Theor Appl Genet (2012) 124:825–833
DOI 10.1007/s00122-011-1747-9

ORIGINAL PAPER

Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data

Junjie Fu · K. Christin Falke · Alexander Thiemann · Tobias A. Schrag · Albrecht E. Melchinger · Stefan Scholten · Matthias Frisch

Received: 1 August 2011 / Accepted: 28 October 2011 / Published online: 19 November 2011

Abstract The performance of hybrids can be predicted with gene expression data from their parental inbred lines. Implementing such prediction approaches in breeding programs promises to increase the efficiency of hybrid breeding. The objectives of our study were to compare the accuracy of prediction models employing multiple linear regression (MLR), partial least squares regression (PLS), support vector machine regression (SVM), and transcriptome-based distances (D_B). For a factorial of 7 flint and 14 dent maize lines, the grain yield of the hybrids was assessed and the gene expression of the parental lines was profiled with a 56k microarray. The accuracy of the prediction models was measured by the correlation between predicted and observed yield employing two cross-validation schemes. The first modeled the prediction of hybrids when testcross data are available for both parental lines (type 2 hybrids), and the second modeled the prediction of hybrids when no testcross data for the parental lines were available (type 0 hybrids). MLR, SVM, and PLS resulted in a high correlation between predicted and observed yield for type 2 hybrids, whereas for type 0 hybrids D_B had greater prediction accuracy. The regression methods were robust to the choice of the set of profiled genes and required only a few hundred genes. In contrast, for an accurate hybrid prediction with D_B, 1,000–1,500 genes were required, and the prediction accuracy depended strongly on the set of profiled genes. We conclude that for prediction within one set of genetic material MLR is a promising approach, and for transfering prediction models from one set of genetic material to a related one, the transcriptome-based distance D_B is most promising.

Communicated by J. Yu.

J. Fu and K. C. Falke contributed equally to this work.

J. Fu
Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

K. C. Falke
Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany

A. Thiemann · S. Scholten
Biocenter Klein Flottbek, Developmental Biology and Biotechnology, University of Hamburg, 22609 Hamburg, Germany

T. A. Schrag · A. E. Melchinger
Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany

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

Introduction

The prediction of the performance of a hybrid with data gathered from its parental inbred lines is expected to increase the efficiency of hybrid breeding. Recently suggested prediction methods using field data, coancestry coefficients, and DNA markers such as AFLPs and SSRs, were reviewed by Schrag et al. (2009). Correlation of heterosis with the average gene expression in the parental inbred lines was suggested by the study of Springer and Stupar (2007). This approach was taken up by Fu et al. (2010) and Thiemann et al. (2010). An alternative approach summarizes the differential gene expression in the parental lines of a hybrid by defining transcriptome-based

distance measures, and uses the distances to predict hybrid performance (Frisch et al. 2010).

The data set employed by Frisch et al. (2010) was used for DNA-marker based hybrid prediction (Schrag et al. 2006). A ceteris paribus comparison showed that using transctriptome data resulted in a considerably more precise prediction than AFLP markers and also than GCA estimates obtained from field trials. In a furhter study in which an extended version of this data set was used (Schrag et al. 2009), DNA-marker based prediction of hybrid performance showed a greater prediction accuracy than BLUP approaches based on pedigree and phenotypic data (Bernardo 1994, 1999). These comparisons indicate that transcriptome-based prediction might outperform GCA, pedigree-based BLUP, and DNA marker-based prediction of hybrid performance.

To address the over-parametrization in prediction models for heterosis and hybrid performance, support vector machine regression (SVM) was suggested by Maenhout et al. (2007) in a study with maize. Partial least squares regression (PLS) was suggested in studies with metabolites in *Arabidopsis* by Gärtner et al. (2009) and Steinfath et al. (2010). A comparative study applying these different prediction methods to one single data set is not yet available.

