

國立臺灣大學生物資源暨農學院農藝系研究所



碩士論文

Graduate Institute of Agronomy

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

利用基因體預測評估雜交組合的表現

Hybrid performance evaluation in plant breeding via
genomic prediction

王珮仙

Pei-Hsien Wang

指導教授：廖振鐸 博士

Chen-Tuo Liao Ph.D

中華民國 111 年 7 月

July, 2022

摘要



基因體預測 (Genomic prediction) 可有效降低育種成本及縮短所需時間，因此在作物育種中，已成為一項評估子代雜交表現強而有力的工具。本次研究中共使用兩筆作物資料，分別為具有 142 個品系的 C. Maxima 南瓜資料和 24 個品系的玉米資料。本研究提出一個同時考慮加性及顯性效應的混合線性模型，以預測雜交後代組合的表現。我們先使用有限制最大概似(restricted maximum likelihood, REML)估計法，來估計出加性效應和顯性效應的變方成份 (variance components)，再利用 Henderdon's 方程式獲得做為訓練集資料之部分雜交後代個體的加性效應和顯性效應，最後結合基因體關聯性矩陣(genomic relationship matrix)，利用基因體最佳線性不偏預測模型(genomic best linear unbiased prediction model, GBLUP model)預測雜交後代表現的育種價(GEBVs)，並進行優良品種的基因組選拔(GS)。而利用育種價得到雜交後代的特殊組合力(SCA)及其親本的一般組合力(GCA)，則可以用來計算雜交優勢(Midparent heterosis, MPH)以及優於親本表現的雜交優勢(Better-parent heterosis, BPH)。根據我們的研究結果，發現在玉米資料中 Mo17, NC350, B73, B97 和 OH7B 為較具潛力的親本，而 P026, P227, P236, P028 和 P235 則為南瓜資料中較具潛力的親本。

關鍵字：基因組預測、基因體最佳線性不偏預測模型、育種價、特殊組合力、一般組合力

Abstract



Genomic prediction has become an increasingly popular tool for hybrid performance evaluation in plant breeding mainly because it can reduce cost and accelerate a breeding program. We used two different crop data sets, one is the pumpkin (*C. Maxima*) data set consisting of 142 parental lines with 4521 filtered single nucleotide polymorphism (SNP) markers, and the other is the maize data set consisting of 24 parental lines with 46,134 filtered SNP markers. In this study, we propose a systematic procedure to predict hybrid performance using a linear mixed effects model that takes both additive and dominance marker effects into account. We first estimated the variance components of additive and dominance effects through restricted maximum likelihood estimation (REML), and used Henderdon's equation to obtain the values of additive and dominance effects of hybrid lines which were used to build training data sets. Finally, we predict genomic estimated breeding values (GEBVs) for individual hybrid combinations and their parental lines through the genomic relationship matrix. The GEBV-based specific combining ability (SCA) for each hybrid and general combining ability (GCA) for its parental lines were then derived to quantify the degree of midparent heterosis (MPH) or better-parent heterosis (BPH) of the hybrid. According to our result, Mo17, NC350, B73, B97 and OH7B are the most potential parental lines in the maize data set; and P026, P227, P236, P028 and P235 are the most potential parental lines in the pumpkin data set.

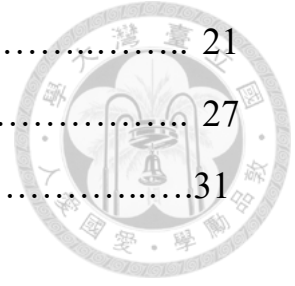
Keywords: Genomic prediction, genomic best linear unbiased prediction model, genomic estimated breeding values, specific combining ability, general combining ability.

Contents



口試委員會審定書.....	i
摘要.....	ii
Abstract.....	iii
List of Tables.....	vi
List of Figures.....	vii
1. Introduction	1
2. Materials and Methods	4
2.1 Pumpkin data set.....	4
2.1.1 Phenotype data.....	4
2.1.2 Genotype data.....	5
2.2 Maize data set.....	6
2.2.1 Phenotype data.....	6
2.2.2 Genotype data.....	6
2.3 Statistical models.....	9
2.4 Estimation for marker effects.....	10
2.5 Prediction for GEBVs.....	11
3. Result	14
3.1 Pumpkin data analysis.....	14
3.1.1 Prediction of potential hybrids and parental lines.....	14
3.1.2 Variance components and heritability.....	17
3.2 Maize data analysis.....	18
3.2.1 Prediction of potential hybrids and parental lines.....	18
3.2.2 Variance components and heritability.....	20

4. Discussion..... 21
References..... 27
Appendix - Rcode..... 31



List of Tables

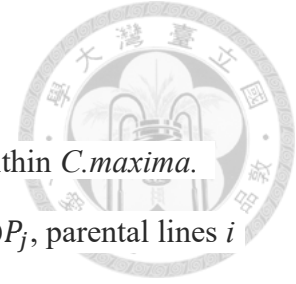


Table 1: The top 25 potential hybrids with GEBVs for trait FWT within *C.maxima*.

Note that $GEBV_{ij}$ represents the GEBVs for hybrid $P_i \otimes P_j$, parental lines i and j , respectively.

Table 2: The top 10 superior parental lines with GCAs for FWT within *C.maxima*.

Table 3: The estimations of the variance components and heritability using REML for trait FWT of *C.maxima*

Table 4. The top 25 potential hybrids with GEBVs for trait grain yield within the maize population. Note that $GEBV_{ij}$ represents the GEBVs for hybrid $P_i \otimes P_j$, parental lines i and j , respectively.

Table 5. The top 10 superior parental lines with GCAs for trait grain yield within the maize population.

Table 6. The estimations of the variance components and heritability using REML for trait grain yield within the maize population.

Table 7. The list of top 25 potential hybrids with GEBVs for trait FWT within *C.maxima* in our study and in Wu et al. The colored lines are both selected in our study and Wu et al.

Table 8. The top 10 superior parental lines with GCAs for FWT within *C.maxima* in our study and in Wu et al. The colored lines are both selected in our study and Wu et al.

List of Figures



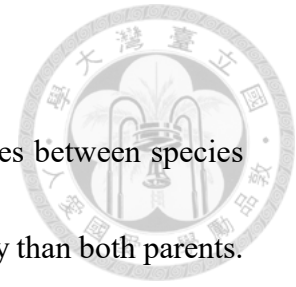
Figure 1: The LD decay results for the ten chromosomes.

Figure 2: the scatter plot of intra-crossing hybrid lines of *C.maxima* for the FWT

values and the GEBVs of the FWT. The colored points indicate those selected hybrids with the highest GEBVs. Note that there are only 2 selected hybrids in the figure because the training set only had 119 hybrid lines.

Figure 3: (a) the scatter plot of inter-crossing group of TM \otimes TS for the grain yield values and the GEBVs of the grain yield. (b) the scatter plot of intra-crossing group of TM \otimes TM for the grain yield values and the GEBVs of the grain yield. (c) the scatter plot of intra-crossing group of TS \otimes TS for the grain yield values and the GEBVs of the grain yield. The colored points indicate those selected hybrids with the highest GEBVs.

Introduction



Heterosis refers to the phenomenon that diverse offspring of crosses between species exhibit greater biological quality, speed of development, and fertility than both parents.

Hybrid breeding is a potential method that employs heterosis to promote genetic divergence among parents, to optimize the exploitation of heterosis and hybrid performance, and to boost yield stability or fruit quality. For example, United States annual corn production had increased from 2 billion bushels to 11.8 billion bushels on 22% less surface area planted since hybrid corn became available for 2000 to 2006 [\(Troyer, 2006\)](#).

