CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

Optimal Designs for Genomic Selection in Hybrid Crops

Tingting Guo[1], Xiaoqing Yu[1], Xianran Li[1], Haozhe Zhang[2], Chengsong Zhu[1], Sherry Flint-Garcia[3], Michael D. McMullen[3], James B. Holland[4], Stephen J. Szalma[4], Randall J. Wisser[5] and Jianming Yu[1,*]

[1]Department of Agronomy, Iowa State University, Ames, IA 50011, USA

[2]Department of Statistics, Iowa State University, Ames, IA 50011, USA

[3]USDA-ARS and Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA

[4]USDA-ARS and Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695, USA

[5]Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, USA

***Correspondence: Jianming Yu (jmyu@iastate.edu)**

https://doi.org/10.1016/j.molp.2018.12.022

ABSTRACT

Improved capacity of genomics and biotechnology has greatly enhanced genetic studies in different areas. Genomic selection exploits the genotype-to-phenotype relationship at the whole-genome level and is being implemented in many crops. Here we show that design-thinking and data-mining techniques can be leveraged to optimize genomic prediction of hybrid performance. We phenotyped a set of 276 maize hybrids generated by crossing founder inbreds of nested association mapping populations for flowering time, ear height, and grain yield. With 10 296 310 SNPs available from the parental inbreds, we explored the patterns of genomic relationships and phenotypic variation to establish training samples based on clustering, graphic network analysis, and genetic mating scheme. Our analysis showed that training set designs outperformed random sampling and earlier methods that either minimize the mean of prediction error variance or maximize the mean of generalized coefficient of determination. Additional analyses of 2556 wheat hybrids from an early-stage hybrid breeding system and 1439 rice hybrids from an established hybrid breeding system validated the approaches. Together, we demonstrated that effective genomic prediction models can be established with a training set 2%–13% of the size of the whole set, enabling an efficient exploration of enormous inference space of genetic combinations.

Key words: data mining, molecular breeding, genomic relationship, genomic selection, optimal design

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

INTRODUCTION

The relationship of genotype to phenotype is a fundamental concept in evolution, biology, and genetics. Among genomics-enabled strategies (Tester and Langridge, 2010; Morrell et al., 2012), genomic selection capitalizes on the genotype–phenotype relationship directly at the whole-genome level and has been implemented in different breeding contexts (Riedelsheimer et al., 2012; Technow et al., 2014; Yu et al., 2016), and human complex traits and diseases (de los Campos et al., 2010).

Different aspects of the selection and breeding process should be examined, given the improved capacity in genomics, biotechnologies, and phenomics. Changes can be proposed at the overall program level or at different stages of a program (Bernardo, 2010; Xu et al., 2014; Kadam et al., 2016). Rather than asking how technologies can enhance existing crop improvement pipelines, we can ask about efficient designs enabled by genomics (Technow et al., 2014; Xu et al., 2014; Zeng et al., 2017). One such question is how to efficiently establish the genotype–phenotype relationship so that reliable predictions can be made to guide the exploration of the enormous genetic space for selection. This is particularly the case for hybrid crops because the number of potential hybrids is prohibitively high for extensive testing (Figure 1A). Many field, vegetable, and flower crops use hybrids, including maize, sorghum, and sunflower. In addition, hybrid rice is being adopted and hybrid

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

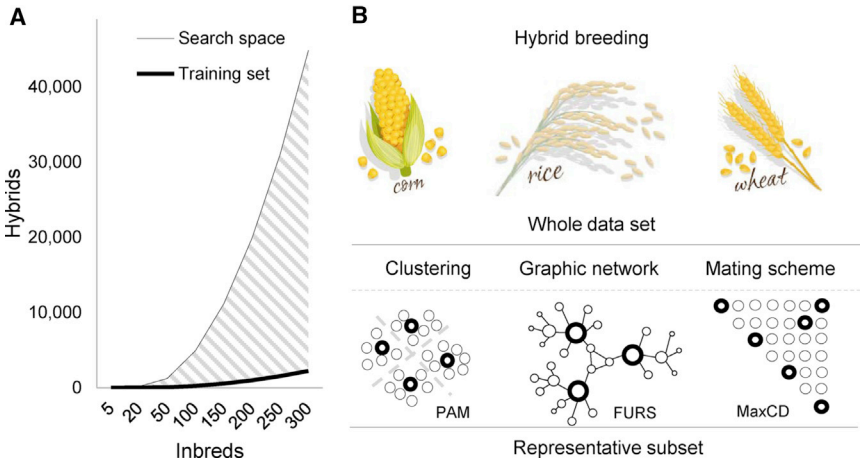

Figure 1. Potential Hybrids and Optimal Designs for Hybrid Prediction.

(A) Training set designs are needed to efficiently explore the large search space of genetic combinations.

(B) Representative subset selection can be conducted by viewing the problem from different angles. PAM is partitioning around medoids. FURS is fast and unique representative subset selection. MaxCD is maximization of connectedness and diversity. The training set is formed by the representative subset shown by boldface circles, and the whole set is shown by all circles.

wheat research has received new attention. With given resources, identifying superior hybrids through genomics-enabled approaches from a huge number of potential combinations is a standing challenge.

Data mining, the process of knowledge discovery from data, has been widely used in many areas by drawing strength from statistics, machine learning, pattern recognition, information retrieval, and network science (Han et al., 2011). Finding an informative subset from a large collection of data objects, or representative subset selection (Pan et al., 2005), is central to many problems in social network, recommender systems, health informatics, and image processing, for which numerous data-mining techniques have been developed to deal with this challenge (Elhamifar et al., 2014). In the context of genomic selection, representative subset selection is a logical choice to exploit the available genotyping data to design the training set, for which the phenotypic data will be collected to establish the genotype–phenotype relationship (Figure 1B).

Previous research efforts were devoted to the feasibility of genomic selection, statistical models of the genotype–phenotype relationship, and empirical testing and implementation (Morota and Gianola, 2014). The accuracy of genomic prediction is influenced by the genetic similarity between the training set and prediction set, and phenotypic diversity of the training set (Albrecht et al., 2011; Rincent et al., 2012; Miedaner et al., 2013). Notably, research in training set design for inbred populations (Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015; Lorenz and Smith, 2015; Marulanda et al., 2015) showed promising results. These studies focused on two parameters derived from the mixed model equation, either minimizing the mean of prediction error variance (PEVmean) or maximizing the mean of coefficient of determination (CDmean), and used either exhaustive search or a genetic algorithm.

Genomic prediction and training population design in hybrids differ from that in inbreds. First, while genomic estimated breeding values are generated for inbreds, predicted genotypic values are generated for hybrids using the covariance matrix containing both additive and dominance genomic relationship matrices. Second, besides genetic relationship among parental inbreds, half-sib relationship is introduced in the crossing process. Third, unlike the inbred population where all individuals are already derived and available, only hybrids in the training set and those chosen based on predicted values need to be derived and phenotyped. Finally, saving the process of obtaining and phenotyping unfavorable hybrids highlights the additional importance of design in genomic prediction for hybrids, the number of which is a function of parental inbreds.

In this study, we examined the training set design for hybrid performance prediction among available inbreds that constitute the overall genetic space from which hybrid combinations need to be selected. For hybrid crops with established heterotic patterns such as maize, rye, and sorghum, the number of potential hybrids is the product of the numbers of inbreds from different heterotic groups ($n_1 \times n_2$, where n_1 is the number for one group and n_2 is the other group). For crops that heterotic patterns need to be developed such as wheat, it is the quadratic ($n(n-1)/2$) of the inbred number (n).

