Plant Breeding, **136**, 331–337 (2017)
© 2017 Blackwell Verlag GmbH
doi:10.1111/pbr.12482

Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme

Carola Zenke-Philippi[1], Matthias Frisch[1], Alexander Thiemann[2], Felix Seifert[2], Tobias Schrag[3], Albrecht E. Melchinger[3], Stefan Scholten[2,3] and Eva Herzog[1,4]

[1]Institute of Agronomy and Plant Breeding II, Justus Liebig University, 35392 Giessen, Germany; [2]Biocenter Klein Flottbek, Developmental Biology and Biotechnology, University of Hamburg, 22609 Hamburg, Germany; [3]Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany; [4]Corresponding author, E-mail: eva.herzog@agrar.uni-giessen.de

With 3 figures and 1 table

Received November 7, 2016 / Accepted March 28, 2017
Communicated by W. Link

Abstract

mRNA transcription profiles are an alternative to DNA markers for predicting hybrid performance. Our objective was to investigate their prediction accuracy in an unbalanced maize data set. We focused on the effectiveness of preselecting a core set of genes for transcription profiling and on the comparison of prediction models. A total of 254 hybrids were evaluated for grain yield and grain dry matter content. The mRNA transcripts of a core set of 2k genes and the genotype of 1k AFLP markers were assessed in the parental lines. Predictions based on transcriptome-based distances determined from the 2k core set of genes resulted in prediction accuracies below 0.5 and could not reach the high accuracies observed with a 46k micro-array in earlier studies. Predictions based on ridge regression resulted in prediction accuracies greater 0.6. Only marginal differences were observed in the prediction accuracies of mRNA transcripts compared with AFLPs. We conclude that mRNA transcription profiles are suitable for hybrid prediction with ridge-regression models in unbalanced designs, even if limited resources allow only transcription profiling of a core set of genes.

Key words: hybrid prediction — genomic prediction — mRNA transcription profiles — transcriptome-based distances — ridge regression

Choosing a suitable training set is crucial for successful prediction of hybrid performance in breeding programmes (Zhao et al. 2015). For prediction models using mRNA transcription profiles, important questions on how to most efficiently use the data generated in earlier breeding cycles are as follows: Which genotypes can be used as the training set? How many and which genes should be profiled? What prediction models have the greatest prediction accuracy?

When genomic selection was introduced for the prediction of plant hybrids, it was already recognized that marker data cannot capture all polygenic effects that might contribute to the traits of interest (Piepho 2009). In the same study, it was suggested that gene expression and metabolomic data might be used in ridge-regression models instead of marker data. Promising results of hybrid prediction have been reported for gene expression profiles (Andorf et al. 2010, Maenhout et al. 2010, Steinfath et al. 2010, Zenke-Philippi et al. 2016), transcriptome-based distances (Frisch et al. 2010, Fu et al. 2012) and metabolomic data (Riedelsheimer et al. 2012, Dan et al. 2016, Xu et al. 2016). Transcriptome-based distances for hybrid prediction were successful when using a 46k micro-array for expression profiling (Frisch et al. 2010). Resource use could be minimized if a small core set of genes related to the traits to be predicted could be used instead of profiling the expression of large sets of genes. The prerequisite is that such a core set is transferable between different experiments in a hybrid breeding programme. The effectiveness of using the transcription profiles of a core set of genes determined in an earlier breeding cycle of a breeding programme for prediction of new hybrids has to our knowledge not yet been investigated.

Experimental and simulation studies on genomic prediction of complex traits with marker data showed that ridge-regression approaches are computationally efficient and yield robust estimates of breeding values with high prediction accuracy (Piepho 2009, Heslot et al. 2012, Riedelsheimer et al. 2012, Technow et al. 2012, Massman et al. 2013). It has therefore been suggested that ridge-regression models could be used for routine prediction of hybrid performance in breeding programmes (Zhao et al. 2015). A combination of ridge-regression models with mRNA transcription profiles for hybrid prediction has been studied recently (Zenke-Philippi et al. 2016). However, the prediction accuracies in this study were estimated by cross-validation with data from one single factorial. A validation with a broader database, consisting of several experiments from one breeding programme, is still lacking.

Our main goal was to investigate how data, generated in earlier cycles of a breeding programme, can be used for transcriptome-based prediction of hybrid performance for grain yield (GY) and grain dry matter content (GDMC) of untested new maize hybrids. We used a data set consisting of 34 dent and 14 flint lines. Four complete factorial crosses of these lines were created in four different years. Taken together, they form an unbalanced incomplete factorial of 254 hybrids. For the parental lines, genotypes for 1k AFLP markers and mRNA transcription profiles for 2k genes were collected.

Our objectives were to (i) investigate whether the transcription profiles of a core set of genes preselected in one factorial can be used in other factorials of the same breeding programme for hybrid prediction with transcriptome-based distances, (ii) explore the prediction accuracy of ridge regression with mRNA transcription profiles in an unbalanced incomplete factorial by cross-validation and (iii) compare the prediction accuracies of mRNA transcription profiles and AFLPs for prediction of hybrid performance of one factorial using data from other factorials of the same breeding programme as the training set.

