Identify Superior Parental Lines for Biparental Crossing via Genomic Prediction: Rice as an Example

Ping-Yuan Chung
National Taiwan University

Chen-Tuo Liao (ctliao@ntu.edu.tw)
Southern Taiwan University https://orcid.org/0000-0001-9777-3701

Original article

Keywords: genomic prediction, genomic selection, mixed effects model, rice breeding

Posted Date: August 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-53364/v1

Identify Superior Parental Lines for Biparental Crossing via Genomic Prediction: Rice as an Example

Ping-Yuan Chung and Chen-Tuo Liao *

Affiliations: PY Chung and CT Liao, Department of Agronomy, National Taiwan University, Taipei, Taiwan. *Corresponding author (ctliao@ntu.edu.tw)

ABSTRACT

Background: A set of superior parental lines is the key to high-performing recombinant inbred lines (RILs) for biparental crossing in a rice breeding program. The number of possible crosses in such a breeding program is often far greater than the number that breeders can handle in the field. A practical parental selection method via genomic prediction (GP) is therefore developed to help breeders identify a set of superior parental lines from a candidate population before field trials.

Results: The parental selection via GP often involves truncation selection, selecting the top fraction of accessions based on their genomic estimated breeding values (GEBVs). However, the truncation selection inevitably causes a loss of genomic diversity in the breeding population. To preserve genomic variation, the selection of closely related accessions should be avoided. We first proposed a new index to quantify the genomic diversity for a set of candidate accessions. Then, we compared the performance of three classes of strategy for the parental selection, including those consider (a) GEBV only, (b) genomic diversity only, and (c) both GEBV and genomic diversity. We analyzed two rice (*Oryza sativa* L.) genome datasets for the comparison. The results show that the strategies considering both GEBV and genomic diversity have the best or second-best performance for all the traits analyzed in this study.

Conclusion: Combining GP with Monte Carlo simulation can be a useful means of parental selection for rice pre-breeding programs. Different strategies can be implemented to identify a set of superior parental lines from a candidate population. In consequence, the strategies considering both GEBV and genomic diversity that can balance the starting GEBV average with maintenance of genomic diversity should be

recommended for practical use.

Keywords: genomic prediction, genomic selection, mixed effects model, rice breeding.

BACKGROUND

Biparental crossing is a commonly used scheme in pure-line breeding for self-pollinated crops such as rice, wheat (*Triticum aestivum* L.), soybean [*Glycine max* (L.) Merr.] and oat (*Avena sativa* L.). Plant breeders cross two inbred parental lines to produce F_1 population, then a subset of diverse individuals of the F_2 population is selected to produce potential RILs after several generations of selfing. Obviously, the parental lines play a fundamental role in the line development and significantly affect the performance of the resulting RILs. However, the identification of superior parental lines from germplasm collections for creating genetic variation to maximize selection response in subsequent cycles is still a challenge for plant breeders (Bernardo 2003; Witcombe et al. 2013). Another practical concern is that the number of possible crosses in such a breeding program is often far greater than the number that breeders can handle in the field. Therefore, it should be of great help to breeders if a limited number of superior parents can be identified before the field trial.

Genomic selection based on the statistical method of GP has been used to improve breeding efficiency in dairy cattle (Hayes et al. 2009) and a variety of crops (Massman et al. 2013; Asoro et al. 2011; Heffner et al. 2011; Lorenz et al. 2012; Spindel et al. 2015). The main concept of GP is to capture all the effects of quantitative trait loci (QTLs) by using dense DNA markers over the whole genome, assuming that the DNA markers are in strong linkage disequilibrium with one or more QTLs (Meuwissen et al. 2001). The most commonly used DNA markers are single nucleotide polymorphisms (SNPs). A GP model is first built using the phenotype and genotype data of a training population. Then, GEBVs for the candidate individuals with known genotype data are predicted through the resulting GP model. There are two kinds of mixed linear model methods are widely employed to obtain the GEBVs: (i) best linear unbiased prediction (BLUP) based on markers and (ii) BLUP based on a genomic relationship matrix. For the BLUP of (i), the marker effects are treated as random effects and the GEBVs of individuals are calculated by multiplying their marker scores by these BLUP estimates. Ridge regression BLUP (rr-BLUP) method (Meuwissen et al. 2001; Piepho 2009) follows this approach. For the BLUP of (ii), the genotypic values of individuals are treated as random effects and estimated through a genomic relationship matrix. The genomic BLUP (GBLUP) method (Habier et al. 2007; VanRaden 2008) follows this approach. For more details regarding the GP models and the estimation methods used for their model parameters, refer to Xavier et al. (2016).

Gaynor et al. (2017) proposed a two-part strategy for implementing genomic selection for line development, addressing the two components: (i) a product development component, to identify inbred lines either for hybrid parent development or cultivar release; (ii) a population improvement, to increase the frequency of favorable alleles through rapid recurrent genomic selection. Conducting a stochastic simulation, they showed that programs using the two-part strategy generated up to 2.5 times more genetic gain than conventional programs, and up to 1.5 times more genetic gain than the best performing standard genomic selection strategy. Also, Yao et al. (2018) combined GP with Monte Carlo simulation to select superior parents in wheat breeding before the field trial. They used the criterion of usefulness function on a selection index, incorporating yield and two quality traits, to evaluate a cross. Their usefulness function took into account both the mean genetic value and genetic variance of progeny populations. Yao et al. (2018) simulated the required progeny populations using the R/qtl package (Broman et al. 2003), and calculated their usefulness function estimates. It was concluded that the use of the usefulness function for parental selection resulted in higher genetic gain than the use of mid-parent GEBV, implying that the strategy for the parental selection cannot only consider GEBVs of the candidate accessions.

Selecting the parental lines with the highest GEBVs (truncation selection), breeders hope to maximally pass favorable properties of the parental lines on to their progeny populations. However, several favorable QTLs can risk being eliminated from the breeding population using the truncation selection (Vanavermaete et al. 2020). We therefore take both GEBV and genomic diversity into account for identifying superior parents in a biparental crossing program. For a specific target trait, we construct a

GBLUP model to predict the GEBVs for the candidate accessions. Furthermore, we propose a new index to quantify the genomic diversity for a set of candidate accessions according to the GBLUP model. We simulate the genotype data for progeny populations over successive generations derived from a cross between two parental lines. The GEBVs of the progeny populations are then predicted by the trained GBLUP model. We further make generation advancement decisions according to the resulting GEBVs. Finally, we assess a set of parental lines based on their F_{10} RILs which are assumed to be a fixed population. Several selection strategies are evaluated within two rice genome datasets.

MATERIALS AND METHODS

The Rice Genome Datasets

Dataset I: We first used the rice genome dataset presented in Zhao et al. (2011) to illustrate our proposed procedure. This dataset was originally collected for genome-wide association study (GWAS). The dataset contains 44,100 SNP variants and 36 traits of 413 *O. sativa* accessions, and has a strong subpopulation structure containing six different groups. We deleted any SNPs with a missing rate of > 0.05 and a minor allele frequency of < 0.05. To reduce redundant collinearity in calculation of the genomic relationship matrix, we only retained about one-third of the SNPs which are evenly distributed over each chromosome. We then imputed a missing SNP marker from its corresponding major homozygous alleles. The final marker matrix consists of 413 accessions and 11,047 SNPs. We here analyzed the six traits: brown rice seed width (BRSW), florets per panicle (FPP), flowering time at Arkansas (FTAA), flowering time at Faridpur (FTAF), plant height (PH), and panicle number per plant (PNPP).

Dataset II: We further analyzed the rice genome dataset presented in Spindle et al. (2015), which was collected for genomic selection study. The dataset contains 73,147 SNP variants and 363 elite breeding lines belonging to *indica* or *indica-admixed* group. The phenotype data include the four years (2009-2012), two seasons per year (dry and wet), of grain yield (YLD), flowering time (FT), and plant height (PH). Note that the PH data in 2009 wet season are not available. The adjusted means for 328 out of the 363 individuals and 10,772 out of the 73,147 SNP markers were used for this study. We here chose one marker every 0.1cM over each chromosome.

Monte Carlo Simulation for the Genotype of Progeny Populations

To simulate the genotype data for progeny populations, we used Gramene Annotated Nipponbare Sequence (Youens-Clark et al. 2011) to estimate recombination rates between two adjacent SNPs. The Gramene Annotated Nipponbare Sequence database contains both the physical and linkage distances between SNPs, which can be downloaded from http://archive.gramene.org. The genetic positions of the SNPs are estimated via linear interpolation between the two markers flanking each SNP. Once the genetic positions were obtained, the recombination rates between adjacent SNPs were estimated via Haldane's mapping function (Haldane 1919):

$$r_{AB} = \frac{1}{2}(1 - e^{-2X_{AB}}),$$

where r_{AB} is the recombination rate and X_{AB} is the linkage distance between SNP markers A and B. Through a series of Bernoulli distributions and the estimated recombination rates, the crossover of each chromosome was simulated to yield the sequence of a gamete, then two gametes were paired to produce the genotype data for the progeny.

