Aus dem Institut für Pflanzenbau und Pflanzenzüchtung II
der Justus-Liebig-Universität Gießen
Professur für Biometrie und Populationsgenetik
Prof. Dr. Matthias Frisch

Prediction of hybrid performance in maize with transcriptome data

Dissertation zur Erlangung des akademischen Grades eines

Doktors der Agrarwissenschaften

- Dr. agr. -

im Fachbereich
Agrarwissenschaften, Ökotrophologie und Umweltmanagement
der Justus-Liebig-Universität Gießen

vorgelegt von

Carola Zenke-Philippi
aus Korbach, Hessen

Gießen, im April 2017

Contents

[1] 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 Genomics* **17**:262.

[2] Zenke-Philippi, C., M. Frisch, A. Thiemann, F. Seifert, T. Schrag, A. E. Melchinger, S. Scholten, and E. Herzog (2017) Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme. *BMC Genomics* **17**:262. *Plant Breeding* **136**:331–337

Abbreviations

AFLP	Amplified fragment length polymorphism
BLUP	best linear unbiased prediction
DNA	desoxyribonucleic acid
GCA	general combining ability
GDMC	grain dry matter content
GWP	genome-wide prediction
GY	grain yield
LD	linkage disequilibrium
MAS	marker-assisted selection
mRNA	messenger ribonucleic acid
QTL	quantitative trait locus
REML	restricted maximum likelihood
RFLP	restriction fragment length polymorphism
RR-BLUP	ridge regression BLUP
SCA	specific combining ability
SNP	single nucleotide polymorphism
SSR	simple sequence repeat

Chapter 1

General introduction

Since the beginning of hybrid production in maize (Shull 1908), a major focus of breeding programs has been to identify the most promising parental inbred lines from all breeding material available. This can be done by conducting field trials to generate phenotypic data and using relationship coefficients to predict the performance of untested lines or hybrids (Crossa et al. 2010). The doubled haploid technology enables breeders to create large numbers of maize inbred lines very fast (Smith et al. 2008), exacerbating the question which of them should be tested in the field. With constant or even increasing phenotyping costs (Desta and Ortiz 2014) and steadily decreasing genotyping costs (Zhao et al. 2015), it suggests itself to use the genotype of a plant, which can be evaluated in a very early developmental stage, to predict its phenotype.

For prediction with genetic markers (Figure 1.1), genotypic and phenotypic data are collected for a training population (training set). Marker effects are estimated with a statistical model to describe the relationship between marker data and phenotype with genotypic data as predictors for the phenotype. The individuals of a new population (validation set) are then genotyped and the phenotypes are predicted with the model that was established previously. In an actual breeding program, promising genotypes would be selected based on these predictions. The importance of field trials for the

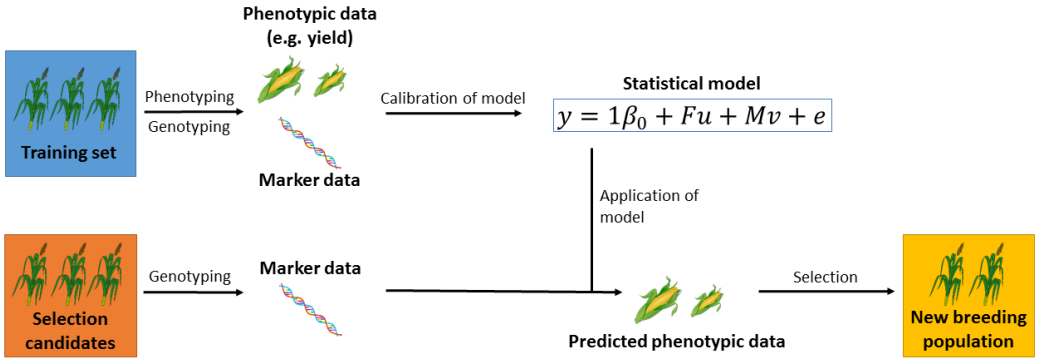

Figure 1.1. Prediction of performance with genetic markers. Individuals in the training set are genotyped and phenotyped. The phenotype ist linked to the genotype with a statistical model. The selection candidates are genotyped and their phenotype is predicted with this model. The best-performing individuals are then transferred to the next stage of the breeding program.

evaluation of genotypes is thereby expanded to updating the model used for prediction with genomic data.

In the prediction of hybrid performance, not the hybrids themselves are genotyped but genotypic data are collected for the parental inbred lines instead. Testcrosses can then either be used to estimate general combining ability (GCA) and specific combining ability (SCA) of those inbred lines or hybrid performance can be targeted directly (Figure 1.2). For the latter approach, parental lines from two distinct heterotic pools, *e.g.*, Flint and Dent, are crossed, the phenotypes of the crosses are evaluated, and a statistical model links the parental genotypes to hybrid performance. Untested crosses can then be predicted from the parental genotypes and the most promising combinations can be tested in the field.

In order to evaluate the quality of the prediction, phenotypic data are also collected for the predicted individuals in the validation set and the correlation between these real phenotypic data and the predicted values is calculated.

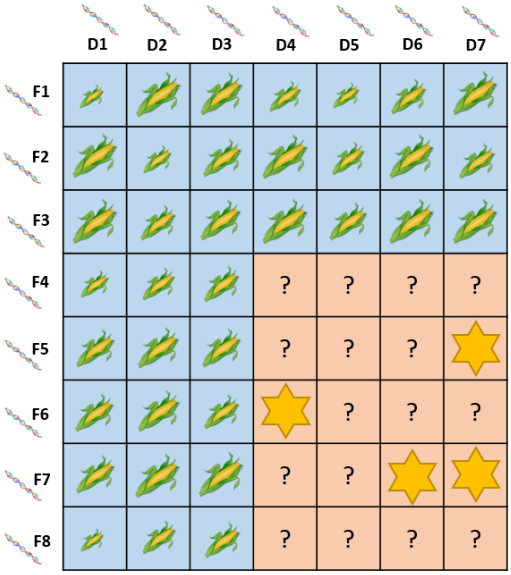

Figure 1.2. Example for hybrid prediction in a set of factorial crosses. All inbred lines are genotyped. Hybrids from the training set (blue) are phenotyped and the statistical model is calibrated based on their performance. The performance of the hybrids in the validation set (orange) is then predicted based on the parental genotypes and the most promising hybrids (marked with stars) are selected for the next stage of the breeding program.

This correlation is called prediction accuracy. In order to successfully identify the most promising individuals, the prediction accuracy has to be as high as possible. It is dependent on the statistical model used for the predictions, on marker type and density, on sizes and compositions of the training and the validation set and their relatedness to one another, on heritability and genetic architecture, on gene effects, on extent and distribution of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL) and on the trait to be predicted (Dan et al. 2016; Desta and Ortiz 2014; Windhausen et al. 2012). In my thesis, I focused on the importance of statistical models, markers types and densities, and the relatedness between training set and validation set for the accuracy of hybrid predictions.

Statistical models for hybrid prediction

Hybrids are created to exploit the heterosis that occurs if two sufficiently different inbred lines are crossed (Shull 1908). While the underlying genetic mechanism for heterosis is subject to an ongoing debate (*cf.* Chen 2013), heterotic traits are generally thought to be controlled by many loci, each with only a small effect on the target trait (infinitesimal model, *cf.* Lorenz et al. 2011). Researchers and breeders are therefore confronted with the situation that the number of markers p (*i.e.*, predictors) exceeds the number of genotypes n (*i.e.*, observations) by far. This is a problem for common linear regression since it only allows $n-1$ predictors in the model. Marker-assisted selection (MAS) emerged in the 1990s to identify markers with a significant influcence on the target trait (Lande and Thompson 1990). It aims to identify relevant loci by investigating the association with the target trait. Only significant markers are selected. Since the selection of the significance threshold is arbitrary, MAS usually leads to the over-estimation of the effects of the selected markers while the influence of all other loci is neglected (Meuwissen et al. 2001). This is called the Beavis effect (Xu 2003). The underlying genetic architecture of polygenic traits like grain yield is therefore not captured

very well by MAS. Under the infinitesimal model, statistical approaches are needed which include all markers in the model.

A different way of making use of marker information was proposed as GBLUP by Bernardo (1994) who used a relationship matrix estimated from genetic markers rather than from pedigree information like in best linear unbiased prediction (BLUP) as often used in animal breeding (Henderson 1975). Meuwissen et al. (2001) then suggested to estimate all marker effects simultaneously. A simple model for such an estimation is ridge regression BLUP (RR-BLUP). RR-BLUP uses a shrinkage factor λ to shrink all marker effects equally towards zero, assuming random marker effects drawn from a normal distribution with a common variance for all markers. This leads to homoscedastic marker variances (Meuwissen et al. 2001). If λ is set to error variance σ_e / genetic variance σ_m at a marker m, the estimates of RR-BLUP are equivalent to those obtained with GBLUP (De Vlaming and Groenen 2015). Variances can be estimated with the restricted maximum likelihood algorithm (REML).

Ridge regression is often criticized for its assumption of homoscedastic marker variances, reflecting a genetic architecture with genetic effects evenly spread throughout the genome. This would lead to reduced prediction accuracies for traits with genetic variance present at few and absent at many loci (Meuwissen et al. 2001). Bayesian models have been proposed to incorporate heteroscedastic marker variances (Meuwissen et al. 2001). However, these methods suffer from their high computational demands (Lorenz et al. 2011) and pose the difficulty of choosing an appropriate prior distribution (Piepho 2009). Additionally, most studies with empirical data showed no significant improvement in prediction accuracy if heterocesdastic instead of homoscedastic marker variances were assumed (*cf.* Heslot et al. 2012; Wimmer et al. 2013). Ridge regression is therefore often recommended for genomic prediction since it is thought to be robust and reliable (Zhao et al. 2015). In addition, new ridge regression methods were developed recently which combine the lower computational demands of ridge regression com-

pared to Bayesian methods with the possibility to include heteroscedastic marker variances (Shen et al. 2013; Hofheinz and Frisch 2014).

Hybrid prediction with ridge regression models has, for example, been done for maize (Technow et al. 2012; Massman et al. 2013; Zenke-Philippi et al. 2016), sunflower (Reif et al. 2013), barley (Philipp et al. 2016), sugar beet (Würschum et al. 2013) and wheat (Zhao et al. 2013c,b, 2014). However, most studies focused on the prediction of testcross performance or crosses from biparental populations rather than on the prediction of factorial hybrids, *e.g.*, in maize (Lorenzana and Bernardo 2009; Albrecht et al. 2011; Riedelsheimer et al. 2012; Windhausen et al. 2012; Zhao et al. 2012a; Riedelsheimer et al. 2013; Zhao et al. 2013a; Albrecht et al. 2014; Lehermeier et al. 2014), canola (Jan et al. 2016), rye (Wang et al. 2014; Auinger et al. 2016), and sugar beet (Hofheinz et al. 2012).

Alternative statistical methods for hybrid prediction in maize include partial least squares regression, support vector machine regression (Fu et al. 2012), and genetic distances. Genetic distances measure the difference between two inbred lines based on their marker profile and aim to predict the performance of the resulting hybrid from this difference. The theoretical background of this approach is the notion that for heterosis to occur, a certain level of genetic dissimilarity between the two parental lines is necessary and that the level of heterosis increases with increasing genetic distance between the parental lines (Lanza et al. 1997; Marsan et al. 1998; Chen 2013). Genetic distances have been used for hybrid prediction since the advent of genetic markers, *e.g.*, in oilseed rape (Diers et al. 1996), sorghum (Jordan et al. 2003), sunflower (Cheres et al. 2000), and maize (Lanza et al. 1997). In these studies, different genetic markers and different measures for the genetic distances were examined. In general, the achieved prediction accuracies were significant but not sufficient for reliable hybrid prediction.