We designed and tested three methods of representative subset selection to establish a training set for genomic prediction in hybrids (Figure 1B). Maximization of connectedness and diversity (MaxCD) was conceived by exploring patterns in genetic relationships and phenotypic variations captured in a genetic mating scheme. Partitioning around medoids (PAM) is a clustering algorithm to classify objects into clusters by minimizing the sum of dissimilarities between the objects labeled in a cluster and a designated center object (medoid) of that cluster (Kaufman and Rousseeuw, 1987, 2009; Jain et al., 1999). Unlike other clustering algorithms, having these medoids identified during the clustering process is desirable for applications when a set of representative objects need to be generated rather than having users select from clusters. Fast and unique representative subset selection (FURS) comes from graphic network analysis (Mall et al., 2013), where many methods have been developed for sampling from graphs with a large number of nodes and edges (Leskovec and Faloutsos, 2006). FURS deterministically selects a set of nodes from a given graph to retain the topology of the graph without explicitly performing community detection in the graph, and was shown to be a better choice than some other techniques (e.g., SlashBurn, Forest-Fire, Metropolis, and Snowball Expansion) (Mall et al., 2013). PAM, FURS, or similar data-mining techniques have not been examined in genomic selection for selecting the representative subsets as the training sets

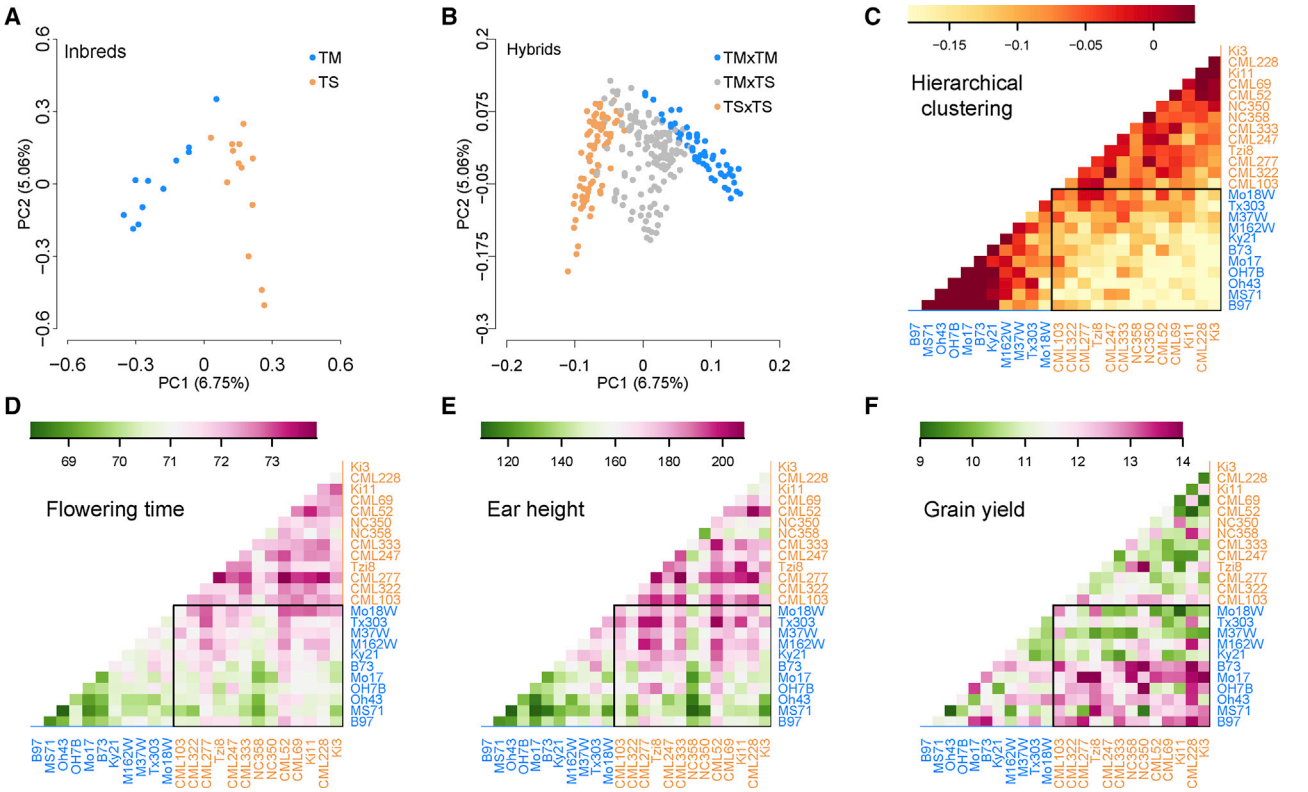

Figure 2. Patterns in Genomic Relationship and Phenotypes in Maize.

(A) PCA plot for 24 parental inbreds. TM, temperate and mixed; TS, tropical and subtropical.

(B) PCA plot for 276 hybrids. Hybrids are color coded into TM × TM, TS × TS, and TM × TS.

(C) Genomic relationship between inbreds in hierarchical cluster order.

(D) Phenotypic values of hybrids for flowering time (days).

(E) Ear height (cm).

(F) Grain yield (Mg ha^{-1}).

For each hybrid in **(D–F)**, inbred parents were ordered by hierarchical clustering. Inter-group hybrids (from factorial) were boxed in **(D–F)**, and the corresponding genomic relationship section in **(C)**.

for prediction. Quantitative genetics is needed to build the genomic relationship matrix and then transform this matrix into data matrices for different data-mining methods (see Methods).

RESULTS

Pattern Discovery in Genomic Relationship and Phenotypic Variation

We started the pattern discovery by first visualizing the genomic relationship matrix of the 24 maize inbreds and 276 hybrids from a half diallel mating scheme (every pair of inbreds crossed). These inbreds were the founders of the nested association mapping (NAM) population, excluding sweet corn and popcorn lines (McMullen et al., 2009). Principal component analysis (PCA) based on 10 million SNPs revealed the major separation, agreeing with the germplasm origins: (1) temperate and mixed (TM), or (2) tropical and subtropical (TS) (Figure 2). The hybrids were also separated into three layers: TM × TM, TM × TS, and TS × TS. Hierarchical cluster analysis of the genomic relationship matrix enabled a direct reorganization of the columns and rows of the matrix, revealing a clear pattern among pairwise relationships, which is difficult to observe when inbreds were randomly ordered. Working with this set of materials with a clear diversity pattern facilitated the method research.

In parallel, we visualized the grouping pattern of 276 single-cross hybrids after analyzing their genotype data inferred from parental inbreds (Supplemental Figure 1). Additive and dominance genomic relationship matrices between hybrids were derived. Each hybrid has a half-sib relationship with 44 other hybrids. When the matrices were organized by sorting parental inbreds in the hierarchical cluster order, a global pattern was readily detected: relationships were closer for hybrids with both parents from the same group than other hybrids.

We phenotyped the 276 hybrids for flowering time (days to anthesis), ear height, and grain yield (Figure 2, Supplemental Figure 2, and Supplemental Table 1). With the ordered inbreds for both row and column, we then visualized the phenotypic values of hybrids for each trait (Figure 2D–2F). Agreeing with

the general knowledge for these traits, the TM × TS inter-group hybrids generally have trait values between two types of intra-group hybrids for flowering time and ear height; but for grain yield, inter-group hybrids have higher values. We estimated variance components of each trait. While additive variance is much higher than dominance variance for flowering time and ear height, it is smaller for grain yield (Supplemental Table 2).

Representative Subset Selection Methods

The patterns detected in inbred genomic relationship matrix, hybrid phenotypic value matrices, and hybrid genomic relationship matrices encouraged us to explore the training set design question from three different angles: genetic mating structure, cluster analysis, and graphic network analysis. For MaxCD, representative subset selection was designed following the pattern detected in the inbred genomic relationship matrix (see Methods). First, we selected a set of hybrids with non-overlapping parental inbreds to ensure they share one parent with those remaining hybrids for which performance needs to be predicted. Next, we selected a set of hybrids from pairs of inbreds most distant from each other. Conceptually, the combination of these two sets ensures a good sampling in the designed training set.

