Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet Angewandte Genetik und Pflanzenzüchtung
Prof. Dr. A.E. Melchinger

# Prediction of hybrid performance in maize using molecular markers

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften

vorgelegt
der Fakultät Agrarwissenschaften

von
Diplom-Agraringenieur
Tobias Schrag
aus Mutlangen

2008

Die vorliegende Arbeit wurde am 18. Juni 2008 von der Fakultät Agrarwissenschaften der Universität Hohenheim als „Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)“ angenommen.

| | |
|---|---|
| Tag der mündlichen Prüfung: | 28. August 2008 |
| 1. Prodekan: | Prof. Dr. W. Bessei |
| Berichterstatter, 1. Prüfer: | Prof. Dr. A.E. Melchinger |
| Mitberichterstatter, 2. Prüfer: | Prof. Dr. H.-P. Piepho |
| 3. Prüfer: | Prof. Dr. R. Blaich |

# Contents



---

[1] Schrag, T.A., A.E. Melchinger, A.P. Sørensen, and M. Frisch. 2006. Theor. Appl. Genet. 113:1037-1047.

[2] Schrag[*], T.A., H.P. Maurer[*], A.E. Melchinger, H.-P. Piepho, J. Peleman, and M. Frisch. 2007. Theor. Appl. Genet. 114:1345-1355.

[3] Schrag T.A., J. Möhring, H.P. Maurer, B.S. Dhillon, A.E. Melchinger, H.-P. Piepho, A.P. Sørensen, and M. Frisch. 2008. Theor. Appl. Genet. *In review*.

[*]Both authors contributed equally.

# Abbreviations

| | |
|---|---|
| AFLP | amplified fragment length polymorphism |
| ANOVA | analysis of variance |
| BLUE | best linear unbiased estimation |
| BLUP | best linear unbiased prediction |
| DH | doubled haploid |
| GCA | general combining ability |
| $GCSM_{HP}$ | general contribution of selected markers to HP |
| GD | genetic distance |
| GDMC | grain dry matter content |
| GY | grain yield |
| HB1/2/3 | haplotype block methods 1/2/3 |
| HP | hybrid performance |
| LD | linkage disequilibrium |
| MH | mid-parent heterosis |
| MLR | multiple linear regression |
| MLR-H | MLR on HP |
| MLR-LM | line *per se* performance plus MLR on MH |
| PP | phenotypic and pedigree data-based methods for prediction of HP |
| PP-GS | PP method with GCA plus SCA |
| PP-L | PP method with line *per se* performance |
| QTL | quantitative trait locus |
| $R^2$ | proportion of explained variance |
| $\Delta R^2$ | difference in $R^2$ between Type 0 and Type 1 hybrids |
| RMSD | square root of the mean square deviation |
| SCA | specific combining ability |
| $SCSM_{HP}$ | specific contribution of selected markers to HP |
| SM | single marker |
| SNP | single nucleotide polymorphism |
| SSR | simple sequence repeat |
| T-BLUP | BLUP based on trait data |

| | |
|---|---|
| TC | testcross |
| TCSM | total contribution of selected markers |
| $TCSM_{HP}$ | TCSM to HP |
| $TCSM_{SCA}$ | TCSM to SCA |
| TEAM | total effects of associated markers |
| TEAM-H | TEAM on HP |
| TEAM-LM | line *per se* performance plus TEAM on MH |
| TM-BLUP | BLUP based on trait and marker data |
| Type 0/1/2 | hybrids between lines of which none/one/both were TC evaluated |

# Chapter 1
# General Introduction

Hybrid maize (*Zea mays* L.) breeders develop a large number of inbred lines and evaluate their performance in cross combinations (Hallauer 1990). In commercial maize breeding programs, identification of single-cross hybrids with superior yield performance is of fundamental importance. However, the number of potential crosses increases rapidly with the number of inbreds. Owing to limited resources, only a small proportion of these crosses is evaluated in field trials. Promising single-cross hybrids can be identified without having them tested in field trials by predicting their hybrid performance (HP) on basis of field trial data available from related crosses.

## Prediction methods for hybrid performance

Methods for performance prediction of single crosses have always been a major issue as successful prediction has the potential to greatly improve the efficiency of commercial breeding programs. Maize germplasm is commonly organised in genetically divergent heterotic groups, therefore, predicting the performance of inter-group hybrids is of greatest interest to maize breeders.

**Line *per se* performance and heterosis.** Predicting the performance of hybrids from the *per se* performance of their parental inbred lines has not been effective due to masking dominance effects (Smith 1986; Hallauer 1990). Thus, line *per se* performance alone does not sufficiently explain the variance of grain yield (GY) of maize hybrids. The difference in performance between a hybrid and the mean of its parents is defined as mid-parent heterosis. Since up to 76% of the GY of maize hybrids (Hallauer and Miranda Filho 1988) is accounted for by mid-parent heterosis, it has to be also considered in prediction of HP.

**General combining ability.** Estimates of general combining ability (GCA) of the parental lines provide an established and simple approach to predict HP (Cockerham 1967; Melchinger et al. 1987). Prediction based on GCA alone ignores specific combining ability (SCA), which is related to specific heterosis and constitutes an important component of HP (Gardner and Eberhart 1966).

**Phenotypic T-BLUP approach.** Best linear unbiased prediction (BLUP) was proposed by Bernardo (1994, 1996) to predict performance of untested single crosses using phenotypic information of related single crosses. In addition to trait data (T-BLUP), this approach uses information about genetic relationships among their parental inbreds, based on coancestry coefficients estimated from pedigree records or molecular marker data. The results of this approach were promising, however, the full potential of molecular markers is not utilised with relationship coefficients. These indicate overall expectations for the whole genome, but ignore specific genomic regions, which may be relevant for the predicted trait.