Diallel method has been extensively used to evaluate the combining ability of parents in hybrids to understand the expression of quantitative traits and to predict the hybrid performances [\(Miller et al., 1980; Kadkol et al., 1984; Sherrif et al., 1985\)](#). Griffing (1956) proposed a method by separating the total genetic variation into general combining ability (GCA) of the parental lines and specific combining ability (SCA) of the hybrid combinations to analyze diallel cross. The GCA is a measure of additive gene activity that relates to the average performance of a particular inbred in a series of hybrid combinations, whereas the SCA is the performance of a parent in reference to general combining ability that linked to the non-additive effects (dominance and epistatic effects), which is a key measure to produce superior hybrid combinations

[\(Sprague and Tatum, 1942; Ali et al., 2014\)](#). The mid-parent heterosis (MPH) is defined as the difference between the hybrid performance and the average of its parental lines (Falconer and Mackay, 1996), and best-parent heterosis (BPH) is defined as the hybrid performance superior to the higher or better parental line (Fonseca and Patterson, 1968).

However, the number of single-cross combinations can increase dramatically as the number of parental lines increase, it will be unrealistic and costly to have all possible hybrids in the field experiment. With the development of genotyping-by-sequencing (GBS) [\(Elshire et al., 2011\)](#), it is achievable to detect enough polymorphic markers covering the entire genome to explore within-species diversity, the most common markers are single nucleotide polymorphisms (SNPs). Due to the high-density SNP markers across an entire genome, genomic selection (GS) becomes a suitable method for plant breeding to reduce cost and accelerate breeding programs (Poland and Trevor, 2012).

The concept of GS is to utilize a training population with known both genotype and phenotype data to build a model that takes untested individuals with known genotypic data only to predict genomic estimated breeding values (GEBVs) [\(Jannink et al., 2010\)](#).

In general, considering both additive and dominance marker effects into a GS model for hybrid performance could provide sufficient prediction accuracy (Wu et al., 2019).

GS has been applied to predict hybrid performance for several crops, such as durum

wheat ([Haile et al., 2018](#)), tropical maize ([Atanda et al., 2021](#)), and barley ([Schmid and Thorwarth, 2014](#)).



The data sets we used in this study are the hybrids selection of pumpkin (*Cucurbita* spp, $2n = 40$) that belongs to *Cucurbitaceae* family (Wu et al., 2019) and the hybrids selection of maize (*Zea mays* L., $2n = 20$) that belongs to *Poaceae* family (Guo et al., 2019). Pumpkin is a major global economic crop in the *Cucurbita* genus, growing on a global scale of around 3 million hectares, and yielding 27.832 million tons in 2021 (<http://faostat.fao.org>). Maize is the most produced and third most consumed cereal crop in the world, after wheat and rice, and maize is also the main staple food crop of more than 300 million Africans, especially in Sub-Saharan Africa.

Genomic best linear unbiased prediction (gBLUP) is a method that uses genomic relationships to estimate the breeding values of an individual ([Clerk and Werf, 2013](#)), and it has been demonstrated in many research that gBLUP has more accurate breeding values than pedigree-based BLUP or has little difference between using gBLUP and the nonlinear models (Moser et al., 2009; VanRaden et al., 2009). In this study, we are going to use two datasets of pumpkin and maize. First of all, we built a gBLUP model for each dataset. Secondly, we used the trained model to predict the GEBVs for all possible hybrid combinations. Lastly, we estimated GCAs for all the parental lines, and SCAs for all the hybrid combinations. In practice, this procedure would provide

advantageous information for breeders to select potential hybrids and superior parental lines.



Materials and Methods

2.1 Pumpkin data set

2.1.1 Phenotype data

We used 119 pumpkin intra-crossing hybrid lines of *C.maxima* as our training data set. The collected phenotypic trait was fruit weight (FWT), which is a quantitative data measured in continuous scale for the hybrids (Wu et al., 2019). The phenotypic data are historical data from 1988 to 2016 and provided by Known-You Seed Co., Ltd. All the trials were conducted in southern Taiwan, so it can be treated as a single location experiment. Every hybrid had six to ten observations at each time points, and the average of them was used as the phenotypic observation for the hybrid of the year. Because the phenotypic values of every hybrid were observed for more than one year, the year effects of phenotypic values need to be removed. Wu. et al (2019) assumed that the year effects were random effects which followed a normal distribution $N(0, \sigma_T^2)$; then the variance component σ_T^2 was estimated by gathering all years sample variances of hybrids as follows:

$$\hat{\sigma}_T^2 = \frac{\sum_{i=1}^n (k_i - 1) s_i^2}{(k_i - 1)},$$

where k_i is the phenotypic values from 1988 to 2016 for hybrid i and s_i^2 is its sample variance. The resulted $\hat{\sigma}_T^2$ is 0.076 for the trait FWT of intra-crossing hybrid lines *C.maxima*; then the estimated year effects will be generated from $N(0, \hat{\sigma}_T^2)$.

2.1.2 Genotype data

The germplasm collection of the pumpkin set we used consisted of 320 parental lines, which were classified into three clusters according to PCA: *C.maxima* with 142 inbreeding lines, *C.pepo* with 60 inbreeding lines and *C.moschata* with 118 inbreeding lines (Wu et al., 2019). We only used the 142 inbreeding lines of *C.maxima* as our parental lines to carry on the following GS analysis. Single-cross hybrids of these inbreds were developed in a half diallel mating scheme, generating 10,011 *C.maxima* × *C.maxima* intra-crossing hybrids. The genomic data for the 324 parental lines were extracted 76,815 SNPs after SNP calling, only 61,179 SNP markers remained after filtered by missing rate ≥ 0.05 . Wu et al. (2019) further filtered the 61,179 SNP by $MAF < 0.05$ and LD blocks within each cluster and obtained 4,521 SNPs remaining for *C.maxima*, 6348 SNPs remaining for *C.moschata* and 8800 SNPs remaining for *C.pepo*. Only the 4,521 SNPs of *C.maxima* were used for building GS model, and the genomic data for 10,011 hybrids were inferred from SNPs data of the 142 inbreeding lines.



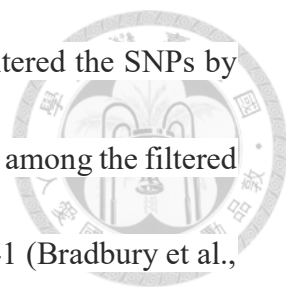
2.2 Maize data set

2.2.1 Phenotype data

Another crop data set we analyzed was maize, which consisted of 24 diverse parents (Flint-Garcia et al., 2005). The 276 single-cross hybrids were obtained at two locations (Columbia, MO and Clayton, NC) from 2005 to 2006, and the used phenotypic trait was grain yield (Mg/ha). The phenotypic observations for hybrids at the year was a single set of best linear unbiased predicted values after combining the data from two different locations (Guo et al., 2019). We chose 50 hybrids at random as our training set in the following analysis to build the GS model.

2.2.2 Genotype data

The 24 diverse parents which were extracted from the Maize HapMap V2 (Chia et al., 2012) at www.panzea.org were classified into two groups according to germplasm origin and PCA: the group of temperate and mixed inbreds (TM) with 11 inbred lines (B73, B97, Ky21, M162W, Mo17, MS71, Oh43, OH7B, M37W, Mo18W, and Tx303) and the group of tropical and sub-tropical inbreds (TS) with 13 inbred lines (CML52, CML69, CML103, CML228, CML247, CML277, CML322, CML333, Ki3, Ki11, NC350, NC358, and Tzi8) (Guo et al., 2019). There were 10,296,310 SNPs for 24 inbreeding lines, after filtered the SNP markers by missing rate ≥ 0.05 and MAF < 0.1 , there was 29,999 SNPs left. Missing genotypes were imputed with major allele. To



screen out the reliable SNPs for building GS models, we further filtered the SNPs by LD blocks. The LD parameter r^2 (the square Pearson's correlation) among the filtered SNPs for each chromosome was then estimated using TASSEL5.2.41 (Bradbury et al., 2007) with sliding window = 10. A smooth function between r^2 and physical distance (bp) was built using R function *loess.smooth*() with second-degree locally weighted polynomial regression. The LD decay of ten chromosomes is displayed in Figure 1. Filtering the 134,726 SNP markers by the resulted LD block sizes if r^2 gets close to 0.2, there were 46,134 SNPs left and were used to build the following GS model. The genotype data for 276 hybrids were generated from SNP data of the 24 parental inbreds. There were $C_2^{11} = 55$ hybrids within TM; $C_2^{13} = 78$ hybrids within TS; and $11 \times 13 = 143$ hybrids between TM and TS.