For PAM and FURS, representative subset selection was approached by treating hybrids as objects to be clustered based on their distances (PAM) or nodes within a complex network (FURS) (see Methods). Specifically, genetic covariance matrix among hybrids was obtained through merging two hybrid genomic relationship matrices (additive and dominance), which can be calculated directly from genotype data, weighted by two variance component estimates. In practice, the weights can be substituted with a ratio based on prior knowledge. Different values were compared to verify the sensitivity of ratio to the representative subset selection. For PAM, a dissimilarity matrix was obtained from the covariance matrix. All hybrids were partitioned into k clusters and the medoid of each cluster was chosen to form the representative subset (k hybrids). For FURS, an undirected and unweighted graph was obtained from the covariance matrix. k nodes from the network of all nodes (hybrids) were chosen to form the representative subset (k hybrids). Notably, the development and validation of designed methods for training sets required phenotypic data, which are not required in actual implementation of these methods.

We also implemented the previously published methods (PEVmean and CDmean through a genetic algorithm) (Akdemir et al., 2015) developed for inbred populations by using the covariance matrix containing both additive and dominance genomic relationship matrices (see Methods). For comparison, we also conducted random sampling of the training set.

Designing Training Set for Maize Hybrids from Diallel

We evaluated training set design methods (MaxCD, PAM, FURS, PEVmean, and CDmean) and random sampling. We performed genomic prediction by selecting a representative subset as the training set for each method, and the remaining hybrids as the testing set. These methods were conducted with multiple subdiallel populations. In each run, 18 inbreds were randomly selected to derive a set of 153 hybrids for method comparison, and each method (including random sampling) generated a single training set with a size of 9.2% of all hybrids, which is a fixed number by the design of MaxCD (Supplemental Table 3).

Prediction accuracy for all representative subset selection methods was higher than random sampling (Figure 3A–3C). They increased 16%–23% for flowering time, 10%–15% for ear height, and 40%–46% for grain yield. In addition, the differences between three designed methods and random sampling were significant based on a Mann–Whitney U test (P values <0.01). Moreover, representative subset selection improved accuracy compared with PEVmean and CDmean for all traits. We also conducted the training percentage analysis for PAM, FURS, PEVmean, CDmean, and random sampling, which have no restriction on the training set size. To achieve the same accuracy, random sampling generally required a larger training set than other methods. For example, for flowering time and ear height, PAM or FURS with 12 hybrids (8%) would achieve about the same level of prediction accuracy as random sampling with 21 hybrids (14%) (Supplemental Figure 3).

To understand the reasons underlying better prediction accuracy by representative subset selection methods, we examined two criteria (Laloë, 1993; Akdemir et al., 2015): (1) connectedness between the training and testing sets and (2) diversity within the training set. First, we quantified connectedness for each hybrid in the testing set using the maximum additive relationship value between this hybrid and all training hybrids. The values for representative subsets were distributed above 0.3, unlike random sampling where the values were negative, indicating that some testing hybrids are not well connected with the training set (Supplemental Figure 4). Second, we quantified diversity of the training set by obtaining the variance of the phenotypic values. With representative subset selection, variances of hybrid performance in the training set were generally larger than those obtained with random sampling (Supplemental Table 4). These results suggest that the joint effects of connectedness and diversity are underlying the higher prediction accuracy of representative subsets relative to randomly sampled subsets.

Designing Training Set for Inter-group Maize Hybrids from Factorial

Next, we focused on the analysis of the phenotype and genotype data from a factorial mating scheme, whereby only hybrids between TM and TS were examined. The 143 inter-group hybrids (TM × TS, Figure 2) were part of all maize hybrids from diallel, which include both inter-group and intra-group hybrids, and a new set of variance components were obtained (Supplemental Table 2). For MaxCD with factorial mating scheme, the same principle of maximizing connectedness and potential phenotypic diversity was followed to obtain a representative subset as the training set; for PAM and FURS, the same procedure was used as with diallel but with updated variance component estimates.

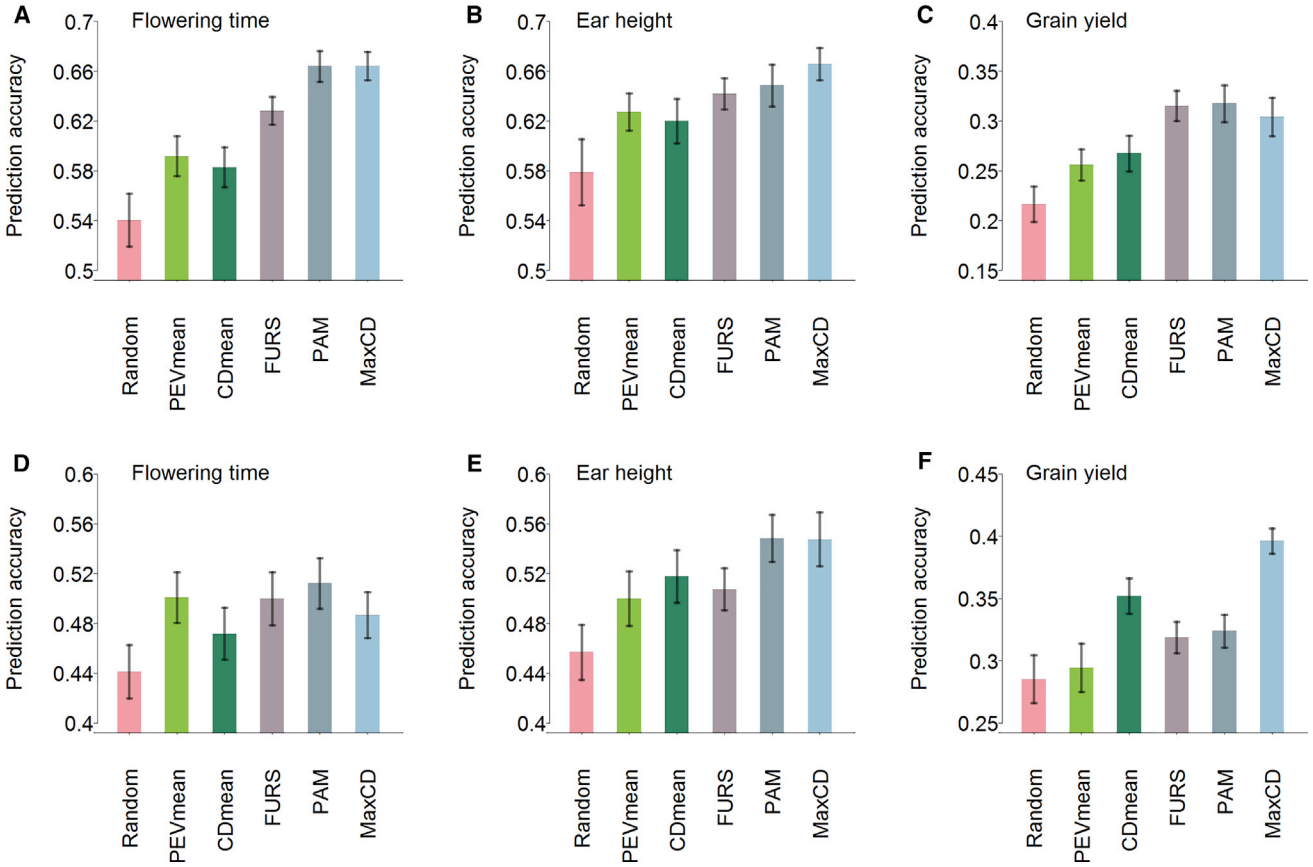

Figure 3. Representative Subset Selection Outperforms Random Sampling in Maize.
(A–C) Prediction accuracy by different sampling methods in maize diallel. **(A)** Flowering time, **(B)** ear height, **(C)** grain yield.
(D–F) Prediction accuracy by different sampling methods in maize factorial. **(D)** Flowering time, **(E)** ear height, **(F)** grain yield.
SE bar is from 50 runs of independent sampling.