GBLUP Model

We considered the following single-trait GBLUP model for GP:

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{g} + \boldsymbol{e}, [1]$$

where $\mathbf{y}$ denotes the vector of phenotypic values of a training population with n individuals; μ is a constant term; $\mathbf{1}_n$ is the vector of order n with all elements equal to 1; $\boldsymbol{g}$ stands for the vector of genotypic values and $\boldsymbol{e}$ is the vector of random errors. It is assumed that $\boldsymbol{g}$ follows a multi-variate normal distribution $\text{MVN}(\mathbf{0}, \sigma_g^2 \boldsymbol{K})$, where $\mathbf{0}$

is a zero vector; σ_g^2 is the genetic variance of additive effects and $\boldsymbol{K}$ is a genomic relationship matrix among the individuals. Furthermore, $\boldsymbol{e}$ follows $\text{MVN}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_n)$, where σ_e^2 is the random error variance and $\boldsymbol{I}_n$ denotes the identity matrix of order n. Here, $\boldsymbol{g}$ and $\boldsymbol{e}$ are assumed to be mutually independent. In this study, we considered the genomic relationship matrix $\boldsymbol{K} = \boldsymbol{M}\boldsymbol{M}^T/p$, where $\boldsymbol{M}$ is the marker score matrix and p is the number of SNP markers. The elements of $\boldsymbol{M}$ are coded as –1, 0, and 1 for the minor homozygous alleles (A_1A_1), the heterozygous alleles (A_1A_2), and the major homozygous alleles (A_2A_2), respectively. The model parameters of the GBLUP model can be estimated through Henderson's equations (Henderson 1984), given by:

$$\begin{bmatrix} n & \boldsymbol{1}_n^T \\ \boldsymbol{1}_n & \boldsymbol{I}_n + \lambda \boldsymbol{K}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}_n^T \boldsymbol{y} \\ \boldsymbol{y} \end{bmatrix}, [2]$$

where the regularization parameter λ is given by $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$. We used the R function mmer( ) in the R package sommer (Covarrubias-Pazaran 2016) to obtain the restricted maximum likelihood estimates (REMLs) for the two variance components of σ_g^2 and σ_e^2, and then plugged the resulting estimates into Eq. [2] to get $\hat{\mu}$ and $\hat{\boldsymbol{g}}$.

Let $\hat{\boldsymbol{g}}_{bp}$ be the vector of estimated genotypic values for a breeding population and $\boldsymbol{K}_{bp}$ be the genomic relationship matrix between the breeding population and the training population. In the case, we have:

$$\hat{\boldsymbol{g}}_{bp} = \boldsymbol{K}_{bp} \boldsymbol{K}^{-1} \hat{\boldsymbol{g}}.$$

The GEBV for the breeding population is $\hat{\boldsymbol{g}}_{bp}$ plus the estimate of the constant term $\hat{\mu}$.

The Index to Quantify Genomic Diversity

Let $\boldsymbol{g}_0$ be the vector of genotypic values and $\boldsymbol{K}_0$ be the genomic relationship matrix for a particular set of accessions with size n_0. According to the GBLUP model of Eq. [1], the covariance matrix for $\boldsymbol{g}_0$ is given by:

$$\text{Var}(\boldsymbol{g}_0) = \sigma_g^2 \boldsymbol{K}_0.$$

The determinant of the covariance matrix represents the overall variability for the genotypic values, which is calculated as:

$$|\mathrm{Var}(\boldsymbol{g}_0)| = (\sigma_g^2)^{n_0}|\boldsymbol{K}_0|. \ [3]$$

Clearly, the determinant of Eq. [3] is proportional to the *D*-score defined below:

$$D\text{-score} = |\boldsymbol{K}_0|. \ [4]$$

The *D*-score of Eq. [4] ranges from 0 to 1. For a fixed number of n_0, a subset of accessions chosen from a breeding population that achieves the maximal D-score will have greater genomic diversity than the competing choices with size n_0. The concept of the *D*-score is adopted from optimum experimental designs (Atkinson and Donev 1992). A simple example is given to illustrate the *D*-score. Suppose that there are $n = 3$ accessions in the candidate set with the genomic relationship matrix:

$$\boldsymbol{K} = \begin{bmatrix} 1 & 0.7 & 0.5 \\ 0.7 & 1 & 0.3 \\ 0.5 & 0.3 & 1 \end{bmatrix}.$$

For $n_0 = 2$, the *D*-score for g_1 and g_2 is calculated as $|\boldsymbol{K}_0| = \begin{vmatrix} 1 & 0.7 \\ 0.7 & 1 \end{vmatrix}$=0.51. Similarly, the *D*-scores for g_1 and g_3, and for g_2 and g_3 are given by 0.75 and 0.91, respectively. Clearly, the two accessions with g_2 and g_3 have greater genomic variation (smaller genomic correlation) than the other competing choices. A set of accessions with the maximal D-score can avoid the selection of closely related individuals.

An Algorithm to Search for Accessions with the Maximal *D*-Score

We required a highly efficient algorithm to search for a subset of accessions within a candidate population so that it can achieve the maximal *D*-score. We used a genetic algorithm to complete this task, which is an exchange algorithm with the three different operators: roulette wheel selection, crossover, and mutation (Whitley 1994). For a given candidate set S_c with n_c accessions, we searched for an optimal subset S_0 with n_0 individuals from S_c. Our algorithm began with a set of m random solutions, each of which is a vector of 0 or 1 with a length equal to n_c. The number of values with a score of 1 in the vector is equal to n_0, corresponding to the chosen accessions at the current stage. Here, we fixed $m = n_0$. We then obtained the elite solutions from the initial m random solutions after a large number of iterations, where each iteration repeated all the three operators. We stopped the algorithm when the maximal *D*-score among the current elite solutions converged.

The Procedure for Selecting Parental Lines

To evaluate a variety of strategies in determining parental lines, we carried out the following steps.

Step 1: For a specific target trait, we used all of the phenotypic values available from the rice genome dataset to build the corresponding GBLUP model of Eq. [1].

Step 2: We predicted the GEBVs for all of the accessions in the dataset through the trained GBLUP model developed in Step 1. Seven strategies were used to select a subset of 10 parental lines according to their GEBVs: (i) the GEBV only (GEBV-O) approach, which chose the top 10 accessions (either maximal or minimal); the genomic diversity only (GD-O) approaches: (ii) GD-O-30, (iii) GD-O-50, and (iv) GD-O-100, which applied the genetic exchange algorithm to search for an optimal subset of 10 accessions from each of the three candidate sets composed of the top 30, 50, and 100 accessions, respectively, such that the chosen subset had the maximal *D*-score; and the approaches (GEBV-GD) considering both GEBV and genomic diversity: (v) GEBV-GD-30, (vi) GEBV-GD-50, and (vii) GEBV-GD-100, which retained the top two accessions, then applied the genetic exchange algorithm to search for another eight accessions from the remainder of each candidate set for GD-O-30, GD-O-50, and GD-O-100, respectively, so that the resulting 10 accessions had the maximal *D*-score.

Step 3: For each subset of 10 accessions determined by the seven strategies, we crossed any two parental lines to produce 45 F_1 hybrids. Here, we started to simulate the genotype data for successive generations of progeny populations through the Monte Carlo simulation. Each of the 45 F_1 hybrids produced 60 individuals of the F_2 population by self-pollination, resulting in 2700 F_2 individuals. After obtaining the GEBVs for the 2700 F_2 individuals via the trained GBLUP model of Step 1, we then retained the top 45 F_2 individuals. Again, we used these 45 F_2 individuals to produce 2700 F_3 individuals (each F_2 individual produced 60 F_3 individuals) and retained the top 45 F_3 individuals. We then repeated the same procedure to produce 2700 F_{10} individuals which are assumed to be a fixed population.

Step 4: For the resulting 2700 F_{10} individuals generated according to each strategy, we found the best F_{10} RIL with the top GEBV.

A flowchart of the procedure is displayed in Figure 1. We repeated this analysis procedure 30 times to obtain the best F_{10} RILs from each repetition for each strategy. The average of the GEBVs for the best F_{10} RILs was then calculated and used as the measure of efficiency for the strategy. Note that for the traits of BRSW, FPP, and PNPP in Dataset I; and YLD in Dataset II, larger GEBVs are preferable (i.e., these traits follow the rule that the larger, the better). The remaining five traits of FTAA, FTAF, and PH in Dataset I; and FT, and PH in Dataset II are those for which the rule is “the smaller, the better”.

Calculation of Genetic Gain

To gain an understanding of the genetic improvement on a target trait using different strategies, we estimated genetic gain as

$$\text{genetic gain} = \overline{GEBV}_{F_{10}} - \overline{GEBV}_{P}, \text{ [5]}$$

where $\overline{GEBV}_{F_{10}}$ denotes the GEBV average among the resulting 2700 F_{10} RILs and $\overline{GEBV}_{P}$ denotes the GEBV average among the 10 selected parental lines for each strategy (Rutkoski 2019). The larger absolute value of the genetic gain indicates the more improvement on the target trait.