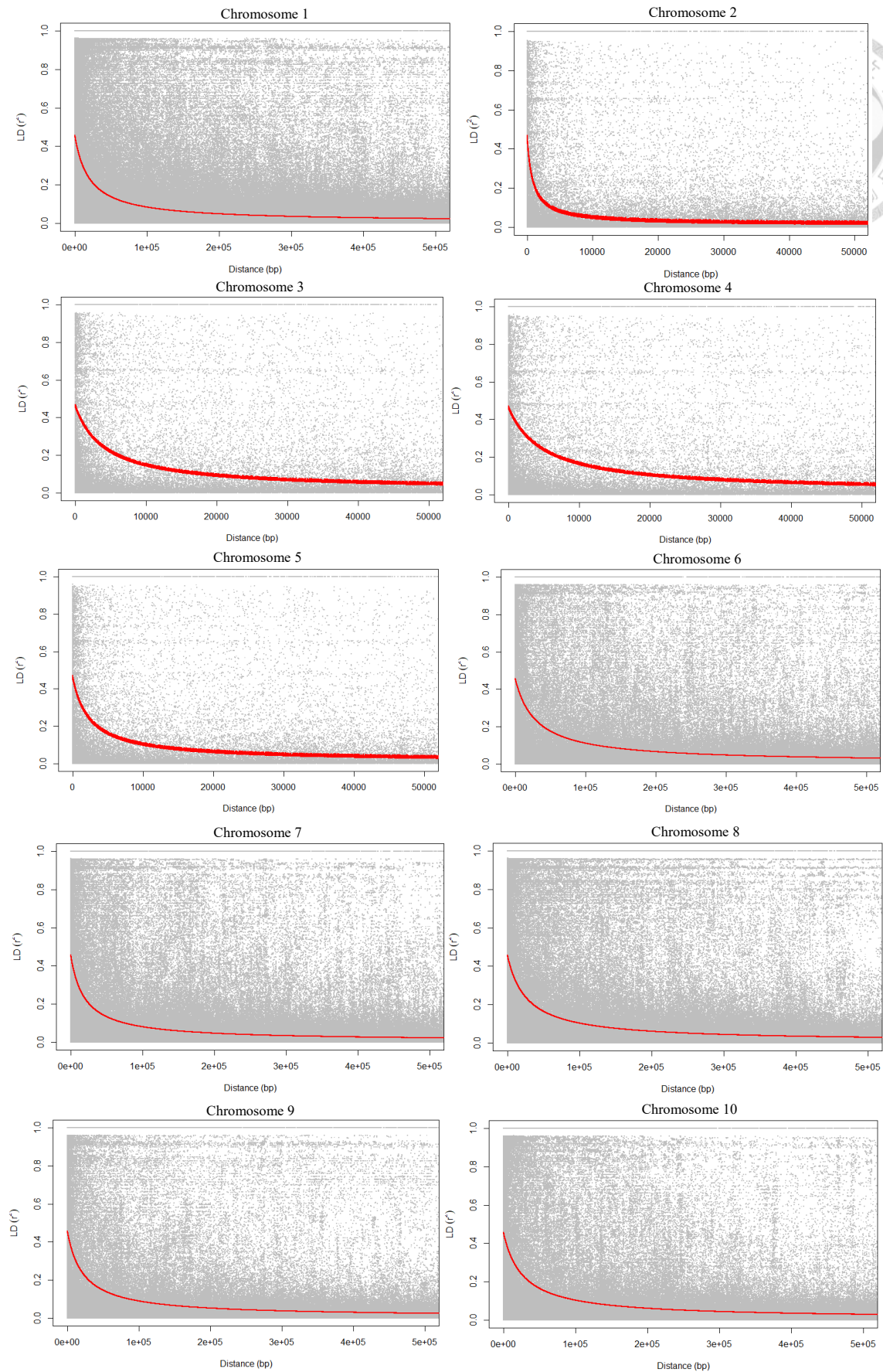
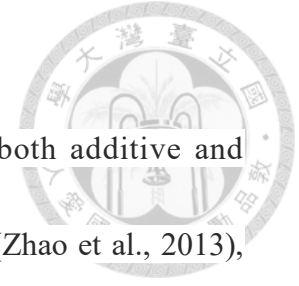


Figure 1: The LD decay results for the ten chromosomes.

2.3 Statistical models



We adopted the genome wide regression model considering both additive and dominance effects for hybrids performance prediction as follows (Zhao et al., 2013),

$$y_{ij} = \mu + \sum_{l=1}^p a_l^{(ij)} \beta_l^A + \sum_{l=1}^p d_l^{(ij)} \beta_l^D + e_{ij}, \quad (1)$$

where y_{ij} is the phenotype value of hybrid $P_i \otimes P_j$; μ is the constant term; $a_l^{(ij)}$ is coded as -1, 0 or 1 if AA, Aa or aa occurs at locus l for the hybrid; β_l^A is the additive effect at locus l ; $d_l^{(ij)}$ is coded as 1, if Aa occurs at locus l for the hybrid, otherwise $d_l^{(ij)}$ is coded as 0 ; β_l^D is the dominance effect at locus l ; and e_{ij} is the random error that follows a normal distribution $N(0, \sigma_e^2)$.

Rewrite the model (1) to the matrix form:

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{X}_A \boldsymbol{\beta}_A + \mathbf{X}_D \boldsymbol{\beta}_D + \mathbf{e}, \quad (2)$$

where \mathbf{y} denotes the vector of the phenotype values; $\mathbf{1}_n$ the unit vector of length n (here n is the number of hybrids); \mathbf{X}_A the additive marker matrix; $\boldsymbol{\beta}_A$ the additive effects vector; \mathbf{X}_D the dominance marker matrix; $\boldsymbol{\beta}_D$ the additive effects vector; and \mathbf{e} the random errors vector.

Because the number of markers p is usually much greater than the number of hybrid observations n , it is challenging to estimate all the parameters in the model above. i.e, $p \gg n$. Thus, it is reasonable to specify a prior on $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_D$ to make the marker effects estimable. Let $\boldsymbol{\beta}_A$ follows normal distribution $N(\mathbf{0}, \sigma_A^2 \mathbf{I}_p)$ and $\boldsymbol{\beta}_D$ follows

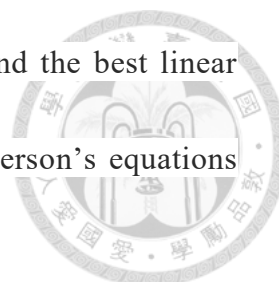
a normal distribution $N(\mathbf{0}, \sigma_D^2 \mathbf{I}_p)$. Furthermore, we reparameterize $\mathbf{g}_A = \mathbf{X}_A \boldsymbol{\beta}_A$ and $\mathbf{g}_D = \mathbf{X}_D \boldsymbol{\beta}_D$ in model (2) to yield the genomic best linear unbiased prediction (gBLUP) model

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{g}_A + \mathbf{g}_D + \mathbf{e}, \quad (3)$$

where \mathbf{g}_A is the vector of additive genomic values following multivariate normal distribution $MVN(\mathbf{0}, \mathbf{K}_A \sigma_A^2)$; and \mathbf{g}_D is the vector of dominance genomic value following multivariate normal distribution $MVN(\mathbf{0}, \mathbf{K}_D \sigma_D^2)$. \mathbf{K}_A is a rescaled variance-covariance matrix for \mathbf{g}_A , which is called as the genomic relationship matrix for the additive effects. Similarly, \mathbf{K}_D is the genomic relationship matrix for the dominance effects. Both \mathbf{K}_A and \mathbf{K}_D are known and derived from the genotypic data of the hybrids. The variance-covariance for \mathbf{g}_A is given by $Cov(\mathbf{g}_A) = \mathbf{X}_A \mathbf{X}_A^T \sigma_A^2$. Similarly, \mathbf{g}_D is given by $Cov(\mathbf{g}_D) = \mathbf{X}_D \mathbf{X}_D^T \sigma_D^2$. So, \mathbf{K}_A is associated with $\mathbf{X}_A \mathbf{X}_A^T$ and \mathbf{K}_D is associated with $\mathbf{X}_D \mathbf{X}_D^T$. Here we have normalized both the additive and dominance marker matrices \mathbf{X}_A and \mathbf{X}_D .

