Genomewide Selection and Prospective Targeted Recombination in Elite Maize Biparental Crosses

A DISSERTATION SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA BY

Sofia Patricia Brandariz Zerboni

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Rex Bernardo, Advisor

August 2018

Acknowledgments

I would like to express a heartfelt thank you to my advisor, Dr. Rex Bernardo, for the incredible support, constant encouragement, dedicated guidance, and kindness that I received from him in my time at the University of Minnesota. He helped me grow as a scientist but also as a person, caring for my professional development but also for my personal happiness. I would like to thank my advisory committee, Drs. Aaron Lorenz, Shengqiang Zhong, Ruth Shaw and Yang Da, for their valuable advice, lessons and support through my Ph.D. process. I am also thankful to Drs. David Butruille, Mike Lohuis, and Sam Eathington for allowing us to use the datasets and for the Ph.D. fellowship provided by Monsanto.

I am very grateful to the instructors from the University of Minnesota and University of the Republic, Uruguay for their input to my development as a scientist. I would also like to thank my previous mentors that helped me be here: Drs. Lucía Gutiérrez, Ariel Castro, Martín Quincke, Omar Borsani, Virginia Gravina, Pablo Speranza and MS. Luis Viega for the great education and support they gave me.

Thank you to my previous and current lab mates: Elizabeth Blissett for her close friendship and support, Emmanuel Adeyemo for his amazing joy and great scientific questions, Nicholas Ames and Sushan Ru for their very valuable help with my analyses, and Joshua Sleper for his encouragement. Thank you to Chona Bernardo for her huge care of all the lab members. I would also like to thank my other friends and colleagues at the University of Minnesota, for their amazing company through this challenging process. A special thank to Mel, Cami, Yinjie, Kayla, Sydney, Adri, Austin, Zenith, Ryan, Nikki, Maddy, Sara, Amrit, Seth, Víctor, Rebekah, Charlie, Ilse, Becky, Cody, James, and Hannah, for making me feel like home.

I am deeply grateful to my parents, José and Matilde, for always believing in me, for listening to me when I need advice, and for their endless support and love. To my eldest sister, Majo, for helping me pass through challenging moments, for her great humor and deep empathy. To my sister Jime, for her amazing support and guidance, for the fun visits and trips together and for always being there for me. To my youngest sister Raque,

for her jokes and great support at late night hours. To Richie for being as family. To my nephew Romeo, for making me feel the proudest auntie in the world. To my extended family, specially to my mother in law Mariela, for her care and love.

To my friends from my home country, Uruguay, because they made me feel as they were with me in every moment of the Ph.D. process even though they were miles, miles away. Thank you for your visits and video calls: Romi Calasso, Maca, Belu, Romi Máscolo, Clauchu, Bettus, Nati, Diego, Santi, Manu, Sarita, Ger, Clemen, Ine, Beluchi, Maga, Flo, Vale, Kari, Agus, Elita, Martín, Vero, Maru, Pau, and Cami.

Last but not least, I am truly thankful to my husband, Guillermo. Thank you for your everlasting patience, care and advice when I need it. For your encouragement, enthusiasm and support in our future. For being my partner in life. Thank you.

Abstract

Genomewide selection is now routine in maize (*Zea mays* L.). The first two studies in this dissertation investigated advanced aspects of genomewide selection, whereas the last study investigated the potential of targeted recombination in elite biparental crosses. The three studies utilized data on 969 maize populations that were phenotyped for yield, moisture, and test weight at multiple years and locations and genotyped with 2911 single nucleotide polymorphism markers. The first study showed that prior selection in the training population reduced the response to selection and predictive ability, but it did not increase the similarity among the best lines. Including a small number of the poorest lines in the training population nearly restored the predictive ability to its original level. The second study showed that in genomewide selection, it is better to use a smaller ad hoc training population than a single, large training population. In particular, the response to selection and predictive ability were lower in a global training population with about 50,000 lines than in a set of about 4500 lines chosen to maximize relatedness with the population undergoing genomewide selection. The third study showed that on average, targeted recombination doubled the predicted gains. For each trait, the gains with targeted recombination were 60% to 400% of the gains from nontargeted recombination. Targeted recombination did not increase gains in only around 4% of the populations. The results indicated that targeted recombination is a potentially powerful sequel to genomewide selection in maize.

Table of Contents

List of Tables

List of Figures

Chapter 1: Maintaining the accuracy of genomewide predictions when selection has occurred in the training population

Routine genomewide selection in maize (*Zea mays* L.) will lead to phenotyping only a subset of the lines in a biparental population between inbreds A and B. If the cross is used as part of the training population for predicting the performance of lines in a future cross, the training population would be a selected rather than a random subset of lines. Our objective was to determine if selection in the training population (i) reduces the response to selection and accuracy of genomewide selection in a biparental (A/B) population, and (ii) increases the genetic similarity of the best lines in the A/B population. A total of 969 biparental maize populations were evaluated at 4 to 12 environments from 2000 to 2008 for grain yield, moisture, and test weight. The parents of the 969 populations were genotyped with 2911 single nucleotide polymorphism (SNP) markers, and marker data were imputed from lower-density screening of the progeny in each biparental cross. Having phenotypic information on only a selected fraction (25%) of the lines significantly reduced the response to selection and predictive ability. However, augmenting the training set with the five poorest lines nearly restored the predictions to their original level of accuracy. Prior selection in the training population did not increase the genetic similarity (calculated from nonimputed SNP data) of the best lines in the A/B population. We concluded that including a small number of the poorest lines in a training population is a practical way to maintain the effectiveness of genomewide selection.

Introduction

Maize (*Zea mays* L.) breeding usually involves crossing two inbreds (A and B) and evaluating the testcross performance of selfed lines, recombinant inbreds, or doubled haploids from the A/B cross (Hallauer, 1990). Genomewide selection (or genomic selection) can be performed in the A/B biparental cross, and previous studies have shown that genomewide selection is effective when the training population is representative of the A/B population (Schulz-Streeck et al., 2012; Riedelsheimer et al., 2013; Jacobson et al., 2014). The general combining ability (GCA) model, where A/* populations (* being a parent from the same heterotic group as A and B) and */B populations are pooled into a training population to predict the performance of progeny in the A/B cross, has been found

useful for routine genomewide selection in elite maize biparental crosses (Jacobson et al., 2014). The GCA model circumvented the need to phenotype progeny of the A/B cross and led to 85% of the gains that eventually would have been obtained with phenotypic selection in the A/B cross (Jacobson et al., 2014). In addition, the predictions from the GCA model were improved by imputing single nucleotide polymorphism (SNP) data from the parents to their progeny (Jacobson et al., 2015a).

The GCA model therefore provides a simple framework for the routine use of genomewide selection in maize breeding programs. As genomewide selection in maize biparental crosses becomes increasingly routine, it is important to consider the long-term effects of the procedure. In particular, routine genomewide selection within a breeding program will lead to only a subset of the lines in a biparental cross being phenotyped prior to the best candidates being eventually released as cultivars. If that cross is used as part of the training population for predicting the performance of lines in a future cross with either of the parents (Fig. 1), the training population would be a selected subset of lines. Having a nonrandom subset of progeny in the training population could in turn affect the accuracy of future predictions and increase the genetic similarity of the lines predicted to be the best (Jacobson et al., 2015b).

Information is currently lacking on how the effectiveness of genomewide selection is affected by prior selection in the training population. Our objective was therefore to determine if selection in the training population (i) reduces the response to selection and accuracy of genomewide selection in an A/B population, and (ii) increases the genetic similarity of the best lines in the A/B population.

Materials and methods

Phenotypic and marker data

Monsanto provided us with testcross phenotypic and SNP marker data for 969 biparental maize populations. The populations were evaluated for grain yield (Mg ha^{-1}), moisture (g H_2O kg^{-1}), and test weight (kg hL^{-1}) at 4 to 12 environments (year-location combinations) in the United States from 2000 to 2008 (Jacobson et al., 2014). A total of 27 F_2 populations were selected as the A/B populations that were subjected to genomewide selection. The 27 A/B populations were chosen on the basis of having at least

four A/* and */B crosses, a minimum population size of 50 lines, and an entry-mean heritability (h^2) significantly greater than zero for each trait (Lian et al., 2014; Jacobson et al., 2015a, 2015b). The lines in the A/B, A/*, and */B populations had the same inbred as the tester. The heritability on an entry-mean basis was estimated as $h^2 = V_G/(V_G + V_R/e)$, where V_G was the genetic variance, V_R was the residual variance, and e was the harmonic mean of the number of locations in the trial (Holland et al., 2003). Only F_2 populations with h^2 significantly different from zero ($P = 0.05$) were used.

The parents of the populations were genotyped with 2911 SNP markers, whereas the progeny in each of the 27 F_2 populations were genotyped with 25 to 123 markers. The genotypes at each locus were coded as 1 if the line was homozygous for the SNP allele from parent A, −1 if the line was homozygous for the SNP allele from parent B, and 0 if the line was heterozygous. Marker loci that were monomorphic between the two parental inbreds or that had a minor allele frequency <0.10 were excluded within each population (Lian et al., 2014; Jacobson et al., 2015a). The SNP data for the progeny were then imputed from the parental SNP data, based on the conditional probability of a nonobserved marker genotype given the two flanking-marker genotypes (Jacobson et al., 2015a). The imputed marker data were used to measure the response to selection and predictive ability when selection occurred in the A/* and */B populations, whereas the nonimputed SNP markers were used to calculate the genetic similarity among the 10% best lines in each A/B population (as described below).

Prior selection in the training population

Suppose A/*, */C, B/*, and */D populations are evaluated in Year 0. A/C and B/D are new biparental crosses available in Year 1. According to the GCA model, the A/* and */C populations can be pooled as a training population to predict the performance of progeny in the A/C cross (Fig. 1; Jacobson et al., 2014). Likewise, the B/* and */D populations can be pooled as a training population for the B/D cross. From genomewide predictions, the A/C progeny and B/D progeny predicted to be the best are subsequently phenotyped in field trials. Assessing the effects of prior genomewide selection therefore requires Year 0 training populations to predict the performance of Year 1 populations, and

the Year 1 populations becoming part of the training population to predict the performance of Year 2 populations.

Only two of the 27 A/B populations had the above conditions necessary for prior selection based on genomewide predictions. We therefore performed two experiments. In Exp. 1, prior selection in the training population was based on phenotypic values. Prior selection was considered for all 27 A/B crosses, and phenotypic selection was therefore considered as a surrogate for genomewide selection. The ramifications of this assumption are discussed in the Results and Discussion. In Exp. 2, prior selection in the training population was based on genomewide predictions, and such prior selection was considered for only the two A/B crosses for which prior selection via the GCA model could be studied.

In both Exp. 1 and 2, genomewide selection was conducted via seven schemes that were functions of the percentage (denoted by P_{Sel}) of lines retained in A/* and */B crosses on the basis of line performance (100, 50, and 25%), whether or not the A/* and */B progeny were a selected subset or a random subset, and whether or not the training population was augmented with the five poorest lines in each of the A/* and */B crosses. The seven schemes were: (i) no selection in the training population (P_{Sel} = 100%), (ii) the top 25% of lines retained (P_{Sel} = 25%), (iii) the top 50% of lines retained (P_{Sel} = 50%), (iv) the top 25% of lines plus the poorest five lines (P_{Sel} = 25% + 5), (v) the top 50% of lines plus the poorest five lines (P_{Sel} = 50% + 5), (vi) the 25% of lines chosen at random (P_{Sel} = 25% random), and (vii) the 50% of lines chosen at random (P_{Sel} = 50% random).

Response to selection and genetic diversity among selected lines

Marker effects were obtained for all 2911 SNP loci (nonimputed and imputed) by ridge regression–best linear unbiased prediction (RR-BLUP) as implemented in the rrBLUP package (Endelman, 2011) in R software (R Core Team, 2018). In Exp. 1, marker effects were estimated separately within each A/* and */B population for each trait. The marker effects were then averaged across the A/* and */B populations (Jacobson et al., 2014; 2015a). In Exp. 2, marker effects were first estimated separately for the (i) A/* and */C populations evaluated in Year 0, (ii) B/* and */D populations evaluated in Year 0, and (iii) A/C and B/D populations evaluated in Year 1. Marker effects were then averaged as

follows: (i) across the A/* and */C populations to predict the performance of progeny in the A/C population, (ii) across B/* and */D to predict the performance of progeny in the B/D population, and (iii) across the A/C and B/D populations to predict the performance of progeny in the A/B population. The performance of all N individuals in the A/C, B/D, and A/B populations was predicted as $\mathbf{y} = \mu\mathbf{1} + \mathbf{Xm}$, where $\mathbf{y}$ was an N x 1 vector of predicted performance, μ was the estimated overall mean, $\mathbf{1}$ was an N x 1 vector with elements equal to 1, $\mathbf{X}$ was an N x N_M incidence matrix with elements of 1, −1, and 0, and $\mathbf{m}$ was an N_M x 1 vector of RR-BLUP marker effects (Jacobson et al., 2014). The response to selection (R) and predictive ability (r_{MP}) were calculated for each A/B population as described by Jacobson et al. (2014, 2015a). For each trait, R was calculated as the phenotypic mean of the 10% of lines with the best predicted performance minus the overall mean of the A/B population. The r_{MP} was calculated as the correlation between the marker-predicted and observed values for the progeny in each A/B population. A t test was used to determine if the R and r_{MP} values across the A/B populations were significantly different ($P = 0.05$) among the seven selection schemes for each trait.

