Published April 28, 2015

RESEARCH

# Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation

John M. Hickey,* Susanne Dreisigacker, Jose Crossa, Sarah Hearne, Raman Babu, Boddupalli M. Prasanna, Martin Grondona, Andres Zambelli, Vanessa S. Windhausen, Ky Mathews, and Gregor Gorjanc

## ABSTRACT

Genomic selection offers great potential for increasing the rate of genetic improvement in plant breeding programs. This research used simulation to evaluate the effectiveness of different strategies for genotyping and phenotyping to enable genomic selection in early generation individuals (e.g., $F_2$) in breeding programs involving biparental or similar (e.g., backcross or top cross) populations. By using phenotypes that were previously collected in other biparental populations, selection decisions could be made without waiting for phenotypes that pertain directly to the selection candidate to be collected, a process that would take at least three growing seasons. If these phenotypes were collected in biparental populations that were closely related to the selection candidates, only a small number of markers (e.g., 200–500) and a small number of phenotypes (e.g., 1000) were needed to achieve effective accuracy of estimated breeding values. If these phenotypes were collected in biparental populations that were not closely related to the selection candidates, as many as 10,000 markers and 5000 to 20,000 phenotypes were needed. Increasing marker density beyond 10,000 markers did not show benefit and in some scenarios reduced the accuracy of prediction. This study provides a guide, enabling resource allocation to be optimized between genotyping and phenotyping investment dependent on the population under development.

J.M. Hickey, The Roslin Institute and Royal (Dick) School of Veterinary Studies, Univ. of Edinburgh, Easter Bush Research Center, Midlothian EH25 9RG, UK; S. Dreisigacker, Global Wheat Program, International Maize and Wheat Improvement Center (CIMMYT), Apdo. 06600 México D.F., México; J. Crossa and K. Mathews, Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. 06600 México D.F., México; S. Hearne, Genetic Resources Program, International Maize and Wheat Improvement Center (CIMMYT), Apdo. 06600 México D.F., México; R. Babu and B.M. Prasanna, Global Maize Program, International Maize and Wheat Improvement Center (CIMMYT), Apdo. 06600 México D.F., México; M. Grondona and A. Zambelli, Center for Biotechnology Research, Advanta Semillas, Ruta 226 Km 60.5, (7620) Balcarce, Argentina; V.S. Windhausen, Saaten Union Recherche, 160 Avenue de Flandre, 60190 Estrées Saint Denis, France; Gregor Gorjanc, Biotechnical Faculty, University of Ljubljana, Domzale, Slovenia. Received 22 Mar. 2013. *Corresponding author (john.hickey@roslin.ed.ac.uk).

**Abbreviations:** BP-G, biparental with one grandparent in common with BP-X; BP-M, phenotypes from mixtures of related and unrelated biparental populations; BP-P, biparental with one parent in common with BP-X; BP-U, biparental population with no pedigree relationship to BP-X; BP-X, focal biparental population; gEBV, genomic estimated breeding values; $N_e$, effective population size; QTL, quantitative trait loci; SNP, single nucleotide polymorphism.

Efficient genetic improvement of plant breeding populations balances genetic gain against its time and cost. Genomic selection offers new opportunities for increasing the efficiency of plant breeding programs (e.g., Bernardo and Yu, 2007; Heffner et al., 2009; Crossa et al., 2010; Lorenz et al., 2011). One of the greatest potentials is in the ability to accurately select individuals of higher

Published in Crop Sci. 54:1476–1488 (2014).
doi: 10.2135/cropsci2013.03.0195
Freely available online through the author-supported open-access option.

breeding value without the requirement of collecting phenotypes pertaining to these individuals. This can facilitate a shortening of the breeding cycle and enable rapid selection and intercrossing of early-generation breeding material. In addition, genomic information could be used to optimize and therefore reduce the overall volume of phenotypic data that is required for making accurate selection decisions.

However, the use of genomic selection in breeding programs is potentially costly. It requires the development of a training population in which the genomic prediction equation can be estimated. Without careful design, setting up such populations may be expensive. For example, a particular training population may only be useful for making breeding decisions about a small group of individuals that are closely related to the training population, and therefore the cost of its development would be spread across only a small number of selection decisions. In addition, the genotypic data itself is expensive, particularly if low-cost genotyping strategies involving low-density single nucleotide polymorphism (SNP) chips and genotype imputation (e.g., Iwata and Jannink, 2010; Huang et al., 2012; Hickey et al., 2012a,b) or low-cost genotyping approaches such as genotyping-by-sequencing (Elshire et al., 2011; Poland et al., 2012) are not used. Also, while accurate selection decisions can be made with a small number of markers and phenotypes within sets of very closely related individuals, such as within biparental populations (Bernardo and Yu, 2007; Lorenzana and Bernardo, 2009; Massman et al., 2013), larger numbers of markers and phenotypes may be required if the training population comprises individuals that are not close relatives of the selection candidates.

For polygenic traits, applications of genomic selection in animal breeding programs have shown that when training populations are small, the training individuals need to be close relatives of the selection candidates if accurate genomic estimated breeding values (gEBV) are to be obtained (e.g., Habier et al., 2007, 2010; Meuwissen, 2009; Clark et al., 2012. Genomic predictions depend on accurate estimates of realized genomic relationships. Distant relatives have small realized genomic relationships because they only share a small proportion of quantitative trait loci (QTL) at specific genomic locations. The genome-wide marker estimates of these small relationships at the QTL can be inaccurate (de los Campos, 2012), inducing unfavorable noise-to-signal ratio. Larger realized genomic relationships are less affected in this way and therefore are more useful for prediction. Use of sufficient numbers of close relatives will result in highly accurate estimates of gEBV. However, as the asymptote of accuracy may be reached with a relatively low number of these individuals, cost inefficiencies will occur beyond this asymptote.

In species where population improvement is relevant, one approach to the implementation of genomic selection may involve setting up several biparental populations, selecting the $F_2$ individuals with the highest genetic merit on the basis of their gEBV, intercrossing the selected $F_2$, and selecting a proportion of the resulting progeny on the basis of their gEBV. This cycle of selection and intercrossing could be repeated a number of times before extracting inbred lines in a conventional way or through double haploid technology. Collecting phenotypic information pertaining to a portion of individuals inside a biparental population (e.g., pertaining to $F_2$) may offer an advantage of reducing the overall phenotyping load for a given biparental population, because when relationships are closer, fewer phenotypes are required. However, collecting phenotypes inside a given biparental population invokes a time cost to enable phenotypic data collection, which delays selection decisions. For example, in maize (*Zea mays* L.) the $F_2$ needs to be selfed and the resulting $F_{2:3}$ crossed to a tester before phenotypic evaluation can commence, a process that takes at least three seasons. An appealing alternative is to use phenotypes that have been previously collected. However, if these phenotypes come from individuals that are too distantly related, selection decisions will have lower accuracy, and there may be an increase in the number of markers and phenotypes needed to obtain acceptable levels of accuracy.

The objectives of this research were to explore the relative value, interaction, and contribution of phenotypic data density (numbers of individual phenotypes), relatedness of training and selection populations, and genetic marker density to the accuracy of genomic estimated breeding values. To explore these dynamics, computer simulations with designs mimicking biparental populations in maize breeding programs and a highly polygenic trait that may reflect grain yield genetic architecture were performed.