**Marker-enhanced TM-BLUP approach.** The T-BLUP approach was extended by Bernardo (1998, 1999) to account for trait and marker data (TM-BLUP) in the prediction of HP. In the extended approach, identity by descent of unobservable quantitative trait locus (QTL) alleles was inferred from molecular marker data and used for modelling the covariances associated with QTL. However, TM-BLUP resulted only in marginal improvement for predicting single-cross performance, compared with the ordinary T-BLUP approach.

**Molecular genetic distances.** Estimates of genetic distances (GD) between the parental lines based on unselected DNA markers alone were not promising for predicting performance of inter-group hybrids (Melchinger 1999). These findings were in agreement with theoretical results of Charcosset and Essioux (1994), who attributed the low correlation between heterosis and GD to (1) no or only loose linkage of heterosis-affecting QTL with the molecular markers employed to estimate GD and (2) different linkage phases between the QTL and

marker alleles in the maternal and paternal gametic arrays, as expected frequently with inter-group hybrids.

**Marker-based prediction of SCA.** Charcosset et al. (1998) evaluated the prediction of HP, comparing different marker-based approaches to account for SCA. Their results for inter-group crosses indicated higher prediction efficiencies with BLUP and factorial regression models compared with a GD model.

**Marker-based prediction of HP.** Associations of amplified fragment length polymorphism (AFLP) markers with HP for GY and SCA across inter-group hybrids were investigated by Vuylsteke et al. (2000). The sum of marker effects across significantly associated markers provided an estimate for the genotypic value of the hybrids. In a linear regression approach, these estimates of genotypic value provided the basis for prediction of HP and SCA. The predictions obtained with this "total sum of selected markers" (TCSM) approach were encouraging, but comparisons with established procedures such as GCA-based methods for prediction of inter-group hybrids are lacking. In addition, the approach does not adjust for multiple testing in the genome scan. Further, it inefficiently uses marker data information, owing to its inability to handle missing data.

## Linkage disequilibrium between markers

Correlation between marker loci can be the result of (1) close linkage between marker loci, particularly with high marker densities, (2) closely related individuals, as occur in breeding programs, and (3) sampling a limited number of genotypes (Flint-Garcia et al. 2003; Stich et al. 2007). As a consequence, the effect of a QTL linked to a series of correlated markers can be inflated and, thereby, the prediction error is increased. In addition, ignoring the correlation of

markers results in an overly stringent adjustment for multiple testing (e.g., with the Bonferroni method) and thereby reduces the power of detecting QTL.

These problems can be addressed by combining highly correlated adjacent markers into haplotype blocks. Simple approaches with fixed block length (Jansen et al. 2003) ignore the correlation structure of the actual marker data. In contrast, data-driven strategies determine haplotype block boundaries by considering linkage disequilibrium (LD) between and within blocks (Gabriel et al. 2002), haplotype diversity within blocks (Patil et al. 2001; Zhang et al. 2002), or both LD decay between blocks and diversity of haplotypes within blocks (Anderson and Novembre 2003). These data-driven approaches were developed to identify haplotype-tagging single nucleotide polymorphisms (SNP) used for association mapping of human disease genes. However, the goal of using haplotype blocks for marker-based performance prediction is to reduce the number of estimated parameters while utilising the total haplotype diversity described by all markers. Such criteria to find haplotype block boundaries have not been investigated hitherto. Haplotype block data are similar to multi-allelic marker systems such as simple sequence repeat (SSR) markers. However, the TCSM prediction method (Vuylsteke et al. 2000) was developed for biallelic AFLP markers and therefore not suitable for multi-allelic marker data. Combining adjacent markers into haplotype blocks only accounts for correlation between tightly linked markers, but not for genome-wide correlation of unlinked markers. Sequential methods for multiple linear regression (MLR) can be used to address multicollinearity among variables, as was discussed by Piepho and Gauch (2001) for mapping of QTL. However, no research has been reported investigating MLR to address genome-wide multicollinearity among markers for prediction of HP.

## Unbalanced data from commercial hybrid breeding programs

The marker-based HP prediction approach devised by Vuylsteke et al. (2000) was only applied to separate experiments with factorial crosses. With the BLUP method (Bernardo 1996), voluminous data from commercial programs, though unbalanced, can be analysed. However, a combination of BLUP with the marker-based genotypic value approach remains to be developed and evaluated. In addition, a combined analysis of hybrids and their parental inbred lines across several trials is possible with BLUP, enabling the efficient determination of heterosis as basis for marker-based heterosis prediction. Evaluating the efficiency of prediction with leave-one-out cross-validation (Bernardo 1996; Vuylsteke et al. 2000) addresses only cases of a few missing hybrids in a factorial, whereas cross-validation with larger proportions of hybrids removed from the complete data set (Bernardo 1994; Charcosset et al. 1998) resembles more closely to the situation of unbalanced data from commercial breeding programs. Likewise, the predicted hybrids considered so far were only crosses between two testcross evaluated lines. However, prediction of hybrids where only one or even none of the parental inbreds were testcross evaluated, was not considered, yet this situation is equally relevant in practice. Prediction efficiency for such hybrids remains to be investigated.

## Objectives

The goal of this thesis research was to develop and evaluate methods for marker-based prediction of HP in unbalanced data from commercial maize hybrid breeding programs. In particular, the objectives were to

(1) identify marker loci associated with QTL for hybrid performance from data of factorial mating experiments,

(2) compare HP prediction by marker-based genotypic value estimates with those based on GCA,

(3) develop models for HP prediction that account for multiple testing and correlated markers (by using haplotype blocks and/or MLR),

(4) develop and examine HP prediction models that complement line *per se* performance with marker-based predicted heterosis, and

(5) evaluate the prediction methods under scenarios that are relevant to practical maize breeding.