Distance measures between parental lines can also be estimated from mRNA transcript abundance levels, *i.e.*, expression levels of genes (Frisch

et al. 2010). First, genes in which differential expression in the training set is associated with the target trait have to be identified. The binary distance D_B between two inbred lines is then estimated from the number of these genes that are differentially expressed between the two lines. Only genes whose expression difference exceeds a certain threshold are included in the binary distance D_B, but all those genes are then included with the same weight. This corresponds well to the infinitesimal model (Frisch et al. 2010). Binary trancriptome-based distances D_B were well-suited to separate Dent and Flint lines in two pools, so they were used as predictors for hybrid performance of a set of 98 factorial crosses in maize (Fu et al. 2012). They were inferior to multiple linear regression, partial least squares regression, and support vector machine regression for predicting grain yield of hybrids with testcross data for both parents but superior for the prediction of hybrids for which no parental testcross data were available (Fu et al. 2012). The question remained whether the method would perform equally well in larger factorials, and with fewer genes to select from. In my thesis, I addressed this issue as well as the question whether it would be possible to transfer a core set of genes for which differential expression was correlated with the trait of interest in one set of factorial crosses to another set of factorial crosses.

Hybrid prediction with genomic, transcriptomic and metabolomic data

Genetic markers like restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs) or single nucleotide polymorphisms (SNPs) are the basis for hybrid prediction with RR-BLUP in which effects are estimated for each marker. Especially SNP markers have the advantage that they are relatively cheap and easy to generate with modern next-generation sequencing methods. In maize, a very high genome coverage can be achieved with the 600 k SNP chip (Unterseer et al. 2014).

Since marker data seem to capture mainly relatedness between individuals rather than additional information, metabolomic data have recently been used for hybrid prediction in maize (Riedelsheimer et al. 2012; Feher et al. 2014) and rice (Dan et al. 2016; Xu et al.). Genomic and metabolomic distances were shown to be only weakly correlated, thus providing access to connected, but nevertheless different layers of infomation (Riedelsheimer et al. 2012). In a different statistical framework of prediction with support vector machine regression, partial least squares regression, and distance-based methods, mRNA transcription profiles were identified as promising predictors for hybrid performance in maize (Frisch et al. 2010; Fu et al. 2012). Transcriptomic data are biologically located between genetic and metabolic information since they are the template for the translation of genes into proteins. Compared to genomic data, they have the advantage that they do not rely on LD between marker and gene and are therefore better suited for the prediction across heterotic pools (Frisch et al. 2010). The low correlation of transcriptome-based distances and genetic distances suggests that the mRNA expression profiles do indeed carry additional information (Frisch et al. 2010). Our goal was to investigate whether transcriptomic data can successfully be used for hybrid prediction in genomic selection models and, if so, if the number of data points required for a successful prediction is equal for mRNA transcription profiles and AFLP markers.

Evaluation of prediction accuracy

The prediction accuracy of a statistical model can be assessed with cross-validation (*cf.* Schrag et al. 2009) in which the data set is randomly divided into a training set and a validation set. The statistical model is calibrated based on the training set and phenotypic values are predicted for the validation set. The correlations between actual and predicted phenotypic performance of hybrids in the validation set are recorded as the prediction accuracies. In order to account for random sampling effects, the cross-validation

procedure is usually repeated many times. A different validation approach is validation with independent factorials (*cf.* Zenke-Philippi et al. 2017). Here, a whole factorial is used as the training set, and another factorial (or several others) as the validation set. This approach more closely resembles a situation in which predictions are made with data from previous breeding cycles and is therefore closer to reality: Genomic prediction is appealing because via genotyping the plants at an early developmental stage, generation intervals can be shortened tremendously. However, this application of genomic selection can only be successful if phenotypic and genotypic data from previous generations can be used.

In hybrid prediction, division of the complete data set into training and validation set leads to different situations regarding the parents of a hybrid that is to be predicted: They can either both be included in the training set (type 2 hybrid), or only one parent is part of the training set but the other one is not (type 1 hybrid), or none of the parental lines is included in the training set (type 0 hybrid) (Fu et al. 2012). If the prediction accuracies for the hybrids are separated based on the hybrid type, it becomes apparent that type 2 hybrids can generally be predicted quite reliably, whereas the prediction of type 0 hybrids is challenging (Fu et al. 2012). Apparently, the prediction accuracy heavily depends on the genetic relatedness between the training and the validation set. Only if the validation set resembles the training set closely, high prediction accuracies can be achieved (Albrecht et al. 2014). This means that if the performance of a hybrid from two parental lines is to be predicted, it is crucial that either both parental lines or close relatives are included in the training set. It would therefore be of great interest to breeders if a method could be established that was able to also predict type 0 hybrids reliably.

Objectives

The main goal of my thesis research was to investigate the efficiency of mRNA transcription profiles for hybrid prediction of maize in a data set originating from an ongoing maize breeding program. Specifically, my objectives were to:

(1) compare the prediction accuracy of AFLP markers with that of mRNA transcription profiles for hybrid prediction,

(2) investigate the number of mRNA transcripts required for accurate prediction,

(3) investigate the transferability of a core set of genes correlated to the trait of interest from one set of factorial crosses to another,

(4) compare prediction accuracy of transcriptome-based distances with ridge regression approaches, and

(5) compare the prediction accuracies of these methods with cross-validation vs. independent validation.

Chapter 2

Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles [1]

[1] 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 Genomics* **17**:262.

Zenke-Philippi *et al. BMC Genomics* (2016) 17:262
DOI 10.1186/s12864-016-2580-y

BMC Genomics

RESEARCH ARTICLE Open Access

Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles

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

Abstract

Background: Ridge regression models can be used for predicting heterosis and hybrid performance. Their application to mRNA transcription profiles has not yet been investigated. Our objective was to compare the prediction accuracy of models employing mRNA transcription profiles with that of models employing genome-wide markers using a data set of 98 maize hybrids from a breeding program.

Results: We predicted hybrid performance and mid-parent heterosis for grain yield and grain dry matter content and employed cross validation to assess the prediction accuracy. Prediction with a ridge regression model using random effects for mRNA transcription profiles resulted in similar prediction accuracies than employing the model to DNA markers. For hybrids, of which none of the parental inbred lines was part of the training set, the ridge regression model did not reach the prediction accuracy that was obtained with a model using transcriptome-based distances.

Conclusion: We conclude that mRNA transcription profiles are a promising alternative to DNA markers for hybrid prediction, but further studies with larger data sets are required to investigate the superiority of alternative prediction models.

Background

The resources for field trials in a hybrid breeding program are restricted and only a fraction of all possible hybrids that could potentially be generated by crossing the inbred lines developed in each cycle of the breeding program can be phenotypically evaluated. The principle of hybrid prediction is to link the performance of phenotypically evaluated hybrids to predictors, such as DNA markers or mRNA transcription profiles, that can be assessed in the parental lines of the hybrids. For each state of the predictor, its effect on the phenotype is estimated and these effects are then used to predict the performance of new hybrids.

DNA markers were employed for hybrid prediction in maize and proved to be superior to prediction approaches based solely on pedigree and phenotypic data [1–5]. First results on using the mRNA transcriptome for hybrid prediction with distance-based approaches [6] or regression-based approaches [7] showed promising results. Genome-wide prediction of general combining ability (GCA) or testcross performance [8–10] can be regarded as a special case of hybrid prediction where one parental component (the tester) is known and the effects of the predictors assessed at the second parental component are used for hybrid prediction. In this context, first results of using metabolites as predictors were successful [10] but showed a lower prediction accuracy than using SNP markers as predictors.

Two important situations can be distinguished in hybrid prediction. The first is that the parental lines of a potential hybrid have already been evaluated for testcross performance with other lines of the breeding pool. If such testcross data are available for both parental lines but the hybrid itself is not yet generated, then we refer to the hybrid as type 2 hybrid (testcross data for two parents available). The second situation is that the parental

*Correspondence: matthias.frisch@uni-giessen.de
[1]Institute of Agronomy and Plant Breeding II, Justus Liebig University, 35392 Giessen, Germany
Full list of author information is available at the end of the article

lines are entirely new and have not yet been evaluated in any test cross. Such hybrids are referred to as type 0 hybrids (testcross data for none of the parents available). The application of ridge regression models in combination with mRNA transcription profiles for the prediction of type 0 and type 2 hybrids has not yet been investigated.

The goal of our study was to investigate the prediction of grain yield and grain dry matter content using field data of 98 maize hybrids and AFLP (amplified fragment length polymorphism) marker data as well as mRNA transcription profiles of their 21 parental lines. In particular, our objectives were to (1) assess the accuracy of predicting hybrid performance with a random effects model using mRNA transcription profiles, (2) investigate the number of mRNA transcripts that are required for precise hybrid prediction, (3) compare the prediction accuracy of a random model employing mRNA with the prediction accuracy obtained with AFLP markers as well as the prediction accuracy of previously published approaches, and (4) draw conclusions on possible application in breeding programs for prediction of hybrid performance and heterosis of type 2 and type 0 hybrids.

Methods

Field data

The field data were presented in detail by [11], where the factorial we used for the present study was referred to as Experiment 1. Here we give only a brief overview. Seven flint and 14 dent elite inbreds developed in the maize breeding program of the University of Hohenheim were used as parental inbreds for $98 = 7 \times 14$ factorial crosses between both groups of inbreds. The inbreds comprised eight dent lines with Iowa Stiff Stalk Synthetic background and six with Iodent background. Four flint lines had a European Flint background and three a Flint/Lancaster background.

The factorial crosses were evaluated in 2002 at six agroecologically diverse locations in Germany (Bad Krozingen, Eckartsweier, Hohenheim, Landau, Sunching, Vechta). The trials were evaluated in two-row plots using α designs with two to three replications. Hybrid performance for grain yield was assessed in Mg ha^{-1} adjusted to 155 g kg^{-1} grain moisture and for grain dry matter content in percent. The mean hybrid performance for grain yield was 11.72 Mg ha^{-1} and for grain dry matter content 67.7 % with broad sense heritabilities of 0.80 (grain yield) and 0.91 (grain dry matter content). The GCA (general combining ability) and SCA (specific combining ability) variance components as well as their interactions with the locations were significantly different from zero ($\alpha = 0.05$) for both traits. The ratios of SCA:GCA variance components were 1.12 (grain yield) and 0.42 (grain dry matter content).

AFLP marker data

The inbred lines were assayed for AFLP markers with 20 primer combinations as described in detail by [11]. 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

Five seedlings of each of the 21 diverse dent and flint maize inbred lines were grown for seven days under controlled conditions (25 °C 16 h day, 21 °C 8 h night, 70 % air humidity). Whole seedling tissue of five biological replicates was frozen in liquid nitrogen, homogenized, and pooled before target labeling and hybridization. Total RNA was isolated, precipitated with LiCl (8M) and purified with the "NucleoSpin RNA Clean-up Kit" (Macherey-Nagel, Düren, Germany) and used to synthesize aminoallyl-labeled RNA (aaRNA) following the "Amino Allyl MessageAmp aRNA" System protocol (Applied Biosystems/Ambion, Austin, USA). aaRNA was coupled with fluorescence dyes Cy3 or Cy5 (GE Healthcare, Chalfont St. Giles, UK) and purified with RNeasy MinElute Kit (Qiagen, Hilden, Germany). The 46k array from the maize oligonucleotide array project [12], GEO platform GPL6438 was hybridized according to the manufacturer instructions. The micro-arrays were scanned (AppliedPrecision ArrayWorx Scanner, Applied Precision Inc., USA) and data was evaluated using GenePix Pro 4.0 (Molecular Devices, Sunnyvale, USA). For the micro-array experiment, an interwoven loop design [13] was applied. It resulted in 63 hybridizations of dent and flint lines by sampling each dent line five times and each flint line eight times.