## MATERIALS AND METHODS

### Genome Simulation

Fifteen replicates of several alternative experimental designs relating to the practical implementation of genomic selection in maize breeding programs that use biparental populations were tested using the plant breeding simulation program AlphaSimPlant. All the designs began with a simulation of sequence data for base gametes of the population for each of 10 chromosomes using the Markovian Coalescence Simulator (Chen et al., 2009). The chromosomes were each 200 cM in length comprising $2 \times 10^8$ base pairs. To roughly mimic the historical changes of effective population size ($N_e$) in maize, these were simulated using an $N_e$ of 100, with piecewise linear increases to an $N_e$ of 2000 at 1000 generations ago and to an $N_e$ of 4000 at 2000 generations ago. A per-site mutation rate of $2.5 \times 10^{-8}$ was used. Each resulting chromosome had on average more than 160,000 segregating sites, resulting in a genome with more than 1.6 million segregating sites. The simulated gametes were then dropped through a pedigree to create several biparental populations that were related to each other in different ways: with one parent in common, with one grandparent in common, or nominally unrelated. These structures varied depending on the

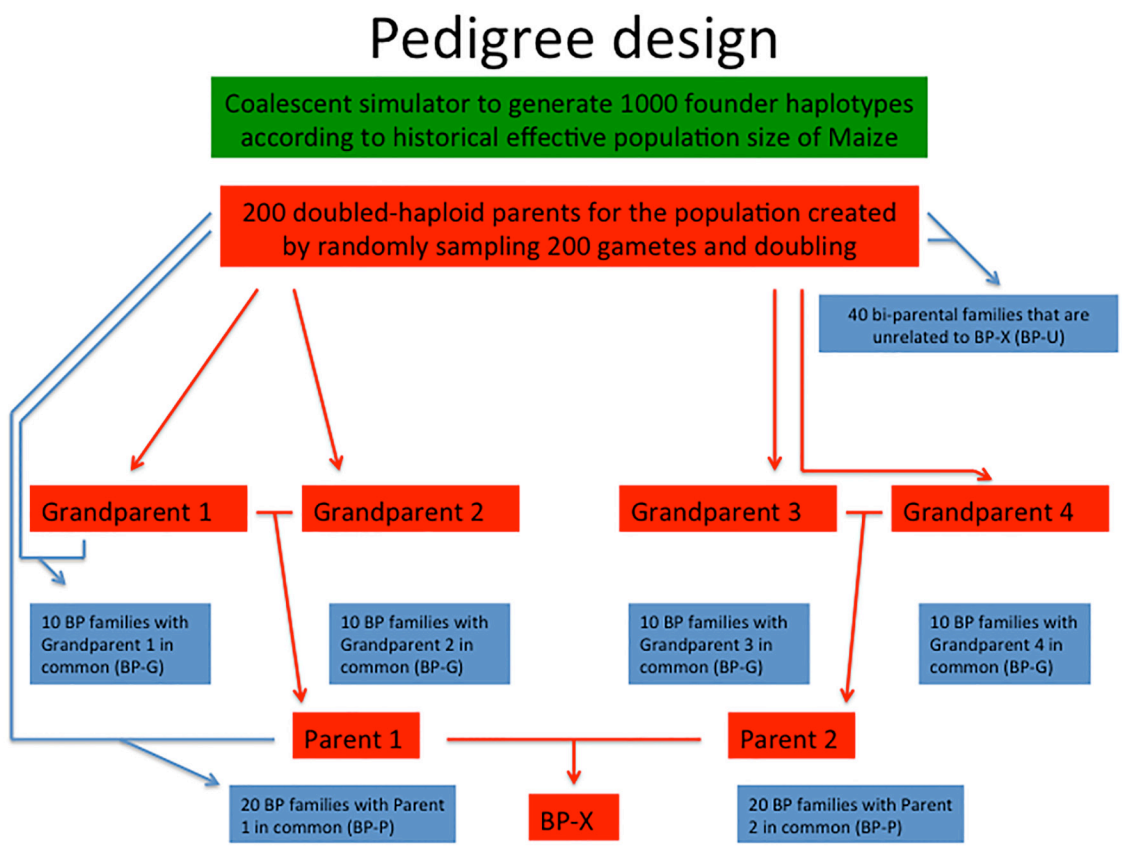


Figure 1. Pedigree relationships between focal biparental population (BP-X) and other biparental populations.

specific research question and are described in detail below. The genomes of tester individuals, which themselves were inbred lines, were simulated by branching a second ancestral population from the first ancestral population 40 generations before the end of the coalescent simulation. This branching ensured that the tester individual came from a different heterotic group.

## Overall Design and Pedigree Structure

The accuracy of selection was evaluated for unphenotyped $F_2$ individuals in a single focal biparental population (BP-X) using the correlation between the true breeding value (the genetic merit of an individual defined as twice the average deviation of its offspring from the population mean) and the gEBV. Phenotypes that were used to train the prediction equation were either generated inside BP-X or inside other biparental populations that were simulated to have certain relationships to BP-X (Fig. 1). Of the simulated biparental populations, 40 had one parent in common with BP-X (BP-P), 20 with Parent 1 in common, and 20 with Parent 2 in common. Another 40 biparental populations shared one grandparent with BP-X (BP-G), 10 biparental populations for each of the four grandparents in common. A further 40 biparental populations had no pedigree relationship to BP-X (BP-U). While individuals in BP-U were nominally unrelated to BP-X, they did have a base level of relationship similar to what would exist between random pairs of individuals from the same breeding population, as opposed to having the much lower levels of relationship that would exist between random pairs of individuals from different breeding populations.

These biparental populations were simulated by first simulating a breeding population of 200 inbred lines by sampling gametes at random with recombination from the gametes simulated by the coalescent process. Four of these inbred lines were randomly selected to be grandparents (G1, G2, G3, G4) of BP-X, two being used to form a biparental population that generated Parent 1 (P1) of BP-X, and two being used to form a biparental population that generated Parent 2 (P2) of BP-X. The two parents of BP-X were generated as a selected F10 individual from the two biparental populations involving the grandparents. The remaining biparental populations were generated by sampling both parents at random when forming the BP-U populations and by sampling one parent at random and the second parent according to the pedigree. For BP-P, the parent selected from the pedigree was selected at random from P1 and P2. For BP-G it was selected at random from G1, G2, G3, and G4. The full pedigree is visualized in Fig. 1.

## Structure of a Biparental Population

The full structure of each biparental population is described in Appendix A. Briefly, each biparental population was simulated to have 550 $F_2$ individuals. To generate phenotypes that pertain to each $F_2$ plant, seeds harvested from each $F_2$ plant were sown in an ear-to-row fashion and test crossed to a common tester individual. The test cross (TC) progenies were evaluated in one-row plots of three replications on the basis of TC progeny performance. Forty of the $F_2$ individuals were selected. The selected $F_2$ were advanced to $F_3$, where more phenotypes were collected analogous to how they were collected for $F_2$. Ten of the $F_3$ were selected and advanced to $F_5$, where phenotypes were again collected analogous to how they were collected for $F_2$. Finally, five selected $F_5$ individuals were advanced to

$F_{10}$. At $F_{10}$ one individual was selected to be the product of the given biparental population. Except for BP-X, all selection was performed at random, because these biparental populations were only used for the purposes of generation of phenotypic and genotypic data to train the prediction equations. In BP-X, selection was performed on the basis of gEBV as opposed to random selection. The accuracy of the gEBV for $F_2$ in BP-X was evaluated for the unphenotyped individuals using the correlation between estimated and true breeding values. The accuracy was not evaluated for phenotyped individuals. The events after the collection of the phenotypes pertaining to the $F_2$ individuals are described here for completeness, but they did not impact the outcomes of this study aside from facilitating the generation of $F_{10}$ individuals that were used as parents of other biparental populations.

## Trait Simulation

A polygenic trait with additive gene action was simulated. To generate QTL effects and true breeding values, 10,000 QTL loci were randomly sampled from the segregating sequence sites. Allele substitution effects for each of these QTL loci were sampled from a normal distribution with a mean of 0 and standard deviation of 1.0 divided by the square root of the number of QTL. True breeding values for each individual were equal to the sum across all QTL loci of the QTL genotype multiplied by the allele substitution effect. A line mean heritability of 0.5 was generated by scaling the residual variance relative to the variance of the breeding values of 200 inbred lines that founded the breeding population, which was given by $\mathbf{a'a}/(n-1)$, where $\mathbf{a}$ is a vector of breeding value of individuals of these inbred lines and $n = 200$.

## Scenarios Designed to Answer Research Objectives

Several scenarios were constructed to study the effects of different degrees of genetic relationship between the training population and selection candidates on the accuracy of gEBV of the selection candidates. Common to each scenario were nine marker densities (50, 100, 200, 300, 400, 500, 1000, 10,000, 100,000) and nine different numbers of $F_2$ phenotypic records collected in each biparental population used for training (5, 10, 20, 50, 100, 150, 200, 300, 500). Explicitly, each phenotypic record is the mean of three replicated rows each with 10 plants. For each marker density a constant set of markers was used across all biparental populations such that in any given family not all markers segregated. The accuracy of the gEBV for the remaining unphenotyped $F_2$ in BP-X was recorded, resulting in the need to simulate 550 $F_2$ per biparental population.