2.4 Estimation for marker effects

To estimate the additive and dominance effects, we use the linear mixed effects model (LMM). The LMM estimation regards both the additive and dominance effects as random effects, which are distributed with $N(\mathbf{0}, \sigma_A^2)$ and $N(\mathbf{0}, \sigma_D^2)$, respectively. We



can obtain the best linear unbiased estimation (BLUE) of μ and the best linear unbiased predictors (BLUPs) of \mathbf{g}_A and \mathbf{g}_D through the Henderson's equations

(Henderson, 1975),

$$\begin{bmatrix} n & \mathbf{1}_n^T & \mathbf{1}_n^T \\ \mathbf{1}_n & \mathbf{I}_n + \mathbf{K}_A^{-1}\lambda_A & \mathbf{I}_n \\ \mathbf{1}_n & \mathbf{I}_n & \mathbf{I}_n + \mathbf{K}_D^{-1}\lambda_D \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}}_A \\ \hat{\mathbf{g}}_D \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n^T \mathbf{y} \\ \mathbf{y} \\ \mathbf{y} \end{bmatrix},$$

w
h
e
r
e

the heritability is obtained as the ratio of genetic variance ($\sigma_A^2 + \sigma_D^2$) to the total phenotypic variance ($\sigma_A^2 + \sigma_D^2 + \sigma_e^2$). $\mathbf{1}_n$ is the unit vector of the number of hybrids n ; $\lambda_A = \sigma_e^2 \sigma_A^2$ and $\lambda_D = \sigma_e^2 \sigma_D^2$ are regularization parameters. For a given training data set with known phenotypic values,

2.5 Prediction for GEBVs

and can be computed directly. The variance components σ_A^2 , σ_D^2 and σ_e^2 can be estimated by restricted maximum likelihood estimation (REML), which are unbiased estimator. For a given breeding population with known genotypic data only, it is now needed to predict the GEBVs of individuals in the breeding populations. The breeding population package `genmer` (Covarrubias-Pazaran, 2016). Subsequently, is assumed to consist of all of the possible hybrids. Let $\mathbf{K}_A^{(bp)}$ denotes the relationship matrix among the additive effects between the breeding population and the training population. Similarly, the corresponding relationship matrix among the dominance

effects is denoted by $\mathbf{K}_D^{(bp)}$. Also, let $\mathbf{g}_A^{(bp)}$ and $\mathbf{g}_D^{(bp)}$ denote the additive and dominance genomic value vectors for the breeding population, respectively. Afterwards from Henderson (1977), the BLUPs for $\mathbf{g}_A^{(bp)}$ and $\mathbf{g}_D^{(bp)}$ are given by

$$\mathbf{g}_A^{(bp)} = \mathbf{K}_A^{(bp)} \mathbf{K}_A^{-1} \hat{\mathbf{g}}_A,$$

and

$$\mathbf{g}_D^{(bp)} = \mathbf{K}_D^{(bp)} \mathbf{K}_D^{-1} \hat{\mathbf{g}}_D.$$

The GEBVs for the breeding population are then predicted by

$$\hat{\mathbf{y}}^{(bp)} = \mathbf{1}_{nF_1} \hat{\mu} + \hat{\mathbf{g}}_A^{(bp)} + \hat{\mathbf{g}}_D^{(bp)},$$

where nF_1 is the number of individuals in the breeding population.

From Werner et al. (2018), model (1) can be rewritten as

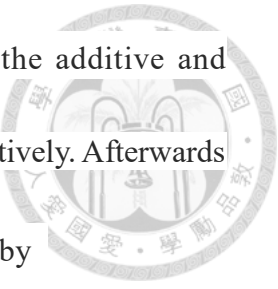
$$y_{ij} = \mu + GCA_i + GCA_j + SCA_{ij} + e_{ij} \quad (4)$$

where GCA_i and GCA_j denote the general combining abilities for P_i and P_j , respectively; and SCA_{ij} denotes the specific combining abilities for the hybrids $P_i \otimes P_j$. As the result, from model (3), we have

$$\mathbf{g}_A^{(ij)} = GCA_i + GCA_j,$$

$$\mathbf{g}_D^{(ij)} = SCA_{ij}.$$

The BLUP of $\mathbf{g}_D^{(ij)}$ is obtained equivalently as





$$\widehat{SCA}_{ij} = \widehat{g}_D^{(ij)}.$$

For a particular parental line i , let

$$\bar{G}_A^{(i)} = \frac{\sum_{j \neq i}^{n_0} \widehat{g}_A^{(ij)}}{n_0 - 1},$$

and

$$\bar{G}_A = \frac{\sum_{i=1}^{n_0} \sum_{j \neq i}^{n_0} \widehat{g}_A^{(ij)}}{nF_1},$$

where $\bar{G}_A^{(i)}$ is the average of \widehat{g}_A of the particular parental line i , \bar{G}_A is the average of all \widehat{g}_A , n_0 is the number of the parental lines, nF_1 is the number of the hybrids.

From the identity $g_A^{(ij)} = GCA_i + GCA_j$, an unbiased estimator for GCA_i is given by

$$\widehat{GCA}_i = \frac{(n_0 - 1)\bar{G}_A^{(i)}}{n_0 - 2} - \frac{n_0 \bar{G}_A}{2(n_0 - 2)}.$$

When n_0 is large enough, we simply let

$$\widehat{GCA}_i = \bar{G}_A^{(i)} - \frac{\bar{G}_A}{2}.$$

Hybrid breeding is a means of heterosis to improve the yield and quality of a crop.

From Wu et al. (2019), the GEBV-based MPH and BPH for $P_i \otimes P_j$ can be estimated

by

$$\widehat{MPH}_{ij} = \widehat{SCA}_{ij},$$

and

$$\widehat{BPH}_{ij} = \widehat{SCA}_{ij} - |\widehat{GCA}_i - \widehat{GCA}_j|.$$



Under the positive heterosis assumption, the value of MPH or BPH is larger, and the heterosis of the hybrid is stronger.

Result

3.1 Pumpkin data analysis

3.1.1 Prediction of potential hybrids and parental lines

Crossing the 119 parental lines, we have $\binom{119}{2} = 10,011$ hybrids within the intra-crossing group of *C.maxima*. Based on the gBLUP model above, we can obtain the GEBVs of all parental lines and hybrid lines. Furthermore, GEBV-based SCA for each hybrid and the GCAs for its parental lines can be calculated, too. For illustration purposes, we only report the top 25 potential hybrids with large GEBVs, together with their SCA, MPH and BPH in Table 1, and the top 10 superior parental lines with large GCAs in Table 2 for FWT within *C.maxima*. Table 1 extracts an important finding that $GEBV_{ij}$ are all greater than both $GEBV_i$ and $GEBV_j$, indicating that the hybrids have better performance than both of their parental lines; and the result can also be presented by the positive and strong BPH_{ij} . Moreover, we can also find all values of MPH_{ij} are equal to the values of SCA_{ij} , which are positive, implying that there exists an obvious heterosis effect for trait FWT within *C.maxima* intra-crossing group. The resulting

GEBVs and GCAs are useful information for plant breeders to select superior parental lines. From Table 2, P026 is involved in the top 6 hybrids having greater GEBVs, and is also the parental line of almost half of the top 25 potential hybrids, demonstrating that P026 could be the most potential line with large trait FWT for its hybrids.