The goal of our study was to investigate the accuracy of predicting grain yield of maize hybrids with gene expression data from their parental inbred lines. In particular, our objectives were to (1) compare models employing binary transcriptome-based distances (D_B), SVM, PLS, and standard multiple linear regression (MLR), (2) investigate the effect of the number of genes used in the prediction models, (3) compare prediction when testcross data for the parental lines of a hybrid are available with prediction when no testcross data for the parental lines are available.

Materials and methods

Field data

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 (S028, S036, S044, S046, S049, S050, S058, and S067) and six with Iodent background (P033, P040, P046, P048, P063, and P066). Four flint lines (F037, F039, F043, and F047) had a European Flint background and three (L024, L035, and L043) a Flint/Lancaster background.

The factorial crosses were evaluated in 2002 at six agroecologically diverse locations in Germany (Bad Krozingen, Eckartsweier, Hohenheim, Landau, Sünching, and 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. The field data were analyzed with a mixed linear model, which was described in detail in a previous study (Schrag et al. 2009), where it was referred to as Experiment 1.

Gene expression data

Five seedlings of each of the 21 maize inbred lines were grown in a climate chamber under regulated growth conditions. RNA was isolated from a mixture of the seedlings of each line when they were 7 days old. The 46k array from the maize oligonucleotide array project (http://www.maizearray.org/, University of Arizona, USA) was used for transcription profiling (Thiemann et al. 2010). For the microarray experiment an interwoven loop design (Kerr and Churchill 2001) 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 microarray experiment, two genes in eight different lines were evaluated by Quantitative RT-PCR, essentially in accordance with the microarray data. The microarray 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 (Smyth 2004) was applied for the tests. A false discovery rate (FDR; Benjamini and Hochberg 1995) of 0.01 for all genes showing a fold change of at least 1.3 was used to detect significant differential expression between inbred lines (Fu et al. 2010). For all differentially expressed genes, we calculated the average of the gene expression level (log2 scale) in the parents of each hybrid.

Prediction methods

The prediction of new hybrids requires a set of related breeding material. With this estimation set, the gene expression of the parental lines and the hybrid performance is assessed, and the parameters for the employed prediction method are estimated. An outline of the prediction procedure is presented in Fig. 1 of Frisch et al. (2010).

We employed the binary transcriptome-based distance D_B that quantifies the number of genes that were differentially expressed in two parental lines. It does not take into account the absolute amount of the differences in gene expression. This avoids a bias in the selection of genes toward those genes for which the difference in gene

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

expression is greatest. Further, prediction with D_B showed a greater prediction acurracy than prediction with the Euclidean transcriptome-based distance D_E (Frisch et al. 2010).

For the calculation of D_B, the genes were ranked according to the p value of a test for the association of differential gene expression with high hybrid performance. A set of genes with small p values was used to estimate the transcriptome-based distance, which was related to hybrid performance with linear regression. For a detailed description see Frisch et al. (2010).

For the regression techniques, the genes were ranked according to the significance of the correlation between the average expression level of a gene in the parental lines and hybrid performance. The significance of the correlations was tested with a t test adjusted for multiple testing using a false discovery rate of 0.01. Sets of genes with small p values were used as predictors in the regression models.

PLS is a regression technique using latent variables that are chosen such that the correlation between predictors and response is maximized. The original gene expression data are projected into a latent lower dimensional space with multidimensional scaling, and the first n latent variables are used for MLR. n was determined with cross-validation such that adding further latent variables to the prediction model did not decrease the mean squared prediction error. The calculations were carried out with the R package *pls* (Mevik and Wehrens 2007).

SVM is a "machine-learning" technique (Drucker et al. 1997) that maps an n-dimensional vector of predictors to the target variable. We used ϵ-insensitive SVR as implemented in the R package *e*1071 (Karatzoglou et al. 2006). The optimal values of the parameters ϵ, C and γ that were required for the Gaussian Radial Basis Function kernel were determined with a grid search (Hsu et al. 2003) using the *tune* function. The parameter space for γ ranged from 10^{-6} to 10^{-3}.