Scenario BP-X (sBP-X) used phenotypes collected on $F_2$ from BP-X only to train the prediction equation. In practice sBP-X would have a time cost (at least three breeding seasons needed to generate the phenotypes) compared with other scenarios, but it was included for the purposes of comparison. The remaining scenarios did not use phenotypes collected in BP-X to train the prediction equation and therefore would not incur this time penalty, as the phenotypes could have been collected in earlier breeding seasons.

To train the prediction equation, the second set of scenarios (sBP-P) used phenotypes that pertained to $F_2$ individuals in biparental populations with one parent in common (BP-P): sBP-P1 used phenotypes collected from 1 BP-P population, sBP-P4 used phenotypes collected from 4 BP-P populations, sBP-P8 used phenotypes collected from 8 BP-P populations, and sBP-P40 used phenotypes collected from 40 BP-P populations. Scenarios sBP-P4, sBP-P8, and sBP-P40 were each balanced with respect to parent P1 and parent P2 of BP-X.

To train the prediction equation, the third set of scenarios (sBP-G) used phenotypes that pertained to $F_2$ individuals in biparental populations with one grandparent in common (BP-G). Scenario sBP-G1 used phenotypes collected from 1 BP-G population, sBP-G4 used phenotypes collected from 4 BP-G populations, sBP-G8 used phenotypes collected from 8 BP-G populations, and sBP-G40 used phenotypes collected from 40 BP-G populations. Scenarios sBP-G4, sBP-G8, and sBP-G40 were each balanced with respect to the four grandparents (G1, G2, G3, G4) of BP-X.

To train the prediction equation, the fourth set of scenarios (sBP-U) used phenotypes that pertained to $F_2$ individuals in biparental populations that were nominally unrelated to BP-X (BP-U). Scenario sBP-U1 used phenotypes collected from 1 BP-U population, sBP-U4 used phenotypes collected from 4 BP-U populations, sBP-U8 used phenotypes collected from 8 BP-U populations, and sBP-U40 used phenotypes collected from 40 BP-U populations.

To train the prediction equation, the fifth set of scenarios (sBP-M) used phenotypes that pertained to $F_2$ individuals in mixtures of BP-P, BP-G, and BP-U populations. Scenario sBP-M-P4G8U40 used phenotypes collected from 4 BP-P, 8 BP-G, and 40 BP-U populations; sBP-M-P4G8U0 used phenotypes collected from 4 BP-P, 8 BP-G, and 0 BP-U populations; sBP-M-P4G0U40 used phenotypes collected from 4 BP-P, 0 BP-G, and 40 BP-U populations; and sBP-M-P0G8U40 used phenotypes collected from 0 BP-P, 8 BP-G, and 40 BP-U populations. The number of phenotypes in the training set for each scenario is given in Table 1. All genomic prediction was performed using ridge regression (Hoerl and Kennard, 1976; Whittaker et al., 1999 Meuwissen et al., 2001) as implemented in the software AlphaBayes2:

$$y_i \sim N\left(\mu_i, \sigma_e^2\right),$$

$$\mu_i = \alpha + \sum_{j=1}^{p} m_j x_{i,j},$$

$$m_j \sim N\left(0, \sigma_m^2\right),$$

Where $y_i$ is the phenotype value of the $i$-th individual, $\alpha$ is the intercept, $m_j$ and $x_{i,j}$ are the allele substitution (additive) effect and genotype code (0, 1, or 2) of the $j$-th genomewide SNP marker, respectively, and $\sigma_e^2$ and $\sigma_m^2$ describe the variance of residuals and allele substitution effects. To minimize the sampling variation the simulated values of the variance components were used instead of estimating them from data. Estimates of allele substitution effects $\left(\widehat{m_j}\right)$ were used to compute individual additive genetic values, that is, $a_i = \sum_{j=1}^{p} \widehat{m_j} x_{i,j}$, of either genotyped and phenotyped individuals or only genotyped individuals (predictions).

Table 1. Total number of phenotypes used for training by the number of phenotypes per population (P) and scenario (s). BP-X, focal biparental population; BP-P, biparental with one parent in common with BP-X; BP-G, biparental with one grandparent in common with BP-X; BPU, biparental population with no pedigree relationship to BP-X; BP-M, phenotypes from mixtures of related and unrelated biparental populations.

| | Number of phenotypes per population | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Scenario | 5 | 10 | 20 | 50 | 100 | 150 | 200 | 300 | 500 |
| sBP-X | 5 | 10 | 20 | 50 | 100 | 150 | 200 | 300 | 500 |
| sBP-P1 | 5 | 10 | 20 | 50 | 100 | 150 | 200 | 300 | 500 |
| sBP-P4 | 20 | 40 | 80 | 200 | 400 | 600 | 800 | 1200 | 2000 |
| sBP-P8 | 40 | 80 | 160 | 400 | 800 | 1200 | 1600 | 2400 | 4000 |
| sBP-P40 | 200 | 400 | 800 | 2000 | 4000 | 6000 | 8000 | 12000 | 20000 |
| sBP-G1 | 5 | 10 | 20 | 50 | 100 | 150 | 200 | 300 | 500 |
| sBP-G4 | 20 | 40 | 80 | 200 | 400 | 600 | 800 | 1200 | 2000 |
| sBP-G8 | 40 | 80 | 160 | 400 | 800 | 1200 | 1600 | 2400 | 4000 |
| sBP-G40 | 200 | 400 | 800 | 2000 | 4000 | 6000 | 8000 | 12000 | 20000 |
| sBP-U1 | 5 | 10 | 20 | 50 | 100 | 150 | 200 | 300 | 500 |
| sBP-U4 | 20 | 40 | 80 | 200 | 400 | 600 | 800 | 1200 | 2000 |
| sBP-U8 | 40 | 80 | 160 | 400 | 800 | 1200 | 1600 | 2400 | 4000 |
| sBP-U40 | 200 | 400 | 800 | 2000 | 4000 | 6000 | 8000 | 12000 | 20000 |
| sBP-M-P4G8U40 | 260 | 520 | 1040 | 2600 | 5200 | 7800 | 10400 | 15600 | 26000 |
| sBP-M-P4G8U0 | 60 | 120 | 240 | 600 | 1200 | 1800 | 2400 | 3600 | 6000 |
| sBP-M-P4G0U40 | 220 | 440 | 880 | 2200 | 4400 | 6600 | 8800 | 13200 | 22000 |
| sBP-M-P0G8U40 | 240 | 480 | 960 | 2400 | 4800 | 7200 | 9600 | 14400 | 24000 |

Table 2. The across-replicate mean (and standard deviation as subscript) number of polymorphic markers in focal biparental population (BP-X) for each of the marker densities.

| Marker platform | 50 | 100 | 200 | 300 | 400 | 500 | 1000 | 10000 | 100000 |
|---|---|---|---|---|---|---|---|---|---|
| No. polymorphic in BP-X | $12_{2}$ | $26_{3}$ | $54_{7}$ | $79_{10}$ | $105_{11}$ | $131_{11}$ | $265_{16}$ | $2688_{88}$ | $26807_{764}$ |

# RESULTS

## Summary Statistics for Simulated Marker Data

In a given biparental population (e.g., BP-X), approximately 26.5% of markers segregated, and the proportion of segregating markers did not change with changing marker density (Table 2). This fact should be kept in mind when interpreting the different marker densities throughout the results and discussion.

## Accuracy When Training with Focal BP-X Phenotypes

The results for sBP-X are shown in Fig. 2. In this scenario the accuracy of the breeding values in $F_2$ increased with the increasing size of the training set. It also increased with increasing marker density up to an asymptote. The asymptote of marker density was different for different training set sizes, but in all cases it appeared to be less than 500 markers, of which an average of 131 or less segregated in BP-X. For the smaller training set sizes (50 or fewer phenotypes) there was a relatively large amount of sampling variation evident across the different phenotype and genotype designs. Training with 5, 20, or 50 phenotypes gave accuracies of approximately 0.2, 0.4, or 0.6, respectively. The differences in accuracy between training with 100 or 500 phenotypes were small. Training with 100 phenotypes gave accuracies of approximately 0.8, while training with 500 phenotypes gave very high accuracy (>0.90). Increasing from 10,000 markers to 100,000 markers did not increase accuracy.

