Theor Appl Genet (2010) 120:441–450
DOI 10.1007/s00122-009-1204-1

ORIGINAL PAPER

# Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize

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

Received: 7 May 2009 / Accepted: 21 October 2009 / Published online: 13 November 2009

**Abstract** Grouping of germplasm and prediction of hybrid performance and heterosis are important applications in hybrid breeding programs. Gene expression analysis is a promising tool to achieve both tasks efficiently. Our objectives were to (1) investigate distance measures based on transcription profiles, (2) compare these with genetic distances based on AFLP markers, and (3) assess the suitability of transcriptome-based distances for grouping of germplasm and prediction of hybrid performance and heterosis in maize. We analyzed transcription profiles from seedlings of the 21 parental maize lines of a 7 × 14 factorial with a 46-k oligonucleotide array. The hybrid performance and heterosis of the 98 hybrids were assessed in field trials. In cluster and principal coordinate analyses for germplasm grouping, the transcriptome-based distances were as powerful as the genetic distances for separating flint from dent inbreds. Cross validation showed that prediction of hybrid performance with transcriptome-based distances using selected markers was more precise than earlier prediction models using DNA markers or general combining ability estimates using field data. Our results suggest that transcriptome-based prediction of hybrid performance and heterosis has a great potential to improve the efficiency of maize hybrid breeding programs.

## Introduction

The prediction of hybrid performance using information from parental inbred lines is of great interest to breeders. If successful, it can increase substantially the efficiency of breeding programs. Prediction methods using the genetic distance between the parental lines failed to predict reliably the hybrid performance of inter-group hybrids in plant breeding programs (Melchinger 1999). In contrast, prediction methods using markers linked to quantitative trait loci affecting the trait under consideration were successfully developed (Vulysteke et al. 2000; Schrag et al. 2006, 2007, 2009a, b).

In hybrid breeding, the germplasm is usually divided into genetically distant heterotic pools. Molecular marker-based genetic distances, graphically displayed by multivariate statistical methods, such as principal coordinate and cluster analyses can be helpful to accomplish this task (Reif et al. 2003, 2005).

With the advent of transcriptome analysis by gene expression profiling, a new lab technology has emerged. It can be employed in studying the molecular basis of heterosis (Bircheler et al. 2003). Stupar et al. (2008) found a correlation of genetic diversity and transcriptional variation. Guo et al. (2006) suggested that differential allele regulation may play an important role for heterosis. Springer and Stupar (2007) suggested that modified levels of gene expression in hybrids may contribute to heterotic

Communicated by A. Charcosset.

Contribution to the special issue “Heterosis in Plants”.

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

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

A. Thiemann · S. Scholten
Biocenter Klein Flottbek, Developmental Biology
and Biotechnology, University of Hamburg,
Ohnhorststrasse 18, 22609 Hamburg, Germany

phenotypes. The applicability of expression profiles, SNP markers, and metabolites for prediction of hybrid performance and heterosis was recently investigated with various approaches (Maenhout et al. 2009; Repsilber et al. 2009; Steinfath et al. 2009).

A primary focus of interest is to use gene expression data in functional analyses for detection of genes underlying agronomic traits (Thiemann et al. 2009). An alternative view on gene expression data is possible by disregarding all functional information on the analyzed genes and considering the transcript abundance levels as quantitative variables characterizing a genotype. These quantitative variables could then be used to construct distance measures between genotypes on the basis of their transcription profiles. In combination with multivariate methods, the distances could be employed for grouping of germplasm and in combination with linear models for prediction of hybrid performance and heterosis. To our knowledge, no previous investigation on transcriptome-based distance measures and their applications is available.

The goal of our study was to investigate the potential application of transcriptome-based distance measures in maize hybrid breeding programs. In particular, our objectives were to (a) investigate distance measures between inbred lines based on gene expression profiles, (b) examine their correlation with molecular marker-based genetic distances, and (c) assess the suitability of transcriptome-based distances for germplasm grouping and prediction of heterosis and hybrid performance.

## Materials and methods

### Field data

Seven flint and 14 dent elite inbred lines developed by the maize breeding program of the University of Hohenheim were used as parental lines for a 7 × 14 factorial mating design. The inbreds comprised eight dent lines with Iowa Stiff Stalk Synthetic (S028, S036, S044, S046, S049, S050, S058, S067) and six with Iodent background (P033, P040, P046, P048, P063, P066). Four flint lines (F037, F039, F043, F047) had an European Flint and three (L024, L035, 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, Vechta). The 21 inbred parents were evaluated for their per se performance in 2003 at four locations (Eckartsweier, Hohenheim, Sünching, Pocking) and in 2004 at three locations (Eckartsweier, Hohenheim, Bad Krozingen). The trials were evaluated in two-row plots using adjacent $\alpha$ designs (generalized lattices) with two to three replications. Grain yield (Mg ha$^{-1}$) adjusted to 155 g kg$^{-1}$, grain moisture, and grain dry matter concentration (%) for the inbred parents and factorial crosses were recorded. The data were analyzed with a mixed linear model as described in detail by Schrag et al. (2009a, b). The factorial set of crosses investigated here is one of the nine factorials analyzed by Schrag et al. (2009a, b) and was also included in the studies of Schrag et al. (2006, 2007), where it was referred to as Experiment 1.

### Molecular marker data

The inbred lines were assayed for AFLP markers with 20 primer combinations as described in detail by Schrag et al. (2006). The AFLP analyses resulted in 1,835 markers.

