Genomic Prediction of Pumpkin Hybrid Performance

Po-Ya Wu, Chih-Wei Tung, Chieh-Ying Lee, and Chen-Tuo Liao*

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. In this study, we propose a systematic procedure to predict hybrid performance using a genomic selection (GS) model that takes both additive and dominance marker effects into account. We first demonstrate the advantage of the additive–dominance effects model over the only additive effects model through a simulation study. Based on the additive–dominance model, we predict genomic estimated breeding values (GEBVs) for individual hybrid combinations and their parental lines. The GEBV-based specific combining ability (SCA) for each hybrid and general combining ability (GCA) for its parental lines are then derived to quantify the degree of midparent heterosis (MPH) or better-parent heterosis (BPH) of the hybrid. Finally, we estimate the variance components resulting from additive and dominance gene action effects and heritability using a genomic best linear unbiased predictor (g-BLUP) model. These estimates are used to justify the results of the genomic prediction study. A pumpkin (Cucurbita spp.) data set is given to illustrate the provided procedure. The data set consists of 320 parental lines with 61,179 collected single nucleotide polymorphism (SNP) markers; 119, 120, and 120 phenotypic values of hybrids on three quantitative traits within C. maxima Duchesne; and 89, 111, and 90 phenotypic values of hybrids on the same three quantitative traits within C. moschata Dechesne.

Abbreviations: BPH, better-parent heterosis; FWT, fruit weight; g-BLUP, genomic best linear unbiased predictor; GBS, genotyping-by-sequencing; GCA, general combining ability; GEBV, genomic estimated breeding value; GS, genomic selection; GBS, genotyping-by-sequencing; GCA, general combining ability; GEBV, genomic estimated breeding value; LS, leaf size; MPH, midparent heterosis; PCA, principal component analysis; QTL, quantitative trait loci; REMI, restricted maximum likelihood estimation; SCA, specific combining ability; SNP, single nucleotide polymorphism; TBV, true breeding value.

CORE IDEAS

- Pumpkin, with rich nutritional compounds, is a staple food in many developing countries.
- Genomic prediction is a practical tool for hybrid performance evaluation in plant breeding.
- Genomic prediction can reduce cost and accelerate a breeding program.
- Genomic prediction applied to select potential hybrid combinations and superior parental lines.

Pumpkin (2n = 40), which belongs to Cucurbitaceae family, consists of at least five domesticated species: C. maxima, C. moschata, C. pepo L., C. argyrosperma C. Huber, and C. ficifolia Bouché (Schaffer and Paris, 2003). The annual production of all Cucurbita crops in the world has exceeded 26.4 Tg in 2016 (http://www.fao.org/faostat/). Pumpkin has been used as a staple food in many developing countries since its mature fruits and seeds are rich in nutritional compounds (Ferriol and Picó, 2008). In this study, we apply genomic prediction to a hybrid pumpkin data set.

Hybrid breeding is a means of using heterosis to improve the yield and quality of a crop. The MPH (Falconcier and Mackay, 1996) is defined as the difference between the hybrid performance and the average of its parental lines. Alternatively, BPH (Fonseca and Patterson, 1968) is defined as the hybrid performance superior to the higher or better parental line. The hybrid breeding is preferred to inbred-line breeding for some reasons such as hybrids can bring greater yield stability, heterozygosity of hybrids allows the combination of dominant...
major genes, and hybrids offer a built-in plant variety protection system (Longin et al., 2012).

A full- or half-diallel design has been conventionally used to estimate SCA for hybrids and GCA for their parental lines in a field trial (Darrudi et al., 2018). From the genetics perspective, the GCA is an indication of additive effects, and the SCA is relevant to nonadditive effects such as dominance and epistasis. In a hybrid breeding program, the main interest of breeders lies in selection of potential hybrid combinations with high SCA and superior parental lines with high GCA (Virmani, 1994). However, the number of crossing combinations can increase dramatically as the number of parental lines increases. This could raise an important issue that the traditional evaluation method for all possible hybrid combinations can be costly or impractical in the field experiment (Fristche-Neto et al., 2018). Recently, GS has been increasingly popular in plant breeding, which can reduce cost and accelerate a breeding program (Jannink et al., 2010; Desta and Ortiz, 2014). In practice, GS has been applied to predict hybrid performance for some important crops such as maize (Zea mays L.) (Technow et al., 2012, 2014), rice (Oryza sativa L.) (Xu et al., 2014; Wang et al., 2016) and wheat (Triticum aestivum L.) (Zhao et al., 2013).

The main concept of GS is to capture all effects of quantitative trait loci (QTL) using dense DNA markers over an entire genome (Meuwissen et al., 2001). The most common DNA markers used in GS are SNPs. Typically, a training population with known genotype and phenotype data is used to build a GS model. Then, the resulting GS model is employed to predict GEBVs for the individuals of a test population with known genotype data only. It is anticipated that incorporating both additive and dominance marker effects into a GS model for hybrid performance prediction could provide better prediction accuracy than the model with additive effects only.