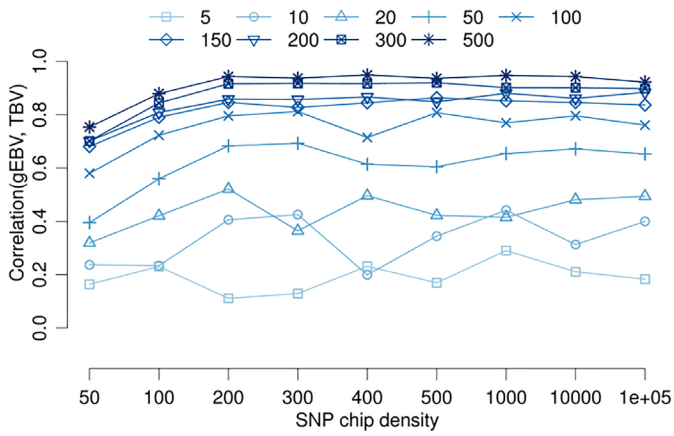


Figure 2. Accuracy of genomic estimated breeding values (gEBV) inside focal biparental population (BP-X) when training in BP-X with different numbers of markers and phenotypes. SNP, single nucleotide polymorphism. TBV, true breeding value.

## Accuracy When Training with BP-P Phenotypes

Key results for sBP-P are shown in Fig. 3, and the complete results are shown in Fig. S1 in the supplementary materials. In this set of scenarios, the accuracy of gEBV increased and the sampling variance decreased as the number of BP-P populations contributing to the training population increased. The accuracy also increased with increasing number of phenotypes recorded inside each BP-P population and with increasing marker density. In comparison with using phenotypes from BP-X, using phenotypes from one BP-P population to construct the training population resulted in very low accuracies. In this case the accuracies were generally lower than 0.4 and displayed a large degree of sampling even with a large number of phenotypes (Fig. 3 and Fig. S1a). Using phenotypes from four (Fig. 3 and Fig. S1b) or eight (Fig. 3 and Fig. S1c) BP-P populations resulted in reasonably high accuracies. For example, using 50 phenotypes from each of four or eight BP-P populations resulted in accuracies of about 0.5 and 0.6, respectively (Fig. 3b), once the asymptote of marker density had been reached. While the asymptote of marker density was generally reached at 10,000 markers, a large proportion of the prediction accuracy could be obtained with as few as 300 markers. With 40 BP-P populations accuracy of prediction in BP-X was generally slightly higher and less variable than with 4 or 8 BP-P populations.

There were no major differences in accuracy when recording many phenotypes in a small number of populations or recording a small number of phenotypes in many populations. Using 100 phenotypes in sBP-P4, 50 phenotypes in sBP-P8, or 10 phenotypes in sBP-P40 resulted in training sets of the same size (400 phenotypes) with the same expected degree of relationship to BP-X. In each case the accuracy ranged between 0.60 and 0.66, once the asymptote of marker density had been reached (Fig. S1).

## Accuracy When Training with BP-G Phenotypes

Key results for sBP-G are shown in Fig. 4 and the complete results are shown in Fig. S2 in the supplementary materials. In this set of scenarios, which used phenotypes from BP-G populations, the accuracies of gEBV were much lower than the accuracies for the corresponding cases of sBP-P, which had parents in common rather than grandparents in common, as was the case for BP-G. For sBP-G1 the accuracies were generally lower than 0.2 (Fig. 4 and Fig. S2a), while for sBP-P1 they were as high as 0.4 (Fig. 3 and Fig. S1a). In sBP-G4 the accuracies were generally lower than 0.4, whereas for sBP-P4 (Fig. 4 and Fig. S2b) they were as high as about 0.8 (Fig. 3 and Fig. S1b). Because of sampling the trend in progression to the asymptote for marker density could only be clearly observed for sBP-G40 (Fig. S2d), which had phenotypes from 40 BP-G

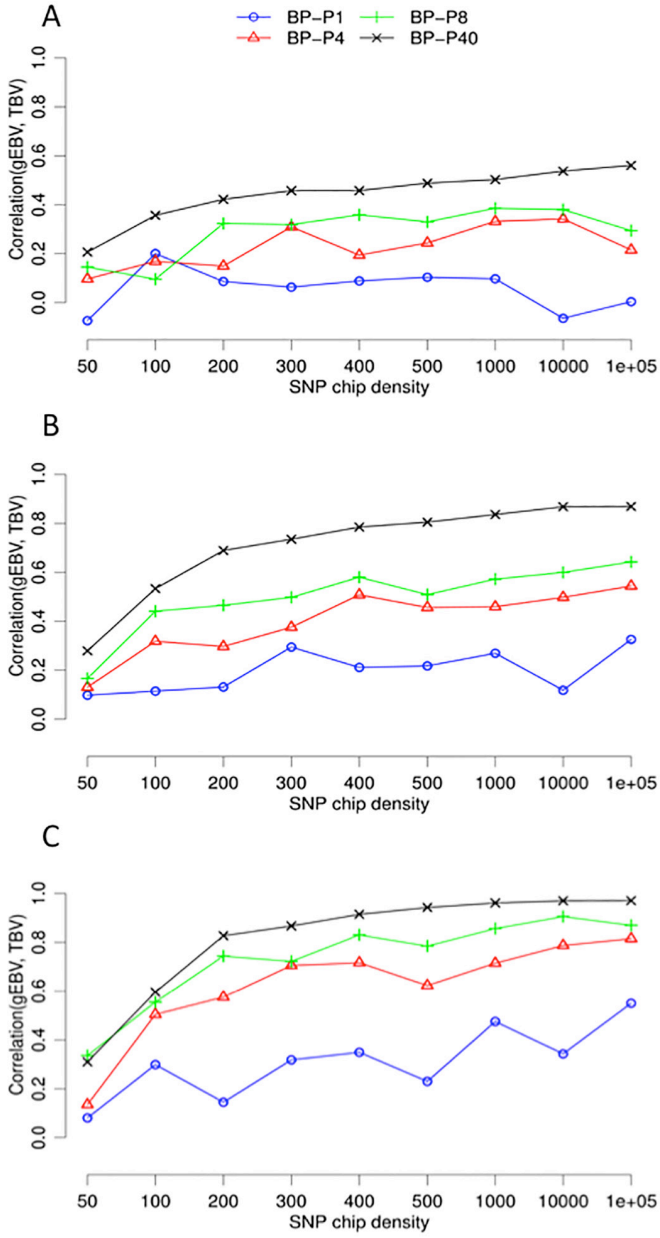


Figure 3. Accuracy of genomic estimated breeding values (gEBV) inside focal biparental population (BP-X) when training in biparental with one parent in common with BP-X (BP-P) with different numbers of markers and phenotypes from 1 BP-P population (BP-P1), 4 BP-P populations (BP-P4), 8 BP-P populations (BP-P8), or 40 BP-P populations (BP-P40). A = 5 phenotypes recorded in each biparental population; B = 50 phenotypes recorded in each biparental population; C = 500 phenotypes recorded in each biparental population. SNP, single nucleotide polymorphism. SNP, single nucleotide polymorphism. TBV, true breeding value.

populations. Approximately linear increases in accuracy were observed up to 10,000 markers. Increasing marker density to 100,000 showed little benefit.