The genetic similarity between pairs of selected lines was calculated as the simple matching coefficient across the SNP loci (Sokal and Michener, 1958; Jacobson et al., 2015b). First, we calculated the within-locus simple matching coefficients by considering the possible combinations of marker genotypes (*MM*, *Mm*, and *mm*) between two lines. The simple matching coefficient was 1 between *MM* and *MM* or between *mm* and *mm*, 0 between *MM* and *mm*, and 0.50 between *Mm* and any other genotype (*MM*, *Mm*, or *mm*). Second, we calculated the mean of the within-locus simple matching coefficients across the SNP loci. The simple matching coefficients were calculated separately for each of the seven selection schemes and the three traits (Jacobson et al., 2015b). We performed t tests to test the significance ($P = 0.05$) of differences between mean simple matching coefficients.

Results and discussion

Response to selection and predictive ability when selection has occurred in the training population

In Exp. 1, phenotypic selection of the best 25% of lines (P_{Sel} = 25%) in the A/* and */B training populations significantly (P = 0.05) reduced R and r_{MP}. For yield, R decreased from 0.22 Mg ha^{-1} with no selection (P_{Sel} = 100%) to 0.07 Mg ha^{-1} with P_{Sel} = 25%, whereas r_{MP} decreased from 0.20 with P_{Sel} = 100% to 0.08 with P_{Sel} = 25%. For all traits, R and r_{MP} were reduced by 50% or more (Table 1).

In contrast, phenotypic selection of the best 50% of lines in the A/* and */B training populations did not significantly reduce R and r_{MP}, except r_{MP} for test weight (Table 1). These results suggested that the effect of selection on subsequent genomewide selection depends on the proportion of lines selected. Among 17 maize populations in a commercial breeding program, the percentage of lines selected during first-year testing had a mean of 25.4% and a maximum of 37% (Bernardo, 1991). These results suggested that the significant (P = 0.05) changes in R and r_{MP} with P_{Sel} = 25% are more relevant than the nonsignificant changes in R and r_{MP} with P_{Sel} = 50%. Therefore, selection in the training population in current maize breeding programs could be compromising long-term gains.

Because selection in the training population decreases the training population size (N), it was important to separate the effects of selection per se from the effects of a lower N on R and r_{MP}. Without selection, the mean size (range in parentheses) of the A/* and */B training populations for the 27 A/B populations was 4525 lines (894–10,171). With P_{Sel} = 25%, the mean size of the training population decreased to 1131 lines (224–2543). Theoretical (Daetwyler et al., 2008, 2010) and empirical (Endelman et al., 2014; Lian et al., 2014) results have shown that the expected prediction accuracy decreases as $\sqrt{N}$ decreases. The lower R and r_{MP} with P_{Sel} = 25% could therefore have been due to selection, a lower $\sqrt{N}$, or a combination of selection and a lower $\sqrt{N}$. When 25% of the lines were chosen at random (P_{Sel} = 25% random), however, R and r_{MP} were not significantly reduced compared with no selection (Table 1). Furthermore, the differences in both R and r_{MP} were significantly different between P_{Sel} = 25% selected and P_{Sel} = 25% random, except for R

for grain yield, which was 0.07 Mg ha^{-1} with P_{Sel} = 25% and 0.14 Mg ha^{-1} with P_{Sel} = 25% random (Table 1). These results indicated that selection itself played a key role in reducing R and r_{MP}. We speculate that if the training populations are smaller, both selection and a lower $\sqrt{N}$ would be important in the effectiveness of genomewide selection.

Including the five poorest lines in the training populations (P_{Sel} = 25% + 5) restored R and r_{MP} (Table 1). For yield, R was initially 0.22 Mg ha^{-1} with no selection, decreased to 0.07 Mg ha^{-1} with P_{Sel} = 25%, and increased to 0.18 Mg ha^{-1} with P_{Sel} = 25% + 5 (Table 1). The corresponding r_{MP} for grain yield was initially 0.20 with no selection, decreased to 0.08 with P_{Sel} = 25%, and increased to 0.18 with P_{Sel} = 25% + 5. Similar trends were found for moisture and test weight (Table 1). These results support previous simulation results which showed that in beef cattle (*Bos taurus* L.), including individuals with extreme yield deviation values (best and poorest) in a reference population can lead to the highest predictive ability of breeding values (Boligon et al., 2012). Including a set of poorer lines in the training population is therefore a simple approach to maintain R and r_{MP}.

If poor lines are to be included in the training population, a relevant question is how many poor lines to include. Including a larger number of lines (e.g., 10 instead of five) would increase the space requirements in subsequent phenotyping. Including only the very poorest line (e.g., P_{Sel} = 25% + 1) would lead to lower space requirements compared with P_{Sel} = 25% + 5. For yield, R was initially 0.22 Mg ha^{-1} with no selection, decreased to 0.07 Mg ha^{-1} with P_{Sel} = 25%, increased to 0.18 Mg ha^{-1} with P_{Sel} = 25% + 5 (Table 1), and was 0.14 Mg ha^{-1} with P_{Sel} = 25% + 1. The corresponding r_{MP} was initially 0.20 with no selection, decreased to 0.08 with P_{Sel} = 25%, increased to 0.18 with P_{Sel} = 25% + 5, and was 0.10 with P_{Sel} = 25% + 1. If only the poorest line is included, it becomes a high-leverage observation that will force the regression line to fit perfectly or nearly perfectly at the poorest value due to the lack of neighboring observations (Chatterjee and Hadi, 1986). Due to the risks associated with a high-leverage observation (which could be a single outlier) and the lack of improvement in R and r_{MP} with P_{Sel} = 25% + 1 over P_{Sel}

= 25% + 5, we do not recommend including only the very poorest line in the training population.

In Exp. 2, in which selection in the A/* and */B training populations was done of the basis of genomewide predictions, the overall trends in R and r_{MP} were similar as with Exp. 1. In Exp. 2, R and r_{MP} for the three traits were generally the largest with no selection, smallest with P_{Sel} = 25%, and intermediate with P_{Sel} = 50%. However, the only significant (P = 0.05) differences were for r_{MP} for test weight, and the few significant differences in r_{MP} were small (Table 2). We attributed this failure to detect significant differences to the low number of A/B populations in Exp. 2. Whereas Exp. 1 involved 27 A/B populations, Exp. 2 had only two A/B populations.

We used phenotypic selection (Exp. 1) as a surrogate for genomewide selection (Exp. 2) because of the larger number of A/B populations available in Exp. 1 than in Exp. 2. The similar trends in both experiments suggested that the effect of selection in the training population did not strongly depend on whether phenotypic selection or genomewide selection was used. In addition, the effect of selection on the subsequent R would depend on the stringency of selection: the more stringent the selection, the greater the effect we should expect. In the maize populations we studied, phenotypic selection was more stringent than genomewide selection, as evidenced by the higher R for phenotypic selection in the Jacobson et al. (2014) study. Hence, phenotypic selection as implemented in Exp. 1 was arguably equivalent to a stringent case of routine genomewide selection in a breeding program.

Diversity of best predicted lines when selection has occurred in the training population

In Exp. 1, the mean genetic diversity of lines predicted to be the best was close to 0.53 for all levels of P_{Sel} and for all traits. In Exp. 2, the mean genetic diversity of lines predicted to be the best ranged from 0.51 to 0.52 for yield and moisture, and from 0.51 to 0.53 for test weight (data not shown). The similar mean genetic diversity across different levels of P_{Sel} indicated a maintenance of genetic diversity among the lines that would be selected. In a previous study with the same 27 A/B populations, genomewide selection (with no selection in the training population) led to only a slight increase in the genetic similarity among selected lines compared with phenotypic selection (Jacobson et al.,

2015b). These previous authors speculated that the increase in the genetic similarity among the selected lines was minimal because of the near-zero probability of having a line that had the favorable SNP allele across all loci. We speculate that the same reason applied to the current study.

Application

Our results suggested that prior selection in the training population reduces R and r_{MP} in maize biparental populations, but it does not increase the genetic similarity of the best lines in the A/B population. To counteract this decrease in R and r_{MP}, we recommend including a small number (e.g., five) of the poorest lines in the A/* and */B crosses that form a training population for the A/B population. Phenotyping the poorest lines, which are unlikely to become cultivars, will require using valuable field space. Phenotyping the five lines predicted to be the poorest should therefore be viewed as an investment to maintain the long-term effectiveness of genomewide selection. Including the five lines with the poorest predicted performance would also serve as a check for the short-term effectiveness of genomewide selection. For example, if field tests later show that the five poorest lines in A/* have the same performance as the P_{Sel} = 25% of lines in A/*, the breeders would then be alerted that genomewide selection might be less effective than expected.

Table 1: Mean and range (in parentheses) of response to selection (R) and predictive ability (r_{MP}) via the general combining ability (GCA) model in Exp. 1, for seven schemes that involved phenotypic selection of different percentages (P_{Sel}) of lines in the A/* and */B training populations (* being a parent from the same heterotic group as A and B).

	Grain yield		Moisture		Test weight	
P_{Sel} (%)	R (Mg ha^{-1})	r_{MP}	R (g kg^{-1})	r_{MP}	R (kg hL^{-1})	r_{MP}
100	0.22a† (−0.16, 0.45)	0.20a (−0.04, 0.36)	−6.41a (−15.07, 0.33)	0.42a (−0.11, 0.67)	0.52a (0.10, 1.01)	0.36a (−0.06, 0.59)
50	0.17a (−0.18, 0.37)	0.15a (−0.08, 0.27)	−4.97a (−11.01, 1.07)	0.36a (−0.07, 0.54)	0.40a (−0.20, 0.88)	0.28b (−0.17, 0.52)
50 + 5‡	0.20a (−0.11, 0.47)	0.19a (−0.02, 0.34)	−6.08a (−13.06, 1.41)	0.41a (−0.09, 0.61)	0.49a (−0.06, 0.96)	0.34a (−0.04, 0.57)
25	0.07b (−0.27, 0.31)	0.08b (−0.13, 0.29)	−2.69b (−13.59, 3.94)	0.21b (−0.09, 0.52)	0.21b (−0.48, 0.83)	0.15c (−0.22, 0.43)
25+ 5	0.18a (−0.14, 0.48)	0.18a (−0.03, 0.34)	−5.80a (−13.31, 1.10)	0.40a (−0.09, 0.61)	0.47a (0.01, 0.96)	0.33a (0.00, 0.56)
50 random	0.19a (−0.19, 0.46)	0.19a (−0.04, 0.37)	−6.13a (−15.70, 0.43)	0.41a (−0.08, 0.63)	0.46a (−0.03, 0.87)	0.35a (−0.04, 0.56)
25 random	0.14ab (−0.21, 0.43)	0.16a (−0.13, 0.35)	−5.79a (−13.88, 0.24)	0.39a (−0.10, 0.63)	0.44a (−0.01, 0.77)	0.30a (−0.03, 0.50)

† Within a column, estimates with a common letter were not significantly different ($P = 0.05$).

‡ + 5 indicates that the five poorest lines were added.

