

Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years

Theresa Albrecht · Hans-Jürgen Auinger · Valentin Wimmer · Joseph O. Ogutu · Carsten Knaak · Milena Ouzunova · Hans-Peter Piepho · Chris-Carolin Schön

Received: 2 November 2013 / Accepted: 24 March 2014 / Published online: 11 April 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract

Key message The calibration data for genomic prediction should represent the full genetic spectrum of a breeding program. Data heterogeneity is minimized by connecting data sources through highly related test units.

Abstract One of the major challenges of genome-enabled prediction in plant breeding lies in the optimum design of the population employed in model training. With highly interconnected breeding cycles staggered in time the choice of data for model training is not straightforward. We used cross-validation and independent validation to assess the performance of genome-based prediction within and across genetic groups, testers, locations, and years. The study comprised data for 1,073 and 857 doubled haploid lines

evaluated as testcrosses in 2 years. Testcrosses were phenotyped for grain dry matter yield and content and genotyped with 56,110 single nucleotide polymorphism markers. Predictive abilities strongly depended on the relatedness of the doubled haploid lines from the estimation set with those on which prediction accuracy was assessed. For scenarios with strong population heterogeneity it was advantageous to perform predictions within a priori defined genetic groups until higher connectivity through related test units was achieved. Differences between group means had a strong effect on predictive abilities obtained with both cross-validation and independent validation. Predictive abilities across subsequent cycles of selection and years were only slightly reduced compared to predictive abilities obtained with cross-validation within the same year. We conclude that the optimum data set for model training in genome-enabled prediction should represent the full genetic and environmental spectrum of the respective breeding program. Data heterogeneity can be reduced by experimental designs that maximize the connectivity between data sources by common or highly related test units.

Communicated by J. Crossa.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2305-z) contains supplementary material, which is available to authorized users.

T. Albrecht · H.-J. Auinger · V. Wimmer · C.-C. Schön (✉)
Plant Breeding, Center of Life and Food Sciences Weihenstephan,
Technische Universität München, Liesel-Beckmann-Straße 2,
85354 Freising, Germany
e-mail: chris.schoen@tum.de

Present Address:

T. Albrecht
Institute for Crop Science and Plant Breeding, Bavarian State
Research Center for Agriculture, 85354 Freising, Germany

V. Wimmer · C. Knaak · M. Ouzunova
KWS SAAT AG, 37555 Einbeck, Germany

J. O. Ogutu · H.-P. Piepho
Bioinformatics Unit, Institute of Crop Science, Universität
Hohenheim, 70599 Stuttgart, Germany

Introduction

In hybrid breeding long-term genetic gain is ensured through recurrent selection. Each selection cycle comprises thousands of progenies derived from a large number of crosses and evaluated for their performance as testcrosses in multi-environment field trials. The best selection candidates in one year may directly serve as recombination units for the next selection cycle or need to be re-evaluated in the following year. In crops with short generation intervals such as maize, multiple cycles of recombination per year can be accomplished, leading to a constant output of

recombined genetic material and highly complex interconnected selection cycles staggered in time within each breeding population (Gordillo and Geiger 2008).

While in the past selection of superior genotypes was mainly based on their phenotypic performance, recent advances in molecular techniques, statistical genetics, and computational algorithms have put genome-enabled selection within reach for quantitatively inherited traits (Meuwissen et al. 2001). In maize breeding, the accuracy of predicting phenotypes from DNA profiles (genomic prediction, GP) has recently been investigated for breeding populations (Albrecht et al. 2011; Crossa et al. 2010), bi- and multiparental populations (Guo et al. 2012; Schulz-Streeck et al. 2012; Riedelsheimer et al. 2013), and diversity panels (Rincent et al. 2012; Riedelsheimer et al. 2012). Estimates of prediction accuracies obtained in these studies were encouraging with respect to the usefulness of GP in practical breeding programs and most of the studies have come to similar conclusions with respect to the factors determining prediction accuracy. The statistical method employed in GP should be chosen wisely depending on the trait under study (Wimmer et al. 2013). For truly quantitative traits, several statistical methods have been shown to perform well over a wide range of populations and experimental settings (Heslot et al. 2012; Pérez-Rodríguez et al. 2012; Lehermeier et al. 2013; Gianola 2013). Major determinants of prediction accuracy in most studies were the sample size and genetic constitution of the population used for model training. As pointed out by Windhausen et al. (2012) and Guo et al. (2014), predictive ability can be highly affected by population structure, when the calibration set comprises genetic groups with significantly different mean performance. The average relatedness of the individuals employed in model training with those individuals on which prediction accuracy is assessed also has a strong effect (Habier et al. 2010). With cross-validation (CV), Albrecht et al. (2011) observed a significant decrease in genomic predictive ability for maize testcross performance when doubled haploid (DH) lines belonging to the same biparental family were mutually excluded from the estimation and test sets.

The body of data available for training of a genome-based prediction model in a hybrid breeding scheme is complex. In plant breeding populations, genetic diversity is maintained by frequent introgressions of genetically diverse, unrelated germplasm which can lead to extensive genetic substructure in the selection cycle in which the material is introduced. The choice of tester may vary over selection cycles and may be attuned to the maturity of the selection candidates such that an early maturing tester is crossed to a group of late maturing selection candidates and vice versa. Model training can be performed with data collected in more than one year to minimize the effect of

genotype \times year interactions. However, by combining data from several selection cycles the average relatedness of the calibration set with the selection candidates is decreased and it is difficult to assess how prediction accuracies will be affected.

Thus, when implementing GP in commercial plant breeding programs the choice of data included in model training is one of the main challenges. Experimental studies based on diversity panels and large bi- or multiparental populations give first insights into the efficiency of GP, but they can only partly reflect the complex structure of a fast cycle hybrid breeding scheme. Therefore, the aim of our study was to assess predictive abilities of GP with genetic material derived from a commercial maize breeding program. The data used here consist of two calibration sets for GP with 1,073 and 857 DH lines exhibiting different degrees of genetic substructure. All lines were evaluated as testcrosses for two quantitative traits, grain dry matter yield and grain dry matter content. DH lines were genotyped with a high-density marker array comprising 56,110 single nucleotide polymorphism markers (SNPs). Our objectives were to (1) assess the relative efficiency of GP over pedigree-based prediction with different calibration sets using cross-validation accounting for genetic substructure, (2) compare predictive abilities across locations with predictive abilities across years, and (3) evaluate the efficiency of prediction across genetic groups and testers.