For experimental validation of the micro-array experiment, two genes in eight different lines were evaluated by Quantitative RT-PCR, essentially in accordance with the micro-array data. For the validation of micro-array expression pattern copy DNA from total RNA of the inbred lines S028, F047, L024, S058, S044, PO33, L043, and F039 was produced with Superscript II (Thermo Fisher Scientific) according to the manufacturer's protocol. Quantitative RT-PCR was conducted for the genes GRMZM2G057829, GRMZM2G021406 and the actin gene (accession number JO1238) with the primer pairs 5-'GAAACCATAACAGACGCGTCATCACATC-3'/5'-CAGCAGGAGCAGAAGAGGGAAAAG-3', 5'-TAGGC TGCTATTTGGGCACTTAGTTTTAC-3'/5'-CCAGTAC GGGAGACATGTAGAGTTC-3', and 5'-TCCTGACACT GAAGTACCCGATTGA-3'/5'-CGTTGTAGAAGGTGT GATGCCAGTT-3', respectively, with the iCycler iQ (BIORAD, Germany) and the qPCR MasterMix Plus for SYBR Green I (Reference: RT-SN2X- 03 + NRFL, Eurogentec, Seraing, Belgium) in triplicates. Actin expression values were used for data normalization before relative

expression levels between lines were calculated. The micro-array data have been deposited in Gene Expression Omnibus (GEO) under the series accession GSE17754.

The gene-oriented probes together with spike-in probes were tested for statistically significant differential expression across all comparisons with a moderated F-test and subsequently with a nested F-test for each comparison of parental lines. The *limma* package [14] was applied for the tests. A false discovery rate [15] of 0.01 for all genes showing a fold change of at least 1.3 and log-2 expression intensity of at least 8 was used to detect significant differential expression between inbred lines [16]. In total, 10,810 genes were differentially expressed in at least one pair of parental lines of the factorial crosses. We refer to this set of predictors as 'mRNA10k', random samples of 1000 out of the 10,810 genes are referred to as 'mRNAr1k'.

Prediction model

To estimate the predictor effects, we used a linear model that relates the phenotype of a hybrid to the marker genotype or mRNA transcription profiles that were observed in the two parental lines of the hybrid:

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{F}\mathbf{u} + \mathbf{M}\mathbf{v} + \mathbf{e} \tag{1}$$

$$u_j \sim \mathrm{N}\left(0, \sigma_f^2\right) \qquad v_j \sim \mathrm{N}\left(0, \sigma_m^2\right) \qquad e_i \sim \mathrm{N}\left(0, \sigma_e^2\right)$$

$\mathbf{y}$ 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 jth predictor at the ith 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, the design matrices contain the gene expression of gene j in the parents of the ith hybrid, the columns of the design matrices $\mathbf{F}$ and $\mathbf{M}$ were normalized. For $\mathbf{F}$ the normalization was carried out according to

$$f_{i,j} = \frac{o_{i,j}}{\max\limits_{k \in \{1 \ldots s\}} \left(o_{k,j}\right)} \tag{2}$$

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). Then the effects $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ were obtained by solving the mixed model equations [17].

With this model the genotypic value of hybrids can be predicted as

$$\hat{\mathbf{y}}^\star = \mathbf{1}\mu + \mathbf{F}^\star\hat{\mathbf{u}} + \mathbf{M}^\star\hat{\mathbf{v}} \tag{3}$$

where $\mathbf{F}^\star$ and $\mathbf{M}^\star$ are the design matrices for the predictors observed at the parental lines of the hybrid. The GCA of inbred lines can be predicted as

$$\hat{\mathbf{g}}_f^\star = \mathbf{F}^\star\hat{\mathbf{u}} \qquad \text{or} \qquad \hat{\mathbf{g}}_m^\star = \mathbf{M}^\star\hat{\mathbf{v}} \tag{4}$$

Assessment of prediction accuracy

The prediction accuracy for type 2 hybrids was evaluated with the cross-validation procedure of [3]. The estimation set consisted of the marker or mRNA data of three randomly chosen flint and five randomly chosen dent lines and the field data of their hybrids, and the validation set consisted of the remaining hybrids of the 7×14 factorial. Both parental lines of an untested hybrid in the validation set are also parents of hybrids belonging to the estimation set. Hence, testcross data are available for both parental lines of a hybrid. The principle is illustrated in Fig. 1a.

For type 0 hybrids, the estimation set consisted of five randomly chosen flint lines and ten randomly chosen dent lines and their hybrids. The validation set consisted of the hybrids of the remaining two flint and four dent lines of the 7×14 factorial. Hence, testcross data were not available for any of the two parental lines of a hybrid (Fig. 1b).

For each prediction model to be evaluated, cross-validation was carried out for 1000 runs. In each run the correlation $r(y, \hat{y})$ between the predicted and the observed hybrid yield and the average prediction error $\sum |\hat{y}_i - y_i|/n$ was assessed. The distribution of these measures over the 1000 replications was then used to compare the prediction models.

Results

For prediction of hybrid performance, the median of the correlations $r(y, \hat{y})$ between observed and predicted values in cross validation with type 2 hybrids was between 0.74 and 0.75 for grain yield and between 0.88 and 0.99 for grain dry matter content (Fig. 2). The differences in the median of the correlation between prediction with AFLPs, with all 10k mRNAs (mRNA10k), and with random samples of 1k out of the 10k mRNAs (mRNAr1k) were negligible. Prediction with mRNAs had a slightly smaller variation around the median than prediction with AFLPs. The average absolute prediction errors $|y - \hat{y}|$ had about the same sizes for prediction with AFLPs, all 10k mRNAs and random samples of 1k out of the 10k mRNAs.

For type 0 hybrids, the correlations between observed and predicted hybrid performance for both traits were lower than for type 2 hybrids. The median of the correlations in cross validation was between 0.54 and 0.56 for grain yield and between 0.29 and 0.41 for grain dry matter content. Differences in the median between the predictor sets AFLP, mRNA10k, and mRNAr1k were small. The ranges of the correlations were very large, and in some

a 3/5 cross validation of type 2 hybrids

	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14
F1	E	E	E	E	E	E	E	E	E	E	E	E	E	E
F2	E	E	E	E	E	E	E	E	E	E	E	E	E	E
F3	E	E	E	E	E	E	E	E	E	E	E	E	E	E
F4	E	E	E	E	E	V	V	V	V	V	V	V	V	V
F5	E	E	E	E	E	V	V	V	V	V	V	V	V	V
F6	E	E	E	E	E	V	V	V	V	V	V	V	V	V
F7	E	E	E	E	E	V	V	V	V	V	V	V	V	V

b 5/10 cross validation of type 0 hybrids

	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14
F1	E	E	E	E	E	E	E	E	E	E				
F2	E	E	E	E	E	E	E	E	E	E				
F3	E	E	E	E	E	E	E	E	E	E				
F4	E	E	E	E	E	E	E	E	E	E				
F5	E	E	E	E	E	E	E	E	E	E				
F6											V	V	V	V
F7											V	V	V	V

Fig. 1 Cross validation schemes. **a** Evaluation of prediction accuracy for untested hybrids in an incomplete factorial. The hybrids in the validation set are of type 2. **b** Evaluation of prediction accuracy for hybrids derived from parental lines of which no testcross data are available. The hybrids in the validation set are of type 0. D01–D14: parental dent lines in random order, F01–F07: parental flint lines in random order, E: hybrids of the estimation set, V: hybrids of the validation set

cross validation runs, even large negative correlations were observed. The average absolute prediction errors were greater than for type 2 hybrids and showed similar values for AFLPs and mRNA.

For prediction of mid-parent heterosis, the median of $r(y, \hat{y})$ with type 2 hybrids was between 0.81 and 0.82 for grain yield and between 0.90 and 0.91 for grain dry matter content (Fig. 3). The differences between the predictor sets AFLP, mRNA10k, mRNAr1k were negligible. The average absolute prediction error $|y-\hat{y}|$ had about the same sizes for the three predictor sets.

For type 0 hybrids, the correlations between observed and predicted mid-parent heterosis were between 0.26 and 0.4 for grain yield. For grain dry matter content no correlation between observed and predicted values in cross validation was observed.

In additional analyses we investigated the effect of further reducing the number of predictor variables below 1000. A decline of the prediction accuracy was observed for both traits (results not shown), which is in line with the results of [6].

We further investigated a ridge regression model in which we included 1000 random mRNAs and in addition the AFLP markers as predictors. We found no situation where combining the predictor sets resulted in a greater prediction accuracy than using them individually (results not shown).

Discussion

Properties of the linear model

In a simple GCA/SCA model $y_{fmr} = \mu + g_f + g_m + s_{fm} + e_{fmr}$ the performance of the rth replication of a hybrid is denoted by y_{fmr}. Factors g_f and g_m describe the GCA values of the parental lines, and s_{fm} is the SCA of the cross. In the linear model of Eq. 1, the GCA values are split into components that can be assigned to individual predictors, **Fv** splits up g_f and **Mu** splits up g_m.

Heterosis, and in consequence high hybrid performance, can be explained by dominant gene action at a large number of loci. Therefore, it is essential that models that attempt to predict hybrid performance include the effect of dominant gene action. The u_j and v_j in Eq. 1 can be interpreted in the sense of average effects (using the terminology of [18] p. 112ff) of the corresponding predictors. Average effects cover the effect a of additive gene action, and in addition they partially cover the effect d of dominant gene action (cf. Eq. 7.4a and 7.4b of [18], p. 113). The amount of the dominant gene action that is captured depends on the differences in the allele frequencies, and takes its minimum of zero for allele frequencies of 1/2. We

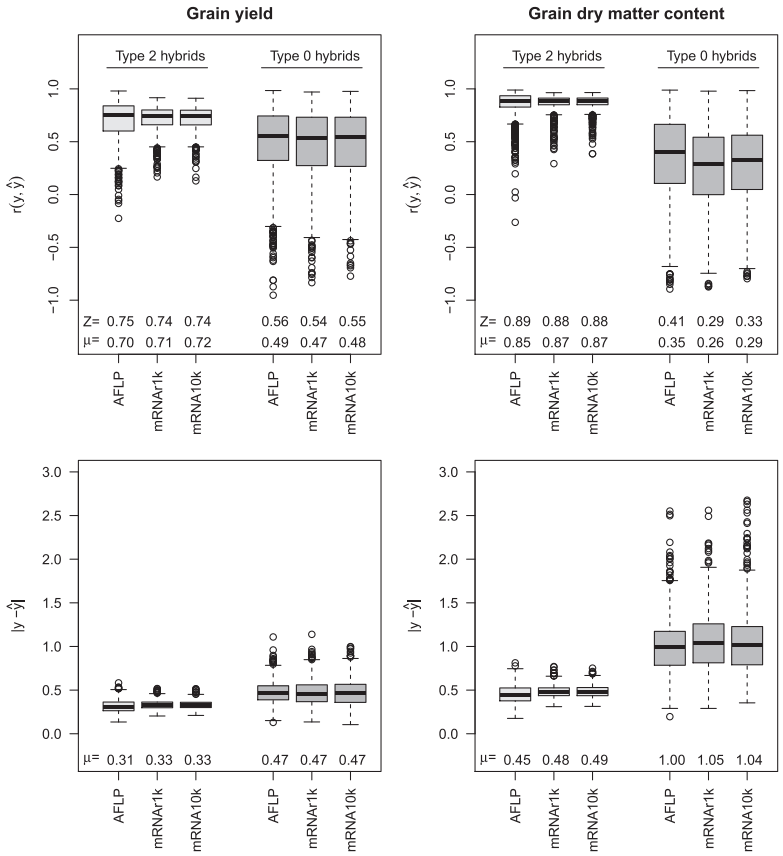

Fig. 2 Prediction accuracy for hybrid performance of type 2 hybrids (*left, in light gray*) and type 0 hybrids (*right, in dark gray*). Correlations $r(y, \hat{y})$ between observed and predicted grain yield and grain dry matter content, and average absolute prediction error $|y - \hat{y}|$ for the predictor sets AFLP (970 AFLP markers), mRNAr1k (1000 random mRNA transcripts), mRNA10k (10,810 mRNA transcripts). The boxplots show the distributions for 1000 cross validation runs, μ are the arithmetic means and Z the medians

hypothesize, that the differences in the allele frequencies in the heterotic pools of our factorial are so large that the average values include to a large extend the effect of dominant gene action. This is supported by the high prediction accuracies observed.