## References

Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336-354

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

Bernardo R (1996) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:50-56

Bernardo R (1998) Predicting the performance of untested single crosses: trait and marker data. In: Lamkey KR and Staub JE (eds) Concepts and breeding of heterosis in crop plants. Crop Science Society of America, Madison, WI, USA, pp 117-127

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

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

Charcosset A, Bonnisseau B, Touchebeuf O, Burstin J, Dubreuil P, Barriere Y, Gallais A, Denis JB (1998) Prediction of maize hybrid silage performance using marker data: comparison of several models for specific combining ability. Crop Sci 38:38-44

Cockerham CC (1967) Prediction of double crosses from single crosses. Der Züchter 37:160-169

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357-374

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229

Gardner CO, Eberhart SA (1966) Analysis and interpretation of the variety cross diallel and related populations. Biometrics 22:439-452

Hallauer AR, Miranda Filho JB (1988) Quantitative genetics in maize breeding. Iowa State University Press, Ames, IA, USA

Hallauer AR (1990) Methods used in developing maize inbreds. Maydica 35:1-16

Melchinger AE, Geiger HH, Seitz G, Schmidt GA (1987) Optimum prediction of three-way crosses from single crosses in forage maize (*Zea mays* L.). Theor Appl Genet 74:339-345

Melchinger AE (1999) Genetic diversity and heterosis. In: Coors JG and Pandey S (eds) The Genetics and Exploitation of Heterosis in Crops. ASA - CSSA, Madison, WI, pp 99-118

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen NP, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719-1723

Piepho HP, Gauch HG (2001) Marker pair selection for mapping quantitative trait loci. Genetics 157:433-444

Smith OS (1986) Covariance between line *per se* and testcross performance. Crop Sci 26:540-543

Stich B, Melchinger AE, Piepho HP, Hamrit S, Schipprack W, Maurer HP, Reif JC (2007) Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. Theor Appl Genet 115:529-536

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

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci U S A 99:7335-7339

# Chapter 2
# Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL

**T.A. Schrag[1], A.E. Melchinger[1], A.P. Sørensen[2], M. Frisch[1]**

[1] *Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, D-70593 Stuttgart, Germany*
[2] *Keygene, P.O. Box 216, 6700 AE Wageningen, The Netherlands*

Theor. Appl. Genet. 113:1037-1047 (2006).
The original publication is available at www.springerlink.com.

**Abstract.** Prediction methods to identify single-cross hybrids with superior yield performance have the potential to greatly improve the efficiency of commercial maize (*Zea mays* L.) hybrid breeding programs. Our objectives were to (1) identify marker loci associated with quantitative trait loci for hybrid performance or specific combining ability (SCA) in maize, (2) compare hybrid performance prediction by genotypic value estimates with that based on general combining ability (GCA) estimates, and (3) investigate a newly proposed combination of the GCA model with SCA predictions from genotypic value estimates. A total of 270 hybrids was evaluated for grain yield and grain dry matter content in four Dent × Flint factorial mating experiments, their parental inbred lines were genotyped with 20 AFLP primer-enzyme combinations. Markers associated significantly with hybrid performance and SCA were identified, genotypic values and SCA effects were estimated, and four hybrid performance prediction approaches were evaluated. For grain yield, between 38 and 98 significant markers were identified for hybrid performance and between zero and five for SCA. Estimates of prediction efficiency ($R^2$) ranged from 0.46 to 0.86 for grain yield and from 0.59 to 0.96 for grain dry matter content. Models enhancing the

GCA approach with SCA estimates resulted in the highest prediction efficiency if the SCA to GCA ratio was high. We conclude that it is advantageous for prediction of single-cross hybrids to enhance a GCAbased model with SCA effects estimated from molecular marker data, if SCA variances are of similar or larger importance as GCA variances.

# Chapter 3
# Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield

**T.A. Schrag[1], H.P. Maurer[1], A.E. Melchinger[1], H.-P. Piepho[2], J. Peleman[3], M. Frisch[1]**

[1] *Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, D-70593 Stuttgart, Germany*
[2] *Bioinformatics Unit of the Institute for Crop Production and Grassland Research, University of Hohenheim, D-70599 Stuttgart, Germany*
[3] *Keygene, P.O. Box 216, 6700 AE Wageningen, The Netherlands*

Theor. Appl. Genet. 114:1345-1355 (2007).
The original publication is available at www.springerlink.com.
T.A. Schrag and H.P. Maurer contributed equally to this work.

**Abstract.** Marker-based prediction of hybrid performance facilitates the identification of untested single-cross hybrids with superior yield performance. Our objectives were to (1) determine the haplotype block structure of experimental germplasm from a hybrid maize breeding program, (2) develop models for hybrid performance prediction based on haplotype blocks, and (3) compare hybrid performance prediction based on haplotype blocks with other approaches, based on single AFLP markers or general combining ability (GCA), under a validation scenario relevant for practical breeding. In total, 270 hybrids were evaluated for grain yield in four Dent × Flint factorial mating experiments. Their parental inbred lines were genotyped with 20 AFLP primer-enzyme combinations. Adjacent marker loci were combined into haplotype blocks. Hybrid performance was predicted on basis of single marker loci and haplotype blocks. Prediction based on variable haplotype block length resulted in an

improved prediction of hybrid performance compared with the use of single AFLP markers. Estimates of prediction efficiency ($R^2$) ranged from 0.305 to 0.889 for marker-based prediction and from 0.465 to 0.898 for GCA-based prediction. For inter-group hybrids with predominance of general over specific combining ability, the hybrid prediction from GCA effects was efficient in identifying promising hybrids. Considering the advantage of haplotype block approaches over single marker approaches for the prediction of inter-group hybrids, we see a high potential to substantially improve the efficiency of hybrid breeding programs.

# Chapter 4
# Haplotype- and marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses

**T.A. Schrag[1], J. Möhring[2], H.P. Maurer[1], B.S. Dhillon[1], A.E. Melchinger[1], H.-P. Piepho[2], A.P. Sørensen[3], M. Frisch[4]**

[1] *Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, D-70593 Stuttgart, Germany*
[2] *Bioinformatics Unit of the Institute for Crop Production and Grassland Research, University of Hohenheim, D-70599 Stuttgart, Germany*
[3] *Keygene, P.O. Box 216, 6700 AE Wageningen, The Netherlands*
[4] *Institute of Agronomy and Plant Breeding II, Justus-Liebig-University Giessen, D-35392 Giessen, Germany*

Theor. Appl. Genet. In review (2008).
The original publication is available at www.springerlink.com.

