Theor Appl Genet (2015) 128:2189–2201
DOI 10.1007/s00122-015-2577-y

ORIGINAL ARTICLE

Forecasting the accuracy of genomic prediction with different selection targets in the training and prediction set as well as truncation selection

Pascal Schopp[1] · Christian Riedelsheimer[1] · H. Friedrich Utz[1] · Chris-Carolin Schön[2] · Albrecht E. Melchinger[1]

Received: 13 March 2015 / Accepted: 26 June 2015 / Published online: 1 August 2015

Abstract
Key message **Deterministic formulas accurately forecast the decline in predictive ability of genomic prediction with changing testers, target environments or traits and truncation selection**.
Abstract Genomic prediction of testcross performance (TP) was found to be a promising selection tool in hybrid breeding as long as the same tester and environments are used in the training and prediction set. In practice, however, selection targets often change in terms of testers, target environments or traits leading to a reduced predictive ability. Hence, it would be desirable to estimate for given training data the expected decline in the predictive ability of genomic prediction under such settings by deterministic formulas that require only quantitative genetic parameters available from the breeding program. Here, we derived formulas for forecasting the predictive ability under different selection targets in the training and prediction set and applied these to predict the TP of lines based on line per se or testcross evaluations. On the basis of two experiments with maize, we validated our approach in four scenarios characterized by different selection targets. Forecasted and empirically observed predictive abilities obtained by cross-validation generally agreed well, with deviations between −0.06 and 0.01 only. Applying the prediction model to a different tester and/or year reduced the predictive ability by not more than 18 %. Accounting additionally for truncation selection in our formulas indicated a substantial reduction in predictive ability in the prediction set, amounting, e.g., to 53 % for a selected fraction $\alpha = 10$ %. In conclusion, our deterministic formulas enable forecasting the predictive abilities of new selection targets with sufficient precision and could be used to calculate parameters required for optimizing the allocation of resources in multi-stage genomic selection.

Communicated by H. Iwata.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-015-2577-y) contains supplementary material, which is available to authorized users.

✉ Albrecht E. Melchinger
melchinger@uni-hohenheim.de

[1] Department of Applied Genetics, Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Fruwirthstr. 21, 70593 Stuttgart, Germany

[2] Plant Breeding, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Liesel-Beckmann-Straße 2, 85354 Freising, Germany

Introduction

Hybrid breeding involves (1) production of new candidates within each heterotic pool, (2) evaluation of their line per se performance (LP) especially for characters related to hybrid seed production and (3) evaluation of their testcross performance (TP) in combination with genotypes from the opposite heterotic pool (Hallauer 1990). With the widespread adoption of the double haploid (DH) technology in major crops such as maize, the number of new lines produced every year in the public and private sector has tremendously increased during the past decade. Hence, highly effective selection schemes are required to identify the most promising lines and hybrid combinations as early and as efficiently as possible (Technow et al. 2014). While maturity and many resistance and quality traits of hybrids can be predicted fairly accurately from the performance of their parents, the correlation between LP and TP for grain yield and other heterotic traits is generally weak (Smith

1986; Hallauer 1990; Mihaljevic et al. 2005a). Therefore, breeders are still searching for alternatives to predict the TP of new untested lines. Marker-assisted selection for TP proved unsatisfactory for most quantitative traits (Bernardo 2008; Heffner et al. 2009), because QTL mapping studies with sample sizes affordable in practical breeding programs warrant detection of only few significant marker–trait associations that explain merely a small to moderate proportion of the genotypic variance for TP (Melchinger et al. 1998; Schön et al. 2004; Mihaljevic et al. 2005a).

A more recent statistical approach for the prediction of genotypic values, termed genomic selection or prediction (Meuwissen et al. 2001), uses genome-wide, dense marker data for fitting models that simultaneously exploit the information provided by all markers. This property enables explaining a substantially larger proportion of the genetic variance of complex traits than with methods that preselect markers based on arbitrary significance thresholds (Meuwissen et al. 2001). Originally developed in the context of animal breeding, many studies found high potential for the use of genomic prediction in plant breeding (e.g., Bernardo and Yu 2007; Heffner et al. 2009; Riedelsheimer et al. 2012; Technow et al. 2014), particularly for prediction of TP (Albrecht et al. 2011; Zhao et al. 2012; Lehermeier et al. 2014). However, integration of genomic prediction into multi-stage selection generally practiced in hybrid breeding warrants further research for balancing the number of candidates evaluated in each breeding stage (Riedelsheimer and Melchinger 2013). To date, only few experimental studies investigated the influence of different testers or environments on genomic prediction of TP (Windhausen et al. 2012; Albrecht et al. 2014). Theoretical approaches for forecasting the predictive ability of genomic prediction based on available training data (i.e., phenotypic and genotypic data for a set of candidates) when altering the selection target by changing the tester, target environments or even target traits are to the best of our knowledge not available.

Dekkers (2007) derived formulas for calculating the response to selection of multiple traits on the basis of phenotypic information as well as genomic estimated breeding values (GEBVs). In contrast to single markers associated with a trait of interest, use of high-density genotypes and GEBVs derived from these data satisfies the assumption of multivariate normality necessary to derive deterministic predictions of response to selection (Lande and Thompson 1990). In hybrid breeding, LP and TP with different testers can be regarded as different selection targets of the candidates (Falconer and Mackay 1996) and, thus, be predicted on the basis of their marker genotypes. The same idea applies likewise to the performance of a candidate in different sets of environments. Exploiting these principles using Dekkers' (2007) theoretical framework motivated us to develop formulas for forecasting the performance of genomic prediction based on preexisting training data and quantitative genetic parameters of the population of interest.

Little attention has also been paid to the influence of truncation selection in the prediction set on the accuracy of the GEBVs. In view of the large number of DH lines produced every year anew, the selection intensity is high and only the genotypes with the highest GEBVs are promoted to the next stage of testing in hybrid breeding. Thus, if the prediction set represents a selected fraction, this must properly be taken into account, because truncation selection alone was found to reduce prediction accuracy (Muir 2007; Jannink 2010; Bastiaansen et al. 2012). As demonstrated here, theoretical results provided by Cochran (1950) can be used to obtain an analytical solution to this problem.

Our objectives were to (1) develop formulas for calculating the predictive ability of genomic prediction based on preexisting training data under altered selection targets in the training and prediction set using a selection index approach, (2) demonstrate the use of these formulas for forecasting the predictive ability of genomic prediction for TP on the basis of LP or TP across different years and/or testers, (3) compare for these scenarios the forecasted and respective empirically observed predictive ability determined by cross-validation based on phenotypic and genomic data from an experiment with maize and (4) assess the influence of different testers and/or years as well as truncation selection on the prediction accuracy in the selected candidates.

Theory

Dekkers (2007) used classical selection index theory to derive for multiple traits the correlations between phenotypic values and GEBVs that are required for calculating the selection gain under direct and indirect phenotypic and genomic selection. Adopting his notation, the theory presented here is based on the following linear model, where the index i refers to the selection target of the candidate, which, for example, might represent a trait or a tester in a specific year

$$P_i = G_i + E_i = Q_i + R_i + E_i = \hat{Q}_i + e_i + R_i + E_i, \qquad (1)$$

where P_i is the phenotypic value of the candidate, G_i the genotypic value of the candidate, E_i the random environmental noise in the phenotypic value of the candidate, Q_i the component of G_i in the candidate associated with markers according to the specification of the prediction model, R_i the part of G_i not associated with markers, $\hat{Q}_i$ the GEBV for Q_i and e_i the prediction error in estimation of Q_i by $\hat{Q}_i$.

