from NCCP. However, GBS markers did not clearly separate the materials of heterotic groups in the present study. Some of the materials from local breeders should belong to the mixed group or others, but they were unable to determine sources for a long time, so they were classified into the unknown group. Although some materials have source information, we are not sure that the information is correct when the materials are collected from third parties. Therefore, it isn't easy to obtain a perfect phylogenetic tree in practical breeding work. In the pedigree, we have conflicts with a phylogenetic tree based on markers, making it hard to decide which one we can believe. At least, the source of inbred lines of NCCP is rich based on the result of DArT markers. However, it may cause trouble for the genomic prediction.
3.4 The genomic prediction accuracy estimated from the DArT and GBS markers

Genomic prediction accuracies were estimated for all the three target traits with the BLUE values and the SNP datasets filtered with different parameters (Fig. 6). Using DArT SNP datasets, the prediction accuracy increased continuously as the TPS increased across all three traits evaluated under a natural drought stress environment (Fig. 6). The prediction accuracy slightly increased across all three traits when the TPS increased from 50% to 90%. The smallest standard error was observed in prediction accuracy when 50% to 60% of the training population size was used to predict the target traits of ER. For the target traits of SPH and GY, 40% of the training population size gained the smallest standard error, which indicated that 40 to 60% of the total genotypes assigned as the training set could achieve good prediction accuracy in the DArT SNP markers. For the GBS SNP markers, a similar trend was observed for GY. With the increase of the TPS, the average and median of prediction accuracy showed slightly different for the target traits of ER and SPH (Fig. 6). The prediction accuracies estimated in a single year are similar trend to using BLUEs across two years (Fig. S1, S2). The H2 of the three traits was the highest in 2019, and the highest prediction accuracy was obtained, but the improvement was very limited. SilicoDArT markers also performed this analysis and got a consistent trend with the DArT SNP markers (Fig. S3).

According to the above results, 50% of the training population size was selected for further analysis.

Genomic prediction accuracies estimated using 2-fold cross-validation (50% TP) for all three traits are shown in Fig. 7. The prediction accuracy between genotypic platforms has reached a significant level for three interesting traits. In contrast, the two types of the marker within the DArT-seq significantly differ in the SPH trait, which may mean the marker's alignment to the reference genome results in a loss to the prediction for some traits. For three traits, the prediction accuracy estimated from the DArT-seq markers was significantly higher than markers from the GBS (P < 0.01). Among the three traits, the prediction accuracy of SPH was the lowest in both the DArT and GBS markers, which consisted of the lowest heritability of SPH. The average prediction accuracies using the SNPs from the DArT-seq
were 0.27, 0.19, and 0.33 for ER, SPH, and GY traits, respectively. The result of SilicoDArT is close to the SNPs from DArT-seq, which were 0.26, 0.22, and 0.33. While, the prediction accuracies estimated using the GBS markers were -0.02, -0.06, and 0.20 for the traits of ER, SPH, and GY, respectively.

3.5 Different prediction accuracy of DArT/GBS marker quality

The result of prediction accuracy estimated in all the traits under three marker sets filtered with combinations of MAF and missing rate was presented in Table 3, Table S2 and Table S3. For the ER trait, the average prediction accuracies range from 0.19 to 0.29, and the highest accuracy corresponds to 2,762 markers in SilicoDArT. Interestingly, higher strict of missing values increases the prediction accuracy, while MAF had little influence on the prediction accuracy. This trend has also been found in ER and SPH traits using three markers; for GY traits, the trend weakens. In most cases, the prediction accuracy is not significantly improved when using more stringent filter markers except SPH. SPH is the only trait that responds to high marker quality among the ER, SPH and GY traits using three marker sets.