The genetic distance $D_A$ between inbred lines $i$ and $j$ was calculated from the banding pattern of $n_m$ AFLP bands as

$$D_A(i,j) = \sqrt{\frac{1}{n_m}\sum_{m=1}^{n_m}[b_m(i) - b_m(j)]^2}, \qquad (1)$$

where $b_m(i)$ and $b_m(j)$ are indicator variables taking the value one, if band $m$ was observed in inbred line $i$ or $j$, respectively, and zero otherwise. $D_A(i, j)$ is related to the single matching coefficient SM$(i, j)$

$$D_A(i,j) = \sqrt{1 - \mathrm{SM}(i,j)}. \qquad (2)$$

$D_A$ has the property of being Euclidean and, therefore, is well suited for principal coordinate and cluster analyses.

### Gene expression data

Five plants of each of the 21 inbred lines were grown for 7 days in a climate chamber under regulated growth conditions. The five biological replicates were pooled (Kendziorski et al. 2005) and RNA was isolated from a mixture of the five seedlings. The 46-k array from the maize oligonucleotide array project (http://www.maizearray.org/, University of Arizona, USA) with 43381 gene-oriented 70-mer maize oligo-spots (in total 46,128 features) printed on a glass-slide was used for hybridization analyses (Thiemann et al. 2009). For the mircroarray analysis we employed an intervoven loop design (Kerr and Churchill 2001) resulting in 62 direct comparisons of dent and flint lines by sampling each dent line five times and each flint line eight times.

Differences in the gene expression were tested with a modified $F$ test using a false discovery rate of 0.01 for all genes showing a fold change of at least 1.3 and expression level (log2) of at least 8.0. All genes, which were differentially expressed in at least one pair of parental lines of the 98 factorial crosses, were assigned to the subset of genes $S_p$.

## Transcriptome-based distance measures

A Euclidean distance $D_E$ between lines $i$ and $j$ can be determined from the gene expression data of $n_g$ genes as

$$D_E(i,j) = \sqrt{\sum_{g=1}^{n_g} \left[l_g(i) - l_g(j)\right]^2}, \tag{3}$$

where $l_g(i)$ and $l_g(j)$ are the base-two logarithms of the transcript abundance of gene $g$ in inbred lines $i$ and $j$.

A binary distance $D_B$ between lines $i$ and $j$ can be determined from the gene expression data as

$$D_B(i,j) = \sqrt{\frac{1}{n_g}\sum_{g=1}^{n_g} \left[x_g(i) - x_g(j)\right]^2} \tag{4}$$

where $x_g(i)$ and $x_g(j)$ are indicator variables taking the values zero and one depending on differential gene expression of gene $g$ in inbred lines $i$ and $j$. If gene $g$ is differentially expressed in lines $i$ and $j$, then

$$x_g(i) = 1 \quad \text{and} \quad x_g(j) = 0 \quad \text{for } l_g(i) > l_g(j), \quad \text{and}$$
$$x_g(i) = 0 \quad \text{and} \quad x_g(j) = 1 \quad \text{for } l_g(i) \leq l_g(j).$$

If gene $g$ is not differentially expressed, then $x_g(i) = x_g(j) = 0$. Equation 4 simplifies to

$$D_B(i,j) = \sqrt{n_s(i,j)/n_g}, \tag{5}$$

where $n_s(i, j)$ is the number of genes differentially expressed in lines $i$ and $j$.

The transcriptome-based distances $D_E$ and $D_B$ were determined for the subset of genes $S_p$. The correlation of $D_E$ and $D_B$ with $D_A$ and the correlation of $D_A$, $D_E$, and $D_B$ with hybrid performance and heterosis was determined. Furthermore, the distances $D_A$, $D_E$, and $D_B$ were subjected to a principal coordinate analysis and a hierarchical cluster analysis using the complete linkage clustering algorithm implemented in the 'hclust' function of statistical software R (Ihaka and Gentleman 1996).

## Association of differential gene expression with hybrid performance and heterosis

To identify genes, which were differentially expressed in the parents of hybrids with high performance, the hybrids are divided in two classes $T$ and $L$ of equal size. The class $T$ consists of hybrids with high and the class $L$ of hybrids with low hybrid performance $Y$:

$$\forall\, Y(i,j) \in T, Y(k,l) \in L : Y(i,j) \geq Y(k,l). \tag{6}$$

Consider gene $g$ and let $o_{gt}$ and $o_{gl}$ denote the numbers of hybrids in class $T$ and $L$, respectively, for which the parents show differential expression of gene $g$. Let

$$P_g = \sum_{k=o_{gt}}^{o_{gt}+o_{gl}} \mathrm{Bin}_{n,p}(k) \tag{7}$$

where $\mathrm{Bin}_{n,p}$ is the probability function of the binomial distribution with parameters $n = o_{gt} + o_{gl}$ and $p = 1/2$. $P_g$ denotes the probability that the count $o_{gt}$, or a larger one, is observed under the condition that differential gene expression occurs with the same probability in the hybrids showing a high and those showing a low hybrid performance. Hence, small values for $P_g$ indicate an association of differential expression with high hybrid performance. A subset of genes, the differential expression of which is associated with mid parent heterosis can be determined in an analogous manner.

On the basis of all 98 hybrids of the factorial, we determined the subsets of genes associated with hybrid performance $S_y$ and heterosis $S_h$. To accomplish this, we compared the probabilities $P_g$ with threshold values determined with the Bonferroni-Holm procedure for a Type-I error of $\alpha = 0.01$. The distances $D_B$ and $D_E$ were calculated for the subsets of genes $S_y$ and $S_h$ and the correlations of these distances with hybrid performance and heterosis, respectively, were determined.