**Abstract.** In hybrid breeding, the prediction of hybrid performance (HP) is extremely important as it is difficult to evaluate inbred lines in numerous cross combinations. Recent developments like doubled haploid production and molecular marker technologies have enhanced the significance of HP prediction. The objectives of our study were to (1) develop methods based on phenotypic, pedigree, and marker data for predicting HP, and (2) compare their efficiency. An unbalanced data set of 400 hybrids from nine factorial crosses tested in different experiments and data of 79 inbred parents were subjected to combined analyses with a mixed linear model. Marker data of the inbreds were obtained with 20 AFLP primer-enzyme combinations. Cross-validation was used to assess the performance prediction of hybrids of which none or only one parental line was testcross evaluated. For HP prediction, the highest proportion of explained variance ($R^2$), 46% for grain yield (GY) and 70% for grain dry matter content

(GDMC), was obtained from line *per se* best linear unbiased prediction (BLUP) estimates plus marker effects associated with mid-parent heterosis (TEAM-LM). For GY, prediction from marker effects associated with HP (TEAM-H) had equally high $R^2$ (45%) as TEAM-LM. For GDMC, which unlike GY displayed only little heterosis, prediction from line *per se* BLUPs using phenotypic and pedigree data (PP-L) had nearly as high $R^2$ (65%) as TEAM-LM. Our study demonstrated that HP was efficiently predicted using molecular markers even for GY when testcross data of both parents are not available. This can help in improving greatly the efficiency of commercial hybrid breeding programs.

# Chapter 5
# General Discussion

Marker-based prediction of hybrid performance (HP) for selection of promising cross combinations can accelerate hybrid breeding programs and improve their efficiency. Recent developments in doubled haploid (DH) and molecular marker technologies have enhanced the significance of marker-based HP prediction.

**Doubled haploids.** Development of inbred lines in maize has been facilitated and accelerated in recent years by the DH technology, which is being increasingly used in commercial hybrid breeding programs (Schmidt 2004; Röber et al. 2005; Seitz 2005). Large numbers of inbred lines can be generated in a short period of time without recurrent selfing over generations. Since DH lines are completely homozygous, they can be used directly for the development of experimental hybrids, i.e., potential hybrid varieties (Gallais and Bordes 2007). However, DH lines represent a random sample of unselected inbred lines from the parental cross. Hence, they must be evaluated for line *per se* and testcross performance (Longin et al. 2007a) and this may favour early testing of S1 lines prior to DH production (Longin et al. 2007b). An attractive alternative is to employ molecular markers to predict HP and thereby accelerate the identification of superior hybrid varieties in an economical manner.

**Molecular markers.** During the last decade, high-throughput genotyping platforms based on single nucleotide polymorphism (SNP) markers were established by large plant breeding programs (Eathington et al. 2007). Routine fingerprinting of lines generates marker data, which have multiple uses such as for quality control, grouping of germplasm, trait mapping, and marker-assisted selection. Therefore, molecular markers are becoming more and more an integral part of commercial breeding programs. Multiplex SNP platforms, which analyse hundreds of SNPs simultaneously in a single reaction, are available for routine application in crop breeding (Hyten et al. 2008). In animal breeding, assays are available comprising more than 54,000 evenly spaced SNP probes that span the

bovine genome (Sellner et al. 2007). For whole-genome genotyping in human genetics, SNP chip platforms such as the Affymetrix GeneChip or Illumina Infinium BeadChip platforms provide about 1,000,000 SNPs on a single chip (Ziegler et al. 2008). Crucial for such high-density chips is the knowledge about a sufficient number of SNPs from the target organism. For maize, the genome sequencing of inbred B73 is nearly complete (www.maizesequence.org) and will serve as a foundation for resequencing the gene space of several maize inbred lines, in both, public and private sectors. This will provide the basis for identification of a huge number of SNPs. In addition, owing to ongoing advances in high-throughput marker technologies and automation, the costs per marker data point are expected to further decrease much faster and more substantially than the costs of phenotypic trait evaluation in field trials. This will ultimately facilitate the routine genotyping of inbred lines in commercial breeding programs with high marker density.

**Marker-based prediction of HP.** The marker-based HP prediction is an efficient tool, which supports the selection of superior hybrids and has great potential to accelerate commercial hybrid breeding programs in a very cost-effective manner. The significance of marker-based HP prediction is further enhanced by recent advances in production of DH lines and high-throughput molecular marker technologies. In this thesis research, methods for prediction of HP were developed, which were based on (1) genotypic value estimated from associated markers only or (2) general combining ability (GCA) or line *per se* estimates from observed phenotypes combined with marker-based estimates of specific combining ability (SCA) or heterosis. Further, in this work, linkage disequilibrium (LD) between markers, missing marker observations, and testing the significance of markers were addressed. The prediction methods were assessed with evaluation procedures, which in a defined way split the entire data set into estimation set for calibration and test set for validation of the prediction models. Marker-based prediction of HP can be integrated in several ways into the breeding scheme for improving the efficiency of commercial hybrid breeding programs.

## Linkage disequilibrium between markers

Strong LD between marker loci was observed in populations of inbred lines developed by private companies or by the University of Hohenheim (Maurer et al. 2006; Stich et al. 2006a; Schrag et al. 2007). This correlation between marker loci can result in an increased prediction error due to inflated quantitative trait locus (QTL) effects and in reduced power of QTL detection due to overly stringent adjustment for multiple testing. To address LD in marker-based prediction, (1) a prediction approach based on multiple linear regression (MLR) was developed and (2) three approaches for detection of haplotype block structure (HB1 to HB3) were applied to the prediction methods both in small populations of separate factorial experiments (Schrag et al. 2007) and in a large population of a combination of several experiments with an extremely unbalanced structure of the data (Schrag et al. 2008).