The LD decay distances of DArT SNP and GBS SNP are 6.16Kb and 3.83Kb, respectively (Fig.S4). The result of prediction accuracy estimated in all the traits under the different DArT/GBS markers filtered by LD is presented in Table 4. SilicoDArT did not perform this filter because the location information on each chromosome was missing. Different LD filtering criteria correspond to different marker densities, and with the strictness of filtering, the number of markers dropped dramatically. SPH obtained higher average prediction accuracy with the original number of markers of about 1/10 using SNPs from DArT-seq. For GBS markers, the traits of ER and GY were filtered by LD0.1, and the prediction accuracy of 2,333 SNPs was higher than that of 91,003 SNPs before strictly filtering. A small amount of strict markers improved the prediction accuracy on both platforms, but the traits were not the same. The filtering criteria of marker quality may vary due to the differences of genotyping platforms, which need to be further explored.
3.6 Prediction accuracy estimated within subgroups

In general, the population with close relationships is more likely to obtain high prediction accuracy. We want to control the population structure and improve the prediction. Due to the NCCP not having too many lines, it should be divided into two subgroups, accounting for both population structure and population size. The PCA plots show the distribution of two subgroups using PC1 and PC2 in the DArT and GBS markers, and the first two PC explained 35% and 31% of the genetic variation of DArT and GBS markers, respectively (Fig.8). The PCA results of the two platforms are very similar. Finally, we chose 180 inbred lines divided into Subgroup 1 and 199 lines for Subgroup 2.

The prediction accuracies for all the traits estimated within subgroups are shown in Table 5. Using the DArT SNP markers, the prediction accuracies in Subgroup 1 with 50% of TPS were 0.17, 0.11, and 0.11 for ER, SPH, and GY traits, respectively. For Subgroup 2, the prediction accuracies were 0.29, 0.20, and 0.33 for ER, SPH and GY traits, respectively. The prediction accuracy of Subgroup 1 was significantly lower than when using the whole population for three traits, whereas Subgroup2’s prediction accuracies were higher than those using the entire population. The result of the GBS marker is different from that of DArT, and the variation of prediction accuracy is not apparent in the two subgroups.

4. Discussion

Drought tolerance, especially in the seedling period, is a complex and inherent trait of maize (Wang et al. 2016). The global climate has changed in recent years, resulting in drastic fluctuations in rainfall patterns and increasing temperature. Sudden climate changes can cause significant economic losses to countries worldwide. Signs of grain yield stagnation in maize, especially in drought-stressed and semi-arid regions, are evident. Fuxin County is a region providing a natural test site located in the main maize producing area of Northeast China that has suffered years of drought. In this region, we can deploy more large-scale experiments without using greenhouses. In the current study, the three traits have found enough phenotypic variation in our NCCP for ER, SPH, and GY traits. The results of correlation analysis show
that ER and SPH potential impact GY under a drought environment. The phylogenetic
tree shows our population has a rich genetic variation for improving drought resistance, although
the two phylogenetic trees are not the same.

GS is an effective breeding tool for improving complex traits in maize (Crossa et al.
2014). The GS can accelerate the genetic gain per unit time and unit cost by reducing the
selection cycle time and the phenotyping cost when the prediction accuracy is high. In the
present study, a natural population genotyped with GBS and DArT markers were used to
estimate the genomic prediction accuracy of ER, SPH, and GY under drought stress in maize.
Results indicated that the prediction is tough across genotyping platforms under a drought
environment. The average prediction accuracies of ER, SPH and GY in the panel estimated
using SNPs of DArT markers were 0.27, 0.19 and 0.33, respectively, SilicoDArT markers
were slightly higher than the DArT SNP; while the prediction accuracies with GBS markers
were -0.02, -0.06, and 0.20. Low prediction accuracy may be mainly affected by the quality of
markers and heritabilities. Our heritabilities and prediction accuracies for the GY trait was
within the range of the previous study that used 22 populations under the water stress
environments (Zhang et al. 2017a). Interestingly, unlike under normal conditions, the
heritabilities of the three traits were more than 0.8 in 2019, but the average prediction
accuracies were only slightly higher than that in 2020 and across two years. This means that
the prediction of drought resistance needs to accumulate more year data to improve the
heritability across years and stable phenotype.