The SCA is neglected in Eq. 1. Extensions that include the SCA are straightforward from a formal point of view (Eq. 4 in [4]). The dissection of the SCA into components that can be assigned to individual predictors results in effects that can be interpreted in the sense of dominance deviations (cf. Table 7.3 of [18], p. 118). Dominance deviations cover the residual part of the effect of dominant gene action d, that is not covered by the average effects. Simulations have shown that the gain in prediction accuracy of models that include dominance deviations is small for divergent heterotic pools [4], because the major part of the effect of dominant gene action d is already covered by the average effects.

It remains open, and requires the analysis of further experimental data sets, whether including SCA in prediction models can actually improve hybrid prediction. In the data set investigated here, the high correlations of up to $r(y, \hat{y}) = 0.9$ between observations and predictions leave only little room for improving the GCA-based approach.

Prediction accuracy compared with older approaches

In earlier investigations on marker-based [11] and transcriptome-based [6, 7] prediction of hybrid performance, we used the same set of hybrids as here. This allows a direct comparison of the accuracy of the different prediction methods.

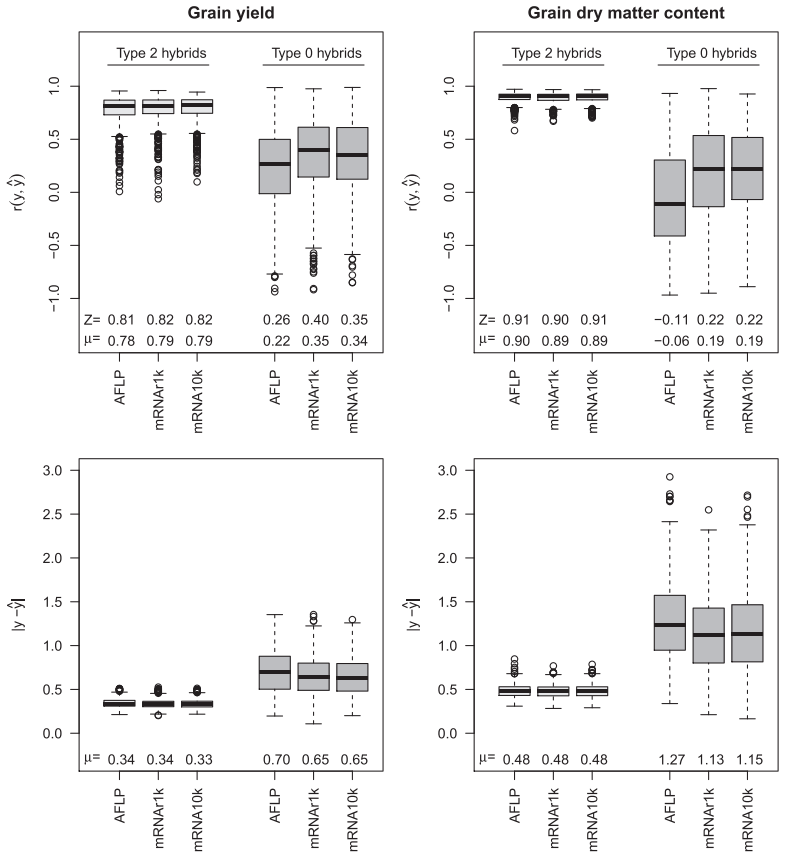

Fig. 3 Prediction accuracy for mid-parent heterosis of type 2 hybrids (*left, in light gray*) and type 0 hybrids (*right, in dark gray*). Correlations $r(y, \hat{y})$ between observed and predicted grain yield and grain dry matter content, and average absolute prediction error $|y - \hat{y}|$ for the predictor sets AFLP (970 AFLP markers), mRNAr1k (1000 random mRNA transcripts), mRNA10k (10,810 mRNA transcripts). The boxplots show the distributions for 1000 cross validation runs, μ are the arithmetic means and Z the medians

In the SM-TEAM approach of [11], first all markers are tested for association with the target trait and then a fixed linear model for the selected markers is fitted. This procedure is in analogy to the QTL-mapping approach, whereas a random model in which all markers remain (Eq. 1) can be regarded as a genome-wide prediction approach, as employed in recent studies on genomic selection. Hence, the theoretical advantages of the genome-wide prediction model, such as less bias in the effect estimates, should result in better statistical properties of the approach presented here compared with the approach of [11]. The correlation between predicted and observed hybrid performance for grain yield of type 2 hybrids obtained by the SM-TEAM approach was 0.65 (Figure 6 in [6]). The random effects model with AFLPs had a median of the correlation of 0.75 (Fig. 2). In consequence, with the present factorial, the ridge regression model applied to DNA marker data had a greater prediction accuracy than the earlier SM-TEAM model.

Transcriptome-based distances reached a prediction accuracy of about 0.8 for hybrid performance and mid parent heterosis of grain yield in type 2 hybrids (Figure 6 in [6]). This value is similar to the prediction accuracy reached by the ridge regression model (Fig. 3) for mid-parent heterosis. However, for hybrid performance, the ridge regression model showed only a correlation of 0.75 (Fig. 2), and, hence could not reach the prediction accuracies of the transcriptome-based distance model.

For prediction of hybrid performance for grain yield of type 0 hybrids, the transcriptome-based distances reached a median of the correlation between observations and predictions of 0.7 (Figure 3 in [7]). This was considerably

greater than the regression-based methods investigated in [7]. For the ridge regression model, a median of the correlation of about 0.55 was reached (Fig. 2). In consequence, for the prediction of type 0 hybrids the transcriptome-based distance model, which employs marker selection, resulted in considerably better predictions than the ridge regression model of this study.

Application in breeding programs

For application of hybrid prediction in breeding programs, it is of central importance that a prediction approach provides a sufficiently high prediction accuracy. For indirect selection approaches, a correlation of 0.7 to 0.9 between the trait for which selection is carried out and the target trait is usually regarded as highly promising and applicable in practice. Hence, the prediction accuracies for type 2 observed in this study can be regarded as suitable for practical applications.

The prediction accuracy of employing the ridge regression model to mRNAs was comparable to that obtained with AFLP markers in the investigated data set. The accuracies for prediction of grain yield and grain dry matter content in type 2 hybrids (Figs. 2 and 3) which were achieved with mRNA data suggest than mRNA can be an alternative to DNA markers in hybrid prediction.

The number of mRNAs required for a high prediction accuracy plays a central role for the costs of assessing the transcription profiles of selection candidates. For both traits and for both types of hybrids, the differences between using 1000 randomly chosen mRNAs or 10,000 mRNAs were negligible. This indicates, that high numbers of mRNA are not necessarily required for hybrid prediction, and that transcription profiling with limited resources might result in prediction accuracies that can be successfully used for indirect selection.

The ridge regression model employed in this study was in summary more precise than the older SM-TEAM prediction model. However it was not superior to the transcriptome based distances suggested by [6]. In particular for prediction of type 0 hybrids, the transcriptome-based distances might be the more promising approach. Further studies with larger data sets are required to verify these trends.

Conclusions

Hybrid prediction has the potential to greatly enhance the efficiency of hybrid breeding. In maize breeding, the doubled haploid technology can generate large numbers of candidate lines that surpass the field capacity by far. Thus, reliable hybrid prediction can be used to increase the selection intensity and hence the response to selection. The data structure of the factorial used in this study is typical for testing experimental hybrids in late stages of a maize hybrid breeding program, and hence the successful application of hybrid prediction with mRNA and ridge regression prediction models can be also expected with other data sets of similar genetic structure.

Abbreviations

AFLP: amplified fragment length polymorphism; GCA: general combining ability; SCA: specific combining ability.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TS collected and analyzed the field data, AT, FS, SS collected the mRNA data, CZP carried out the predictions, AEM, MF, SS conceived the study, CZP, MF wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

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

Author details

[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.

Received: 27 August 2015 Accepted: 8 March 2016
Published online: 29 March 2016

References

1. Bernardo R. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 1994;34:20–5.
2. Vuylsteke M, Kuiper M, Stam P. Chromosomal regions involved in hybrid performance and heterosis: Their AFLP-based identification and practical use in prediction models. Heredity. 2000;85:208–18.
3. Schrag TA, Möhring JM, Maurer HP, Dhillon BS, Melchinger AE, Piepho HP, Sorensen AP, Frisch M. Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theor Appl Genet. 2009;118:741–51.
4. Technow F, Riedelsheimer C, Schrag TA, Melchinger AE. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet. 2012;125:1181–1194.
5. Massman JM, Gordillo A, Lorenzana RE, Bernardo R. Genomewide predictions from maize single-cross data. Theor Appl Genet. 2013;126: 13–22.
6. Frisch M, Thiemann A, Fu J, Schrag T, Scholten S, Melchinger AE. Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theor Appl Genet. 2010;120: 441–50.
7. Fu J, Falke KC, Thiemann A, Schrag TA, Melchinger AE, Scholten S, Frisch M. 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. 2012;124:825–33.
8. Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön CC. Genome-based prediction of testcross values in maize. Theor Appl Genet. 2011;123:339–50.
9. Hofheinz N, Borchardt D, Weissleder K, Frisch M. Genome-based prediction of test cross performance in two subsequent breeding cycles. Theor Appl Genet. 2012;125:1639–45.
10. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet. 2012;44:217–20.
11. Schrag TA, Melchinger AE, Sorensen AP, Frisch M. 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. 2006;113:1037–47.

12. Gardiner JM, Buell CR, Elumalai R, Galbraith DW, Henderson DA, Iniguez AL, Kaeppler SM, Kim JJ, Liu J, Smith A, Zheng L, Chandler VL. Design, production, and utilization of long oligonucleotide microarrays for expression analysis in maize. Maydica. 2005;50:425–35.
13. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics. 2001;2:183–201.
14. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:3.
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.
16. Fu J, Thieman A, Schrag TA, Melchinger AE, Scholten S, Frisch M. Dissecting grain yield pathways and their interactions to grain dry matter content through a two-step correlation approach with maize seedling transcriptome. BMC Plant Biol. 2010;10:63.
17. Henderson CR. Applications of Linear Models in Animal Breeding. Guelph, Canada: University of Guelph; 1984.
18. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. Harlow Essex UK: Longman Group; 1996.