With single marker data, MLR did not generally show superiority over approaches based on total effects of associate markers (TEAM) when applied to separate experiments. There was a tendency that HB2 or HB3 did increase prediction efficiencies for TEAM. Combining both approaches addressing LD (namely MLR with HB) did not clearly increase the prediction efficiency. Rather, when combining MLR with HB1, the prediction efficiencies were often reduced due to the problem of missing marker observations within haplotype blocks. Although the TEAM approaches, when applied to the original marker data, did not account for LD among marker loci, higher prediction efficiencies were obtained than with MLR in the combined analysis of experiments (Schrag et al. 2008). In addition, the use of HB2 or HB3 instead of the original marker data had only marginal effect on prediction efficiencies. Disadvantages of the haplotype block-based approaches were (1) the reduced robustness of the approach due to the requirement of defining several parameters, and (2) that changes in LD structure and amount of missing data due to the addition of new lines to the set of parental inbreds (even if only used for prediction of new hybrids), may require the computationally demanding re-analysis of the

haplotype block structure and re-estimation of their genetic effects. An interesting topic for further research is to investigate, whether a considerable increase in marker density will result in a larger impact of LD on prediction efficiency, so that haplotype blocks will be more advantageous.

## Missing marker observations

Markers, which are affected by one or more missing observations, cannot be used for prediction in the approaches based on total contribution of selected markers (TCSM, Schrag et al. 2006) and MLR (Schrag et al. 2007). With a drop out rate of 3% and N = 79 samples genotyped, under the assumption of equal failure probabilities for all markers and all samples, the expected proportion of usable markers is 9%, which is similar to that observed by Schrag et al. (2008). In commercial breeding programs that have large numbers of inbred lines included in the analysis, this limitation will be even more pronounced. Even with a very low drop out rate of 0.5%, with N = 1000 genotyped samples only 0.7% of all markers would be usable with MLR and TCSM, rendering these approaches extremely inefficient. To overcome such information loss, the TEAM approach was developed, where the average of marker effect estimates was used as a substitute if marker data were missing for the hybrid to be predicted (Schrag et al. 2007). As an interesting alternative to be investigated, best linear unbiased prediction (BLUP) instead of best linear unbiased estimation (BLUE) could be employed to obtain shrinkage estimates of random marker effects (H.-P. Piepho, personal communication). In case of missing marker data, this results in a random marker effect equal to zero. In addition, shrinkage would also address the problem of unequal number of observations in different marker genotype classes.

Missing marker data represent also a major problem for haplotype block analysis, because either (1) missing observations of an affected marker locus carry over to the corresponding haplotype block or (2) affected marker loci do

not join into haplotype blocks, thus, chopping the block structure. Missing data of a single marker can be predicted from observed genotypes of tightly linked markers (Balding 2006). This data imputation can be straightforwardly carried out as single imputation (Scheet and Stephens 2006) or as multiple imputations, which better reflect the uncertainty about the true values (Anderson and Novembre 2003; Souverein et al. 2006). Altogether, data imputation promises to be a simple and effective solution for utilising marker data for HP prediction in the presence of missing marker data points.

## Significance testing of marker effects

Much work has been done on statistical significance methodology for genomewide studies in a more general manner (Storey and Tibshirani 2003; Fernando et al. 2004), specific for QTL mapping in segregating populations (Churchill and Doerge 1994; Cheverud 2001), and for association mapping studies (Dudbridge and Koeleman 2004). The problem of an increased rate of false-positives due to multiple testing of markers in a genome scan was addressed (Schrag et al. 2007) by controlling the false discovery rate at 5% (Benjamini and Hochberg 1995). However, testing for significance of QTL in a genome scan generally bears the risk of false positives and false negatives. In addition, defining appropriate significance levels for marker testing is difficult, because marker-assisted recurrent selection tends to yield a higher selection response at more relaxed significance levels (Bernardo and Charcosset 2006).

Strong LD can be expected between markers and QTL when genotyping with very high marker density. Based on this idea, Meuwissen et al. (2001) suggested to omit least square analysis with given thresholds for significance testing and instead use BLUP and Bayesian approaches for prediction of random effects associated with each chromosome segment of the genome. In simulations studies, response to selection with genomic selection approaches was investigated for development of inbred lines (Bernardo and Yu 2007; Piyasatian

et al. 2007). However, approaches, that adopt the idea of genomic selection for prediction of HP remain to be developed and evaluated for their usefulness in identifying the pairs of inbreds that produce single-cross hybrids with outstanding performance.

## Evaluation of prediction efficiency

The prediction efficiency was evaluated to compare the various prediction methods and judge whether they provide predictions efficient enough to be applicable in breeding programs. This was achieved by cross-validation, in which data were divided into an estimation set for calibration and a test set for validation of the prediction methods. Prediction efficiency was assessed by two statistics, namely the proportion of explained variance ($R^2$) and the square root of the mean square deviation (RMSD) between predicted and observed HP values (Charcosset et al. 1998; Kobayashi and Salam 2000). Higher estimate of $R^2$ and lower estimate of RMSD indicated better efficiency. Whereas $R^2$ is a measurement for fitting of the model to the observations and related to the response to indirect (i.e., marker-based) selection, the RMSD not only considers error due to lack of correlation, but also error due to bias, and, therefore, is adequate for more direct comparison of observed and predicted values.