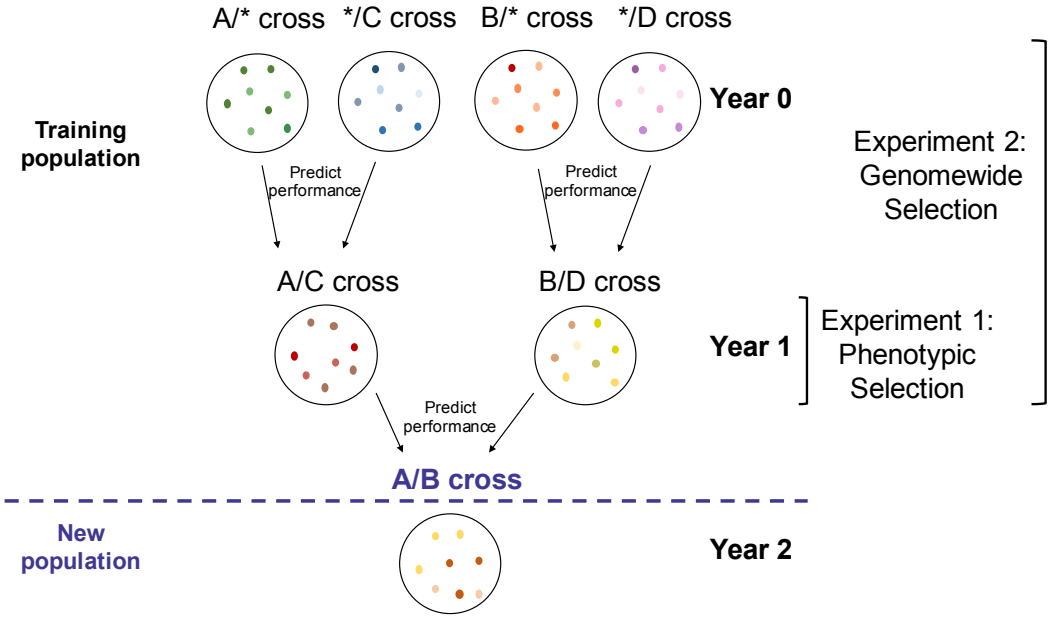
Table 2: Mean and range (in parentheses) of response to selection (R) and predictive ability (r_{MP}) via the general combining ability (GCA) model in Exp. 2, for five schemes that involved genomewide selection of different percentages (P_{Sel}) of lines in the A/* and */B training populations (* being a parent from the same heterotic group as A and B).

	Grain yield		Moisture		Test weight	
P_{Sel} (%)	R (Mg ha^{-1})	r_{MP}	R (g kg^{-1})	r_{MP}	R (kg hL^{-1})	r_{MP}
100	0.21a† (0.19, 0.22)	0.17a (0.10, 0.24)	−3.78a (−7.78, 0.22)	0.28a (0.09, 0.47)	0.39a (0.34, 0.44)	0.22a (0.20, 0.23)
50	0.14a (0.07, 0.21)	0.17a (0.10, 0.24)	−4.39a (−8.01, −0.78)	0.25a (0.08, 0.42)	0.27a (0.08, 0.46)	0.12ab (0.12, 0.12)
50 + 5‡	0.17a (0.13, 0.22)	0.19a (0.15, 0.24)	−3.96a (−8.14, 0.22)	0.25a (0.08, 0.42)	0.15a (−0.06, 0.37)	0.12b (0.11, 0.13)
25	0.07a (0.02, 0.12)	0.16a (0.06, 0.26)	−3.62a (−5.96, −1.29)	0.20a (0.09, 0.31)	0.19a (−0.01, 0.39)	0.12b (0.11, 0.12)
25 + 5	0.14a (0.14, 0.14)	0.18a (0.11, 0.25)	−3.32a (−4.54, −2.10)	0.18a (0.04, 0.31)	0.17a (−0.05, 0.38)	0.11ab (0.08, 0.13)

† Within a column, estimates with a common letter were not significantly different ($P = 0.05$).

‡ + 5 indicates that the five poorest lines were added.

Figure 1: General combining ability (GCA) model and prior selection in the training population on the basis of phenotypic values (Exp. 1) and genomewide predictions (Exp. 2). In Exp. 1, all available A/* and */B crosses (* being a parent from the same heterotic group as A and B) were pooled to predict the performance of progeny in each of 27 biparental (A/B) crosses. In Exp. 2, the A/* and */C populations evaluated in Year 0 were pooled to predict the performance of progeny in the A/C cross, whereas the B/* and */D populations evaluated in Year 0 were pooled to predict the performance of progeny in the B/D cross. The A/C and B/D progeny with the best predicted performance were phenotyped in Year 1. The A/C and B/D progeny phenotyped in Year 1 were then pooled to predict the performance of progeny in the A/B cross in Year 2.



Chapter 2: Small ad hoc versus large general training populations for genomewide selection in maize biparental crosses

In genomewide selection, different types of training populations can be used for a biparental population made from homozygous parents (A and B). Our objective was to determine if the response to selection (R) and predictive ability (r_{MP}) in an A/B population are higher with a large training population that is used for all biparental crosses, or with smaller ad hoc training populations highly related to the target A/B population. We studied 969 biparental maize (*Zea mays* L.) populations phenotyped at four to 12 environments. Parent-offspring marker imputation was done for 2911 single nucleotide polymorphism (SNP) loci. For 27 A/B populations, training populations were constructed by pooling: (i) all prior populations with A as one parent (A/*, where * is a related inbred) and with B as one parent (*/B) [General Combining Ability (GCA) model]; (ii) A/* or */B crosses only; and (iii) all */* crosses (Same Background model, SB). The SB model training population was 450–6000% as large as the GCA model training populations, but the mean coefficient of coancestry between the training population and A/B population was lower for the SB model (0.44) than for the GCA model (0.71). The GCA model had the highest R and r_{MP} for all traits. For yield, R was 0.22 Mg ha^{-1} with the GCA model, 0.15 Mg ha^{-1} with the SB model, and 0.15–0.16 Mg ha^{-1} with the A/* and */B models. We concluded that it is best to use an ad hoc training population for each A/B population.

Introduction

In maize (*Zea mays* L.) breeding, genomewide selection (or genomic selection) is typically performed among the progeny within a biparental cross. Suppose a biparental population is formed from the cross between two maize inbreds (A and B) that belong to the Iowa Stiff Stalk Synthetic (BSSS) heterotic group. A training population for the A/B cross, as well as for all other BSSS biparental crosses, can be made by pooling all prior biparental crosses that belong to the same BSSS genetic background. On the other hand, because genomewide selection is most effective when the training population is representative of the A/B population (Schulz-Streeck et al., 2012; Riedelsheimer et al., 2013; Jacobson et al., 2014), the response to selection (R) and predictive ability (r_{MP})

within the A/B population may be higher if the training population includes only those prior biparental populations that are most related to the A/B cross.

The General Combining Ability (GCA) model, which estimates the mean effects of marker alleles from parents A and B in a series of prior crosses, has been found effective for genomewide selection in an A/B population (Jacobson et al., 2014). In particular, the training population in the GCA model is formed by pooling all prior A/* populations (* being a parent from the same heterotic group as A and B) and all prior */B populations to predict the performance of progeny in the A/B cross. Among 30 maize A/B populations, the mean *R* for grain yield, moisture, and test weight was 72% higher with the GCA model than with a Same Background model that used prior */* crosses (with A/* and */B populations being excluded) as the training population (Jacobson et al., 2014). The GCA model therefore provides an easy rule for choosing an ad hoc training population that is smaller but is highly related to each A/B population undergoing genomewide selection.

In a previous study of the GCA model (Jacobson et al., 2014), the number of crosses (N_X) in the training population was kept constant between the GCA model and the Same Background model, and the number of lines (N) were kept generally similar to that in the GCA model. This assumption was needed to make an equal-resources comparison between the two models. However, this assumption did not reflect the reality that N_X and N are naturally higher in the Same Background model than in the GCA model because a breeding program has more */* crosses than A/* and */B crosses. In practice, all prior */* crosses can be used in Same Background model. As shown later in this study, the training population in the GCA model may consist of 4500 lines from all prior A/* and */B biparental crosses, but the training population in the Same Background model may consist of 50,000 lines. Because r_{MP} increases as N increases (Daetwyler et al., 2008, 2010; Endelman et al., 2014; Lian et al., 2014), the larger training population may compensate for the lower level of relatedness between the training population and A/B population in the Same Background model.

Choosing only those */* crosses that meet a specified threshold of similarity with the A/B population may increase the relatedness between the Same Background training population and the A/B population. Moreover, pooling together */* crosses with A/* and

*/B crosses may also improve the relatedness between the training population and A/B population compared to the SB model, while also increasing *N*. Furthermore, the GCA model assumed that both A and B had previously been used as parents of biparental crosses. While a new inbred obtained from an external source (Parra and Hallauer, 1996; Phillips, 2010) may immediately be used as one of the parents of an A/B cross, the new inbred would not have prior A/* or */B data available. Having either A/* crosses only or */B crosses only may decrease the effectiveness of the GCA model.

Our main objective was to determine if R and r_{MP} in an A/B population are higher with a single, large training population that is used for all biparental crosses (Same Background model), or with smaller ad hoc training populations that have a high level of relatedness with a given A/B population (GCA model). Our specific goals were to determine if the usefulness of the GCA model is diminished in comparison with having prior data on only A or B, or in comparison with a Same Background model that has a larger N_X and N as well as different levels of similarity between the */* populations and the A/B population.

Materials and methods

Phenotypic and marker data

The data and procedures have been previously described (Jacobson et al., 2014, 2015a, b; Lian et al., 2014; Brandariz and Bernardo, 2018) but are also briefly described here for the readers' convenience. Monsanto provided us with testcross phenotypic and SNP marker data for 969 biparental maize populations. The populations were evaluated for grain yield (Mg ha^{-1} at 155 g H_2O kg^{-1}), moisture (g H_2O kg^{-1}), and test weight (kg hL^{-1}) at four to 12 environments in the U.S. from 2000 to 2008 (Jacobson et al., 2014). A total of 27 F_2 populations were selected as the A/B populations according to criteria described by Jacobson et al. (2014, 2015a). The lines in the A/B and training populations had the same inbred as the tester. The A and B parents belonged to the same heterotic group whereas the tester belonged to the opposite heterotic group. Among the 969 biparental crosses, 485 A/B populations were in one heterotic group whereas 484 A/B

populations were in an opposite heterotic group. For a given trait, F_2 populations with nonsignificant ($P = 0.05$) h^2 estimates were excluded from the analysis.

The parents of the populations were genotyped with 2911 SNP markers, whereas the progeny in each of the 27 F_2 populations were genotyped with 25 to 123 markers. The genotypes at each locus were coded as 1 if the line was homozygous for the SNP allele from parent A, –1 if the line was homozygous for the SNP allele from parent B, and 0 if the line was heterozygous. Marker loci that were monomorphic between the two parental inbreds or that had a minor allele frequency less than 0.10 were excluded within each population (Lian et al., 2014; Jacobson et al., 2015a). The SNP data for the progeny were then imputed from the parental SNP data, based on the conditional probability of a non-observed marker genotype given the two flanking-marker genotypes (Jacobson et al., 2015a). A previous study with the same data sets showed that marker imputation improved the predictive ability and response to selection, and that around 500 SNP markers were sufficient for yield and 1000 SNP markers were sufficient for moisture and test weight (Jacobson et al., 2015a).

Training populations

The training populations were constructed as follows: (i) GCA model, wherein all A/* populations and all */B populations were pooled in the training population for predicting the performance of progeny in the A/B population; (ii) A/* populations only; (iii) */B populations only; (iv) Same Background (SB) model, wherein the training population comprised all available */* populations within each heterotic group; (v) SB model with the same N_X and a similar N as the GCA model (SB_{Equal}); (vi) SB model with a coefficient of similarity greater than 0.60 between the */* crosses and the A/B population ($SB_{0.60}$); (vii) SB model with a coefficient of similarity greater than 0.70 between the */* crosses and the A/B population ($SB_{0.70}$); and (viii) a combination of the SB and GCA models, with */*, A/* and */B populations pooled together (SB + GCA). The coefficient of similarity between the parents of an A/B population and training population was calculated as the simple matching coefficient across the SNP loci (Sokal and Michener, 1958), as described by Jacobson et al. (2015b).

Coefficient of coancestry

For each A/B population, we estimated the coefficient of coancestry among A, B, and the parents denoted by * to which A and B were crossed. As shown in the Results section, these coefficients of coancestry (f_{AB}, f_{A*}, and f_{*B}) determined the coefficient of coancestry between an individual in the training population and an individual in the A/B population. The marker-based coefficient of coancestry between any two individuals (X and Y) was estimated as $f_{XY} = (S_{XY} - \Theta)/(1 - \Theta)$, where S_{XY} was the marker similarity between X and Y, and Θ was the probability that two individuals share alleles that are alike in state but are not identical by descent (Lynch, 1988; Bernardo, 1993; Bernardo, 2010). Given that SNP loci are biallelic, we assumed Θ was equal to 0.50. This value of Θ further assumed allele frequencies of 0.50 among unrelated individuals. Nevertheless, Θ was expected to remain close to 0.50 as long as marker allele frequencies ranged from about 0.40 to 0.60. With the latter allele frequencies, Θ was [1 – 2(0.60)(0.40)] = 0.52 instead of 0.50.