Materials and methods

Plant material

The experimental data in this study consist of two sets of genetic material from the maize dent pool comprising 1,073 and 857 DH lines. The two sets of material represent two selection cycles of the same maize breeding program and form two calibration sets for GP (Fig. 1). In calibration set 1 (CS1), the 1,073 DH lines were derived from 192 crosses among 55 parents (43 inbred lines and 12 single-crosses). In calibration set 2 (CS2), the 857 DH lines were obtained from 294 crosses among 75 parents (55 inbred lines and 20 single-crosses). The two calibration sets were connected by 25 parents (Supplementary Figure S1a). The number of DH lines derived from each cross was highly variable, ranging from 1 to 63 DH lines with an average of 6 DH lines in CS1 and 1 to 26 DH lines and an average of 3 DH lines in CS2. The number of progeny per parent ranged from 1 to 203 with an average of 39 DH lines in CS1 and from 1 to 130 with an average of 23 DH lines in CS2. Pedigree records included a minimum of three generations with 479 ancestors representing germplasm of different breeding populations (Supplementary Figure S2). DH

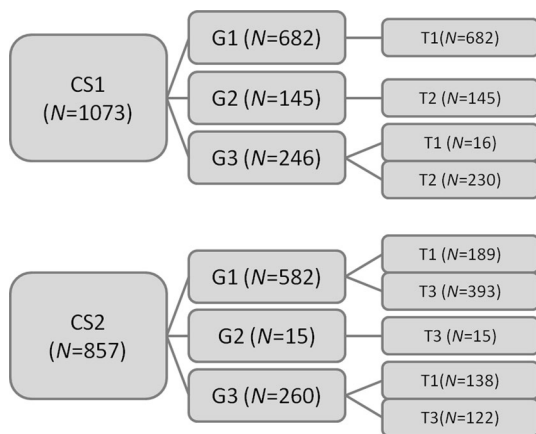


Fig. 1 Data structure and number of DH lines (N) in calibration sets CS1 and CS2 allocated to genetic groups G1, G2, and G3 and crossed to testers T1, T2, and T3

lines were assigned to one of three genetic groups depending on whether they were derived from crosses between parents originating from the current dent gene pool (G1) or from crosses of lines from the current gene pool with dent germplasm from two different unrelated sources (G2, G3). Supplementary Figures S1 and S3 show the connection between genetic groups G1, G2, and G3 based on parental lines used in the crosses from which DH lines were derived. All materials described in this study are proprietary to KWS SAAT AG.

Genotypic data analysis

Both calibration sets were genotyped with 56,110 SNP markers using the Illumina[®] MaizeSNP50 BeadChip (Ganal et al. 2011). Only high-quality SNPs with a GenTrain score ≥ 0.7 and a call frequency ≥ 0.9 were used for further analysis. We discarded SNPs with a minor allele frequency (MAF) < 0.01 and redundant SNPs, resulting in 15,732 and 16,846 useful, polymorphic SNPs for CS1 and CS2, respectively. For the joint analysis of both calibration sets ($N = 1,930$ DH lines), the same SNP selection steps were performed resulting in 17,734 polymorphic SNPs. Marker genotypes were consistently coded as the number of copies of the minor allele, i.e., 0 or 2. Missing values were reconstructed using the function ‘codeGeno’ from the R package ‘synbreed’ with the option ‘beagleAfterFamily’ (Wimmer et al. 2012).

Field experiments and phenotypic analysis

Phenotypic performance of the DH lines of CS1 was evaluated from testcrosses with a single-cross (T1) or double-cross (T2) tester in 2010. DH lines of CS2 were evaluated as testcrosses with either of two single-cross testers (T1 or

T3) in 2011. The three testers were derived from six parental inbred lines (A, B, C, D, E, F) from the opposite heterotic flint pool, with T1 being a cross of inbred lines A \times B, T2 of lines C.D \times B.E, and T3 of lines B \times F. In CS1, four group/tester combinations were evaluated and in CS2 five combinations. Each DH line was crossed to only one of the three testers and each of the 192 crosses in CS1 and 294 crosses in CS2 was assigned to only one of the nine possible group/tester combinations of the two calibration sets. A summary of the distribution of DH lines across groups and testers in each calibration set is given in Fig. 1.

Field trials were conducted in six German locations in 2010 and 2011, with four locations represented in both years. In 2010, entries were distributed across 16 trials. Each trial was conducted in four of the six locations resulting in 6–16 trials per location. In 2011, entries were distributed across 12 trials and each trial was conducted in three or four of the six locations, resulting in 4–12 trials per location. Each trial was laid out in a 10×10 lattice design with two replications containing 94 entries and six hybrid checks in 2010 and 95 entries and five hybrid checks in 2011. Entries comprised the 1,930 genotyped DH lines and 714 additional DH lines for which no genotypic data were available.

Measured phenotypic traits were grain dry matter yield (GDY, dt/ha) and grain dry matter content (GDC, %). Outliers were identified and removed based on maximum deviant residuals according to Grubbs (1950). Adjusted entry means across locations within years were calculated in a two-stage approach according to Piepho et al. (2012). Variance components were estimated separately for genotyped and non-genotyped entries and their corresponding genotype \times location interactions (Piepho et al. 2006) using restricted maximum likelihood with the software ASReml 3.0 (Gilmour et al. 2009). Significance of variance components was tested according to Stram and Lee (1994). Trait heritabilities were approximated based on genotyped DH lines for both traits and both years according to Holland et al. (2003).