Method comparison was conducted with multiple subfactorial populations. In each run the total number of hybrids was 80, and each method generated a single training set with a size of 12.5% of all the hybrids (Supplemental Table 3). Prediction accuracy values of representative subset selection methods were higher than random sampling for all three traits, although the difference was not always significant (Figure 3D–3F). Compared with random sampling, the accuracy of MaxCD, PAM, and FURS increased 10%–16% for flowering time, 11%–20% for ear height, and 12%–39% for grain yield. The reduction in the degree of advantages with factorial (inter-group hybrids) versus diallel (both inter- and intra-group hybrids) may be partially due to the reduced overall sample size. In addition, less pronounced patterns among inter-group hybrids played a role. For grain yield, heritability for factorial (0.60) was higher than diallel (0.51), which may partially explain the difference in prediction accuracy. However, because factorial is part of diallel, a direct comparison of two mating schemes would not be appropriate, and should be examined in detail (Fritsche-Neto et al., 2018).

Designing a Training Set for Wheat Hybrids

A dataset of 2556 hybrids from a diallel with 72 wheat inbred lines adapted to Central Europe (Zhao et al., 2015) was available to test these training set designs for a crop in which hybrid breeding is still at the early stage. These inbreds have been separated into two sets of 36 lines belonging to two heterotic groups. The PCA plot based on genotype data indicated that the separation between these two groups was not strong (Figure 4A and 4B), which agreed with the lack of intensive divergent selection. After ordering the inbreds by hierarchical clustering, a detectable grouping pattern emerged from the genomic relationship matrix (Figure 4C). Phenotypically, the mean grain yield of the inter-group hybrids was 10.833 Mg ha^{-1}, slightly higher than the means of crosses within each of the two groups (10.827 Mg ha^{-1} and 10.670 Mg ha^{-1}) (Figure 4D), in general agreement with what was expected from the genomic relationship matrix.

Method comparison was conducted with multiple subdiallel populations. In each run, the total number of hybrids was 1770, and each method (including random sampling) generated a single training set with a size of 2.5% of all the hybrids (Supplemental Table 3). The representative subset with each method resulted in a higher prediction accuracy value than random sampling (Figure 4E). The value increased 18% for MaxCD, 11% for PAM, and 4% for FURS. Variance of prediction accuracy was lower for these methods, which is desirable, than for random sampling. While PEVmean, CDmean, and FURS had higher accuracy than random sampling, the *P* values >0.01 showed no statistically significant difference. In the training percentage

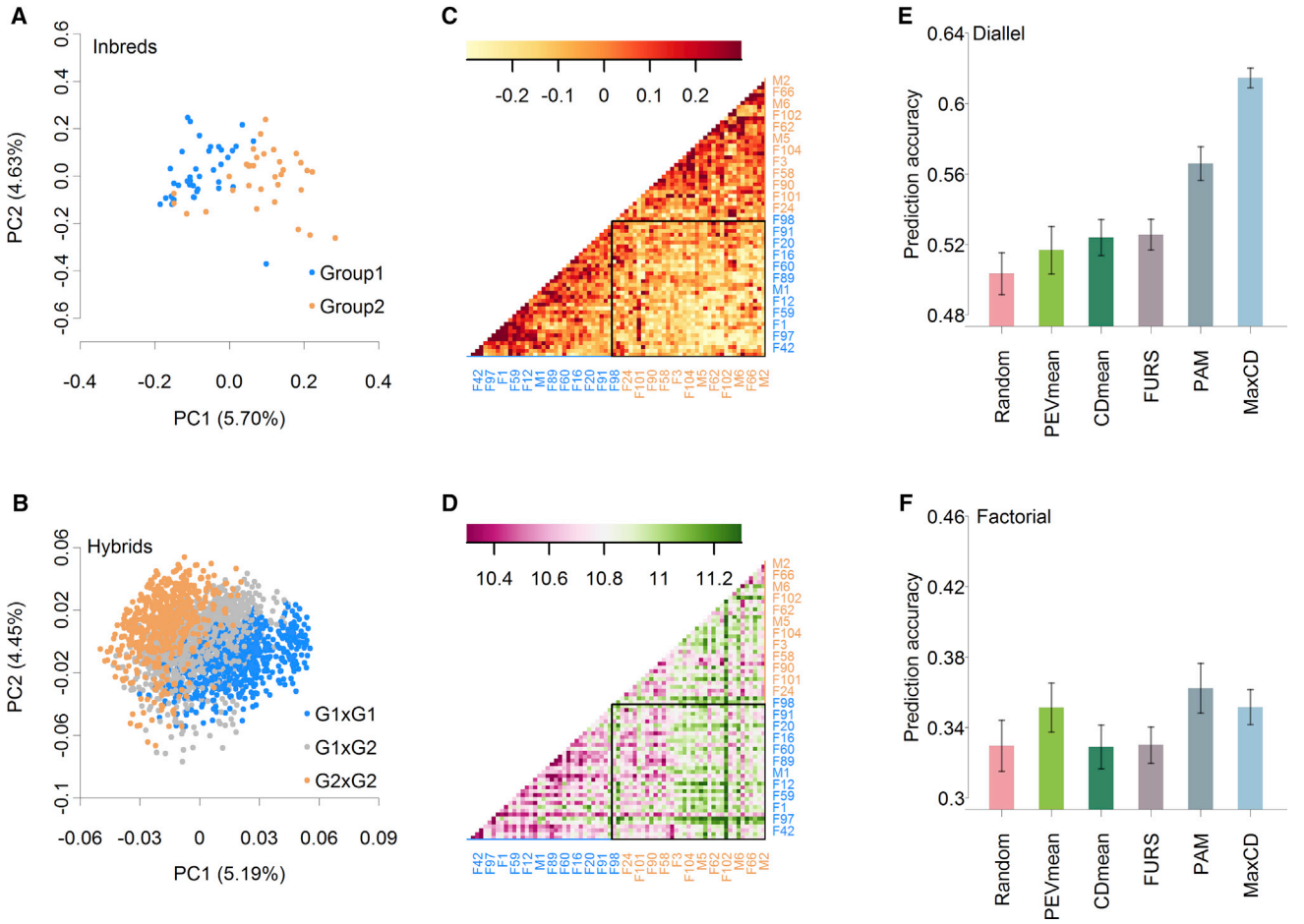

Figure 4. Representative Subset Selection Methods Applied to Wheat Hybrids.

(A) PCA plot for inbreds.

(B) PCA plot for hybrids.

(C) Genomic relationships among inbreds which were ordered by hierarchical clustering.

(D) Grain yield for hybrids (Mg ha^{-1}). Inbred parents of the hybrids were ordered by hierarchical clustering for rows and columns.

(E) Mean prediction accuracy by different sampling methods in diallel.

(F) Mean prediction accuracy by different sampling methods in factorial.

SE bar is from 50 runs of independent sampling. In **(C)** and **(D)**, only one of every three inbreds was shown to avoid the overlapping of labels, and the section corresponding to inter-group hybrids (from factorial) is boxed.

analysis, the advantages for PAM and FURS were observed at the lower sample size end (Supplemental Figure 5).

We also extracted a factorial mating scheme for 420 hybrids (35 females by 12 males) with the original, observed phenotype data (Zhao et al., 2015). Applying representative subset selection to this dataset, PEVmean, PAM, and MaxCD showed slightly better prediction accuracy than random sampling (Figure 4F), which may be explained by less strong patterns observed in genotype relationships in factorial compared with diallel. Restricting the method comparison to the wheat diallel with the original observed phenotype data of 1604 hybrids (Zhao et al., 2015), we obtained similar results. There was no significant difference between random sampling and methods of PEVmean, CDmean, and FURS, but PAM had significantly higher accuracy than other methods (Supplemental Figure 6). Collectively, these results suggest that representative subsets work well even in populations without strong subdivision, and that pattern detection and exploitation should be conducted with multiple methods.

Designing a Training Set for Rice Hybrids

Finally, we examined a rice dataset of 1495 elite hybrids (Huang et al., 2015). Of these hybrids, most (1439) were from *indica* × *indica* crosses, and we chose to study this set for its dominant sample size (Figure 5A and 5B). Because these hybrids constituted an incomplete factorial design, the representative subset was selected by PAM and FURS, but not MaxCD. Sixty percent of 1439 hybrids were randomly sampled for multiple subpopulation analysis. In each run, a single training set was constructed for each method with a size of 5% of all the hybrids (Supplemental Table 3).