Most of the genomic prediction studies focus on quantitative traits measured in continuous scale. To save the cost of phenotyping, ordinal scores are sometimes recorded for a quantitative trait such as the degree of disease resistance, growth vigor, fruit set ability, etc. The prediction accuracy could be reduced if the methods for analyzing continuous traits are directly adopted to ordinal traits, which are measured by positive integer numbers. Iwata et al. (2009) developed a Bayesian method based on the threshold model for association study on ordinal traits. Similarly, Wang et al. (2013) extended the GS study for ordinal traits and called their proposed methods BayesT methods. Kizilkaya et al. (2014) further quantified the reduction in prediction accuracy of a GS study for ordinal traits vs. continuous traits through simulation studies. In this study, we apply three BayesT methods to analyze two ordinal traits under the additive–dominance effects model.

**MATERIALS AND METHODS**

**Pumpkin Data Set**

The genomic DNA of 320 parental lines was extracted from fresh leaves using a Qiagen Plant DNeasy kit. The genotyping-by-sequencing (GBS) libraries were prepared by the protocol proposed in Elshire et al. (2011) and were sequenced on Illumina system HiSeq 2500 machine. The GBS sequenced reads from 320 parental lines of pumpkin were processed using TASSEL-GBS pipeline for SNP calling (Glubitz et al., 2014) and aligned to the genome of C. maxima downloaded from Cucurbit Genomics Database (http://cucurbitgenomics.org/). After SNP calling, 76,815 SNPs were obtained and filtered by missing rate ≥0.05 for the sites of each SNP. The missing markers were then imputed by the software LinkImpute (Money et al., 2015). There were 61,179 resulting SNPs for further analysis. Principal component analysis (PCA) is performed on the genomic correlation matrix among markers to investigate the population structure among the 320 parental lines. The PCA result is presented in Supplemental Fig. S1 of supplementary information online. The top two PCs explain ~90% (PC1, 64.72%; PC2, 26.85%) of the total genetic variation. Accordingly, the accessions are classified into three species including 142 accessions in C. maxima, 118 accessions in C. moschata, and 60 accessions in C. pepo. We further filter the 61,179 SNPs by minor allele frequency <0.05 within each species. This leads to 4521 SNPs remaining for C. maxima; 6348 SNPs for C. moschata, and 8800 SNPs for C. pepo, which will be used to obtain the SNP genotypes for hybrid combinations of intracrossing groups. The hybrid combinations of an intracrossing group mean that they are F1 hybrids within the domesticated species.

The phenotypic data of hybrids comprise one continuous trait of fruit weight (FWT) and two ordinal traits of growth vigor (GV) and leaf size (LS). The number of phenotypic values of the traits for two intracrossing groups of C. maxima and C. moschata is given in Table 1. Growth vigor is related to the vitality and exuberance of a plant, which was measured at the beginning of reproductive stage and recorded by five ordinal scores. Leaf size was classified into five ordinal scores according to the area of fully expanded leaf.

The phenotypic data were summarized from a collection of historical data provided by Known-You Seed Co., Ltd., which were scattered over the period of 1988 to 2016. All the trials were conducted at the experimental field of the seed company located in southern Taiwan, which can be treated as a single location. Every hybrid had six to ten observations at each time point, and the average of the observations is used as the phenotypic value for the hybrid at the time point. The phenotypic values of one hybrid could be observed across different years, so we remove the year effects from the phenotypic values. It is assumed that the year effects are random effects following the normal distribution N(0,σ2y). Then, the variance component $σ^2_y$ is estimated by pooling all possible sample variances of hybrids as follows:
Table 1. The number of phenotypic values for the traits in the two intracrossing groups.

| Group | Trait† | C. maxima | C. moshata | No. of levels | Description
|-------|--------|-----------|-------------|--------------|----------------|
| FWT   | 119    | 89        | —           |                | kg
| GV    | 120    | 111       | 5           | weak (1) to strong (5) |
| LS    | 120    | 90        | 5           | small (1) to large (5) |

† FWT, fruit weight (continuous trait); GV, growth vigor (ordinal trait); LS, leaf size (ordinal trait).

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(k_i-1)s_i}{\sum_{i=1}^{n}(k_i-1)}
\]

where \(k_i\) is the number of phenotypic values over the years for hybrid \(i\) and \(s_i\) is its sample variance. The estimates of year effects are then generated from \(N(0,\hat{\sigma}^2)\). The \(\hat{\sigma}^2\) are obtained as 0.076 and 0.243 for trait FWT of the intracrossing groups \(C. maxima\) and \(C. moshata\), respectively. The adjusted mean for each hybrid is the average of the phenotypic values, which have been adjusted to the intracrossing groups and mates of year effects are then generated from \(N(0,\sigma^2)\).

Statistical Models

The following additive-effects model, abbreviated as the A model hereinafter, is typically used in a GS study:

\[
y_{ij} = \mu + \sum_{l=1}^{L} a_{ijl}^A + e_{ij}
\]  

where \(y_{ij}\) is a continuous phenotypic value of hybrid \(P_i \times P_j\) (both \(P_i\) and \(P_j\) are parental lines); \(\mu\) is the constant term; \(a_{ij}^A\) coded as -1, 0, or 1 if \(M_1\) or \(M_2\), or \(M_1M_2\) occurs at locus \(l\) for the hybrid; \(a_{ijl}^A\) the additive effect at locus \(l\); \(p\) the number of markers and \(e_{ij}\) a random error following the normal distribution \(N(0,\sigma^2)\). However, considering the heterozygosity effects for a hybrid, we extend the following additive plus dominance effects model, abbreviated as the A+D model hereinafter:

\[
y_{ij} = \mu + \sum_{l=1}^{L} a_{ijl}^A + \sum_{l=1}^{L} d_{ijl}^D + e_{ij}
\]

where \(d_{ijl}^D\) coded as 1 if \(M_1\) or \(M_2\) occurs at locus \(l\) for the hybrid; otherwise \(d_{ijl}^D=0\), and \(d_{ijl}^D\) denotes the dominance effect at locus \(l\).