Response to selection and predictive ability

Marker effects were obtained for all 2911 SNP loci (non-imputed and imputed) by ridge regression–best linear unbiased prediction (RR-BLUP) as implemented in the rrBLUP package (Endelman, 2011) in R software (R Core Team, 2018). We used RR-BLUP because previous studies have shown that more complex models did not substantially improve the prediction accuracy for yield (Lorenzana and Bernardo, 2009; Heffner et al., 2011; Heslot et al., 2012; Riedelsheimer et al., 2012). Marker effects were estimated separately within each cross for each trait according to procedures described by Jacobson et al. (2014). The performance of all N individuals in the A/B population was predicted as $\mathbf{y} = \mu\mathbf{1} + \mathbf{Xm}$, where $\mathbf{y}$ was an $N \times 1$ vector of predicted performance; μ was the estimated overall mean; $\mathbf{1}$ was an $N \times 1$ vector with elements equal to 1; $\mathbf{X}$ was an $N \times N_M$ incidence matrix with elements of 1, –1, and 0; and $\mathbf{m}$ was an $N_M \times 1$ vector of RR-BLUP marker effects averaged across the biparental populations in the training population, e.g., A/* and */B populations in the GCA model, and */* populations in the SB model (Jacobson et al., 2014). The R and r_{MP} were calculated for each A/B population. For each trait, R was calculated as the phenotypic mean of the 10% of lines with the best

predicted performance minus the overall mean of the A/B population. The r_{MP} was calculated as the correlation between the marker-predicted and observed values for the progeny in each A/B population. A *t* test was used to determine if the *R* and r_{MP} values across the test populations were significantly different (P = 0.05) among the training population models for each trait.

Results

Training population size, genetic similarity, and coancestry

The number of crosses (N_X) and lines (N) varied among the different types of training populations. The ranking of the models in terms of size of the training population (largest to smallest) was as follows: SB + GCA, SB, $SB_{0.60}$, $SB_{0.70}$, GCA and SB_{Equal}, and A/* and */B (Fig. 2). The SB model had a mean (range in parenthesis) N_X of 320 (289–352) and a mean N of 49,941 (45,422–54,691, Fig. 2). The GCA model had a mean N_X of 27 (5–59) and a mean N of 4,525 (894–10,171). The SB + GCA model resulted in a mean N_X of 347 (338–357) and a mean N of 54,466 (53,406–55,627). In terms of N, the SB model training population was 450 to 6000% as large as the training population in the GCA model. The A/* model training population was 60% of the size of the GCA model training population and the */B model training population was 40% of the size of the GCA model training population.

The marker similarity between the training population and the A/B population ($S_{TP,A/B}$) varied among the different models for constructing the training population. The $S_{TP,A/B}$ was highest in the GCA model, with a mean $S_{TP,A/B}$ (range in parenthesis) of 0.80 (0.73–0.86, Fig. 2). The A/* and */B models had a mean $S_{TP,A/B}$ of 0.80 (0.73–0.85). The SB_{Equal} model had the lowest $S_{TP,A/B}$, with a mean of 0.72 (0.63–0.77), but the $S_{TP,A/B}$ values were close among the SB_{Equal}, SB, and $SB_{0.60}$ and SB + GCA models.

The coefficient of coancestry between the training population and the A/B population ($f_{TP,A/B}$) likewise varied among the different models. First, the marker-based coefficients of coancestry had a mean (range in parentheses) of 0.47 (0.21–0.67) for f_{AB}, 0.65 (0.47–0.94) for f_{A*}, and 0.70 (0.45–0.92) for f_{*B} in the GCA model. Second, we found that the expected value of $f_{TP,A/B}$ was $(1 + f_{AB} + f_{A*} + f_{*B})/4$ in the GCA model. The

estimated $f_{TP,A/B}$ in the GCA model had a mean of 0.71 (0.59–0.80). Third, the marker-based coefficients of coancestry had a mean (range in parentheses) of 0.44 (0.26–0.55) for f_{A*}, and 0.43 (0.26–0.52) for f_{*B} in the SB model. Fourth, we found that the expected value of $f_{TP,A/B}$ was $(f_{A*} + f_{*B})/2$ in the SB model. The estimated $f_{TP,A/B}$ in the SB model had a mean of 0.44 (0.28–0.51). Fifth, we found that the expected value of $f_{TP,A/B}$ was $(1 + f_{AB} + f_{A*} + f_{*B})/4$ in the A/* and */B models. This expected value was the same as that for the GCA model, and the estimated $f_{TP,A/B}$ had a mean of 0.73 (0.57–0.81) in the A/* model and 0.67 (0.57–0.78) in the */B model. Sixth, the marker-based coefficients of coancestry had a mean (range in parentheses) 0.47 (0.21–0.67) for f_{AB}, 0.45 (0.28–0.56) for f_{A*}, and 0.44 (0.27–0.52) for f_{*B} in the SB + GCA model. In the SB + GCA model, the expected value of $f_{TP,A/B}$ is intermediate to the expected $f_{TP,A/B}$ values in the SB model and GCA model.

Response to selection and predictive ability with different training populations

A single, large training population (SB model and SB + GCA model) led to lower R and r_{MP} across the 27 A/B populations compared to the GCA model, but the differences were statistically significant only for moisture and test weight for the SB model, and only for test weight for the SB + GCA model ($P = 0.05$, Table 1). For grain yield, R was 0.22 Mg ha^{-1} with the GCA model, 0.15 Mg ha^{-1} with the SB model and 0.17 Mg ha^{-1} with the SB + GCA model. The corresponding r_{MP} for yield was 0.20 with the GCA model, 0.16 with the SB model and 0.18 with the SB + GCA model. For moisture and test weight, the SB model and SB + GCA model led to larger decreases in R and r_{MP} (Table 3).

Restricting the SB model to having the same number of randomly selected crosses (SB_{Equal}) as the GCA model significantly decreased R and r_{MP} for all traits ($P = 0.05$, Table 1). For grain yield, R was initially 0.22 Mg ha^{-1} with the GCA model, decreased to 0.15 Mg ha^{-1} with the SB model, and decreased further to 0.11 Mg ha^{-1} with the SB_{Equal} model. The corresponding r_{MP} for grain yield was 0.20 with the GCA model, 0.16 with the SB model, and 0.12 with the SB_{Equal} model. Larger reductions were found for moisture and test weight (Table 3). The values of R and r_{MP} with the SB_{Equal} model differed from those reported by Jacobson et al. (2014) because of three reasons: (1) the previous study used unimputed marker data, while in this study marker data were imputed from lower-density

screening of the progeny in each biparental cross (see Materials and methods); (2) the previous study used 30 instead of 27 A/B populations; and (3) the crosses for the SB_{Equal} training population were selected at random, which meant that the */* crosses used were not the same in the two studies.

Filtering the */* crosses (used in the SB model) to increase similarity with the A/B population ($SB_{0.60}$ or $SB_{0.70}$) was ineffective for improving R and r_{MP} compared to the SB model (Table 3). For grain yield, R was 0.22 Mg ha^{-1} with the GCA model and 0.15–0.16 Mg ha^{-1} with the SB, $SB_{0.60,}$ and $SB_{0.70}$ models. Similar trends were found for moisture and test weight (Table 3).

When the training population included only the A/* crosses or only the */B crosses, R and r_{MP} were reduced although the differences were significant ($P = 0.05$) only for test weight (Table 3). For yield, R was 0.22 Mg ha^{-1} with the GCA model and 0.15–0.16 Mg ha^{-1} with the A/* and */B models, whereas the corresponding r_{MP} was 0.20 with the GCA model and 0.15–016 with the A/* and */B model. Similar trends were found for moisture and test weight (Table 3).

Discussion

Our main finding was that in genomewide selection in maize, it is better to use an ad hoc training population for each A/B biparental population (GCA model) than to use a single, large training population for all A/B populations (SB model and SB + GCA model). The training population should then be put together only after each A/B population has been chosen. The results also showed the importance of both the relatedness between the training population and A/B population, and the size of the training population.

The SB model and SB + GCA model had, to our knowledge, the largest training population ever described in the literature for plants (mean N of about 50,000; Fig. 2). Despite this very large N in the SB model and SB + GCA model, the R and r_{MP} were higher with the GCA model, which had a mean N of about 4500. This result indicated that the relatedness between the A/B population and training population is more important than the size of the training population. Previous studies likewise highlighted the importance of including related crosses in the training population rather than increasing N by adding

unrelated or less-related crosses (Riedelsheimer et al., 2013; Jacobson et al., 2014; Lorenz and Smith, 2015).

The higher relatedness in the GCA model was evidenced by the higher coefficient of coancestry between the training population and the A/B population ($f_{TP,A/B}$) in the GCA model (0.71) than in the SB model (0.44) and SB + GCA model (0.45). In accordance with theoretical expectations, the estimated $f_{TP,A/B}$ was equal between the GCA model and the A/* or */B models. Furthermore, $f_{TP,A/B}$ is expected to be highest when individuals in an A/B population are used as the training population for other individuals in the same A/B population [i.e., A/B model, Jacobson et al. (2014)]. For the 27 A/B populations used in this study, the estimated $f_{TP,A/B}$ for the A/B model had a mean (range in parentheses) of 0.74 (0.60–0.84). The mean $f_{TP,A/B}$ of 0.71 in the GCA model was therefore close to the value of $f_{TP,A/B}$ in the A/B model.

The expected values of $f_{TP,A/B}$ were $(1 + f_{AB} + f_{A*} + f_{*B})/4$ in the GCA model and A/* and */B models, versus $(f_{A*} + f_{*B})/2$ in the SB model. The high $f_{TP,A/B}$ in the GCA and A/* and */B models was partly due to the higher values of f_{A*} and f_{*B} in these three models than in the SB model. These higher values of f_{A*} and f_{*B} in the GCA model and A/* and */B models were likely due to A/B crosses being made primarily within subgroups of parental inbreds. Suppose that the Iowa Stiff Stalk Synthetic maize heterotic group includes three subgroups: B14-type inbreds, B37-type inbreds, and B73-type inbreds. Inbreds within each subgroup are more closely related than inbreds in different subgroups. Furthermore, suppose that A/B crosses are most often (but not always) made within each subgroup. If A and B are both B73-type inbreds, the A/* and */B populations that are pooled into the GCA model training population will be mostly of the B73 background. In contrast, the SB model, as therefore the SB + GCA model, for a B73-type A/B population will also include crosses within the B14 and B37 subgroups. The f_{A*} and f_{*B} will consequently be lower in the SB, intermediate in the SB + GCA model and higher model than in the GCA model, as was observed in this study.

The GCA model as well as the A/*, */B and SB + GCA models obviously assume that the A and B parents have been used as parents of prior biparental crosses. There may be situations in which neither parent has been used in inbred development in previous

years. In this situation, the GCA model cannot be used and the SB model will need to be used. The lower R and r_{MP} with the SB model than with the GCA model underscores the difficulty in predicting the performance of progeny of two untested parental inbreds.

Smaller sizes of the training population have been shown to decrease prediction accuracy (r_{MG}), which is defined as the correlation between predicted and true genotypic values and is equal to r_{MP}/h, where h is the square root of heritability (Daetwyler et al., 2008, 2010; Endelman et al., 2014; Lian et al., 2014). The expected prediction accuracy is $E(r_{MG}) = r^2 [(Nh^2 / r^2Nh^2 + M_e)]^{1/2}$, where r^2 is the linkage disequilibrium between a marker and a quantitative trait locus, and M_e is the effective number of chromosome segments (Lian et al., 2014). This $E(r_{MG})$ applies only if the training population is genetically identical to the population undergoing genomewide selection (i.e., A/B model). For the 969 maize biparental crosses used in this study, Lian et al. (2014) reported mean values of $r^2 = 0.46$, $h^2 = 0.46$, and $M_e = 82$ for grain yield. The mean N in the SB model (49,941) was about 10 times larger than the mean N (4525) in the SB_{Equal} model. However, $E(r_{MG})$ [with the r^2, h^2, and M_e values reported by Lian et al. (2014)] increases by less than 4% when N increases from 4525 to 49,941. In contrast, if N increases 10-fold from 450 to 4500, $E(r_{MG})$ increases by 31%. Therefore, when N is already large, the increase in r_{MG} due to further increases in N is minor. The training population was also already large with the A/* model (mean N of 2726) and */B model (mean N of 1777). This phenomenon of diminishing returns when N is large also explained why a 50% decrease in N led to less than a 50% decrease in r_{MP} when the training population included prior crosses for only one of the two parents (A/* and */B models) in comparison with the GCA model.