Chapter 3

Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme [1]

[1] Zenke-Philippi, C., M. Frisch, A. Thiemann, F. Seifert, T. Schrag, A. E. Melchinger, S. Scholten, and E. Herzog (2017) Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme. *Plant Breeding* **136**:331–337.

Plant Breeding, **136**, 331–337 (2017)

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

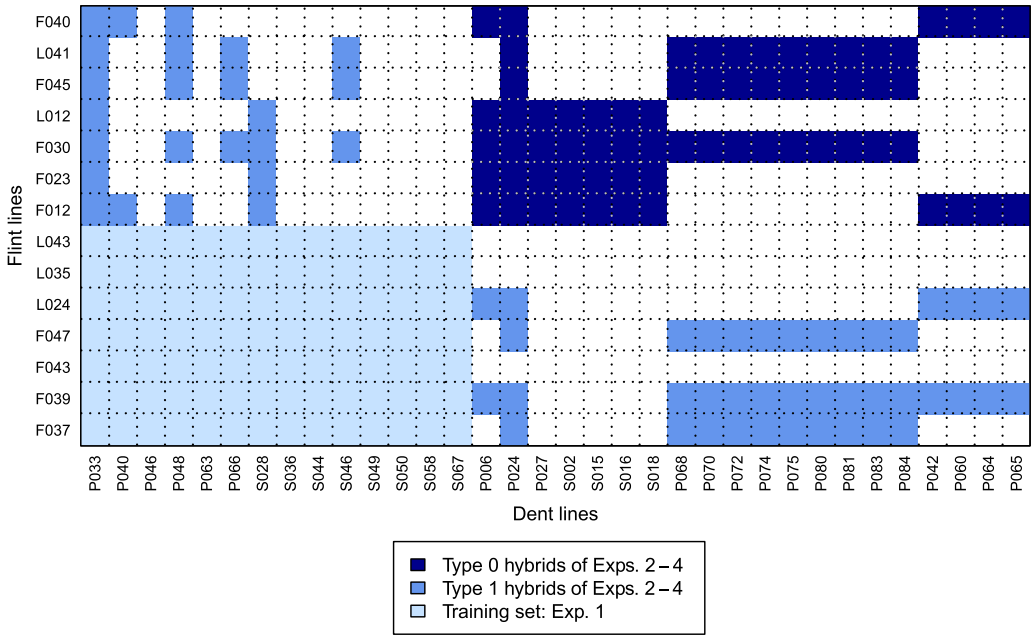

Fig. 1: The 34 dent and 14 flint lines of our data set and the hybrids generated from them. The display illustrates prediction of type 0 (dark blue) and type 1 (medium blue) hybrids of exps. 2–4 using the factorial of Exp. 1 (light blue) as training set [Color figure can be viewed at wileyonlinelibrary.com]

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 $\mathbf{y}$ 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):

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

$\mathbf{y}$ 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):

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

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,

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

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

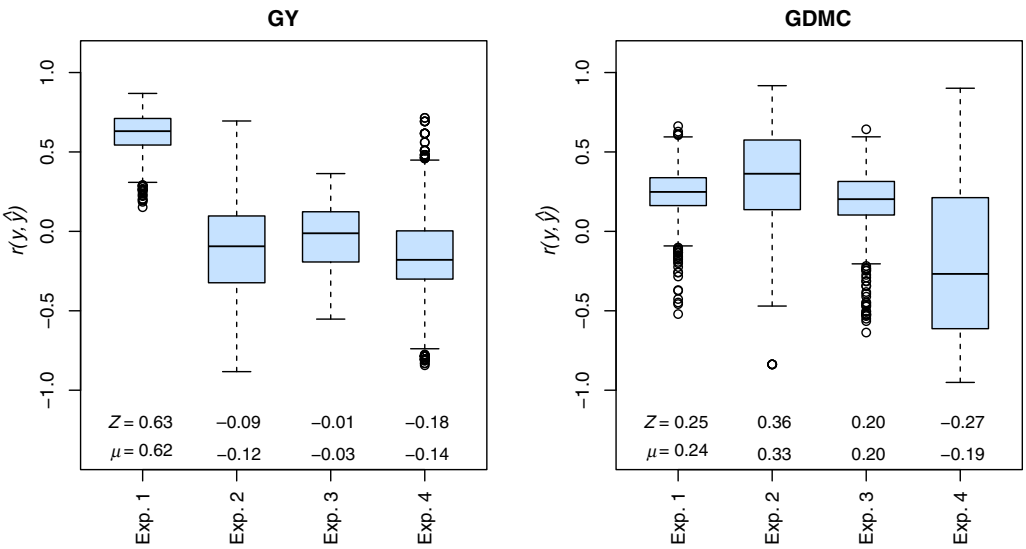

Fig. 2: Cross-validation within experiments for assessing the accuracy $r(y, \hat{y})$ of predicting hybrid performance for GY and GDMC with transcriptome-based distances using the 2k core set of mRNA transcripts. The boxplots show the distributions for 1000 cross-validation runs, μ are the arithmetic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

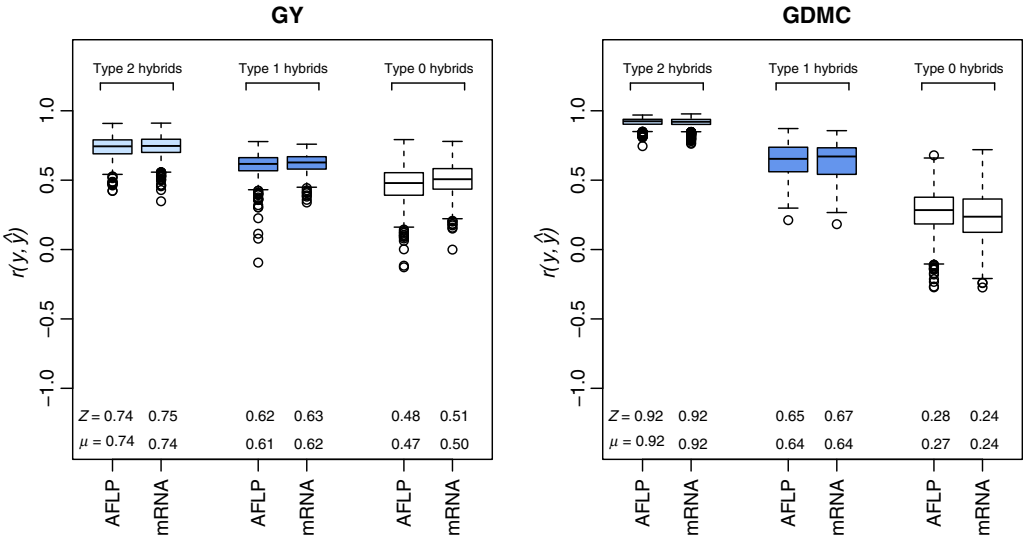

Fig. 3: Cross-validation across experiments for assessing the accuracy $r(y, \hat{y})$ of predicting hybrid performance for GY and GDMC with ridge regression using AFLPs and mRNA transcripts. The boxplots show the distributions for 1000 cross-validation runs, μ are the arithmetic means and Z are the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

Chapter 4

General discussion

Genetic markers vs. mRNA transcription profiles in ridge regression

In addition to DNA markers, hybrid prediction can be carried out with trancriptomic data (Fu et al. 2012; Guo et al. 2016). Compared to DNA markers, they have the advantage that they do not rely on linkage between the marker allele and the functional allele responsible for the phenotype (Frisch et al. 2010). Moreover, they are one step closer to the protein as the end product of gene expression and should therefore carry additional information about the plant as well as about epistatic interactions (Guo et al. 2016).

mRNA transcription profiles were employed in a ridge regression prediction model to predict hybrid performance of grain yield and grain dry matter content in a balanced set of factorial crosses with 98 maize hybrids (Zenke-Philippi et al. 2016) and in an unbalanced data set with 230 maize hybrids (Zenke-Philippi et al. 2017), both from ongoing breeding programs. Prediction accuracies were similar for RR-BLUP with AFLP marker data and mRNA transcription profiles (Zenke-Philippi et al. 2016, 2017), indicating

that the information content was comparable but not greater for mRNAs than for AFLPs. There are several possible explanations of this finding.

Inconsistent LD between Dent and Flint lines would have had no effect on the estimates of hybrid performance in the data set because average marker effects were modeled separately for the two parental breeding pools (Zenke-Philippi et al. 2016). Furthermore, a study on 100 Dent and 97 Flint lines in maize revealed a high proportion of SNP markers with consistent linkage phases across the Flint and Dent heterotic pools (Technow et al. 2013). Consequently, the independence of mRNA trancription profiles from linkage phases did not yield any advantage.

Relatedness was apparently the main source of information utilized by both AFLP markers and mRNA transcription profiles when predictions were made with RR-BLUP. Principal coordinate analyses with mRNAs and AFLPs resulted in the same separation of the maize lines into the different heterotic pools, indicating that both types of markers identify the same information on relatedness between the lines (Frisch et al. 2010). Prediction of breeding values in cattle with RR-BLUP was found to use relationship information to a great extent whereas LD information could not be exploited (Habier et al. 2007). Contradictingly, in a study in maize, prediction accuracy for hybrid prediction decreased only slightly when all close relationships between inbred lines were removed (Riedelsheimer et al. 2012). The authors argue, however, that genetic markers make use of "baseline relationships" caused by long chromosomal segments which are transmitted intact over generations. Comparable prediction accuracies for prediction of inbred lines with GBLUP in a maize diversity panel were reported for SNP markers and mRNA transcripts (Guo et al. 2016). Even though only weak correlations were found between the genomic relationships estimated from SNP markers and mRNA transcripts, the authors hypothesize that SNP markers as well as mRNA transcripts make use of the same source of genetic information, even if it is not relatedness directly. Additionally, transcription profiles are highly dependent on the tissue and the developmental stage the material is

collected from (Guo et al. 2016). The inbred line seedlings in the present study were grown under controlled conditions and harvested at 7 days. The mRNA expression profiles are therefore "standardized" in the way that they most likely resemble a basic state of gene expression and therefore contain the same information as genetic markers. However, with genotyping of plants that have to be grown under field conditions, phenotyping could be done directly and genomic prediction would lose its appeal of increasing genetic gain per unit time and cost compared to phenotypic selection (Heslot et al. 2015).

The mRNA transcription profiles were assessed in the parental inbred lines and the RR-BLUP model estimated the additive marker effects for mRNAs separately for male and female parents so that only epistasis occurring *within* an inbred line was captured whereas epistasis *between* to inbred lines, *i.e.*, in a hybrid, was not accounted for. However, mRNA expression levels in maize hybrids were shown to be largely additive when compared to the parental levels (Thiemann et al. 2014), which corresponds to the set-up of the RR-BLUP model. Additionally, explicitly modeling additive $\times$ additive epistasis did not result in an increased prediction accuracy in two maize data sets (Jiang and Reif 2015), confirming that epistasis is unlikely to be a major cause of heterosis in maize (Garcia et al. 2008).

Summarizing, the hypothesis that mRNA transcription profiles contain more information than AFLP markers could not be confirmed by the data.