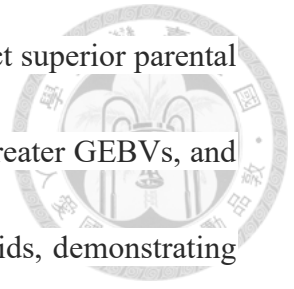
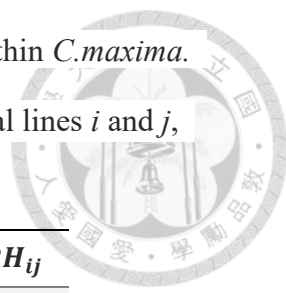


Table 1. The top 25 potential hybrids with GEBVs for trait FWT within *C.maxima*.

Note that $GEBV_{ij}$ represents the GEBVs for hybrid $P_i \otimes P_j$, parental lines i and j , respectively.



$P_i \otimes P_j$	$GEBV_{ij}$	SCA_{ij}	MPH_{ij}	BPH_{ij}
P026:P236	3.5247	1.22387	1.22387	1.16776
P026:P234	3.46522	1.19235	1.19235	1.10848
P026:P235	3.45985	1.17050	1.1705	1.103
P026:P027	3.43926	1.17168	1.17168	1.08259
P026:P028	3.42251	1.13012	1.13012	1.06567
P026:P237	3.37263	1.15218	1.15218	1.01627
P227:P235	3.23955	1.04271	1.04271	1.01831
P227:P236	3.23017	1.02186	1.02186	0.98606
P227:P234	3.21376	1.03341	1.03341	1.02537
P028:P227	3.1845	0.98495	0.98495	0.95749
P026:P302	3.15434	0.96857	0.96857	0.79823
P007:P026	3.15141	0.96106	0.96106	0.79529
P027:P227	3.1126	0.93785	0.93785	0.93503
P227:P237	3.08472	0.95678	0.95678	0.91278
P234:P313	3.06315	1.08743	1.08743	0.87622
P235:P313	3.05542	1.06322	1.06322	0.83564
P028:P313	3.03891	1.04399	1.04399	0.81336
P236:P313	2.99477	0.99109	0.99109	0.75211
P026:P233	2.97518	0.87697	0.87697	0.61969
P027:P313	2.95479	0.98468	0.98468	0.77868
P008:P026	2.93982	0.78612	0.78612	0.58396
P026:P254	2.93148	0.7519	0.7519	0.5754
P007:P227	2.93007	0.83255	0.83255	0.75869
P227:P302	2.9299	0.83664	0.83664	0.7582
P100:P234	2.92776	0.90815	0.90815	0.74063

Table 2. The top 10 superior parental lines with GCAs for FWT within *C.maxima*.

P_i	$GEBV_i$	GCA_i
P026	2.35171	0.39086
P236	2.23949	0.33475
P028	2.2228	0.3264
P235	2.21671	0.32335
P234	2.18397	0.30698
P027	2.17354	0.30177
P227	2.1679	0.29895
P252	2.08416	0.25708
P237	2.07989	0.25495
P324	2.05612	0.24306



3.1.2 Variance components and heritability

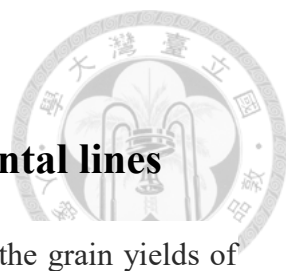
Table 3 presents the estimations of the variance components and heritability using REML for trait FWT within intra-crossing group of *C.maxima*. We can see that there exists some non-negligible dominance effects in the *C.maixma* intra-crossing group, and it explains why the values of MPH_{ij} and BPH_{ij} are all positive.

Table 3. The estimations of the variance components and heritability using REML for trait FWT of *C.maxima*

σ_A^2	σ_D^2	σ_e^2	h^2
0.244	0.256	0.064	0.887

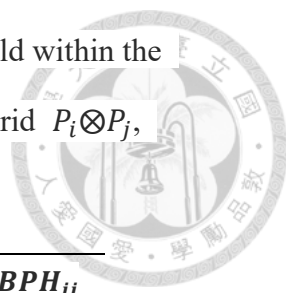
3.2 Maize data analysis

3.2.1 Prediction of potential hybrids and parental lines



There are $\binom{24}{2} = 276$ hybrids of the germplasm of maize, and all the grain yields of the hybrids are known. We chose 50 hybrids as training set at random to build the gBLUP model. For illustration purposes, we also only report the top 25 potential hybrids with large GEBVs, together with their SCA, MPH and BPH in Table 4, and the top 10 superior parental lines with large GCAs in Table 5. As the table shows, the $GEBV_{ij}$ are all greater than both $GEBV_i$ and $GEBV_j$, which represents that the performances of the hybrids are all better than their parental lines; and the positive BPH_{ij} can also support the result. Besides, the positive SCA_{ij} , which are equal to MPH_{ij} , indicates that there also exist heterosis effects within this groups. From Table 5, we can find that the top 5 parental lines with the greatest GEBVs: Mo17, NC350, B73, B97, and OH7B are all the parental lines of the top 5 hybrids with large GEBVs. It means that the five inbred lines are potential parental lines for high grain yield, which can be a worthy information for plant breeders.

Table 4. The top 25 potential hybrids with GEBVs for trait grain yield within the maize population. Note that $GEBV_{ij}$ represents the GEBVs for hybrid $P_i \otimes P_j$, parental lines i and j , respectively.



$P_i \otimes P_j$	$GEBV_{ij}$	SCA_{ij}	MPH_{ij}	BPH_{ij}
B97:MO17	19.419	1.766	1.766	1.651
MO17:CML277	18.974	1.704	1.548	1.454
B73:NC350	18.788	1.536	1.536	1.507
OH7B:NC350	18.142	1.535	1.535	1.418
B73:CML228	17.788	1.480	1.480	1.325
MO17:NC358	17.569	1.341	1.341	1.198
MS71:KI3	16.681	1.537	1.537	1.267
NC358:TZI8	16.613	1.340	1.340	1.322
B97:CML69	16.600	1.468	1.468	1.282
B73:NC358	16.369	1.420	1.420	0.982
B73:MO17	15.892	1.460	1.460	1.181
MO17:CML69	15.429	1.340	1.340	1.155
MO17:KI3	15.343	0.956	0.926	0.817
CML333:NC358	15.326	0.906	0.906	0.747
MO17:CML228	15.241	0.915	0.915	0.872
B73:B97	15.237	0.807	0.807	0.615
KI11:NC350	15.134	1.086	1.086	0.732
B97:NC358	15.032	1.129	1.129	0.721
B97:TX303	15.027	0.620	0.620	0.524
OH7B:CML228	14.826	0.794	0.794	0.688
MO17:CML52	14.723	0.812	0.812	0.757
CML69:CML228	14.422	0.608	0.608	0.595
B97:M37W	14.204	1.024	1.024	0.953
MS71:NC358	13.697	0.923	0.923	0.711
M162W:CML228	13.142	0.811	0.811	0.732

Table 5. The top 10 superior parental lines with GCAs for trait grain yield within the maize population.



P_i	GEV_i	GCA_i
MO17	18.223	0.5563
NC350	18.061	0.5356
B73	17.821	0.5055
B97	17.692	0.4410
OH7B	16.592	0.4111
CML69	15.945	0.3996
KI3	15.746	0.3155
CML277	15.449	0.3050
MS71	13.418	0.2034
TZI8	12.403	0.2027

3.2.2 Variance components and heritability

Table 6 reveals the estimations of the variance components and heritability using REML for trait grain yield with the maize population. There exist strong dominance effects apparently, and it also explains why the values of MPH_{ij} and BPH_{ij} are all positive.

Table 6. The estimations of the variance components and heritability using REML for trait grain yield within the maize population.