The marker-based prediction approach of Vuylsteke et al. (2000) was extended by Schrag et al. (2006) and compared with the established GCA-based prediction method. To allow comparisons of results for grain yield (GY) prediction between both studies, a similar leave-one-out evaluation was carried out. By this procedure only one hybrid in each validation run is excluded from the estimation set and because ample information on both parents is still available from other cross combinations, this results in rather high estimates of $R^2$. This approach provides results relevant to a situation, where only a few hybrids drop out of a full factorial mating experiment, e.g., due to crossing failure. In a typical situation of a factorial mating scheme to evaluate

experimental hybrids, the lines from each heterotic group are crossed with only a small subset of lines from the opposite group. This was addressed in a cross-validation procedure, where a constant number of lines was randomly chosen as testers for all opposite lines (Schrag et al. 2007). Although this procedure is different from that employed by Schrag et al. (2006), efficiencies obtained for GCA-based predictions for the same experiments were similar in both studies. Owing to the rather small number of lines in the experiments (especially Exps. 2 and 4), the estimation set represented a large proportion of the entire data as was also the case in Schrag et al. (2007). In addition, for both studies the predicted hybrids were crosses between two testcross evaluated lines (Type 2 hybrids). In these two studies, the prediction efficiency of GCA was higher compared with methods exclusively based on markers.

The number of tested parents of hybrids to be predicted had strong influence on the prediction efficiency. Thus, a cross-validation procedure was developed to evaluate prediction of hybrids of which only one parent (Type 1 hybrids) or even no parent (Type 0 hybrids) was testcross evaluated (Schrag et al. 2008). Prediction efficiency in general was lower for Type 0 than Type 1 hybrids as expected due to lesser information on the parents. In addition, the prediction methods exclusively based on markers associated with HP (e.g., TEAM-H) achieved higher GY prediction efficiency than the phenotypic GCA+SCA method for Type 1 hybrids ($R^2$ 0.49 *vs*. 0.37) and this difference was even more pronounced for Type 0 hybrids ($R^2$ 0.37 *vs*. 0.04). Thus, it was demonstrated that HP can be efficiently predicted, even if only one or none of the parental lines were testcross evaluated, with the marker-based methods developed in the study.

## Proposals for application of marker-based hybrid prediction

Owing to the complete homozygosity of DH lines after chromosome doubling, they can immediately be used for (1) marker genotyping, (2) seed increase,

(3) evaluation of line *per se* performance, (4) production and evaluation of test-crosses for GCA estimation, and (5) production and evaluation of experimental hybrids. Although technically feasible, the instant and broad-scale production of hybrids is not efficient in terms of optimum use of resources because DH lines, unlike conventionally developed inbred lines, have not undergone any previous evaluation and selection for line *per se* performance or GCA. Thus, marker genotyping the DH lines immediately after their development followed by marker-based prediction of HP will particularly enhance the efficiency of DH-based hybrid breeding programs. Marker-based HP prediction can be introduced at various stages of such a DH-based breeding program, e.g.

(1) after field evaluation of experimental hybrids,

(2) after evaluation of line *per se* and testcross performance, but before selection and evaluation of experimental hybrids,

(3) and immediately after development and marker genotyping of DH lines.

These three stages will be described in more detail in the following proposals, which refer to commercial breeding programs for single cross hybrids between DH lines and should be regarded as a starting point for further research to investigate their selection gain and technical feasibility.

**Marker-based HP prediction after evaluation of experimental hybrids.** This procedure involves marker-based HP prediction of untested crosses between inbreds, of which hybrid combinations were already field evaluated. Here, the marker-based prediction is an add-on to complement but not modify established procedures:

(1) Collect field trial data of experimental hybrids from the current breeding cycle, and optionally from previous cycles as well as line *per se* performance of the parental inbreds.

(2) Collect routine marker data of all parental inbreds involved (or genotype them with markers *de novo* if adequate marker data are not available). Use pedigree data or marker data to estimate coefficients of coancestry to be used in the prediction procedure.

(3) Estimate marker effects from the collected data.

(4) Predict the hybrids (they are Type 2 hybrids) with TEAM-H or, if line *per se* data are available, with TEAM-LM (Schrag et al. 2008). In addition, cross-validation of Type 2 hybrids can be carried out to obtain an estimate of the prediction efficiency for the analysed material and trait.

(5) Select the superior but untested hybrids, produce their seed and evaluate them in field trials. Use these results to decide about the development of additional hybrid varieties.

The goal of such a low-key scenario is the selection of superior experimental hybrids, which were not identified with standard procedures such as GCA-based prediction. Its advantages are, that hybrids selected with the classical procedure are not rejected, hybrids between the most current lines are included in the estimation set improving prediction, and the risk of missing superior hybrids is reduced due to additional use of marker-based HP prediction. On the other hand, probably additional genotyping of the parental inbreds and expenses for the prediction analyses are necessary, additional costs arise from production and field evaluation of hybrids selected as a results of marker-based predictions, and the hybrids selected that way will be available only with a delay, which might be the biggest disadvantage keeping in view the continuous gain in selection with each breeding cycle.

**Marker-based HP prediction after DH line evaluation.** This procedure involves marker-based HP prediction of untested crosses between inbred lines, which were evaluated for line *per se* performance and in testcrosses to estimate GCA but not in experimental hybrids:

(1) Develop DH populations from intra-group crosses within heterotic groups and marker genotype all DH lines.

(2) Increase seed of all DH lines by selfing, and evaluate line *per se* performance in field trials. For lines with satisfactory line *per se* performance, produce testcrosses using opposite inbred or F1 testers and evaluate their GCA.

(3) Estimate marker effects from field data of testcrosses. Field data of testcrosses and experimental hybrids from previous cycles, if available, can also be included in the estimation of marker effects.

(4) Collect marker data of all inbred lines included for HP prediction.

(5) Predict the performance of all potential hybrids with TEAM-H and TEAM-LM (Schrag et al. 2008). All predicted hybrids will be of Type 2. However, only a few of the involved parental lines will be extensively tested (either because they were testers in testcrosses or already selected as parents of experimental hybrids). For the current DH lines, evaluation data are only available from the testcrosses.

(6) With marker-based prediction, identify the superior hybrids, produce their seed and evaluate them in field trials. The results from testcross evaluations may be considered additionally for selection of hybrids or hybrid parents.

