Theor Appl Genet (2012) 125:1639–1645
DOI 10.1007/s00122-012-1940-5

ORIGINAL PAPER

Genome-based prediction of test cross performance in two subsequent breeding cycles

Nina Hofheinz · Dietrich Borchardt · Knuth Weissleder · Matthias Frisch

Received: 26 June 2012 / Accepted: 29 June 2012 / Published online: 20 July 2012

Abstract Genome-based prediction of genetic values is expected to overcome shortcomings that limit the application of QTL mapping and marker-assisted selection in plant breeding. Our goal was to study the genome-based prediction of test cross performance with genetic effects that were estimated using genotypes from the preceding breeding cycle. In particular, our objectives were to employ a ridge regression approach that approximates best linear unbiased prediction of genetic effects, compare cross validation with validation using genetic material of the subsequent breeding cycle, and investigate the prospects of genome-based prediction in sugar beet breeding. We focused on the traits sugar content and standard molasses loss (ML) and used a set of 310 sugar beet lines to estimate genetic effects at 384 SNP markers. In cross validation, correlations >0.8 between observed and predicted test cross performance were observed for both traits. However, in validation with 56 lines from the next breeding cycle, a correlation of 0.8 could only be observed for sugar content, for standard ML the correlation reduced to 0.4. We found that ridge regression based on preliminary estimates of the heritability provided a very good approximation of best linear unbiased prediction and was not accompanied with a loss in prediction accuracy. We conclude that prediction accuracy assessed with cross validation within one cycle of a breeding program can not be used as an indicator for the accuracy of predicting lines of the next cycle. Prediction of lines of the next cycle seems promising for traits with high heritabilities.

Communicated by M. Sillanpää.

N. Hofheinz · M. Frisch (✉)
Institute of Agronomy and Plant Breeding II,
Justus Liebig University, 35392 Giessen, Germany
e-mail: matthias.frisch@agrar.uni-giessen.de;
matthias.frisch@uni-giessen.de

D. Borchardt · K. Weissleder
KWS Saat AG, 37555 Einbeck, Germany

Introduction

Prediction of genetic values with genome-wide dense marker maps was proposed in an animal breeding context by Meuwissen et al. (2001). Simulation studies (Bernardo and Yu 2007; Bernardo 2009; Wong and Bernardo 2008; Xu 2003; Zhong et al. 2009) suggested that it can overcome shortcomings limiting the application of QTL mapping and marker assisted selection in plant breeding.

In a study with maize, test cross performance for kernel dry weight of 208 doubled haploid lines was assessed in five locations (Piepho 2009). The lines were genotyped with 136 SNP and SSR markers and the model fit of various ridge regression models was assessed. It was suggested that genotype × environment interactions and genetic effects not captured by markers should be included in genome-based prediction models. Parametric and semi-parametric models for genome-based prediction were compared in a study using phenotypic data of 599 wheat lines grown in four environments and 300 maize lines grown under two different conditions (Crossa et al. 2010). 1,447 markers were used for the genotyping of the wheat lines and 1,148 markers for the maize lines. In cross validation, correlations between observed and predicted performance in the range of 0.4–0.5 were observed for grain yield and up to 0.79 for flowering time. Genome-based prediction with mixed linear models was investigated in a study with 1,380 doubled haploid maize lines grown in seven environments and phenotyped for the traits grain dry matter yield and grain dry matter content (Albrecht et al. 2011). The lines were genotyped with 1,152 SNP markers. In cross validation, correlations between predicted and observed test cross performance up to 0.74 were observed. While Piepho (2009) used the model fit in the estimation set as a measure to compare alternative models, Crossa et al. (2010) and Albrecht et al. (2011) used cross validation in the estimation set to assess prediction accuracy. However, no results are available investigating the accuracy of genome-based prediction when the set of lines to be predicted belongs to the breeding cycle that follows the breeding cycle to which the estimation set belongs.