## Transcriptome based prediction of hybrid performance and heterosis

For prediction of new hybrids, a reference set of related breeding material (estimation set) is required. The estimation set consists of the expression profiles of parental inbred lines $u$, $v$ and the field data of their factorial crosses $Y(u, v)$. From these data, the set of genes $S$ is determined, of which differential expression is associated with hybrid performance. The (Euclidean or binary) distances $D^S(u, v)$ on basis of $S$ between the parental lines are determined. The distances between parents and the performance data of the hybrids in the estimation set are used to estimate the regression parameters $\beta_0$ and $\beta_1$ with the linear regression model

$$Y(u,v) = \beta_0 + \beta_1 D^S(u,v). \tag{8}$$

To predict the performance $Y(i, j)$ of a new hybrid, the gene expression profiles of the parental lines $i$ and $j$ are assessed. The set of genes $S$ is used to determine the (binary or Euclidean) distance $D^S\ (i, j)$ between the parental lines. From the distance $D^S\ (i, j)$ and the regression parameters $\beta_0$ and $\beta_1$ the performance $Y(i, j)$ of the new hybrid is predicted with Eq. 8. The prediction procedure is summarized in Fig. 1. Mid-parent heterosis $H(i, j)$ is predicted in an analogous manner.

## Assessment of prediction efficiency

The prediction efficiency was evaluated with a cross-validation procedure in which we divided our data in an

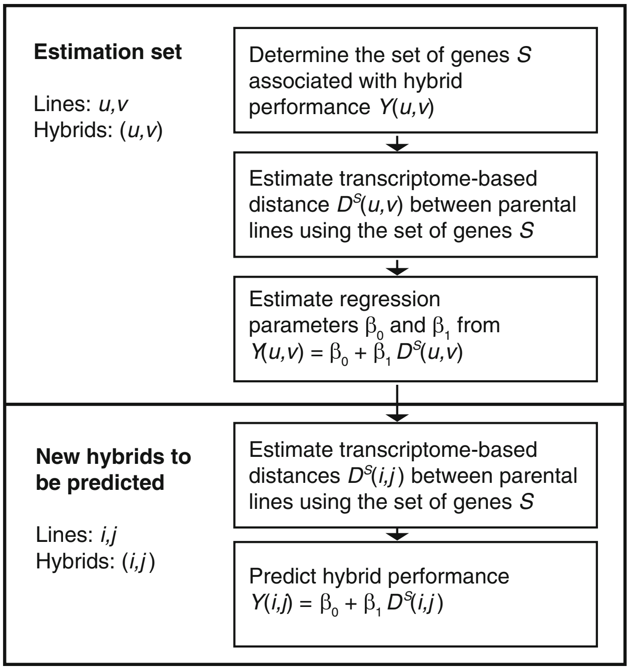

Estimation set
Lines: u,v
Hybrids: (u,v)
Determine the set of genes S associated with hybrid performance Y(u,v)
Estimate transcriptome-based distance D^S(u,v) between parental lines using the set of genes S
Estimate regression parameters β0 and β1 from Y(u,v) = β0 + β1 D^S(u,v)
New hybrids to be predicted
Lines: i,j
Hybrids: (i,j)
Estimate transcriptome-based distances D^S(i,j) between parental lines using the set of genes S
Predict hybrid performance Y(i,j) = β0 + β1 D^S(i,j)

**Fig. 1** Prediction of hybrid performance

estimation set, used for estimation of prediction parameters, and a validation set for which prediction was carried out. The estimation set consisted of the transcriptome data of five randomly chosen dent and three randomly chosen flint lines and the field data of their hybrids. The validation set consisted of the remaining inbreds and hybrids of the factorial. The subsets of genes associated with hybrid performance and heterosis were determined in the estimation set by comparing the binomial probability $P_g$ with a threshold of $\alpha = 0.05$ (employing no adjustment for multiple testing). With these distances, the regression parameters $\beta_0$ and $\beta_1$ were determined in the estimation set (Eq. 8). Employing $S$, $\beta_0$, and $\beta_1$, the hybrid performance and heterosis of the hybrids in the validation set were predicted. The correlation coefficient $r$ of observed with predicted values of $Y(i, j)$ and $H(i, j)$ was assessed in 100 cross validation rounds. It was employed as a measure to assess the prediction efficiency.

In this cross validation scheme, less than half of the lines are used to estimate the prediction parameters and more than half of the lines for validation. Therefore, it models the reality of maize breeding programs much better than the older approach of a "one-leave-out" cross validation (Vuylsteke et al. 2000; Schrag et al. 2006).

The efficiency of the transcriptome-based prediction was compared with that of prediction based on the general combining ability (GCA) estimated from field trials and that of prediction with the 'single-marker total effects of associated markers' (SM-TEAM) approach of Schrag et al. (2007).

## 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. Therefore, the GCA is expected to explain only partially the variation in hybrid performance in this factorial. The results were presented in detail by Schrag et al. (2006).

The subset $S_p$ of genes, which were differentially expressed in at least one pair of parental lines of the factorial crosses consisted of 10,810 genes. The ranges of the distances between lines from the same heterotic pool (intra-pool distances) were $0.22 \le D_A \le 0.57$, $0.16 \le D_B \le 0.53$, and $25.7 \le D_E \le 76.5$. The ranges of the distances between lines from opposite heterotic pools (inter-pool distances) were $0.56 \le D_A \le 0.61$, $0.39 \le D_B \le 0.69$, and $38.6 \le D_E \le 89.0$. The genetic distance $D_A$ was strongly correlated with the transcriptome-based distances $D_B$ and $D_E$ for intra-pool crosses, but only loosely for inter-pool crosses (Fig. 2).

The first principal coordinate clearly separated the flint from the dent lines for all three distance measures (Fig. 3). The only exception was flint line L024, which had a first principal coordinate closer to the dent than to the flint pool for distance $D_E$. The second principal coordinate separated the dent lines with Stiff Stalk from those with Iodent background (Fig. 3).