MLR with forward selection of regressors on basis of Akaike information criterion (AIC) was carried out with the *step* and *lm* functions of R (Ihaka and Gentleman, 1996).

Assessment of prediction accuracy

The accuracy of hybrid performance prediction was evaluated for (a) untested hybrids of a partial factorial and (b) hybrids derived from parental lines for which no testcross data are available.

For prediction of untested hybrids of a partial factorial, we employed the cross-validation procedure of Schrag et al. (2009). The estimation set consisted of three randomly chosen flint and five randomly chosen dent lines and their hybrids, and the validation set consisted of the remaining hybrids of a 7×14 factorial. The principle is illustrated in Fig. 1a. Both parental lines of an untested hybrid in the validation set are also parents of hybrids belonging to the estimation set. In the terminology of Schrag et al. (2009), the hybrids of the validation set are called type 2 hybrids, because testcross data are available for both parental lines of a hybrids.

For prediction of hybrids derived from parental lines for which no testcross data were available, we employed a cross-validation procedure in which 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 (Fig. 1b). The hybrids of the validation set are called type 0 hybrids, because for none of the parental lines testcross data are available.

For each scenario to be evaluated, cross-validation was carried out for 100 runs and the prediction accuracy in the validation set was measured by the correlation $r(y, \hat{y})$ between the predicted and the observed hybrid yield.

Sets of genes used for prediction

The genes were ranked according to the p value of the test for association with hybrid yield as described above. For

convenience, we call sets of genes consisting of the n genes with the smallest p value, as the "n best genes".

To investigate the number of genes required for the different prediction methods, hybrid performance was predicted with D_B, SVR, and PLS employing the best 50, 100, 200, 500, 1,000, 1,500, 2,000, 3,000, 4,000, and 5,000 genes. For MLR, sets of the best 50, 100, and 200 genes were used. Due to the employed forward selection algorithm, only genes explaining a large proportion of the variance are included in the model. Therefore, the larger sets of genes are expected to result in the same final MLR models as these three sets.

To investigate how important an optimal ranking of the genes is, prediction on basis of D_B, PLS, and SVM was carried out with (1) the best 200 genes, (2) the second best 200 genes, (3) the third best 200 genes, (4) 200 random genes out of the best 1,000 genes, (5) 200 random genes out of the best 5,000 genes, (6) the best 1,000 genes, (7) the second best 1,000 genes, (8) the third best 1,000 genes, (9) 1,000 random genes out of the first 5,000 genes, and (10) 1,000 random genes out of all genes with differential gene expression. For prediction with MLR only the sets (1)–(5) were used.

Results

The mean grain yield of the 98 hybrids was 11.72 Mg ha^{-1} with a broad sense heritability of 80.3%. The GCA and SCA variance components, as well as their interactions with the locations were significantly different from zero ($\alpha = 0.05$). The ratio of SCA:GCA variance components was 1.12. The field data were presented in detail by Schrag et al. (2006).

10,810 genes were differentially expressed in at least one pair of parental lines of the factorial crosses. Thiemann et al. (2010) as well as Fu et al. (2010) presented lists of genes of which the average gene expression in the parental lines was correlated with heterotic traits and provided a functional characterization of heterosis.

For untested hybrids in an incomplete factorial (type 2 hybrids), prediction with the regression methods resulted in a greater correlation between predicted and observed yield than prediction with transcriptome-based distances (Fig. 2). The differences between the regression methods were small, and the correlations were almost not affected by the number of genes employed for prediction. In contrast, for prediction with transcriptome-based distances, the correlation between predicted and observed yield was greatest when 1,000–1,500 genes were used for prediction.

The correlation between predicted and observed yield was smaller for hybrids derived from parental lines of which no testcross data are available (type 0 hybrids) than for type 2 hybrids (Fig. 3). Furthermore, the correlation coefficients determined in the cross-validation runs had a considerable greater range for type 0 hybrids than for type 2 hybrids.