Genomic prediction was conducted for heading date, height, grain weight, and grain length. For all four traits, both PAM and

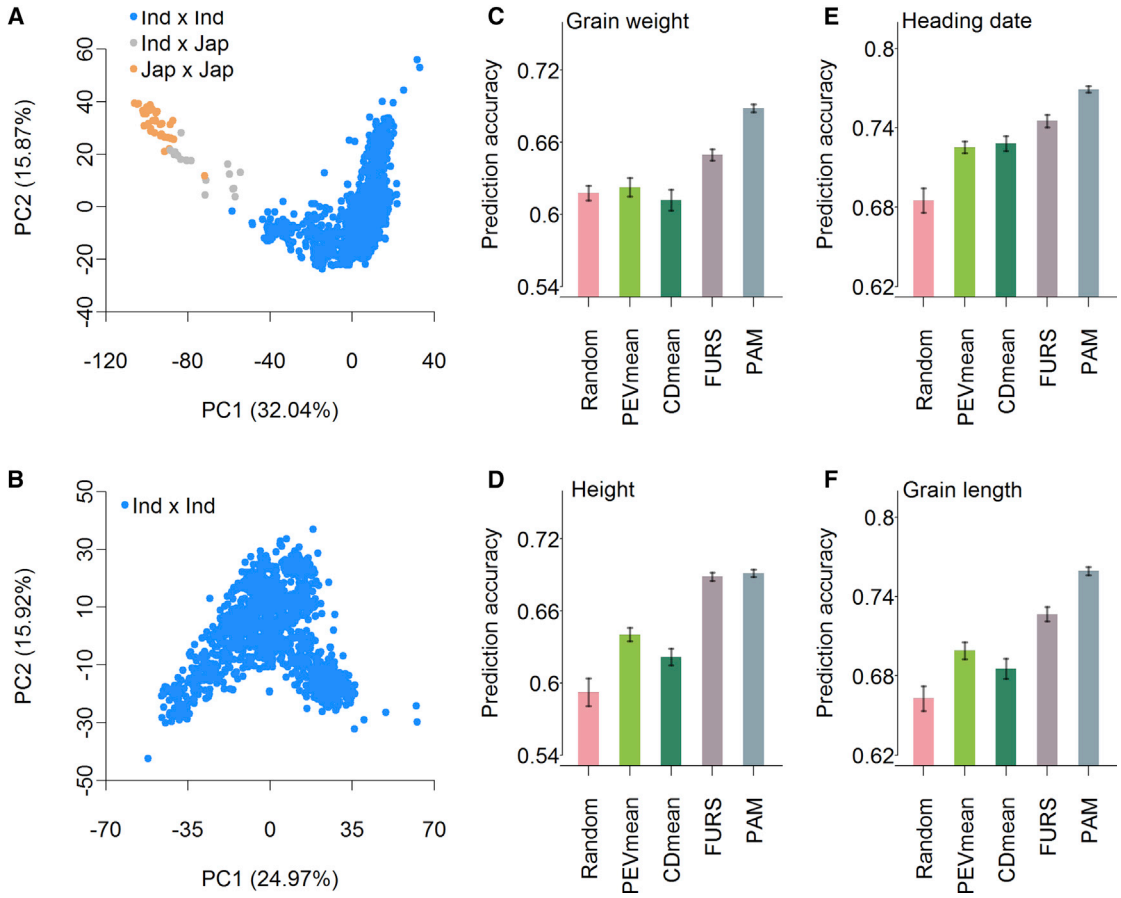

Figure 5. Representative Subset Selection Methods Applied to Rice Hybrids.
(A) PCA plot for all rice hybrids.
(B) PCA plot for *indica* × *indica* hybrids.
(C) Prediction accuracy by different sampling methods for grain weight (g).
(D) Height (cm).
(E) Heading date (day).
(F) Grain length (mm).
Ind, *indica*; Jap, *japonica*. SE bar is from 50 runs of independent sampling.

FURS gave significantly better predictions than random sampling (Figure 5C–5F). The prediction accuracy increased 9%–13% for grain length, 5%–11% for grain weight, 21%–25% for heading date, and 11%–12% for height. PEVmean and CDmean performed better than random sampling but not as well as PAM and FURS. In the training percentage analysis, the advantage of PAM and FURS was generally evident (Supplemental Figure 7). These results demonstrated that representative subset selection by either PAM or FURS could improve prediction efficiency over the use of a random training set of hybrids. Because of these hybrids were from a nationwide variety registration (Huang et al., 2015), further extensive study of the representative subset is likely generate informative findings.

Using these rice data, we also investigated how different variance component ratios affect the final prediction accuracy. This represents a scenario in which no empirical estimate is available. We applied a sequential number of 0.1–1 with an interval of 0.1 for ratio of dominance to additive variances (Supplemental Table 5). Results across four traits were consistent in showing the insensitivity of the variance ratio.

Overall Assessment

Through comparing different training set design methods, we showed that representative subset selection methods performed better than random sampling and that MaxCD, PAM, and FURS performed better than PEVmean and CDmean. On the other hand, different methods outperformed other ones in different species–trait combinations. Among three newly examined methods, MaxCD has the advantages of an easy layout, being non-specific to trait, and no requirement of a ratio between additive and dominance variance. The actual size of the representative subset can be expanded to ensure adequate sampling of the overall landscape of the genomic relationship matrix after the ordering. This method is also a good choice if a defined set of inbreds is to be crossed to establish an initial training set to explore the hybrid space. MaxCD may also be examined for training set design in inbreds and testcrosses, where every *k*th

individual is selected based on the order from the hierarchical cluster analysis of the genomic relationship.

In other scenarios where only a subset of all possible combinations of inbreds are relevant for investigation due to physiological or agronomic constraints, PAM would be a good choice. PAM and FURS enjoy the flexibility to generate a representative subset with different sizes. Visualization of the selection on top of the inbred genomic relationship matrix can help check the space coverage. To implement PAM, FURS, PEVmean, and CDmean for genomic prediction in hybrids, the need of a ratio between additive and dominance variance can be satisfied based on the domain knowledge, the analysis of existing data, or the preliminary experiment to estimate the ratio. This requirement also stresses the iterative nature of research in training set design as breeding and agronomic practices evolve. For practical applications, different selected sets made for different traits, through different methods (PAM, FURS, PEVmean, and CDmean), can be tabulated to identify a final common set.

DISCUSSION

The findings of this study were first uncovered through the empirical maize research with a diallel, then confirmed through analysis of inter-group hybrids and two previously released datasets in wheat and rice (Huang et al., 2015; Zhao et al., 2015). Overall, the designed methods showed promising prediction results. The advantages of designed methods using data mining and design thinking are mainly driven by the underlying patterns in genomic relationships between materials, and patterns in relationships between genomic and phenotype.

Conceptually, the divergence of temperate from tropical and subtropical materials is analogous to the divergence of two heterotic groups (typically stiff stalk and non-stiff stalk) in actual breeding materials. Even though NAM founders were used in the initial investigation, we found that pattern discovery with the proof-of-concept small data was helpful in exploring different methods, which were further investigated and validated with large datasets from more relevant germplasm such as in wheat and rice. In theory, estimation of ratio between additive and dominance variances is needed for all methods exploring the genetic relationship of hybrids. If no phenotype data can be mined for estimation, the prior knowledge of relevant traits available in the literature can be used. On the other hand, this also highlights the iterative process of updating training sets as breeding and agronomic practices evolve.

In this study, we investigated using small sets of hybrids in three species to predict the performance of the large genotype space of all hybrids. Our choice of the percentage was based on preliminary investigation to obtain reasonable prediction accuracy, and the generally recommended 15% from data-mining literature (Han et al., 2011). The upper bound of percentage we examined was 32%, which provided a comparison context. The exact optimal percentage of training set would vary with the size of the overall set and can be data specific. Given the large number of hybrid combinations, focus needs to be on optimal design with a small training set, rather than trying to identify the optimal percentage. Our analyses did show that the increase in predication accuracy generally started leveling off around or before 15% for eight species–trait combinations. In practice, available resources will have to be considered, and the initial training set may be gradually expanded as more data are collected and mined.