RESULTS

Strategies Comparison Based on the best F_{10} RILs

The GEBV averages of the best F_{10} RILs from the 30 repetitions using each of the seven strategies are displayed in Tables 1 and 2 for the two datasets. The results in the tables show that the strategies considering both GEBV and genomic diversity (GEBV-GD-30, -50, -100) generally have satisfactory efficiency, because they achieve the best or second-best performance for all the traits. Therefore, this kind of strategies could be a reliable means of determining the parental lines. On the other hand, the strategies accounting for genomic diversity only (GD-O-30, -50, -100) don't have satisfactory efficiency for all the traits, with the exception of GD-O-100 for YLD in Dataset II. For the strategy based on GEBV only, the GEBV-O has the best or second-best performance for FPP, and PH in Dataset I; and PH, and FT in Dataset II, but also has the worst or second-worst performance for the remaining four traits in Dataset I and YLD in Dataset II. Thus, the GEBV-O could be a high-risk strategy.

We also displayed the GEBV averages with the plus and minus one unit of their corresponding standard deviations for the best individuals from the 30 repetitions over consecutive generations in Figures 2 and 3. From the figures, the four strategies of GEBV-O, GEBV-GD-30, -50, -100 selected the same best individual from the 30 repetitions at parental generation, and also at F_1 generation, so there is no standard deviation shown with the corresponding GEBV averages. The GEBV averages of the best selected parental lines by the strategies can be ranked as GEBV-O = GEBV-GD-30 = GEBV-GD-50 = GEBV-GD-100 > GD-O-30 > GD-O-50 > GD-O-100 in decreasing desirability. The desirability at parental generation decreases as the degree of diversity increases for the three strategies considering the genomic diversity only. Also, the desirability declines from parental generation to F_1 generation for every strategy, due to the heterogenous alleles in F_1 hybrids.

To explore the extent to which the top two accessions contribute to the subset of ten parental lines determined by the four strategies of GEBV-O, GEBV-GD-30, -50, -100, we compared each subset with a reduced group consisting of F_1 hybrids whose parental lines contain at least one of the top two accessions for each subset. Every reduced group consists of 17 F_1 hybrids. Similarly, we followed the analysis procedure to obtain the GEBV averages for the best F_{10} RILs from 30 repetitions based on the reduced group. The results are displayed in Table 3 with the corresponding GEBV averages based on the group of the original 45 F_1 hybrids. From the table, there is no practical significant difference between these two groups for all the traits using the four strategies.

Genetic Gains for the Strategies

The average among the genetic gains on a target trait for each strategy calculated by Eq. [5] from the 30 repetitions is displayed in Tables 4 and 5 for Datasets I and II, respectively. It is reasonable to compare the performance of the strategies according to the endpoint of $\overline{GEBV}_{F_{10}}$. From the tables, we found that the comparison results based on $\overline{GEBV}_{F_{10}}$ are consistent with the above results based on the best F_{10} RILs. Also, the strategies considering genomic diversity (GD-O-30, -50, -100; GEBV-GD-30, -50, -100) have greater genetic gain than the GEBV-O for all the traits except PH in Dataset I (Table 4). As expected, the genetic gain usually increases with the increase of the genomic diversity (GD-O-100 outperforms both GD-O-50 and GD-O-30 for all the traits except BRSW, and FTAF in Dataset I; GEBV-GD-100 outperforms both GEBV-GD-50 and GEBV-GD-30 for all the traits). In addition, GEBV-O has the best $\overline{GEBV}_P$; GEBV-GD-30 has better $\overline{GEBV}_P$ than GD-O-30; GEBV-GD-50 has better $\overline{GEBV}_P$ than GD-O-50 and GEBV-GD-100 has better $\overline{GEBV}_P$ than GD-O-100 for all the traits. Namely, a strategy has a relatively good starting point as it considers more degree of GEBV.

DISCUSSION

From the results for comparing the proposed strategies, those considering both GEBV and genomic diversity or considering GEBV only can be recommended for practical use. Furthermore, from the results for exploring the extent to which the top two accessions contribute to the parental lines determined by the four strategies of GEBV-O, GEBV-GD-30, -50, -100, we have the conclusion: the economical strategies with 17 F_1

hybrids whose parental lines contain at least one of the top two accessions for each selected subset can be a practical alternative to those with 45 F_1 hybrids composed of all of the possible crosses.

From Tables 4 and 5, the strategies considering genomic diversity only (GD-O-30, -50, -100) generally have greater genetic gain, mainly due to their more genomic variation but less favorable $\overline{GEBV_P}$, so they have more room to improve. Also, the GEBV-O has the best starting $\overline{GEBV_P}$ but the least genomic diversity in the base population, so it has less potential to improve. The strategies considering both GEBV and genomic diversity (GEBV-GD-30, -50, -100) could balance the tradeoff between starting $\overline{GEBV_P}$ and genomic variation of the base population.

Dataset II was specifically collected for genomic selection. All of the available accessions in the dataset belong to *indica* or *indica-admixed* group. From the results of the performance based on the best F_{10} RILs in Table 2, all the seven strategies seem to have close performance for the three target traits. The resulting GEBV averages of the best F_{10} RILs range from 6472 to 6546 kg/ha for YLD, from 85.889 to 91.852 cm for PH, and from 77.725 to 78.410 days for FT. This could be due to the fact that the candidate accessions in Dataset II are elite breeding lines which have limited genomic diversity and similar phenotypic values for the target traits. However, the two strategies with greater genomic diversity, GD-O-100 and GEBV-GD-100 for YLD (their corresponding GEBV averages are 6546 and 6539 kg/ha), led to larger YLD than the other five strategies (their corresponding GEBV averages range from 6472 to 6506 kg/ha). The four strategies of GEBV-O, GEBV-GD-30, -50, -100 performed equally well for PH (their corresponding GEBV averages range between 85.817 and 86.062 cm), but slightly better than GD-O-30, -50, -100 (their corresponding GEBV averages are 87.517, 89.920, and 91.799 cm). The consistent results based on the $\overline{GEBV_{F_{10}}}$ can be found in Table 5.

It is known that Dataset I contains more genomic diversity than Dataset II, since it consists of five subpopulations and one admixed group. The more genomic diversity of Dataset I could lead to a bigger difference between the strategies considering both GEBV and genomic diversity, and the strategy considering GEBV only for some traits.

For example, the difference of the GEBV averages among the best F_{10} RILs between GEBV-GD-50 and GEBV-O is about -9.06 days for FTAA, and -2.55 days for FTAF in Dataset I (Table 1), but the corresponding difference is just -0.09 days for FT in Dataset II (Table 2). However, the flowering time is very sensitive to environments, so the genomic diversity cannot solely amount to the different results between these two datasets. More interestingly, the more genomic diversity of Dataset I could lead to a larger genetic gain for a specific trait. From Table 4, the mean of the genetic gains using the seven strategies for PH in Dataset I is given by -42.15 cm. But, from Table 5, the corresponding mean in Dataset II is just -13.79 cm.

Daetwyler et al. (2015) and Goiffon et al. (2017) highlighted that an increase in rare favorable alleles in a population can help improve selection responses. Selecting only parental lines with the highest GEBVs can result in a loss of rare favorable alleles for some target traits, thus missing potential RILs over future generations. From the results of BRSW, FTAA, FTAF, and PNPP in Figure 2; and YLD in Figure 3, the performance of GEBV-O appears to be inferior to GEBV-GD-30, -50, -100. This indicates that an increase in genomic diversity in parental lines could compensate for this possible deficiency, and then improve the long-term response to the target traits. The greater genomic diversity could increase the possibility of containing favorable alleles in parental lines, and we therefore expect that the chance of harboring the favorable alleles would increase in RIL populations.

Apparently, the numbers of accessions fixed in the proposed strategies seem to be a little arbitrary, such as those of selecting 10 parental lines, retaining the top 2 accessions, and searching 10 or another 8 accessions from the three candidate sets composed of the top 30, 50 and 100 accessions, respectively. A user certainly can adjust these numbers in the strategies for her/his own study. Also, it was required to have historical phenotypic data used to build the GP model. If the historical phenotypic data are not available, then a pilot experiment is needed to phenotype a set of accessions, which can be determined using an optimization algorithm (Ou and Liao 2019). An R function for performing the proposed procedure of selecting parental lines is available from the authors upon request.

As mentioned earlier, Yao et al. (2018) evaluated genetic gain using the usefulness function in parental selection, and showed that their selection strategy outperformed the strategy using the mid-parent GEBV. In this study, we emphasized both GEBV and genomic diversity in parental selection, and we made generation advancement decisions for each selection strategy according to the GEBVs of the top individuals. Finally, we compared different strategies based on the performance of the best F_{10} RILs, and discussed genetic gain for target traits using the strategies. Moreover, Yao et al. (2018) showed that applying a selection index incorporating multiple traits can simultaneously improve both yield and quality in wheat than the individual trait selection. Also, Jia and Jannink (2012), Hayashi and Iwata (2013), and Guo et al. (2014) highlighted that multiple-trait GP models can provide better prediction accuracy than single-trait GP models for those traits with low heritability. We will consider the selection index and the multiple-trait GP models into the framework of the current study, so as to investigate the multiple-trait situations in a future study.