Materials and Methods

Field data: The field data were presented in detail by Schrag et al. (2006). In total, 48 maize elite inbred lines developed in the breeding

programme of the University of Hohenheim were used as parental lines for the factorial crosses under evaluation. The inbreds comprised 34 dent lines with Iodent or Iowa Stiff Stalk Synthetic background, and 14 flint lines with European flint or flint/Lancaster background. Four dent × flint factorial mating experiments (14 × 7, 11 × 4, 14 × 6, 11 × 4), further referred to as exps. 1–4, were produced, providing a total of 270 hybrids. Thereby, eight dent lines and six flint lines were included in more than one factorial. Each factorial was evaluated in a 1-year experiment (2002, 1999, 2003, 2001) with field trials at four to six locations in Germany under diverse agroecological conditions. The trials were evaluated in two-row plots using adjacent alpha designs with two to three replications. The hybrid performance of the crosses was recorded for GY in Mg/ha adjusted to 155 g/kg grain moisture and for GDMC in percentage. When combined, the four experiments can be regarded as an unbalanced incomplete factorial (Fig. 1).

Statistical analysis of the field data: The statistical analysis of the field data was presented in detail by Schrag et al. (2009). A mixed linear model was employed, in which main effects for years, locations and check varieties were treated as fixed. This allowed to account for performance differences between experiments. Genotypic effects, all interactions and block effects for trials, replications within trials and incomplete blocks within replications were treated as random. The residual error variance was assumed to be specific for each trial. All other block variances were assumed to be homogeneous. Mixed linear model analyses were performed with ASReml (Gilmour et al. 2002).

AFLP marker data: The inbred lines were assayed for AFLP markers with 20 primer combinations as described in detail by Schrag et al. (2006). After removing markers with more than 10% missing values and a gene diversity smaller than 0.2, the number of 970 high-quality markers remained for the analysis.

Gene expression data: For our '2k core set' of differentially expressed genes, we used a custom 2k micro-array (GEO Platform accession number: GPL22267) with 2232 oligonucleotide sequences (50–70 nt) of the maize oligonucleotide array project (University of Arizona, USA; http://www.maizearray.org). The oligonucleotides were synthesized by Ocimum Biosolutions (Ijsselstein, the Netherlands) and printed on poly-L-lysine-coated glass slides with a Microgrid II printer (BioRobotics, Boston, MA, USA). The selection of oligonucleotides for the 2k core set was based on 46k array expression data from Exp. 1 (GEO Platform accession number: GPL6438). The main fraction of oligonucleotides (1639) represents genes that showed differential expression between the parental genotypes of Exp. 1 and consistent association with hybrid performance for GY in cross-validation runs to estimate prediction accuracies for this trait (Frisch et al. 2010). In addition, the array contains partially overlapping fractions of genes that correlated with hybrid performance for GY (378), hybrid performance for GDMC (200) or mid-parent heterosis for GY (345), and 205 representatives of the six most overrepresented biological processes among genes correlated with hybrid performance for GY in Exp. 1 (Thiemann et al. 2010).

To obtain the plant material for the gene expression analysis, the parental inbred lines of the hybrids were grown for 7 days under controlled conditions. We did not use plants from the field experiment. For the parental lines of exps. 2, 3 and 4, four seedlings were grown, and for the parental lines of Exp. 1, five seedlings were grown, to obtain biological replicates. The temperature under which the seedlings were grown was 25°C for 16 h per day and 21°C for 8 h at night; the air humidity was 70%. The plants were grown with randomized plate position. The whole 7-day-old seedlings were sampled and frozen in liquid nitrogen. As we aimed for the identification of genotype-dependent expression differences, the biological replicates were pooled and homogenized prior to RNA extraction. Total RNA was isolated with mirVana miRNA isolation kit (Ambion, Thermo Scientific, Waltham, Massachusetts, USA). Two control lines, one from the dent and one from the flint pool, were included in each of the experiments if they were not part of the factorial anyway. For exps. 2 and 4, only 9 dent lines were included in the micro-array experiment, reducing the size of the factorials to 9 × 4, the total number of inbred lines to 48 and the total number of hybrids to 254, of which 230 were different. An interwoven loop design of two-colour