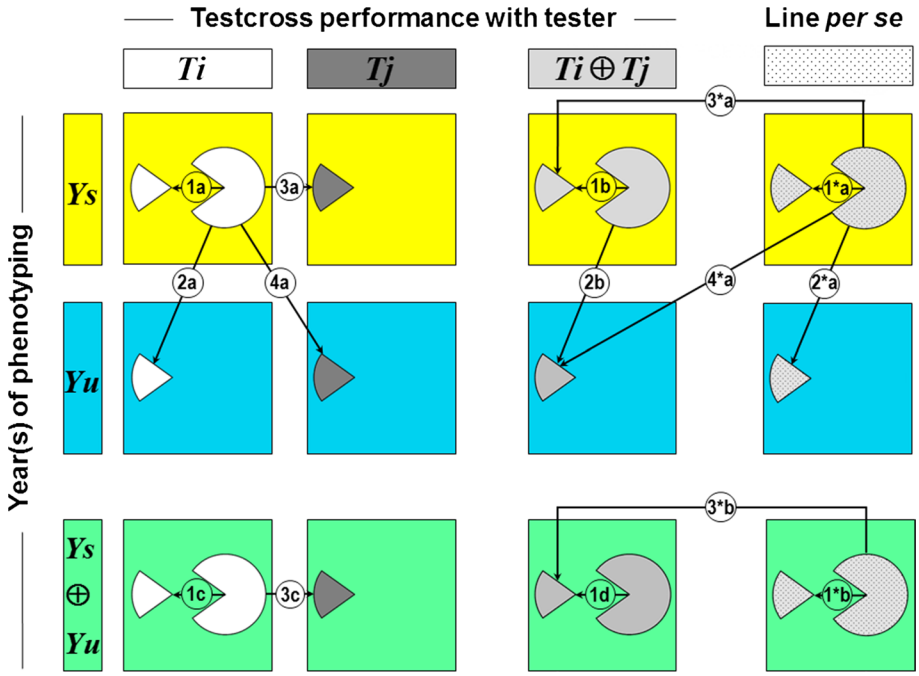

Fig. 1 Schematic representation of the predictions analyzed in this study. The large piece of the pie represents the training set (TS) and the small one the prediction set (PS). The first two columns and rows refer to the same (indices i and s, respectively) and different (indices j and u, respectively) testers and years in the TS and PS, respectively. The third column and row refer to the mean (indicated by $\oplus$) across both testers and/or years in the TS and PS, respectively. The first three columns refer to testcross performance evaluated in the TS; the fourth column refers to line per se performance evaluated in TS

For two selection targets i and j, we can express the correlations of the variables P_j, G_j and $\hat{Q}_j$ with variables Q_j, P_i, G_i and $\hat{Q}_i$ as a function of various quantitative genetic parameters (Supplementary Table S1). Unlike the heritabilities h_i^2 and h_j^2 for selection target i and j, respectively, many parameters in this table such as $q_i = r(G_i, Q_i)$ and $q_j = r(G_j, Q_j)$, as well as $r(Q_j, Q_i)$, cannot be estimated directly from experimental data. The formula given in Supplementary Table S1

$$r\left(P_j, \hat{Q}_i\right) = \frac{h_j}{h_i}\frac{q_j}{q_i} r\left(P_i, \hat{Q}_i\right) r\left(Q_j, Q_i\right) \tag{2}$$

can be simplified under certain assumptions. If q_i and q_j are large, then $r(Q_j, Q_i)$ approaches $r(G_j, G_i)$. Moreover, if the ratios h_j:h_i and q_j:q_i are close to 1, we obtain from Eq. (2) the following approximation

$$r\left(P_j, \hat{Q}_i\right) \approx r\left(P_i, \hat{Q}_i\right) r\left(G_j, G_i\right). \tag{3}$$

One goal of our study is to adopt and apply Eq. (3) to the problem of predicting TP of lines based on training set (TS) data evaluated with a tester or in a year different from the prediction set (PS). In the following part, we demonstrate how preexisting training data and quantitative genetic parameters can be used to forecast the predictive ability in a PS with an altered selection target prior to obtaining the corresponding phenotypes required to fit its own prediction model. For this purpose, we established four scenarios (Fig. 1) in which the TS and PS included the same or different tester(s) and were evaluated in the same or different year(s). Further, depending on whether one or both tester(s) [or year(s)] were included in the TS data, we specified four sub-scenarios **a**–**d** (Fig. 1). In detail, we predicted TP (with tester T1 or T2 or the mean across both testers) in the PS with GEBVs obtained from (1) TS data for TP with the same or different tester(s) evaluated in the same or different year(s) (Scenarios 1–4 in Fig. 1) or (2) TS data on LP evaluated in the same or different year(s) (Scenarios 1*–4* in Fig. 1).

Let P_{ju} denote the LP or TP of a candidate with tester(s) j (where $j = 1, 2$ for TP with tester T1 and T2, respectively, $j = 3$ for the mean TP across both testers and $j = 4$ for LP) averaged over locations in year(s) u (where $u = 1, 2$ for year Y1 and Y2, respectively, and $u = 3$ for the mean across both years). Likewise, let $\hat{Q}_{is}$ denote the GEBV of a candidate based on LP or TP with tester(s) i (where $i = 1, 2$ for TP with tester T1 and T2, $i = 3$ for the mean TP across both testers and $i = 4$ for LP) and over locations in year(s) s (where $s = 1, 2$ for year Y1 and Y2, respectively, and $s = 3$ for the mean over both years).

For Scenarios 1–4, the genetic trait correlation $r(G_j, G_i)$ in Eq. (3) must be replaced by the corresponding correlation $r(G_{ju}, G_{is})$ between the selection targets, i.e., the correlation between TP in the TS data (where $i = 1, 2, 3$ and $s = 1, 2, 3$ as defined above) and TP in the PS data (where $j = 1, 2, 3$ and $u = 1, 2, 3$ as defined above) given by

$$r\left(G_{ju}, G_{is}\right) = \frac{\sigma_g^2 + \delta_{ij}^t \sigma_{gt}^2/T + \delta_{su}^y \sigma_{gy}^2/Y + \delta_{ij}^t \delta_{su}^y \sigma_{gty}^2/TY}{\sigma_g^2 + \sigma_{gt}^2/T + \sigma_{gy}^2/Y + \sigma_{gty}^2/TY}, \tag{4}$$

where σ_g^2 is the genotypic (i.e., general combining ability) variance among candidates, σ_{gt}^2 the genotype × tester

(i.e., specific combining ability) interaction variance, σ^2_{gy} the genotype × year interaction variance, σ^2_{gty} the genotype × tester × year interaction variance, T the number of testers ($T = 1$ or 2), Y the number of years ($Y = 1$ or 2), $\delta^t_{ij} = 1$ in Scenarios 1_ and 2_ [same tester(s) ($i = j$) in TS and PS] and $= 0$ in Scenarios 3_ and 4_ [different testers ($i \neq j$) in TS and PS], $\delta^y_{su} = 1$ in Scenarios 1_ and 3_ [TS and PS evaluated in the same year(s) ($s = u$)] and $= 0$ in Scenarios 2_ and 4_ [TS and PS evaluated in different years ($s \neq u$)], where the underscores represent letters **a**–**d** used in Fig. 1 for characterizing the various sub-scenarios.

For Scenarios 1* and 2*, $r(G_{ju}, G_{is})$ is the correlation between LP in the TS data (where $i = 4$ and $s = 1, 2, 3$ as defined above) and LP in the PS data (where $j = 4$ and $u = 1, 2, 3$ as defined above) given by

$$r\left(G_{ju}, G_{is}\right) = \frac{\sigma^{2*}_g + \delta^y_{su}\sigma^{2*}_{gy}/Y}{\sigma^{2*}_g + \sigma^{2*}_{gy}/Y}, \quad (5)$$

where σ^{2*}_g is the genotypic variance among candidates and σ^{2*}_{gy} the genotype × year interaction variance. The asterisk distinguishes the genotypic variance for LP from the general combining ability variance for TP (no asterisk).

For Scenarios 3* and 4*, $r(G_{ju}, G_{is})$ is the correlation between LP in the TS data (where $i = 4$ and $s = 1, 2, 3$ as defined above) and TP in the PS data (where $j = 3$ and $u = 1, 2, 3$, as defined above) given by

$$r\left(G_{ju}, G_{is}\right) = \frac{\mathrm{cov}_{g,g^*} + \delta^y_{su}\frac{\mathrm{cov}_{gy,gy^*}}{Y}}{\sqrt{\sigma^2_g + \frac{\sigma^2_{gt}}{T} + \frac{\sigma^2_{gy}}{Y} + \frac{\sigma^2_{gty}}{TY}}\sqrt{\sigma^2_{g^*} + \frac{\sigma^2_{gy^*}}{Y}}}, \quad (6)$$

where cov_{g,g^*} is the genotypic covariance between GCA and LP among the candidates, cov_{gy,gy^*} the genotype × year covariance between GCA and LP, and all other terms are as defined above.