For type 0 hybrids, prediction with transcriptome-based distances showed a greater correlation between predicted and observed yield than prediction with the regression methods (Fig. 3). Within the regression methods, PLS and SVM showed a greater correlation between observed and predicted yield than MLR. The low correlation with MLR was accompanied by very large absolute values of the prediction errors (results not shown).

For prediction with transcriptome-based distances, the choice of the set of genes employed as predictors greatly affected the observed correlations between predicted and observed yield. Sets of genes with a highly significant association with hybrid performance showed greater prediction accuracy than sets of genes with lower significances and sets of randomly selected genes (Figs. 2, 3). In contrast, for the regression-based methods the set of genes employed for prediction had only a marginal effect on the correlation between observed and predicted yield.

Discussion

Multiple linear regression

MLR with forward selection of predictor variables was included in our study to investigate whether the more sophisticated regression methods PLS and SVM were superior to standard methods. Forward selection includes predictor variables to the model until no further improvement of the fit is detected by the AIC. Therefore, the number of variables included is solely controlled by the algorithm but not by the user. This property distinguishes MLR from the other investigated prediction methods, where the number of predictor variables in the model can be determined by the user. With high multicollinearity, the number of predictor variables included by forward selection into a MLR model is expected to be low. This was observed with our data set, where the final model consisted of seven to ten predictor variables in most of the cross-validation runs.

It had only marginal effect on the prediction accuracy whether the best 50 or 200 genes, or even random sets of 200 genes among the best 1,000 or 5,000 genes were used as starting set for the model selection (Figs. 2, 3). This can be explained by the low number of genes included in the final model and the high multicollinearity of gene expression in our data set. The combination of both resulted in final models with comparable prediction accuracy that were found for different starting sets of genes. We