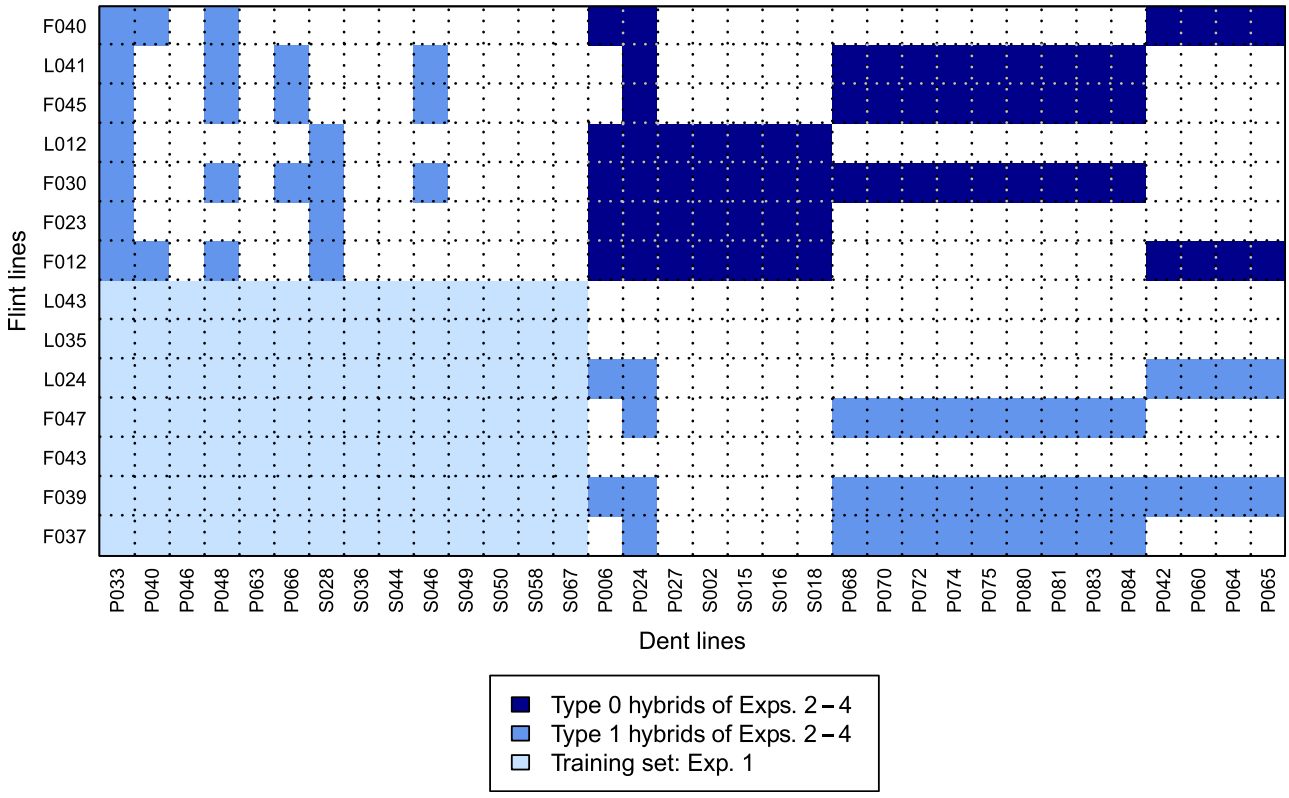

Fig. 1: The 34 dent and 14 flint lines of our data set and the hybrids generated from them. The display illustrates prediction of type 0 (dark blue) and type 1 (medium blue) hybrids of exps. 2–4 using the factorial of Exp. 1 (light blue) as training set [Color figure can be viewed at wileyonlinelibrary.com]

hybridizations striving for equal sampling and minimal distance between pairs of genotypes (Kerr and Churchill 2001) was developed for each factorial to minimize average variance. Sixty-three, 21, 57 and 21 hybridizations were performed for exps. 1, 2, 3 and 4 including 21, 15, 22 and 15 inbred lines, respectively. Both dyes (Cy3 or Cy5) were alternately used for each genotype to reduce systematic bias. RNA labelling and hybridizations were performed according to the protocols of the maize oligonucleotide array project (http://www.maizearray.org). The micro-arrays were scanned (AppliedPrecision ArrayWorx Scanner; Applied Precision Inc., Issaquah, Washington, USA), and the data were evaluated using the Software GENEPIX PRO 4.0 (Molecular Devices, Sunnyvale, CA, USA). The 2k micro-array was used for exps. 2–4. For Exp. 1, the raw files from the 46k micro-array were reduced to the oligos from the 2k micro-array. The data for exps. 1–4 have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002) and are accessible through GEO Series accession numbers GSE17754, GSE85286, GSE85287 and GSE85288, respectively.

The limma package (Ritchie et al. 2015) was applied for the tests. For each experiment, $n-1$ of the arrays were chosen as coefficients, with n being the number of lines investigated in that experiment and the coefficients describing the interconnections between all arrays. A background correction, a normalization within arrays, and a normalization between arrays was carried out. An ordinary least squares model was fit for each gene with the coefficients describing differences between the RNA sources hybridized on the corresponding arrays. These differences were tested for significance with a moderated F-test (Smyth 2004). A false discovery rate (Benjamini and Hochberg 1995) of 0.01 was used to adjust for multiple testing (Fu et al. 2012). The micro-array data were first analysed separately for each experiment. In total, 2122, 104, 542 and 140 genes of the 2k core set were found to be differentially expressed in exps. 1–4, respectively. In a second step, all micro-arrays of the four experiments were analysed together, resulting in 985 differentially expressed genes. For all differentially expressed genes, we calculated the expression level (log2 scale) of each gene for each inbred line from the coefficients from the linear model.

Transcriptome-based distances: The binary transcriptome-based distance D_B between two inbred lines i and j for n_g genes was calculated as:

$$D_B(i,j) = \sqrt{\frac{n_s(i,j)}{n_g}}, \tag{1}$$

with n_s (i, j) being the number of genes differentially expressed in inbred lines i and j (Frisch et al. 2010). Two genes were considered to be differentially expressed if the difference in their gene expression level exceeded a threshold of 1.3. The calculated transcriptome-based distances D_B were then used in a linear regression model:

$$\mathbf{y} = \beta_0 + \beta_1 D_B(u,v), \tag{2}$$

with $\boldsymbol{y}$ as the response vector consisting of the hybrid performance of the $i = 1 \ldots n$ hybrids, β_0 as a fixed intercept, β_1 as a regression coefficient and D_B (u,v) as a vector with the binary transcriptome-based distances between all $u = 1 \ldots n_u$ female and $v = 1 \ldots n_v$ male parents (Frisch et al. 2010). For a hybrid with parents u and v in the training set, D_B between the two parents was calculated and Eq. (2) was used to predict the performance $\hat{y}$ of the resulting hybrid.