## Accuracy When Training with BP-U Phenotypes

Key results for sBP-U are shown in Fig. 5, and the complete results are shown in Fig. S3 in the supplementary

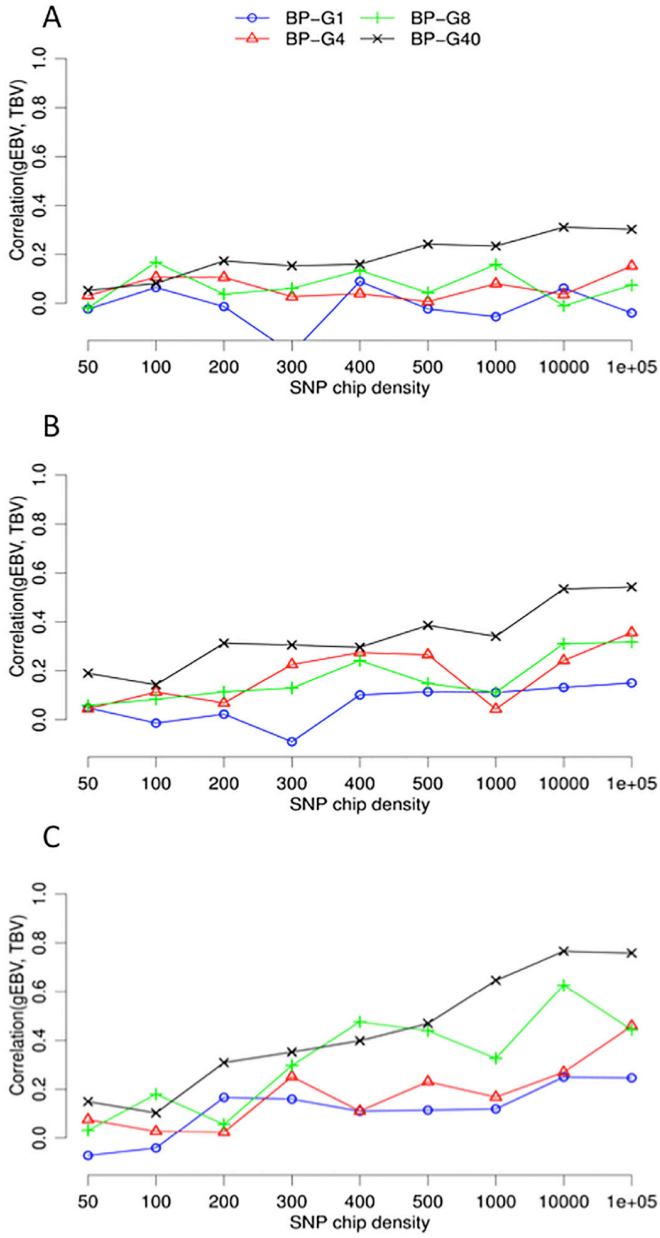


Figure 4. Accuracy of genomic estimated breeding values (gEBV) inside focal biparental population (BP-X) when training in biparental with one grandparent in common with BP-X (BP-G) with different numbers of markers and phenotypes from 1 BP-G population (BP-G1), 4 BP-G populations (BP-G4), 8 BP-G populations (BP-G8), or 40 BP-G populations (BP-G40). A = 5 phenotypes recorded in each biparental population; B = 50 phenotypes recorded in each biparental population; C = 500 phenotypes recorded in each biparental population. SNP, single nucleotide polymorphism. TBV, true breeding value.

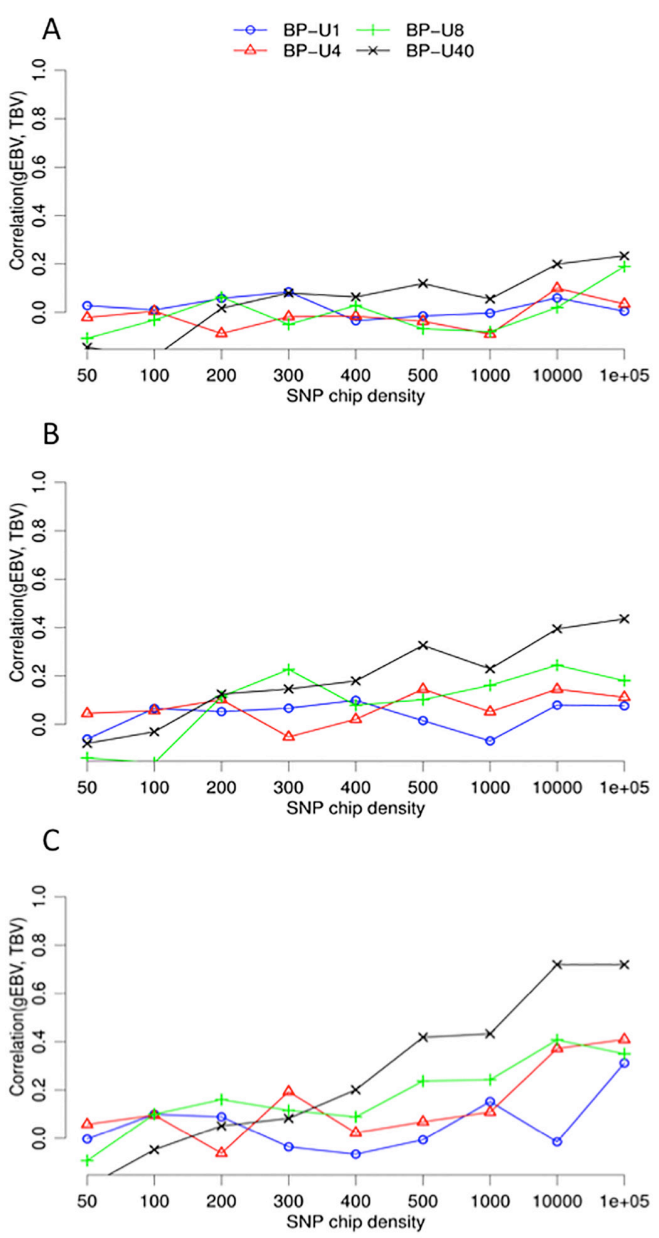


Figure 5. Accuracy of genomic estimated breeding values (gEBV) inside focal biparental population (BP-X) when training in biparental population with no pedigree relationship to BP-X (BP-U) with different numbers of markers and phenotypes from 1 BP-U population (BP-U1), 4 BP-U populations (BP-U4), 8 BP-U populations (BP-U8), or 40 BP-U populations (BP-U40). A = 5 phenotypes recorded in each biparental population; B = 50 phenotypes recorded in each biparental population; C = 500 phenotypes recorded in each biparental population. SNP, single nucleotide polymorphism. TBV, true breeding value.

materials. In this set of scenarios, which used phenotypes from BP-U populations, the accuracies of gEBV were lower than the accuracies for the corresponding cases of sBP-P (Fig. 3) and sBP-G (Fig. 4), which used phenotypes from related biparental populations. The accuracies for sBP-U1 (Fig. 5 and Fig. S3a) and sBP-U4 (Fig. 5 and Fig. S3b) were generally less than 0.2 and were subject to a large degree of sampling variation. High accuracies could be obtained using many phenotypes from unrelated populations. In sBP-U40 (Fig. 5c and Fig. S3d) the accuracies reached up to 0.7 with 20,000 phenotypes, that is, with 40 populations each contributing 500 phenotypes. Increasing marker density up to 10,000 markers resulted in increased accuracy, but there was little benefit when increasing marker density to 100,000. Comparing sBP-U40 (Fig. 5c [black line] and Fig. S3d) with sBP-G40 (Fig. 4c [black

line] and Fig. S2d) showed that, while for sBP-G40 the accuracy increased approximately linearly with increasing marker density up to 10,000 markers such that reasonable accuracies (e.g., 0.6) could be obtained with as few as 1000 SNP; the use of fewer than 10,000 markers in sBP-U40 resulted in much lower accuracies.

### Accuracy When Training with Phenotypes from Mixtures of Related and Unrelated Biparental Populations

Key results for sBP-M are shown in Fig. 6, and the complete results are shown in Fig. S4 in the supplementary materials. In this set of scenarios, which used phenotypes from mixtures of related and unrelated biparental populations, high accuracies could be achieved (>0.80), and major benefits were obtained by increasing the marker density up to 10,000 markers. When marker density was low and information on close relatives was available, the addition of phenotypes from less related biparental populations resulted in reduced accuracy. This can be clearly seen by comparing scenario phenotypes from sBP-P4 (Fig. 3 [red line] and Fig. S1b) with sBP-M-P4G0U40 (Fig. 6 [red line] and Fig. S4b). Both of these scenarios had in their training sets phenotypes from 4 biparental populations with one parent in common, but the second scenario also included phenotypes from 40 unrelated biparental populations. However, when the marker density was high (e.g., 10,000 markers), the accuracy increased with the addition of additional phenotypes that were less related. This trend could also be observed by comparing sBP-M-P4G8U0 (Fig. 6 [blue line] and Fig. S4a) with sBP-M-P4G8U40 (Fig. 6 [black line] and Fig. S4d). The accuracy from sBP-M-P4G8U0 increased over that of sBP-M-P4G8U0 only when the marker density was high. Scenario sBP-M-P4G8U40 gave high accuracies when the marker density was high. When 50 phenotypes were recorded within each biparental population, the accuracy reached approximately 0.7. This was a similar level of accuracy to what could be obtained by recording 100 phenotypes within BP-X itself, but sBP-M-P4G8U40 offers the advantage of a much-reduced time and cost per selection decision (because the costs can be spread over many more selection decisions) needed to record the phenotypes.