Prediction models

To predict testcross performance of DH lines, we used two mixed models differing in the definition of the variance–covariance structure for random testcross effects. For the first model, pedigree-based estimates of kinship coefficients were used to model the variance–covariance structure of testcross effects. The model is denoted as PBLUP and is described as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{t} + \mathbf{e},$$

where \mathbf{y} is a $N \times 1$ vector of adjusted means for N DH lines obtained from the phenotypic analysis; $\boldsymbol{\beta}$ is a $c \times 1$ vector of fixed effects with $c = 4$ (CS1) and $c = 5$ (CS2)

factor levels for group/tester combinations (Fig. 1). The fixed effect was included to correct for genetic substructure within calibration sets. Random testcross effects were modeled with the $N \times 1$ vector $\mathbf{t} \sim N(\mathbf{0}, \mathbf{K}\sigma_t^2)$, where \mathbf{K} is a $N \times N$ matrix of expected kinship coefficients derived from pedigree information estimated with the ‘kin’ function of the R package ‘synbreed’ (Wimmer et al. 2012) and σ_t^2 is the testcross variance pertaining to the PBLUP model. The design matrices \mathbf{X} and \mathbf{Z} assign the adjusted means to the fixed and random effects, respectively. The $N \times 1$ residual vector \mathbf{e} is assumed to be independent and normally distributed with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_{ep}^2)$, where \mathbf{I} is an identity matrix and σ_{ep}^2 is the residual variance.

In the second prediction model, the variance–covariance structure of testcross effects was derived from genome-wide marker data. The model is denoted as GBLUP and can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where the vectors \mathbf{y} , $\boldsymbol{\beta}$ and the design matrices \mathbf{X} and \mathbf{Z} are defined as in the PBLUP model and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_{eG}^2)$, with σ_{eG}^2 being the residual variance pertaining to the GBLUP model. The random testcross effects \mathbf{u} are assumed to be normally distributed with $\mathbf{u} \sim N(\mathbf{0}, \mathbf{U}\sigma_u^2)$, where \mathbf{U} is a $N \times N$ matrix of realized kinship coefficients based on the marker data and σ_u^2 is the testcross variance pertaining to the GBLUP model. We calculated \mathbf{U} according to Astle and Balding (2009) as identical by descent (IBD) estimates between pairs of individuals from genome-wide SNP data. The elements of \mathbf{U} were estimated as

$$U_{ij} = \frac{1}{2M} \sum_{m=1}^M \frac{(w_{im} - 2p_m)(w_{jm} - 2p_m)}{2p_m(1 - p_m)},$$

where w_{im} and w_{jm} are genotype scores of individuals i and j at marker locus m , p_m is the MAF of marker m in the population under study, and M the number of marker loci.

Analysis of genetic substructure

In CS2, we analyzed the relatedness of DH lines within and between the four group/tester combinations G1/T1, G1/T3, G3/T1, and G3/T3 based on pedigree information. Given the a priori definition of substructure in a data set, the maximum kinship between a DH line i and all other lines from the same group/tester combination ($k_{\max,i}$ within) should be significantly higher than the maximum kinship between the DH line and all lines from the other group/tester combinations ($k_{\max,i}$ between) (Saatchi et al. 2011). Mean kinship coefficients (\bar{k}) were calculated as average of pairwise kinship coefficients between DH lines from the same or different group/tester combinations.

Cross-validation

We used ten times replicated fivefold cross-validation (CV) as described in Albrecht et al. (2011) to assess the predictive abilities of PBLUP and GBLUP. Model parameters were estimated in estimation sets (ES) sampled according to procedures described below (CV1, CV2, and CV3). Predictive abilities were estimated as Pearson’s correlation coefficient r_{TS} between the predicted and the observed testcross values of DH lines in test sets (TS).

In CV1, we employed random genotypic sampling of ES and TS for assessing the predictive abilities within calibration sets, within genetic groups, or within group/tester combinations (Supplementary Figure S4a). The predictive ability for each TS was calculated based on vectors $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{t}}$ and $\hat{\mathbf{u}}$ estimated from the corresponding ES as $r_{TS} = r(\mathbf{y}_{TS} - \mathbf{X}_{TS}\hat{\boldsymbol{\beta}}, \mathbf{Z}_{TS}\hat{\mathbf{t}})$ for PBLUP and $r_{TS} = r(\mathbf{y}_{TS} - \mathbf{X}_{TS}\hat{\boldsymbol{\beta}}, \mathbf{Z}_{TS}\hat{\mathbf{u}})$ for GBLUP. Here, the vector \mathbf{y}_{TS} is a $N_{TS} \times 1$ vector of observed phenotypes in the TS and \mathbf{X}_{TS} and \mathbf{Z}_{TS} are $N_{TS} \times c$ and $N_{TS} \times N$ design matrices for fixed and random effects, respectively.

In CV2, predictive abilities were assessed across genetic groups in CS1 and across group/tester combinations in CS2 conditional on a given TS (Supplementary Figure S4b). Here, the same TS as sampled for CV1 were used. However, the ES were sampled from DH lines not belonging to the same genetic group or group/tester combination as the TS or from all remaining DH lines belonging to the same calibration set (ES_{All}). For CS1, sampling conditional on a TS from G3 ($N_{TS} = 49$) is illustrated in Supplementary Figure S5. Here, ES_{All} comprised the remaining 1024 DH lines from CS1, while ES_{G1G2} comprised only DH lines from the groups G1 and G2 ($N_{ES} = 827$). When different group/tester combinations occurred in the ES and TS, a correction for the fixed group/tester effect was not possible and predictive abilities were calculated based on the unadjusted observed testcross performance \mathbf{y}_{TS} .

In CV3, the predictive ability of GBLUP across locations was assessed to analyze the influence of genotype \times location interactions. For this procedure, only the subset of DH lines belonging to genetic group G1 in CS1 ($N = 682$) was used. The data set was divided into five genotypic subsets each evaluated in the same four locations resulting in 20 disconnected subsets (Supplementary Figure S6). The ES comprised adjusted means from three locations and four genotypic subsets. Following Utz et al. (2000), three different TS were defined for each ES taking genotypic sampling (TS_g), sampling of locations (TS_{loc}) and both factors simultaneously (TS_{g \times loc}) into account.

Validation across years

To investigate predictive abilities of GBLUP across years, data from both calibration sets were analyzed jointly ($N = 1,930$) with a realized kinship matrix U calculated from $M = 17,734$ SNPs. The vector of fixed effects β in the joint model included $c = 9$ factor levels for each year/group/tester combination. To predict testcross performance of DH lines from CS2, phenotypic observations from CS2 were masked and predicted from data in CS1 and vice versa. Predictive abilities were measured as the correlation of predicted testcross values and observed testcross performance across all DH lines in each calibration set and for DH lines within each group/tester combination separately. Differences across group/tester subsets were visualized with elliptic contours. The shape of each ellipse represents the 95 % confidence region of a bivariate normal distribution with mean and variance–covariance structure corresponding to the mean and variance–covariance matrix of the predicted and observed testcross values within each group/tester combination.