In the $SB_{0.60}$ and $SB_{0.70}$ models, filtering the */* crosses to increase the similarity between the training population and the A/B population ($S_{TP,A/B}$) increased $f_{TP,A/B}$ while maintaining a large N (mean N > 36,000, Fig. 2). However, R and r_{MP} were not significantly higher in the $SB_{0.60}$ and $SB_{0.70}$ models than in the SB model (Table 3). Because the mean $S_{TP,A/B}$ in the GCA model was 0.80, we tried a stricter threshold of 0.80 ($SB_{0.80}$ model) in 20 of the 27 A/B populations for which such a stricter threshold was possible. The $SB_{0.80}$ model had a mean (range in parentheses) N of 4983 (164, 12483) and

a mean N_X of 34 (1, 88), but the R and r_{MP} in the $SB_{0.80}$ model were lower than those in the $SB_{0.70}$ and $SB_{0.60}$ models (results not shown).

This lack of improvement was noteworthy because the mean N, N_X, and $S_{TP,A/B}$ were all roughly equal between the $SB_{0.80}$ model and the GCA model. Suppose A_1 and A_2 are the parents of A, and B_1 and B_2 are the parents of B. A grandparental training population can be formed by pooling the A_1/*, A_2/*, */B_1, and */B_2 biparental crosses, but such a grandparental training population was found ineffective for predicting the performance of progeny in the A/B biparental population (Hickey et al., 2015). While all of the individual alleles in the A/B population are found among the four grandparents, chromosomal blocks found in the A/B biparental population are represented better in the A/* and */B crosses than in the grandparental crosses.

These previous results (Hickey et al., 2015), as well as those in the current study, suggest that the usefulness of the GCA model may be due to having large blocks of chromosomes in common between the A/* and */B crosses used as the training population and the A/B population undergoing genomewide selection. The current study used data on 2911 SNP markers and the Hickey et al. (2015) study simulated up to 100,000 markers. Given the agreement between our empirical results and the Hickey et al. (2015) simulation results, we speculate that our overall results would remain the same even if a larger number of SNP markers are used. In the maize populations we studied, the mean r^2 between adjacent SNP markers was 0.93 within each of the 27 A/B populations, 0.49 across the pool of A/* and */B populations used in the GCA model, and 0.23 across the pool of */* populations used in the SB model. Having a larger number of markers may increase the linkage disequilibrium in the training populations for both the GCA and SB models, but the effect of this higher linkage disequilibrium is unclear given that the linkage disequilibrium in the A/B populations was already very high (0.93) with 2911 SNP markers. The current study and the Hickey et al. (2015) studies both used RR-BLUP, and it remains to be seen whether the combined use of Bayesian models and higher marker densities would improve R and r_{MP} for traits with no major QTL (such as yield) in maize biparental crosses.

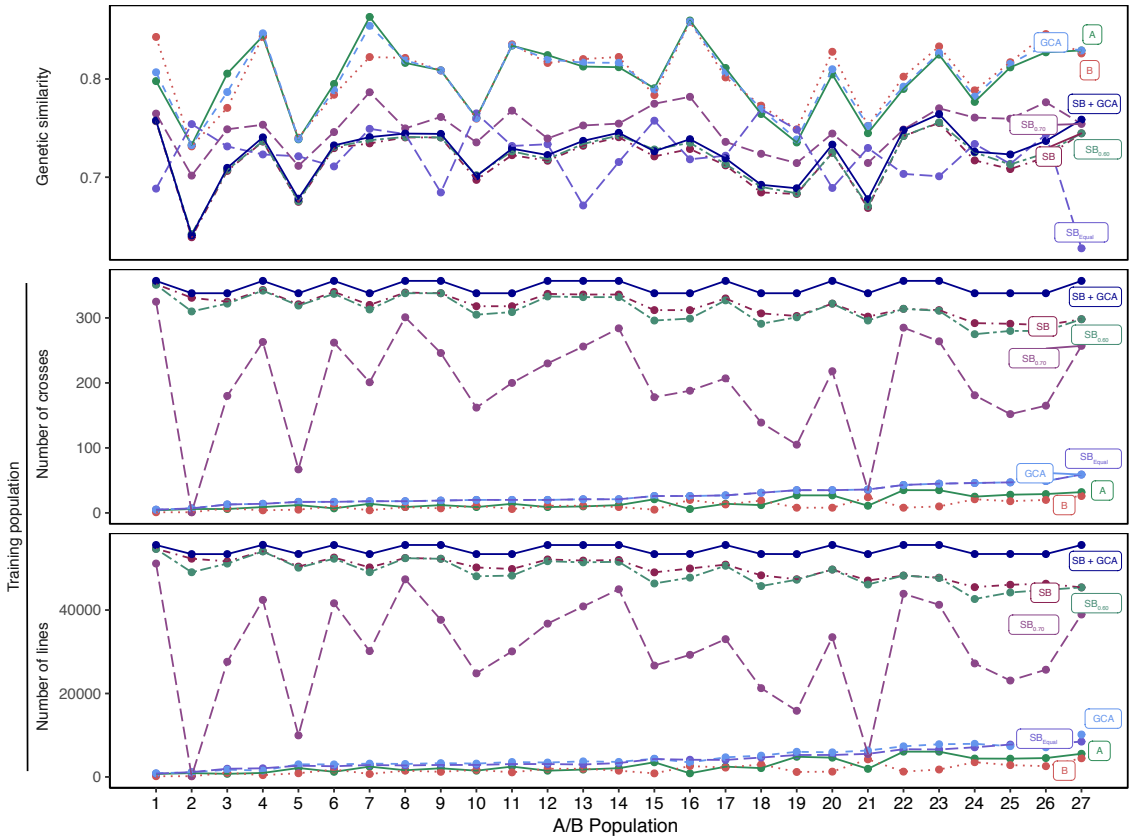
Table 3: Mean and range (in parentheses) of response to selection (R) and predictive ability (r_{MP}) for grain yield, moisture, and test weight across 27 A/B maize populations, for the following training population models: (1) General Combining Ability (GCA) model with A/* and */B crosses; (2) A/* crosses only (A/*); (3) */B crosses only (*/B); (4) Same Background model with all */* crosses (SB); (5) SB with the same number of randomly selected */* crosses as the GCA model (SB_{Equal}); (6) SB with a coefficient of similarity between the */* crosses and A/B population greater than 0.60 ($SB_{0.60}$); (7) SB with a coefficient of similarity between the */* crosses and A/B population greater than 0.70 ($SB_{0.70}$); and (8) SB + GCA model with */*, A/* and */B crosses.

	Grain yield		Moisture		Test weight	
Model	R (Mg ha^{-1})	r_{MP}	R (g kg^{-1})	r_{MP}	R (kg hL^{-1})	r_{MP}
GCA	0.22a[a] (-0.16, 0.45)	0.20a (-0.04, 0.36)	-6.41a (-15.07, 0.33)	0.42a (-0.11, 0.67)	0.52a (0.10, 1.01)	0.36a (-0.06, 0.59)
A/*	0.15ab (-0.10, 0.36)	0.15ab (-0.04, 0.43)	-5.23ab (-12.01, 1.03)	0.35ab (-0.13, 0.63)	0.42ab (0.07, 0.73)	0.28b (-0.04, 0.47)
*/B	0.16ab (-0.15, 0.47)	0.16ab (-0.09, 0.39)	-5.17ab (-14.19, 0.44)	0.35ab (-0.02, 0.64)	0.34bcd (-0.41, 0.65)	0.26b (-0.17, 0.43)
SB	0.15ab (-0.26, 0.38)	0.16ab (-0.05, 0.42)	-4.38b (-9.36, -0.98)	0.31bc (-0.03, 0.58)	0.31cd (-0.05, 0.74)	0.24bc (-0.07, 0.43)
SB_{Equal}	0.11b (-0.20, 0.38)	0.12b (-0.10, 0.39)	-3.84b (-10.69, 2.36)	0.25c (-0.06, 0.46)	0.23d (-0.33, 0.53)	0.18c (-0.02, 0.38)
$SB_{0.60}$	0.15ab	0.16ab	-4.31b	0.31bc	0.31cd	0.24bc

	(-0.26, 0.38)	(-0.05, 0.43)	(-9.36, -0.98)	(-0.03, 0.58)	(-0.05, 0.74)	(-0.06, 0.43)
$SB_{0.70}$	0.16ab (-0.19, 0.42)	0.16ab (-0.04, 0.41)	-4.40b (-10.33, -0.13)	0.32bc (-0.02, 0.58)	0.32bcd (-0.05, 0.73)	0.24bc (-0.06, 0.44)
SB + GCA	0.17ab (-0.30, 0.41)	0.18ab (-0.05, 0.45)	-5.00ab (-13.53, -0.98)	0.35ab (0.01, 0.60)	0.40bc (-0.05, 0.96)	0.29b (-0.07, 0.48)

[a] Within a column, estimates with a common letter were not significantly different (P = 0.05).

Figure 2: Mean genetic similarity between training and A/B populations (top), number of crosses in the training population (middle), and number of lines in the training population (bottom) for seven training population models across the 27 A/B populations.



Chapter 3: Predicted genetic gains from targeted recombination in elite biparental maize populations

Targeted recombination is the ability to induce or select for specific recombination points on chromosomes. A first study with the intermated B73 × Mo17 maize (*Zea mays* L.) population showed that targeted recombination doubles the predicted gains for yield and other agronomic traits. Our objective was to assess the predicted gains from targeted recombination for quantitative traits in multiple, elite maize populations. A total of 969 biparental maize populations were phenotyped at 4–12 environments in the U.S. from 2000 to 2008. Positions of one and two targeted recombinations per chromosome were determined from genomewide marker effects for 2911 single nucleotide polymorphism (SNP) loci. Relative efficiency (RE_{Targeted}) was calculated as the predicted response to targeted recombination divided by the predicted response to nontargeted recombination. On average, targeted recombination doubled the predicted genetic gains for yield, moisture, and test weight. For each trait, RE_{Targeted} ranged from around 60% to 400% among the populations, and targeted recombination did not increase gains in around 4% of the populations. The RE_{Targeted} tended to decrease as the similarity between the parents increased. Having targeted recombination on three chromosomes (for yield and test weight) to seven chromosomes (for moisture) led to the same or greater predicted gain than nontargeted recombination. Marker intervals for targeted recombination varied across populations and traits. Overall, our results for multiple, elite maize populations indicated that targeted recombination is a most promising breeding approach.

Introduction

Maize (*Zea mays* L.) breeding usually involves crossing two inbreds (A and B) to form a breeding population, developing selfed lines or doubled haploids from the A/B population, and evaluating the lines for their testcross performance (Hallauer, 1990). Breeders therefore choose which crosses to make and which progeny to select. However, breeders have not attempted to control recombination among loci for quantitative traits and have simply relied on the results of random meiosis and fertilization during the development of breeding lines (Bernardo, 2017).

Targeted recombination is the ability to induce or select for specific recombination points so that genetic gains can be maximized. The marker intervals where targeted recombination should occur can be determined from genomewide marker effects (Bernardo, 2017; Ru and Bernardo, 2018). Suppose a chromosome has three SNP (single nucleotide polymorphism) markers and the effects of marker alleles (in order on the chromosome) carried by two parental inbreds are as follows: [0.2, 0.4, –0.5] in parent A, and [–0.2, –0.4, 0.5] in parent B. If recombination occurs between the second and third markers, a [0.2, 0.4, 0.5] homologue that maximizes the genetic gain can be recovered and converted into a doubled haploid. The same process can be done with many SNP loci on a chromosome and across multiple chromosomes.

An initial study in the intermated B73 × Mo17 maize population indicated that targeted recombination can be a powerful breeding approach (Bernardo, 2017). In particular, predicted gains with one targeted recombination per chromosome ($x = 1$) in this classic maize population were twice the predicted gains with nontargeted recombination for yield and other agronomic traits (Bernardo, 2017). Predicted gains were higher with two targeted recombinations per chromosome ($x = 2$). While this initial Bernardo (2017) study indicated that targeted recombination is a promising breeding approach, we do not know if the predicted gains will also be doubled in newer, elite maize germplasm. We also do not know the extent of variation in the usefulness of targeted recombination among multiple breeding populations. Lastly, information is lacking on factors that influence the usefulness of targeted recombination for quantitative traits.

Our main objective of this study was to assess the predicted gains from targeted recombination for maize yield, moisture, and test weight in multiple, elite breeding populations. Our specific goals were to: (i) determine the extent of variation in predicted gains from targeted recombination across elite maize populations; (ii) identify factors that cause targeted recombination to be ineffective; and (iii) determine the number of chromosomes on which targeted recombination needs to be performed to achieve equal or greater predicted gain as nontargeted recombination.