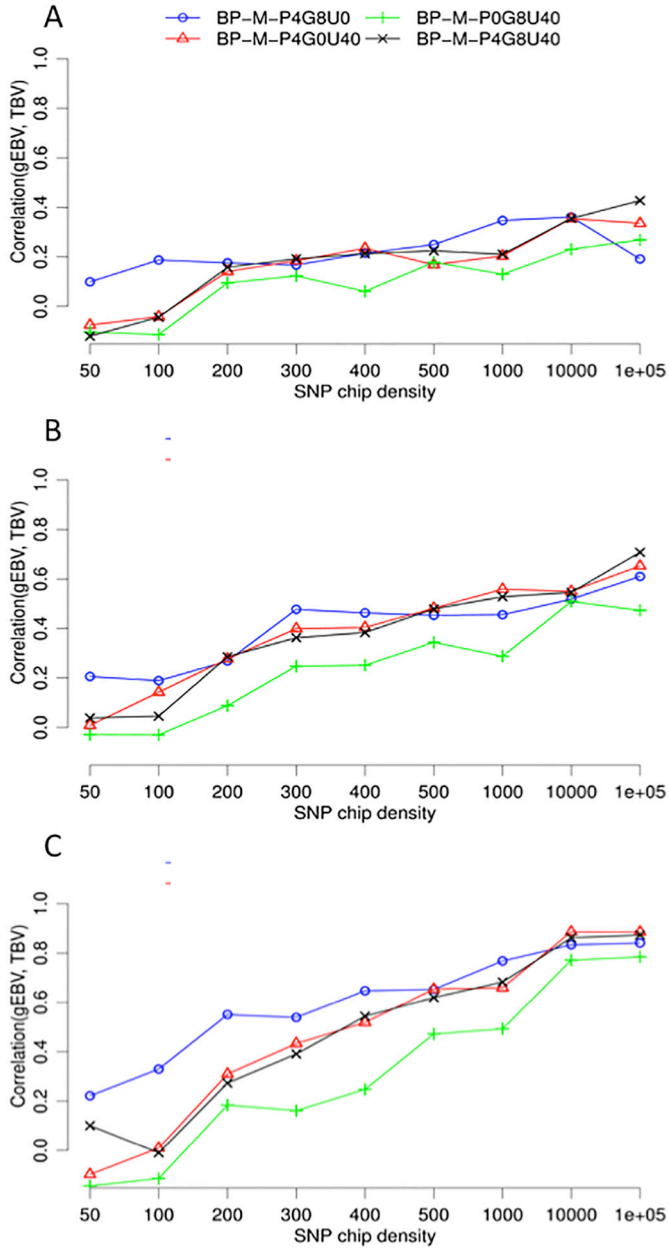


Figure 6. Accuracy of genomic estimated breeding values (gEBV) inside focal biparental population (BP-X) when training in phenotypes from mixtures of related and unrelated biparental populations (BP-M) with different numbers of markers and phenotypes from 4P + 8G + 0U (BP-M-P4G8U0), 4P + 0G + 40U (BP-M-P4G0U40), 0P + 8G + 40U (BP-M-P0G8U40), and 4P + 8G + 40U (BP-M-P4G8U40). A = 5 phenotypes recorded in each biparental population; B = 50 phenotypes recorded in each biparental population; C = 500 phenotypes recorded in each biparental population. SNP, single nucleotide polymorphism. TBV, true breeding value.

## DISCUSSION

In plant breeding programs that use biparental populations, genomic selection can facilitate accurate selection (e.g., correlation of 0.6 between true breeding value and gEBV) among early-generation breeding material, such as $F_2$ individuals. This research showed that at least four different approaches can be used to generate reasonably accurate gEBV: (i) collecting phenotypic information in the biparental population of interest; (ii) collecting phenotypes in closely related biparental populations; (iii) collecting phenotypic information in unrelated biparental populations; or (iv) using combinations of phenotypic information collected in a mix of related and unrelated biparental populations. Collecting a small number of phenotypes in the biparental population of interest gave the highest accuracy and required only a small number of markers (e.g., 200–500 SNP). However, practical

application of this approach would incur a time and cost per selection decision penalty because at least three seasons would be required to generate these phenotypes. This time penalty would also occur with similar designs, such as when a recombinant inbred line population is generated first to be used as a training set (e.g., Bernardo and Yu, 2007). Using phenotypes from other biparental populations would not incur this time and cost penalty (because the costs of training set construction can be spread over many more selection decisions), since in a typical breeding program this phenotypic data could have been previously generated. Nevertheless, it is encouraging to see robust prediction accuracies within a given biparental population, which is very relevant and applicable in the context of doubled haploid-based (DH) maize breeding, where large numbers of homozygous lines are routinely obtained from elite × elite crosses. In such instances, collecting phenotypes of only a portion of DH lines obtained from a particular cross would be sufficient to facilitate prediction of the rest of the lines within the same cross. Besides, such an approach is also helpful if $F_2$ individuals are to be selected or culled on the basis of prediction for per se performance such as disease resistance or nutritional quality.

Using information from related biparental populations can give high accuracies and does not require large numbers of markers (e.g., 300–500 markers). However, for this approach to be competitive in terms of accuracy, a greater number of phenotypes need to be recorded compared with when recording phenotypes directly in the biparental population of interest. For example, 400 to 1000 phenotypes were needed to achieve an accuracy of 0.6 when using phenotypes from biparental populations with parents in common, whereas only 50 phenotypes may be needed to achieve an accuracy of 0.6 when using phenotypes from BP-X. In addition, the genomes of both parents need to be represented in the individuals that are phenotyped. Using information from unrelated biparental populations generally gave low accuracies unless very large numbers of phenotypes were available (e.g., 4000 to 20,000) and a large number of markers were used (e.g., 10,000). Using all available phenotypic information (i.e., data from combinations of related and unrelated biparental populations) was better than ignoring the unrelated phenotypic information if the marker density was high (e.g., 10,000). However, if the marker density was too low (e.g., 200 markers) it was best not to include the distantly related phenotypes in the training population.

Close relatives share long haplotypes because they are separated by small numbers of meiosis and therefore there is little opportunity for recombination to break apart shared haplotypes. This leads to two advantages for genomic prediction that can explain the trend of higher accuracies when information from related rather than unrelated individuals was used to train the prediction equation. First, longer haplotypes mean that the total number of independent haplotype effects that need to be estimated is smaller, resulting in more phenotypes per haplotype, which leads to more precise estimates of each haplotype effect. Second, in contrast to short haplotypes, when longer haplotypes are shared between individuals, the marker-based estimate of the realized genomic relationship is closer to the true relationship at the QTL because many more markers track QTL. As relationships become smaller the noise or bias in the estimate of the realized genomic relationship increases (de los Campos, 2012), which induces an increasingly unfavorable noise-to-signal ratio. This reduces the power of small relationships for accurately estimating gEBV.

In addition to the ability to use already existing phenotypes and thus not incurring a time or cost penalty, there are at least two further potential benefits from using phenotypes from different biparental populations. First, it could mean that a greater number of environments and years could be represented in the training population, inducing a greater level of robustness to genotype-by-environment interaction in the estimation of gEBV. Second, one of the major potential benefits of genomic selection to a population improvement program based on recurrent selection is the ability to do a number of rounds of selection and intercrossing of early generation material. If the training population consisted of $F_2$ individuals from BP-X itself, each round of intercrossing would result in decreasing relationships between the training and subsequent prediction population resulting in reduced accuracies of gEBV. If the training population were comprised of partly unrelated material, the reduction in accuracy with increasing rounds of intercrossing would likely be much less. This is because marker associations that would underlie the prediction equation in this circumstance would parameterize short haplotype effects and linkage disequilibrium information rather than long haplotype effects and large linkage blocks. Long haplotype effects and large linkage blocks are broken down more rapidly by meiosis with increasing generations of intercrossing than short haplotype effects.