Marker number and size of training set in ridge regression

In order to reduce genotyping costs, it is interesting to know how far the marker density can be reduced without losing prediction accuracy. Marker densities in other studies only had a minor influence on prediction accuracies (Zhao et al. 2015) and reached plateaus at a few hundred (Lorenzana and Bernardo 2009; Zhao et al. 2012a, 2013a) or a few thousand markers (Technow et al. 2012; Crossa et al. 2014), depending on the populations used for the predictions. It was investigated whether these numbers were sufficient for prediction with mRNA transcripts as well and similar prediction accuracies were found for prediction of grain yield with 10k and 1k mRNAs in a set of factorial crosses with 98 hybrids (Zenke-Philippi et al. 2016) and comparable medians for additional complete and incomplete sets of factorial crosses predicted with 2k mRNAs or less (Zenke-Philippi et al. 2017). In all cases, prediction with 970 AFLP markers resulted in similar prediction accuracies (Zenke-Philippi et al. 2017, 2016). These results indicate that a relatively low number of markers, no matter whether genomic or transcriptomic, is sufficient for successful predictions in a data set with similar structure as ours.

Predictions became more accurate in terms of less variation across cross-validation runs with the data set of 230 hybrids compared to 98 hybrids (Zenke-Philippi et al. 2016, 2017). In the unbalanced data set with 230 hybrids, 58 hybrids were on average selected as parents of the training set for type 0 hybrids. Others studies found increasing prediction accuracies for type 2 and type 1 hybrids with training set sizes ranging from 300-450 hybrids, a plateau of prediction accuracy for type 0 hybrids with a training set size of 300 hybrids (Technow et al. 2014) and stable predictive abilities for 108 genotypes in the training set (Windhausen et al. 2012). These numbers indicate that an increase in the size of the data set might further improve prediction accuracies.

Results from maize (Riedelsheimer et al. 2012; Zhao et al. 2012b) and rye (Wang et al. 2014) suggest that it is not the sheer number of hybrids but more the genetic variation that accounts for high prediction accuracies, either directly or via the number of polymorphic markers. This might be the explanation for the finding in independent validation that comparable training set sizes differ considerably in the prediction accuracies (Zenke-Philippi et al. 2017). In conclusion, it is benefical to use training sets which are genetically diverse but nevertheless represent the validation sets as closely as possible. If enough emphasis is put on this point, reliable and stable prediction results might already be achievable with relatively low numbers of hybrids in the training set.

In the two studies presented here, prediction accuracies for grain dry matter content were substantially higher for type 2 hybrids and substantially lower for type 0 hybrids compared to grain yield (Zenke-Philippi et al. 2016, 2017). This finding is in line with results from other studies in maize (Massman et al. 2013; Technow et al. 2014) and shows that general conclusions about the importance of single parameters for hybrid prediction can be difficult even within the same species since the prediction of different traits may have different requirements.

Hybrid prediction with a core set of mRNAs

A core set of transcriptomic markers which are not randomly selected but correlated to the trait of interest would be an interesting option to reduce the number of markers needed for prediction. Prediction of type 0 hybrids was successful with transcriptome-based distances for gene numbers of around 1000 to 1500 when genes were selected from 10k mRNAs based on association of differential gene expression with hybrid performance (Fu et al. 2012). It seemed possible, therefore, to develop a core set of genes based on one set

of factorial hybrids and use it for the prediction of hybrid performance with transcriptome-based distances in different factorials. In order to check the applicability of this approach, a 2k subset from 47k genes was selected based on association with hybrid performance of grain yield and grain dry matter content in a set of 98 factorial crosses (Thiemann et al. 2010). This core set of genes was then used in microarray analyses of three additional sets of factorial crosses. 2122, 104, 542, and 140 genes of the 2k core set were found to be differentially expressed in the four sets of factorial crosses, respectively. 985 genes were differentially expressed in the unbalanced data set with 230 hybrids. These were employed for hybrid prediction with transcriptome-based distances and RR-BLUP with cross-validation both within each of the four factorials and within the unbalanced data set as well as with independent validation, *i.e.*, with one or two factorials as the training set for the prediction of the rest of the unbalanced data set (Zenke-Philippi et al. 2017).

First, cross-validation for prediction within the single experiments was carried out for predictions with transcriptome-based distances estimated from the 2k core set. In a set of 98 factorial crosses, based on which the genes were selected, a median of 0.63 was found in correlations between actual and predicted hybrid performance for grain yield for prediction with D_B based on the 2k core set. For cross-validation within the other factorial crosses and within the unbalanced data set, correlations were negligibly small (Zenke-Philippi et al. 2017). When one or two sets of factorial crosses were used as the training set and the remaining crosses as the validation set, prediction accuracies ranged around zero (Zenke-Philippi et al. 2017). No increase in prediction accuracies was found when predictions were made with transcriptome-based distances only based on genes correlated with either grain yield or grain dry matter content (Zenke-Philippi et al. 2017), even though these traits are known to be negatively correlated.

In conclusion, no core set of genes for prediction of a particular trait could be identified that can be selected based on one factorial and then be transferred to other data sets. A sufficient number of genes has to be available

to estimate meaningful transcriptome-based distances separately for each data set. With the high throughput of modern next-generation sequencing techniques that can be applied to mRNA, however, it is realistic to generate the needed data for a reasonable price so that transcriptome-based distances are still a promising approach for hybrid prediction.

When the 2k core set was used in a ridge regression model, prediction accuracies were comparable with those obtained with AFLPs (Zenke-Philippi et al. 2017). For cross-validation in the unbalanced data set with 230 hybrids, the medians of prediction accuracies for grain yield were ≈ 0.75, ≈ 0.60 and ≈ 0.50 for type 2, type 1 and type 0 hybrids, respectively. The medians of prediction accuracies for grain dry matter content were ≈ 0.90, ≈ 0.65 and ≈ 0.30 for type 2, type 1 and type 0 hybrids, respectively. When one or two factorial subsets of crosses were used for prediction of the remaining factorials, average prediction accuracies were near 0.6 for type 1 and considerably lower than 0.5 for type 0 hybrids. In all cases, prediction accuracies for RR-BLUP with mRNAs were comparable to those for RR-BLUP with AFLPs which were regarded as a baseline across the two studies. Pre-selecting a core set of genes therefore did not affect prediction accuracies.

The selection of genes based on one factorial can be regarded as a severe case of ascertainment bias, *i.e.*, when marker data are not obtained from a random sample of the polymorphisms in the population of interest (Heslot et al. 2013). In that context, the 2k microarray is severely affected by ascertainment bias as the samples are selected based on one factorial. In wheat, prediction accuracies were identical for prediction with biased SNPs compared to random SNPs obtained with genotyping-by-sequencing (Heslot et al. 2013). This is in line with the finding that the selection of mRNAs based on a set of factorial crosses yields comparable prediction results in other sets of factorial crosses and indicates that prediction with RR-BLUP is robust towards biased marker selection.

Ridge regression vs. transcriptome-based distances

The prediction of performance of type 0 hybrids without parental testcross data is especially interesting for breeders since with *e.g.*, 1000 inbred lines available for each heterotic group and 100 of them in the training set, there are 10,000 type 2 hybrids, 180,000 type 1 hybrids and 810,000 type 0 hybrids (Technow et al. 2014). Promising hybrids are therefore most likely to be found among type 0 hybrids (Technow et al. 2014) and their discovery requires accurate prediction. Ridge regression approaches were promising for prediction of type 2 hybrids in a set of 98 factorial crosses but failed to achieve the desired results for type 0 hybrids (Zenke-Philippi et al. 2016). Transcriptome-based distances emerged as a powerful alternative for the prediction of grain yield in type 0 hybrids, with medians in prediction accuracies of up to 0.7 (Fu et al. 2012).

In a set of 98 factorial crosses, the predictive potential of transcriptome-based distances was confirmed for hybrid performance in grain yield but not in grain dry matter content (Zenke-Philippi et al. 2017). Grain yield has a high level of heterosis in hybrids whereas the mid-parent heterosis for grain dry matter content is low (Thiemann et al. 2010). Since transcriptome-based distances measure the dissimilarity between the parental lines which is regarded as a prerequisite for heterosis (Lanza et al. 1997; Marsan et al. 1998; Chen 2013), it is consequent that a heterotic trait like grain yield can be predicted with transcriptome-based distances whereas a trait with a low level of heterosis like grain dry matter content cannot. Transcriptome-based distances are therefore only useful for the prediction of traits with a sufficient level of heterosis.

In the other sets of factorial crosses and in the combined, unbalanced data set, prediction accuracies for predictions with transcriptome-based distances ranged around zero, even for type 2 hybrids (Zenke-Philippi et al. 2017).

Predictions with RR-BLUP, on the other hand, resulted in prediction accuracies of ≈ 0.75 and ≈ 0.50 for grain yield and ≈ 0.90 and ≈ 0.25 for grain dry matter content in type 2 and type 0 hybrids, respectively. In this study, less than 2k mRNAs were used for the estimation of transcriptome-based distances instead of a selection from 10k mRNAs (Fu et al. 2012) based on the association of differential gene expression with high hybrid performance (Frisch et al. 2010). This indicates that transcriptome-based distances need a sufficient number of mRNAs to be successfully estimated and that no preselection based on different sets of factorial crosses is possible.

Cross-validation vs. independent validation

Genomic prediction is appealing because it has the potential to shorten generation intervals considerably. This is achieved via genotyping individuals at an early developmental stage and predicting their performance instead of having to wait for the field data. However, resources are needed to create training sets. Thus, the possibility to use material from previous breeding cycles to predict hybrid performance instead of having to design specific training sets is of particular interest to breeders. It is therefore surprising (Jonas and de Koning 2013) that predictions are usually made with cross-validation within the same breeding generation and relatively few studies are available for prediction across generations (*cf.* Hofheinz et al. 2012; Auinger et al. 2016; He et al. 2016; Michel et al. 2016) or with data already available from previous years. Independent validation, *i.e.*, the prediction of sets of factorial crosses with a model calibrated with a different set of factorial crosses, resembles the latter situation.

Slightly higher prediction accuracies were found with cross-validation than with independent validation for both type 0 and type 1 hybrids (Zenke-Philippi et al. 2017). In cross-validation, the data set is randomly divided into

training and validation set, ensuring an overall more balanced representation of the validation set by the training set. In independent validation, on the contrary, the factorial used as the training set may substantially differ from the validation set. It therefore reflects the prediction across breeding cycles better, indicating that prediction accuracies achieved with cross-validation over-estimate the potential of genomic prediction to a certain degree. Even if high prediction accuracies are obtained with cross-validation, prediction of subsequent breeding cycles can, depending on the trait, result in a large decrease in prediction accuracy (Hofheinz et al. 2012).

The use of already available data for genomic prediction therefore seems possible as long as the genetic pools and sub-pools in the validation set are represented well by the training set and relatedness is ensured (Zenke-Philippi et al. 2017), as already shown for rye (Auinger et al. 2016). This further emphasizes the fact that the prediction accuracy heavily depends on the genetic relatedness between the training and the validation set. Only if the validation set is closely related to the training set, high prediction accuracies can be achieved (Albrecht et al. 2014; Technow et al. 2014; Albrecht et al. 2011). Relatedness between training and validation set is *the* determining factor for the success of hybrid prediction and genomic prediction in general. When there is a choice between increasing the degree of relatedness between training and validation set vs. increasing the size of the training set, breeders should aim for more closely related sets with comparable genetic composition.