Materials and methods

Phenotypic and marker data

The data have been previously described (Jacobson et al., 2014, 2015a, b; Lian et al., 2014; Brandariz and Bernardo, 2018) but are also briefly described here for the readers' convenience. Monsanto provided us with testcross phenotypic and SNP marker data for 969 biparental maize populations. The populations were evaluated for yield (Mg ha^{-1} at 155 g H_2O kg^{-1}), moisture (g H_2O kg^{-1}), and test weight (kg hL^{-1}) at four to 12 environments in the U.S. from 2000 to 2008 (Jacobson et al., 2014). Only the F_2 populations with heritability (h^2) significantly different from zero ($P = 0.05$) were used in this study.

The parents of the populations were genotyped with 2911 SNP markers, whereas the progeny were genotyped with 25 to 123 markers. The genotypes at each locus were coded as 1 if the line was homozygous for the SNP allele from parent A, –1 if the line was homozygous for the SNP allele from parent B, and 0 if the line was heterozygous. Marker loci that were monomorphic between the two parental inbreds or that had a minor allele frequency less than 0.10 were excluded within each population (Lian et al., 2014; Jacobson et al., 2015a). The SNP data for the progeny were then imputed from the parental SNP data, based on the conditional probability of a non-observed marker genotype given the two flanking-marker genotypes (Jacobson et al., 2015a). Monsanto provided us with a consensus map for all populations. The linkage map comprised 1668 cM, and the chromosome sizes ranged from 103 cM for chromosome 6 to 245 cM for chromosome 1.

Genomewide marker effects

Marker effects were obtained by ridge regression–best linear unbiased prediction (RR–BLUP) as implemented in the rrBLUP package (Endelman, 2011) in R software (R Core Team, 2018). Two training population models were used for estimating the marker effects: the A/B model, and the General Combining Ability (GCA) model. In the A/B model, marker effects were estimated from the A/B population itself. For each trait, the performance of an individual was predicted from information on the rest of the individuals ($N - 1$) as $y_p = \mu + \mathbf{Xm}$, where y_p was the predicted performance of the individual; μ was the estimated overall mean from RR-BLUP analysis of the $N - 1$ individuals used in the

training population; **X** was an $1 \times N_M$ incidence matrix with elements of 1, –1, and 0; and **m** was an $N_M \times 1$ vector of marker effects estimated from the remaining $N - 1$ individuals (Jacobson et al., 2014). The marker effects and predictive ability (r_{MP}), the latter defined as the correlation between marker-predicted genotypic values and phenotypic values, were estimated by deleting one individual at a time as described above and with cross-validation across environments as described by Jacobson et al. (2014).

In the GCA model, the training population was obtained by pooling all prior A/* populations (* being a parent from the same heterotic group as A and B) and all prior */B populations to predict the performance of progeny in the A/B cross. Marker effects were estimated separately within each cross for each trait (Jacobson et al., 2014). The performance of all N individuals in the A/B population was predicted as $\mathbf{y} = \mu\mathbf{1} + \mathbf{Xm}$, where **y** was an $N \times 1$ vector of predicted performance; **1** was an $N \times 1$ vector with elements equal to 1; **X** was an $N \times N_M$ incidence matrix with elements of 1, –1, and 0; and **m** was an $N_M \times 1$ vector of marker effects averaged across the A/* and */B populations (Jacobson et al., 2014). Marker effects were estimated with data from all environments within the A/* and */B populations, and r_{MP} was estimated using cross-validation across environments (Jacobson et al., 2014; 2015a).

A total of 27 F_2 populations were selected as the A/B populations on the basis of having at least four A/* and */B crosses, a minimum population size of 50 lines, and an entry-mean heritability (h^2) significantly greater than zero for each trait as described by Jacobson et al. (2014, 2015a). The conditions previously described allowed us to compare the A/B and GCA models only for these 27 A/B populations. The model that achieved greater r_{MP} for each trait and A/B population was selected. For the rest of the populations, only the A/B model was used. Once predictions were obtained, populations were filtered according to having (i) r_{MP} greater than 0.40, or (ii) a correlation between marker-predicted genotypic values and true genotypic values (r_{MG}) greater than 0.65. The r_{MG} values were estimated as r_{MP}/h (Dekkers, 2007; Lee et al., 2008).

The number of F_2 populations with $h^2 > 0$ was 706 for yield, 707 for moisture, and 698 for test weight. The number of F_2 populations with $r_{MP} > 0.40$ was 30 for yield, 314

for moisture, and 187 for test weight. The number of F_2 populations with $r_{MG} > 0.65$ was 6 for yield, 51 for moisture, and 17 for test weight.

Predicted gains from targeted recombination

Marker effects were used to determine the marker intervals where one and two targeted recombinations should occur on each chromosome, as described by Bernardo (2017). First, one targeted recombination was considered at each marker interval on each chromosome. The performance of two doubled haploids, induced from each of the two resulting homologues, was calculated as the sum of the effects of the alleles carried by each homologue. The homologue with superior performance (higher yield and test weight, and lower moisture) was identified. The procedure was repeated for all 10 chromosomes. The predicted performance of the doubled haploid resulting from targeted recombination was calculated as the sum of μ from RR-BLUP analysis plus the sum of the gains from targeted recombination within each chromosome. Second, two targeted recombinations were considered between all pairs of marker intervals within each chromosome, assuming independent recombinations on each chromosome. Subsequent procedures were the same as those with one targeted recombination per chromosome.

For each F_2 population, the response to selection with targeted recombination (R_{Targeted}) was calculated as the predicted performance of the best doubled haploid from targeted recombination minus the estimate of μ from RR-BLUP analysis. The response with nontargeted recombination ($R_{\text{Nontargeted}}$) was estimated as the marker-predicted performance of the best observed line minus the the estimate of μ from RR-BLUP analysis (Bernardo, 2017).

The relative efficiency of selection with targeted recombination compared to nontargeted recombination was calculated for each population as $RE_{\text{Targeted}} = (R_{\text{Targeted}} / R_{\text{Nontargeted}}) \times 100$ (Bernardo, 2017). Confidence intervals ($P = 0.05$) for the difference in the genetic gain with targeted recombination compared to nontargeted recombination were conducted by obtaining the difference in the genetic gain for each chromosome and performing 1,000 bootstrap samples within each chromosome (Ru and Bernardo, 2018; Bernardo, 2017).

Factors associated with the relative efficiency of targeted recombination

The correlation was calculated between RE_{Targeted} and r_{MP}, r_{MG}, h^2, and the marker similarity (S_{AB}) between the parents of the biparental cross. The S_{AB} between parents was estimated as the simple matching coefficient across the SNP loci (Sokal and Michener, 1958; Jacobson et al., 2015b; Brandariz and Bernardo, 2018). First, we calculated the within-locus simple matching coefficients by considering the possible combinations of marker genotypes (*MM*, *Mm*, and *mm*) between the parents. The simple matching coefficient was 1 between *MM* and *MM* or between *mm* and *mm*, 0 between *MM* and *mm*, and 0.50 between *Mm* and any other genotype (*MM*, *Mm*, or *mm*). Second, we calculated the mean of the within-locus simple matching coefficients across the SNP loci.

Results

Mean and variability of the relative efficiency of targeted versus nontargeted recombination

On average, having one (x = 1) or two (x = 2) targeted recombinations per chromosome doubled the predicted genetic gain for yield, moisture, and test weight. For yield, the mean RE_{Targeted} was 217% with x = 1 and 233% with x = 2 for populations with $h^2 > 0$; 230% with x = 1 and 258% with x = 2 for populations with $r_{MP} > 0.40$; and 243% with x = 1 and 292% with x = 2 for populations with $r_{MG} > 0.65$ (Fig. 3). The mean RE_{Targeted} for moisture and test weight was also around 200% for all subsets of populations meeting the h^2, r_{MP}, and r_{MG} criteria (Fig. 3).

While the mean RE_{Targeted} was around 200%, the individual RE_{Targeted} values differed among the populations. For the populations with $h^2 > 0$, RE_{Targeted} for yield ranged from 60 to 454% with x = 1, and from 72 to 452% with x = 2. However, for populations with $r_{MP} > 0.40$ or with $r_{MG} > 0.65$, RE_{Targeted} for yield was always greater than 100% (Fig. 3). For moisture, RE_{Targeted} ranged from 63 to 422% with x = 1, and from 80 to 449% with x = 2 for populations with $h^2 > 0$. Similar variation in RE_{Targeted} was found among populations with $r_{MP} > 0.40$ or with $r_{MG} > 0.65$ (Fig. 3). The RE_{Targeted} for test weight ranged from 60 to 415% with x = 1, and from 81 to 431% with x = 2 with $h^2 > 0$. Some populations with $r_{MP} > 0.40$ had RE_{Targeted} less than 100%, but none of the populations with $r_{MG} > 0.65$ had RE_{Targeted} less than 100% (Fig. 3).

The variation in RE_{Targeted} (i.e., ratio between predicted gains) was accompanied by variation in the predicted gain itself (i.e., numerator of RE_{Targeted}). For the populations with $h^2 > 0$, R_{Targeted} for yield ranged from <0.1 to 3.0 Mg ha^{-1} with $x = 1$, and from <0.1 to 3.7 Mg ha^{-1} with $x = 2$. These R_{Targeted} values resulted in predicted yields ranging from 9 to 17 Mg ha^{-1} with $x = 1$ and $x = 2$ (Fig. 4). About 97% of predicted yields were lower than 15.7 Mg ha^{-1} (250 bushels ac^{-1}) with either $x = 1$ and $x = 2$ with $h^2 > 0$ (Fig. 4).

The R_{Targeted} for moisture ranged from less than –0.1 to –64 g kg^{-1} with $x = 1$ and from less than –0.1 to –69.3 g kg^{-1} with $x = 2$, resulting in predicted moisture values of 123 to 313 g kg^{-1} with $x = 1$ and $x = 2$ (Fig. 4). For test weight, the R_{Targeted} ranged from <0.1 to 5 kg hL^{-1} with $x = 1$ and from <0.1 to 5.7 kg hL^{-1} with $x = 2$, resulting in predicted test weights of 68 to 82 kg hL^{-1} with $x = 1$ and $x = 2$ (Fig. 4).

Frequency of populations in which targeted recombination is likely ineffective

Targeted recombination was predicted to be ineffective when (i) RE_{Targeted} was less than 100%, or (ii) RE_{Targeted} was numerically greater than 100% but was not statistically different ($P = 0.05$) from 100%. The numbers of populations with $RE_{\text{Targeted}} < 100\%$ were as follows: 13 for yield (1.8%) with $x = 1$ and 8 (1.1%) with $x = 2$ (5 in common); 6 for moisture (0.85%) with $x = 1$ and 8 (1.1%) with $x = 2$ (2 in common); and 10 for test weight (1.4%) with $x = 1$ and 7 (1.0%) with $x = 2$ (3 in common). In these populations, the best line had more than two nontargeted recombinations on some chromosomes (results not shown). The numbers of populations with RE_{Targeted} exceeding 100%, but not significantly greater than 100%, were as follows: 20 for yield (2.8%) with $x = 1$ and 19 (2.7%) with $x = 2$ (14 in common); 15 for moisture (2.3%) with $x = 1$ and 21 (3.0%) with $x = 2$ (11 in common); and 17 for test weight (2.4%) with $x = 1$ and 19 (2.7%) with $x = 2$ (11 in common). Adding the frequencies of $RE_{\text{Targeted}} < 100\%$ and RE_{Targeted} not significantly greater than 100% led to an overall frequency of around 4% of populations for which targeted recombination for a given trait was ineffective.

The correlations were low (–0.04 to 0.14) and mostly nonsignificant between RE_{Targeted} and values of h^2, r_{MP}, and r_{MG} for all traits. However, the correlation was significant between RE_{Targeted} for each trait and the marker similarity (S_{AB}) between the parents of the biparental cross. The correlations between RE_{Targeted} and S_{AB} were as

follows: –0.23 for yield with both $x = 1$ and $x = 2$; –0.25 for moisture with $x = 1$ and –0.27 with $x = 2$; and –0.24 for test weight with $x = 1$ and –0.25 with $x = 2$.

Chromosome contributions and targeted-recombination positions

In general, chromosomal contributions to the total predicted gain (averaged across populations with $h^2 > 0$) were proportional to the sizes (in cM) of the chromosomes (Fig. 5). However, chromosomal contributions were slightly larger than expected for chromosomes 1, 5 and 2 and were slightly smaller than expected for chromosomes 8 and 9. The mean and range (in parentheses) of the minimum number of chromosomes needed for targeted recombination to reach equal or greater predicted gains compared with nontargeted recombination was: 3 (1, 8) with $x = 1$ and 3 (1, 9) with $x = 2$ for yield; 7 (1, 9) with $x = 1$ and 7 (3, 9) with $x = 2$ for moisture; and 3 (1, 9) with both $x = 1$ and $x = 2$ for test weight.