Following Dekkers (2007), prediction accuracies $r(G_{ju}, \hat{Q}_{is})$ were obtained from predictive abilities $r(P_{ju}, \hat{Q}_{is})$ by multiplying the latter with $1/h_{ju}$, where h^2_{ju} is the heritability on a progeny-mean basis of a genotype in the PS over T testers and Y years calculated as

$$h^2_{ju} = \frac{\sigma^2_g + \sigma^2_{gt}/T + \sigma^2_{gy}/Y + \sigma^2_{gty}/TY}{\sigma^2_g + \frac{\sigma^2_{gt}}{T} + \frac{\sigma^2_{gtl}+\sigma^2_{gtl}/T}{L} + \frac{\sigma^2_{gy}+\sigma^2_{gty}/T}{Y} + \frac{\sigma^2_{gly}+\sigma^2_{gtly}/T}{LY} + \frac{\sigma^2_\varepsilon}{TLYR}}, \quad (7)$$

where all variance components are defined as above and in addition σ^2_{gl} is the genotype × location interaction variance, σ^2_{gtl} the genotype × tester × location interaction variance, σ^2_{gtly} the genotype × tester × location × year interaction variance, σ^2_ε the residual error variance, L the number of locations ($L = 3$ in our experimental example), and R the number of replications in each location ($R = 2$ in our experimental example).

The heritabilities for LP were calculated analogously. In subsequent sections, all values estimated directly from experimental data are labeled with a hat [e.g., $\hat{r}(P_{ju}, \hat{Q}_{is})$ for the empirically determined predictive ability], whereas values forecasted by deterministic formulas are labeled with a tilde [e.g., $\tilde{r}(P_{ju}, \hat{Q}_{is})$ for the forecasted predictive ability].

Materials and methods

Genetic materials, phenotyping and genotyping

The data for our study were taken from two experiments with maize for biogas production described in detail by Grieder et al. (2012a, b) and Riedelsheimer et al. (2012). These experiments were re-analyzed with new objectives for the present study. Both experiments were based on a diversity panel of 285 inbred lines from the dent heterotic pool (including European dent, US Corn Belt dent, tropical germplasm) covering a broad range of variation for maturity and other agronomic traits. Based on prior information about their flowering date, the inbred lines were divided into three maturity groups, each comprising 95 genotypes.

In Experiment 1, 570 testcrosses were developed by crossing all 3 × 95 lines with two single-cross flint testers (T1 and T2). Together with five common check hybrids, the six groups of 95 testcrosses were evaluated in separate but adjacent trials laid out as 20 × 5 alpha designs with two replications. In Experiment 2, the three groups of 95 inbred lines together with 5 common check inbred lines were evaluated for their LP in adjacent trials also laid out as 20 × 5 alpha designs with two replications. All trials were conducted with two row plots in 2008 (Y1) and 2009 (Y2) at three locations in Germany with diverse agro-ecological conditions. Details on the experimental procedures including plant densities and data recording were described by Grieder et al. (2012a, b). Here, we concentrated on the traits dry matter yield (DMY, in t ha^{-1}) and dry matter concentration (DMC, in %) at silage harvest.

Genotyping of all 285 inbred lines was performed using the Illumina SNP chip MaizeSNP50 containing 56,110 SNPs (Ganal et al. 2011). Quality control and preprocessing of the genotypic data were performed as described in detail by Riedelsheimer et al. (2012). A total of 35,564 SNPs with a minor allele frequency (MAF) ≥5 % remained after filtering and were used for further analyses.

Statistical analyses

Phenotypic analysis

Phenotypic data were analyzed in multiple steps as detailed by Riedelsheimer et al. (2012). First, a lattice analysis of

each individual trial (corresponding to a combination of group of candidates, location, year and type of evaluation) was conducted to calculate adjusted-entry means of all 100 entries in each group and effective error mean squares using software PLABSTAT (Utz 2005). Second, we calculated a combined ANOVA jointly across both testcross series (i.e., across all trials for both testers and both years) to estimate variance components ($\hat{\sigma}_g^2, \hat{\sigma}_{gt}^2, \hat{\sigma}_{gy}^2$ etc.), treating genotypes and all other effects as random. Importantly, we followed the standard procedure in the analysis of testcross series and assumed homogeneous variances for locations, testers and years. This allowed us to insert the same estimates of variance components into Eqs. (4)–(7), irrespective of the tester and year used in each scenario. Analogously, variance components were calculated for LP. Third, we calculated a multivariate ANOVA with the data of both TP and LP to obtain estimates of their covariances. The last two steps were conducted with ASREML (Gilmour et al. 2009) and results are shown in Table 1. For fitting the genomic prediction models in the second step of the two-stage approach, we followed, e.g., Zhao et al. (2012), and calculated best linear unbiased estimates (BLUEs) of the entries by treating genotypes as fixed effects, while the assumptions for all other model terms remained unchanged.

Genomic analyses

Linkage disequilibrium (LD) between markers in the diversity panel was calculated as the squared correlation (r^2) between alleles at two loci (Hill and Robertson 1968). For graphical representation of LD, we grouped pairs of SNPs in consecutive bins of 0.05 mega base pairs (Mbp) width and calculated the average LD across all pairs within each bin. LD between adjacent SNPs was determined additionally. For all LD statistics, we used the preprocessed marker data described above.

Genomic prediction model and cross-validation schemes

Genomic prediction was performed using genomic best linear unbiased prediction (GBLUP) for each scenario shown in Fig. 1. The GBLUP model can be written as

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{Za} + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{y}$ is the vector of BLUEs of entries, μ the overall mean, $\boldsymbol{Z}$ an incidence matrix allocating adjusted entry means to breeding values, $\boldsymbol{a}$ the vector of breeding values and $\boldsymbol{\varepsilon}$ the vector of residuals. Vectors of random variables $\boldsymbol{a}$ and $\boldsymbol{\varepsilon}$ are assumed to be normally distributed as $N\left(0, \boldsymbol{K}\sigma_a^2\right)$ and $N\left(0, \boldsymbol{I}\sigma_\varepsilon^2\right)$, respectively, where $\boldsymbol{K} = \boldsymbol{WW}'/4\sum_{m=1}^{M} p_m(1-p_m)$ is the genomic relationship matrix (Habier et al. 2007; VanRaden 2008), $\boldsymbol{W}$ the centered version of the genotype matrix $\boldsymbol{X}$ coded as $\{-1,0,1\}$, p_m the sample frequency of the minor allele at marker locus m, σ_a^2 the additive genetic variance, $\boldsymbol{I}$ an identity matrix and σ_ε^2 the residual error variance. To account for the variance of fully homozygous inbred lines, the original denominator $2\sum_{(m=1)}^{M} p_m(1-p_m)$ of $\boldsymbol{K}$ (Habier et al. 2007) was multiplied by two. Restricted maximum-likelihood estimates of variance components as well as mixed model solutions $\hat{a}$ were calculated using the *mixed.solve* function implemented in the R-package *rrBLUP* (Endelman 2011). Note that breeding values a and GEBVs $\hat{a}$ correspond to Q and $\hat{Q}$, respectively, as used in the theory section.

Habier et al. (2010) showed that the accuracy of genomic prediction is strongly correlated with the maximum additive genetic relationship of each individual v in the PS with all individuals in the TS. Analogously, we calculated for each individual the maximum pairwise genomic kinship with all

Table 1 Estimates of variance and covariance components and their standard errors for TP and LP of the 285 maize inbred lines evaluated for dry matter yield (DMY, in t ha^{-1}) and dry matter content (DMC, in %)

Component[a]	DMY			DMC		
	Var(TP)	Var(LP)	Cov(TP,LP)	Var(TP)	Var(LP)	Cov(TP, LP)
g	183.4 ± 18.61	519.31 ± 51.81	206.47 ± 25.60	5.45 ± 0.54	33.70 ± 3.24	11.25 ± 1.20
gt	20.49 ± 4.28	–	–	0.33 ± 0.05	–	–
gy	8.92 ± 5.04	57.71 ± 11.34	14.74 ± 5.65	0.00	2.64 ± 0.46	0.00
gty	11.99 ± 4.37	–	–	0.05 ± 0.05	–	–
gl	0.00	75.24 ± 12.33	0.00	1.22 ± 0.16	6.72 ± 0.64	2.35 ± 0.26
gtl	0.00	–	–	0.00	–	–
gyl	52.53 ± 5.06	131.79 ± 12.07	19.14 ± 7.37[b]	1.83 ± 0.13	4.59 ± 0.40	1.30 ± 0.20[b]
gtyl	19.68 ± 5.34	–	–	0.34 ± 0.07	–	–
ε	177.8 ± 5.46	93.41 ± 4.47	–	2.21 ± 0.07	2.05 ± 0.09	–

[a] The meaning of the letters is detailed in "Materials and methods"

[b] Estimates refer to $\text{cov}_{gyl} + \text{cov}_{\varepsilon}/R$, because residual errors in TP and LP were assumed to be uncorrelated

other individuals in the population. For this purpose, we calculated genetic similarity indices as $\boldsymbol{S} = \frac{1}{2M}\boldsymbol{X}\boldsymbol{X}' + \boldsymbol{J}$ (Astle and Balding 2009), where $\boldsymbol{X}$ is defined as above, M is the number of markers and $\boldsymbol{J}$ is a $n \times n$ matrix with all elements being equal to 0.5. These values were then transformed into genetic kinship coefficients following Eding and Meuwissen (2001) as $f_{vw} = (S_{vw} - S_{min})/(1 - S_{min})$. For each line v, we then selected the maximum of values f_{vw} (for $v \neq w$) with all other lines, termed $\max(f_v)$, which served as an indicator of the impact of genetic relationships contributing to prediction accuracy in our population.