Prediction accuracies in type 0 hybrids vary the most between cross-validation runs compared to type 1 and type 2 hybrids (Zenke-Philippi et al. 2016, 2017), presumably because they are more prone to changes in the training set due to random sampling (Technow et al. 2014). However, the high prediction accuracies achieved for single cross-validation runs also mean that even the prediction of type 0 hybrids can be successful for certain compositions of the training set. The careful design of training sets might therefore contribute to improvements in the prediction of hybrids without parental testcross data and requires further investigation.

Conclusions

Prediction of heterotic traits like grain yield with transcriptome-based distances can be superior to prediction with ridge regression models for type 0 hybrids without parental lines in the training set. However, these predictions then cannot be based on a small core set of genes. A sufficient number of genes has to be available to for precise prediction with transcriptome-based distances. The performance of ridge regression models, on the other hand, was robust towards changes in the selection of the genes used, regarding both number and ascertainment bias. Ridge regression models might therefore be favorable if relatively few genes are available and/or if the heterosis of a trait is low, as for grain dry matter content. Transcriptome-based distances, on the other hand, could be advantageous if data for many genes are available, if the relationship between training set and validation set is low, and if the traits show high heterosis like grain yield. One strength of mRNA transcription profiles compared to DNA markers could be that they can be used in transcriptome-based distances as well as ridge regression models which allows choosing the appropriate model for different situations. This makes them more flexible in the application and facilitates the use of data already available from earlier breeding cycles.

Chapter 5

Summary

Most studies on genomic prediction of hybrids employ genetic markers as the main carrier of information. Very few use transcriptomic or metabolomic data despite the fact that the end product of gene expression, *i.e.*, the protein, might carry more information than genetic markers. The main goal of the present study was therefore to investigate whether gene expression profiles can be employed successfully for hybrid prediction in maize.

With RR-BLUP, similar accuracies were found for ALFP markers and mRNA transcription profiles for prediction of hybrid maize grain yield and grain dry matter content within a set of 98 factorial crosses and within an unbalanced set of 230 maize hybrids. This indicates that either the mRNA transcription profiles do not carry additional information or that this information cannot be exploited by the model since the prediction accuracy of RR-BLUP was shown to be based largely on the relatedness between training and validation set.

No investigations on the number of required mRNA transcripts for reliable predictions had been conducted so far. Comparable prediction accuracies were found for 10k, 2k and 1k mRNA transcripts, and 1k AFLP markers. This means that also in terms of the number of mRNA transcripts required, the mRNA transcription profiles are comparable to genetic markers.

Summary

A major challenge is the successful prediction of hybrids whose parents are not in the training set. Transcriptome-based binary distances D_B based on 10k mRNA transcripts had been shown to be advantageous in this situation. However, prediction with D_B based on 2k mRNA transcripts was found to be inferior to prediction with RR-BLUP in most cases, especially in grain dry matter content, a trait with low heterosis. Apparently, a large number of mRNAs must be available to select from for meaningful transcriptome-based distances and prediction can only be successful in heterotic traits.

Pre-selection of a core set of genes for hybrid prediction with transcriptome-based distances would save resources since only evaluation of this reduced subset of genes would be required. A core set of 2k genes was identified in a set of 98 factorial crosses. Genes were selected for the core set if they either showed differential expression between the parental genotypes and consistent association with hybrid performance for grain yield or if they were correlated with hybrid performance for grain yield or grain dry matter content or mid-parent heterosis for grain yield. The core set was then used for hybrid prediction in three other factorials and in an unbalanced data set of 230 hybrids. Prediction accuraries were negligibly small for prediction with D_B. Prediction accuracies for prediction with RR-BLUP were comparable to those achieved with 1k ALFP markers but not higher. This indicates that using gene subsets is not a promising approach to save resources in applied breeding programs.

Even more than the generation of marker data, the calibration of models based on appropriate training sets is very resource-intensive. It would therefore be beneficial for breeders if material from previous breeding cycles could be employed for that purpose. Prediction accuracies of RR-BLUP with ALFP markers and mRNA transcription profiles were evaluated when one or two of the four sets of factorial crosses formed the training set and the remaining factorials formed the validation set. Mean prediction accuracies in grain yield and grain dry matter content were higher than 0.55 in type 1 hybrids and 0.16 to 0.38 in type 0 hybrids. Thus, prediction with models calibrated

with material from previous breeding cycles seems to be possible if sufficient relatedness is ensured.

In conclusion, mRNA transcription profiles can be regarded as promising predictors of hybrid performance in maize. Their possible application in RR-BLUP as well as in transcriptome-based distances makes them more versatile than DNA markers and their use is recommended over that of DNA markers, if possible.

Chapter 6

Zusammenfassung

Genomweite Vorhersagen der Hybridleistung werden meistens auf Basis genetischer Marker durchgeführt. Das Potenzial von Transkriptom- und Metabolomdaten für die Hybridvorhersage wurde bislang nur in wenigen Studien untersucht, obwohl in diesen Daten durch ihre größere Nähe zum Protein, dem Endprodukt der Genexpression, zusätzlich verwertbare Information enthalten sein könnte. Das Hauptziel der vorliegenden Arbeit war daher, zu überprüfen, ob Genexpressionsprofile für die Hybridvorhersage in Mais genutzt werden können. Mit RR-BLUP waren die Vorhersagegenauigkeiten für Kornertrag und Korntrockenmassegehalt von Maishybriden in einem Datensatz mit 98 faktoriellen Kreuzungen sowie in einem unbalancierten Datensatz mit 230 Hybriden vergleichbar für ALFP-Marker und mRNA-Transkriptionsprofile. Entweder war in den mRNA-Transkriptionsprofilen also keine zusätzliche Information enthalten oder diese Information konnte nicht genutzt werden, da die Vorhersagegenauigkeit von RR-BLUP zum Großteil auf der Verwandtschaft zwischen Trainingsset und Validierungsset beruht.

Auch zur für aussagekräftige Vorhersagen nötigen Anzahl an mRNA-Transkripten gab es bislang keine Untersuchungen. Die in der vorliegenden Arbeit erzielten Vorhersagegenauigkeiten waren vergleichbar für 10000, 2000 und 1000 mRNA-Transkripte und 1000 AFLP-Marker. mRNA-Transkriptionsprofile entsprechen also auch bei der Anzahl der zur Vorhersage benötigten Datenpunkte DNA-Markern.

Ein wichtiges Ziel der Hybridvorhersage ist die erfolgreiche Prognose der Leistung von Hybriden, deren Elternlinien nicht Teil des Trainingssets sind. Transkriptombasierte binäre Distanzen D_B hatten sich dafür in vorhergehenden Studien als vorteilhaft erwiesen. Die Vorhersagegenauigkeiten für die auf 2000 mRNA-Transkripten basierenden Distanzen waren in der vorliegenden Arbeit in den meisten Fällen geringer als die für RR-BLUP, besonders bei Merkmalen mit einer geringen Heterosis, wie z. B. dem Korntrockenmassegehalt. Transkriptombasierte Distanzen sind also offensichtlich nur für die Vorhersage heterotischer Merkmale geeignet und erfordern, dass eine ausreichende Zahl von mRNA-Transkripten zu ihrer Schätzung zur Verfügung steht.

Die Auswahl eines Core Sets von Genen, das dann routinemäßig zur Hybridvorhersage mit transkriptombasierten Distanzen genutzt werden könnte, würde Ressourcen sparen, da nur die Evaluierung des reduziertes Satzes an Genen nötig wäre. Ein Core Set von Genen wurde in einem faktoriellen Kreuzungsschmema mit 98 Hybriden ausgewählt. Kriterien für die Auswahl eines Gens waren entweder die differentielle Genexpression in den Elternlinien gepaart mit einem Zusammenhang zum Kornertrag der Hybriden oder eine Korrelation des Gens mit der Hybridleistung im Kornertrag oder in der Korntrockenmasse oder mit der Heterosis im Vergleich zum Elternmittel im Kornertrag. Anschließend wurden Hybridvorhersagen in drei weiteren faktoriellen Kreuzungsschemata sowie in einem unbalancierten Datensatz mit 230 Hybriden vorgenommen. Die Vorhersagegenauigkeiten für transkriptombasierte Distanzen waren vernachlässigbar gering. Mit RR-BLUP konnten für diese 2000 Gene ähnliche, aber keine höheren Vorhersagegenauigkeiten erreicht werden wie mit 1000 AFLP-Markern. Die Hybridvorhersage mit vorausgewählten Sätzen von Genen ist also kein erfolgversprechender Ansatz, um Ressourcen zu sparen.

Noch ressourcenintensiver als die Erstellung von Markerdaten ist die Schätzung von Vorhersagemodellen basierend auf passenden Trainingsets. Daher wäre es vorteilhaft für Züchter, Material aus vergangenen Zuchtzyklen

dafür nutzen zu können. In der vorliegenden Arbeit wurden die Vorhersagegenauigkeiten von RR-BLUP mit mRNA-Transkriptionsprofilen und mit AFLP-Markern, wenn ein oder zwei faktorielle Kreuzungsschemata als Trainingset und die restlichen Kreuzungsschemata als Validierungsset genutzt wurden, untersucht. Die Ergebnisse waren mit mittleren Vorhersagegenauigkeiten für Kornertrag und Korntrockenmassegehalt von mehr als 0,55 für Typ-1-Hybriden und 0,16 bis 0,38 für Typ-0-Hybriden vielversprechend. Die Anpassung von Modellen basierend auf Material von vorhergehenden Zuchtzyklen ist also offenbar möglich, wenn eine ausreichende Verwandtschaft zu den vorherzusagenden Hybriden besteht.

Insgesamt erscheinen mRNA-Transkriptionsprofile als erfolgversprechende Prädiktoren für die Hybridleistung von Mais. Sie sind sowohl mit RR-BLUP als auch in transkriptombasierten Distanzen zu verwenden, sind somit vielseitiger als DNA-Marker und sollten daher an ihrer Stelle für die Hybridvorhersage genutzt werden, sofern möglich.

Chapter 7

Literature

T. Albrecht, V. Wimmer, H. J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, and C. C. Schön. Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, 123(2):339–350, 2011.

T. Albrecht, H.-J. Auinger, V. Wimmer, J. O. Ogutu, C. Knaak, M. Ouzunova, H.-P. Piepho, and C.-C. Schön. Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theoretical and Applied Genetics*, 127(6):1375–1386, 2014.

H.-J. Auinger, M. Schönleben, C. Lehermeier, M. Schmidt, V. Korzun, H. H. Geiger, H.-P. Piepho, A. Gordillo, P. Wilde, E. Bauer, et al. Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). *Theoretical and Applied Genetics*, 129 (11):2043–2053, 2016.

R. Bernardo. Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, 34(1):20–25, 1994.

Z. J. Chen. Genomic and epigenetic insights into the molecular bases of heterosis. *Nature Reviews Genetics*, 14(7):471–482, 2013.

M. Cheres, J. Miller, J. Crane, and S. Knapp. Genetic distance as a predictor of heterosis and hybrid performance within and between heterotic groups in sunflower. *Theoretical and Applied Genetics*, 100(6):889–894, 2000.

J. Crossa, G. de Los Campos, P. Pérez, D. Gianola, J. Burgueño, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker, J. Yan, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186(2):713–724, 2010.

J. Crossa, P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, et al. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, 112(1):48–60, 2014.

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

R. De Vlaming and P. J. Groenen. The current and future use of ridge regression for prediction in quantitative genetics. *BioMed Research International*, 2015, 2015.