Cluster analyses based on the genetic distance $D_A$ and the binary distance $D_B$ resulted in separate clusters for the flint and dent lines (Fig. 3). The binary distance $D_B$ even separated sub-clusters of flint lines having European flint and Flint/Lancaster background. The Euclidean distances $D_E$ did not separate the flint and dent lines. A clear separation between the different dent backgrounds (Stiff Stalk vs. Iodent) was not observed for any of the distances.

The genetic distance $D_A$ between parental lines was neither significantly correlated with hybrid performance nor with heterosis for grain yield (Fig. 4). In contrast, the distances $D_B$ and $D_E$, determined from the subset $S_p$ of differentially expressed genes, were correlated with hybrid performance and heterosis.

The subset of genes $S_y$ whose differential expression was associated with hybrid performance consisted of 1,424 genes, and the subset $S_h$ of genes associated with heterosis of 1,763 genes. The distances $D_B$ and $D_E$, determined from the subsets of genes $S_y$ and $S_h$ were strongly correlated with hybrid performance and heterosis for grain yield (Fig. 5).

In cross validation, the correlation of observed with predicted values for hybrid performance and heterosis was

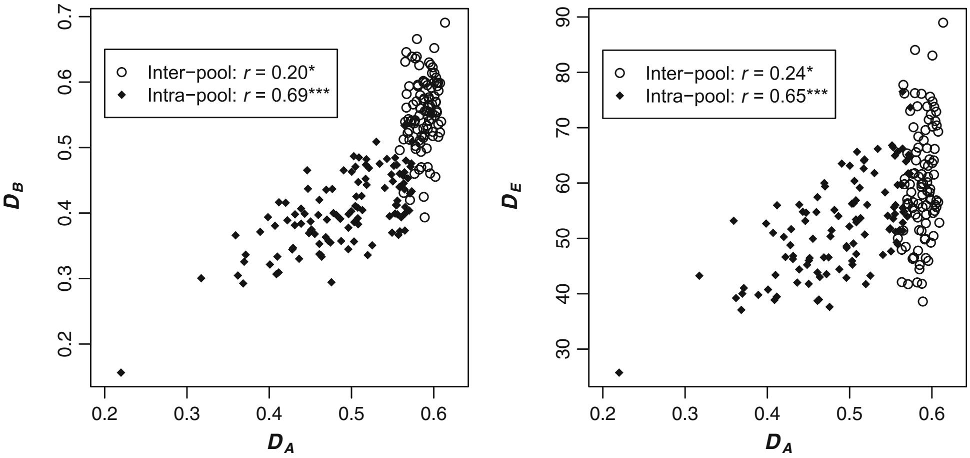


**Fig. 2** Correlation $r$ of the genetic distance $D_A$ with the transcriptome-based distances $D_B$ and $D_E$. The distances $D_B$ and $D_E$ were determined from the subset of genes $S_p$, comprising 10,810 differentially expressed genes. *$P \leq 0.05$, ***$P \leq 0.001$

greater for prediction with the distances $D_B$ and $D_E$ than for prediction with the earlier prediction methods GCA and SM-TEAM (Fig. 6). Prediction with the binary distance $D_B$ resulted in stronger correlations than prediction with the Euclidean distance $D_E$ for both, hybrid performance and heterosis.

For prediction of grain yield with the binary distance $D_B$, the correlations of observed with predicted values had the smallest ranges and inter-quartile distances if 1,000–1,500 genes were selected for prediction (Fig. 6). For fewer than 1,000 genes, the ranges and inter-quartile distances increased and for more than 1,500 genes the median decreased.

## Discussion

### Transcriptome analysis in the seedling stage

The RNA for expression profiling was extracted from entire seedlings 7 days after sowing. From a biological point of view, this raises the question whether there are justifications for the assumption that the transcript levels in seedlings are related to agronomic performance. There may be specific development stages and specific tissues, in which gene expression is functionally more closely related to grain yield, grain dry matter concentration, or other important traits. From a practical point of view, expression profiling in the seedling stage has the big advantage that the data can be generated quickly and with limited resources, compared with growing plants for a longer time period and analyzing specific tissues. It remains open to further investigations, whether a possible gain in information content of the transcriptome data at later development stages may outweigh this time advantage in breeding programs, in particular because establishing a new variety one season earlier in the market can be the key to its economic success.