The most frequent position(s) where one and two targeted recombinations should occur varied across populations. For yield, the most frequent position with $x = 1$ on chromosome 1 had an incidence of 15% (Table 4). In other words, 15% of the populations had the same targeted-recombination position for $x = 1$ on chromosome 1. Chromosomes 6 and 9 were the chromosomes that shared the greatest incidence of the most frequent positions for each of the traits (Table 4). For these two chromosomes, around 40% of the populations had the same targeted-recombination position for $x = 1$, and around 30% of the populations shared a position for $x = 2$.

Targeted-recombination positions also varied across traits. For $x = 1$ on chromosome 1, only 3% of the populations shared the same targeted-recombination position for all three traits, 24% of the populations shared the same position for two traits, and 73% had unique positions for each trait (Table 5). More than 50% of the populations did not share targeted-recombination positions across traits, except for chromosomes 6 and 9 with $x = 1$ (Table 5).

Discussion

Targeted recombination as a promising breeding approach

Our results strongly indicated that targeted recombination can substantially improve genetic gains for yield, moisture, and test weight in elite maize breeding

populations. On average, targeted recombination doubled the predicted gains for all traits. The RE_{Targeted} values in this study were similar to those found for yield in the B73 × Mo17 maize population (212% with $x = 1$ and 254% with $x = 2$; Bernardo, 2017). Thus, targeted recombination is a most promising breeding approach for quantitative traits in maize. As discussed in the next section, this assumes that inducing artificial targeted recombination or pyramiding natural recombinations that already occur at the desired marker intervals becomes feasible.

While the predicted gains were doubled on average, the RE_{Targeted} for each trait ranged from around 60 to 400% among the populations, showing that predicted gains from targeted recombination may be population dependent. More importantly, targeted recombination was ineffective in only around 4% of the populations per trait. This low frequency suggests that targeted recombination will usually be superior to nontargeted recombination. The infrequent values of $RE_{\text{Targeted}} < 100\%$ were due to having more than two nontargeted recombinations on some chromosomes. A larger number of targeted recombinations ($x > 2$ per chromosome) is expected to increase the gains from targeted recombination, but achieving more than 1–2 targeted recombinations per chromosome will likely be difficult. The few ineffective RE_{Targeted} values were associated with crossing highly similar parents. However, the effectiveness of targeted recombination was not associated with variation in h^2, r_{MP}, and r_{MG} (assuming $h^2 > 0$).

Although we compared the R_{Targeted} for the best double haploid with the $R_{\text{Nontargeted}}$ for the best F_3 line, we do not expect a difference in the $R_{\text{Nontargeted}}$ if we would have been able to estimate it from the best double haploid. We based this on empirical results that showed similar ranges between mean testcross performances of 50 S_1 and 50 S_8 lines crossed to five different testers (Lopez-Perez, 1979). In particular, the range of the mean testcross performances when crossed to Mo17 as a tester, which was the suggested tester to use as an elite inbred from opposite heterotic group, was higher with S_1 (31.8 q ha^{-1}) than S_8 lines (27.34 q ha^{-1}) for yield, and was similar for moisture (see Table 16 from Lopez-Perez, 1979). Therefore, we expect a similar $R_{\text{Nontargeted}}$ if we would have estimated it with the best predicted double haploid.

Designing a maize breeding program that incorporates targeted recombination

Incorporating targeted recombination will alter key stages in breeding a hybrid species such as maize, as well as in breeding self-pollinated species for which predicted gains from targeted recombination also doubled the gains from nontargeted recombination (Ru and Bernardo, 2018). Our results indicated that if targeted recombination is to be used, the parents of a biparental cross should not be highly similar to each other. Therefore, the S_{AB} between the parents should be assessed.

Once the parents are chosen, the training population for obtaining the marker effects should be selected. In the A/B model (Jacobson et al., 2014), the training population is genetically identical to the population in which targeted recombination will be done. Therefore, the A/B model is likely preferable as long as the number of lines in the A/B cross is large enough to obtained high values of r_{MP}. However, a disadvantage of the A/B model is that progeny in the A/B cross need to be phenotyped to estimate the marker effects. An alternative is to use the GCA model (Jacobson et al., 2014; Jacobson et al., 2015a). In the GCA model, previous A/* and */B crosses need to be available and marker data should be imputed from lower-density screening of the progeny (Jacobson et al., 2015a). Genomewide marker effects can then be used to estimate the $R_{\text{Nontargeted}}$ and the R_{Targeted} with $x = 1$ and $x = 2$; more than two targeted recombinations per chromosome might be infeasible. Populations with significant $RE_{\text{Targeted}} > 100\%$ should be used for targeted recombination. Given that the position of a targeted recombination affects the performance of multiple traits, a selection index can be used and one or two targeted-recombination positions per chromosome can be identified for the selection index value (Bernardo, 2017).

Targeted recombination can be achieved in two ways. The most feasible way consists on screening and pyramiding recombination events with a similar procedure to marker-assisted backcrossing. This approach involves performing foreground selection for chromosomes with a targeted-recombination position, and background selection across the rest of the chromosomes (Bernardo, 2014; Bernardo, 2017). Our results indicated that targeted recombination do not need to occur on all chromosomes to achieve equal or greater gains compared to nontargeted recombination. As chromosomes

contributed in different proportions to the total predicted gains, having targeted recombination on about three (for yield and test weight) to seven (for moisture) chromosomes was predicted to achieve equal or greater gains than nontargeted recombination. Hence, for an efficient use of resources the breeders should prioritize chromosomes and combine targeted with natural recombination. Further research is needed on breeding schemes to pyramid natural targeted recombinations as efficiently as possible.

A second way of achieving targeted recombination involves CRISPR technology (clustered regularly interspaced short palindromic repeats; Cong et al., 2013; Ran et al., 2013; Hsu et al., 2014). CRISPR system has been used to target recombination for building a fine map for manganese sensitivity in yeast (Sadhu et al., 2016). In addition, it has been used to target recombination between homologous chromosomes in tomato (Filler Hayut et al., 2017). A protocol that involves a multiplex CRISPR system for inducing homologous recombinations, screening cells with the targeted recombinations, and regenerating cells into plants is yet to be developed (Bernardo, 2017). Thus, the feasibility of CRISPR technology for routinely inducing targeted recombination in plants is still undetermined.

Table 4: Incidence (%) of the most frequent position(s) for one ($x = 1$) and two ($x = 2$) targeted recombinations per chromosome, across maize F_2 populations with heritability on an entry-mean basis (h^2) greater than 0.

	Yield		Moisture		Test weight	
Chromosomes	$x = 1$	$x = 2$	$x = 1$	$x = 2$	$x = 1$	$x = 2$
1	15	4, 3	15	4, 4	14	4, 4
2	25	10, 10	20	9, 9	24	10, 10
3	25	14, 14	29	12, 12	27	14, 14
4	22	8, 8	24	9, 9	20	6, 6
5	23	9, 9	22	9, 9	22	9, 9
6	45	35, 35	42	32, 32	39	31, 31
7	28	16, 16	27	14, 14	27	15, 15
8	29	16, 16	27	14, 14	26	13, 13
9	42	33, 33	44	34, 34	41	33, 33
10	30	15, 15	26	11, 11	27	14, 14

Table 5: Percentage (%) of maize F_2 populations for which targeted-recombination positions were common for different traits.

		Chromosome									
Targeted recombinations per chromosome	Positions shared across traits	1	2	3	4	5	6	7	8	9	10
$x = 1$	All traits with same position	3	4	8	3	5	19	6	8	19	7
	Two traits with same position	24	33	36	30	33	42	38	31	46	37
	All traits with unique position	73	63	56	66	62	39	55	61	35	55
$x = 2$	All traits with same position	3, 3	5, 3	7, 6	2, 2	5, 5	17, 18	5, 7	7, 7	19, 18	6, 7
	Two traits with same position	26, 27	33, 32	33, 36	25, 31	34, 30	38, 39	35, 31	32, 34	41, 42	38, 38
	All traits with unique position	71, 70	62, 64	59, 58	73, 67	62, 65	45, 43	60, 62	61, 59	39, 40	56, 54

Populations with heritability on an entry-mean basis (h^2) greater than 0 were used. Targeted-recombination positions were evaluated for one targeted recombination ($x = 1$) and two targeted recombinations ($x = 2$) per chromosome.

Figure 3: Box-plot of relative efficiency (RE_{Targeted}, %) of selection with one targeted recombination ($x = 1$) and two targeted recombinations ($x = 2$) per chromosome compared to nontargeted recombination, for maize F_2 populations with: heritability on an entry-mean basis (h^2) greater than 0; correlation between marker-predicted genotypic values and phenotypic values (r_{MP}) greater than 0.40; and correlation between marker-predicted genotypic values and true genotypic values (r_{MG}) greater than 0.65.

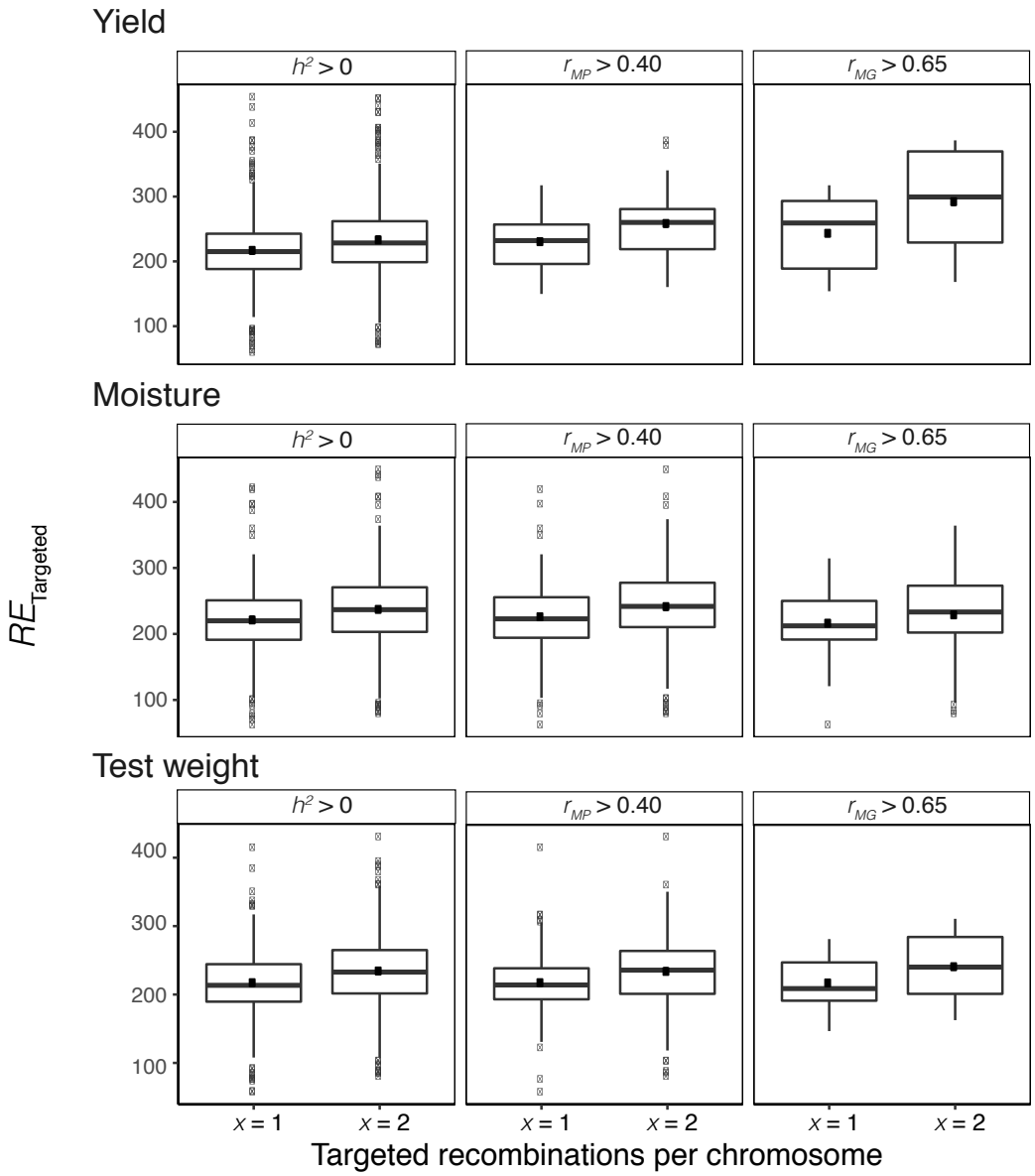


Figure 4: Predicted performance to selection with nontargeted recombination, one targeted recombination ($x = 1$) and two targeted recombinations ($x = 2$) per chromosome, for maize F_2 populations with: heritability on an entry-mean basis (h^2) greater than 0; correlation between marker-predicted genotypic values and phenotypic values (r_{MP}) greater than 0.40; and correlation between marker-predicted genotypic values and true genotypic values (r_{MG}) greater than 0.65.