If \(y_{ij}\) is an ordinal phenotypic value, we proceed with the threshold model. Suppose that \(y_{ij}\) is recorded as one of \(m\) ordinal scores, and the threshold model is given by the following:

\[
y_{ij} = \begin{cases} 1, & \text{if } -\infty < y_{ij} < \gamma_1, \\ 2, & \text{if } \gamma_1 < y_{ij} < \gamma_2, \\ \vdots & \vdots \\ m, & \text{if } \gamma_{m-1} < y_{ij} < -\infty. \\ \end{cases}
\]

where \(\gamma_1, \ldots, \gamma_{m-1}\) are threshold parameters used to define the ordinal categories, and \(w_{ij}\) denotes a latent variable. This means that the original phenotypic value is continuous but now recorded as a positive integer. In other words, \(w_{ij}\) can be thought of as a scaled continuous phenotypic value of Model 2 (Eq. [2]). Let \(w_{ij} = \mu + \sum_{l=1}^{L} a_{ijl}^A + \sum_{l=1}^{L} d_{ijl}^D\), then this relationship can be described by \(w_{ij} = \eta_{ij} + \varepsilon_{ij}\) where \(\varepsilon_{ij}\) is distributed as the standard normal distribution \(N(0,1)\). Also, let \(\gamma\) be the threshold parameters vector. The conditional probability of \(y_{ij}\) given \(\eta_{ij}\) and \(\gamma\) is thus obtained as follows:

\[
\begin{align*}
Pr(y_{ij}=1 | \eta_{ij}, \gamma) &= Pr(\varepsilon_{ij} < w_{ij} < \gamma) \\
Pr(y_{ij}=2 | \eta_{ij}, \gamma) &= Pr(\varepsilon_{ij} < \gamma - \eta_{ij}) \\
&\vdots \\
Pr(y_{ij}=m | \eta_{ij}, \gamma) &= \Phi(\gamma - \eta_{ij}) - \Phi(\gamma_{m-1} - \eta_{ij})
\end{align*}
\]

where \(\Phi()\) is the cumulative probability function of \(N(0,1)\). This conditional probability is called as the probit link that is a function of the unknown marker effects and threshold parameters.

Estimation Methods for Marker Effects

It is challenging to estimate all the parameters in the GS Model 1 or 2, because that the number of the additive effects in Model 1 or both the additive and dominance effects in Model 2 is usually much greater than the number of phenotypic values. We thus adopt linear mixed-effects model (LMM) and three Bayesian estimation methods—Bayes A, Bayes B, and Bayes C—to tackle this problem.

Let \(\beta_A\) and \(\beta_D\) be the additive and dominance effects vectors, respectively. The LMM estimation treats \(\beta_A\) and \(\beta_D\) as two sets of random effects distributed with \(N(0,\sigma^2)\) and \(N(0,\sigma^2)\), respectively. The BLUPs for \(\beta_A\) and \(\beta_D\) are then obtained through the following Henderson’s equations (Henderson, 1984):

\[
\begin{bmatrix}
\hat{\beta}_A \\
\hat{\beta}_D
\end{bmatrix} =
\begin{bmatrix}
I_p - \lambda_A I_n & \lambda_A X_p \\
X_p' & \lambda_A X_p' X_p + \lambda_D I_n
\end{bmatrix}^{-1}
\begin{bmatrix}
I_p X_p' y \\
X_p' y
\end{bmatrix}
\]

where \(I\) is the identity matrix of order \(p\) (here \(p\) is the number of markers), and \(1_n\) the unit vector of length \(n\) (here \(n\) is the number of the hybrids). The regularization parameters are given by \(\lambda_A = \sigma_A^2 / \sigma^2\) and \(\lambda_D = \sigma_D^2 / \sigma^2\). The variance components \(\sigma_A^2\), \(\sigma_D^2\), and \(\sigma^2\) are estimated using restricted maximum likelihood estimation (REML). This method can be performed with the built-in function `mmre()` in R package sommer (Covarrubias-Pazaran, 2016).