Assembling optimal training sets enables a forward-looking, designed approach so that a prediction-guided genotype search can be conducted with the most relevant germplasm and cultivation for data generation and knowledge discovery. This complements the efforts to mine existing performance data for current and historical hybrids to develop the genotype–phenotype prediction model, although consideration must be given to the changes of planting density and nitrogen application rate associated with the historical records (Duvick, 2005). Meanwhile, a few studies explored efficient genomic mating (Kinghorn and Shepherd, 1994; Shepherd and Kinghorn, 1998; Kinghorn, 2011; Akdemir and Isidro Sanchez, 2016), most of which had successful results from simulation studies. The hybrids training set selection using data-mining methods incorporated with efficient genomic mating algorithms will provide new thinking with regard to the overall hybrid breeding process. Furthermore, different data-mining techniques (Han et al., 2011) can be applied to questions beyond the initial training set design—for example, designing the validation set to balance prediction improvement and commercial hybrid identification, or finding the optimal balance between short-term gain and long-term potential of the program.

Streamlining genomic selection using analytical methods could greatly reduce costs and maximize profits (Riedelsheimer et al., 2012; Technow et al., 2014; Xu et al., 2014). Pattern discovery, prediction-guided search, and accurate prediction based on sound pattern discovery are likely to become more critical to reducing time and cost when these resources are in short supply. Data-mining methods we introduced in this and earlier studies (Rincent et al., 2012; Akdemir et al., 2015; Isidro et al., 2015) could be further improved and extended. More broadly, data-mining tools could also be tested for identifying the most important and relevant genetic materials for different genomics-enabled processes: parental selection and inbred development, assembling core phenotyping panels from gene bank collections, or narrowing down the candidate list for comprehensive phenotyping. Extensive genomic profiling with transcriptomics and metabolomics across time points and tissues would greatly benefit from a well-designed set of genetic materials. Research findings through mining genomic and phenomic data can then inform the designing decisions in future studies of complex traits in crops and human diseases (de los Campos et al., 2010; Wray et al., 2013; Gamazon et al., 2015).

METHODS

Maize Hybrids, Phenotyping, and Genotype Information

The germplasm collection used in the present study consisted of 24 diverse parents (Flint-Garcia et al., 2005). These parental lines were classified into two main groups according to germplasm origin and PCA: TM group with 11 inbred lines (B73, B97, Ky21, M162W, Mo17, MS71, Oh43, OH7B, M37W, Mo18W, and Tx303) and TS group with 13 inbred lines (CML52, CML69, CML103, CML228, CML247, CML277, CML322, CML333, Ki3, Ki11, NC350, NC358, and Tzi8). Single-cross hybrids of these inbreds were developed in a half diallel mating scheme, including TM × TM and TS × TS intra-group hybrids, and TM × TS inter-group hybrids.

We evaluated 276 single-cross hybrids at two locations (Columbia, MO and Clayton, NC) in 2005 and 2006 for three traits: flowering time (day), ear height (cm), and grain yield (Mg ha^{-1}). Flowering time was measured as the number of days to pollen shed from planting (i.e., days to anthesis), ear height as the distance from the ground to the primary ear-bearing node, and grain yield as harvest weight corrected to 15.5% moisture. After combining data from two locations, a single set of best linear unbiased prediction values was calculated using PROC MIXED in SAS 9.3 software (SAS Institute).

Genotype data for inbreds were extracted from the Maize HapMap V2 (Chia et al., 2012) at www.panzea.org. There were 10 296 310 SNPs for the 24 inbred lines. Missing genotypes were imputed with Beagle 4.1 (Browning and Browning, 2016). Genotype data for hybrids were inferred from SNP data of the parental inbreds.

Additive and dominance variance components in the diallel were estimated by using $\sigma_A^2 = 2\sigma_{GCA}^2$ and $\sigma_D^2 = \sigma_{SCA}^2$, where σ_{GCA}^2 and σ_{SCA}^2 are variances of general and specific combining ability. Additive and dominance variance components in the factorial design were estimated as $\sigma_{A\_P1}^2 = \sigma_{GCA\_P1}^2$, $\sigma_{A\_P2}^2 = \sigma_{GCA\_P2}^2$, and $\sigma_D^2 = \sigma_{SCA}^2$, where P_1 is one group of individuals and P_2 is the second group of individuals. These component estimates were used in generating the covariance among hybrids.

Pattern Discovery and Visualization

For inbreds, the genomic relationship matrix (**G**) was calculated as $\mathbf{G} = \frac{\mathbf{WW}'}{2\sum p_i q_i}$, where **W** is a matrix of centered genotypes, p_i is the major allele frequency of locus i q_i is the minor allele frequency. PCA was conducted with **G** as the input using the function prcomp in R (R Core Team, 2013). A dissimilarity matrix was obtained by transforming **G** using the agglomeration method with the option of "centroid," and this dissimilarity matrix was used for hierarchical cluster analysis (HCA) to obtain the sorting order of inbreds using the function hclust in R (R Core Team, 2013).

The additive genomic relationships (**A**) among hybrids was calculated as $\mathbf{A} = \frac{\mathbf{ZZ}'}{2\sum p_i q_i}$ where **Z** is a matrix of centered genotypes with $(0–2p_i)$, $(1–2p_i)$, and $(2–2p_i)$ for three different genotypes, p_i is the major allele frequency of locus i q_i is the minor allele frequency. The dominance genomic relationship matrix (**D**) was calculated as $\mathbf{D} = \frac{\mathbf{HH}'}{2\sum p_i q_i (1 - 2p_i q_i)}$, where **H** is a matrix of centered genotypes with $(0–2p_iq_i)$ and $(1–2p_iq_i)$ for homozygous and heterozygous at locus i. The covariance matrix among hybrids was obtained as $\sigma_A^2\mathbf{A} + \sigma_D^2\mathbf{D}$, where σ_A^2 and σ_D^2 are additive and dominance variances. **A**, **D**, and the covariance matrix can be organized by the order of inbreds from HCA.

Genomic Prediction

Genomic prediction was conducted with the established genomic best linear unbiased prediction approach considering both additive and dominance effects (Bernardo, 1994; Xu et al., 2014). Preliminary analysis with other methods (Morota and Gianola, 2014) did not identify sizable differences. The training sets were established by either the representative subset selection methods or random sampling for the genotype–phenotype prediction model. The remaining, non-selected hybrids were used as the testing set. Prediction accuracy was calculated as the correlation between the predicted genotypic values and the observed phenotypes. Preliminary analysis indicated that the number of hybrids in the training set determined by MaxCD is a reasonable number for comparison, while the other two methods have the flexibility to select an arbitrary number.

Five sets of population data were analyzed: maize diallel, maize factorial, wheat diallel, wheat factorial, and rice incomplete factorial. Genomic prediction was conducted by selecting 2.5%–12.5% of the whole set of hybrids as the training set (Supplemental Table 3). The choice of this range followed the suggested value in representative subset selection (<15%) (Han et al., 2011). For maize diallel, we conducted additional training percentage analyses for random sampling, PEVmean, CDmean, PAM, and FURS by varying the training sample size from 2% to 32% of the original whole set. Training percentage analyses for wheat diallel and rice incomplete factorial were also conducted for PAM, FURS, and random sampling, but not for PEVmean and CDmean due to the high computational cost.

Multiple subpopulation analysis was conducted. A total of 50 subdiallel or subfactorial populations were randomly sampled, and the size of these subpopulations were from 55% to 70% of the original whole populations (Supplemental Table 3). This size was chosen to ensure that the resulting populations are not too small, but still vary among them. Accordingly, 50 prediction accuracy values were generated for different training set selection methods with the same training set size: random sampling, PEVmean, CDmean, PAM, FURS, and MaxCD.