When using information from close relatives, marker associations are due to linkage between markers and QTL, whereas when using information from distant relatives the marker associations are due to linkage disequilibrium between markers and QTL. When individuals that are distantly related to each other dominate the training data set, a large proportion of the resulting estimated marker associations are likely due to linkage disequilibrium rather than linkage. This may explain the lower levels of accuracy that were observed when distant relatives were added to a training population already containing close relatives, when marker density was low. Models that explicitly partition information from linkage and linkage disequilibrium

(e.g., Meuwissen and Goddard, 2004; Legarra and Fernando, 2009; Schulz-Streeck et al., 2012) could allow both types of information to be used more optimally.

The marker density required to obtain accurate gEBV depended on the degree of relationships between the training and prediction populations. With close relatives accurate predictions could be obtained with as few as 200 markers, and increasing the marker density up to 10,000 did not improve the accuracy of prediction. Small numbers of markers are sufficient because the shared haplotypes and linkage blocks are long and their effects can be captured with a small number of markers. When using distant relatives or mixtures of close and distant relatives, 10,000 markers were required because the prediction equation depends more on shared short haplotypes or linkage disequilibrium between markers and QTL. Increasing the marker density from 10,000 to 100,000 showed little increase in accuracy and often caused slightly reduced accuracy. This trend was particularly obvious when the numbers of phenotypes in the training population was small. When the number of markers to be estimated greatly exceeds the number of training phenotypes, the genomic prediction equation overfits the data, resulting in nongenetic effects (e.g., environmental variance) being attributed to marker effects. While this results in an increase in the accuracy of prediction within the training data set, it reduces the predictive power in other data sets where the nongenetic effects are different. In this research, ridge regression (Hoerl and Kennard, 1976; Whittaker et al., 1999 Meuwissen et al., 2001), which is one of the simplest statistical methods for genomic prediction, was used to estimate the prediction equation. More complex statistical methods (e.g., BayesB; Meuwissen et al., 2001) or other types of mixture models (de los Campos et al., 2013) could be used, which make use of statistical or biological prior information, and these may reduce the effect of overfitting, but the benefit would be marginal. Several studies have shown that, in comparison with ridge regression types of methods, more complex statistical methods give little increase in the accuracy of genomic prediction for polygenic traits (Bernardo and Yu, 2007; Clark et al., 2012; Cleveland et al., 2012; Daetwyler et al., 2013).

Variance in response to selection is a risk to a breeding program (Nicholas, 1980). In breeding programs that use designs similar to those in this research, variance in response to selection may be caused by variance in the accuracy of the gEBV. When the training populations were small there was a large degree of sampling variation across replicates; increasing the training population size reduced this sampling variation. This level of sampling variation observed means that individual field trials of this nature will also be subject to sampling. Sampling is known to affect both marker assisted recurrent selection (Johnson, 2004) and genomic selection in biparental populations (Massman et al., 2013).

In plant breeding programs genomic predictions can comprise at least three components: the effects of population structure, family effects (the parent average term [PA]), and the effect of Mendelian segregation within a family (the Mendelian sampling term [MS]). Generally, breeders' knowledge or pedigree information can be used to make accurate predictions of the effect of population structure or PA (Windhausen et al., 2012). However, without lots of phenotypes pertaining directly to the individual plant (e.g., the individual $F_2$), breeders' knowledge or pedigree information can make no prediction about the MS term, and collecting these phenotypes takes time and resources. Response to selection per unit time depends on the accuracy and time taken to evaluate the MS term of each selection candidate (Woolliams et al., 1999). In this research the accuracy of the genomic predictions were calculated within the focal biparental population (BP-X). It therefore reflects the accuracy of the estimate of the MS term within this family. This estimate of accuracy was not inflated by the effect of population structure or PA and reflects the potential for making genetic improvement. When the prediction accuracy reflects population structure or PA it reflects the potential for population or family replacement rather than population improvement. Under these circumstances, if the prediction accuracy is high, useful genetic resources will be quickly discarded and little sustainable genetic improvement will occur.

The phenotyping and genotyping alternatives tested in this study suggest that accurate selection among early-generation breeding material, such as $F_2$ individuals, can be made using genomic selection. On the basis of the results, one approach to the practical application of genomic selection to plant breeding programs could involve the collection of 50 $F_2$ phenotypes in several biparental populations each breeding season. Until this process leads to generation of a relatively large training population (e.g., 5000 to 20,000 individuals), accurate genomic predictions will only be possible for closely related individuals. However, once sufficient volumes of data are available, accurate genomic predictions could be available for all individuals in a breeding program. When the completely unrelated material, such as individuals from exotic landraces for the purposes of introgression (e.g., Bernardo, 2009), is added to the breeding pool phenotypic data pertaining to this material needs to be collected to update the prediction equation. While this approach of phenotyping $F_2$ individuals is powerful, it does invoke an extra phenotyping step that is not always performed in breeding programs for hard-to-measure traits such as grain yield. Preliminary yield trials are routinely performed in breeding programs and have the advantage of more replication at more locations than could be done at the $F_2$ stage. However, preliminary yield trials would be performed on a selected subset of the population and therefore would represent less of the total genetic variation in a population than a random selection of $F_2$ individuals would represent. In addition,

individuals with preliminary yield trials, (e.g., for $F_6$ individuals) would be less related to $F_2$ individuals in other populations than $F_2$ individuals, because $F_6$ individuals go through more meiosis than $F_2$ individuals. The interaction between more replication, and consequently higher heritability, of preliminary yield trials, the representation of less genetic diversity, and the lower levels of relatedness in comparison with the $F_2$ phenotypes that were used in this study needs to be explored.

One of the major costs of implementing this design for genomic selection in plant breeding programs will be the genotyping costs. However, these can be reduced using genotype imputation (e.g., Iwata and Jannink, 2010; Hickey et al., 2012a,b; Huang et al., 2012). One approach to genotype imputation could involve genotyping the parents of each biparental population with 10,000 or more markers and genotyping the individuals in the training and prediction sets with fewer markers and then using these markers to impute the parental genotypes. This approach generally has little negative impact on the resulting gEBV accuracy (e.g., Cleveland and Hickey, 2013) and could reduce the costs per individual to $11, if only 200 markers were needed, and $21, if 400 markers were needed. This assumes DNA extraction costs of $3 and each set of 50 markers cost $2. Some plant breeding applications choose to use marker panels that are known to segregate in a given biparental population. In this research the same simulated marker platform was used in different populations, meaning that in a given population only 30% of the markers segregated. Therefore, if the option to construct population-specific marker platforms were available, such that only markers known to segregate in the population were included on the marker platform, even fewer markers than suggested by this research would be required, but sharing of data from different biparental populations would require imputation to a common set of markers.

While the example in this research focuses on the implementation of genomic selection in biparental populations in maize, the messages pertain to other plant species and breeding schemes. For example, in wheat (*Triticum aestivum* L.), use of genomic selection for early generation breeding decisions could be made by collecting phenotypes that pertain to $F_2$ individuals using $F_{2:4}$ families and/or individuals. Other types of phenotyping could also be used, such as phenotypes pertaining to later generation material (e.g., $F_6$), or inbred lines. However, sufficient numbers of these phenotypes would be required, and they would have to be sufficiently related to the material for which the gEBV were being estimated.

There is an increasing interest in the use of genomic selection as an enabling technology for strengthening the crop breeding programs of the National Agricultural Research Systems as well as those of the small and medium enterprise private sector companies in the developing countries. International agricultural research centers that provide elite breeding materials as public goods (e.g., CIMMYT) could potentially facilitate the establishment and coordination of such genomic selection networks in specific crop species either globally or regionally (Babu et al., 2012). Such a network could effectively link the elite breeding materials and genotypic data (provided by the coordinating international center) with the phenotypic data generated by the network members on relevant traits in the target agroecologies. Within the context of the results of this study, these networks may facilitate the collection of the required 5000 to 20,000 phenotypes for making genomic selection work when using unrelated breeding material, something that is often beyond the capacity of most individual small and medium crop breeding programs in the short term. Also, to make such genomic selection networks highly effective, common genotyping platforms that behave consistently across time are needed so that genomic estimated breeding values can be shared across time and among the members of the network.