We employed the binary transcriptome-based distance D_B, because in a previous analysis of Exp. 1, predictions with D_B showed greater correlations to the observed values than predictions with the Euclidean distance D_E, which is based on the quantitative expression levels (Frisch et al. 2010).

Ridge-regression model: To estimate the predictor effects, we used a linear model that relates the phenotype of a hybrid to the marker genotypes or mRNA transcription profiles that were observed in the two parental lines of the hybrid as described in Zenke-Philippi et al. (2016):

$$\begin{aligned} &\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{Fu} + \mathbf{Mv} + \mathbf{e} \\ &u_j \sim N(0, \sigma_f^2) \quad v_j \sim N(0, \sigma_m^2) \quad e_i \sim N(0, \sigma_e^2) \end{aligned} \tag{3}$$

$\mathbf{y}$ is the response vector consisting of the hybrid performance of the $i = 1 \ldots$ n hybrids, $\mathbf{1}$ is a vector of 1's and β_0 a fixed intercept. $\mathbf{u}$ and $\mathbf{v}$ are the vectors of the genetic effects of the $j = 1 \ldots p$ predictors in the female and male parent, respectively. The design matrices $\mathbf{F}$ and $\mathbf{M}$ consist of values $f_{i,j}$ and $m_{i,j}$ that code the observation of the j-th predictor at the i-th hybrid. For marker data, $f_{i,j}$ or $m_{i,j}$ is 1 if the AFLP band was observed in a parent and 0 otherwise. For mRNA transcripts, the design matrices contain the gene expression of gene j in the parents of the i-th hybrid. The columns of the design matrices $\mathbf{F}$ and $\mathbf{M}$ were normalized. For $\mathbf{F}$, the normalization was carried out according to Frisch et al. (2010):

$$f_{i,j} = \frac{o_{i,j}}{\max_{k \in \{1 \ldots s\}}(o_{k,j})}, \tag{4}$$

where $o_{i,j}$ are non-normalized original values for gene expression, and s is the number of parental lines used as female parents. For $\mathbf{M}$, the normalization was carried out analogously. The variances $\hat{\sigma}_f^2$, $\hat{\sigma}_m^2$, and $\hat{\sigma}_e^2$ were estimated by restricted maximum likelihood (REML). The effects $\hat{u}$ and $\hat{v}$ were obtained by solving the mixed model equations (Henderson 1984). With this model, the genotypic value of hybrids can be predicted as,

$$\hat{\mathbf{y}}^* = \mathbf{1}\hat{\beta}_0 + \mathbf{F}^*\hat{u} + \mathbf{M}^*\hat{v}, \tag{5}$$

where $\mathbf{F}^*$ and $\mathbf{M}^*$ are the design matrices for the predictors observed at the parental lines of the hybrid.

The components of $\mathbf{u}$ and $\mathbf{v}$ are additive main effects of the polymorphisms indicated by the respective design matrices. Genetically, they can be interpreted as effects for testcross performance if only the lines of the investigated experiment are considered. If the lines of the investigated experiment are considered as a representative sample from all lines of the opposite heterotic pool, the effects can be considered as estimates for the general combining ability. Technically an extension of the model to include the interaction effects between components of the parameter vectors of $\mathbf{u}$ and $\mathbf{v}$ is straightforward. By some authors these interactions are considered as dominance effects (Eq. 4 of Technow et al. 2012). The interaction effects could also be interpreted as effects for special combining ability. We chose not to include the interaction effects in the model, because it cannot be expected that interaction effects could be estimated with sufficient precision from the data set.

Assessment of prediction accuracy: For comparing the models, we determined prediction accuracies as the correlation $r(y, \hat{y})$ between predicted and observed hybrid performance. Some authors refer to this correlation as 'predictive ability' (*cf* Albrecht et al. 2011).

We used cross-validation, in which the data were split into training and validation sets on the basis of a random assignment. Cross-validation was carried out for 1000 replications, and in each run, the prediction accuracy was assessed. In addition, we validated the prediction accuracy by dividing the data into training and validation set on the basis of the four experiments.

For evaluating prediction accuracies, we distinguished three types of hybrids. For type 2 hybrids, both parental lines of an untested hybrid were part of the training set, for type 1 and type 0 hybrids, one or none, respectively. The structure of training and validation set for type 0 and type 1 hybrids for cross-validation within experiments is illustrated in Fig. 1 of Fu et al. (2012). Cross-validation across experiments is illustrated in Fig. 1 of Schrag et al. (2009). Validation using Exp. 1 as training set and exps. 2–4 as validation set is illustrated in Fig. 1.

Cross-validation within experiments was carried out to evaluate the prediction accuracy of transcriptome-based distance prediction following the scheme described by Fu et al. (2012).