In the same trait, the prediction accuracy estimated from the DArT markers was higher
than that estimated from the GBS marker dataset, and the difference was significantly. Both
DArT and GBS are cost efficient and high-throughput genotyping platforms that can be used
to implement GS. The GBS markers were implemented successfully in maize to improve
various traits with different levels of genetic complexity (Crossa et al. 2013; Wang et al.
2020). Comparatively, only a few studies were reported using DArT markers to perform GS
in wheat (Crossa et al. 2016; Liu et al. 2020). We spent a similar amount of money on two
genotyping platforms, about $30 per sample, to control the budget within an acceptable range
for breeders. This price can obtain DArT markers with normal quality, but a little bit lower
quality for GBS markers. That means the genome coverage from GBS has not reached 1×,
which is probably the reason why the prediction accuracies are not suitable when using GBS markers in this research. Our research indicates that DArT genotyping platform is more suitable than GBS when working with a lower budget. GS using DArT markers is also being implemented in CIMMYT maize breeding programs for improving grain yield and the other major agronomic traits.

This study showed that the average prediction accuracy increased as the TPS increased in all the traits in both genotyping platforms. In contrast, the standard error was first increased and then decreased (Fig. 6). Relatively high prediction accuracies with the slightest standard error were observed in all traits when 50% to 60% of the total genotypes were used as a training set. Results of this study are consistent with previous reports (Guo et al. 2020) that can be applied in the real breeding program. The marker density is an essential effect on prediction accuracy if population sets and heritabilities are the same. Our research found that the prediction accuracy of about 8,000 markers from DArT is significantly higher than that of 90,000 markers from GBS, which emphasized the importance of marking quality and indicated 8,000 markers is enough to implicate genomic prediction in maize. Low-density marker populations appeal to GS because they decrease multicollinearity and computational time consumption and allow more individuals to be genotyped for the same cost. However, this prediction was made in a breeding program's initial stage, and higher marker density could play a role in the following breeding cycle for the long term. As a mature genotyping platform, GBS markers have a superb performance on genomic prediction (Liu et al. 2021). The higher marker density of GBS also has a high application prospect in the multi-cycle breeding process.

We tried to increase the prediction accuracy by improving the marker quality and controlling the population structure, but at the same time, the marker density was also reduced. Our results show that the prediction accuracy improved in some cases (Table 3, 4, 5), even if the density of the marker was reduced. However, this strategy doesn't always work, and may be due to the fact that some complex traits require markers with specific effects. Also, there is a tradeoff between the number of markers and marker quality or population structure because marker quality decreases as the number of markers increase in a specific marker dataset. This study suggests that genomic prediction can also be performed using high-quality and
low-density markers in the initial breeding stage under the drought environment, especially when heritability is low.

The low-cost DArT platform increases the possibility of integrating genomic selection strategies into practical breeding programs for small breeding companies and the private sector. The SilicoDArT markers It is possible to capture lost effects using a single reference genome for a diversity-rich population, which is essential for challenging drought prediction.

The key to the genomic prediction of drought resistance is to improve heritability and stabilize the phenotype in maize. Natural drought experimental sites can increase the heritability of planting materials by accumulating data for many years without a greenhouse.

Fuxin county provides a chance to study genomic prediction for drought tolerance. The experimental design and model for drought resistance also need to be further developed.

References


Bernardo R (2016) Bandwagons I, too, have known. Theoretical and Applied Genetics 129:2323-2332


Combs E, Bernardo R (2013) Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. The Plant Genome 6


Cui ZH, Dong HX, Zhang A, Ruan YY, He Y, Zhang ZW (2020) Assessment of the Potential for Genomic Selection To Improve Husk Traits in Maize. G3:Genes Genomes Genetics 10:g3.401600.402020


Team R (2021) R: A language and environment for statistical computing


Vivek BS, Krishna GK, Vengadessan V, Babu R, Zaidi PH, Kha LQ, Mandal SS, Grudloyma P,


Statements & Declarations

Funding

This work was supported by Shanghai Agriculture Applied Technology Development Program, China (Z20190101), the National Science Foundation for Young Scientists of China (31801442), the CIMMYT-China Specialty Maize Research Center (KF201802), the Natural Sciences Foundation of Liaoning Provincial Department of Education (LJKZ0657).