The goal of our study was to assess the accuracy of genome-based prediction of test cross performance for sugar content (SC) and standard molasses loss (ML) in sugar beet (*Beta vulgaris* L.) by using data from two subsequent cycles of a breeding program. In particular our objectives were to (1) compare ridge regression employing preliminary estimates of the heritability (RIR) with best linear unbiased prediction (BLUP) for predicting marker effects, (2) compare cross validation for assessing prediction accuracy of genome-based prediction with validation using data from a subsequent breeding cycle, (3) draw conclusions on the potential of genome-based prediction in sugar beet breeding.

Methods

Plant material

The estimation set consisted of 310 inbred lines randomly derived from 34 crosses among 9 diploid sugar beet lines. The number of progenies from each cross ranged from two to seven. The 56 lines of the validation set were derived from 8 crosses among 6 lines of the estimation set. The number of progenies from each cross ranged from 3 to 11. The line development included selection between crossing parents as well as selection between lines. The lines were selected for high performance and to maintain the genetic diversity within the breeding pool.

Field data

Test cross performance of the lines of the estimation set was evaluated for SC (%) and ML (%) in field trials at six European locations with one tester. The lines were a subset of a larger trial that was set up in 10 × 10 lattices with two replications. The lines of the validation set were evaluated as part of a larger trial that employed alpha lattices with block size 10 at six European locations. A two-replicate design was employed. The first replicate was assigned to the first of two testers and the second replicate to the second. Four standard genotypes were included. The field trials were analysed with a two-stage analysis. In the first stage the adjusted entry means were calculated for each environment. These were combined in an analysis of series of experiments. The error variance for the analysis of the series was obtained by pooling the individual error variances. Following the practice of commercial sugar beet breeding, relative values were calculated that refer to the average of the standard lines. The relative values were calculated from the means across environments. The use of relative values might not be totally consistent with assumptions implicitly made by our further analyses; however, in the present study, our focus is on practical applicability.

Marker data

Genotyping was carried out with the same marker set of 384 SNPs in the estimation set and in the validation set. The nine chromosomes of sugar beet had lengths of about 1 M and the total map length was 10.25 M. Hence, the average map distance between two adjacent markers was 3.6 cM. Markers with more than two alleles, more than 20 % missing values, or a low degree of polymorphism ($1 - \sum f_i^2 < 0.1$, where f_1, f_2 are the allele frequencies at a marker) were discarded. This resulted in 300 SNPs for the estimation set and 198 SNPs for the validation set that were used for the calculations.

The marker data were used to investigate the relatedness of the material (Fig. 1) and the decay of linkage disequilibrium depending between pairs of loci depending on their map distance (Fig. 2).

Linear model

For estimating the genetic effects of the SNPs we used the linear model

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the vector of N phenotypic values, β_0 a fixed intercept, $\mathbf{Z}$ the design matrix relating the marker data to genotypes, $\mathbf{u}$ the vector of genetic effects, and $\mathbf{e}$ the vector of residuals. The genetic effects $u_l (l = 1\ldots m)$ at the m SNPs were assumed to follow a normal distribution with expectation 0 and variance σ_u^2. The residuals were assumed to follow a normal distribution with expectation 0 and variance σ_e^2. It was assumed that $\text{cov}(u_i,u_j) = 0\ (i \neq j)$ and $\text{cov}(e_k,e_l) = 0\ (k \neq l)$.

We assume that the possible allele effects at each locus follow a distribution with a common variance. An alternative model takes the allele frequencies at the individual loci into account and assumes that in the estimation set each locus contributes equally to the genetic variance (Crossa et al. 2010). For predicting the genetic values in a