Results

Phenotypic analysis

Testcross means for GDY and GDC differed significantly ($p < 0.01$) between calibration sets, genetic groups, and group/tester combinations (Supplementary Table S1). In 2010 (CS1), adjusted means for GDY ranged from 95.13 to 148.20 dt/ha with a mean of 126.71 dt/ha. Climatic conditions in 2011 were more favorable for maize production than in 2010 and GDY and GDC in CS2 were significantly ($p < 0.01$) higher than in CS1 with a range of 108.10 to 165.30 dt/ha and a mean of 144.31 dt/ha for GDY and a mean of 71.81 % for GDC. Mean GDY in G2 was substantially reduced compared to G1 and G3 in both calibration sets. In CS2, significant differences ($p < 0.01$) in the variance and mean performance were observed between DH lines crossed to testers T1 and T3 for both traits (Supplementary Figure S7).

In both calibration sets, genotypic and genotype \times location interaction variances were highly significant ($p < 0.01$) for both traits. Trait heritabilities on a progeny mean basis were $\hat{h}_{\text{GDY}}^2 = 0.72$ and $\hat{h}_{\text{GDC}}^2 = 0.94$ for CS1 and $\hat{h}_{\text{GDY}}^2 = 0.71$ and $\hat{h}_{\text{GDC}}^2 = 0.95$ for CS2.

Prediction within calibration sets and group/tester combinations

In both calibration sets, GBLUP outperformed PBLUP for both traits (Table 1). Consistent with expectation, absolute

predictive abilities of GBLUP and the relative advantage of genome-enabled over pedigree-based prediction were reduced in CS2 compared to CS1 due to the DH lines being derived from a larger number of crosses derived from more parents and the lower average number of progeny per cross. In CS1, mean GBLUP predictive abilities were $r_{\text{TS}} = 0.59$ for GDY in the complete set of lines and ranged between $r_{\text{TS}} = 0.40$ for the smallest group G2 ($N_{\text{G2}} = 145$) and $r_{\text{TS}} = 0.65$ for G1, the group with the largest sample size ($N_{\text{G1}} = 682$). Although the sample size of the ES was increased when the complete set of lines was analyzed ($N = 1,073$), there was no gain in predictive ability over lines from G1 for GDY. For GDC, however, the mean predictive ability was $r_{\text{TS}} = 0.87$ for CS1 and higher than within individual groups. Although G2 and G3 were small ($N_{\text{G2}} = 145$ and $N_{\text{G3}} = 246$), mean predictive abilities for GDC were still high ($r_{\text{TS}} = 0.77$ and 0.78 , respectively).

For CS2, the mean GBLUP predictive ability for GDY was $r_{\text{TS}} = 0.49$ for the complete set of lines and ranged between $r_{\text{TS}} = 0.30$ for G3/T3 ($N_{\text{G3/T3}} = 122$) to $r_{\text{TS}} = 0.59$ for G1/T3 ($N_{\text{G1/T3}} = 393$). Predictive abilities within G1/T3 were higher than in the complete CS2 ($N = 857$). For GDC, the mean predictive ability was $r_{\text{TS}} = 0.86$ in CS2 and decreased when predictions were performed within individual group/tester combinations.

Prediction across group/tester combinations

Results for predictions across genetic groups and group/tester combinations obtained with GBLUP and CV2 are illustrated in Figs. 2 and 3 and Supplementary Figure S8.

In CS1, tester T1 was crossed mainly to DH lines from group G1 whereas T2 was crossed only to lines of groups G2 and G3 (Fig. 1). Thus, the effects of genetic substructure and tester cannot be separated within CS1. For GDY, predictive abilities in group G1 slightly decreased when the ES was augmented with DH lines from G2 and G3 even though the sample size of the ES was almost doubled (Fig. 2a). The same effect was observed for group G3 even though the increase in sample size of the ES was more dramatic than for group G1. Only for G2, there was an increase in predictive ability from $r_{\text{TS}} = 0.40$ to 0.44 when the ES included also lines from groups G1 and G3. In this case, the increase in sample size was almost tenfold, with $N_{\text{ES}} = 1,044$ for ES_{All} compared to $N_{\text{ES}} = 116$ for ES_{G2} . To account for the effect of sample size, we predicted testcross values of DH lines from G2 with 116 randomly sampled DH lines from the 1,044 DH lines of ES_{All} . This led to a substantial decrease of average predictive abilities from $r_{\text{TS}} = 0.44$ to 0.23 (results not shown). For all groups in CS1, predictive abilities decreased substantially ($r_{\text{TS}} = 0.26$ – 0.30) when the ES consisted only of DH lines from genetic groups that were not included in the TS. It is

Table 1 Predictive ability of PBLUP and GBLUP with CV1 for grain dry matter yield (GDY) and grain dry matter content (GDC) for calibration sets CS1 and CS2 as well as for genetic groups (G1, G2, and G3) of CS1 and for group/tester combinations (G1/T1, G1/T3, G3/T1, and G3/T3) of CS2

Trait Model	CS1	Group		
		G1	G2	G3
GDY				
PBLUP	0.383 ± 0.004 ^a	0.396 ± 0.011	0.360 ± 0.037	0.259 ± 0.027
GBLUP	0.594 ± 0.006	0.650 ± 0.007	0.402 ± 0.035	0.432 ± 0.034
GDC				
PBLUP	0.597 ± 0.004	0.527 ± 0.008	0.586 ± 0.017	0.510 ± 0.016
GBLUP	0.872 ± 0.003	0.840 ± 0.005	0.766 ± 0.021	0.776 ± 0.013
	CS2	Group/tester		
		G1/T1	G1/T3	G3/T1
GDY				G3/T3
PBLUP	0.400 ± 0.011 ^a	0.405 ± 0.022	0.448 ± 0.013	0.140 ± 0.046
GBLUP	0.489 ± 0.012	0.442 ± 0.029	0.592 ± 0.009	0.331 ± 0.048
GDC				
PBLUP	0.594 ± 0.005	0.591 ± 0.026	0.544 ± 0.013	0.597 ± 0.025
GBLUP	0.861 ± 0.004	0.724 ± 0.029	0.832 ± 0.007	0.804 ± 0.016