Bayes A implements three hierarchical priors. It is assumed that \(\beta_A \sim N(0, G_A)\) and \(\beta_D \sim N(0, G_D)\), where \(G_A\) and \(G_D\) are diagonal matrices with diagonal elements \(\{\sigma_A^2, \sigma_A^2, \ldots, \sigma_A^2\}\) and \(\{\sigma_D^2, \sigma_D^2, \ldots, \sigma_D^2\}\), respectively. Here \(\sigma_A^2\) and \(\sigma_D^2\) stand for the variance parameters of additive and dominance effects at locus \(l\), respectively. Second, the variance parameters \(\sigma_A^2\) and \(\sigma_D^2\) are assumed to be distributed as scaled-inverse chi-square distributions. Third, the scale parameters in the scaled-inverse chi-square distributions are further assumed to follow gamma distributions. Alternatively, Bayes B is modified from Bayes A but considers a mixture priors that a proportion \(\pi\) of marker effects is nonzero, and the remaining \(1 - \pi\) is null effect. The nonzero effects have the same priors structure as Bayes A.
Bayes C also implements the mixture priors same as Bayes B, except the variances of the nonzero effects are assumed to be homogenous. We refer to Perez and de los Campos (2014) for more details regarding these Bayesian methods and use their R package BGLR to perform data analysis.

The Bayesian estimation methods Bayes TA, Bayes TB, and Bayes TC are based on the probit link function of Eq. [3], in which their hierarchical priors are according to Bayes A, Bayes B, and Bayes C, respectively. The unknown marker effects together with the threshold parameters are estimated via Gibbs sampling using the R package BGLR.

Simulation Study

A simulation study is conducted to compare the two GS models and evaluate the four estimation methods. We use the genotype data of 142 accessions within C. maxima as a template. The 142 parental lines can produce \( \binom{142}{2} = 10,011 \) possible hybrids. We first perform linkage disequilibrium (LD) decay analysis on the original 4521 SNPs. The detail regarding the LD decay analysis can be found in Supplemental information. From the LD decay analysis result, we determine 1174 LD blocks over the genome. We thus arbitrarily fix the number of QTL as 1174 for the 10,011 hybrids. The sample variance for the TBVs is calculated as the estimate for genetic variance that is fixed at 0.3, 0.5, and 0.7. The observed values of heritability are calculated as 0.299, 0.500, and 0.699 from the 10,011 simulated data.

Step 3: Determine the training population and build the GS models. For each simulated data set, we randomly choose 150 hybrids from the 10,011 hybrids as the training population to build the two GS models by the four estimation methods. Note that the marker matrix consists of the original 4521 SNPs in the GS model building.

Step 4: Predict GEBVs for the individuals in the test population and calculate Pearson’s correlation between the GEBVs and simulated phenotypic values. The remaining 9861 hybrids not selected in the training population in Step 3 serve as the test population. The GEBVs are exactly the fitted values from the two GS models.

There are nine simulation scenarios (three different cases for true dominance effects by three different values of \( h^2 \)). The procedure of Steps 2 to 4 is repeated 100 times for each simulation scenario. The average of the resulting 100 Pearson’s correlation coefficients will be reported to measure prediction accuracy for the simulation scenario.

Pumpkin Data Analysis

We analyze the pumpkin data for each of C. maxima and C. moschata intraspecific groups with the following steps:

Step 1: Evaluate the GS models and estimation methods through data cross-validation. We employ the leave-one-out procedure, in which the two GS models are built using \( n-1 \) hybrids and validated with the remaining hybrid. This process is repeated \( n \) times such that all the hybrids have been validated individually. For the continuous trait, Pearson’s correlation is still used as a measure of prediction accuracy. However, Pearson's correlation is unable to exactly measure the prediction accuracy for ordinal traits because the phenotypic values are discrete levels. Here we propose an index to evaluate the cross-validation. From the probit link function of Eq. [3], we simply replace the latent value \( Y_i \) with its GEBV to yield the estimated prediction probability of the test hybrid that falls into each ordinal category. The test hybrid is thus designated to be the category with the highest prediction probability. The correct prediction probability is defined as the probability of correct prediction for the \( n \) hybrids.

Step 2: Derive SCA for each hybrid and GCAs for its parental lines to quantify the degree of MPH and BPH of the hybrid combination. Let SCA(\( ij \)) be the SCA for hybrid \( P_iP_j \). Also, let GCA(\( ij \)) and GCA(\( ij \)) be the GCAs for its parental lines \( i \) and \( j \), respectively. From Werner et al. (2018), the A+D model of Eq. [2] can be rewritten as follows:

\[
y_{ij} = \mu + GCA_{(i)} + GCA_{(j)} + SCA_{(ij)} + e_{ij}
\]

where \( GCA_{(i)} = \sum_{l=1}^{p} g_{il}\beta_{il}^{(0)} \), \( GCA_{(j)} = \sum_{l=1}^{p} g_{jl}\beta_{jl}^{(0)} \) and \( SCA_{(ij)} = \sum_{l=1}^{p} g_{il}g_{jl}\beta_{iljl}^{(0)} \) (the dominance effects part). Here \( \beta_{il}^{(0)} \) is coded as ~0.5, 0, or 0.5 if \( M_iM_i \), \( M_iM_j \), or \( M_iM_j \) occurs at locus \( l \) for parental line \( i \) and similarly \( \beta_{jl}^{(0)} \) is for parental line \( j \), so that \( GCA_{(i)} + GCA_{(j)} = \sum_{l=1}^{p} g_{il}g_{jl}\beta_{iljl}^{(0)} \) (the additive effects part).
Moreover, let GEBV_{hyb}, GEBV_{pa1}, and GEBV_{pa2} be the GEBVs for the hybrid and its two parental lines. Then, from Model 4, there exist the following relationships:

\[ E[\text{GEBV}_{hyb}] = \mu + GCA_{pa1} + GCA_{pa2} + SCA_{pa1} + SCA_{pa2} \]
\[ E[\text{GEBV}_{pa1}] = \mu + 2GCA_{pa1} \]
\[ E[\text{GEBV}_{pa2}] = \mu + 2GCA_{pa2} \]

where \( E[\cdot] \) denotes the expectation operator. Furthermore, from the definition of MPH and BPH for a hybrid combination, we have the following:

\[ \text{MPH}_{hyb} = E[\text{GEBV}_{hyb}] - \frac{E[\text{GEBV}_{pa1}] + E[\text{GEBV}_{pa2}]}{2} = SCA_{pa1} \]
\[ \text{BPH}_{hyb} = E[\text{GEBV}_{hyb}] - \max \{ E[\text{GEBV}_{pa1}], E[\text{GEBV}_{pa2}] \} = SCA_{pa2} - |GCA_{pa1} - GCA_{pa2}| \]

where \( |GCA_{pa1} - GCA_{pa2}| \) denotes the absolute value of \( GCA_{pa1} - GCA_{pa2} \). So the degree of MPH for a hybrid can be directly quantified by its GEBV-based SCA, and the degree of BPH must be measured by not only its GEBV-based SCA but also its parents’ GCAs. Under the positive heterosis assumption, the value of MPH or BPH is larger, and the heterosis of the hybrid is stronger.

Step 3: Estimate the variance components and heritability for a continuous trait. Based on Model 4, we treat all the GCA effects as a set of random effects and all the SCA effects as another set of random effects. Then, we express the genomic BLUP (g-BLUP) model as the following matrix form:

\[ y_t = 1_n \mu + Z_{GCA} u_{GCA} + Z_{SCA} u_{SCA} + e_t \]  \[ \quad [5] \]

where \( y_t \) denotes the phenotypic values vector; \( 1_n \) is the unit vector of length \( n \) (here \( n \) is the number of the hybrids); \( \mu \) is the constant term; \( u_{GCA} \) and \( u_{SCA} \) are vectors of GCA and SCA effects, respectively; \( Z_{GCA} \) and \( Z_{SCA} \) are incidence matrices of GCA and SCA effects, respectively; and \( e_t \) is the vector of residuals following the normal distribution \( N(0, \sigma_e^2) \). It is assumed that \( u_{GCA} \sim N(0, A_{GCA} \sigma_{GCA}^2) \) and \( u_{SCA} \sim N(0, D_{SCA} \sigma_{SCA}^2) \), where \( A \) and \( D \) denote the additive and dominance genomic relationship matrices, respectively. The \( A \) and \( D \) matrices are calculated using the methods presented in Endelman and Jannink (2012) and Su et al. (2012), respectively. We use the sommer package (Covarrubias-Pazaran, 2016) to obtain the REML estimates for the variance components \( \sigma_{GCA}^2 \), \( \sigma_{SCA}^2 \), and \( \sigma_e^2 \). Subsequently, the heritability is obtained as the ratio of genetic variance \( \sigma_{GCA}^2 + \sigma_{SCA}^2 \) to the total phenotypic variance \( \sigma_{GCA}^2 + \sigma_{SCA}^2 + \sigma_e^2 \).

### RESULTS

#### Simulation Study

The averages of the 100 simulated Pearson’s correlation coefficients estimated from the four estimation methods for the nine simulation scenarios based on the two GS models are displayed in Fig. 1. The A model is found to have slightly better prediction accuracy than the A+D model under Case 1 of the null dominance effects situation. This could be due to the redundant dominance effect parameters in the A+D model. However, the A+D model indeed appears to have better prediction accuracy than the A model for both Cases 2 and 3 of the mediate and strong dominance effects situations. The difference of the Pearson’s correlation averages between the two GS models increases as the true dominance effects get stronger. Interestingly, the Pearson’s correlation average at a fixed heritability and GS model combination decreases as the true dominance effects get stronger. For example, at LMM, \( h^2 = 0.7 \) and the A+D model, the average of the 100 simulated Pearson’s correlation coefficients, is 0.765 for Case 1, 0.689 for Case 2, and 0.637 for Case 3. This could be due to that the bias in estimation of marker effects increases as the true dominance effects get stronger. In consequence, the A+D model could have robust prediction ability for hybrid performance, even the trait under study is mainly controlled by additive effects. Moreover, the LMM and three Bayesian estimations are found to have quite similar prediction accuracy under the same scenario and the GS model. As expected, prediction accuracy increases with the increase in heritability.

#### Pumpkin Data Analysis

### Cross-Validation

We first evaluate the two GS models and the four estimation methods through the leave-one-out cross-validation to determine the most appropriate method for each trait. The resulting Pearson’s correlation for the continuous trait FWT is displayed in Table 2. From the table, the A+D model can have slightly better prediction accuracy than the A model in the \( C. maxima \) intracrossing group. In contrast to the simulation results of Fig. 1, this might imply that the variation of trait FWT in the \( C. maxima \) intracrossing group is due to both the additive and dominance effects. On the other hand, the A model slightly outperforms the A+D model in the \( C. moschata \) intracrossing group, implying that trait FWT in this group could be almost completely controlled by the additive effects, and the dominance effects can be negligible. The Pearson’s correlation for the \( C. maxima \) group is much greater than the \( C. moschata \) group, indicating that the heritability for FWT of the \( C. maxima \) intracrossing group should be significantly larger than the \( C. moschata \) intracrossing group.