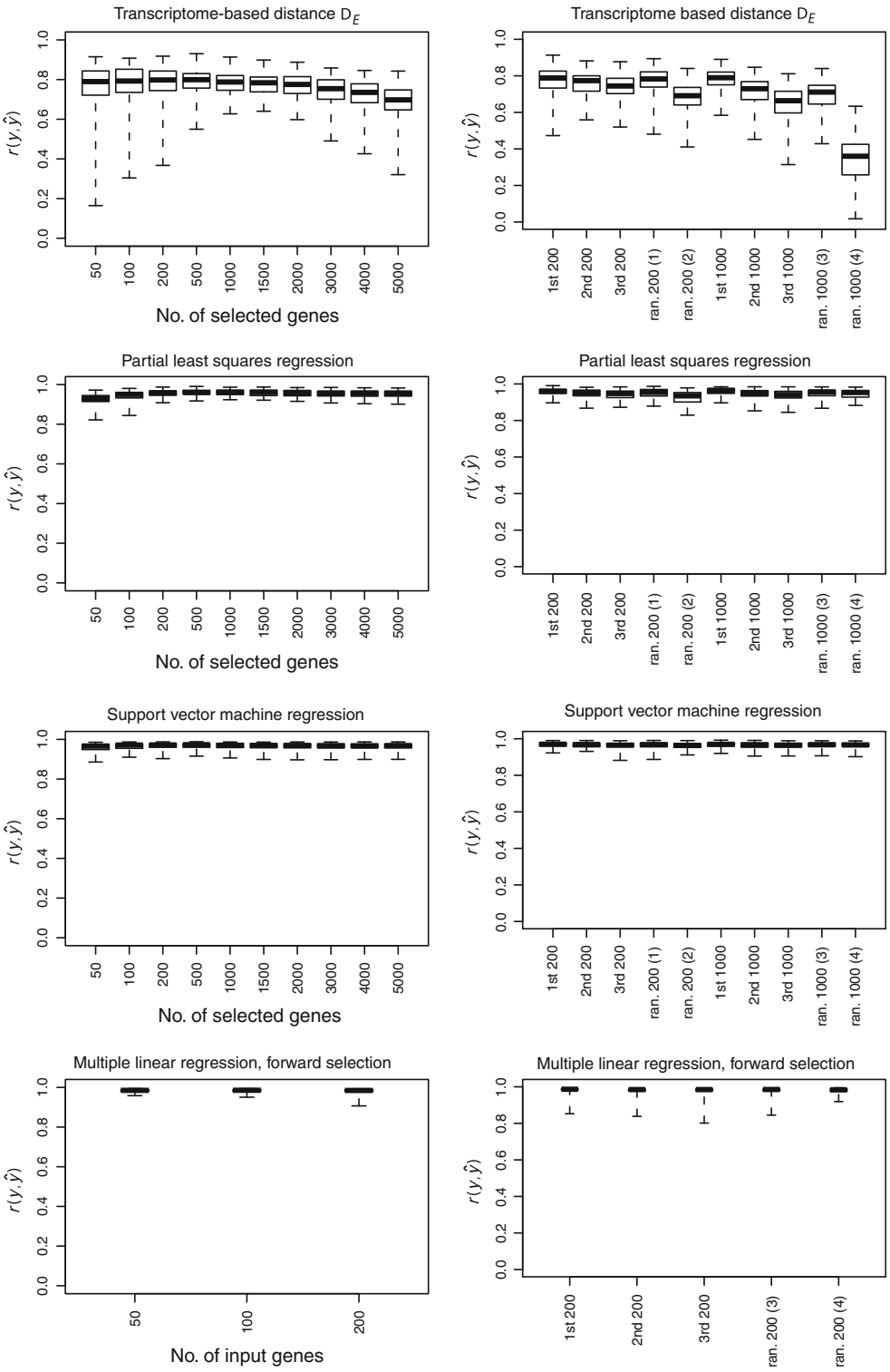

Fig. 2 Prediction accuracy for untested hybrids in an incomplete factorial (3/5 cross validation of type 2 hybrids). Correlation $r(y,\hat{y})$ between predicted and observed hybrid yield for the different prediction methods. *Left* best 50 to 5,000 genes. *Right* sets consisting of the best, second best, and third best 200 genes, 200 random genes among the best 1,000 (1), 200 random genes among the best 5,000 (2), best, second best, and third best 1,000 genes, 1,000 random genes among the best 5,000 (3), and among all differentially expressed genes (4)

conclude that MLR is robust to the choice of genes employed as starting set for model selection.

The prediction accuracy of MLR was considerably lower for type 0 hybrids than for type 2 hybrids. While a lower prediction accuracy for type 0 hybrids was also observed with the other prediction methods, the difference between the two types of hybrids was greatest for MLR (Figs. 2, 3). This can be interpreted as an indicator for a low transferability of the genes selected for prediction with MLR from one set of genetic material to another set. The coincidence of the low number of predictor variables in MLR and the large differences in