CONCLUSIONS

Combining GP with Monte Carlo simulation can be a useful means of detecting superior parents for rice pre-breeding programs. Different strategies can be implemented to identify a set of superior parental lines from a candidate population. The strategy considering GEBV only can have a better starting GEBV average but less genomic diversity in the base population. On the other hand, the strategies considering genomic diversity only can have greater genomic diversity but a less favorable starting GEBV average in the base population. The strategies considering both GEBV and genomic diversity that can balance the starting GEBV average with maintenance of genomic diversity should be recommended for practical use.

Abbreviations: BLUP, best linear unbiased predictor; BRSW, brown rice seed width; GBLUP, genomic best linear unbiased predictor; GEBV, genomic estimated breeding value; GEBV-GD, algorithms considering both GEBV and genomic diversity; GEBV-O, algorithms considering GEBV only; FPP, florets per panicle; FT, flowering time; FTAA, flowering time at Arkansas; FTAF, flowering time at Faridpur; GD-O, algorithms considering genomic diversity only; PH, plant height; PNPP, panicle number per plant; RIL, recombinant inbred line; SNP, single nucleotide polymorphism; YLD, grain yield.

DECLARATIONS

Ethical Approval and Consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of supporting data: Not applicable.

Competing interests: The authors declare that there is no conflict of interest.

Funding: Not applicable.

Authors' contributions: PY analyzed the datasets, wrote the R functions, prepared the tables and figures, and drafted the manuscript. CT supervised the research, derived the analysis approach, and drafted the manuscript.

Acknowledgements: This research was supported by the Ministry of Science and Technology, Taiwan (grant number: MOST 108-2118-M-002-002-MY2).

REFERENCES

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, and J.L. Jannink. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite north American oats. Plant Genome 4: 132–144.

Atkison, A.C., and A.N. Donev. 1992. Optimum experimental designs. New York:

Oxford University Press.

Bernardo, R. 2003. Parental selection, number of breeding populations, and size of each population in inbred development. Theor. Appl. Genet. 107: 1252-1256.

Broman, K.W., H. Wu, S. Sen, and G.A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889-890.

Covarrubias-Pazaran, G. 2016. Genome-assisted prediction of quantitative traits using the R package sommer. PLOS One 11: e0156744.

Daetwyler, H.D., M.J. Hayden, G.C. Spangenberg, and B.J. Hayes. 2015. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics 200: 1341–1348.

Gaynor, R.C., G. Gorjanc, A.R. Bentley, E.S. Ober, P. Howell, R. Jackson, et al. 2017. A two-part strategy for using genomic selection to develop inbred lines. Crop Sci. 57: 2372–2386.

Goiffon, M., A. Kusmec, L. Wang, G. Hu, and P.S. Schnable. 2017. Improving response in genomic selection with a population-based selection strategy: Optimal population value selection. Genetics 206: 1675–1682.

Guo, G., F. Zhao, Y. Wang, Y. Zhang, L. Du, and G. Su. 2014. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genetics 15: 30.

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

Haldane, J.B.S. 1919. The combination of linkage values and the calculation of distance between the loci for linked factors. Genetics 8: 299–309.

Hayashi, T., and H. Iwata. 2013. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC bioinformatics 14: 1–14.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433–443.

Heffner, E.L., J.L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in wheat breeding program. Plant Genome 4: 65–75.

Henderson, C.R. 1984. Applications of linear models in animal breeding. Univ. of Guelph, Guelph, Ontario.

Jia, Y., and J.L. Jannink. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192: 1513-1522.

Lorenz, A.J., K.P. Smith, and J.L. Jannink. 2012. Potential and optimization of genomic selection for *Fusarium* head blight resistance in six-row barley. Crop Sci. 52: 1609–1621.

Massman, J.M., H.J.G. Jung, and R. Bernardo. 2013. Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. 53: 58–66.

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

Ou, J.H., and C.T. Liao. 2019. Training set determination for genomic selection. Theor. Appl. Genet. 132: 2781–2792.

Piepho, H.P. 2009. Ridge regression and extensions for genome-wide selection in maize. Crop Sci. 49: 1165–1176.

Rutkoski, J.E. 2019. A practical guide to genetic gain. Adv. Agron. 157: 217-249.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redona, et al. 2015. Genomic selection and association mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLOS Genetics 11: e1005350.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Vanavermaete, D., J. Fostier, S. Maenhout, and B. De Baets. 2020. Preservation of genetic variation in a breeding population for long-term genetic gain. G3: Genes|Genomes|Genetics doi:10.1534/g3.120.401354.

Whitley, D. 1994. A genetic algorithm tutorial. Stat. Comput. 4: 65–85.

Witcombe, J.R., S. Gyawali, M. Subedi, D.S. Virk, and K.D. Joshi. 2013. Plant breeding can be made more efficient by having fewer, better crosses. BMC Plant Biol. 13: 22.

Xavier, A., W.M. Muir, B. Craig, and K.M. Rainey. 2016. Walking through the statistical black boxes of plant breeding. Theor. Appl. Genet. 129: 1933–1949.

Yao, J., D. Zhao, X. Chen, Y. Zhang, and J. Wang. 2018. Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). Crop J. 6: 353-365.

Youens-Clark, K., E. Buckler, T. Casstevens, C. Chen, G. DeClerck, P. Derwent, et al. 2011. Gramene database in 2010: updates and extensions. Nucleic Acids Res. 39: D1085–D1094.

Zhao, K., C.W. Tung, G.C. Eizenga, M.H. Wright, M.L. Ali, A.H. Price, et al. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nature Communications 2: 467.

Table 1: The ranking and the GEBV average (in parentheses) for the best F_{10} RILs from the 30 repetitions using the seven proposed strategies in Dataset I.

	BRSW	FPP	FTAA	FTAF	PH	PNPP
GEBV-O	6 (3.418)	**2 (5.961)**	6 (56.521)	6 (61.856)	**1 (42.185)**	6 (4.125)
GD-O-30	7 (3.408)	5 (5.951)	3 (51.564)	3 (59.353)	5 (49.337)	3 (4.188)
GD-O-50	3 (3.576)	6 (5.916)	5 (53.348)	5 (60.126)	6 (49.801)	5 (4.138)
GD-O-100	4 (3.496)	7 (5.882)	7 (56.835)	7 (61.967)	7 (51.788)	7 (4.086)
GEBV-GD-30	5 (3.419)	3 (5.954)	**1 (47.136)**	**1 (59.216)**	**2 (42.699)**	**1 (4.225)**
GEBV-GD-50	**1 (3.656)**	**1 (5.964)**	**2 (47.457)**	**2 (59.304)**	3 (43.232)	**2 (4.214)**
GEBV-GD-100	**2 (3.634)**	4 (5.953)	4 (51.382)	4 (59.634)	4 (43.498)	4 (4.171)

(i) The best and second-best strategies are indicated in bold text, and the worst and second-worst strategies are indicated by underlining.

(ii) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores.

(iii) BRSW: brown rice seed width; FPP: florets per panicle; FTAA: flowering time at Arkansas; FTAF: flowering time at Faridpur; PH: plant height; PNPP: panicle number per plant.

Table 2: The ranking and the GEBV average (in parentheses) for the best F_{10} RILs from the 30 repetitions using the seven proposed strategies in Dataset II.

	YLD	PH	FT
GEBV-O	7 (6472)	**1 (85.817)**	**2 (77.818)**
GD-O-30	4 (6491)	5 (87.517)	7 (78.410)
GD-O-50	5 (6489)	6 (89.920)	5 (78.164)
GD-O-100	**1 (6546)**	7 (91.799)	6 (78.359)
GEBV-GD-30	3 (6506)	**2 (85.976)**	4 (77.883)
GEBV-GD-50	6 (6485)	3 (85.917)	**1 (77.725)**
GEBV-GD-100	**2 (6539)**	4 (86.062)	3 (77.873)

(i) The best and second-best strategies are indicated in bold text, and the worst and second-worst strategies are indicated by underlining.

(ii) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores.

(iii) YLD: yield; PH: plant height; FT: flowering time.

Table 3: The GEBV averages for the best F_{10} RILs from the 30 repetitions based on the group of the original 45 F_1 hybrids and the reduced group of 17 F_1 hybrids using the four strategies of GEBV-O, GEBV-GD-30, GEBV-GD-50, and GEBV-GD-100.

	GEBV-O		GEBV-GD-30		GEBV-GD-50		GEBV-GD-100	
Dataset I	45 F_1	17 F_1	45 F_1	17 F_1	45 F_1	17 F_1	45 F_1	17 F_1
BRSW	3.418	3.423	3.419	3.418	3.656	3.652	3.634	3.650
FPP	5.961	5.965	5.954	5.957	5.964	5.958	5.953	5.943
FTAA	56.521	57.513	47.136	46.961	47.457	47.421	51.382	51.734
FTAF	61.856	61.850	59.216	59.123	59.304	59.232	59.634	59.713
PH	42.185	43.409	42.699	43.271	43.232	43.791	43.498	43.854
PNPP	4.125	4.129	4.225	4.226	4.214	4.204	4.171	4.161
Dataset II	45 F_1	17 F_1	45 F_1	17 F_1	45 F_1	17 F_1	45 F_1	17 F_1
YLD	6472	6476	6506	6499	6485	6484	6539	6534
PH	85.817	85.991	85.976	85.844	85.917	86.092	86.062	86.060
FT	78.818	77.834	77.883	77.750	77.725	77.778	77.873	77.690

(i) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores.