The estimation set for evaluating the prediction accuracy for type 2 hybrids in Exp. 1 consisted of three randomly chosen flint and five randomly chosen dent lines and their hybrids, and the validation set consisted of the remaining part of the factorial. For exps. 2–4, we used three flint and three dent lines; for Exp. 3, five flint and two dent lines; and for Exp. 4, three flint and three dent lines and the corresponding hybrids as training set. The remaining part of the factorial was used as validation set. For the evaluation of the prediction of type 0 hybrids, ten and five, six and three, ten and four, and six and three flint and dent lines were used in exps. 1–4, respectively.

Cross-validation across experiments was carried out following the scheme of Schrag et al. (2009), in which seven flint and 17 dent lines were randomly chosen. Their marker genotype or transcription profiles, together with the hybrids that were actually available in the unbalanced data set, were used as training set and the remaining hybrids as validation set.

For validation on the basis of the four experiments, the subdivisions of the data set into training and validation sets are listed in Table 1.

Results

Cross-validation within experiments with transcriptome-based distances determined from the 2k core set of mRNA transcripts resulted in prediction accuracies $r(y, \hat{y})$ with large ranges and mean values around zero for exps. 2–4 for GY and GDMC (Fig. 2). Only for Exp. 1, which was used to define the 2k core set of genes, the average prediction accuracy reached a value of 0.63 for GY.

Cross-validation across experiments for assessing the prediction accuracies for GY and GDMC with ridge regression resulted in small differences between AFLPs and mRNA transcripts (Fig. 3). The average prediction accuracy for hybrid performance of type 1 hybrids was greater than $r(y, \hat{y}) = 0.6$ for both GY and GDMC. For type 0 hybrids, the prediction accuracies amounted to 0.5 for GY and 0.25 for GDMC. The variances of the prediction accuracies among the cross-validation runs were small.

For validation by splitting the data into training and validation set on the basis of the four experiments, and predicting hybrid performance with ridge regression, average prediction accuracies of around 0.6 were observed for type 1 hybrids for both traits for AFLPs as well as for mRNA transcripts (Table 1). For type 0 hybrids, the prediction accuracies were considerably smaller than 0.5 on average.

Table 1: Accuracy $r(y, \hat{y})$ of predicting hybrid performance for GY and GDMC with ridge regression using AFLPs and mRNA transcripts. One or two of the experiments were used as the training set and the remaining experiments were used as the validation set

Training set Exps.	Validation set Exps.	GY Type 0/Type 1	GDMC Type 0/Type 1
		$r(y, \hat{y})$	
Ridge regression with 1k AFLPs			
1	2,3,4	0.25/0.58	−0.19/0.27
2	1,3,4	0.36/0.71	0.14/0.74
3	1,2,4	0.33/0.51	0.36/0.69
4	1,2,3	0.22/0.57	−0.10/0.26
1,2	3,4	0.26/0.50	0.55/0.72
1,3	2,4	0.02/0.51	−0.09/0.66
1,4	2,3	0.15/0.64	0.02/0.40
2,3	1,4	0.56/0.55	0.28/0.59
2,4	1,3	0.34/0.65	0.07/0.62
3,4	1,2	0.54/0.66	0.53/0.66
Mean		0.30/0.59	0.16/0.56
Ridge regression with the 2k core set of mRNA transcripts			
1	2,3,4	0.30/0.56	−0.24/0.25
2	1,3,4	0.52/0.65	0.15/0.81
3	1,2,4	0.49/0.56	0.47/0.72
4	1,2,3	0.25/0.50	0.08/0.32
1,2	3,4	0.26/0.42	0.36/0.71
1,3	2,4	0.13/0.57	0.37/0.73
1,4	2,3	0.07/0.58	−0.07/0.34
2,3	1,4	0.69/0.63	0.50/0.57
2,4	1,3	0.60/0.61	0.02/0.74
3,4	1,2	0.50/0.69	0.77/0.73
Mean		0.38/0.58	0.24/0.59

Discussion

The efficient use of previously generated data as training set is essential for the successful implementation of hybrid prediction, as the assembly and data generation of training sets can be costly and time-consuming. We discuss approaches to re-use data from factorial crosses originally conducted to select among experimental hybrids as training set for the prediction of hybrid performance for GY and GDMC of related breeding material.

In general, the gene expression data showed a high level of statistical robustness with respect to the developmental stage of the plant. The prediction accuracies were high, even if the gene expression in early seedling stages might not be the same as in later developmental stages that determine agronomic performance, and even if the 7-day-old plants might not be in exactly the same developmental stage. This high level of robustness might be explained by gene expression patterns that stay constant within the developmental stages of a certain genotype but vary between genotypes.

Transcriptome-based distances

Employing the gene expression of a 46k micro-array for hybrid prediction with transcriptome-based distances resulted in prediction accuracies of up to $r(y, \hat{y}) = 0.8$ for GY of type 2 hybrids in cross-validation with the data set of Exp. 1 (Frisch et al. 2010). Creating a core set of genes with a good ability to predict hybrid performance could considerably reduce the resources required and therefore contribute to establishing the method in breeding programmes. This was our motivation to build a core set of 2k genes, which were selected on the basis of the association of differential gene expression and hybrid performance in Exp. 1.