For each scenario and combination of tester(s) and year(s), we applied fivefold cross-validation (CV) using 228 lines (80 %) of the 285 inbred lines as TS for calibrating the prediction model and 57 lines (20 %) as PS. This procedure was repeated 200 times using different random splits of the 285 genotypes, yielding a total of 1000 replications per scenario and combination of tester(s) and year(s). Using data from the PS, we estimated the predictive ability $r(P_{ju}, \hat{Q}_{is})$ as the correlation between GEBVs $\hat{a}$ and the observed phenotypic values y in each CV run. Subsequently, we calculated means ($\hat{r}_{y,\hat{a}}$) and standard deviations (SDs) of these values over the 1000 replications and reciprocally equivalent scenarios in terms of tester(s) and year(s). Accordingly, we calculated $\hat{r}(\hat{Q}_{ju}, \hat{Q}_{is})$, which represents the correlation among GEBVs obtained using (a) the TS taken from the same data set as the PS [i.e., same tester(s) and year(s)] and (b) the TS corresponding to a different data set as the PS depending on the scenario (e.g., different tester and year in Scenario 4**a**).

Estimates of prediction accuracy $\hat{r}_{g,\hat{a}} = \hat{r}(P_{ju}, \hat{Q}_{is})/\hat{h}_{ju}$ were calculated for comparing the efficiency of genomic prediction across different scenarios, where the heritability in the PS differed between sub-scenarios (**a**–**d**) depending on the number of years and testers used for evaluation. For simplicity, we ignored the relatively small standard error of heritability estimates in calculating the SD of $\hat{r}_{g,\hat{a}}$ as $\mathrm{SD}(\hat{r}(P_{ju}, \hat{Q}_{is}))/\hat{h}_{ju}$.

Forecasting the predictive ability

For Scenarios 2–4 and 2*–4*, we calculated the forecasted predictive abilities $\tilde{r}_{y,\hat{a}}$ based on Eq. (3). The correlation $r(G_j, G_i)$ required in this formula reflects the expected reduction in predictive ability compared with Scenario 1, i.e., if prediction of TP is based on data from a different year and/or tester or LP instead of TP. Estimates of $r(G_j, G_i)$, denoted as $\hat{r}(G_{ju}, G_{is})$, were obtained by inserting estimates of variance and covariance components obtained from the combined ANOVA of TP and LP data into Eqs. (4)–(6). For $r(P_i, \hat{Q}_i)$, also required in Eq. (3), we inserted the mean of the empirically observed predictive ability $\hat{r}(P_{is}, \hat{Q}_{is})$ obtained via CV for Scenario 1 using the same number of years Y and testers T as the scenario to be forecasted, i.e., we used the sub-scenario of Scenario 1 having the same letter as the sub-scenario to be predicted. For example, for calculating $\tilde{r}(P_{ju}, \hat{Q}_{is}) = \hat{r}(P_{is}, \hat{Q}_{is})\hat{r}(G_{ju}, G_{is})$ in Scenario 3**c**, we inserted for $\hat{r}(P_{is}, \hat{Q}_{is})$ the observed predictive ability of Scenario 1**c**. The same approach was used for forecasting the predictive ability in all sub-scenarios of Scenarios 2–4 and 2*–4*.

Results

Linkage disequilibrium and genomic relationships

As already shown in Riedelsheimer et al. (2012), average LD (measured as r^2) for markers on the same chromosome showed a steep decline up to a physical distance Δ of approximately 0.5 Mbp (Supplementary Fig. S1a). For very short distances $\Delta < 0.05$ Mbp, we observed strong LD ($r^2 > 0.40$), while LD declined below $r^2 = 0.1$ for $\Delta = 0.5$ Mbp. The average long-range LD ($\Delta > 5$ Mbp) amounted to $r^2 = 0.02$–0.04 for markers on the same chromosome, slightly above the average of $r^2 = 0.01$ among markers on different chromosomes. LD between pairs of adjacent markers was generally high with an average of $r^2 = 0.35$. However, the values varied strongly and showed a U-shaped distribution (Supplementary Fig. S1b), with a considerable proportion of adjacent markers being in weak LD ($r^2 < 0.1$). The $\max(f_v)$ values showed substantial variation among lines ranging from 0.351 to 0.997 with multiple peaks (Supplementary Fig. S2).

Variance components and correlations

For both TP and LP, estimates of the genotypic variance, $\hat{\sigma}_g^2$ and $\hat{\sigma}_g^{2*}$, respectively, were highly significant ($P < 0.01$) for both traits (Table 1). Compared with $\hat{\sigma}_g^2$ for TP, $\hat{\sigma}_g^{2*}$ for LP was 2.8 times higher for DMY and 6.2 times higher for DMC. For TP, the magnitude of individual interaction variances was much smaller than $\hat{\sigma}_g^2$ for both traits. Noteworthy interaction variances for DMY were $\hat{\sigma}_{gt}^2, \hat{\sigma}_{gyl}^2$ and $\hat{\sigma}_{gtyl}^2$, which reached 11, 28 and 11 % of $\hat{\sigma}_g^2$, respectively. For DMC, $\hat{\sigma}_{gl}^2$ and $\hat{\sigma}_{gyl}^2$ reached 22 and 34 % of $\hat{\sigma}_g^2$, respectively. For LP, we found similar results with $\hat{\sigma}_{gl}^{2*}$ and $\hat{\sigma}_{gyl}^{2*}$ ranging from 14 to 25 % of $\hat{\sigma}_g^{2*}$ for both traits. The size of $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_\varepsilon^{2*}$ relative to $\hat{\sigma}_g^2$ and $\hat{\sigma}_g^{2*}$, respectively, was about twice as large for DMY compared with DMC.

The analysis of covariance revealed medium to high estimates of the genetic correlations between LP and TP for DMY (0.67) and DMC (0.83) (Table 1). For interaction

effects, we found estimated correlations among TP and LP of 0.65 for genotype × year (DMY) and 0.82 for genotype × location (DMC) interactions, while the two remaining correlations were set to zero, because the respective variance components of the TP data were estimated as zero (Table 1).

Observed and forecasted predictive abilities

Forecasted predictive abilities $\tilde{r}_{y,\hat{a}} = \tilde{r}(P_{ju}, \hat{Q}_{is})$ using TP (Scenarios 2–4) or LP (Scenarios 2*–4*) in the TS were always in the range of the empirically observed predictive abilities $\hat{r}_{y,\hat{a}} = \hat{r}(P_{ju}, \hat{Q}_{is})$ and their SDs (Tables 2, 3). In general, the magnitude of $\tilde{r}_{y,\hat{a}}$ was similar or slightly smaller than $\hat{r}_{y,\hat{a}}$. In the majority of scenarios, the correlation $\hat{r}(G_{ju}, G_{is})$ slightly exceeded $\hat{r}(\hat{Q}_{ju}, \hat{Q}_{is})$, but the differences were generally small.

Prediction accuracies

Estimates of prediction accuracies of GEBVs calculated on the basis of TP in the TS ranged from 0.59 (Scenarios 2**a** and 4**a**) to 0.71 (Scenario 4**a**) for DMY and from 0.73 (Scenario 4**a**) to 0.76 (Scenarios 1**b** and 1**d**) for DMC (Fig. 2a). Differences among scenarios were larger for DMY compared with DMC. Prediction accuracies increased slightly (0.02–0.07) from Scenario 1**a**–1**d** as a result of more reliable phenotypic information underlying genomic prediction by averaging data over testers (1**b**) or years (1**c**) or both (1**d**). Similar trends and arguments applied to the various sub-scenarios of Scenarios 2 and 3. Using training data from a different tester (Scenario 2) or a different year (Scenario 3) or both (Scenario 4) for fitting the models reduced $\hat{r}_{g,\hat{a}}$ in the PS for DMY only to a small extent (<0.05). In contrast, $\hat{r}_{g,\hat{a}}$ for DMC hardly varied and ranged from 0.74 to 0.76 for all scenarios.