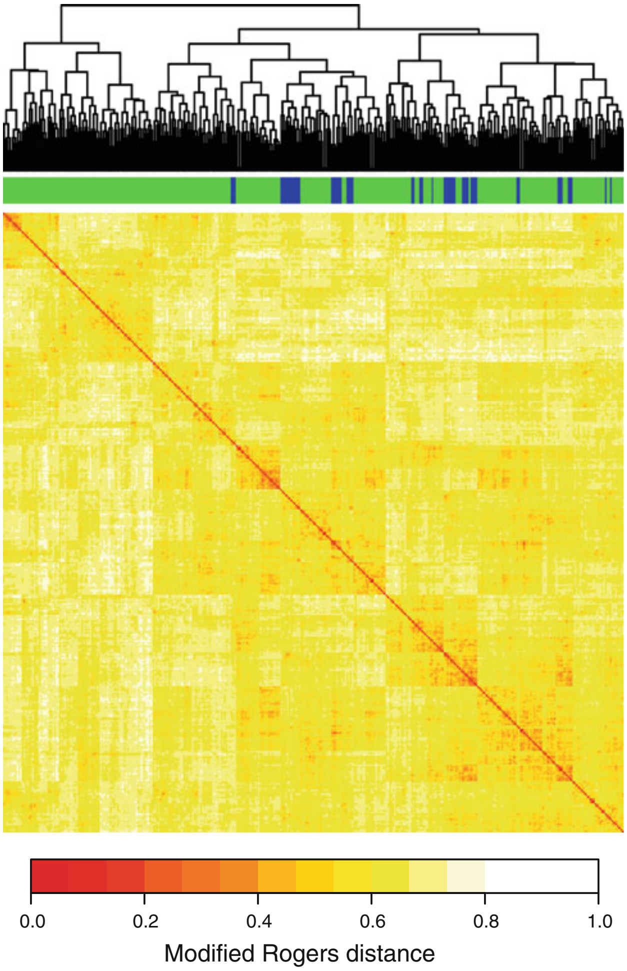

Fig. 1 Relatedness of the employed inbred lines based on the Modified Rogers distance determined from the SNP marker data. Average linkage clustering was used for ordering the distance matrix. Lines of cycle n are marked in *green* and lines of cycle $n + 1$ in *blue* (color figure online)

new validation set, our approach seems more suitable, because allele frequencies in the estimation and validation sets are most likely different.

The assumption of independent residuals in Eq. 1 is simplifying, because adjusted means are neither uncorrelated nor necessarily homoscedastic. It remains open to further research, whether more advanced linear models, that combine the analysis of the field design and the modeling of marker effects are able to increase the accuracy of prediction of genetic values.

Best linear unbiased prediction

We used an expectation-maximization (EM) algorithm to obtain restricted maximum likelihood (REML) estimates of the variance components σ_u^2 and σ_e^2 (Searle et al. 1992, p. 303). The EM algorithm is known to be slow in convergence and commercial software implements more sophisticated numerical approaches. However, it showed good performance for our data set. Convergence was reached with less than 10 iterations and computing times less than one second were required when using starting values determined on basis of Eq. 7. The algorithm showed high numerical stability and similar performance for other data sets from sugar beet and maize breeding programs.

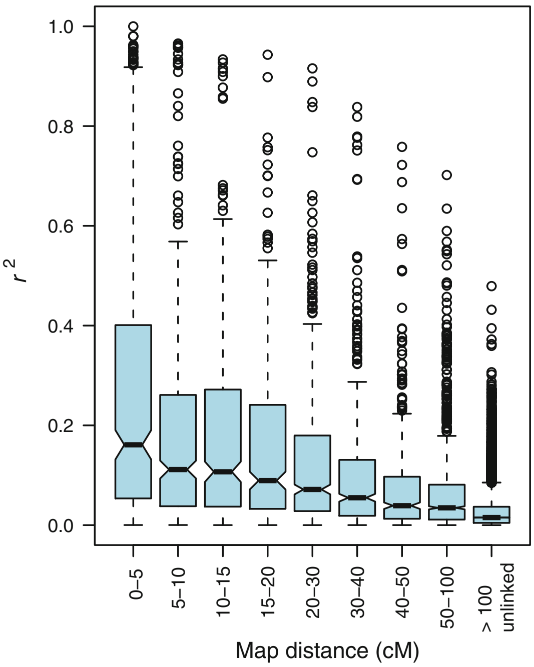

Fig. 2 Distribution of the pairwise linkage disequilibrium measure r^2 depending on the map distance between SNPs

To obtain best linear unbiased predictions (BLUP) of the genetic effects we solved (Searle 1987, p. 509)

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \lambda^2\mathbf{I} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix} \tag{2}$$

for **u** where

$$\lambda^2 = \sigma_e^2 / \sigma_u^2 \tag{3}$$

An LU decomposition with back substitution (Press et al. 1992, p. 44) was used for solving Eq. 2.