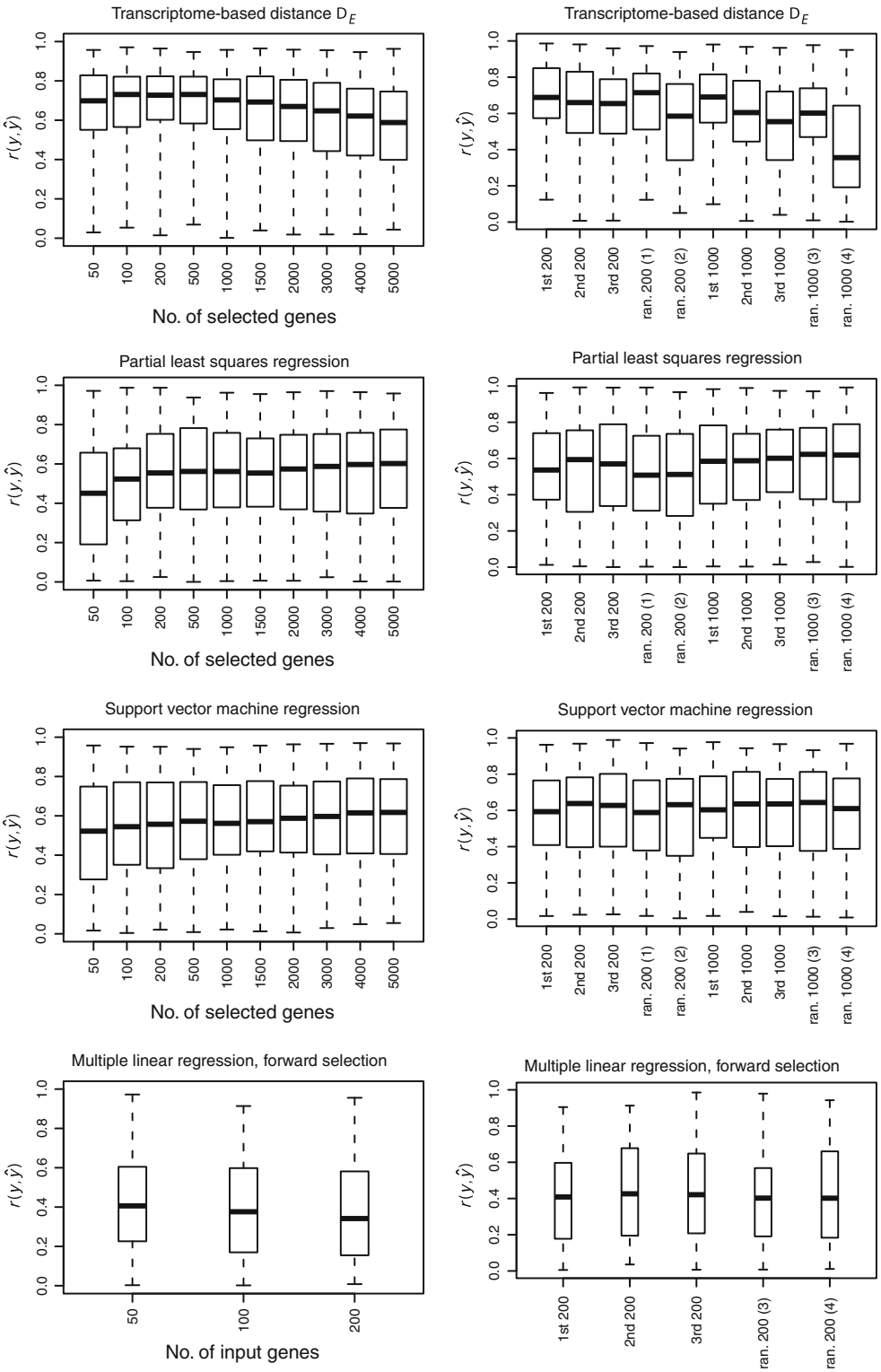

Fig. 3 Prediction accuracy for hybrids derived from parental lines of which no testcross data are available (5/10 cross validation of type 0 hybrids). Correlation $r(y, \hat{y})$ between predicted and observed hybrid yield for the different prediction methods. *Left* best 50 to 5,000 genes. *Right* sets consisting of the best, second best, and third best 200 genes, 200 random genes among the best 1,000 (1), 200 random genes among the best 5,000 (2), best, second best, and third best 1,000 genes, 1,000 random genes among the best 5,000 (3), and among all differentially expressed genes (4)

prediction accuracy between type 0 and type 2 hybrids suggests the hypothesis that the low transferability may be a consequence of the low number of predictors in the model. Analysis of further data sets is needed to investigate this hypothesis.