Prediction accuracy for LP evaluated in the TS and PS (Scenarios 1***a** and 2***a**) amounted to about 0.61 for DMY and 0.80 for DMC (Fig. 2b). Thus, these values were slightly lower for DMY and slightly higher for DMC compared with the corresponding values of Scenarios 1**a** and 2**a**. However, when TP was predicted from LP in the TS (Scenarios 3* and 4*), the reduction in $\hat{r}_{g,\hat{a}}$ amounted to 0.19 for DMY and 0.08 for DMC (Scenario 1***a** vs. 4***a**) compared with the $\hat{r}_{g,\hat{a}}$ values for Scenarios 1* and 2*, respectively.

Prediction accuracy $\hat{h}$ of the phenotypic data, calculated as the square root of heritability $\hat{h}^2_{ju}$ in the PS, displayed almost identical trends for TP and LP. Values ranged from 0.89 to 0.96 for both traits and all scenarios (Fig. 2), where scenarios that used data of two years instead of one year (and/or testers) always showed slightly higher values. In Scenarios 1–4, $\hat{h}$ exceeded $\hat{r}_{g,\hat{a}}$ by 0.24–0.31 for DMY and 0.14–0.19 for DMC. For Scenarios 1* and 2*, $\hat{h}$ was superior to $\hat{r}_{g,\hat{a}}$ by 0.34–0.37 for DMY and 0.16–0.19 for DMC, while larger differences were observed for Scenarios 3* and 4*, amounting to 0.53 for DMY and 0.22 for DMC.

Table 2 Estimates[a] (and SD) of observed and forecasted predictive ability $\hat{r}_{y,\hat{a}} = \hat{r}(P_{ju}, \hat{Q}_{is})$ and $\tilde{r}_{y,\hat{a}} = \tilde{r}(P_{ju}, \hat{Q}_{is})$, respectively, as well as correlations among GEBVs $\hat{r}_{\hat{a}} = \hat{r}(\hat{Q}_{ju}, \hat{Q}_{is})$ for prediction of TP from data of TP (Scenarios 1–4, see Fig. 1) for DMY and DMC. The estimated genotypic correlation $\hat{r}_g = \hat{r}(G_{ju}, G_{is})$ for each scenario is also given

Description of scenarios						Dry matter yield (DMY)				Dry matter content (DMC)			
Characterization of the prediction set	δ^t_{ij}	δ^y_{su}	T	Y	$\delta^t_{ij}T; \delta^y_{su}Y$	$\hat{r}_{y,\hat{a}} \pm$ SD	$\tilde{r}_{y,\hat{a}}$	$\hat{r}_{\hat{a}}$	$\hat{r}_g$	$\hat{r}_{y,\hat{a}} \pm$ SD	$\tilde{r}_{y,\hat{a}}$	$\hat{r}_{\hat{a}}$	$\hat{r}_g$
Scenario 1													
1a: $u = 1 \vee 2; j = 1 \vee 2$	1	1	1	1	1; 1	0.57 ± 0.10	–	–	–	0.67 ± 0.08	–	–	–
1b: $u = 1 \vee 2; j = 3$	1	1	2	1	2; 1	0.63 ± 0.09	–	–	–	0.69 ± 0.07	–	–	–
1c: $u = 3; j = 1 \vee 2$	1	1	1	2	1; 2	0.63 ± 0.07	–	–	–	0.69 ± 0.08	–	–	–
1d: $u = 3; j = 3$	1	1	2	2	2; 2	0.68 ± 0.06	–	–	–	0.71 ± 0.06	–	–	–
Scenario 2													
2a: $u = 1 \vee 2; j = 1 \vee 2$	1	0	1	1	1; 0	0.53 ± 0.09	0.52	0.80	0.91	0.66 ± 0.08	0.66	0.95	0.99
2b: $u = 1 \vee 2; j = 3$	1	0	2	1	2; 0	0.59 ± 0.08	0.58	0.85	0.93	0.68 ± 0.07	0.69	0.96	1.00
Scenario 3													
3a: $u = 1 \vee 2; j = 1 \vee 2$	0	1	1	1	0; 1	0.54 ± 0.11	0.49	0.82	0.86	0.66 ± 0.08	0.62	0.97	0.94
3c: $u = 3; j = 1 \vee 2$	0	1	1	2	0; 2	0.61 ± 0.08	0.55	0.88	0.88	0.69 ± 0.07	0.65	0.98	0.94
Scenario 4													
4a: $u = 1 \vee 2; j = 1 \vee 2$	0	0	1	1	0; 0	0.53 ± 0.09	0.47	0.77	0.82	0.65 ± 0.08	0.62	0.97	0.94

[a] Estimates refer to the mean of individual estimates across equivalent combinations of i, j; s, u, where $i, j = 1, 2, 3$ refer to the TP with tester T1, T2 and their mean, respectively, and $s, u = 1, 2, 3$ refer to the mean over locations in year Y1, Y2 and across both years, respectively. The mathematical operator $\vee$ represents a "logical or"

Discussion

Theoretical aspects on forecasting predictive ability based on selection index theory

The first objective of our study was to derive deterministic formulas that enable forecasting the predictive ability of genomic prediction under altered selection targets in the TS and PS using a selection index approach as employed by Dekkers (2007). In the original publication, the author provided a theoretical framework that allows incorporating GEBVs of multiple traits obtained by genomic prediction into selection indices. For transferring this approach to forecasting the predictive ability, we borrowed the basic idea of Falconer and Mackay (1996) and treated in our application LP and TP from different years and/or testers as different selection targets of the candidates. The central parameter required for forecasting the predictive ability is the correlation $r(G_j, G_i)$ in Eq. (3), which, in the absence of selection, corresponds to the correlation between the original and new selection target in the TS and PS, respectively. For instance, when forecasting the TP of lines based on a different tester, the genotypic correlation among the testers has to be inserted for $r(G_j, G_i)$ in Eq. (3) (Windhausen et al. 2012). Likewise, forecasting TP with LP data requires an estimate of the genotypic correlation between TP and LP for the trait of interest. In conclusion, we provide formulas that enable the application of Eq. (3) to three situations in plant breeding, namely forecasting the predictive ability of (1) LP based on LP data, (2) TP based on LP data and (3) TP based on TP data, where the year(s) and/or tester(s) of the forecasted PS can be identical or different from that of the TS. Note that we exemplarily demonstrate the use of our formulas for two testers and two years as different selection targets in the TS and PS, but extensions to arbitrary numbers or alternative selection targets such as different traits or locations are straightforward.

Several assumptions must be made when applying the theory of Dekkers (2007) to the situation considered in the present study. First, the full equation for forecasting the predictive ability [Eq. (2)] includes the correlation $r(Q_j, Q_i)$ between all QTL effects captured by markers under both selection targets (e.g., TP and LP or TP in Y1 and TP in Y2). Since $r(Q_j, Q_i)$ is generally unknown, we substitute it by the correlation $r(G_j, G_i)$, which can be directly estimated from variance components [Eqs. (4)–(6)] for the various scenarios investigated in detail. This approximation works well if the proportion of the total genotypic variance explained by markers (q_i^2 and q_j^2) is close to 1 in both selection targets. The deviation of this proportion from 1 has been termed the "still missing heritability" and is partially caused by genetic variants not well tagged by markers (Wray et al. 2013). By using a high marker density, the "still missing heritability" found for complex traits in breeding populations with small effective population size was indeed close to zero (Jensen et al. 2012). Moreover, deviations of q_i^2 and q_j^2 from 1 can be caused by insufficient specification of the prediction model, e.g., by ignoring dominance and epistatic effects contributing to the genotypic value of an individual when modeling breeding values only [Eq. (8)]. Since the lines in our study were

Table 3 Estimates[a] (and SD) of observed and forecasted predictive ability $\hat{r}_{y,\hat{a}} = \hat{r}(P_{ju}, \hat{Q}_{is})$ and $\tilde{r}_{y,\hat{a}} = \tilde{r}(P_{ju}, \hat{Q}_{is})$, respectively, as well as correlations among GEBVs $\hat{r}_{\hat{a}} = \hat{r}(\hat{Q}_{ju}, \hat{Q}_{is})$ for prediction of (1) LP (Scenarios 1* and 2*, see Fig. 1) or (2) mean TP over testers T1 and T2 (Scenarios 3* and 4*, see Fig. 1) from data of LP for DMY and DMC. The estimated genotypic correlation $\hat{r}_g = \hat{r}(G_{ju}, G_{is})$ for each scenario is also given