^a ± standard deviation

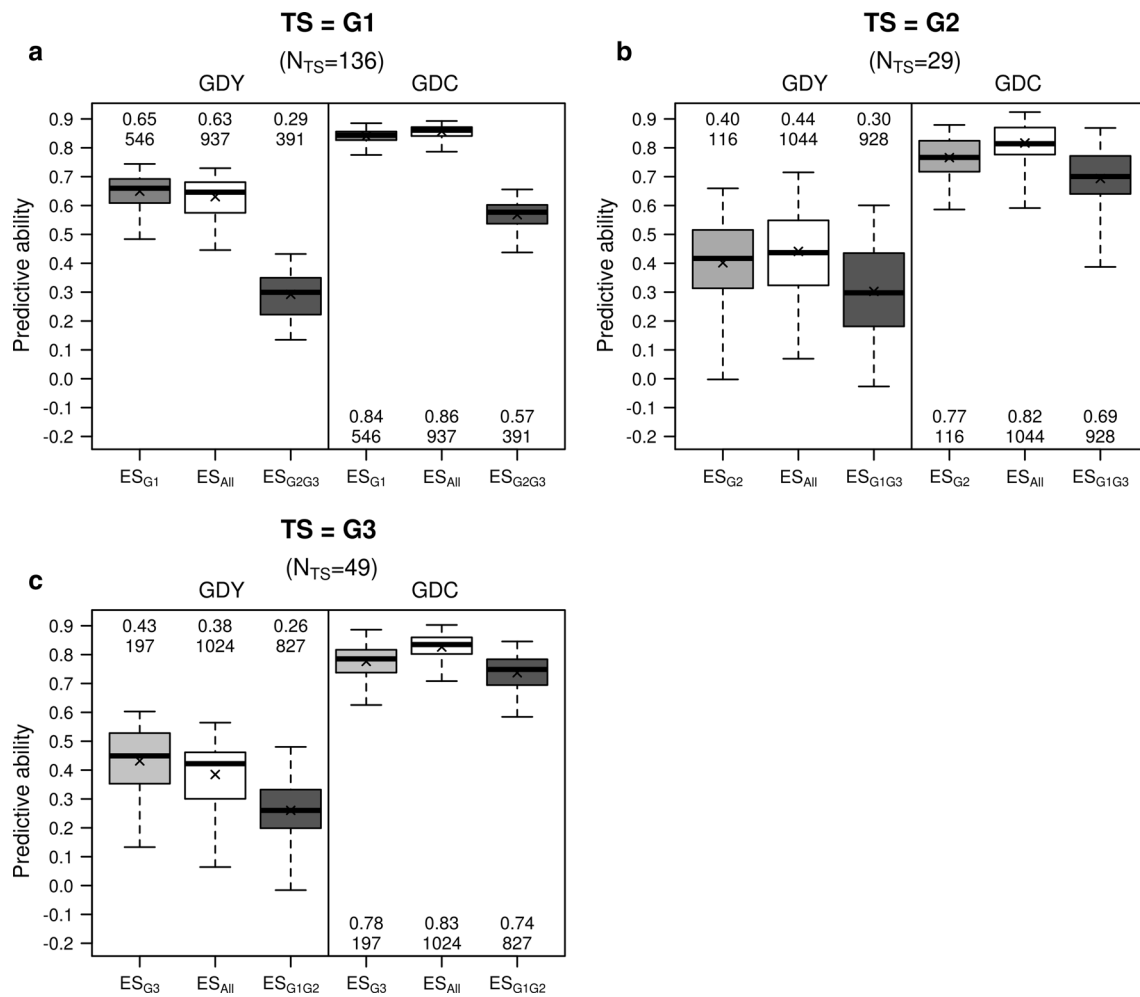


Fig. 2 Predictive ability within and across genetic groups obtained with GBLUP and CV1 and CV2, respectively. Specific estimation sets (ES) were sampled for test sets (TS) of genetic groups **a** G1, **b** G2, and **c** G3 in calibration set CS1. These specific ES comprised DH lines from the same group, all groups (ES_{All}), or different groups

(ES_{G2G3}, ES_{G1G3}, ES_{G1G2}). Boxplots show the range, median (bar) and mean (×) of 50 CV runs for grain dry matter yield (GDY) and grain dry matter content (GDC). Numbers above or below boxplots indicate average predictive abilities and size of the estimation sets, respectively

noteworthy, however, that the decrease in predictive abilities in CS1 might have been slightly overestimated because of the different testers used for production of testcrosses in groups G1, G2, and G3.

For GDC, augmenting the ES with lines from the other genetic groups led to higher predictive abilities for all groups. Predictions of genetic values of DH lines from groups G2 and G3 even performed well ($r_{TS} = 0.69$ and 0.74 , respectively) when the ES comprised only DH lines from the other genetic groups. For the TS of G1, predictive abilities decreased markedly ($r_{TS} = 0.57$) when only the groups G2 and G3 were included in the ES ($N_{ES} = 391$).

For CS2, it was possible to evaluate the effects of genetic groups and testers separately, because DH lines from both groups were crossed to both testers. When the

sample size of the ES was kept constant at $N_{ES} = 98$, predictive abilities obtained with CV for both traits were significantly ($p < 0.01$) correlated with the parameter of relatedness \bar{k}_{max} calculated for DH lines within and between group/tester combinations (Fig. 3). Correlations between the mean pairwise kinship parameter \bar{k} and predictive ability were only significant ($p < 0.01$) for GDY.

When the sample size of the ES was not fixed, predictions with genetic group G1 crossed to tester T3 (ES_{G1/T3}) outperformed predictions within G3/T3 probably due to the much larger sample size of the ES_{G1/T3} (Supplementary Figure S8d). It was surprising though that the prediction across groups (e.g., G1/T3 predicted with G3/T3) performed better than the prediction across testers (e.g., G1/T3 predicted with G1/T1) in all cases (Supplementary Figure S8).