According to the ranking in Table 2, we choose Bayes C and LMM using the A+D model for trait FWT in the intracrossing groups of \( C. maxima \) and \( C. moschata \), respectively. The resulting correct prediction probability of the cross-validation is summarized in Table 3 for ordinal traits. From the ranking in the table, we simply choose Bayes TB for trait GV and Bayes TA for trait LS using the A+D model in both intracrossing groups.

#### Combining Ability and Heterosis

There are \( \text{GEBV}_{hyb} = 2 \times 10^{10} \text{ kg} \) and \( \text{GEBV}_{pa} = 6.903 \) possible hybrid combinations within the intracrossing groups of \( C. maxima \) and \( C. moschata \), respectively. The GEBV-based SCA for each hybrid and GCAs for its parental lines are calculated using the most appropriate estimation methods determined from the cross-validation results. For
Fig. 1. The averages of the 100 simulated Pearson’s correlation coefficients (prediction accuracy) under the four estimation methods (LMM, Bayes A, Bayes B, and Bayes C) and nine simulation scenarios (three different heritability [0.3, 0.5, and 0.7] crossing three cases [Case 1, null dominance effects; Case 2, mediate dominance effects; and Case 3, strong dominance effects]) using the two genomic selection models (A model and A+D model). Note that A model means additive effects model and A+D model is additive plus dominance effects model.
illustration purposes, we report only the top 25 potential hybrids with large GEBV\(_{(ij)}\) together with their associated SCA, GCA, MPH, and BPH in Supplemental Table S1, and the top 10 superior parental lines with large GEBV-based GCAs in Supplemental Table S7 for trait FWT within the \(C. \text{maxima}\) intracrossing group. The corresponding results for trait FWT within the \(C. \text{moschata}\) are summarized in Supplemental Tables S2 and S8.

From Supplemental Table S1, there is an important finding that GEBV\(_{(ij)}\) are all greater than both GEBV\(_{A}\) and GEBV\(_{A+D}\), meaning that the hybrids have superior performance over the better parental line (all the corresponding BPH\(_{(ij)}\) > 0). Besides, the top two hybrids with high phenotypic values in the training data (P026\(\times\)P236 = 3.73; P026\(\times\)P234 = 3.60) appear at the first and second rankings among the top 25 potential hybrids. The top five superior parental lines P236, P235, P026, P234, and P028 in Supplemental Table S7 are found to involve the top five potential hybrids P026\(\times\)P236, P026\(\times\)P234, P026\(\times\)P235, P026\(\times\)P027, and P026\(\times\)P028 in Supplemental Table S1. Particularly, P026 involves 10 of the top 25 potential hybrids, so it can be a top potential parental line. From Supplemental Table S2, all values for MPH\(_{(ij)}\) equal values for SCA\(_{(ij)}\), which all equal zero, implying that there is no heterosis effect for trait FWT within the \(C. \text{moschata}\) intracrossing group. Also, all the values for BPH\(_{(ij)}\) are less than zero in Supplemental Table S2, showing that every hybrid has worse performance than at least one of its parental lines. From Supplemental Table S8, P308 could be the best parental line with large FWT, and the hybrids involving P308 are expected to have better performance (Supplemental Table S2). The corresponding results for ordinal traits GV and LS within both intracrossing groups are available in Supplemental Tables S3 through S6 and Supplemental Tables S9 through S12.

**Variance Components and Heritability**

We further estimate the variance components and heritability for trait FWT within each intracrossing group using the g-BLUP model of Eq. [5]. The results are summarized in Table 4. From the table, there exist some non-negligible dominance effects in the \(C. \text{maxima}\) intracrossing group but none in the \(C. \text{moschata}\) intracrossing group. The estimates of the variance components of the GCA and SCA effects reflect the results in Table 2 that the two GS models do not have consistent prediction accuracy comparisons.

**DISCUSSION**

In this study, we have demonstrated that the additive–dominance effects model is advantageous over the additive effects model for genomic prediction of hybrid performance. Not only does it have credible prediction accuracy, but it also enables breeders to quantify the degree of MPH or BPH for individual hybrids according to GEBV-based SCAs and GCAs. In practice, we have developed a systemic procedure to carry out genomic prediction for hybrid performance. The proposed analysis method could provide useful information for breeders to select promising hybrid combinations and parental lines from a hybrid breeding program.