In our data set, a complete regression model including expression of all 10,800 differentially expressed genes would be heavily over-parameterized. Therefore, it might be suspected that just by chance there could be a high correlation between hybrid performance and the expression

of some genes (or linear combinations of them). We checked this hypothesis with randomized data sets, where the yield values were randomly assigned to the hybrids. For these randomized data sets no genes were found that showed a significant correlation with yield (results not shown). Hence, our data provide no evidence that the heavy over-parameterization in combination with high multicollinearity in gene expression data exclude MLR in general as a suitable method for transcriptome-based prediction of hybrid performance.

Partial least squares regression and support vector machine regression

Hybrid prediction with PLS and SVM was as accurate as prediction with MLR for type 2 hybrids. For type 0 hybrids, PLS and SVM were clearly superior to MLR and the prediction accuracy increased slightly with increasing number of genes used for prediction (Figs. 2, 3). Hence, for PLS and SVM the transferability of estimation parameters from one set of genetic material to another is better than for MLR, and selecting only small numbers of genes for prediction seems not to be desirable. In line with the hypothesis that the low number of genes in MLR models is responsible for the poor transferability, the greater number of predictors with SVM and PLS might be responsible for the better transferability.

In conclusion, the regression-based methods provided for type 2 hybrids a better prediction accuracy than transcriptome-based distances (Fig. 2) and DNA marker-based prediction (Schrag et al. 2006). Moreover, the possibility of specifying the numbers of genes in the model can be regarded as an advantage of SVR and PLS compared with MLR, resulting in a more robust prediction with respect to the transferability to new genetic materials.

Transcriptome-based distances

For transcriptome-based distances, the correlation of predicted with observed hybrid yield depended strongly on the number of genes included in the model. The optimum number was 1,000–1,500 genes for type 0 hybrids and 500 for type 2 hybrids. The set of genes employed for prediction strongly affected the prediction accuracy (Figs. 2, 3). Hence, in contrast to the regression methods, the selection of a suitable set of genes for prediction is of high importance with transcriptome-based distances. With MLR most of the genes available for prediction were not included in the model, and with SVR and PLS some genes may have a low weight, even if they were included in the model. In contrast, with binary transcriptome-based distances, all genes included in the model contribute with equal weights to the prediction. Therefore, including genes that explain only a small part of the variation of the target variable results in a decline of the prediction accuracy. This explains the sensitivity of prediction with transcriptome-based distances with respect to the employed set of genes.

For prediction of type 2 hybrids, the regression methods provided a greater prediction accuracy than transcriptome-based distances, whereas for type 0 hybrids transcriptome-based distances were superior. Prediction of type 2 hybrids can be regarded as prediction within one set of genetic material. Within one set of genetic material, only some of the loci that are underlying a quantitative trait may be highly polymorphic, whereas at other loci underlying the trait only a low level of polymorphism may occur. In consequence, only those loci that show a high degree of polymorphism explain a large proportion of the variation in the target variable and were, therefore, included into regression models with a high weight. Excluding genes with small degree of polymorphism from the model has no consequences on the prediction accuracy when prediction is carried out within one set of genetic material. However, excluding them when the prediction model is transferred to a new set of genetic material may affect the prediction accuracy. The variability of these genes in the new set of breeding material may be greater than in the estimation set, and hence, excluding them from the model may reduce the proportion of variance explained by the model and result in a lower accuracy.

The stringent selection of genes with MLR and the weighing of predictor genes with PLS and SVM is expected not to include genes in the prediction models that are underlying the trait but have a low degree of polymorphism in the estimation set. In contrast, the lower stringency in selection of genes for calculating transcriptome-based distances is expected to include such genes. If their degree of polymorphism in the new set of breeding material to which the prediction parameters are transferred is greater than in the estimation set, then these genes contribute to a precise prediction of new hybrids. This effect could explain the better transferability of prediction models with transcriptome-based distances to new sets of breeding materials and at the same time the superiority of regression methods for type 2 hybrids.