The use of a few DH testers is preferable over an F1 tester as it has advantages for the estimation of marker effects and, in addition, results in the development of testcrosses that are potential hybrid varieties. However, additional costs arise from the production and evaluation of a considerably larger number of testcrosses, which may only partially be compensated by a reduction in the number of replicates per trial. The goal of this procedure is to accelerate the development of hybrid varieties by early identification and selection of promising hybrids while still carrying out phenotypic evaluation of the lines under selection.

**Marker-based HP prediction immediately after DH line development.** This procedure involves marker-based HP prediction of crosses between unevaluated DH lines. Hence, the first selection step among DH lines fully relies on prediction based on marker effects estimated from data of the previous breeding cycles:

(1) Develop DH populations from intra-group crosses within heterotic groups and marker genotype all DH lines.

(2) Collect marker data of all inbred lines included for HP prediction.

(3) Predict the performance of all hybrids with TEAM-H (Schrag et al. 2008), using marker effects estimated from field data of previous cycles. The predicted hybrids will be of Type 0 (inter-group crosses between new lines in both heterotic groups), Type 1 (crosses between new lines from a heterotic group with tested lines from the opposite group), or Type 2 (crosses between tested lines, which are usually parents of the intra-group crosses or parents of experimental hybrids of previous cycles). Identify the superior hybrids.

(4) Select lines that are parents of the identified superior hybrids. As an additional criterion, GCA can be estimated from the "virtual" factorial mating experiment consisting of all predicted hybrids.

(5) For selected lines, increase seed by selfing and evaluate line *per se* performance in field trials.

(6) Reject lines with unsatisfactory line *per se* performance and produce testcrosses for the remaining lines. The best lines from the opposite heterotic group (identified in step 4) can be used as testers, which already provides potential hybrid varieties.

(7) Evaluate testcrosses and include also these data to repeat the estimation of marker-effects. Based on the refined marker-effects, repeat the prediction of HP, and select hybrids for seed production and field evaluation.

The goal of this procedure, which relies heavily on marker-based prediction, is to select DH lines based on their marker genotypes well in time to (1) reduce expenditures for labour and field capacities devoted to seed increase, evaluation of line *per* se performance, production and evaluation of testcrosses and experimental hybrids, and (2) accelerate variety development by identifying and producing superior hybrids at a very early stage.

## Studies with larger data sets

Materials investigated in this thesis research were from an academic breeding program and were similar in structure with materials from commercial programs. However, commercial hybrid breeding programs have considerably larger dimensions, generating hundreds or even thousands of inbred lines every year. Large data sets have the advantage of being "real world examples", which by subsampling allow to study precisely the effects of the investigated factors. This was demonstrated, for example, with a very large experimental data set composed of 976 $F_5$ maize testcross progenies from a commercial breeding program, in which the effect of sample size and number of test environments on detection and validity of QTL was investigated (Schön et al. 2004). In analogy, a large data base would enable further investigations of factors influencing the efficiency of marker-based HP prediction methods:

(1) With higher number of observations in the estimation set, an increase in prediction efficiency is expected (Schrag et al. 2008). This effect can be studied by sampling estimation sets of different size from the entire data set.

(2) Hybrids between two testcross evaluated parents (Type 2) are used for the estimation set, and therefore, in a limited data set, evaluation of their prediction efficiency in the test set is difficult (Schrag et al. 2008). However, in a large data set, enough observations are available for both estimation set and test set. This would allow to study the prediction efficiency for Type 2 hybrids, even separately for crosses between inbreds evaluated within the same factorial or in different factorials even tested in different years.

(3) Evaluation data of inbred lines from previous breeding cycles provide a large phenotypic information resource for estimation of marker effects at no additional costs for field data. However, across multiple cycles of a breeding program, LD between markers and QTL is affected by recombination and can reduce the prediction efficiency. This effect can be

investigated by including materials from different breeding cycles in a large data set.

(4) Identifying an optimum marker density is crucial for an efficient application of the approach in breeding programs, even with ongoing reduction of costs per marker data point. In data sets based on very large numbers of molecular markers, the effect of marker density and distribution on the prediction efficiency can be investigated by using different subsets of markers.

Other factors such as the number of QTL for a given trait are difficult to investigate with such studies, if these factors do not vary within the data set, are confounded with other factors, or may not easily be controlled by sampling from the observed data. Computer simulation software for plant breeding programs such as Plabsoft (Maurer et al. 2008) and QU-GENE (Podlich and Cooper 1998) provide powerful tools to investigate such questions. Although simulations can only be carried out under simplifying assumptions, their usefulness was demonstrated with investigations on HP prediction (Bernardo 1999), marker-assisted backcrossing (Frisch and Melchinger 2001; Prigge et al. 2008), LD (Stich et al. 2007), association mapping (Stich et al. 2006b), marker-assisted recurrent selection (Bernardo and Charcosset 2006), and genomic selection (Meuwissen et al. 2001; Wong and Bernardo 2008). Thus, simulation studies provide the means for further investigations on marker-based HP prediction.

## References

Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336-354

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781-791

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

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

Bernardo R, Charcosset A (2006) Usefulness of gene information in marker-assisted recurrent selection: A simulation appraisal. Crop Sci 46:614-621

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

Charcosset A, Bonnisseau B, Touchebeuf O, Burstin J, Dubreuil P, Barriere Y, Gallais A, Denis JB (1998) Prediction of maize hybrid silage performance using marker data: comparison of several models for specific combining ability. Crop Sci 38:38-44

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52-58

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963-971

Dudbridge F, Koeleman BPC (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75:424-435

Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK (2007) Molecular markers in a commercial breeding program. Crop Sci 47:S154-S163

Fernando RL, Nettleton D, Southey BR, Dekkers JCM, Rothschild MF, Soller M (2004) Controlling the proportion of false positives in multiple dependent tests. Genetics 166:611-619

Frisch M, Melchinger AE (2001) Marker-assisted backcrossing for simultaneous introgression of two genes. Crop Sci 41:1716-1725

Gallais A, Bordes J (2007) The use of doubled haploids in recurrent selection and hybrid development in maize. Crop Sci 47:S190-S201