σ_A^2	σ_D^2	σ_e^2	h^2
1.69	2.24	0.753	0.839

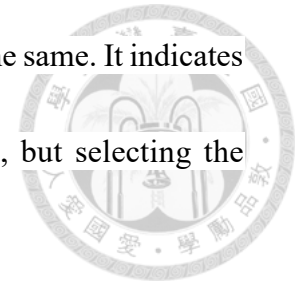
Discussion



Previous GS surveys such as that conducted by Wu et al., (2019) used a two-step method, which needs to evaluate the GS models and estimation methods through data cross-validation to obtain the GEBVs of all the hybrid combinations first, and to adapt another gBLUP model rewritten by GCA and SCA to estimate the variance components and heritability. In this study, we propose another best linear unbiased prediction approach which can get all information of interest through a single additive plus dominance effects gBLUP model. To evaluate the genomic relationship between the training set and all the possible hybrid lines can estimate their additive and dominance genomic values, and the REMLs of variance components can be used to obtain the heritability. In brief, this procedure will be more efficient and practical than before.


The variance components of additive and dominance effects estimated by Wu et al., (2019) are 0.033 and 0.202, respectively. The different variance components will affect the BLUPs of \mathbf{g}_A and \mathbf{g}_D and further impact the final estimated GEBVs. In this study we used another method to evaluate the kinship matrixes of additive and dominance effects, and we believe the result is more reliable and closer to real situation. Comparing the top 25 potential hybrid lines and top 10 superior parental lines of *C.maxima* in our study with those in Wu et al. (Table 7 and Table 8), both of the results have high consistency with each other. Especially the hybrid lines, even though the

orders are different, the top 9 hybrid lines with highest GEBVs are the same. It indicates that different estimated methods may result in different GEBVs, but selecting the potential lines for breeders is the most important goal.



The scatter plots of the GEBVs and the phenotypic values for both the two data sets are in Figure 2 and Figure 3, respectively. The colored points indicate those selected hybrids with the highest GEBVs. Most of the colored points gather on the upper right corner, it means that the selected hybrid line with higher GEBV also has the higher phenotypic value in reality. It's a valuable result because phenotypic selection is usually costly and time-consuming for selective breeding, if there is great consistency between the result of genomic selection and phenotypic selection, genomic prediction will be a good choice for breeders.

Comparing Figure 3.a with Figure 3.b and Figure 3.c, we can find that the number of selected hybrid lines in $TM \otimes TS$ is much more than which in $TM \otimes TM$ and $TS \otimes TS$, it's in our expectation because the heterosis of inter-crossing group $TM \otimes TS$ is stronger. Additionally, the GEBVs of both intra-crossing hybrid lines of $TM \otimes TM$ and $TS \otimes TS$ are more approximate to the observed grain yield values than the GEBVs of inter-crossing hybrid lines of $TM \otimes TS$, because their genetic relationships between training and testing populations are much closer.



The results above highlight that the prediction accuracy of GS model might be influenced by various different factors such as training population sizes, varying degrees of relationship from reference populations, genetic distance and genetic relationship between training and testing populations, different estimations of genetic variance and the number of SNP markers ([Lee et al., 2017](#), [Scutari et al., 2016](#), [Estaghvirou et al., 2013](#)). Although we can select the most potential hybrids and superior parental lines for breeders, as explained earlier, there are still some defects about predict accuracy need to be improved. In particular, there is often a large number of possible hybrid lines need to be evaluated in a hybrid breeding. As a result, the training population size is often much smaller than the testing population size. As suggested by Akdemir et al. (2015), determination of an optimal training population could be an efficient key to a successful GS. The readers interested in this issue are referred to Ou and Liao (2019) and Sanchez and Akdemir (2021).

Table 7. The list of top 25 potential hybrids with GEBVs for trait FWT within *C.maxima* in our study and in Wu et al. The colored lines are both selected in our study and Wu et al.



our study	Wu <i>et al.</i> (2019)
P026:P236	P026:P236
P026:P234	P026:P027
P026:P235	P026:P235
P026:P027	P026:P234
P026:P028	P026:P237
P026:P237	P026:P028
P227:P235	P227:P236
P227:P236	P227:P235
P227:P234	P227:P234
P028:P227	P026:P233
P026:P302	P227:P237
P007:P026	P028:P227
P027:P227	P027:P227
P227:P237	P026:P138
P026:P233	P026:P253
P026:P254	P026:P255
P100:P234	P026:P254
P026:P253	P026:P302
P100:P235	P007:P026
P008:P026	P008:P026
P026:P255	P026:P252
P234:P313	P026:P243
P235:P313	P138:P241
P028:P100	P026:P241
P026:P252	P100:P236

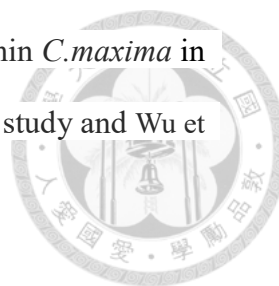


Table 8. The top 10 superior parental lines with GCAs for FWT within *C.maxima* in our study and in Wu et al. The colored lines are both selected in our study and Wu et al.

our study	Wu <i>et al.</i> (2019)
P026	P026
P227	P138
P236	P236
P028	P235
P235	P234
P234	P237
P027	P027
P324	P028
P252	P227
P237	P255

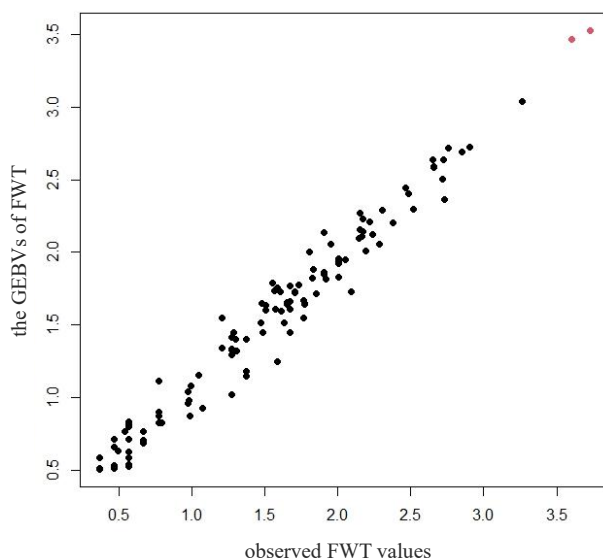


Figure 2: the scatter plot of intra-crossing hybrid lines of *C.maxima* for the FWT values and the GEBVs of the FWT. The colored points indicate those selected hybrids with the highest GEBVs. Note that there are only 2 selected hybrids in the figure because the training set only had 119 hybrid lines.

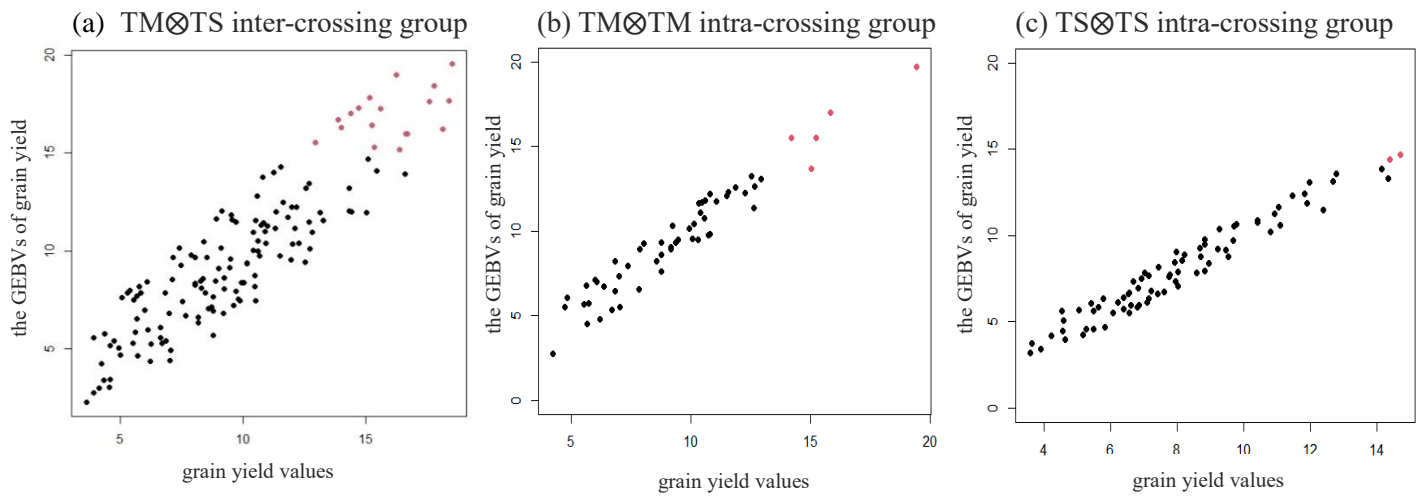
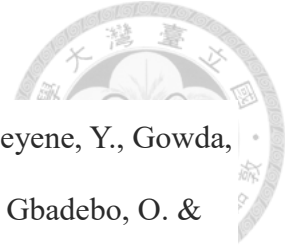


Figure 3: (a) the scatter plot of inter-crossing group of $TM \otimes TS$ for the grain yield values and the GEBVs of the grain yield. (b) the scatter plot of intra-crossing group of $TM \otimes TM$ for the grain yield values and the GEBVs of the grain yield. (c) the scatter plot of intra-crossing group of $TS \otimes TS$ for the grain yield values and the GEBVs of the grain yield. The colored points indicate those selected hybrids with the highest GEBVs.