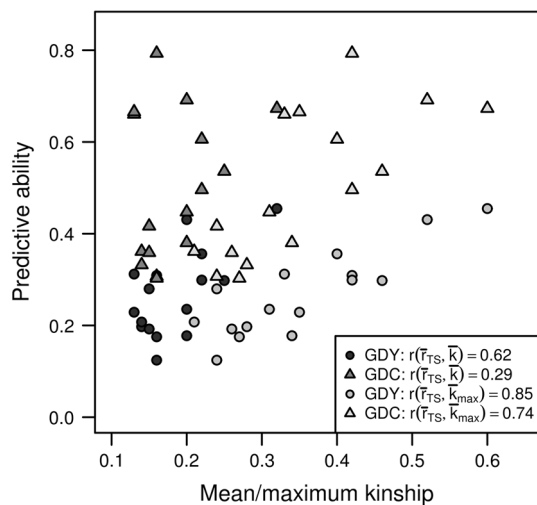


Fig. 3 Mean predictive ability (\bar{r}_{TS}) of GBLUP for traits GDY and GDC obtained with CV1 and CV2 for the prediction within and across group/tester combinations (G1/T1, G1/T3, G3/T1, and G3/T3) in calibration set CS2 plotted against mean (\bar{k}) and maximum kinship (\bar{k}_{max}) within and across group/tester combinations. The sample size of the estimation sets was fixed to $N_{ES} = 98$ for each of the 16 possible combinations. The correlations (r) between the predictive ability and mean and maximum kinship are given in the legend

Prediction across locations and years

Results of the prediction across locations obtained with CV3 within the subset of group G1 of CS1 are given in Table 2. When predicting GDY testcross performance in an independent location (TS_{loc}), mean predictive abilities were markedly reduced compared to predicting independent genotypes (TS_g) except for location 1. Reducing the number of locations in the ES from four to three decreased the predictive ability across genotypes depending on how much the estimated heritability in the ES was reduced. Predictive abilities for GDC were not affected as strongly by the reduced number of locations in the ES as were those for GDY. Moreover, prediction across locations generally performed well for GDC with a minimum predictive ability of $r_{TS} = 0.78$ for location 1.

Table 2 Heritability of the estimation set (h_{ES}^2) and predictive ability obtained for three different test sets (TS_g , TS_{loc} , and $TS_{g \times loc}$) of CV3 for grain dry matter yield (GDY) and grain dry matter content (GDC) for four possible combinations of estimation and test sets of the subset G1 in calibration set CS1

Trait	Location in TS	h_{ES}^2	Predictive ability \pm standard deviation		
			TS_g	TS_{loc}	$TS_{g \times loc}$
GDY	1	0.64	0.581 \pm 0.008	0.620 \pm 0.001	0.534 \pm 0.008
	2	0.64	0.618 \pm 0.007	0.586 \pm 0.001	0.477 \pm 0.008
	3	0.69	0.667 \pm 0.008	0.427 \pm 0.001	0.326 \pm 0.010
	4	0.71	0.641 \pm 0.008	0.448 \pm 0.001	0.362 \pm 0.009
GDC	1	0.93	0.840 \pm 0.005	0.782 \pm 0.000	0.681 \pm 0.005
	2	0.90	0.833 \pm 0.006	0.865 \pm 0.000	0.754 \pm 0.006
	3	0.91	0.838 \pm 0.006	0.860 \pm 0.000	0.748 \pm 0.005
	4	0.91	0.819 \pm 0.005	0.853 \pm 0.000	0.770 \pm 0.007

For GDY, Fig. 4 shows correlations between the observed testcross performance of DH lines in one year and their genome-based predicted values derived from model training with the entire calibration set of the other year. When calculated across group/tester combinations, predictive abilities were higher in CS1 ($r_{CS1} = 0.58$) than in CS2 ($r_{CS2} = 0.38$) and ranged between 0.30 and 0.44 within group/tester combinations except for DH lines from group G2 that could not be predicted accurately ($r_{TS} = 0.28$ in CS1 and -0.15 in CS2, latter not shown). For GDC, predictive abilities across years ranged between 0.50 and 0.74 within group/tester combinations (Supplementary Figure S9). Predictions across years for both traits were in a similar range as predictions obtained by CV3 when sampling independent genotypes (Table 2, $TS_{g \times loc}$).

Discussion

When implementing new technologies in commercial breeding programs it is crucial to optimize the allocation of resources between new and established breeding procedures. However, successful optimization of breeding scenarios requires a realistic framework of assumptions based on population and trait-specific parameters derived from experimental data. In this study, we provide estimates of prediction accuracies of GP specific to hybrid maize breeding populations accounting for the effects of genetic heterogeneity of the breeding population and genotype \times environment interactions.

Optimizing populations for model training

The two calibration sets employed here represent two successive cycles of selection of an advanced cycle maize breeding program. In both calibration sets, the DH lines assigned to genetic group G1 form the core set of germplasm representing the breeding population. However, in a breeding scheme based on doubled haploid technology the genetic variance of the breeding population gets

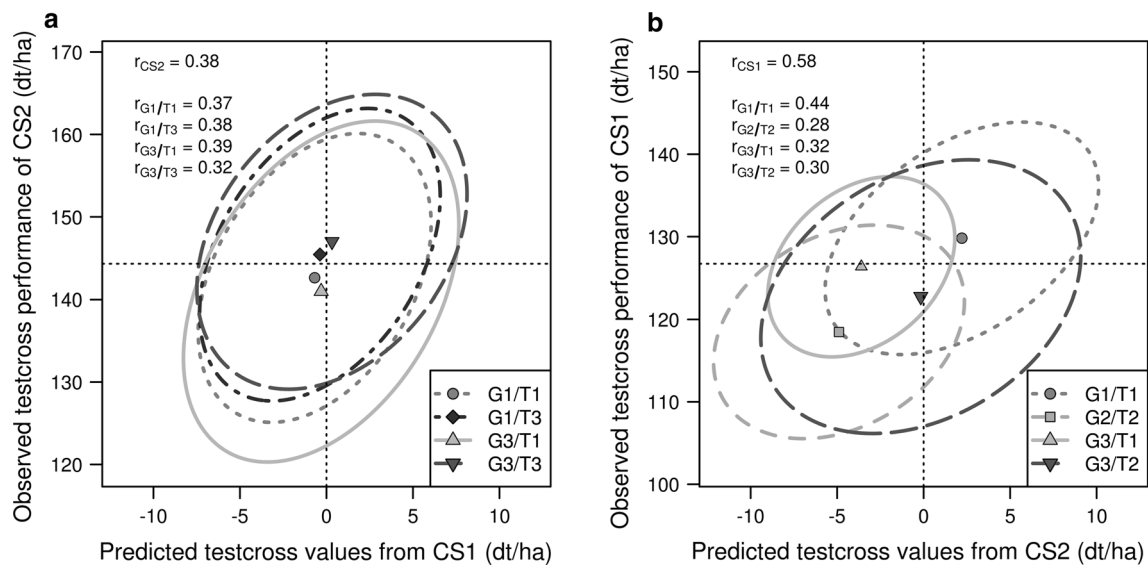


Fig. 4 Observed testcross performance in calibration sets **a** CS2 and **b** CS1 plotted against predicted testcross values derived with GBLUP from calibration set **a** CS1 and **b** CS2 for grain dry matter yield visualized as elliptical contours representing 95 % confidence intervals of the respective bivariate normal distributions for four group/tester