While this simulation study provides excellent guidelines for optimizing various parameters of a practical genomic selection–based cereal breeding program, certain deviations resulting in altered prediction accuracies of empirical phenotypes may occur as a result of phenomena such as epistasis, epigenetic modifications, and various metabolic and system-level interactions. However, simulations using polygenic models with additive genetic effects similar to those used in this study have given results that are similar to those from real data (e.g., Hickey and Gorjanc, 2012; Clark et al., 2012; Daetwyler et al., 2013).

## CONCLUSIONS

Alternative phenotyping and genotyping strategies to facilitate genomic selection for early generation materials (e.g., $F_2$) in plant breeding programs were evaluated through simulation. By using phenotypes that would have been previously collected, accurate (e.g., 0.6 correlation between true breeding value and gEBV) selection could be made without having to wait for phenotypes that pertain directly to these individuals to be collected. The relationships between the individuals in the training population and the individuals in the prediction population had a major impact on the accuracy of the gEBV. Accurate predictions could be obtained with a small number of markers (e.g., 300–500) and a small number of phenotypes (e.g., 200–1000) when the phenotypes were collected from closely related biparental populations. To generate accurate predictions from nominally unrelated individuals many more phenotypes (e.g., 20,000) and many more markers (e.g., 10,000) were required. Further work is required to evaluate the extent to which genotype imputation can be used to reduce the costs of genotyping and how phenotyping trials and phenotyping costs can be optimized within the context of the designs used in this study.

## APPENDIX A. SIMULATION OF A BIPARENTAL POPULATION

1. Cross inbred parent A with inbred parent B to produce $F_1$ seed
2. Sow $F_1$ seed and self to produce $F_2$ seed
3. Sow $F_2$ seed and self to produce $F_{2:3}$ seed
4. Sow some $F_{2:3}$ seed and cross to Tester
5. Sow $F_{2:3}$ × Tester seed to generate phenotypes
6. Select best $F_{2:3}$ families to advance on the basis of testcross performance phenotypes. This essentially advances the remaining $F_{2:3}$ seeds of the best $F_2$
7. Sow $F_{2:3}$ seeds and self to produce $F_4$ seed
8. Sow some $F_4$ seed and cross to Tester to produce $F_4$ × Tester seed
9. Plant $F_4$ × Tester seed and phenotype
10. Advance best $F_4$ families on the basis of phenotype
11. Self $F_4$ seed and self to produce $F_5$ seed
12. Sow $F_5$ seed and self to produce $F_6$ seed
13. Sow some of the $F_6$ seed and cross to Tester to produce $F_6$ × Tester seed
14. Sow $F_6$ × Tester seed and phenotype
15. Advance the best $F_6$ families
16. Sow best $F_6$ seed and self to produce $F_7$ seed
17. Sow and self $F_7$ seed to produce $F_8$ seed
18. Sow and self $F_8$ seed to produce $F_9$ seed
19. Sow and self $F_9$ seed to produce $F_{10}$ seed


## Acknowledgments

This work was funded by CIMMYT, Seeds of Discovery, the Maize CRP, the Wheat CRP, and Advanta Semillas. Kevin Pixley is acknowledged for his reading and making useful suggestions about an earlier draft of this paper.


## References


Babu, R., S.K. Nair, B.S. Vivek, F. San Vicente, and B.M. Prasanna. 2012. Integrating marker-assisted selection in the DH-based breeding pipeline for rapid development and delivery of superior parental lines and cultivars. In: B.M. Prasanna, V. Chaikam, and G. Mahuku, editors, Doubled haploid technology in maize breeding: Theory and practice. CIMMYT, D.F., Mexico. p. 39–44.

Bernardo, R. 2009. Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Sci. 49(2):419-425.

Bernardo, R., and J. Yu. 2007. Prospects for genome-wide selection for quantitative traits in maize. Crop Sci. 47:1082–1090. doi:10.2135/cropsci2006.11.0690

Chen G.K., P. Marjoram, J.D. Wall. 2009. Fast and flexible simulation of DNA sequence data. Genome Res. 19:136–142.

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4. doi:10.1186/1297-9686-44-4

Cleveland, M.A., J.M. Hickey, and S. Forni. 2012. A common dataset for genomic analysis of livestock populations. G3 2: 429–35.

Cleveland, M.A., and J.M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci.

Crossa, J., G. de los Campos, P. Perez, D. Gianola, J. Burgueno, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521

Daetwyler, H.D., M.P.L. Calus, R. Pong-Wong, G. de los Campos, and J.M. Hickey. 2013. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365. doi:10.1534/genetics.112.147983

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2013. Whole genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. doi:10.1534/genetics.112.143313

de los Campos, G. 2012. Whole-genome prediction of human complex traits using data from related and unrelated individuals. Proceedings of the International Conference on Quantitative Genetics, Edinburgh, June 17–22, 2012.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379 doi:10.1371/journal.pone.0019379

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.

Habier, D., J. Tetens, F.R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5. doi:10.1186/1297-9686-42-5

Heffner, E.L., M.E. Sorrells, and J.L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512

Hickey, J.M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654–663. doi:10.2135/cropsci2011.07.0358

Hickey, J.M., B.P. Kinghorn, B. Tier, J.H.J. van der Werf, and M.A. Cleveland. 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation method. Genet. Sel. Evol. 44:9. doi:10.1186/1297-9686-44-9

Hickey, J.M., and G. Gorjanc. 2012. Simulated data for genomic selection and GWAS using a combination of coalescent and gene drop methods. Genes, Genomes. Genetics 2:425–427.

Hoerl, A.E., and R.W. Kennard. 1976. Ridge regression iterative estimation of the biasing parameter. Communications in Statistics—Theory and Methods. 5(1):77-88.

Huang, Y., J.M. Hickey, M.A. Cleveland, and C. Maltecca. 2012.

Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. 44:25. doi:10.1186/1297-9686-44-25

Iwata, H., and J.-L. Jannink. 2010. Marker genotype imputation in a low-marker density panel with a high-marker-density reference panel: Accuracy evaluation in barley breeding lines. Crop Sci. 50:1269–1278. doi:10.2135/cropsci2009.08.0434

Johnson, R. 2004. Marker-assisted selection. Plant Breed. Rev. 24:293–309.

Legarra, A., and R.L. Fernando. 2009. Linear models for joint association and linkage QTL mapping. Genet. Sel. Evol. 41:43–60. doi:10.1186/1297-9686-41-43

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J.L. Jannink. 2011. Genomic selection in plant breeding: Knowledge and prospects. Adv. Agron. 110:77–123. doi:10.1016/B978-0-12-385531-2.00002-5

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151–161. doi:10.1007/s00122-009-1166-3

Massman, J.M., H.J.G. Jung, and R. Bernardo. 2013. Genome-wide 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. 2009. Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41:35. doi:10.1186/1297-9686-41-35

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.

Meuwissen, T.H.E., and M.E. Goddard. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261–279. doi:10.1186/1297-9686-36-3-261

Nicholas, F.W. 1980. Size of population required for artificial selection. Genet. Res. 35:85–105. doi:10.1017/S0016672300013951

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 5:103–113 10.3835/plantgenome2012.06.0006. doi:10.3835/plantgenome2012.06.0006

Schulz-Streeck, T., J.O. Ogutu, Z. Karaman, C. Knaak, and H.P. Piepho. 2012. Genomic selection using multiple populations. Crop Sci. 52:2453–2461. doi:10.2135/cropsci2012.03.0160

Whittaker, J.C., R. Thompson, and M.C. Denham. 1999. Marker-assisted selection using ridge regression. Annals of Human Genetics 63(4):366.

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J-L. Jannink, M.E. Sorrells, R. Babu, J.E. Cairns, A. Tarekegne, K. Segman, Y. Beyene, P. Grudloyma, F. Technow, C. Riedelsheimer, and A. E. Melchinger. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2:1427–1436.

Woolliams, J., P. Bijma, and B. Villanueva. 1999. Expected genetic contributions and their impact on gene flow and genetic gain. Genetics 153:1009–1020.