Description of scenarios				Dry matter yield (DMY)				Dry matter content (DMC)			
Characterization of the prediction set	δ_{su}^y	Y	$\delta_{su}^y Y$	$\hat{r}_{y,\hat{a}} \pm$ SD	$\tilde{r}_{y,\hat{a}}$	$\hat{r}_{\hat{a}}$	$\hat{r}_g$	$\hat{r}_{y,\hat{a}} \pm$ SD	$\tilde{r}_{y,\hat{a}}$	$\hat{r}_{\hat{a}}$	$\hat{r}_g$
Scenario 1*											
1*a: $u = 1 \vee 2; j = 4$	1	1	1	0.55 ± 0.08	–	–	–	0.74 ± 0.06	–	–	–
1*b: $u = 3; j = 4$	1	2	2	0.59 ± 0.08	–	–	–	0.76 ± 0.05	–	–	–
Scenario 2*											
2*a: $u = 1 \vee 2; j = 4$	0	1	0	0.53 ± 0.09	0.49	0.87	0.90	0.72 ± 0.06	0.69	0.94	0.93
Scenario 3*											
3*a: $u = 1 \vee 2; j = 3$	1	1	1	0.39 ± 0.11	0.35	0.61	0.64	0.64 ± 0.07	0.58	0.88	0.78
3*b: $u = 3; j = 3$	1	2	2	0.41 ± 0.11	0.38	0.61	0.64	0.66 ± 0.07	0.61	0.88	0.80
Scenario 4*											
4*a: $u = 1 \vee 2; j = 3$	0	1	0	0.38 ± 0.10	0.33	0.56	0.60	0.63 ± 0.07	0.58	0.86	0.78

[a] Estimates refer to the mean of individual estimates across equivalent combinations of s and u, where $s, u = 1, 2, 3$ refer to the mean over locations in year Y1, Y2 and across both years, respectively; $j = 4$ refers to LP and $j = 3$ to the mean TP with testers T1 and T2. The mathematical operator $\vee$ represents a "logical or"

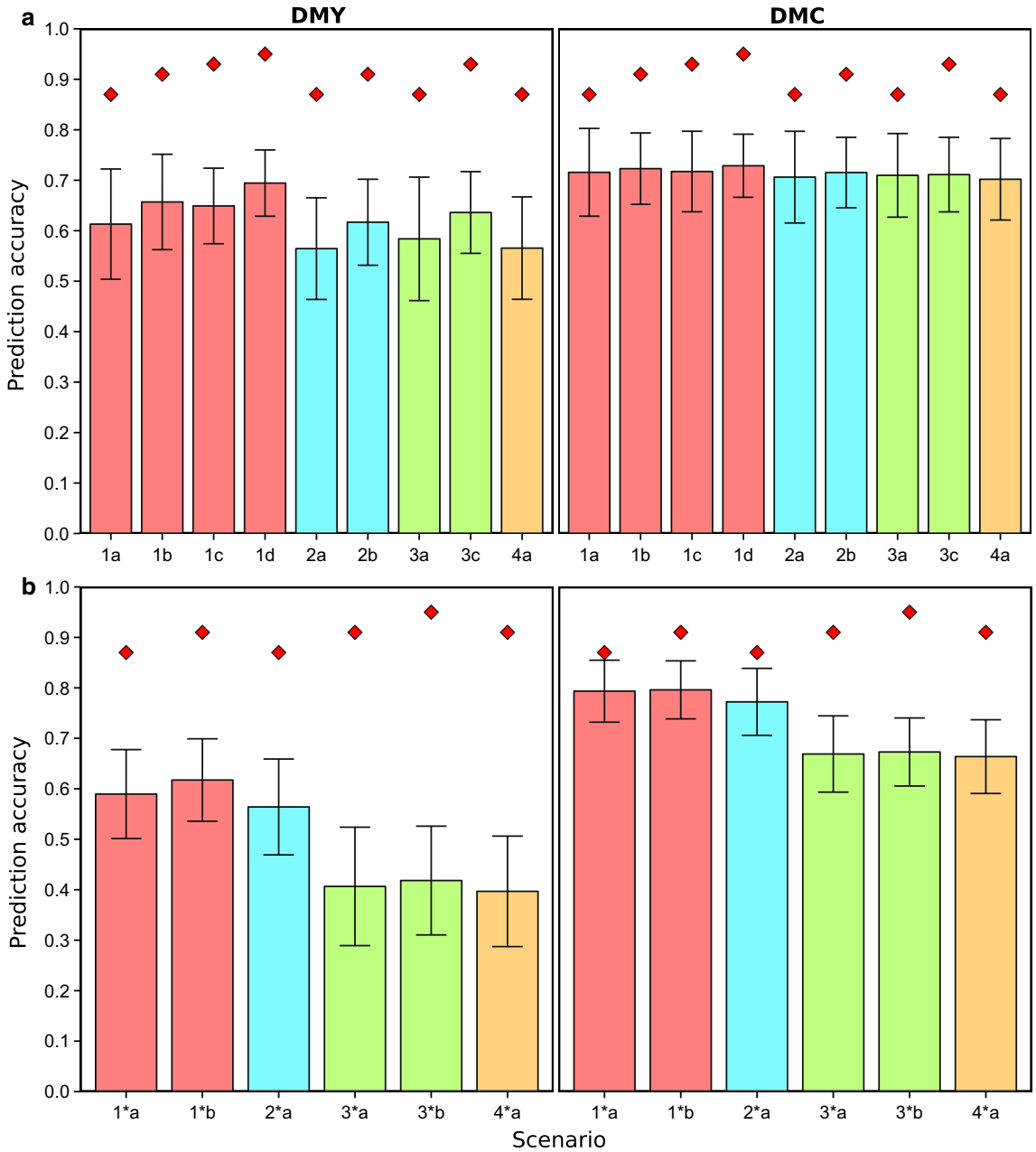

Fig. 2 Estimates of prediction accuracy, $\hat{r}_{g,\hat{a}}$, of GEBVs in the prediction set (PS), calculated as $\hat{r}(P_{ju}, \hat{Q}_{is})/\hat{h}_{ju}$, with their SDs for prediction of testcross performance (TP) for dry matter yield (DMY) and dry matter content (DMC) from **a** TP (Scenarios 1–4) or **b** line per se performance (LP) in the training set data (Scenarios 1*–4*) (see Fig. 1 for details). The red diamonds represent the accuracy of phenotypic selection, calculated as the square root of heritability $\hat{h}^2_{ju}$ in the PS

nearly homozygous and epistasis is generally assumed to have a small influence on the total genotypic value of complex traits in maize (Lamkey et al. 1995; Mihaljevic et al. 2005b), the substitution of $r(Q_j, Q_i)$ by $r(G_j, G_i)$ is expected to provide a good approximation in the present context as also indicated by the high similarity between $\hat{r}(G_{ju}, G_{is})$ and $\hat{r}(\hat{Q}_{ju}, \hat{Q}_{is})$ (Tables 2, 3).

Second, simplification of Eqs. (2)–(3) rests on the assumption that both the ratios h_j:h_i and q_j:q_i are approximately 1. In our study, these assumptions were closely met, because for most scenarios to be predicted (e.g., Scenarios 2–4) the corresponding Scenario 1 had exactly the same heritability (Fig. 2) according to Eq. (7). Exceptions were Scenarios 3* and 4*, where LP and TP data had slightly different heritabilities (Fig. 2b), but the ratios h_j:h_i were still close to 1. In practice, however, caution must be exercised, because heritabilities between the TS and the PS can differ depending on the stage of a

breeding program. For example, the heritability in an early stage (corresponding to evaluation of the TS) is likely to be different from later stages (corresponding to evaluation of the PS). If prior knowledge about h_j and h_i is available at each stage and one assumes that both q_i^2 and q_j^2 approach 1, then different heritabilities in TS and PS can be accounted for in Eq. (3) as

$$\tilde{r}\left(P_j, \hat{Q}_i\right) \approx \frac{h_j}{h_i} r\left(P_i, \hat{Q}_i\right) r\left(G_j, G_i\right). \quad (9)$$

Additional errors in forecasting might be introduced by violation of the assumptions made in the statistical analyses of the phenotypic data. Here, we assumed homogeneous interaction variances of genotypes with testers and years as well as homogeneous error variances. However, depending on the trait, testers and individual target environments, this assumption might be violated in practice. If so, an over- or underestimation of the forecasted predictive ability $\tilde{r}_{y,\hat{a}}$ can occur, because (1) the precision of $\hat{r}(G_{ju}, G_{is})$ might be affected and (2) the true ratios h_j:h_i and q_j:q_i might deviate from 1. A typical example in which this might happen are disease-related traits such as resistance against *Fusarium* head blight, where breeders make use of both weak and strong testers to screen the material for resistance and general combining ability, respectively. Similar arguments apply to years differing strongly in climatic conditions. Further research is warranted to investigate the effect of such possible violations on the accuracy of forecasted predictive abilities using data with more heterogeneous interaction and error variances.