(ii) BRSW: brown rice seed width; FPP: florets per panicle; FTAA: flowering time at Arkansas; FTAF: flowering time at Faridpur; PH: plant height; PNPP: panicle number per plant.

(iii) YLD: yield; PH: plant height; FT: flowering time.

Table 4: The average of genetic gains from the 30 repetitions for Dataset I.

	BRSW			FPP		
	$\overline{\text{GEBV}}_{\text{P}}$	$\overline{\text{GEBV}}_{\text{F}_{10}}$	genetic gain	$\overline{\text{GEBV}}_{\text{P}}$	$\overline{\text{GEBV}}_{\text{F}_{10}}$	genetic gain
GEBV-O	3.17	3.42	0.25	5.51	5.96	0.45
GD-O-30	3.10	3.41	0.31	5.48	5.95	0.47
GD-O-50	3.00	3.57	0.57	5.41	5.91	0.50
GD-O-100	2.94	3.49	0.55	5.31	5.88	0.57
GEBV-GD-30	3.12	3.42	0.30	5.48	5.95	0.47
GEBV-GD-50	3.04	3.65	0.61	5.43	5.96	0.53
GEBV-GD-100	3.00	3.63	0.63	5.34	5.95	0.61

	FTAA			FTAF		
	$\overline{\text{GEBV}}_{\text{P}}$	$\overline{\text{GEBV}}_{\text{F}_{10}}$	genetic gain	$\overline{\text{GEBV}}_{\text{P}}$	$\overline{\text{GEBV}}_{\text{F}_{10}}$	genetic gain
GEBV-O	64.30	56.57	-7.73	63.45	61.87	-1.58
GD-O-30	72.25	49.26	-22.99	64.93	59.40	-5.53
GD-O-50	75.41	53.54	-21.87	65.82	60.16	-5.66
GD-O-100	80.01	57.00	-23.01	67.34	62.01	-5.33
GEBV-GD-30	71.09	47.31	-23.78	64.68	59.25	-5.43
GEBV-GD-50	72.86	47.64	-25.22	65.40	59.35	-6.05
GEBV-GD-100	77.16	51.53	-25.63	66.46	59.68	-6.78

	PH			PNPP		
	$\overline{\text{GEBV}}_{\text{P}}$	$\overline{\text{GEBV}}_{\text{F}_{10}}$	genetic gain	$\overline{\text{GEBV}}_{\text{P}}$	$\overline{\text{GEBV}}_{\text{F}_{10}}$	genetic gain
GEBV-O	83.77	42.52	-41.25	3.93	4.12	0.19
GD-O-30	89.50	49.69	-39.81	3.86	4.19	0.33
GD-O-50	90.11	50.13	-39.98	3.80	4.14	0.34
GD-O-100	92.10	52.10	-40.00	3.64	4.08	0.44
GEBV-GD-30	87.26	42.99	-44.27	3.90	4.22	0.32
GEBV-GD-50	87.95	43.50	-44.45	3.84	4.21	0.37
GEBV-GD-100	89.27	43.95	-45.32	3.70	4.17	0.47

(i) $\overline{GEBV_P}$: the GEBV average among the 10 selected parental lines. $\overline{GEBV_{F_{10}}}$: the GEBV average among the resulting 2700 F_{10} RILs.

(ii) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores.

(iii) BRSW: brown rice seed width; FPP: florets per panicle; FTAA: flowering time at Arkansas; FTAF: flowering time at Faridpur; PH: plant height; PNPP: panicle number per plant.

Table 5: The average of genetic gains from the 30 repetitions for Dataset II.

YLD

	$\overline{\mathrm{GEBV}}_{\mathrm{P}}$	$\overline{\mathrm{GEBV}}_{\mathrm{F}_{10}}$	genetic gain
GEBV-O	5571.61	6468.60	896.99
GD-O-30	5452.39	6488.02	1035.63
GD-O-50	5436.58	6484.58	1048.00
GD-O-100	5289.74	6540.72	1250.98
GEBV-GD-30	5538.44	6501.23	962.79
GEBV-GD-50	5522.45	6482.13	959.68
GEBV-GD-100	5454.37	6535.79	1081.42

PH

	$\overline{\mathrm{GEBV}}_{\mathrm{P}}$	$\overline{\mathrm{GEBV}}_{\mathrm{F}_{10}}$	genetic gain
GEBV-O	97.75	85.89	-11.86
GD-O-30	102.20	87.59	-14.61
GD-O-50	103.66	89.99	-13.67
GD-O-100	106.83	91.85	-14.98
GEBV-GD-30	99.00	86.01	-12.99
GEBV-GD-50	99.39	85.99	-13.40
GEBV-GD-100	101.15	86.13	-15.02

FT

	$\overline{\mathrm{GEBV}}_{\mathrm{P}}$	$\overline{\mathrm{GEBV}}_{\mathrm{F}_{10}}$	genetic gain
GEBV-O	83.14	77.84	-5.30
GD-O-30	83.98	78.73	-5.25
GD-O-50	84.57	78.19	-6.38
GD-O-100	85.62	78.39	-7.23
GEBV-GD-30	83.44	77.90	-5.54
GEBV-GD-50	83.69	77.76	-5.93
GEBV-GD-100	84.16	77.89	-6.27

(i) $\overline{GEBV}_P$: the GEBV average among the 10 selected parental lines. $\overline{GEBV}_{F_{10}}$: the

GEBV average among the resulting 2700 F_{10} RILs.

(ii) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores.

(iii) YLD: yield; PH: plant height; FT: flowering time.

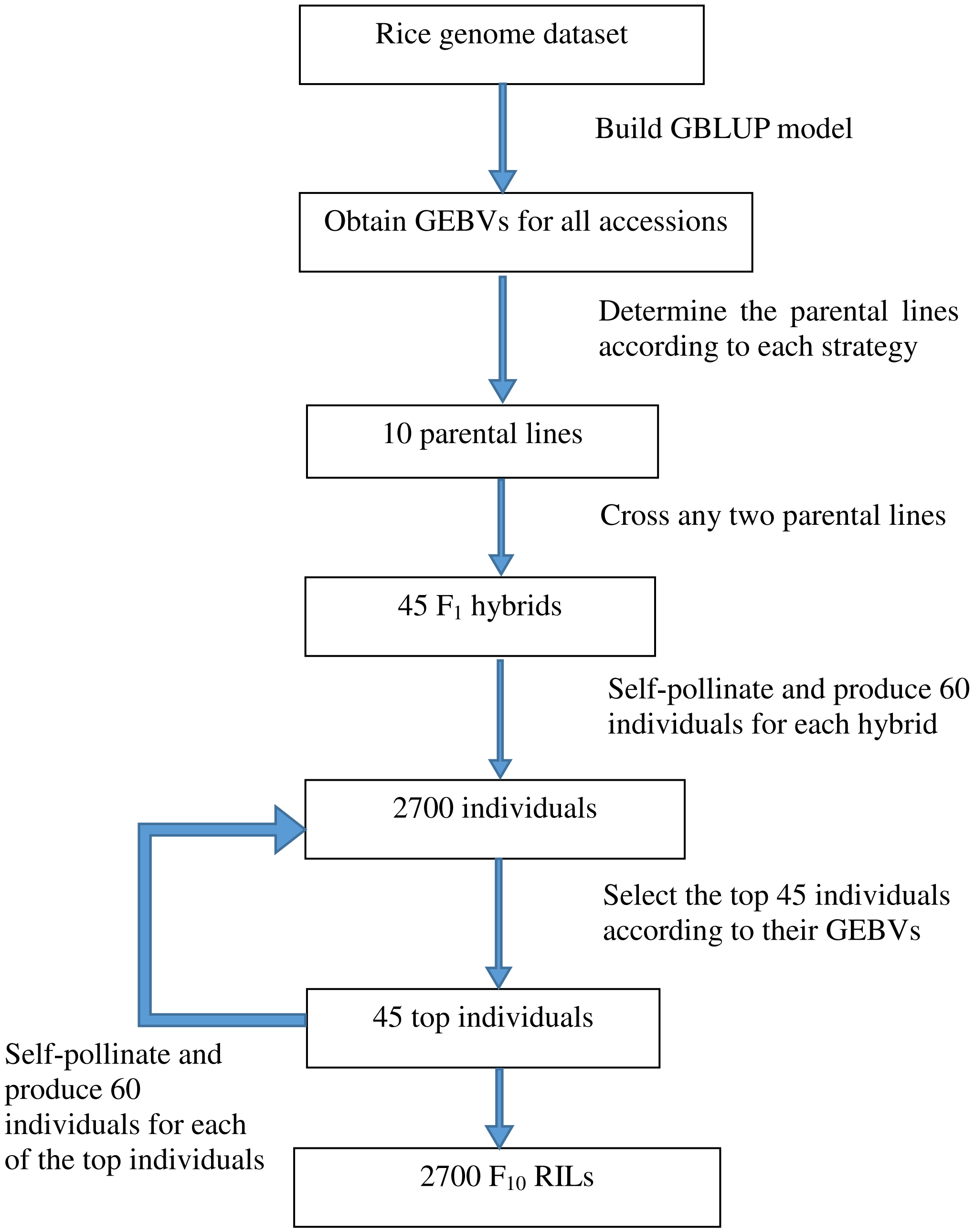
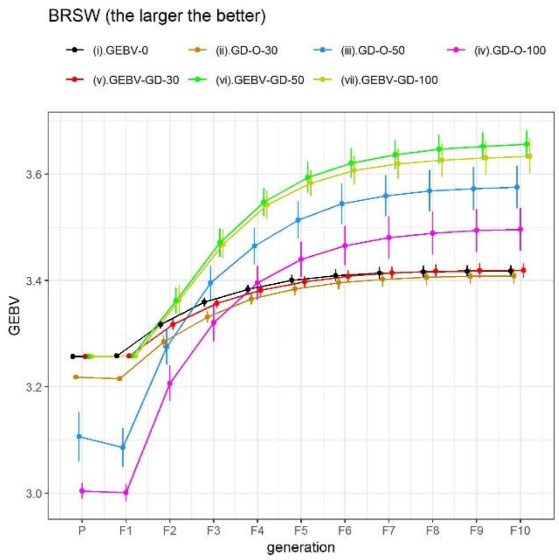
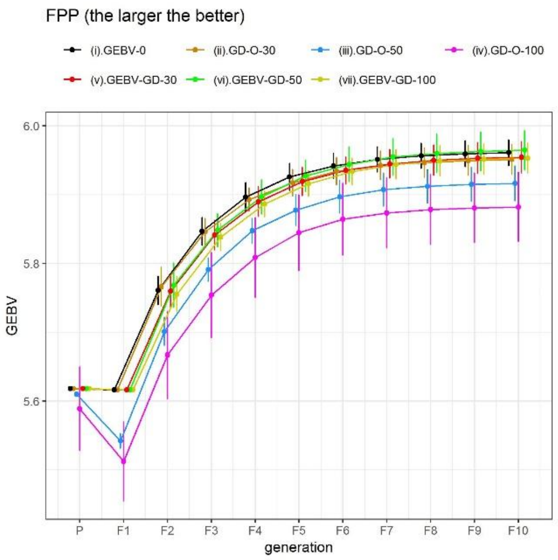
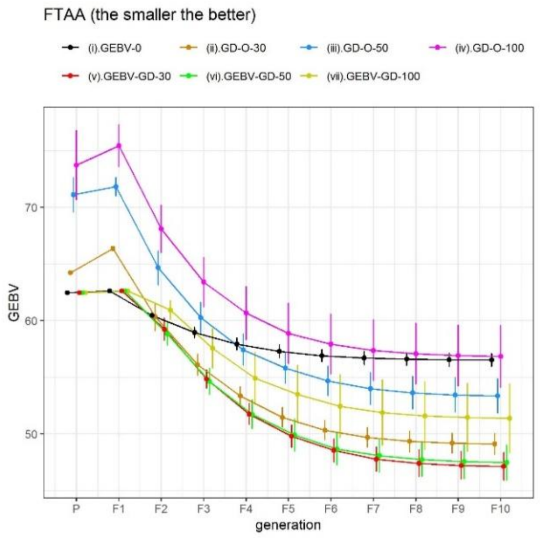