The high transferability of transcriptome-based distance models from one set of genetic material to another corresponds well to the infinitesimal model of quantitative genetics. This model postulates that a large number of genes with small effects are underlying quantitative traits like yield, which agrees with the observation that 1,000–1,500 genes with small effects of equal size explain a high proportion of the variation in maize yield. This observation has an important consequence on the functional modeling of complex traits like grain yield. If models are set up with genetic material in which only a part

of the genes underlying the trait are varying, then these models may have a high explanatory value for the data set under consideration, as demonstrated by the high prediction accuracy of MLR with type 2 hybrids. However, drawing general conclusions on the mechanisms underlying a trait from such a study could be difficult, because only a part of the actual genetic architecture of the trait can be assessed and modeled.

Application in breeding programs

For the possible application of transcriptome-based prediction models in breeding programs, the prediction of type 0 and type 2 hybrids needs to be considered separately. For prediction of type 2 hybrids, the MLR approach combines several favorable properties. It assures a high prediction accuracy with standard statistical methods that are implemented in any statistical software and does not need advanced programs or software packages. The approach is very robust with respect to the number of genes employed for prediction and also the set of genes of which transcript abundance was assessed. Hence, application could be possible even with only a few hundred genes. To determine such a set of genes, preliminary experiments seem to be suitable in which the expression of a large number of genes is assessed and the genes related to the trait under consideration are determined. From those a small subset is selected. While it needs to be assured that the genes of this subset are related to the trait under consideration, high efforts for selecting the best among them seems not necessary.

For prediction of type 0 hybrids, transcriptome-based distances provide a sufficient prediction accuracy, whereas the regression-based methods lack the transferability of models from one set of genetic material to another. The approach requires expression profiling of considerably more genes than for prediction of type 2 hybrids with MLR and, moreover, the selection of a suitable set of genes is important for a high prediction accuracy. This suggests that the estimation set should be genetically related to the new set of hybrids, which should be predicted. The inbred lines of a previous cycle of a breeding program seem to be a suitable choice.

Our results confirmed that transcriptome-based approaches have a high potential for prediction of hybrid performance. However, it is questionable whether microarray technology can be used to generate sufficient information on the transcriptome economically. Analysis of gene expression by next generation RNA sequencing approaches is expected to replace microarray experiments and the costs per sequence read of these technologies are dropping rapidly. This might provide the possibility to implement transcriptome-based prediction and exploit the additional information that is present in transcriptome data in plant breeding programs.

Acknowledgments This research was funded by the Deutsche Forschungsgemeinschaft (grants no. FR 1615/4-1, ME 2260/5-1, SCHO 764/6-1). We thank Gregory Mahone for helpful comments on the manuscript.

References

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

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

Bernardo R (1999) Marker-assisted best linear unbiased prediction of single-cross performance. Crop Sci 39:1277–1282

Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, pp 155–161

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

Fu J, Thiemann A, Scholten S, Schrag T, Melchinger AE, Frisch M (2010) 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 10:63

Gärtner T, Steinfath M, Andorf S, Lisec J, Meyer R, Altmann T, Willmitzer L, Selbig J (2009) Improved heterosis prediction by combining information on DNA- and metabolic markers. PLoS One 4:e5220

Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Department of Computer Science and Information Engineering. National Taiwan University, Taipei

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput Graph Stat 5:299–314

Karatzoglou A, Meyer D, Hornik K (2006) Support vector machines in R. J Stat Softw 15:1–28

Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2:183–201

Maenhout S, Baets BD, Haesaert G, Bockstaele EV (2007) Support vector machine regression for the prediction of maize hybrid performance. Theor Appl Genet 115:1003–1013

Mevik B, Wehrens R (2007) The pls package: principal component and partial least squares regression in R. J Stat Softw 18:1–23

Schrag TA, Melchinger AE, Sorensen AP, Frisch M (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 TA, Möhring JM, Maurer HP, Dhillon BS, Melchinger AE, Piepho HP, Sorensen AP, Frisch M (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 GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:3

Springer NM, Stupar RM (2007) Allelic variation and heterosis in maize: how do two halves make more than a whole? Genome Res 17:264–275

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

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