Z. A. Desta and R. Ortiz. Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9):592–601, 2014.

B. Diers, P. McVetty, and T. Osborn. Relationship between heterosis and genetic distance based on restriction fragment length polymorphism markers in oilseed rape (*Brassica napus* L.). *Crop Science*, 36(1):79–83, 1996.

K. Feher, J. Lisec, L. Römisch-Margl, J. Selbig, A. Gierl, H.-P. Piepho, Z. Nikoloski, and L. Willmitzer. Deducing hybrid performance from parental metabolic profiles of young primary roots of maize by using a multivariate diallel approach. *PLOS ONE*, 9(1):e85435, 2014.

M. Frisch, A. Thiemann, J. Fu, T. Schrag, S. Scholten, and A. E. Melchinger. Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theoretical and Applied Genetics*, 120(2):441–450, 2010.

J. Fu, K. Falke, A. Thiemann, T. A. Schrag, M. A. E., S. Scholten, and M. Frisch. Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. *Theoretical and Applied Genetics*, 124(5):825–833, 2012.

A. A. F. Garcia, S. Wang, A. E. Melchinger, and Z.-B. Zeng. Quantitative trait loci mapping and the genetic basis of heterosis in maize and rice. *Genetics*, 180(3):1707–1724, 2008.

Z. Guo, M. M. Magwire, C. J. Basten, Z. Xu, and D. Wang. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics*, 129(12):2413–2427, 2016.

D. Habier, R. Fernando, and J. Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.

S. He, A. W. Schulthess, V. Mirdita, Y. Zhao, V. Korzun, R. Bothe, E. Ebmeyer, J. C. Reif, and Y. Jiang. Genomic selection in a commercial winter wheat population. *Theoretical and Applied Genetics*, 129(3): 641–651, 2016.

C. R. Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–447, 1975.

N. Heslot, H.-P. Yang, M. E. Sorrells, and J.-L. Jannink. Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52(1):146–160, 2012.

N. Heslot, J. Rutkoski, J. Poland, J.-L. Jannink, and M. E. Sorrells. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLOS ONE*, 8(9):e74612, 2013.

N. Heslot, J.-L. Jannink, and M. E. Sorrells. Perspectives for genomic selection applications and research in plants. *Crop Science*, 55(1):1–12, 2015.

N. Hofheinz and M. Frisch. Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3: Genes— Genomes— Genetics*, 4(3):539–546, 2014.

N. Hofheinz, D. Borchardt, K. Weissleder, and M. Frisch. Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theoretical and Applied Genetics*, 125(8):1639–1645, 2012.

H. U. Jan, A. Abbadi, S. Lücke, R. A. Nichols, and R. J. Snowdon. Genomic prediction of testcross performance in canola (*Brassica napus*). *PLOS ONE*, 11(1):e0147769, 2016.

Y. Jiang and J. C. Reif. Modeling epistasis in genomic selection. *Genetics*, 201(2):759–768, 2015.

E. Jonas and D.-J. de Koning. Does genomic selection have a future in plant breeding? *Trends in Biotechnology*, 31(9):497–504, 2013.

D. Jordan, Y. Tao, I. Godwin, R. Henzell, M. Cooper, and C. McIntyre. Prediction of hybrid performance in grain sorghum using RFLP markers. *Theoretical and Applied Genetics*, 106(3):559–567, 2003.

R. Lande and R. Thompson. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756, 1990.

B. L. L. Lanza, L. C. de Souza Jr., M. L. M. Ottoboni, C. M. L. Vieira, and P. A. de Souza. Genetic distance of inbred lines and prediction of maize single-cross performance using RAPD markers. *Theoretical and Applied Genetics*, 94(8):1023–1030, 1997.

C. Lehermeier, N. Krämer, E. Bauer, C. Bauland, C. Camisan, L. Campo, P. Flament, A. E. Melchinger, M. Menz, N. Meyer, et al. Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. *Genetics*, 198(1):3–16, 2014.

A. J. Lorenz, S. Chao, F. G. Asoro, E. S. Heffner, T. Hayashi, H. Iwata, K. P. Smith, M. E. Sorrells, and J.-L. Jannink. Genomic selection in plant breeding: Knowledge and prospects. *Advances in Agronomy*, 110:77–123, 2011.

R. E. Lorenzana and R. Bernardo. Accuracy of genotypic predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, 120(1):151–161, 2009.

P. A. Marsan, P. Castiglioni, F. Fusari, M. Kuiper, and M. Motto. Genetic diversity and its relationship to hybrid performance in maize as revealed by RFLP and AFLP markers. *Theoretical and Applied Genetics*, 96(2): 219–227, 1998.

J. M. Massman, A. Gordillo, R. E. Lorenzana, and R. Bernardo. Genomewide predictions from maize single-cross data. *Theoretical and Applied Genetics*, 126(1):13–22, 2013.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4): 1819–1829, 2001.

S. Michel, C. Ametz, H. Gungor, D. Epure, H. Grausgruber, F. Löschenberger, and H. Buerstmayr. Genomic selection across multiple breeding cycles in applied bread wheat breeding. *Theoretical and Applied Genetics*, 129(6):1179–1189, 2016.

N. Philipp, G. Liu, Y. Zhao, S. He, M. Spiller, G. Stiewe, K. Pillen, J. C. Reif, and Z. Li. Genomic prediction of barley hybrid performance. *Plant Genome*, 9(2), 2016.

H.-P. Piepho. Ridge regression and extensions for genomewide selection in maize. *Crop Science*, 49(4):1165–1176, 2009.

J. Reif, Y. Zhao, T. Würschum, M. Gowda, and V. Hahn. Genomic prediction of sunflower hybrid performance. *Plant Breeding*, 132(1):107–114, 2013.

C. Riedelsheimer, A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer, and A. E. Melchinger. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics*, 44(2):217–220, 2012.

C. Riedelsheimer, J. B. Endelman, M. Stange, M. E. Sorrells, J.-L. Jannink, and A. E. Melchinger. Genomic predictability of interconnected biparental maize populations. *Genetics*, 194(2):493–503, 2013.

T. A. Schrag, J. M. Möhring, H. P. Maurer, B. S. Dhillon, A. E. Melchinger, H.-P. Piepho, A. P. Sorensen, and M. Frisch. Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. *Theoretical and Applied Genetics*, 118(4):741–751, 2009.

X. Shen, M. Alam, F. Fikse, and L. Rönnegård. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, 2013.

G. H. Shull. The composition of a field of maize. *Journal of Heredity*, os-4 (1):296–301, 1908. doi: 10.1093/jhered/os-4.1.296.

J. Smith, T. Hussain, E. Jones, G. Graham, D. Podlich, S. Wall, and M. Williams. Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops. *Molecular Breeding*, 22(1):51–59, 2008.

F. Technow, C. Riedelsheimer, T. A. Schrag, and A. E. Melchinger. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theoretical and Applied Genetics*, 125(6):1181–1194, 2012.

F. Technow, A. Bürger, and A. E. Melchinger. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3: Genes— Genomes— Genetics*, 3(2):197–203, 2013.

F. Technow, T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer, and A. E. Melchinger. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics*, 197(4): 1343–1355, 2014.

A. Thiemann, J. Fu, T. A. Schrag, A. E. Melchinger, M. Frisch, and S. Scholten. Correlation between parental transcriptome and field data for the characterization of heterosis in *Zea mays* L. *Theoretical and Applied Genetics*, 120(2):401–413, 2010.

A. Thiemann, J. Fu, F. Seifert, R. T. Grant-Downton, T. A. Schrag, H. Pospisil, M. Frisch, A. E. Melchinger, and S. Scholten. Genome-wide meta-analysis of maize heterosis reveals the potential role of additive gene expression at pericentromeric loci. *BMC Plant Biology*, 14(1):88, 2014.

S. Unterseer, E. Bauer, G. Haberer, M. Seidel, C. Knaak, M. Ouzunova, T. Meitinger, T. M. Strom, R. Fries, H. Pausch, et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*, 15(1):823, 2014.

Y. Wang, M. F. Mette, T. Miedaner, M. Gottwald, P. Wilde, J. C. Reif, and Y. Zhao. The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. *BMC Genomics*, 15(1):556, 2014.

V. Wimmer, C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang, and C.-C. Schön. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, 195(2):573–587, 2013.

V. S. Windhausen, G. N. Atlin, J. M. Hickey, J. Crossa, J.-L. Jannink, M. E. Sorrells, B. Raman, J. E. Cairns, A. Tarekegne, K. Semagn, et al. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3: Genes— Genomes— Genetics*, 2(11):1427–1436, 2012.

T. Würschum, J. C. Reif, T. Kraft, G. Janssen, and Y. Zhao. Genomic selection in sugar beet breeding populations. *BMC Genetics*, 14(1):85, 2013.

S. Xu. Theoretical basis of the Beavis effect. *Genetics*, 165(4):2259–2268, 2003.

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

C. Zenke-Philippi, A. Thiemann, F. Seifert, T. Schrag, A. E. Melchinger, S. Scholten, and M. Frisch. Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles. *BMC Genomics*, 17(1):262, 2016.

C. Zenke-Philippi, M. Frisch, A. Thiemann, F. Seifert, T. Schrag, A. E. Melchinger, S. Scholten, and E. Herzog. Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding program. *Plant Breeding*, In Press., 2017.

Y. Zhao, M. Gowda, W. Liu, T. Würschum, H. P. Maurer, F. H. Longin, N. Ranc, and J. C. Reif. Accuracy of genomic selection in European maize elite breeding populations. *Theoretical and Applied Genetics*, 124(4):769–776, 2012a.

Y. Zhao, M. Gowda, F. H. Longin, T. Würschum, N. Ranc, and J. C. Reif. Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theoretical and Applied Genetics*, 125(4):707–713, 2012b.

Y. Zhao, M. Gowda, W. Liu, T. Würschum, H. P. Maurer, F. H. Longin, N. Ranc, H.-P. Piepho, and J. C. Reif. Choice of shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations. *Plant Breeding*, 132(1):99–106, 2013a.

Y. Zhao, M. Gowda, T. Würschum, C. F. H. Longin, V. Korzun, S. Kollers, R. Schachschneider, J. Zeng, R. Fernando, J. Dubcovsky, and J. C. Reif.

Dissecting the genetic architecture of frost tolerance in Central European winter wheat. *Journal of Experimental Botany*, 64(14):4453–4460, 2013b.

Y. Zhao, J. Zeng, R. Fernando, and J. C. Reif. Genomic prediction of hybrid wheat performance. *Crop Science*, 53(3):802–810, 2013c.

Y. Zhao, M. F. Mette, M. Gowda, C. F. H. Longin, and J. C. Reif. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity*, 112(6):638–645, 2014.

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

Acknowledgments

I am grateful to my academic supervisor Prof. Dr. Matthias Frisch for his continuous support, many suggestions and his advise during my thesis work.

Many thanks to Prof. Dr. Rod Snowdon for being my second supervisor.

Thanks to all my colleagues at the institute for Biometry for the nice working atmosphere and their help in all circumstances, especially to Dr. Eva Herzog for proof-reading this manuscript, to Renate Schmidt for her help with everything to do with organisation and to my office mate Dr. Nathalie Steiner.

Finally, I would like to thank my husband, Sebastian Philippi, for his support during the last years, my mother, Heike Wille-Zenke, for her patience with my changes of direction, and my father, Georg Zenke, as well as my stepfather, Karl Wille, and my grandparents for awakening my interest in agriculture.

Eidesstattliche Erklärung

Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe.

Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.

Gießen, 12. April 2017 Carola Anna Luise Zenke-Philippi