Rice genome dataset
Build GBLUP model
Obtain GEBVs for all accessions
Determine the parental lines according to each strategy
10 parental lines
Cross any two parental lines
45 F1 hybrids
Self-pollinate and produce 60 individuals for each hybrid
2700 individuals
Select the top 45 individuals according to their GEBVs
45 top individuals
Self-pollinate and produce 60 individuals for each of the top individuals
2700 F10 RILs

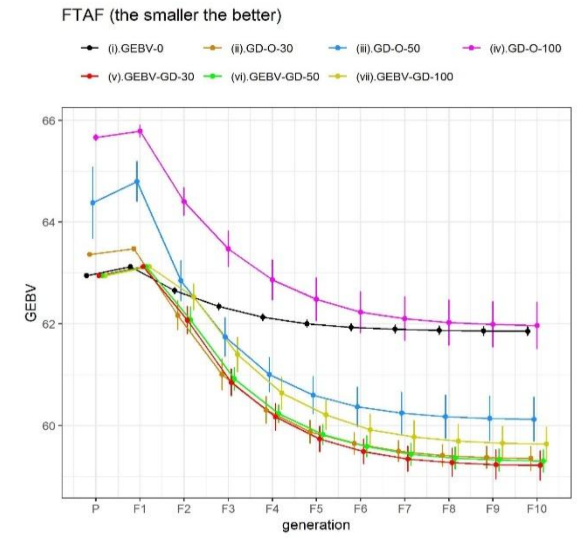
Figure 1: The working flow for the Monte Carlo simulation.

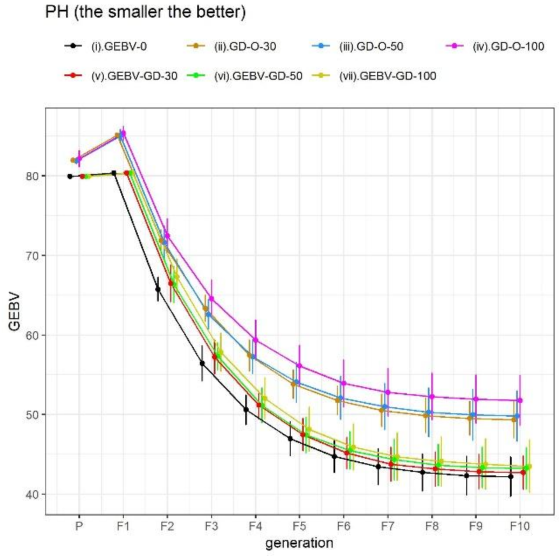
GEBV: genomic estimated breeding value; GBLUP: genomic best linear unbiased predictor; RIL: recombinant inbred line.


BRSW (the larger the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
GEBV
3.0
3.2
3.4
3.6
generation
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10


FPP (the larger the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
GEBV
5.6
5.8
6.0
generation
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10


FTAA (the smaller the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
GEBV
50
60
70
generation
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10


FTAF (the smaller the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
GEBV
60
62
64
66
generation
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10


PH (the smaller the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
GEBV
40
50
60
70
80
generation
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10

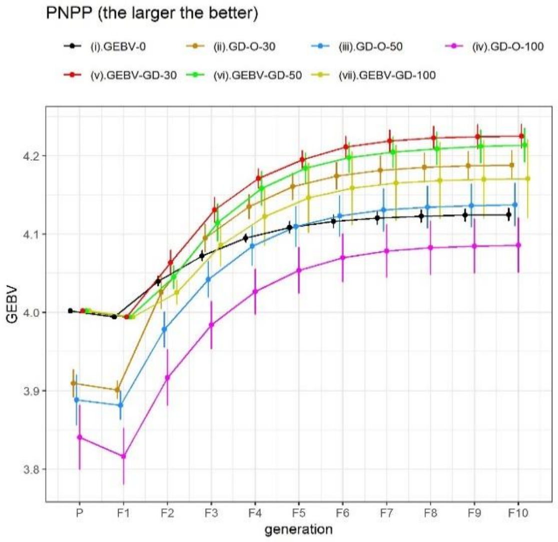
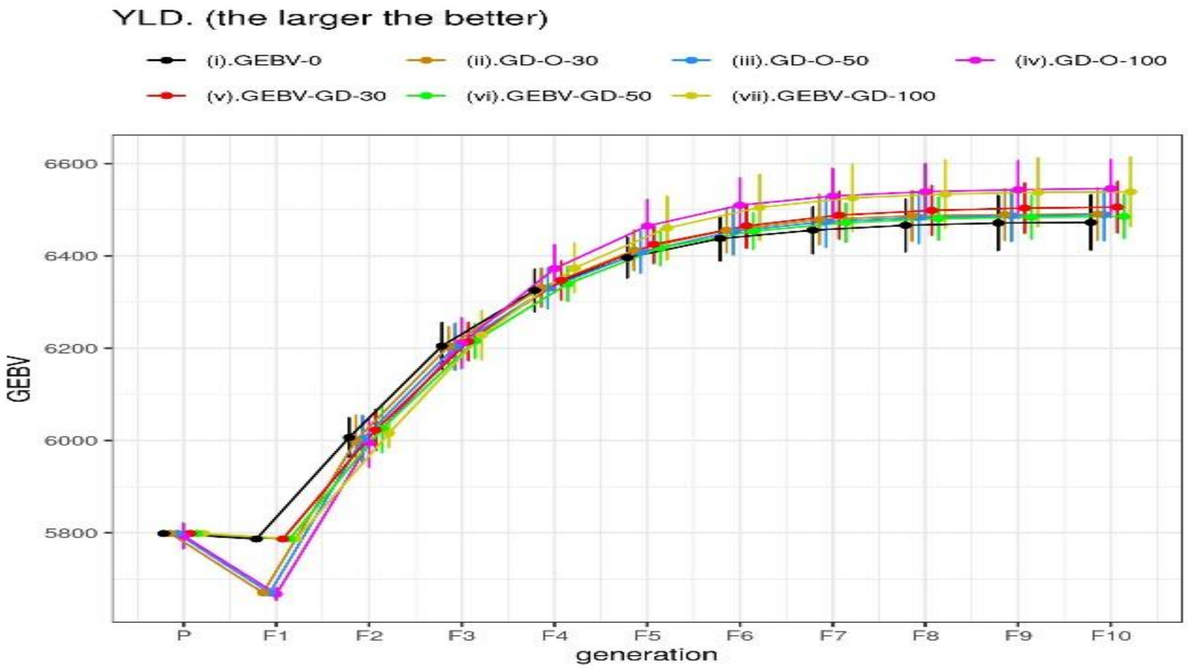

PNPP (the larger the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
GEBV
3.8
3.9
4.0
4.1
4.2
generation
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10

Figure 2: The GEBV averages for the best individuals from the 30 repetitions at consecutive generations for the six chosen traits in Dataset I.