To explore how the relatedness between training and test populations might influence on the prediction accuracy, we further consider the following studies. Let \(T_{0}^\alpha\) be the test population and \(T_{0}^\beta\), \(T_{0}^\gamma\), and \(T_{0}^\delta\) be three exclusive validation subgroups of \(T_{0}^\alpha\). The hybrids in \(T_{0}^\alpha\), \(T_{0}^\beta\), and \(T_{0}^\gamma\) have 0, 1, and 2, respectively, of their parental lines appearing in hybrid combinations of the training population. In the simulation study, we consider Case 2 of the mediate dominance effects situation, the average

---

**Table 2.** Pearson’s correlation and ranking (in parentheses) for the cross-validation of trait fruit weight.

<table>
<thead>
<tr>
<th>Group</th>
<th>Model†‡</th>
<th>LMM‡</th>
<th>Bayes A</th>
<th>Bayes B</th>
<th>Bayes C</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C. \text{maxima})</td>
<td>A</td>
<td>0.761 (3)</td>
<td>0.761 (4)</td>
<td>0.762 (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A+D</td>
<td>0.761 (4)</td>
<td>0.774 (3)</td>
<td>0.771 (1)</td>
<td></td>
</tr>
<tr>
<td>(C. \text{moschata})</td>
<td>A</td>
<td>0.482 (1)</td>
<td>0.472 (4)</td>
<td>0.472 (2)</td>
<td>0.475 (3)</td>
</tr>
<tr>
<td></td>
<td>A+D</td>
<td>0.481 (1)</td>
<td>0.446 (4)</td>
<td>0.447 (2)</td>
<td>0.437 (3)</td>
</tr>
</tbody>
</table>

† A, additive effects model; A+D, additive plus dominance effects model.
‡ LMM, linear mixed-effects model.

**Table 3.** Correct prediction probability and ranking (in parentheses) for the cross-validation of ordinal traits growth vigor (GV) and leaf size (LS).

<table>
<thead>
<tr>
<th>Group</th>
<th>Trait</th>
<th>Model†‡</th>
<th>Bayes TA</th>
<th>Bayes TB</th>
<th>Bayes TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C. \text{maxima})</td>
<td>GV</td>
<td>A</td>
<td>0.892 (2)</td>
<td>0.901 (1)</td>
<td>0.875 (3)</td>
</tr>
<tr>
<td></td>
<td>A+D</td>
<td>0.883 (2)</td>
<td>0.892 (1)</td>
<td>0.883 (2)</td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>A</td>
<td>0.875 (1)</td>
<td>0.875 (1)</td>
<td>0.867 (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A+D</td>
<td>0.883 (1)</td>
<td>0.875 (2)</td>
<td>0.875 (2)</td>
<td></td>
</tr>
<tr>
<td>(C. \text{moschata})</td>
<td>GV</td>
<td>A</td>
<td>0.685 (1)</td>
<td>0.685 (1)</td>
<td>0.685 (1)</td>
</tr>
<tr>
<td></td>
<td>A+D</td>
<td>0.658 (3)</td>
<td>0.685 (1)</td>
<td>0.676 (2)</td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>A</td>
<td>0.711 (1)</td>
<td>0.711 (1)</td>
<td>0.711 (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A+D</td>
<td>0.711 (1)</td>
<td>0.711 (1)</td>
<td>0.711 (1)</td>
<td></td>
</tr>
</tbody>
</table>

† A, additive effects model; A+D, additive plus dominance effects model.

---

**Table 4.** Estimates of variance component and heritability for trait fruit weight (FWT) in the two intracrossing groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>(\sigma_{GCA}^2)†</th>
<th>(\sigma_{SCA}^2)‡</th>
<th>(\sigma_{E}^2)§</th>
<th>(h^2)¶</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C. \text{maxima})</td>
<td>0.033</td>
<td>0.020</td>
<td>0.073</td>
<td>0.764</td>
</tr>
<tr>
<td>(C. \text{moschata})</td>
<td>0.170</td>
<td>0.694</td>
<td>0.197</td>
<td></td>
</tr>
</tbody>
</table>

† \(\sigma_{GCA}^2\), variance component due to general combining ability (GCA).
‡ \(\sigma_{SCA}^2\), variance component due to specific combining ability (SCA).
§ \(\sigma_{E}^2\), variance component due to nongenetic effects.
¶ \(h^2\), heritability, which is calculated by the ratio of genetic variance \((\sigma_{GCA}^2 + \sigma_{SCA}^2)\) to the total phenotypic variance \((\sigma_{GCA}^2 + \sigma_{SCA}^2 + \sigma_{E}^2)\).
prediction accuracy for the validation subgroups are displayed in Table 5. In the pumpkin data analysis, we randomly choose one quarter of observations of trait FWT as a training population within *C. maxima* and *C. moschata* intracrossing groups, the remaining observations not selected in the training population serve as the test population. The procedure was repeated 30 times and the average accuracy for the validation subgroups are also displayed in Table 5. As expected, the prediction accuracy of the three validation subgroups for all scenarios under study is ranked as $T_0 < T_1 < T_2$, which reflects the relatedness and indicates the more precise prediction accuracy among the different validation subgroups. The simulation studies (Technow et al., 2012) and empirical data analysis in maize (Technow et al., 2014) and wheat (Zhao et al., 2013; Kadam et al., 2016) also lead to the consistent results.