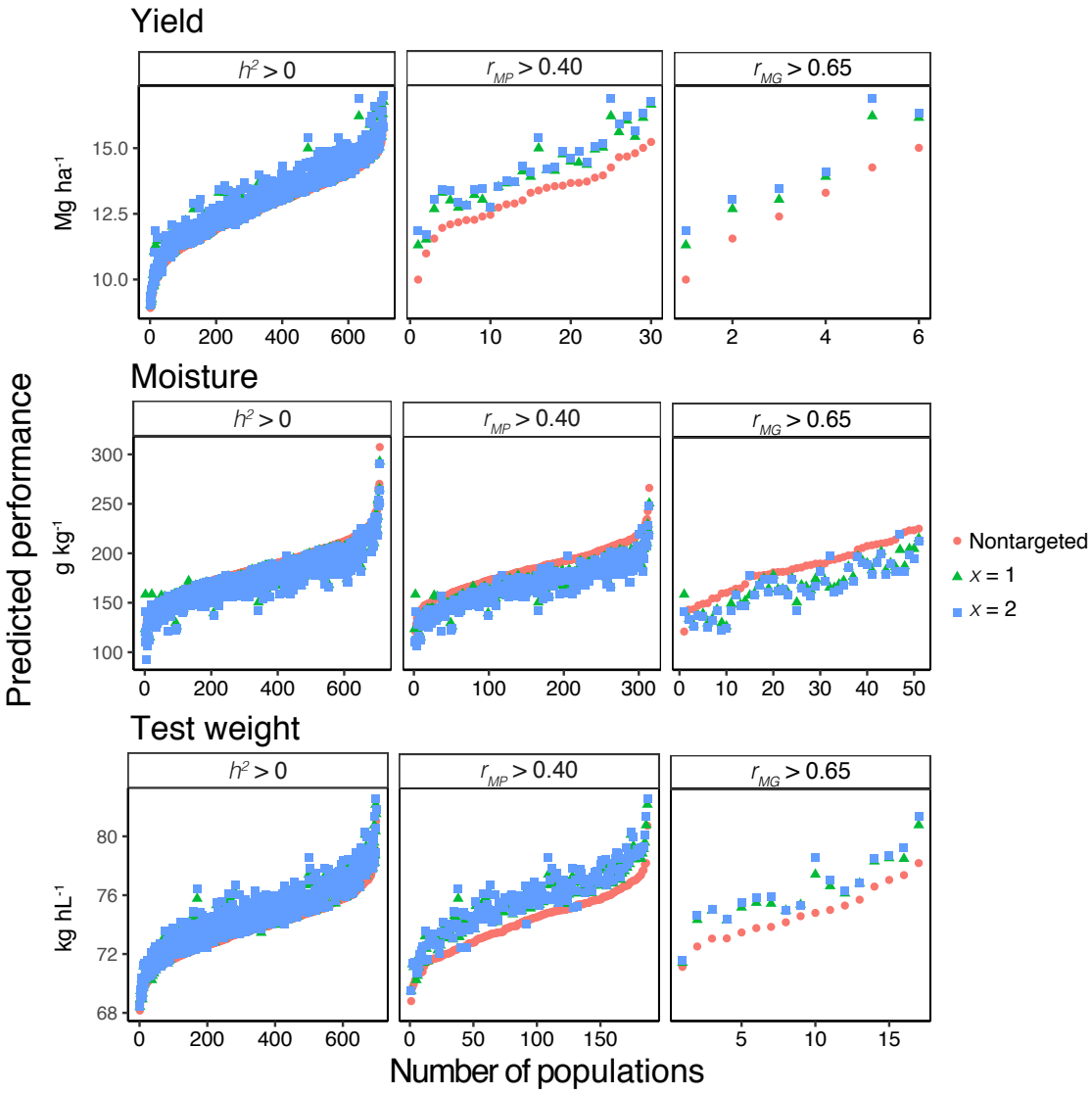
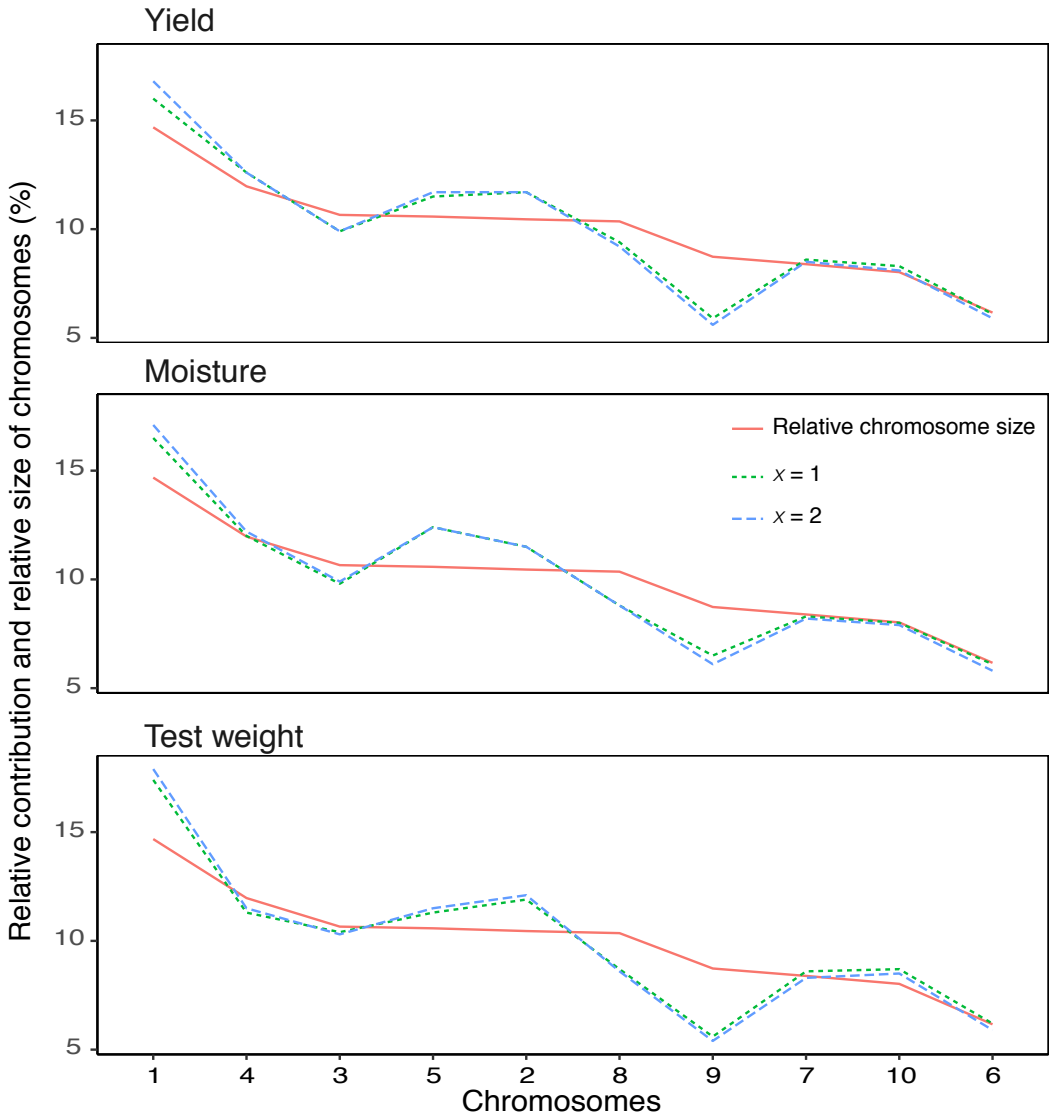


Figure 5: Mean percentage (%) of: (i) chromosome contribution relative to the total predicted gain with one targeted recombination (x = 1) and two targeted recombinations (x = 2) per chromosome; and (ii) size of each chromosome relative to the total linkage map, across maize populations with heritability on an entry-mean basis (h^2) greater than 0.



Bibliography

Bernardo, R. 1991. Retrospective index weights used in multiple trait selection in a maize breeding program. Crop Sci. 31:1174–1179.

Bernardo, R. 1993. Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet. 85:1055–1062.

Bernardo, R. 2010. Breeding for quantitative traits in plants. 2nd ed. Stemma Press, Woodbury, MN. ISBN 978-0-9720724-1-0.

Bernardo, R. 2014. Essentials of plant breeding. Stemma Press, Woodbury, MN. ISBN 978-0-9720724-2-7.

Bernardo, R. 2017. Prospective targeted recombination and genetic gains for quantitative traits in maize. Plant Genome 10.

Boligon, A.A., N. Long, L.G. Albuquerque, K.A. Weigel, D. Gianola, and G.J.M. Rosa. 2012. Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. J. Anim. Sci. 90:4716–4722.

Brandariz, S.P., and R. Bernardo. 2018. Maintaining the accuracy of genomewide predictions when selection has occurred in the training population. Crop Sci. 58:1–6.

Chaterjee, S., and A.S. Hadi. 1986. Influential observations, high leverage points, and outliers in linear regression. Stat. Sci. 1:379-416.

Cong, L., F.A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P.D. Hsu, X. Wu, W. Jiang, L.A. Marraffini, and F. Zhang. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science 339:819–823.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 1031:1021–1031.

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395.

Dekkers, J.C.M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet. 124:331–341.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R

package rrBLUP. Plant Genome 4:250–255.

Endelman, J.B., G.N. Atlin, Y. Beyene, K. Semagn, X. Zhang, M.E. Sorrells, and J.L. Jannink. 2014. Optimal design of preliminary yield trials with genome-wide markers. Crop Sci. 54:48–59.

Filler Hayut, S., C.M. Bessudo, and A.A. Levy. 2017. Targeted recombination between homologous chromosomes for precise breeding in tomato. Nat. Commun. 8: 15605.

Hallauer, A.R. 1990. Methods used in developing maize inbreds. Maydica 35:1–16.

Heffner, E.L., J.L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:1–11.

Heslot, N., H.P. Yang, M.E. Sorrells, and J.L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. Crop Sci. 52:146–160.

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B.M. Prasanna, M. Grondona, A. Zambelli, V.S. Windhausen, K. Mathews, and G. Gorjanc. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci. 54:1476–1488.

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez. 2003. Estimating and interpreting heritability for plant breeding: An update. In: J. Janick, editor, Plant Breed Reviews. John Wiley & Sons, New York. Vol. 22. p. 9–112.

Hsu, P.D., E.S. Lander, and F. Zhang. 2014. Development and applications of CRISPR-Cas9 for genome engineering. Cell 157:1262–1278.

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

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2015a. Marker imputation before genomewide selection in biparental maize populations. Plant Genome 8.

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2015b. Minimal loss of genetic diversity after genomewide selection within biparental maize populations. Crop Sci. 55:783–789.

Lee, S.H., J.H.J. van der Werf, B.J. Hayes, M.E. Goddard, and P.M. Visscher. 2008.

Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. 4:e1000231.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo. 2014. Genomewide prediction accuracy within 969 maize biparental populations. Crop Sci. 54:1514–1522.

Lopez-Perez, Eleuterio. 1979. Comparisons among five different testers for the evaluation of unselected lines of maize (*Zea mays* L.). Retrospective Theses and Dissertations. 6655.

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

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.

Lynch, M. 1988. Estimation of relatedness by DNA fingerprinting. Mol Biol Evol. 5:584–599.

Parra, J.R., and A.R. Hallauer. 1996. Utilization of exotic maize germplasm. In: J. Janick, editor, Plant Breed Reviews. John Wiley & Sons, Oxford, UK. Vol 14. p 165–187.

Phillips, R.L. 2010. Mobilizing science to break yield barriers. Crop Sci. S-99–S-108.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ran, F.A., P.D. Hsu, J. Wright, V. Agarwala, D.A. Scott, and F. Zhang. 2013. Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. 8.

Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J.-L. Jannink, and A.E. Melchinger. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503.

Riedelsheimer, C., F. Technow, and A.E. Melchinger. 2012. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13:452.

Ru, S., and R. Bernardo. 2018. Targeted recombination to increase genetic gain in self-

pollinated crops. Poster P0167 presented at: XXVI Plant & Animal Genome Conference, San Diego, CA. 13-17 Jan.

Sadhu, M.J., J.S. Bloom, L. Day, and L. Kruglyak. 2016. CRISPR-directed mitotic recombination enables genetic mapping without crosses. Science 352:1113–1116.

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.

Sokal, R.R., and C.D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38:1409–1438.