(i) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores.

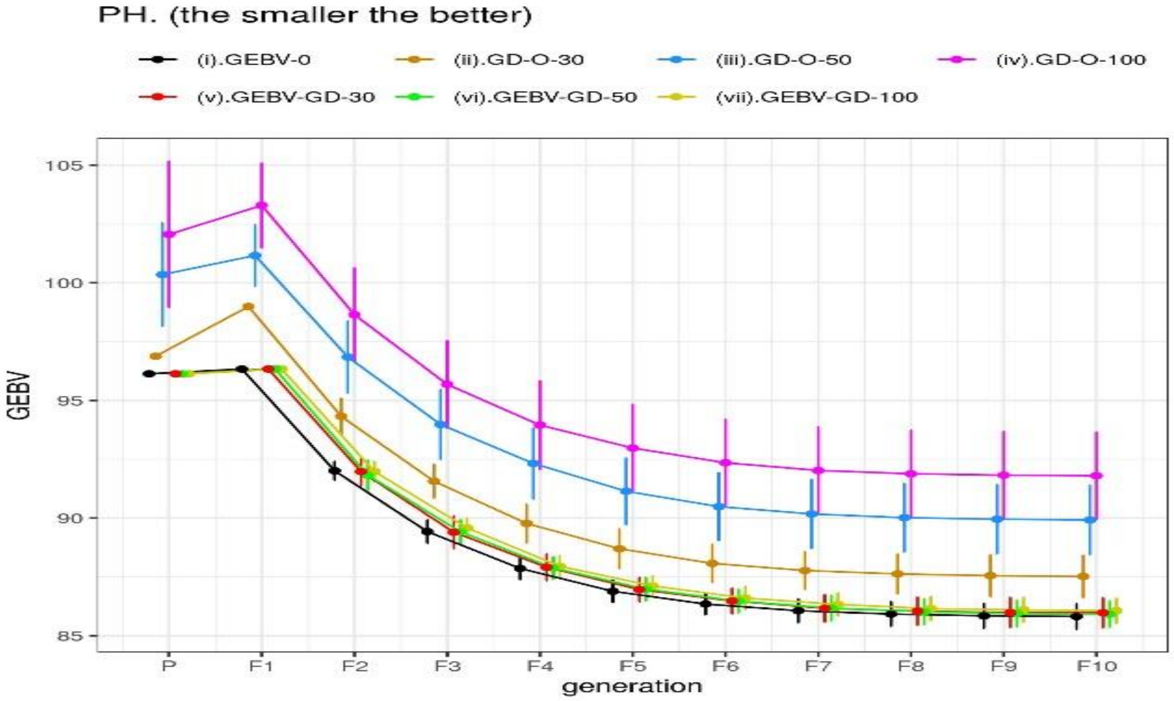
(ii) BRSW: brown rice seed width; FPP: florets per panicle; FTAA: flowering time at Arkansas; FTAF: flowering time at Faridpur; PH: plant height; PNPP: panicle number per plant.



YLD. (the larger the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
6600
6400
6200
6000
5800
GEBV
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
generation

644

645



PH. (the smaller the better)
(i).GEBV-0
(ii).GD-O-30
(iii).GD-O-50
(iv).GD-O-100
(v).GEBV-GD-30
(vi).GEBV-GD-50
(vii).GEBV-GD-100
105
100
95
90
85
GEBV
P
F1
F2
F3
F4
F5
F6
F7
F8
F9
F10
generation

646

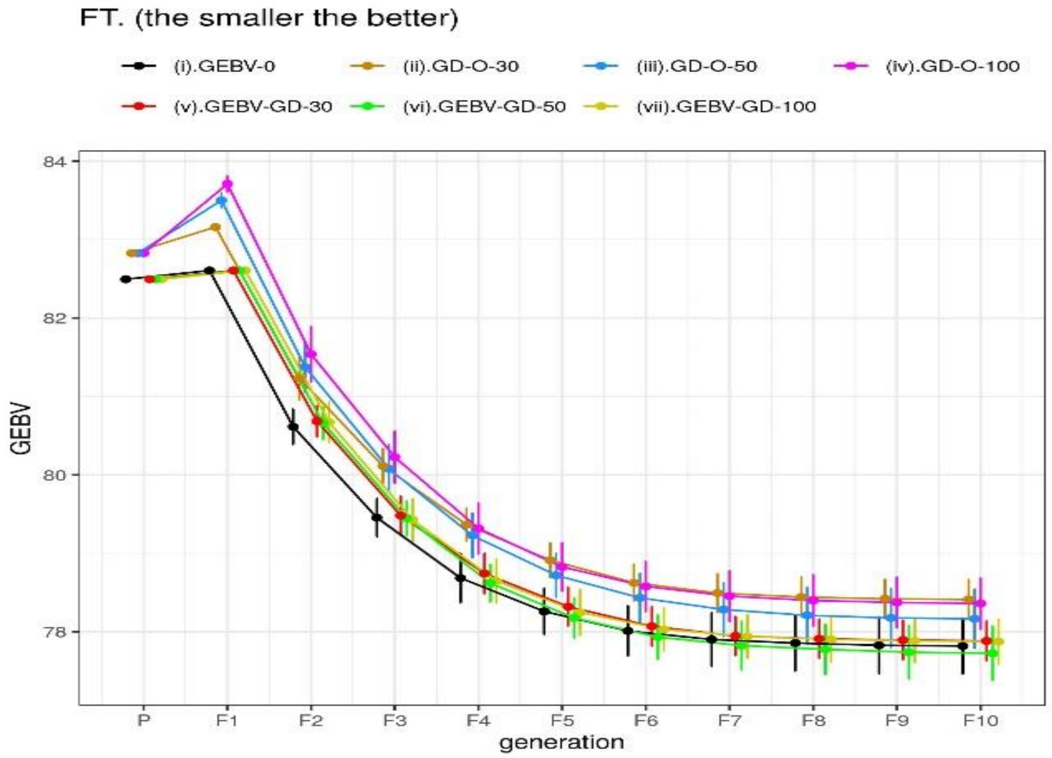


Figure 3: The GEBV averages for the best individuals form the 30 repetitions at consecutive generations for the three target traits in Dataset II.

(i) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores.

(ii) YLD, yield; PH, plant height; FT, flowering time.

Figures

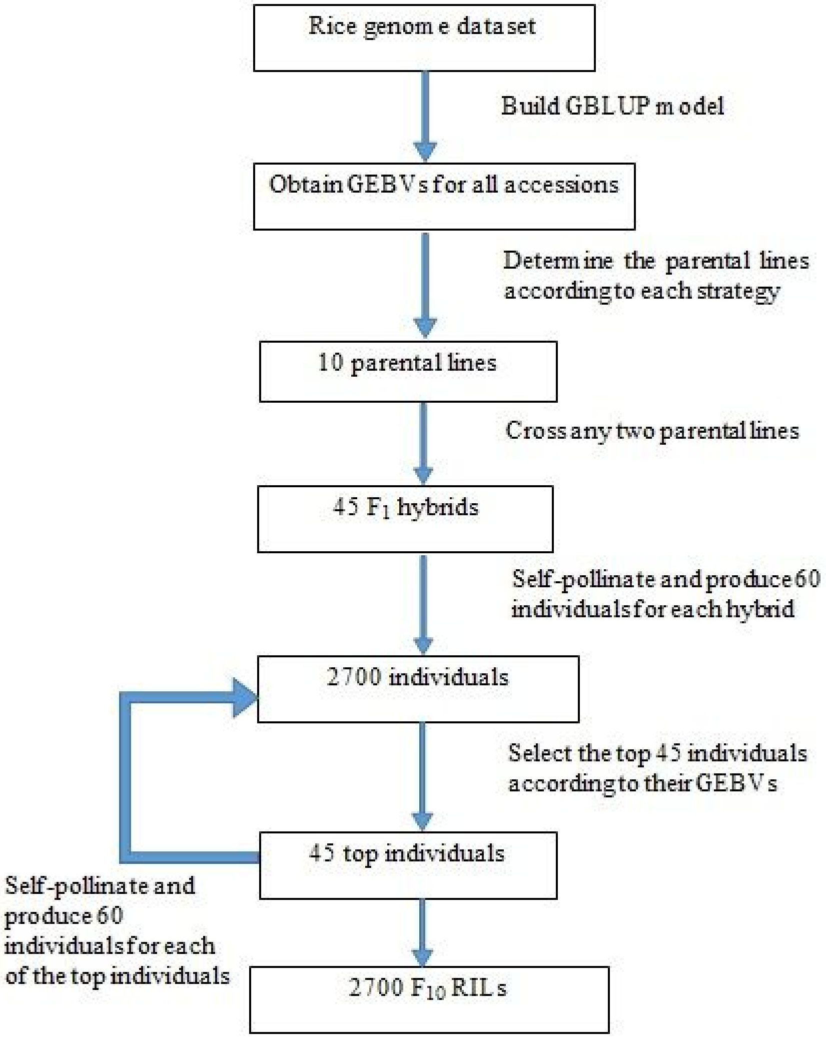

Rice genom e dataset
Build GBLUP m odel
Obtain GEBVs for all accessions
Determ ine the parental lines according to each strategy
10 parental lines
Cross any two parental lines
45 F1 hybrids
Self-pollinate and produce 60 individuals for each hybrid
2700 individuals
Select the top 45 individuals according to their GEBVs
45 top individuals
Self-pollinate and produce 60 individuals for each of the top individuals
2700 F10 RILs

Figure 1

The working flow for the Monte Carlo simulation. GEBV: genomic estimated breeding value; GBLUP: genomic best linear unbiased predictor; RIL: recombinant inbred line.