Cross-validation within exps. 2–4 resulted in low prediction accuracies for type 2 hybrids (Fig. 2) and prediction accuracies near zero for type 0 hybrids (results not shown). These values cannot be regarded as useful for indirect selection. The results of the cross-validation consequently suggest that using a core set of genes for hybrid prediction with transcriptome-based distances is not effective.

Establishing the 2k core set was based on the association of differential gene expression with hybrid performance for GY and GDMC. As these two traits are negatively correlated, including genes related to both traits in the 2k core set could serve as an explanation for the low prediction accuracies. To investigate this hypothesis, we carried out an additional analysis, in which we divided the genes of the 2k core set into two subsets. One subset contained genes associated with GY, and the second contained genes associated with GDMC. Hybrid prediction with these subsets did not result in prediction accuracies that were greater than with the complete 2k core set (results not shown). Hence, having genes related to both traits in the 2k core set does

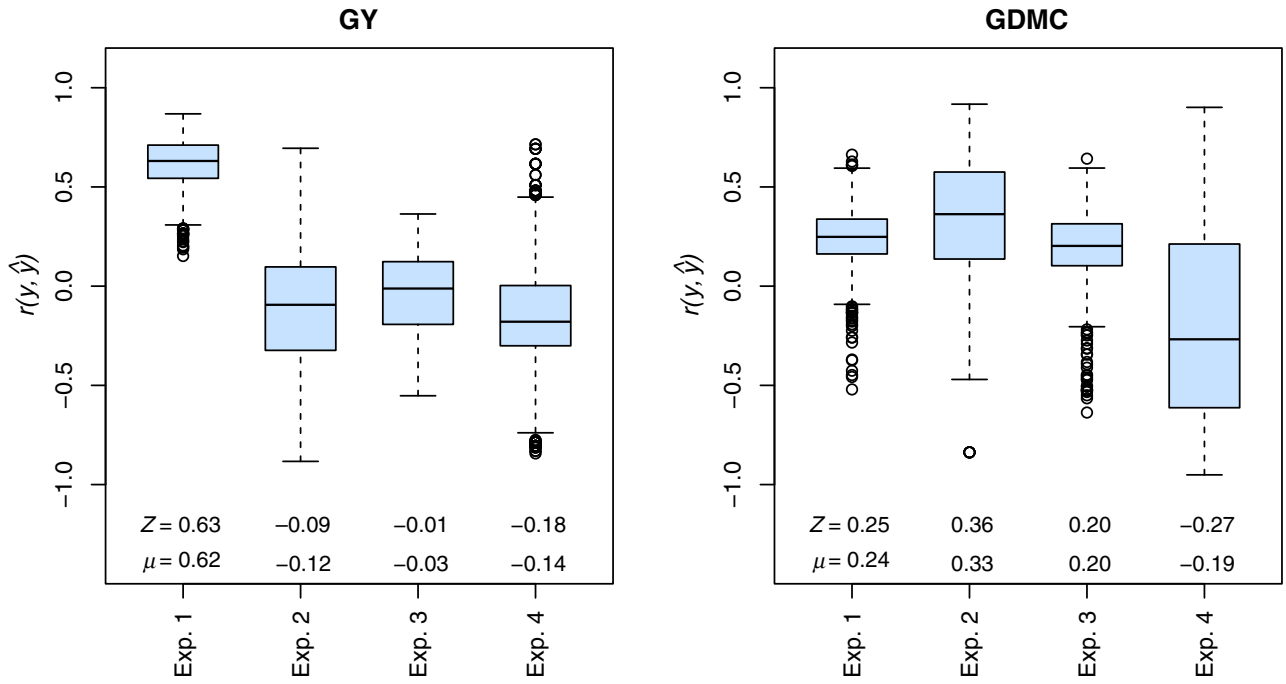

Fig. 2: Cross-validation within experiments for assessing the accuracy $r(y, \hat{y})$ of predicting hybrid performance for GY and GDMC with transcriptome-based distances using the 2k core set of mRNA transcripts. The boxplots show the distributions for 1000 cross-validation runs, μ are the arithmetic means and Z are the medians [Color figure can be viewed at wileyonlinelibrary.com]