Author's contributions

AZ, SY, and ZC collected materials; AZ, SC, DD analyzed data; AZ, SC, CL, JN, XD, YG, YZ and JF carried out the field experiments; HZ and YR designed the study; SC, AZ and YL wrote the manuscript; XZ, JC, LZ, YR, XD and JF revised the manuscript. All authors have read and approved the final version of the manuscript.

Conflict of interest

The authors have no conflicts of interest to declare that are relevant to the content of this
Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.
**Figures**

**Figure 1**

Distribution of temperature and rainfall in maize planting plots of Fuxin Mongolian Autonomous County in 2019 and 2020. The horizontal axis shows the date from sowing to harvest. The panels of a and c are Temperature (°C) information, including maximum temperature, minimum temperature, and average temperature. The panels of b and d are rainfall information, and the unit is mm.
Figure 2

Frequency distributions and correlations of BLUE (best unbiased linear estimator) as phenotype values were calculated from the maize seedling emergence rate (ER), seedling plant height (SPH), and grain yield (GY) of maize measured in 19FX (2019 Fuxin) and 20FX (2020 Fuxin) drought environment. The plots on the diagonal represent the phenotypic distribution frequency of ER, SPH, GY. The values above the diagonal line are the Pearson's correlation coefficients between every two traits. The values below the
diagonal line are scattered plots for every two traits. *Represents a significant difference at the 0.05 level; **Represents a significant difference at the 0.01 level.

**Figure 3**

Distribution of MAF and missing rate before filtering in three marker datasets: (a) MAF distribution of the SilicoDArT marker dataset; (b) MAF distribution of the DA'T marker dataset; (c) MAF distribution of the GBS marker dataset; (d) missing rate distribution of the SilicoDArT marker dataset (e) missing rate distribution of the DArT marker dataset; (f) missing rate distribution of the GBS marker dataset.

**Figure 4**

The markers density thermogram of DArT and GBS datasets. (a) the markers density thermogram of DArT's SNP; (b) the markers density thermogram of GBS.

**Figure 5**

Phylogenetic Trees for SNP markers from DArT-seq and GBS platforms. The letters a and b represent NCCP's phylogenetic trees from DArT's SNPs and GBS's SNPs. The hex RGB color values of #F44336, #8FCE00, #FFFFFF, #2A329F, #2986CC, #CE7E00, #D3D3FF, #A3C4C9, #EA9999, #F9CB9C, #B6D7A8, #744700, and #EEEEEE represent the hybrid groups of NSS, SS, Huanglvxi, Lvxi, Reid, Mixed, France, PA, Jidan, Longdan, Lvdahonggu, Tangsipingtou, and Unkonwn, respectively.

**Figure 6**

Genomic prediction accuracies of ER, SPH, and GY in NCCP across two natural drought conditions, when the training population size was set from 10 to 90% of total genotypes, with an interval of 10%. Panel (a) ER estimated with DArT markers; (b) ER estimated with GBS markers; (c) SPH estimated with DArT markers; (d) SPH estimated with GBS markers; (e) GY estimated with DArT markers; (f) GY estimated with GBS markers.
Figure 7

Genomic prediction accuracies of ER, SPH, and GY estimated with the SilicoDArT, DArT and GBS marker datasets under 2-fold cross-validation. The significant difference was analyzed by one-way ANOVA of SPSS.

Figure 8

The first two principal components of DArT and GBS datasets. (a) the first two principal components of DArT. (b) the first two principal components of GBS. Subgroup 1 is marked with black and subgroup 2 is marked with red.

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

- SupplementaryFiles.docx