There are several factors that might affect the accuracy of genomic prediction such as heritability, marker density, estimation methods for marker effects, training population size, and genomic relationship between training and test populations (Desta and Ortiz, 2014; Zhao et al., 2015; Wang et al., 2018). In this study, we have shown that the prediction accuracy could increase with the increase in heritability and in the degree of relatedness between training and test populations. In addition, we also evaluate some shrinkage estimation methods such as ridge regression, least absolute shrinkage and selection operator (LASSO), and elastic net methods through the same simulation study. The averages of the 100 simulated Pearson’s correlation coefficients for these shrinkage estimation methods together with the LMM and three Bayesian methods under Case 2 of the mediate dominance effects situation are displayed in Fig. 2. It is shown that ridge regression appears to have similar prediction accuracy with the LMM and three Bayesian methods. However, LASSO and elastic net methods seem not to perform well. This could be due to the fact that the $\ell_1$ norm penalty specified for both LASSO and elastic net methods often leads to a sparse solution with just few nonzero marker effects in the model (Wimmer et al., 2013) and their performances are nonsufficient in genomic prediction.

We now investigate how the marker density may affect the prediction accuracy. As mentioned earlier, the numbers of SNP markers used in the analysis for *C. maxima* and *C. moschata* are given by 4521 and 6348 SNPs.

---

**Table 5.** The average prediction accuracy for $T_0$, $T_1$, $T_2$, and $T_{all}$ validation groups in the simulation study and pumpkin data analysis based on the A+D model and Bayes A for the simulation study, Bayes C for *C. maxima*, and linear mixed-effects model for *C. moschata*. Note that the A+D model is the additive plus dominance effects model.

<table>
<thead>
<tr>
<th>Validation group</th>
<th>$h^2$</th>
<th>$T_0$</th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_{all}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation study</td>
<td>0.3</td>
<td>0.147</td>
<td>0.292</td>
<td>0.394</td>
<td>0.377</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.274</td>
<td>0.451</td>
<td>0.565</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.408</td>
<td>0.611</td>
<td>0.717</td>
<td>0.701</td>
</tr>
<tr>
<td>Pumpkin data analysis</td>
<td><em>C. maxima</em></td>
<td>0.636</td>
<td>0.731</td>
<td>0.773</td>
<td>0.728</td>
</tr>
<tr>
<td></td>
<td><em>C. moschata</em></td>
<td>0.052</td>
<td>0.350</td>
<td>0.392</td>
<td>0.331</td>
</tr>
</tbody>
</table>

† $T_0$, no parent of hybrids appears in the combinations of the training population; $T_1$, only one parent of hybrids appears in the combinations of the training population; $T_2$, both parents of hybrids appear in the combinations of the training population; $T_{all}$, all hybrids of the test population.

---

**Fig. 2.** The averages of the 100 simulated Pearson’s correlation coefficients under Case 2 of the mediate dominance effects situation using the A+D model. Note that A+D model is additive plus dominance effects model.
respectively. We fit trait FWT by different marker densities, which are determined according to various numbers of SNPs evenly spaced over each chromosome. We also collected phenotypic values of 51 hybrids of the intercrossing group between *C. maxima* and *C. moschata*. There were 16,275 SNP markers collected for such a more diverse intercrossing group. Similarly, we fit trait FWT by different marker densities for this intercrossing group using Bayes A. The prediction accuracy for the two intra- and the intercrossing groups are displayed in Fig. 3. From the figure, the prediction accuracy generally appears to improve as the marker density increases from the beginning, which could enhance the chance of marker–QTL associations to capture more accurate marker effects (Desta and Ortiz, 2014; Norman et al., 2018). However, the prediction accuracy gradually reaches the plateau as marker density increases.
exceeds 1000, 2000, and 4000 SNPs for the *C. maxima* intracrossing group, the *C. moschata* intracrossing group, and their intercrossing group, respectively. This indicates that more diverse populations might need more markers to achieve better performance. However, the prediction accuracy tends to slightly decrease as the number of SNPs gets greater than 8000 for the intercrossing group. This could be due to the model overfitting caused from redundant markers (Hickey et al., 2014).

Although the use of ordinal score as a measure can save the cost for phenotyping, this could lead to a loss of information in data analysis. It has been verified that the increase in the number of ordinal levels can enhance prediction accuracy of a GS model (Kizilkaya et al., 2014). Thus, one should increase the number of levels as many as possible when using the ordinal score as a measure. As discussed earlier, Pearson’s correlation could be an unsuitable index for assessing the accuracy of genomic prediction for ordinal traits. There is still room for improvement to develop methods for evaluating genomic prediction for ordinal traits. The ranking method proposed by Blondel et al. (2015) could be a promising one.

There is often a large number of possible hybrid combinations needed to be evaluated in a hybrid breeding program. Namely, the training population size can be much smaller than the test population size. Thus, it is desired to investigate how the size of training population may affect the prediction accuracy. Under Case 2 of the mediate dominance effects situation in the simulation study, we display the prediction accuracy using Bayes C estimation for the training population size ranging from 10 to 1000 in Fig. 4. From the figure, the prediction accuracy gradually reaches the plateau as the training population size exceeds 300. As suggested by Akdemir et al. (2015), determination of an optimal training population could be an economical and efficient way to a successful GS with the limited resources. This will be an issue for future study.

**Conflict of Interest**

The authors declare that there is no conflict of interest.

**Supplemental Material Available**

Supplemental material is available online for this article.

**ACKNOWLEDGMENTS**

This research was supported by the Ministry of Science and Technology Taiwan (grant number: MOST 106-2118-M-002-002-MY2).

**REFERENCES**