References

- Atanda, S.A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., Gowda, M., Kate, D., Xuecai, Z., Boddupalli, M.P., Pangirayi, T., Eric, Y.D., Gbadebo, O. & Kelly, R.R. (2014). Genomic Selection in Barley Breeding. *Biotechnological Approaches to Barley Improvement*, 69, 367-378.
- Akdemir, D., Sanchez, J.I., Jannink, J.L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, 47, 38.
- Clark, S.A., van der Werf, J. (2013). Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding. *Genome-Wide Association Studies and Genomic Prediction*, 1019, 321-330.
- Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., N. de Leon, Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*, 44, 803-807.
- Estagvirou, S.B.O., Joseph, O.O., Torben, S.-S., Carsten, K., Milena, O., Andres, G. & Piepho, H.-P. (2013). Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC Genomics*, 14, 860.
- Falconer, D.S., and Mackay, T.F.C. (1996). *Introduction to quantitative genetics*.
- Fonseca, S., and Patterson, F.L. (1968). Hybrid vigor in a seven-parent diallel cross in common winter wheat (*Triticum aestivum* L.). *Crop Sci.*, 8, 85–88.
- Griffing, B. (1956). Concept of general and specific combining ability in relation to diallel crossing system. *Aust. J. Biol. Sci.*, 9, 463-493.



Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., Michael, D. McMullen, J.B., Holland, S.J., Szalma, R.J., Yu, W.J. (2019). Optimal Designs for Genomic Selection in Hybrid Crops. *Molecular Plant*, 12, 390-401.

Gerhard, M., Bruce, T., Crump, R.E., Mehar, S.K. & Herman, W.R. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, 41(1), 56.

Giovanny, C.-P. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *Plos One*, 11(6): e0156744.

Haile, J.K., N'Diaye, A., Clarke, F., Clarke, J., Knox, R., Rutkoski, J., Bassi, F.M., Pozniak, C.J. (2018). Genomic selection for grain yield and quality traits in durum wheat. *Molecular Breeding*, 38, 75.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 32, 69-84.

Henderson, C.R. (1977). Best linear unbiased prediction of breeding values not in the model for records. *Journal of Dairy Science*, 60, 783-787.

Henderson, C.R. (1977). Best linear unbiased prediction of breeding values not in the model for records. *Journal of Dairy Science*, 60, 783-787.

Jesse, A. Poland, T.W.R. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome*. 5.

Jean-Luc, J., Aaron, J., Iwata, L.H. (2010). Genomic selection in plant breeding: from theory to practice. *BRIEFINGS IN FUNCTIONAL GENOMICS*, 9, 166 -177.

Kadkol, G.P., Anand, I.J. & Sharma, R.P. (1984). Combining ability and heterosis in sunflower. *Indian J. Genet.*, 44, 447-451.



Lee, S.H., Sam, C., Julius, H.J., van der Werf (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. PLOS ONE, 12(12): e0189775.

Miller, J.F., Hammond, J.J. & Roath, W.W. (1980). Comparison of inbred vs. single-cross testers and estimation of genetic effects in sunflowers. *Crop Sci.*, 20, 703-706.

Marco, S., Ian, M., David, B. (2016). Using Genetic Distance to Infer the Accuracy of Genomic Prediction. PLOS GENETICS, 12(9): e1006288.

Ou, J.H., Liao, C.T. (2019). Training set determination for genomic selection. *Theoretical and Applied Genetics*, 132, 2781–2792.

Qurban, A., Arfan, A., Awan. M.F., Tariq M., Sajed, A., Samiullah, T.R., Saira, A., Salah ud D., Mukhtar, A., Muhammad, N.S., Muhammad, S., Nazar, H.K., Muhammad, A., Idrees, A.N. and Tayyab, H. (2014) Combining ability analysis for various physiological, grain yield and quality traits of *Zea mays* L. *Life Science Journal*, 11, 540-551.

Robert, J., Elshire, J.C., Glaubitz, Q.S., Jesse, A., Poland, K.K., Edward, S.B., Sharon, E.M. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLOS ONE, 6.

Sherrif, N.M., Appadurai, R. & Rangasamy, M. (1985). Combining ability in sunflower. *Indian J. Agric. Sci.*, 55, 315-318.

Sprague, G.F., Tatum, L.A. (1942) General vs. specific combining ability in single crosses of corn. *Agronomy Journal*, 34, 923-932.

Sikiru, A.A., Michael, O., Burgueño, J., Jose, C., Dzidzienyo, D., Beyene, Yoseph., Gowda, M., Dreher, K., Xuecai, Z., Boddupalli, M.P., Pangirayi, T., Eric, Y.D.,



Gbadebo, O. & Kelly, R.R. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theoretical and Applied Genetics*, 134, 279–294.

Sánchez, J.I.Y. and Deniz, A. (2021). Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview. *Front. Plant Sci.*, 12: 715910.

Troyer, A. F. (2006). Adaptedness and Heterosis in Corn and Mule Hybrids. *Crop Sci.*, 46, 528-543.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S. (2009). Invited Review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, 92, 16-24.

Wu, P.-Y., Tung, C.-W., Lee, C.-Y., Liao, C.-T. (2019). Genomic Prediction of Pumpkin Hybrid Performance. *The Plant Genome*, 12 (2).

Werner, C.R., Qian, L., Voss-Fels, K.P., Abbadi, A., Leckband, G., Frisch, M., Snowdon, R.J. (2018). Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. *Theor. Appl. Genet.*, 131(2), 299–317.

Zhao, Y., Zeng, J., Fernando, R., Reif, J.C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci.* 53, 802.