Empirical results on forecasting predictive ability

We compared the forecasted predictive abilities $\tilde{r}_{y,\hat{a}}$ with the empirically observed predictive abilities $\hat{r}_{y,\hat{a}}$ obtained by CV to determine the precision of the forecasts (Tables 2, 3). With few exceptions, we obtained values of $\tilde{r}_{y,\hat{a}}$ that were within the range of $\hat{r}_{y,\hat{a}}$ and its SD, suggesting that Eq. (3) is suitable for forecasting the predictive ability from preexisting TS data. Generally, $\tilde{r}_{y,\hat{a}}$ tended to be smaller than $\hat{r}_{y,\hat{a}}$ (e.g., Table 2, Scenario 3 and DMY). Since CV yields unbiased estimates, smaller values of $\tilde{r}_{y,\hat{a}}$ compared with $\hat{r}_{y,\hat{a}}$ are most likely attributable to an overestimation of the interaction variances or underestimation of the genotypic variance entering Eqs. (4)–(6). For example in Scenario 3, overestimation of $\hat{\sigma}^2_{gy}$ and/or $\hat{\sigma}^2_{gty}$ might have caused an underestimation of $\hat{r}(G_{ju}, G_{is})$, which would explain the trend of stronger underestimation of $\tilde{r}_{y,\hat{a}}$ when predicting across years instead of across testers (Table 2).

As with most deterministic predictions, the successful application of our theory in practice depends on the precision of the quantitative genetic parameters required in our formulas. Accurate knowledge about these parameters, such as variance components and genetic correlations among different testers, TP and LP or different target locations are usually available in large-scale breeding programs. In the case of different years, however, accurate estimates of the required genotype by year interaction are only obtainable if the candidates were phenotyped in the target year too. Thus, one must rely on a priori available estimates from previous years, which are likely to be less accurate and therefore decrease the efficiency of our approach under such situations. Moreover, estimates of genetic correlations between LP and TP as well as variance components are available in the literature for many agronomically important traits and could be employed in the absence of germplasm-specific estimates.

A specific property of our data are the generally high values of $\hat{r}(G_{ju}, G_{is})$, which for most scenarios were above 0.80 (Tables 2, 3). Thus, the expected differences between $\tilde{r}(P_{ju}, \hat{Q}_{is})$ in Scenarios 2–4 (or 2*–4*, respectively) and $\hat{r}(P_{is}, \hat{Q}_{is})$ in Scenario 1 (or 1*, respectively) were small. Nevertheless, this reflects the situation usually faced in practice, where high correlations among TP with different testers are to be expected in hybrid breeding with genetically distant heterotic groups owing to the low importance of specific combining ability effects under this setting (Reif et al. 2007; Technow et al. 2014). Albrecht et al. (2014) reported a strong decline in $\hat{r}_{y,\hat{a}}$ when prediction was performed across testers. However, in their study this effect was confounded with the effect of two different groups of lines crossed to each tester as generally encountered in practice. Here, we crossed the very same set of DH lines to each tester, which enabled us to (1) estimate genetic correlations not available in the aforementioned study and (2) assess the effect of specific combining ability on the decline of predictive ability across testers. The good agreement of observed and forecasted predictive abilities in our study suggests that the strong decline found by Albrecht et al. (2014) was indeed caused by the different genetic groups as conjectured by the authors.

The impact of genotype × year interaction variances can be more severe than in the present study, leading to smaller values of $\hat{r}(G_{ju}, G_{is})$ and a stronger decline in $\hat{r}_{y,\hat{a}}$. A similar situation compared with ours was observed by Windhausen et al. (2012), who found a minor influence of different locations for TS and PS on the prediction accuracy. In principle, this finding holds also true for the situation of different years (Albrecht et al. 2014). In conclusion, it will be necessary to validate our approach with data displaying larger genotype × year (or tester) interactions, preferably from several years and testers. A first indication for successful application of the formulas with smaller values of $\hat{r}(G_{ju}, G_{is})$ is in Scenarios 3* and 4* (e.g., 0.60 for DMY in Scenario 4***a**), where the forecasted predictive abilities were still fairly accurate.

Including the effect of selection in forecasting the predictive ability

Let $\tilde{r}_{y,\hat{a}}$ be the forecasted predictive ability in one of the scenarios described above and suppose we practice truncation selection on the basis of GEBVs $\hat{a}$ and retain only a fraction α of the candidates in the PS for further evaluation. Then, using Eq. (10) in Cochran (1950), the expected predictive ability $\tilde{r}'_{y,\hat{a}}$ in this subset can be calculated as

$$\tilde{r}'_{y,\hat{a}} = \tilde{r}_{y,\hat{a}} \sqrt{\frac{[1 - i(i-z)]}{[1 - (\tilde{r}_{y,\hat{a}})^2 i(i-z)]}}, \quad (10)$$

where i is the selection intensity and z the standardized truncation point corresponding to α. For example, Scenario 4**a** had a forecasted predictive ability $\tilde{r}_{y,\hat{a}} = 0.47$ for DMY in the absence of selection based on TS data evaluated with a different tester in a different year. This corresponds to an expected reduction of the predictive ability of 18 % compared with a PS evaluated with the same tester in the same year ($\tilde{r}_{y,\hat{a}} = 0.57$ in Scenario 1**a**). Assuming that only $\alpha = 10\,\%$ of all lines in the PS are selected, the expected $\tilde{r}'_{y,\hat{a}}$ calculated with the aid of Eq. (10) equals 0.21, and the total reduction of predictive ability amounts to 63 %. Thus, for α values commonly employed in breeding practice (e.g., $\alpha < 20\,\%$), truncation selection reduces the predictive ability in the PS to a much larger extent than using a different tester or year (Muir 2007; Jannink 2010; Bastiaansen et al. 2012).

The accuracy of genomic predictions across scenarios and comparison to phenotypic selection

Overall, we found medium to high values for prediction accuracy $\hat{r}_{g,\hat{a}}$, which was surprising considering the diverse collection of lines incorporated in the population. In view of the latest findings on factors contributing to genomic prediction accuracy (Habier et al. 2010, 2013), such high values are unlikely if only population-wide linkage disequilibrium (Supplementary Fig. S1a and b) among markers and QTL are exploited with the sample size available in our study. Thus, the observed values most likely resulted from hidden population structure caused by closely related genetic subgroups incorporated in the diversity panel as indicated by the wide distribution of the $\max(f_v)$ values (Supplementary Fig. S2) and the vast range of long-range LD (Supplementary Fig. S1a) (Riedelsheimer et al. 2012). Thus, for the majority of individuals in the PS, close relatives with varying degrees of kinship were available in the TS, which presumably represents the main cause of the high $\hat{r}_{g,\hat{a}}$ values observed here (Habier et al. 2010; Albrecht et al. 2011; Windhausen et al. 2012).

The reasons for the relatively small impact of different testers and/or years on TP and the resulting high values of $\hat{r}(G_{ju}, G_{is})$ discussed in the previous section apply also to the comparison of prediction accuracy $\hat{r}_{g,\hat{a}}$ across the different scenarios. Generally, among the sub-scenarios within each scenario (e.g., Scenario 1**a**–1**d**), a change in $\hat{r}_{g,\hat{a}}$ can be explained mostly by a change in heritability attributable to incorporating data from more than one tester or year in the TS. However, the absolute increase in accuracy was only small as a result of the small magnitude of the interaction variances. Comparing $\hat{r}_{g,\hat{a}}$ of Scenario 1 to the corresponding Scenarios 2–4, a stronger decline was observed for DMY compared with DMC. This is most likely a result of the more complex genetic architecture of DMY, even though estimated heritabilities were similar for both traits (Fig. 2a).