combinations of CS1 and CS2, respectively. The respective groups are indicated by the *symbol* in the center of the ellipse and different types of *lines*. Correlations (r) are given for each calibration set and within group/tester combinations

depleted rapidly (Gordillo and Geiger 2008). To maintain sufficient genetic variance and thus secure long-term selection gain the genetic diversity of the breeding population must be broadened by introgression of unrelated genetic material. In GP, Hayes et al. (2009) argued that predictions should be more persistent over generations when multiple groups are included in the reference population. Furthermore, Rincent et al. (2012) recommended including genetic material derived from a diverse set of parents for model training to maintain high genetic variation within the estimation set. On the other hand, Habier et al. (2010) showed that for complex traits like milk yield and somatic cell score in cattle, prediction accuracy in GP was highly dependent on the number of closely related individuals in the estimation and the test set. Thus, a fundamental question in GP concerns how the accuracy of genome-enabled prediction is affected when progeny derived from crosses with unrelated material are used for model training.

When fixing the sample size of the ES in CV, there was a close association between the pedigree-based kinship of the DH lines assigned to the ES and TS (\bar{k}_{\max}) and the predictive ability of GBLUP for GDY and GDC (Fig. 3). However, when new genetic material is integrated into an existing breeding population, there is generally a strong imbalance in the number of lines derived from adapted and new genetic material and inferences on the association of the degree of relatedness and predictive ability are not as straightforward.

In both calibration sets, G1 represented the largest of the three genetic groups with the highest performance level. Even though the sample size of the estimation set was substantially increased by adding progeny from crosses with unrelated material, predictive abilities for GDY were reduced for G1 in CS1 representing the core set of lines. Only when the sample size of the ES was small, as was the case for G2 in CS1, predictive abilities slightly improved by adding progenies from crosses with unrelated material. Our results corroborate the findings of Riedelsheimer et al. (2013) for multiparental crosses of maize that prediction accuracy may not improve or may even decrease when genetic material from an unrelated family is included in the estimation set. Thus, when the genetic diversity of a breeding population is expanded by introducing new unrelated genetic material, it may be advisable to perform predictions for GDY separately within a priori defined genetic groups, in particular if the level of connectivity between groups is low. We conclude that (1) the optimum data set for model training in GP should encompass the full genetic spectrum of a breeding population and that (2) data heterogeneity should be minimized by experimental designs that maximize the connectivity between data sources through common and/or highly related test units.

Cross-validation and validation

In most plant breeding studies, predictive abilities obtained with GP under different experimental scenarios and

statistical methods have been assessed by resampling methods such as cross-validation. However, predictive abilities derived from CV are known to be optimistically biased if subgroups exist that differ significantly in their mean performance and this difference is not accounted for in the statistical model (Guo et al. 2014). This effect was also demonstrated by Windhausen et al. (2012) for prediction of the performance of 255 maize hybrids derived from eight breeding populations. Predictive abilities obtained with random sampling in CV were biased by the differences in population means and could not be confirmed by independent validation. In this study, we also observed this phenomenon. When the population substructure introduced by the three genetic groups was not accounted for by fitting fixed group effects in the prediction model, predictive abilities in CV increased from $r_{TS} = 0.59$ to 0.73 for GDY in CS1 (data not shown), because testcross effects captured the variation across groups in addition to within groups.

Thus far, only few studies have compared predictive abilities obtained from CV with those realized in independent validation. Hofheinz et al. (2012) pointed out that CV with genetic material from the same selection cycle can overestimate the potential of GP for traits with medium heritability compared to validation with recombined genetic material evaluated in a different year. Similar results were also reported by Utz et al. (2000) for marker-assisted selection of complex traits where CV also overestimated predictive abilities relative to independent validation. With heritabilities in CS1 and CS2 being almost identical, we could compare predictive abilities from cross-validation in CS1 with validation in CS2 (Fig. 4a and Supplementary Figure S9a) and cross-validation in CS2 with validation in CS1 (Fig. 4b and Supplementary Figure S9b). As expected, we observed a decrease in predictive ability for validation compared to CV for both traits. However, estimates of the correlation between the predicted and observed genetic values can be inflated or deflated in validation just as much as in CV. While the overall correlation of observed and predicted testcross values of DH lines in CS1 was overestimated for GDY ($r_{CS1} = 0.58$, Fig. 4b) due to group mean differences in the predicted and the observed values, it was substantially underestimated for GDC in CS2 ($r_{CS2} = 0.51$, Supplementary Figure S9a) due to large differences in means between testers within groups for the observed but not for the predicted values.

We conclude that accounting for genetic heterogeneity in the estimation and the test sets is crucial for both CV and validation but it is a non-trivial task. Mean centering of input data or modeling of group means as fixed effects can only be performed when distinct groups are known a priori. However, when breeding cycles are highly interconnected, pedigree-based definition of genetic groups becomes difficult. Marker data can support the identification of

substructure within the population employed in model training (Supplemental Figure S2, Guo et al. 2014). But how to efficiently correct for the effect of (hidden) population heterogeneity in genome-based prediction across selection cycles remains a problem that warrants further research.