### Theoretical properties of the transcriptome-based distances

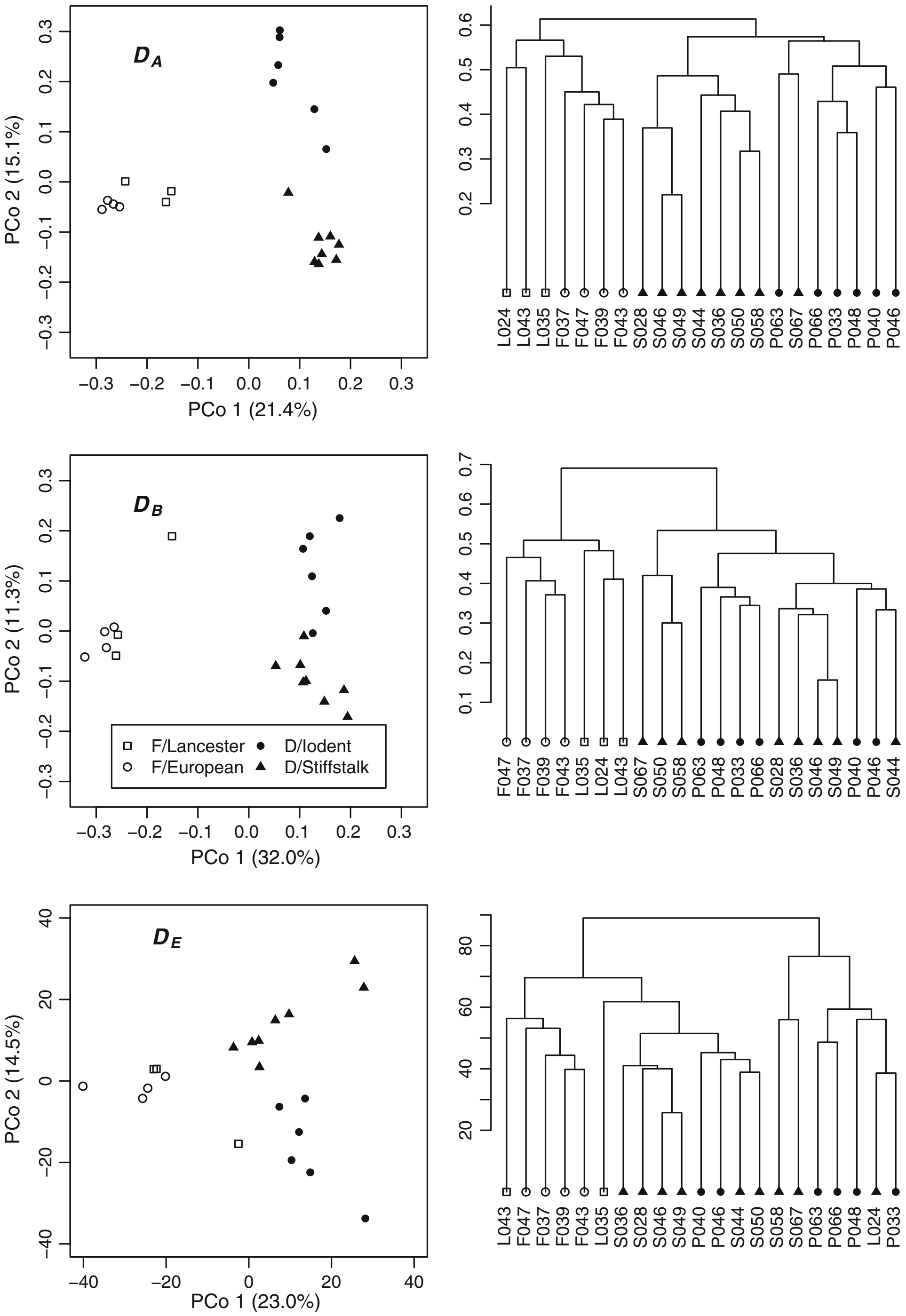
The distance $D_E$ has the desirable property of being Euclidean and, hence, from a mathematical point of view, is suitable for a broad range of multivariate methods. From a genetical point of view, it has the shortcoming that genes with a big difference in the transcript abundance influence more the final value of the distance than genes with a smaller difference. Transformations, such as the base-two logarithmic transformation employed in this study, reduce the numerical effect. However, genetical or physiological models justifying such transformations are lacking.

The binary distance $D_B$ discards the quantitative information on the transcription levels and assigns equal effects to all differentially expressed genes. This corresponds to a quantitative genetic model with many genes having small effects of similar size. Employing $D_B$ for prediction of heterosis with the linear regression model of Eq. 8 corresponds well to the hypothesis, that small dominance effects at a large number of loci are resulting in the heterosis observed in a hybrid.

### Correlation of the genetic distances with the transcriptome-based distances

The transcriptome-based distances showed a significant correlation of about 0.7 with the genetic distance for intra-pool crosses, whereas for inter-pool crosses only a loose correlation of about 0.2 was observed (Fig. 2). The high correlation of the transcriptome-based distances with the genetic distance for intra-pool crosses supports the hypothesis that within one heterotic pool both measures contain to a large part similar information. For inter-pool crosses, the low correlation suggests that the information content of the transcriptome-based and genetic distances differs. The range of inter-pool genetic distance was considerably smaller than the range of inter-pool transcriptome-based distances. This

**Fig. 3** Principle coordinate analyses and hierarchical cluster analyses based on the genetic distance $D_A$ and the transcriptome-based distances $D_B$ and $D_E$. The distances $D_B$ and $D_E$ were determined from the subset of genes $S_p$ comprising 10,810 differentially expressed genes



supports the hypothesis that the inter-pool transcriptome-based distances carry more information than the inter-pool genetic distances. The transcriptome-based distances are directly quantifying the expression of genes, which may be responsible for the phenotype and do not rely on the linkage between markers and genes. In consequence, transcriptome data should be preferable to marker data when different heterotic pools are considered.

## Germplasm grouping

The multivariate analyses employing the binary distance $D_B$ grouped the flint and dent pools as clearly as did the genetic distance $D_A$. In addition, cluster analysis with $D_B$ separated two subgroups within the flint pool (Fig. 3). In principal coordinate analysis on basis of $D_E$, the flint line L024 was close to the dent lines, and in the cluster analyses