When we predicted LP with LP data (Scenarios 1* and 2*), the values of $\hat{r}_{g,\hat{a}}$ were of similar magnitude as for the corresponding TP Scenarios 1 and 2. In conclusion, genomic prediction seems promising to select for LP among lines. However, further research is needed to determine whether the merit of genomic selection for seed or pollen production traits commonly assessed during the LP stage in hybrid breeding is as great as for DMY and DMC investigated here. Moreover, in practice, hundreds to thousands of DH lines generated anew every year are usually tested less extensively than in our LP experiments, and the genotypic variance of elite breeding material is likely to be smaller than in the diversity set of the present study, which might further impair the potential of genomic selection at this stage.

Prediction accuracies dropped substantially when TP was predicted based on data from LP for DMY, but not for DMC (Fig. 2b). Again, this observation can be explained most likely by the more complex genetic architecture of this heterotic trait (Smith 1986). Thus, compared with DMC, the genetic correlation among LP and TP is smaller (Table 3) and the resulting drop in accuracy from Scenario 1* to Scenarios 3* and 4* is larger (Fig. 2b). Nevertheless, in comparison with previous studies aiming at predicting TP based on marker loci significantly associated with QTL (Melchinger et al. 1998; Schön et al. 2004; Mihaljevic et al. 2005a), the accuracies achievable with genomic prediction are encouraging. In harmony with our findings, Foiada et al. (2015) predicted successfully TP of resistance against the European corn borer in maize based on LP of the same lines.

For the diversity panel investigated here, the accuracy of phenotypic selection (shown as $\hat{h}_{ju}$ in Fig. 2) exceeded considerably that of genomic prediction. This deviates from results of biparental and multi-cross populations, where the gap is generally much smaller for most traits (Riedelsheimer et al. 2013; Lehermeier et al. 2014).

Nevertheless, even with a considerable gap, genomic selection can be advantageous by reducing the generation interval (e.g., by usage of an off-season nursery) and increasing selection intensity, which becomes possible by increasing entry numbers owing to the lower costs of genotyping compared with phenotyping each candidate (Jannink et al. 2010; Riedelsheimer et al. 2013).

Conclusions

Numerous studies about the use of genomic prediction in the various phases of hybrid breeding have been conducted in recent years, including the prediction of LP (Technow et al. 2013), TP (Albrecht et al. 2011; Riedelsheimer et al. 2012; Zhao et al. 2012; Lehermeier et al. 2014) and hybrid performance (Technow et al. 2012, 2014; Massman et al. 2013). In agreement with the present investigation, the general conclusion of these studies was that genomic prediction holds great promise for the prediction of untested genotypes at each selection stage as long as the selection targets in the TS and PS are identical. However, in the course of a hybrid breeding program, the selection target of the candidates change due to the type of evaluation (LP vs. TP) or the use of new testers and/or different test environments. Here, we presented formulas that allow forecasting the predictive ability under such scenarios with information available from the breeding program. Without such formulas, the predictive ability for the desired selection target can only be assessed after phenotyping for the new selection target in the PS. Application of the deterministic formulas to experimental data from maize yielded close agreement of the forecasted and empirically observed predictive abilities, but further research is warranted to corroborate our findings with other experimental data.

The results of our study provide the basis for addressing the optimization of multi-stage genomic selection under a fixed budget and given costs for phenotyping and genotyping. This task has not been addressed hitherto, in spite of the longstanding history of multi-stage phenotypic selection, where both theory (Cochran 1950; Utz 1969) and software tools (Mi et al. 2014) have been developed to calculate the selection response and determine the optimum allocation of resources. With our formulas, breeders can forecast the predictive abilities at each selection stage, which allows them to apply these powerful tools also to multi-stage genomic selection.

Author contribution statement C.C.S. perceived the idea of investigating the topic. P.S. and A.E.M. wrote the manuscript. A.E.M., H.F.U. and P.S. developed the quantitative genetic theory. C.R. planned and conducted the field trials and pre-processed the marker data. H.F.U. conducted the phenotypic data analyses. P.S. performed genomic data analysis and cross-validation. P.S. and A.E.M. created all figures and tables. All authors read and reviewed the manuscript.

Acknowledgments The authors thank F. Mauch, J. Jesse, H. Poeschel, R. Lutz, T. Schmidt, S. Pluskat and R. Volkhausen for their assistance in conducting the field experiments. Funding for this research came from the German Federal Ministry of Education and Research (BMBF) within the framework of the projects GABI-Energy (FK 0315045B) and Cornfed (FK 03115461A) as well as from Syngenta under a Ph.D. fellowship for PS.

Compliance with ethical standards

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

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

References

Albrecht T, Wimmer V, Auinger H et al (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350

Albrecht T, Auinger H-J, Wimmer V et al (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor Appl Genet 127:1375–1386

Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451–471

Bastiaansen JWM, Coster A, Calus MPL et al (2012) Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet Sel Evol 44:3

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48:1649–1664

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Cochran WG (1950) Improvement by means of selection. In: Proceedings of second Berkeley symposium on mathematical statistics and probability, University of California Press, pp 449–470

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. Genetics 124:331–341

Eding H, Meuwissen T (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J Anim Breed Genet 118:141–159

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longmans Green, Harlow

Foiada F, Westermeier P, Kessel B, Ouzunova M, Wimmer V, Mayerhofer W, Presterl T, Dilger M, Kreps R, Eder J, Schön CC (2015) Improving resistance to the European corn borer: a comprehensive study in elite maize using QTL mapping and genome-wide prediction. Theor Appl Genet 128:875–891

Ganal MW, Durstewitz G, Polley A et al (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

Gilmour A, Gogel B, Cullis B, Thompson R (2009) ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead

Grieder C, Dhillon BS, Schipprack W, Melchinger AE (2012a) Breeding maize as biogas substrate in Central Europe: I. Quantitative-genetic parameters for testcross performance. Theor Appl Genet 124:971–980

Grieder C, Dhillon BS, Schipprack W, Melchinger AE (2012b) Breeding maize as biogas substrate in Central Europe: II. Quantitative-genetic parameters for inbred lines and correlations with testcross performance. Theor Appl Genet 124:981–988

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Habier D, Tetens J, Seefried F et al (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:5

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194:597–607

Hallauer AR (1990) Methods used in developing maize inbreds. Maydica 35:1–16

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Jannink J-L (2010) Dynamics of long-term genomic selection. Genet Sel Evol 42:35

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics Proteomics 9:166–177

Jensen J, Su G, Madsen P (2012) Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genet 13:44

Lamkey KR, Schnicker BJ, Melchinger AE (1995) Epistasis in an elite maize hybrid and choice of generation for inbred line development. Crop Sci 35:1272–1281

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Lehermeier C, Krämer N, Bauer E et al (2014) Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. Genetics 198:3–16

Massman JM, Gordillo A, Lorenzana RE, Bernardo R (2013) Genome-wide predictions from maize single-cross data. Theor Appl Genet 126:13–22

Melchinger AE, Utz H, Schön C-C (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. 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

Mi X, Utz HF, Technow F, Melchinger AE (2014) Optimizing resource allocation for multistage selection in plant breeding with r package. Crop Sci 54:1413–1418

Mihaljevic R, Schön C-C, Utz HF, Melchinger AE (2005a) Correlations and QTL correspondence between line per se and testcross performance for agronomic traits in four populations of European maize. Crop Sci 45:114–122

Mihaljevic R, Utz HF, Melchinger AE (2005b) No evidence for epistasis in hybrid and per se performance of elite European flint maize inbreeds from generation means and QTL analyses. Crop Sci 45:2605–2613

Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet 124:342–355

Reif JC, Gumpert F-M, Fischer S, Melchinger AE (2007) Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176:1931–1934

Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet 126:2835–2848

Riedelsheimer C, Czedik-Eysenberg A, Grieder C et al (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217–220

Riedelsheimer C, Endelman JB, Stange M et al (2013) Genomic predictability of interconnected bi-parental maize populations. Genetics 194:493–503

Schön C-C, Utz H, Groh S et al (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 498:485–498

Smith O (1986) Covariance between line per se and testcross performance. Crop Sci 2:540–543

Technow F, Riedelsheimer C, Schrag T, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:1181–1194

Technow F, Bürger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 3:197–203

Technow F, Schrag T, Schipprack W et al (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics. doi:10.1534/genetics.114.165860

Utz H (1969) Mehrstufenselektion in der Pflanzenzüchtung. Dissertation Thesis, University of Hohenheim

Utz H (2005) PLABSTAT—a computer program for statistical analysis of plant breeding experiments. University of Hohenheim, Stuttgart

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Windhausen VS, Atlin GN, Hickey JM et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2:1427–1436

Wray NR, Yang J, Hayes BJ et al (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14:507–515

Zhao Y, Gowda M, Liu W et al (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776