Prediction across environments

In most breeding programs, the large majority of selection candidates are tested in one year and only candidates that perform exceedingly well are re-tested in subsequent years. To compensate for the lack of information on performance in multiple years selection candidates are grown in a suite of locations chosen to represent target environments and to optimally capture genotype \times environment interactions. We evaluated if predictive abilities obtained in across year predictions can be anticipated with predictions across locations. Predictive abilities obtained with CV3 can account for genotypic and environmental sampling ($TS_{g \times loc}$), but not for the advance in selection cycle from CS1 to CS2. Nevertheless, predictions with CV3 in $TS_{g \times loc}$ ranged from $r_{TS} = 0.33$ to 0.53 for GDY (Table 2) and were quite similar to results from validation across years. The same was true for GDC and thus we conclude that the effect of across year prediction was captured quite well by CV3. To obtain a more complete picture on the effects of genotype \times environment interactions on predictive abilities and the usefulness of accounting for them in GP will require more than two calibration sets. For example, predictions across years might benefit if data sets tested in different years are connected by common lines, or if the effects of genotype \times environment interactions are modeled as suggested by Burgueño et al. (2012).

Prediction across testers

When the phenotypic evaluation of DH lines is performed with an inbred or single-cross tester, the effects of general and specific combining ability cannot be separated in GBLUP (Albrecht et al. 2011). However, if heterotic groups were established a long time ago as in hybrid maize breeding, then the relative importance of specific relative to general combining ability is expected to be small (Reif et al. 2007) and genetic correlations between DH lines crossed to different testers should be medium to high. In CS2, predictive abilities within and across groups were comparable for the two testers even though testcrosses with tester T1 and T3 differed significantly in their means and variances for both traits (Supplementary Table S1 and Figure S7). We could not estimate genetic correlations between testers because each DH line was only crossed to one of the two testers. If we assume a lower bound of $r_g = 0.60$ for

the genetic correlation between two testers based on results from the literature (e.g., Melchinger et al. 1998), predictive abilities obtained within the same group but with a different tester were disappointingly low in this study, especially because the two flint single-cross testers T1 and T3 shared a common parent. Even with non-random assignment of DH lines to the two testers it is not obvious why predictive abilities for the within group/across tester comparison were consistently lower than those for the across group/within tester comparison for both traits. This remained true even when the sample size of the ES was fixed at $N_{ES} = 98$ (data not shown). For this data set, it cannot be determined whether the low predictive abilities across testers are due to the non-random assignment of DH lines to the tester subsets or the effects of specific combining ability. Thus, in the context of GP the choice of tester and the magnitude of specific combining ability effects merit further attention.

Conclusions

The integration of genome-enabled prediction into commercial plant breeding schemes has to face many challenges and is not as straightforward as reported for cattle populations (Jonas and de Koning 2013). Our results show that realistic estimates for prediction accuracies can only be obtained by adequately accounting for the effects of genetic heterogeneity and genotype \times environment interactions. The magnitude of correlations between the predicted and observed testcross performance for grain yield and maturity in validation across years was encouraging and argues for implementing GP in commercial breeding programs considering that the correlations were calculated for subsequent cycles of selection, different years, and different testers. We conclude that the accuracy of GP may be increased by optimizing experimental designs so that groups distinguished by genetic substructure, testers, or environmental effects are connected by common or highly related test units. Thus, to more effectively implement GP, it will be necessary to revisit the current breeding schemes with a view to re-allocating resources based on trait and population parameter estimates obtained from large experimental studies such as the one presented here.

Acknowledgments We thank Daniel Gianola, Sofia daSilva, and Torben Schulz-Streeck for helpful comments on the manuscript. We also thank Ruedi Fries and Hubert Pausch for processing of SNP arrays. This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Syntbreed—Synergistic plant and animal breeding” (FKZ: 0315528A).

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The authors declare that the experiments comply with the current laws of Germany.

References

- Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. *Theor Appl Genet* 123:339–350
- Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. *Stat Sci* 24:451–471
- Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype \times environment interaction using pedigree and dense molecular markers. *Crop Sci* 53:707–719
- Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banzinger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713–724
- Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schön CC, Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* 6:e28334
- Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194:573–596
- Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0. VSN International Ltd. Hemel Hempstead
- Gordillo GA, Geiger HH (2008) Alternative recurrent selection strategies using doubled haploid lines in hybrid maize breeding. *Crop Sci* 48:911–922
- Grubbs FE (1950) Sample criteria for testing outlying observations. *Ann Math Stat* 21:27–58
- Guo Z, Tucker DM, Lu J, Kishore V, Gay G (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. *Theor Appl Genet* 124:261–275
- Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G (2014) The impact of population structure on genomic prediction in stratified populations. *Theor Appl Genet* 127:749–762
- Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet Sel Evol* 42:5
- Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet Sel Evol* 41:51
- Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. *Crop Sci* 52:146–160
- Hofheinz N, Borchardt D, Weissleder K, Frisch M (2012) Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theor Appl Genet* 125:1639–1645
- Holland JB, Nyquist WE, Cervantes-Martínez CT (2003) Estimating and interpreting heritability for plant breeding: an update. *Plant Breed Rev* 22:9–112
- Jonas E, de Koning DJ (2013) Does genomic selection have a future in plant breeding? *Trends Biotechnol* 31:497–504
- Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. *Stat Appl Genet Mol Biol* 12:375–391

- Melchinger AE, Utz HF, Schön CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149:383–403
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3* 2:1595–1605
- Piepho HP, Williams ER, Fleck M (2006) A note on the analysis of designed experiments with complex treatment structure. *HortScience* 41:446–452
- Piepho HP, Möhring J, Schulz-Streeck T, Ogutu JO (2012) A stage-wise approach for the analysis of multi-environment trials. *Biom J* 54:844–860
- Reif J, Gumpert FM, Fischer S, Melchinger AE (2007) Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics* 176:1931–1934
- Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisek J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* 44:217–220
- Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. *Genetics* 194:493–503
- Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez VM, Moreno-Gonzalez J, Melchinger AE, Bauer E, Schön CC, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192:715–728
- Saatchi M, McClure MC, McKay SD, Rolf MM, Kim JW, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, Bauck S, Woodward B, Dekkers JCM, Fernando RL, Schnabel RD, Garrick DJ, Taylor JF (2011) Accuracies of genomic breeding values in American Angus beef cattle using k-means clustering for cross-validation. *Genet Sel Evol* 43:40
- Schulz-Streeck T, Ogutu JO, Karaman Z, Knaak C, Piepho HP (2012) Genomic selection using multiple populations. *Crop Sci* 52:2453–2461
- Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics* 50:1171–1177
- Utz HF, Melchinger AE, Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross-validation and validation with independent samples. *Genetics* 154:1839–1849
- Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28:2086–2087
- Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Schön CC (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195:573–587
- Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, Melchinger AE (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3* 2:1427–1436