With respect to terminology, we follow the literature on linear models (Searle 1987, Searle et al. 1992) and Meuwissen et al. (2001), and use the abbreviation BLUP for the best linear unbiased prediction of the elements of the **u** vector. Albrecht et al. (2011) employed the term random regression (Model RR) for a similar model.

Prediction with ridge regression

RIR was carried out by solving the mixed model equations (Eq. 2) with a fixed shrinkage parameter λ^2. As a starting point, we used the convenient but incorrect assumption (Bernardo and Yu 2007), that the variance due to each marker can be approximated by dividing an estimate of the genotypic variance by the number of markers. As pointed out by Piepho (2009), estimates of the genotypic variance are usually obtained from models assuming independent genotype effects. This is in contrast with the marker-based ridge regression model that implies correlation among genotypic effects. Due to this fundamental difference in the models, we do not claim mathematical rigour for the RIR approach suggested in the following.

To determine λ^2 we used preliminary estimates of the heritability h_p^2 that are typically available for the traits under selection in a breeding program. In a simplified model, these can be interpreted as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \tag{4}$$

where σ_u^2 is the genetic variance and σ_e^2 the residual variance. This approximation is a second point where we do not claim mathematical rigour for our approach: The masking variance used to obtain the heritability estimate typically includes not only the residual variance but further variance components. These are totally ignored in our interpretation of the heritability. This is expected to result in an inflated value for the error variance, resulting in a stronger shrinkage of the genotypic effects. However, if we make this simplification, we can write

$$\frac{\sigma_e^2}{\sigma_g^2} = \frac{1}{h^2} - 1, \tag{5}$$

and together with the assumption of equal variances of the marker effects, we can use the approximation

$$\sigma_u^2 \approx \frac{1}{m}\sigma_g^2 \tag{6}$$

to define

$$\lambda^2 = \frac{\sigma_e^2}{\sigma_u^2} = m\left(\frac{1}{h_\mathrm{p}^2} - 1\right). \tag{7}$$

Using a shrinkage factor as defined in Eq. 7 can be regarded as an approximation of the BLUP approach. The difference between RIR and BLUP is that with RIR the shrinkage factor is determined from genetic and residual variances that were approximated from results on preliminary estimates of heritability, while in the BLUP approach these variances are estimated from the data. Hence, if the variance components correspond to the marker data, as is the case in the simulation example of Shepherd et al. (2010), then Eq. 7 results in BLUP. If preliminary estimates for the heritability are used, then it approximates BLUP. To determine λ^2 for our experimental data, we used preliminary estimates of the heritabilities of $h_\mathrm{p}^2 = 0.9$ and 0.4 for the traits SC and ML. These values are not estimated for the particular set of material under consideration, nor approaches were employed to obtain the most precise heritability estimates possible for unbalanced data (Piepho and Möhring 2007). The appeal of the method lies in the fact that it employees rule-of-thumb estimates of the heritability that are easily available in breeding programs.

Validation

For assessing the prediction accuracy we carried out (a) cross validation within one breeding cycle and (b) validation with lines of the next breeding cycle. In each of 100 cross validation runs, the lines of the first breeding cycle were divided randomly to two parts, 254 lines were used to estimate marker effects and 56 lines to validate the effects. The correlations between observed and predicted test cross performance for RIR and BLUP were averaged over the 100 runs. For validation with lines from the next breeding cycle, we estimated the marker effects with the lines from the first breeding cycle and predicted the test cross performance of the lines of the subsequent breeding cycle. Then we assessed the correlation between the predicted and observed test cross performance.