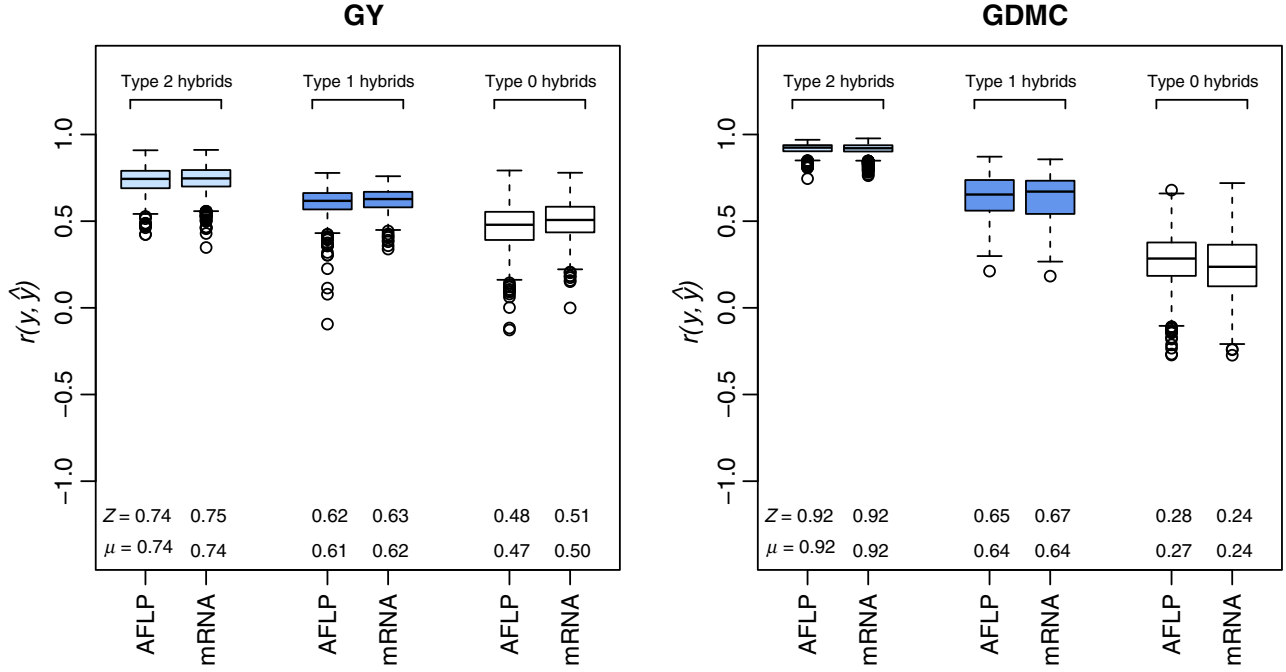

Fig. 3: Cross-validation across experiments for assessing the accuracy $r(y, \hat{y})$ of predicting hybrid performance for GY and GDMC with ridge regression using AFLPs and mRNA transcripts. The boxplots show the distributions for 1000 cross-validation runs, μ are the arithmetic means and Z are the medians [Color figure can be viewed at wileyonlinelibrary.com]

not seem to be the reason for the low prediction accuracies in our data set.

The cross-validation was complemented by a validation using one or two experiments as training set for predicting the performance of the type 0 and type 1 hybrids of exps. 1–4 with transcriptome-based distances determined with the 2k core set. The correlation between observed and predicted hybrid performance was close to zero for both traits (results not shown).

To summarize, neither cross-validation within experiments nor validation across experiments convincingly demonstrated that a core set of genes determined in one experiment can be used for hybrid prediction with transcriptome-based distances in other experiments. In particular, it was not possible with the 2k core set to reach the high prediction accuracies that were observed with the full 46k micro-array for type 0 hybrids in earlier studies (Frisch et al. 2010, Fu et al. 2012). We therefore conclude that

preselecting a core set of genes is not a useful strategy for saving resources in hybrid prediction with transcriptome-based distances.

Ridge regression

The transcriptome-based distance approach attempts to identify genes of which differential gene expression in parental lines is associated with high hybrid performance. Even if the idea of identifying 2k genes of which the differential gene expression is functionally related to hybrid performance for GY and GDMC was not successful with our data set, the gene expression data of the 2k core set can be employed in a ridge-regression model in the sense of marker data (Zenke-Philippi et al. 2016). In this case, similar expression of a certain gene in two parental lines can be regarded as an indicator for a common genomic region, and the prediction accuracies of ridge-regression models with mRNA transcription profiles and AFLP markers can be compared.

Our data set can be regarded as an 'incomplete factorial' (see Fig. 1 of Schrag et al. 2009, for a graphical illustration), and the 1k AFLPs or 2k mRNA transcripts can be used as predictors for ridge regression. This allows cross-validation to investigate hybrid prediction with unbalanced data, employing the cross-validation procedure described by Schrag et al. (2009). In cross-validation, the average prediction accuracy for performance of type 1 hybrids was greater than $r(y, \hat{y}) = 0.6$ for both traits, irrespective of whether AFLPs or mRNA transcription profiles were used as predictors in the ridge-regression approach (Fig. 3). For type 0 hybrids, the prediction accuracies were around 0.5 for GY and 0.25 for GDMC.

To complement the cross-validation, we used the data of either one or two of the four experiments as training set and predicted the hybrid performance of the remaining factorials (Table 1). Prediction of exps. 2–4 using Exp. 1 as training set is illustrated in Fig. 1. For type 1 hybrids, a mean prediction accuracy of about 0.6 was reached for both traits. For type 0 hybrids, prediction accuracies that were on average smaller than 0.5 were observed, with small differences between AFLPs and mRNA transcripts. This confirms that the ridge-regression approach, which resulted in high prediction accuracies for the balanced data of Exp. 1 (Zenke-Philippi et al. 2016), has the potential to be successfully applied with unbalanced data sets.

The motivation for using transcriptome data in hybrid prediction is that mRNA transcripts might be able to capture gene interactions and epistatic effects that cannot be captured by DNA markers. However, prediction accuracies of the ridge-regression model reached similar values for mRNA transcripts and AFLP data (Fig. 3). From this we conclude that, with our data set, the mRNA transcripts have about the same level of information content as ALFPs, and the confirmation of the hypothesis that additional information content of mRNA transcripts can be used to increase prediction accuracy remains open for further research.