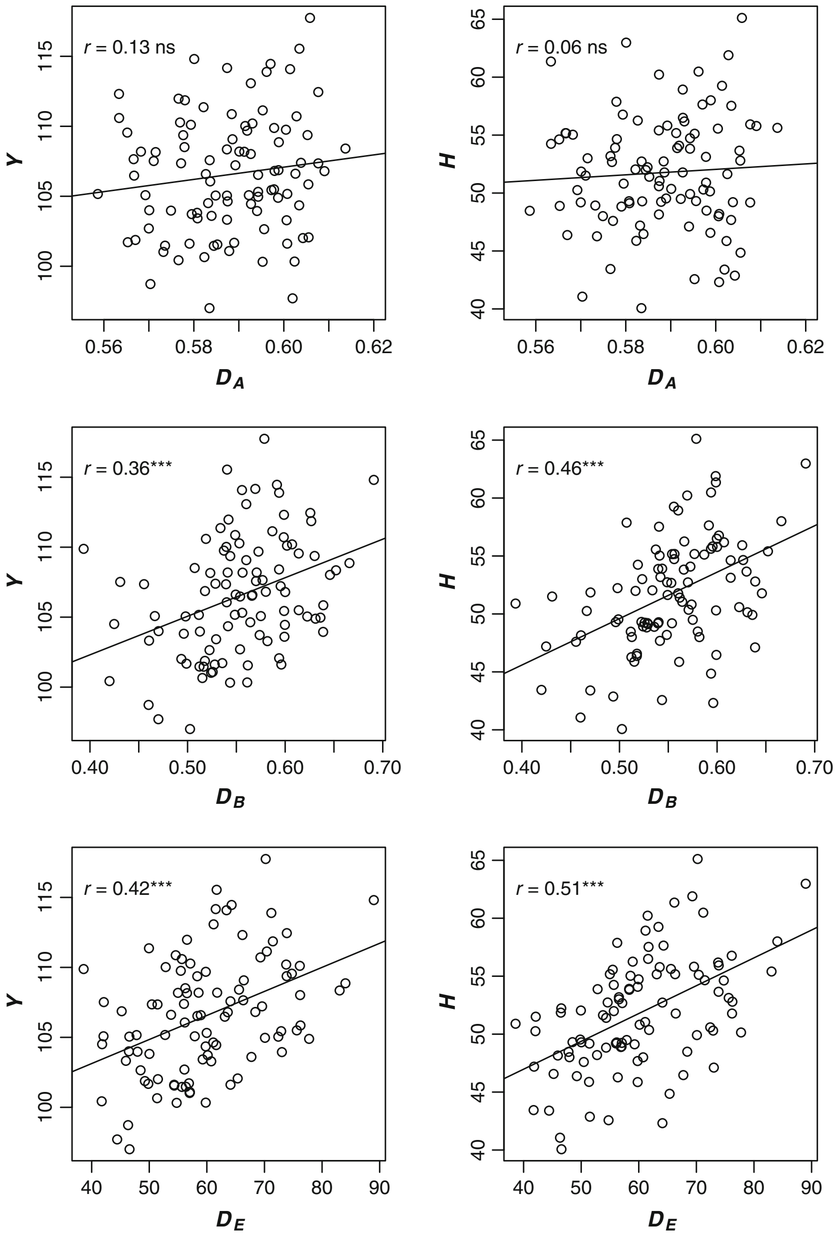


**Fig. 4** Correlation $r$ of hybrid performance $Y$ (*left hand side*) and mid-parent heterosis $H$ (*right-hand side*) for grain yield with the genetic distance $D_A$ (*top*), binary distance $D_B$ (*center*), and Euclidean distance $D_E$ (*bottom*). The distances $D_B$ and $D_E$ were determined from the subset of genes $S_P$ comprising 10,810 differentially expressed genes. ns: $P > 0.05$, $^{***}P \leq 0.001$

no separate clusters for flint and dent lines were formed. In conclusion, multivariate analyses based on $D_B$ were more effective in grouping the germplasm than analyses based on the genetic distance $D_A$, whereas grouping on basis of $D_E$ was slightly less effective than grouping on basis of $D_A$.

## Correlation of the transcriptome-based distances with hybrid performance and heterosis

In many studies in maize, it has been observed that the genetic distance between the parents of inter-pool hybrids was not correlated with the hybrid performance or heterosis (cf. Melchinger 1999). Our data on the AFLP-based genetic distance $D_A$ confirms this result (Fig. 4). The significant correlations of the transcriptome-based distances with hybrid performance and heterosis may be explained by the high density of investigated loci, the analysis of the genes rather than markers, and the inclusion of additive–additive interactions in the analysis.

Transcription profiling resulted in 10,810 differentially expressed genes in the factorial. Such a high number of loci investigated means a good coverage of the genes underlying grain yield and, therefore, resulted not only in significant but very strong correlations.

Founder effects, selection, and random genetic drift can result in differences in the linkage disequilibrium between marker alleles and functional alleles in different heterotic pools (Boppenmeier et al. 1992; Charcosset and Essioux