Results

For SC the correlation between observed and predicted test cross values in the estimation set was $r = 0.94$ with RIR (employing a $h_p^2 = 0.9$) and $r = 0.93$ for BLUP. In cross validation, correlations of on average 0.82 were observed for both prediction models. Prediction of the test cross values of lines of the next breeding cycle resulted in correlations $r = 0.79$ (RIR) and 0.80 (BLUP, Fig. 3).

For ML the correlation between predicted and observed test cross values in the estimation set was slightly greater for BLUP ($r = 0.94$) than for RIR ($r = 0.90$). However, in cross validation similar average correlations of $r = 0.85$ (RIR) and 0.86 (BLUP) were observed for both prediction models. Despite these high correlations in cross validation, that were even greater than those observed for SC, the transferability of the effect estimates to the next breeding cycle was low. A correlation of $r = 0.41$ was observed for RIR (employing a $h_\mathrm{p}^2 = 0.4$) and $r = 0.39$ was observed for BLUP.

Discussion

Accuracy of prediction methods

Bayesian methods provided better prediction accuracy than BLUP in the study that initially suggested genome-based

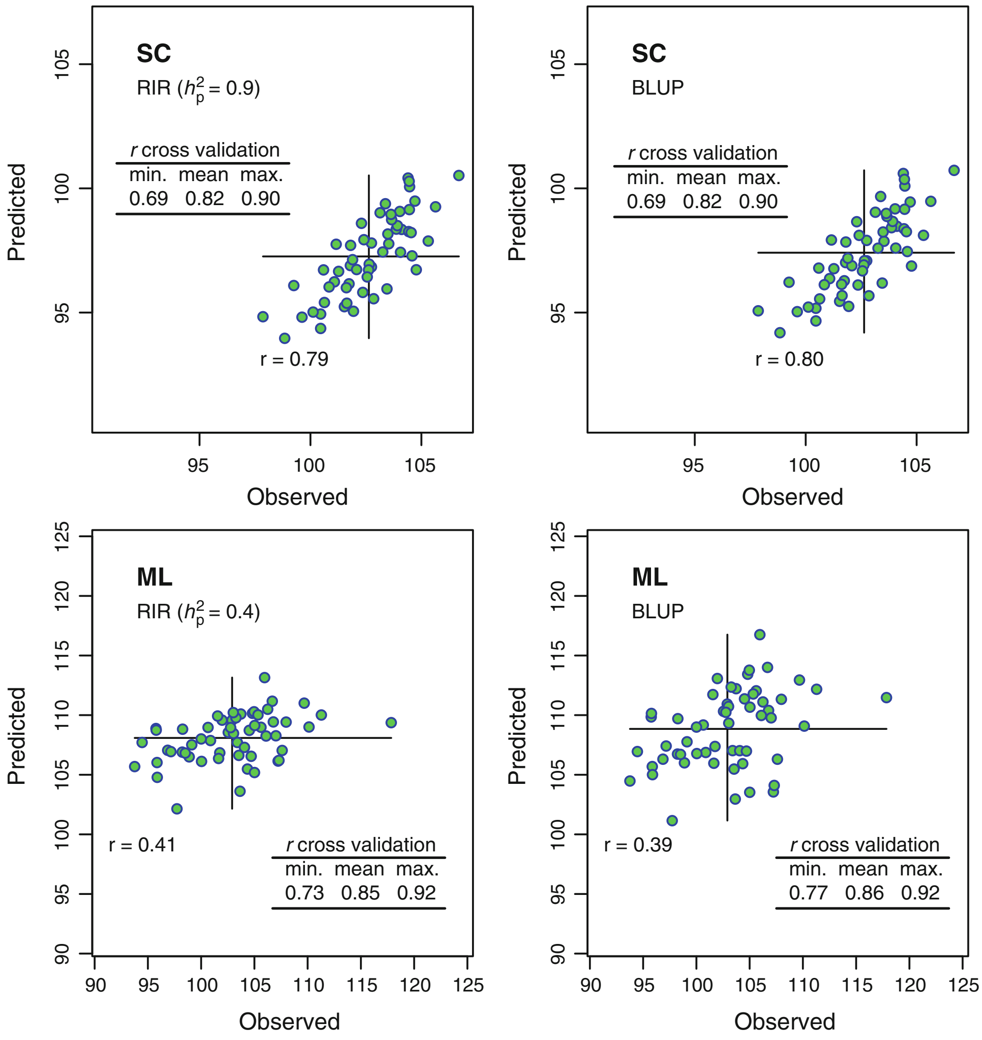

Fig. 3 Prediction of test cross performance for SC and ML. Observed versus predicted test cross performance in the validation set for prediction with RIR and BLUP. h_p^2 is the preliminary estimate of the heritability employed in RIR and *r* the correlation between predicted and observed values. In the tables, the minimum, the mean, and the maximum of the correlations between predicted and observed test cross performance in the cross validation runs with the estimation set are presented

prediction of genetic values (Meuwissen et al. 2001). Since then much effort was invested in Bayesian estimation methods (Gianola and van Kaam 2008; Park and Casella 2008; Gianola et al. 2006) that allow for distributions of the genetic effects with unequal variances, because it was expected that they provide improved prediction accuracy. However, as pointed out by Piepho (2009) and Bernardo and Yu (2007), the fact that all genetic effects are modelled as realizations of random variables with the same variance does not imply that all loci contribute equally to the genetic value. It was suggested by Piepho (2009) and Goddard and Hayes (2007) that the advantage of Bayesian estimation over BLUP observed by Meuwissen et al. (2001) might be a consequence of the effect distributions in the employed simulation model. Bernardo and Yu (2007) concluded that for plant models, Bayesian methods would provide little, if any, advantage and Zhong et al. (2009) found BLUP to outperform Bayesian estimation. For grain yield in maize Albrecht et al. (2011) and Crossa et al. (2010) found that prediction accuracy of BLUP was similar