For the cross-validation within the unbalanced data set, 17×7 parental lines were selected as parents of the training set (following Schrag et al. 2009). On average, the training set consisted of 58 hybrids obtained from crosses of these parental lines. Technow et al. (2014) reported that the prediction accuracy for type 2 and type 1 hybrids increased when the size of the training set increased from 300 to 450 hybrids. For type 0 hybrids, a plateau of prediction accuracy was reached at a training set size of 300 hybrids. This indicates that increasing the size of the training set compared to our data might further improve prediction accuracies. Nevertheless, reliable and stable prediction results could already be achieved in the present study with relatively low numbers of hybrids in the training set. Close relatives in training and validation set (Albrecht et al. 2011) and a good resemblance of the validation set and the training set (Albrecht et al. 2014) are prerequisites for successful predictions. We conclude that with relatively narrow breeding pools, as in our experiment, hybrid prediction with ridge regression is promising with small training sets. This enables hybrid prediction even in situations where only limited resources are available.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (grants no. FR 1615/4-1, ME 2260/5-1, SCHO 764/6-1).

Conflict of interest

The authors declare that they have no competing interests.

References

Albrecht, T., V. Wimmer, H. J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, and C. C. Schön, 2011: Genome-based prediction of testcross values in maize. Theor. Appl. Genet. **123**, 339—350.

Albrecht, T., H. J. Auinger, V. Wimmer, J. O. Ogutu, C. Knaak, M. Ouzunova, H.-P. Piepho, and C. C. Schön, 2014: Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. **127**, 1375—1386.

Andorf, S., J. Selbig, T. Altmann, K. Poos, H. Witucka-Wall, and D. Repsilber, 2010: Enriched partial correlations in genome-wide gene expression profiles of hybrids (*A. thaliana*): a systems biological approach towards the molecular basis of heterosis. Theor. Appl. Genet. **120**, 249—259.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. **57**, 289—300.

Dan, Z., J. Hu, W. Zhou, G. Yao, R. Zhu, Y. Zhu, and W. Huang, 2016: Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). Sci. Rep. **6**, 732.

Edgar, R., M. Domrachev, and A. E. Lash, 2002: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**, 207—210.

Frisch, M., A. Thiemann, J. Fu, T. A. Schrag, S. Scholten, and A. E. Melchinger, 2010: Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theor. Appl. Genet. **120**, 441—450.

Fu, J., K. C. Falke, A. Thiemann, T. A. Schrag, A. E. Melchinger, S. Scholten, and M. Frisch, 2012: Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. Theor. Appl. Genet. **124**, 825—833.

Gilmour, A. R., B. R. Cullis, S. J. Welham, and R. Thompson, 2002: ASReml Reference Manual. Release 1.0. VSN International, Hempstead.

Henderson, C., 1984: Applications of Linear Models in Animal Breeding. University of Guelph, Guelph.

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

Kerr, M. K., and G. A. Churchill, 2001: Experimental design for gene expression microarrays. Biostatistics **2**, 183—201.

Maenhout, S., B. De Baets, and G. Haesaert, 2010: Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theor. Appl. Genet. **120**, 415—427.

Massman, J. M., A. Gordillo, R. E. Lorenzana, and R. Bernardo, 2013: Genomewide predictions from maize single-cross data. Theor. Appl. Genet. **126**, 13—22.

Piepho, H. P., 2009: Ridge regression and extensions for genomewide selection in maize. Crop Sci. **49**, 1165—1176.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer, and A. E. Melchinger, 2012: Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. **44**, 217—220.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, 2015: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43**, e47.

Schrag, T. A., A. E. Melchinger, A. P. Sorensen, and M. Frisch, 2006: Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor. Appl. Genet. **113**, 1037—1047.

Schrag, T. A., J. M. Möhring, H. P. Maurer, B. S. Dhillon, A. E. Melchinger, H.-P. Piepho, A. P. Sorensen, and M. Frisch, 2009: Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theor. Appl. Genet. **118**, 741—751.

Smyth, G. K., 2004: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. **3**, 3.

Steinfath, M., T. Gärtner, J. Lisec, R. C. Meyer, T. Altmann, L. Willmitzer, and J. Selbig, 2010: Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theor. Appl. Genet. **120**, 239—247.

Technow, F., C. Riedelsheimer, T. A. Schrag, and A. E. Melchinger, 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., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer, and A. E. Melchinger, 2014: Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics **197**, 1343—1355.

Thiemann, A., J. Fu, T. A. Schrag, A. E. Melchinger, M. Frisch, and S. Scholten, 2010: Correlation between parental transcriptome and field data for the characterization of heterosis in *Zea mays* L. Theor. Appl. Genet. **120**, 401—413.

Xu, S., Y. Xu, L. Gong, and Q. Zhang, 2016: Metabolomic prediction of yield in hybrid rice. The Plant Journal, **88**(2), 219—227.

Zenke-Philippi, C., A. Thiemann, F. Seifert, T. Schrag, A. E. Melchinger, S. Scholten, and M. Frisch, 2016: Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles. BMC Genom. **17**, 262.

Zhao, Y., M. F. Mette, and J. C. Reif, 2015: Genomic selection in hybrid breeding. Plant Breeding **134**, 1—10.