Rcode

```
##### The needed data #####
## hybrid.pheno: the observed phenotypic data of the 10011 hybrids
## Xa.bp: the additive effects of the 10011 hybrids
## Xd.bp: the dominance effects of the 10011 hybrids
## Xa.train = the additive effects of the 119 training set
## Xd.train = the dominance effects of the 119 training set
#####
index = which(!is.na(hybrid.pheno$F1.weight))
Xa.train = Xa.bp[index,] ;dim(Xa.train)
Xd.train = Xd.bp[index,] ;dim(Xd.train)

##### calculate the relation matrix #####
##### normalize the Xa.bp and Xd.bp #####
#####
relation.matrix.additive = function(Xa.bp){
  Xa.bp.normal = matrix(0, nrow = nrow(Xa.bp), ncol = ncol(Xa.bp))
  for (i in 1: ncol(Xa.bp)){
    cat(paste(i),sep = "\n")
    for (j in 1: nrow(Xa.bp)){
      cat(paste(j),sep = "\n")
      Xa.bp.normal[j,i] = (Xa.bp[j, i] - mean(Xa.bp[,i]))/sd(Xa.bp[,i])
    }
  }
  rownames(Xa.bp.normal) = rownames(Xa.bp)
  colnames(Xa.bp.normal) = colnames(Xa.bp)
  Xa.train.normal = Xa.bp.normal[index,]
  Xa.train.normal = as.matrix(Xa.train.normal)
  Ka = Xa.train.normal %*% t(Xa.train.normal) / ncol(Xa.train.normal)
  return(Ka)
}
Ka = relation.matrix.additive(Xa.bp)
```



```
relation.matrix.dominance = function(Xd.bp){
  Xd.bp.normal = matrix(0, nrow = nrow(Xd.bp), ncol = ncol(Xd.bp))
  for (i in 1: ncol(Xd.bp)){
    cat(paste(i),sep = "\n")
    for (j in 1: nrow(Xd.bp)){
      cat(paste(j),sep = "\n")
      Xd.bp.normal[j,i] = (Xd.bp[j, i] - mean(Xd.bp[,i]))/sd(Xd.bp[,i])
    }
  }
  rownames(Xd.bp.normal) = rownames(Xd.bp)
  colnames(Xd.bp.normal) = colnames(Xd.bp)
  Xd.train.normal = Xd.bp.normal[index,]
  Xd.train.normal = as.matrix(Xd.train.normal)
  Kd = Xd.train.normal %*% t(Xd.train.normal) / ncol(Xd.train.normal)
  return(Ka)
}
Kd = relation.matrix.dominance(Xd.bp)
```

```
#####
#### sommer package
#####
install.packages("sommer")
library(sommer)

#### create incidence matrices of gA and gD (ZA and ZD)
Za = Zd = diag(1, nrow(Xa.train), nrow(Xa.train))
rownames(Za) = rownames(Zd) = colnames(Za) = colnames(Zd) =
rownames(Xa.train)

#### fit the model based on GBLUP model to get the REMLof variance components
fit.GBLUP = mmer(F1.weight ~ 1,
                 random = ~vs(list(Za), Gu = Ka) +vs(list(Zd), Gu = Kd),
                 rcov = ~vs(units),
                 data = hybrid.pheno)
variance.REML = unlist(fit.GBLUP$sigma) ;variance.REML
```



```
##### extract ga.BLUP, gd.BLUP from GBLUP model
ga.BLUP = fit.GBLUP$U$`u:Za`[[1]]
ga.BLUP = as.matrix(ga.BLUP)
gd.BLUP = fit.GBLUP$U$`u:Zd`[[1]]
gd.BLUP = as.matrix(gd.BLUP)

#####
#### SCA(ij) = gd_hat_bp
#### GCA(i)
#####
SCA = function(Xd.bp.normal, Xd.train.normal, Kd, gd.BLUP){
  Xd.bp.normal = as.matrix(Xd.bp.normal)
  Kd.bp = Xd.bp.normal %*% t(Xd.train.normal) / ncol(Xd.bp.normal)
  gd.bp = Kd.bp %*% solve(Kd) %*% gd.BLUP
  return(gd.bp)
}
SCA = SCA(Xd.bp.normal, Xd.train.normal, Kd, gd.BLUP)
colnames(SCA) = "SCA"

GCA = function(Ka, Xa.bp.normal, Xa.train.normal, ga.BLUP){
  Ka = Ka + diag(10^(-10), nrow = nrow(Ka), ncol = ncol(Ka))
  Xa.bp.normal = as.matrix(Xa.bp.normal)
  Ka.bp = Xa.bp.normal %*% t(Xa.train.normal) / ncol(Xa.bp.normal)
  ga.bp = Ka.bp %*% solve(Ka) %*% ga.BLUP
  ga.bp = data.frame(ga.bp)
  ga.bp$P1 = rownames(ga.bp)
  ga.bp$P2 = rownames(ga.bp)
  ga.bp$P1 = substr(ga.bp$P1, 1, 4)
  ga.bp$P2 = substr(ga.bp$P2, 6, 9)
  GCA = c()
  for(i in 1: length(unique(c(ga.bp$P1, ga.bp$P2)))){
    a = c(which(ga.bp$P1 == unique(c(ga.bp$P1, ga.bp$P2))[i]), which(ga.bp$P2
== unique(c(ga.bp$P1, ga.bp$P2))[i]))
    GCA[i] = mean(ga.bp[a,1]) - mean(ga.bp[,1])/2
  }
  GCA = data.frame(GCA)
  rownames(GCA) = unique(c(ga.bp$P1, ga.bp$P2))
}
```



```

    colnames(GCA) = "GCA"
    return(GCA)
}
GCA = data.frame(GCA(Ka, Xa.bp.normal, Xa.train.normal, ga.BLUP))

```



```

#####
## MPH(ij) and BPH(ij)
#####

```

```

MPH = SCA
colnames(MPH) = "MPH"

```

```

BPH.F1 = substr(rownames(SCA), 1, 4)
BPH.F2 = substr(rownames(SCA), 6, 9)
BPH = matrix(0, nrow = nrow(SCA), ncol = 1)
for (i in 1: nrow(SCA)) {
  BPH[i, 1] = as.numeric(SCA[i, 1]) - abs(GCA[which(rownames(GCA) ==
BPH.F1[[i]]), 1] - GCA[which(rownames(GCA) == BPH.F2[[i]]), 1])
}
rownames(BPH) = rownames(MPH)
colnames(BPH) = "BPH"

```

```

#####
## GEBV(ij) = μ + ga.bp + gd.bp
## GEBV(i) = μ + 2GCA(i)
#####

```

```

GEBV.bp = function(μ, Ka, Xa.bp.normal, Xa.train.normal, ga.BLUP, Xd.bp.normal,
Xd.train.normal, Kd, gd.BLUP){
  Ka = Ka + diag(10^(-10), nrow = nrow(Ka), ncol = ncol(Ka))
  Xa.bp.normal = as.matrix(Xa.bp.normal)
  Ka.bp = Xa.bp.normal %*% t(Xa.train.normal) / ncol(Xa.bp.normal)
  ga.bp = Ka.bp %*% solve(Ka) %*% ga.BLUP
  Xd.bp.normal = as.matrix(Xd.bp.normal)
  Kd.bp = Xd.bp.normal %*% t(Xd.train.normal) / ncol(Xd.bp.normal)
  gd.bp = Kd.bp %*% solve(Kd) %*% gd.BLUP

  GEBV.bp = matrix(0, 10011, 1)

```



```
for (i in 1:10011) {
  GEBV.bp[i, 1] =  $\mu[i, 1] + ga.bp[i, 1] + gd.bp[i, 1]$ 
}
rownames(GEBV.bp) = rownames(ga.bp)
colnames(GEBV.bp) = "GEBV"
F1 = rownames(GEBV.bp)
F1 = as.matrix(F1)
GEBV.bp = cbind(F1, GEBV.bp)
colnames(GEBV.bp) = c("F1", "GEBV")
GEBV.bp[, 2] = as.numeric(GEBV.bp[, 2])
GEBV.bp = GEBV.bp[order(GEBV.bp[, 2], decreasing = T),]
return(GEBV.bp)
}
 $\mu$  = matrix(as.numeric(round(fit.GBLUP$fitted[[1]], 2)), 10011, 1)
GEBV.bp = GEBV.bp( $\mu$ , Ka, Xa.bp.normal, Xa.train.normal, ga.BLUP, Xd.bp.normal,
Xd.train.normal, Kd, gd.BLUP)
```

```
GEBV.parental = function( $\mu$ , GCA){
  GEBV.parental = matrix(0, 142, 1)
  for (i in 1:142) {
    GEBV.parental[i, 1] =  $\mu[i, 1] + 2 * GCA[i, 1]$ 
  }
  rownames(GEBV.parental) = rownames(GCA)
  colnames(GEBV.parental) = "GEBV"
  P = rownames(GEBV.parental)
  P = as.matrix(P)
  GEBV.parental = cbind(P, GEBV.parental)
  colnames(GEBV.parental) = c("parental lines", "GEBV")
  GEBV.parental = data.frame(GEBV.parental)
  GEBV.parental[, 2] = as.numeric(GEBV.parental[, 2])
  GEBV.parental = GEBV.parental[order(GEBV.parental[, 2], decreasing = T), ]
  return(GEBV.parental)
}
GEBV.parental = GEBV.parental( $\mu$ , GCA)
```