to that of Bayesian estimation with varying variances. However, for flowering time superiority of Bayesian estimation was observed (Crossa et al. 2010). In accordance with Daetwyler et al. (2010), a possible conclusion from these studies is that approaches with variable variances might be superior for traits that are controlled by a few major genes. In contrast, for polygenic traits that follow closely the infinitesimal model of quantitative genetics, models with constant variances might be more appropriate. Schneider et al. (2002) detected five QTLs for SC on five chromosomes. They also found several QTLs for potassium, sodium, and alpha-amino nitrogen, which account for the trait ML. These results suggest that, due to the polygenic inheritance of SC and ML, BLUP is an appropriate method for genome-wide prediction in our data set.

Average correlations between predicted and observed test cross performance from the cross validation of BLUP were 0.82 for SC and 0.86 for ML (Fig. 3). These values confirm the results of Albrecht et al. (2011) and Piepho (2009) that BLUP can provide precise predictions, and support the hypothesis that for polygenic traits BLUP with constant variances is a suitable prediction method. RIR based on preliminary estimates of the heritability provided the same prediction accuracy as BLUP for both traits. With the present data set consisting of roughly 300 lines and 300 markers, obtaining BLUPs was not technically challenging. However, with large data sets convergence problems could occur. For such data sets an RIR approach might prove useful. In conclusion, BLUP provided genome-based predictions of high accuracy, and approximating BLUP on basis of preliminary estimates of heritabilities with RIR is a computationally simple alternative that was not accompanied with losses in prediction accuracy.

Cross validation and validation with the subsequent breeding cycle

The average correlations between predicted and observed test cross performance in cross validation were 0.82 (SC) and 0.86 (ML). Compared with results from maize and wheat (Crossa et al. 2010; Albrecht et al. 2011) these values are high. An explanation for the high correlations might be the homogeneity of the material in the investigated breeding pool. With an average distance between two adjacent markers of ≈ 3 cM, prediction of genetic values still relies on gametic disequilibrium between marker and QTL alleles. If the breeding material in a pool is homogeneous, then the linkage phase of marker and QTL alleles is expected to be the same for large parts of the material, resulting in high prediction accuracy. In more diverse breeding material, however, more dense marker maps, ideally to the point that each gene underlying a trait can be directly traced by a SNP, are expected to improve prediction accuracy.