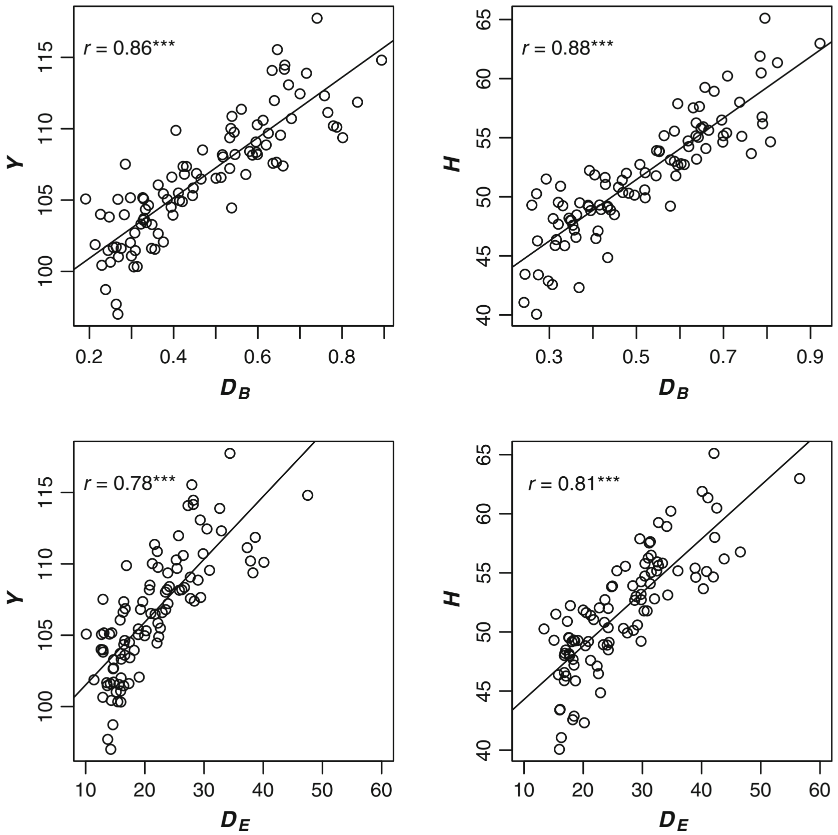


**Fig. 5** Correlation *r* of hybrid performance *Y* (*left-hand side*) and mid-parent heterosis *H* (*right-hand side*) for grain yield with the binary distance $D_B$ (*top*) and Euclidean distance $D_E$ (*bottom*). The distances were determined from the subset of genes $S_y$ comprising 1,424 genes of which differential expression is associated with hybrid performance (*left-hand side*), and the subset $S_h$ comprising 1,763 genes of which differential expression is associated with heterosis (*right-hand side*). ***$P \leq 0.001$

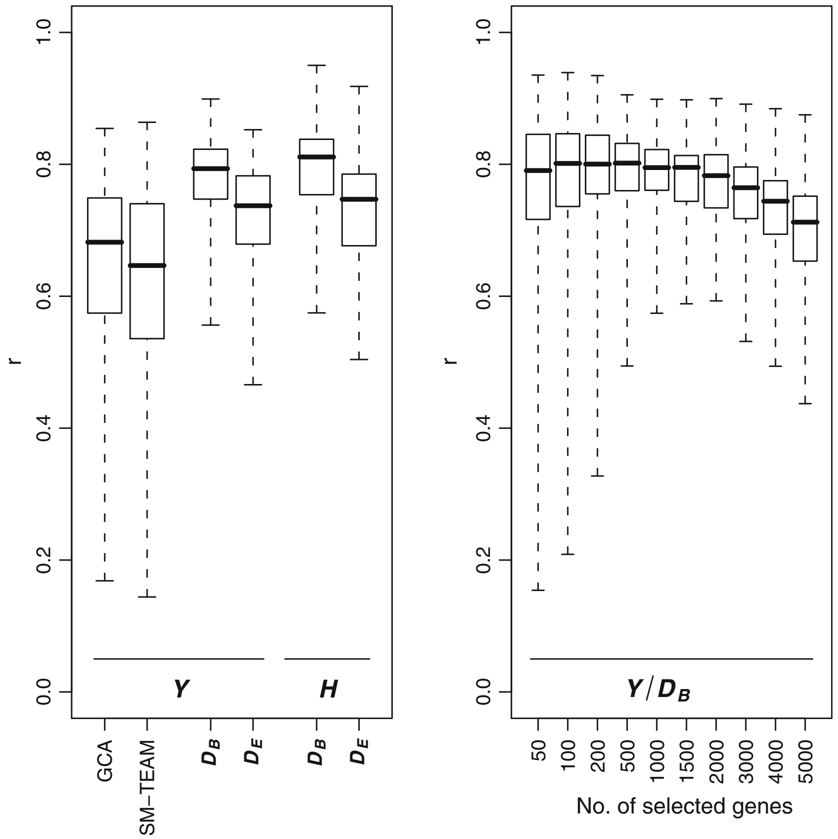


**Fig. 6** Correlations *r* of the observed with predicted hybrid performance *Y* and mid-parent heterosis *H* for grain yield obtained by cross validation with 100 rounds. The *bold lines* in the *boxes* denote the median, the borders of the boxes the quartiles, and the ends of the whiskers the extreme values of *r*. *Left-hand side*: The genes for prediction were selected in the estimation set with an exact binomial test ($\alpha = 0.05$), resulting in approx. 800–1,000 genes for each trait in the individual cross validation rounds. $D_B$ denotes prediction with the binary and $D_E$ with the Euclidean transcriptome-based distances. Results for GCA and SM-TEAM were taken from Schrag et al. (2007). *Right hand side*: Fixed numbers of 50 to 5,000 genes were selected in each cross validation round on basis of small probabilities $P_g$

1994). While in one heterotic pool, a certain marker allele can be in coupling phase linkage with a certain allele at a functional locus, in the opposite pool the marker allele may be in repulsion phase linkage with this functional allele. Therefore, inter-pool genetic distances at marker loci may provide only a poor estimate for the differences at functional genes between two lines belonging to different heterotic pools. Expression profiling investigates directly the genes, and does not rely on linkage disequilibrium between marker alleles and functional alleles. Therefore, it is not affected by different linkage phases in different heterotic pools and quantifies directly the differences at functional genes between two lines. This seems to be the main reason why the transcriptome-based distances are strongly correlated with hybrid performance and heterosis, whereas the inter-pool genetic distances are not correlated.

Furthermore, additive–additive interactions responsible for increased RNA transcription are accounted for in the transcription profiling and, hence, contribute to the transcriptome-based distances. These may increase the proportion of phenotypic variance explained by the distances and, thus, can also contribute to the high correlation of transcriptome-based distances with hybrid performance and heterosis.

The correlation of transcriptome-based distances with hybrid performance and heterosis for selected genes (subsets $S_y$ and $S_h$, Fig. 5) was considerably larger than for unselected genes (subset $S_p$, Fig. 4). Hence, using a set of selected genes for prediction models has the potential to increase prediction efficiency.