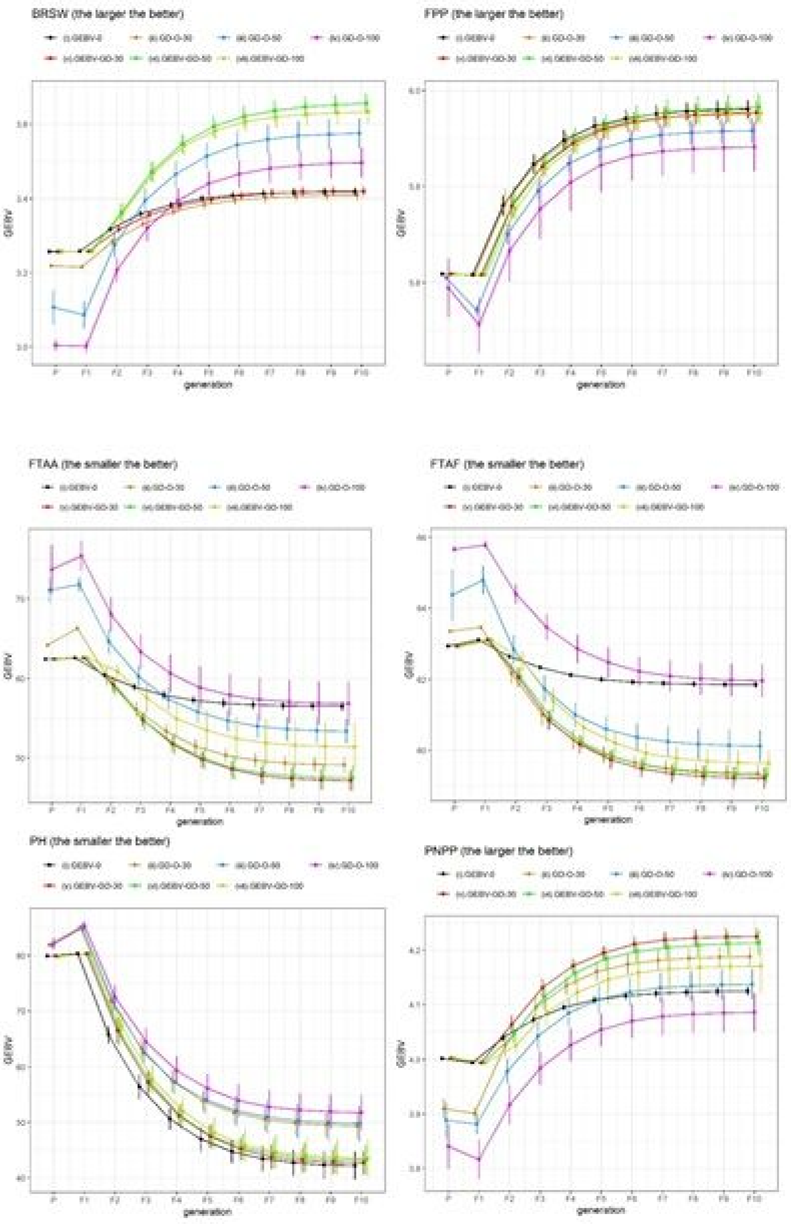


Figure 2

The GEBV averages for the best individuals from the 30 repetitions at consecutive generations for the six chosen traits in Dataset I. (i) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores. (ii) BRSW: brown rice seed width; FPP: florets per panicle; FTAA: flowering time at Arkansas; FTAF: flowering time at Faridpur; PH: plant height; PNPP: panicle number per plant.

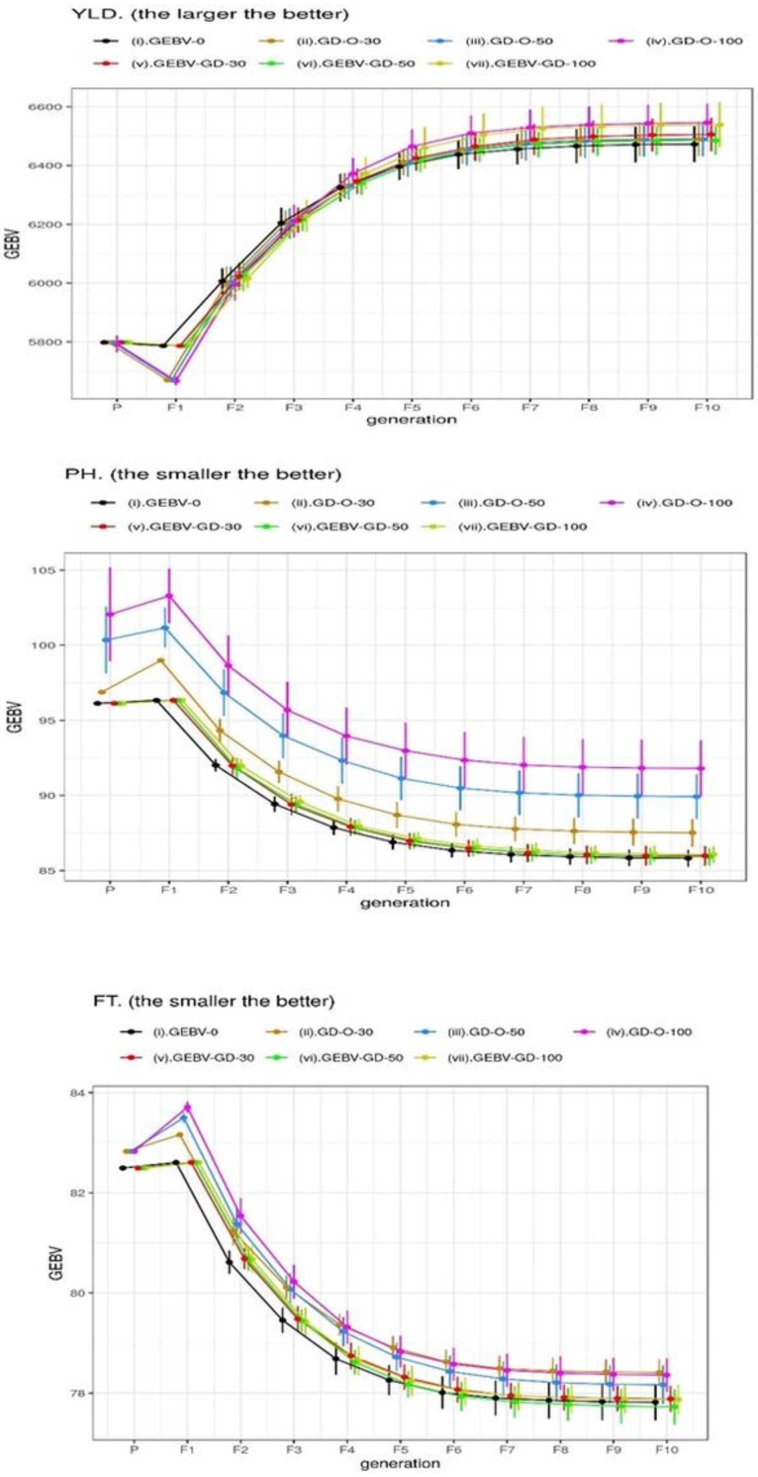


Figure 3

The GEBV averages for the best individuals form the 30 repetitions at consecutive generations for the three target traits in Dataset II. (i) GEBV-O: the subset of the top 10 accessions with the minimal or maximal GEBVs; GD-O-30, -50, -100: the subsets of 10 accessions with the maximal D-scores chosen from the candidate sets composed of the top 30, 50, and 100 accessions, respectively; GEBV-GD-30, -50, -100: the subsets of the top 2 accessions plus 8 accessions chosen from the remainder of the candidate sets composed of the top 30, 50, and 100 accessions, respectively, which have the maximal D-scores. (ii) YLD, yield; PH, plant height; FT, flowering time.