Partitioning around Medoids

As a clustering method, PAM classifies objects into clusters by minimizing the sum of dissimilarities between the objects labeled to be in a cluster and a designated center object (medoid) of that cluster (Kaufman and Rousseeuw, 1987, 2009). PAM is a good candidate for training set design as it provides both the partition of all objects into clusters and representative objects of clusters. In practice, in addition to genotype data, a ratio between additive and dominance variance components is needed to derive the covariance matrix and can be fulfilled with prior knowledge of the target trait. The covariance matrix for hybrids was then transformed to a dissimilarity matrix.

Input: a user defined k number of clusters (k medoids) to be found among n objects; a dissimilarity matrix **D** among n objects.

Output: a set of k clusters and their medoids.

Steps:

(1) arbitrarily choose k objects in D as the initial representative objects (o_j, $j = 1$ to k);
(2) assign each remaining object to the cluster with the nearest representative object;
(3) randomly select a non-representative object o_{random};
(4) compute the total cost of swapping the representative object o_j with o_{random};
(5) if the cost of swapping $S < 0$ then swap o_j with o_{random} to form a new set of k representative objects;
(6) repeat steps 2–5 until no change.

The total cost is defined as $T = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i)$, where T is the sum of the absolute error for all objects p in the dataset; C_i includes objects in individual cluster; and o_i is the representative object of C_i. Cost of swapping (S) is the difference between current total cost and past total cost.

Fast and Unique Representative Subset Selection

As a graphic network method, FURS selects a subset of nodes at the center of the communities in a large-scale network without explicitly performing community detection (Mall et al., 2013). In practice, the need and solution to a variance component ratio to obtain the covariance matrix are the same as for PAM. The covariance matrix for hybrids was transformed into a correlation matrix. Applying a threshold value of 0.2 for either linked or unlinked pairwise connection to the correlation matrix, we obtained a matrix of an undirected and unweighted graph G = (V, E), where V is a finite set of nodes (hybrids) and E is a finite set

of edges (links). The tuning parameter was applied in this method with the number of 0.2 to transform weighted matrix to the unweighted graph, which was selected after multiple practices.

Input: a user defined size *k for* representative nodes to be chosen among all nodes; a network G = (V, E), a pair (*u*, *v*) is defined as an edge from node *u* to node *v*.

Output: a subset of representative nodes S.

Steps:
(1) compute a list of nodes with their corresponding degree centrality values L = (V, D(V)), where D is the degree distribution function of V;
(2) order the nodes based on their degree centrality values in descending order;
(3) choose the node with highest degree centrality as one object in S;
(4) deactivate the immediate neighbors of this node because they can be reached directly from this node;
(5) select the node with highest degree centrality among the active nodes and place it into S;
(6) repeat steps 4 and 5 until all nodes are deactivated; stop the loop if |S| reaches *k*; otherwise
(7) reactivate the deactivated nodes and repeat steps 3–6 until |S| = *k*.

Maximization of Connectedness and Diversity

We proposed MaxCD based on the characteristics of half diallel and patterns embedded in a genomic relationship matrix among inbreds. It selects hybrids for the training set by ensuring that every hybrid to be predicted shares at least one parent with a hybrid in the training set, and that both intra-group and inter-group hybrids are sampled (Supplemental Figure 8). Earlier research in genomic prediction (Jacobson et al., 2014; Xu et al., 2014; Kadam et al., 2016) and mating design (Stich, 2009) provided relevant findings for us to explore this representative subset sampling design.

Input: inbred genomic relationship matrix.

Output: a set of representative hybrids.

Steps:
(1) order inbreds on genomic relationship matrix by HCA;
(2) remove reciprocals and selfs to obtain an isosceles right triangle;
(3) select every other hybrid on the long edge as representative;
(4) starting from vertex, select half of the consecutive hybrids along the height of the triangle as representative.

Pattern examination in step 1 for genomic relationship matrix among inbreds guides the selection of representative hybrids. After step 2, two properties are found in the isosceles right triangle: (1) two neighboring hybrids have a closer relationship than each with a non-neighboring hybrid; and (2) hybrids are interconnected through their shared parental inbreds and genetic relationship of parental inbreds. Step 3 aims for strong connectedness between the selected hybrids and leftovers (because the selected hybrids are half sibs of unselected ones) and diversity (because the inbreds are sorted and one hybrid was sampled for each inbred). Step 4 aims for adding diversity because hybrids from less distant inbreds need to be sampled and their phenotypic values can be different from hybrids from adjacent inbreds. Applying MaxCD to a half diallel with the number of *n* parental inbreds, the size of training set is ceiling ($^3/_4 n$), the size of testing set is floor ($^1/_2 n(n-1) - {}^3/_4 n$). The same ordering and sampling principles can be extended to a diallel with more than two groups of inbreds. MaxCD for factorial was designed based the same sampling principles.

Maximization of Connectedness and Diversity for Factorial

A factorial mating scheme is commonly used in crop-breeding programs when two known groups of inbreds are used as either males or females. The inter-group hybrids within a diallel are equivalent to hybrids from a factorial mating scheme. MaxCD in a factorial mating scheme is designed as follows (Supplemental Figure 8).

Input: inbred genomic relationship matrix.

Output: a set of representative hybrids.

Steps:
(1) order inbreds on genomic relationship matrix by HCA and retain only one section corresponding to the inter-group hybrids (factorial part);
(2) draw both diagonal lines of the rectangle;
(3) identify all hybrids in cells that are crossed by the diagonal lines;
(4) select every other hybrid along each diagonal line as representative.

Step 1 is made to reveal the general pattern of hybrids from four areas: upper left, upper right, lower left, and lower right. The closer the hybrids are to one of the four corners, the more typical they represent other hybrids from that area. For a factorial mating scheme with one set of n_1 inbreds and the second set of n_2 inbreds, MaxCD results in a training set of n_1 (given that $n_1 > n_2$) and a testing set of $n_1 n_2 - n_1$.

PEVmean and CDmean

Methods of PEVmean and CDmean were conducted by using the R package STPGA version 4.0 (Akdemir, 2014; Akdemir et al., 2015) and function GenAlgForSubsetSelectionNoTest. It uses a genetic algorithm to select training individuals so that optimality criterion is reached. We used optimality criterion of "PEVMEAN" for method PEVmean and criterion of "CDMEAN" for method CDmean. The number of iterations in implementing the genetic algorithm was 100 when population size was larger than 1000, and 200 otherwise. All other parameters were set as default.

Wheat Hybrids

A set of 1604 wheat hybrids produced from crosses among 120 female lines and 15 male lines were evaluated for grain yield in 11 environments (Zhao et al., 2015). Grain yield data for all 9045 unique hybrids from these 135 parental lines were predicted based on those of the phenotyped individuals (Zhao et al., 2015). Following earlier findings, we extracted a half diallel of 2566 hybrids, which were crosses among 72 inbreds (SI Dataset). Genotypes of hybrids were inferred from 72 inbreds with 17 372 SNPs. In addition, we also extracted a factorial mating scheme for 420 hybrids (35 females by 12 males) with the original observed phenotype data. The same procedures as in maize were used to obtain the variance components (Supplemental Table 6), which were then used in constructing the covariance among hybrids.

Rice Hybrids

The rice hybrid panel consisted of 1495 hybrids (Huang et al., 2015). Most of these hybrids (1,439) were *indica* × *indica* crosses. The rest were 18 *indica* × *japonica* crosses and 38 *japonica* × *japonica* crosses. The 1495 hybrids were evaluated for heading date (day), height (cm), grain weight (g), and grain length (mm) for 2 years at two locations in the original publication. We focused on the analysis of 1439 *indica* × *indica* hybrids and the phenotype data from Sanya (presented in the main text of the original manuscript) for genomic prediction in this study. Analysis of the second environment was also conducted. These hybrids were directly genotyped with 1 654 030 SNPs, and no genotype information of parental inbreds was available (Huang et al., 2015). We estimated the additive and dominance variance components through the mixed model (Supplemental Table 6).