The correlation between observed and predicted values in cross validation was smaller for SC ($h_\mathrm{p}^2 = 0.9$) than for ML ($h_\mathrm{p}^2 = 0.4$). This result indicates that even for traits with low heritabilities, good correlations between observed and predicted performance can be obtained in cross validation. The relatedness of the genotypes within a breeding pool can be a reason for such high correlations. The following example illustrates the problem. Assume several full sib lines that share common marker alleles at several loci not underlying the trait under consideration. In addition, they share a high performance. Some lines are part of the estimation set in a cross validation run and others are part of the validation set. As a consequence, high effect estimates are assigned to the common marker alleles, and these effects are validated by the sister lines in the validation set. An important conclusion from these results is that cross validation in breeding pools of related material does not necessarily correct prediction models for over-fitting. In consequence, high correlations between predicted and observed performance in cross validation do not guarantee a good transferability of the estimated effects to a different set of breeding material.

In contrast to cross validation, where the correlations between predicted and observed performance were high for both traits traits, in independent validation large differences were observed. While for SC correlations amounted to 0.8, only correlations of 0.4 were observed for ML. These correlations correspond well to the preliminary estimates of the heritability $h_\mathrm{p}^2 = 0.9$ (SC) and 0.4 (ML). This indicates that cross validation can only provide limited information on the accuracy of predicting line performance with effects estimated from a previous breeding cycle. In particular it remains open to further research whether results comparing the accuracy of different prediction models are robust with respect to the difference between cross validation and independent validation.

Application in breeding programs

Test cross performance of lines in hybrid breeding can be predicted either with effects estimated from related lines of the same breeding cycle or with effects estimated in a previous breeding cycle. Prediction of untested lines with an estimation set from the same breeding cycle can be implemented by generating more candidate lines than will be evaluated in field trials. After having evaluated a portion of the lines in field trials, the performance of the second portion of lines is predicted, and the lines with the best predictions were included in the second stage of line testing. Employing genome-based selection in such a scenario is conceptually similar to the assessment of prediction accuracy with cross validation. Due to the relatedness of the breeding material, even random associations between markers and phenotypes can be exploited by genome-based prediction. The high correlations in cross validation suggest that a considerable gain in response to selection can be realized with such applications.

Prediction of lines with an estimation set from the previous breeding cycle can be implemented as follows. More candidate lines are generated than will be evaluated in the field trials. All of these are genotyped and those with the best predicted test cross values were evaluated in the field. This can be regarded as indirect selection where the correlation ρ between the trait under selection and the trait to be improved is the correlation between the gene effects in the estimation set and the gene effects in the validation set (which could be called in this context more appropriately prediction set). The upper bound of this correlation is limited by a measure for the heritability, that takes into account not only the variance components of the field trial, but in addition the genetic change through recombination. We conclude that for assessing the accuracy of genome-based prediction with effects estimated in previous breeding cycles, cross validation within one cycle is not sufficient, but independent validation is required. Our results suggest that such predictions are only promising for traits with high heritabilities.

Acknowledgments We thank Gregory Mahone for proof reading the manuscript. We thank two anonymous reviewers for their helpful comments.

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

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

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

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Bänzinger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

Daetwyler HD, Pong-Wong R, Villanueva B, Wooliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031

Gianola D, van Kaam JBCHM (2008) Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776

Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124:323–330

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103:681–686

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165–1176

Piepho HP, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

Schneider K, Schäfer-Pregl R, Borchardt DC, Salamini F (2002) Mapping QTLs for sucrose content, yield and quality in a sugar beet population fingerprinted by EST-related markers. Theor Appl Genet 104:1107–1113

Searle SR (1987) Linear models for unbalanced data. Wiley, New York

Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York

Shepherd RK, Meuwissen THE, Woolliams JA (2010) Genomic selection and complex trait prediction using a fast EM algorithm applied to genomewide-markers. BMC Bioinformatics 11:529

Wong CK, Bernardo R (2008) Genome wide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor Appl Genet 116:815–824

Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163:789–801

Zhong S, Dekkers JCM, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364