### Transcriptome-based prediction of hybrid performance and heterosis

Selection of the set of genes $S$, employed for prediction, was based on the binomial probability (Eq. 7). It was determined separately for each cross validation run, employing a threshold of $\alpha = 0.05$ and no adjustment for multiple testing. This was an arbitrary choice, and the number of genes selected with such a procedure depends strongly on the size of the estimation set and chosen thresholds. For the present data set, however, it resulted in approximately 800–1,000 genes for each run, which was near the optimum value of 1,000–1,500 genes (Fig. 6). An alternative strategy would be to determine once a ‘core’ set of genes responsible for a given trait and use these genes subsequently for prediction. It is an interesting area for further research, whether using a fixed set of genes improves the prediction efficiency or not.

Predicted values of heterosis are of use in practical breeding programs only in combination with the mid-parent value of line per se performance, whereas the predicted hybrid performance can be applied directly for making selection decisions. Therefore, the prediction of heterosis is of use, if it can be accomplished with higher precision than that of hybrid performance. In our study, the predictions of hybrid performance and heterosis were equally precise (Fig. 6). Consequently, we recommend prediction of the hybrid performance rather than that of heterosis.

Predictions with the binary distance $D_B$ showed greater correlations to the observed values than predictions with the Euclidean distance $D_E$. Further, $D_B$ outperformed by far the GCA-based prediction and also the prediction with a linear model using selected AFLP markers (Fig. 6). In consequence, prediction models employing the transcriptome-based distance $D_B$ should provide the most precise predictions of hybrid performance available to date.

### Applications in breeding programs

Expression profiling of seedlings can be conducted directly after producing new inbred lines. From the transcriptome data, the performance of possible hybrids can be predicted and the promising hybrids can be produced and tested in field trials. This indirect pre-selection step based on expression profilecs can enhance the response to selection. At present transcriptome analysis is expensive, but a decrease in lab costs is expected and with such a cost decrease, the suggested pre-selection can increase the cost efficiency of breeding programs.

Assuming a heritability near one of the transcriptome data ($h_I^2 = 1$), the response $G_I$ to indirect selection based on transcriptome-based distances is $G_I = i\, h_I\, r\, \sigma_g$, whereas that for direct selection based on field trials is $G_D = i\, h_D\, \sigma_g$ (assuming constant selection intensity $i$ and genetic variance $\sigma_g^2$). Consequently, with correlations of $r \approx 0.8$ (Fig. 6) indirect selection has the same efficiency as that of direct selection with a heritability of $h_D^2 = (0.8)^2 = 0.64$. This demonstrates the potential utility of the transcriptome-based prediction of hybrid performance in hybrid breeding programs.

**Acknowledgments** This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the priority program SPP 1149 “Heterosis in Plants” (grant no. FR 1615/4-1). We thank Prof. Dr. B.S. Dhillon for helpful suggestions on our manuscript. This article is dedicated to Professor Dr. H. Friedrich Utz on the occasion of his 70th anniversary.

## References

Birchler JA, Auger DL, Riddle NC (2003) In search of the molecular basis of Heterosis. Plant Cell 15:2236–2239

Boppenmeier J, Melchinger AE, Brunklaus-Jung E, Geiger HH, Herrmann RG (1992) Genetic diversity for RFLPs in European maize inbreds. I: relation to performance of Flint x Dent crosses for forage traits. Crop Sci 32:895–902

Charcosset A, Essioux L (1994) The effect of population-structure on the relationship between heterosis and heterozygosity at marker loci. Theor Appl Genet 89:336–343

Guo M, Rupe M, Yang X, Crasta O, Zinselmeier C, Smith O, Bowen B (2006) Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. Theor Appl Genet 113:831–845

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

Kendziorski C, Irizarry RA, Chen K-S, Haag JD, Gould MN (2005) On the utility of pooling biological samples in microarray experiments. Proc Nat Acad Sci 102:4252–4257

Kerr MK, Churchill GA (2001) Statistical design and the analysis of gene expression microarray data. Genet Res 77:123–128

Maenhout S, De Baets B, Haesaert G (2009) Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theor Appl Genet (in press)

Melchinger AE (1999) Genetic diversity and heterosis. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. ASA-CSSA, Madison, pp 99–118

Reif JC, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA, Vasal SK, Srinivasan G, Bohn M, Frisch M (2003) Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci 43:1275–1282

Repsilber D, Andorf S, Selbig J, Altmann T, Witucka-Wall H (2009) 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 (in press)

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, Maurer HP, Melchinger AE, Piepho HP, Peleman J, Frisch M (2007) Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. Theor Appl Genet 114:1345–1355

Schrag TA, Möhring J, Maurer HP, Dhillon BS, Melchinger AE, Piepho HP, Sorensen AP, Frisch M (2009a) Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theor Appl Genet 118:741–751

Schrag TA, Möhring J, Kusterer B, Dhillon BS, Melchinger AE, Piepho HP, Frisch M (2009b) Hybrid performance prediction in maize using molecular markers and joint analyses of hybrids and parental inbreds. Theor Appl Genet (in press)

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 RC, Altmann T, Willmitzer L, Selbig J (2009) Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theor Appl Genet (in press)

Stupar RM, Gardiner JM, Oldre AG, Haun WJ, Chandler VL, Springer NM (2008) Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. BMC Plant Biol 8:33

Thiemann A, Fu J, Schrag TA, Melchinger AE, Frisch M, Scholten S (2009) Correlation between parental transcriptome and field data for the characterization of heterosis in *Zea mays* L. Theor Appl Genet (in press)

Vuylsteke M, Kuiper M, Stam P (2000) Chromosomal regions involved in hybrid performance and heterosis: their AFLP-based identification and practical use in prediction models. Heredity 85:208–218