Variance Component Ratio

For PAM, FURS, PEVmean, and CDmean, a variance ratio is needed to construct the covariance matrix. In our method comparison, where

phenotype data are available, variance component ratios were directly obtained through either the standard analysis procedures of mating designs when the layout of inbred crossing is known (maize and wheat) or the mixed model procedure by fitting marker-derived additive and dominance relationship matrices simultaneously when the layout of the inbred crossing is not available (rice).

To examine the scenario in which no phenotype data are used and no empirical estimates are available, we conducted computer stimulations based on 40% of the rice data by setting up the ratio of dominance to additive variance to vary from 0.1 to 1.0 to construct the input covariance matrix for representative subset selection. Once the training set was selected, genomic prediction was conducted to assess the prediction accuracy. This procedure was repeated 50 times for each trait and each method (PAM, FURS, and random sampling) to obtain the average prediction accuracy.

Data and Code Access

Original and processed phenotype data for maize hybrids, processed phenotype data for wheat hybrids, and original rice hybrids are included as Supplemental Data 1, 2, and 3, respectively. R scripts for representative subset selection through MaxCD, PAM, and FURS are available at GitHub (https://github.com/TingtingGuo0722/OptimalDesign).

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work is supported by NSF Plant Genome Research Program [IOS-1238142] and ISU Plant Sciences Institute Faculty Scholar Fund.

AUTHOR CONTRIBUTIONS

J.Y., T.G., and R.J.W. designed the overall project; T.G., X.Y., X.L., H.Z., C.Z., R.J.W., and J.Y. conducted method development and data analysis; S.F.-G., M.D.M., and J.B.H. conducted the maize experiment and provided comments; T.G. and J.Y. wrote the paper.

ACKNOWLEDGMENTS

We appreciate late Steve Szalma for trait evaluation and data collection in the maize experiment. No conflict of interest declared.

Received: October 8, 2018
Revised: December 14, 2018
Accepted: December 24, 2018
Published: January 5, 2019

REFERENCES

Akdemir, D. (2014). Training population selection for (breeding value) prediction. ArXiv, arXiv:1401.7953.

Akdemir, D., and Isidro Sanchez, J. (2016). Efficient breeding by genomic mating. Front. Genet. **7**:210.

Akdemir, D., Sanchez, J.I., and Jannink, J.L. (2015). Optimization of genomic selection training populations with a genetic algorithm. Genet. Sel. Evol. **47**:38.

Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M., Simianer, H., and Schön, C.-C. (2011). Genome-based prediction of testcross values in maize. Theor. Appl. Genet. **123**:339–350.

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. **34**:20–25.

Bernardo, R. (2010). Breeding for Quantitative Traits in Plants, 2nd edn (Woodbury, MN: Stemma Press).

Browning, B.L., and Browning, S.R. (2016). Genotype imputation with millions of reference samples. Am. J. Hum. Genet. **98**:116–126.

Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., and Glaubitz, J.C. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. **44**:803–807.

de los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. **11**:880–886.

Duvick, D.N. (2005). The contribution of breeding to yield advances in maize (Zea mays L.). Adv. Agron. **86**:83–145.

Elhamifar, E., Sapiro, G., and Sastry, S. (2014). Dissimilarity-based sparse subset selection. ArXiv, arXiv:1407.6810.

Flint-Garcia, S.A., Thuillet, A.C., Yu, J., Pressoir, G., Romero, S.M., Mitchell, S.E., Doebley, J., Kresovich, S., Goodman, M.M., and Buckler, E.S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. **44**:1054–1064.

Fritsche-Neto, R., Akdemir, D., and Jannink, J.L. (2018). Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. Theor. Appl. Genet. **131**:1153–1162.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, and Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. **47**:1091–1098.

Han, J., Pei, J., and Kamber, M. (2011). Data Mining: Concepts and Techniques (Waltham, MA: Elsevier).

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., Li, W., Zhan, Q., Cheng, B., and Xia, J. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat. Commun. **6**:6258.

Isidro, J., Jannink, J.L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M.E. (2015). Training set optimization under population structure in genomic selection. Theor. Appl. Genet. **128**:145–158.

Jacobson, A., Lian, L., Zhong, S.Q., and Bernardo, R. (2014). General combining ability model for genomewide selection in a biparental cross. Crop Sci. **54**:895–905.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data clustering: a review. ACM Comput. Surv. **31**:264–323.

Kadam, D.C., Potts, S.M., Bohn, M.O., Lipka, A.E., and Lorenz, A.J. (2016). Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. G3 (Bethesda) **6**:3443–3453.

Kaufman, L., and Rousseeuw, P. (1987). Clustering by Means of Medoids (Amsterdam: North-Holland).

Kaufman, L., and Rousseeuw, P.J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis (New York: John Wiley & Sons).

Kinghorn, B., and Shepherd, R. (1994). A tactical approach to breeding for information-rich designs. In: Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production, 7–12 August, Guelph. pp. 255–261.

Kinghorn, B.P. (2011). An algorithm for efficient constrained mate selection. Genet. Sel. Evol. **43**:4.

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. Genet. Sel. Evol. **25**:1.

Leskovec, J., and Faloutsos, C. (2006). Sampling from large graphs. In KDD '06, pp. 631–636. https://doi.org/10.1145/1150402.1150479.

Lorenz, A.J., and Smith, K.P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Sci. **55**:2657–2667.

Mall, R., Langone, R., and Suykens, J.A.K. (2013). FURS: Fast and Unique Representative Subset selection retaining large-scale community structure. Soc. Netw. Anal. Min. **3**:1075–1095.

Marulanda, J.J., Melchinger, A.E., and Würschum, T. (2015). Genomic selection in biparental populations: assessment of parameters for optimum estimation set design. Plant Breed. **134**:623–630.

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., and Bottoms, C. (2009). Genetic properties of the maize nested association mapping population. Science **325**:737–740.

Miedaner, T., Zhao, Y., Gowda, M., Longin, C.F.H., Korzun, V., Ebmeyer, E., Kazman, E., and Reif, J.C. (2013). Genetic architecture of resistance to *Septoria tritici* blotch in European wheat. BMC Genomics **14**:1.

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. Front. Genet. **5**:363.

Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2012). Crop genomics: advances and applications. Nat. Rev. Genet. **13**:85–96.

Pan, F., Wang, W., Tung, A.K., and Yang, J. (2005). Finding representative set from massive data. In: Fifth IEEE International Conference on Data Mining (ICDM '05): IEEE. 8 pp.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A.E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. **44**:217–220.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodriguez, V.M., Moreno-Gonzalez, J., Melchinger, A., and Bauer, E. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics **192**:715–728.

Shepherd, R., and Kinghorn, B. (1998). A tactical approach to the design of crossbreeding programs. In: Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production: 11–16 January, Armidale. pp. 431–438.

Stich, B. (2009). Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. Genetics **183**:1525–1534.

R Core Team. (2013). R: A Language and Environment for Statistical Computing. http://www.r-project.org/.

Technow, F., Schrag, T.A., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A.E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics **197**:1343–1355.

Tester, M., and Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. Science **327**:818–822.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. **14**:507–515.

Xu, S., Zhu, D., and Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc. Natl. Acad. Sci. U S A **111**:12456–12461.

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S.E., Roozeboom, K.L., Wang, D., Wang, M.L., and Pederson, G.A. (2016). Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants **2**:16150.

Zeng, D., Tian, Z., Rao, Y., Dong, G., Yang, Y., Huang, L., Leng, Y., Xu, J., Sun, C., Zhang, G., et al. (2017). Rational design of high-yield and superior-quality rice. Nat. Plants **3**:17031.

Zhao, Y., Li, Z., Liu, G., Jiang, Y., Maurer, H.P., Würschum, T., Mock, H.-P., Matros, A., Ebmeyer, E., and Schachschneider, R. (2015). Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc. Natl. Acad. Sci. U S A **112**:15624–15629.