Hyten DL, Song Q, Choi IY, Yoon MS, Specht JE, Matukumalli LK, Nelson RL, Shoemaker RC, Young ND, Cregan PB (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945-952

Kobayashi K, Salam MU (2000) Comparing simulated and measured values using mean squared deviation and its components. Agron J 92:345-352

Longin CFH, Utz H, Melchinger A, Reif JC (2007a) Hybrid maize breeding with doubled haploids: II. Optimum type and number of testers in two-stage selection for general combining ability. Theor Appl Genet 114:393-402

Longin CFH, Utz HF, Reif JC, Wegenast T, Schipprack W, Melchinger AE (2007b) Hybrid maize breeding with doubled haploids: III. Efficiency of early testing prior to doubled haploid production in two-stage selection for testcross performance. Theor Appl Genet 115:519-527

Maurer HP, Knaak C, Melchinger AE, Ouzunova M, Frisch M (2006) Linkage disequilibrium between SSR markers in six pools of elite lines of an European breeding program for hybrid maize. Maydica 51:269-279

Maurer HP, Melchinger AE, Frisch M (2008) Population genetic simulation and data analysis with Plabsoft. Euphytica 161:133-139

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

Piyasatian N, Fernando RL, Dekkers JCM (2007) Genomic selection for marker-assisted improvement in line crosses. Theor Appl Genet 115:665-674

Podlich DW, Cooper M (1998) QU-GENE: a platform for quantitative analysis of genetic models. Bioinformatics 14:632-653

Prigge V, Maurer HP, Mackill DJ, Melchinger AE, Frisch M (2008) Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice. Theor Appl Genet 116:739-744

Röber FK, Gordillo GA, Geiger HH (2005) *In vivo* haploid induction in maize - Performance of new inducers and significance of doubled haploid lines in hybrid breeding. Maydica 50:275

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629-644

Schmidt W (2004) Hybridmaiszüchtung bei der KWS SAAT AG. (In German). Bericht über die 54. Tagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs 2003, Gumpenstein. pp 1-6

Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485-498

Schrag TA, Melchinger AE, Sørensen 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 H-P, 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, Sørensen AP, Melchinger AE, Piepho H-P, Frisch M (2008) Haplotype- and marker-based prediction of hybrid performance in maize utilising incomplete data from different factorial experiments. Theor Appl Genet. *In review*

Seitz G (2005) The use of doubled haploids in corn breeding. Proceedings of the 41st annual Illinois corn Breeders' School 2005, Urbana-Champaign. pp 1-7

Sellner EM, Kim JW, McClure MC, Taylor KH, Schnabel RD, Taylor JF (2007) Board-invited review: Applications of genomic information in livestock. Journal of Animal Science 85:3148-3158

Souverein OW, Zwinderman AH, Tanck MWT (2006) Multiple imputation of missing genotype data for unrelated individuals. Annals of Human Genetics 70:372-381

Stich B, Maurer HP, Melchinger AE, Frisch M, Heckenberger M, van der Voort JR, Peleman J, Sørensen AP, Reif JC (2006a) Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Mol Breed 17:217-226

Stich B, Melchinger AE, Piepho HP, Heckenberger M, Maurer HP, Reif JC (2006b) A new test for family-based association mapping with inbred lines from plant breeding programs. Theor Appl Genet 113:1121-1130

Stich B, Melchinger AE, Piepho HP, Hamrit S, Schipprack W, Maurer HP, Reif JC (2007) Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. Theor Appl Genet 115:529-536

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440-9445

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

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

Ziegler A, König IR, Thompson JR (2008) Biostatistical aspects of genome-wide association studies. Biometrical Journal 50:8-28

# Chapter 6
# Summary

Maize breeders develop a large number of inbred lines in each breeding cycle, but, owing to resource constraints, evaluate only a small proportion of all possible crosses among these lines in field trials. Therefore, predicting the performance of hybrids by utilising the data available from related crosses to identify untested but promising hybrids is extremely important. The objectives of this thesis research were to develop and evaluate methods for marker-based prediction of hybrid performance (HP) in unbalanced data as typically generated in commercial maize hybrid breeding programs.

For HP prediction, a promising approach uses the sum of effects across quantitative trait loci (QTL) as predictor. However, comparison of this approach with established prediction methods based on general combining ability (GCA) was lacking. In addition, prediction of specific combining ability (SCA) is also possible with this approach, but was so far not used for HP prediction. The objectives of the first study in this thesis were to identify QTL for grain yield and grain dry matter content, combine GCA with marker-based SCA estimates for HP prediction, and compare marker-based prediction with established methods. Hybrids from four Dent × Flint factorial mating experiments were evaluated in field trials and their parental inbreds were genotyped with amplified fragment length polymorphism (AFLP) markers. Efficiency for prediction of hybrids, of which both parents were testcross evaluated (Type 2), was assessed by leave-one-out cross-validation. The established GCA-based method predicted HP better than the approach exclusively based on markers. However, with greater relevance of SCA, combining GCA with marker-based SCA estimates was superior compared with HP prediction based on GCA only.

Linkage disequilibrium between markers was expected to reduce the prediction efficiency due to inflated QTL effects and reduced power. Thus, in the second study, multiple linear regression (MLR) with forward selection was

employed for HP prediction. In addition, adjacent markers in strong linkage disequilibrium were combined into haplotype blocks. An approach based on total effects of associated markers (TEAM) was developed for multi-allelic haplotype blocks. Genome scans to search for significant QTL involve multiple testing of many markers, which increases the rate of false-positive associations. Thus, the TEAM approach was enhanced by controlling the false discovery rate. Considerable loss of marker information can be caused by few missing observations, if the prediction method depends on complete marker data. Therefore, the TEAM approach was improved to cope with missing marker observations. Modification of the cross-validation procedure reflected, that often only a subset of parental lines is crossed with all lines from the opposite heterotic group in a factorial mating design. The prediction approaches were evaluated with the same field data as in the previous study. The results suggested that with haplotype blocks instead of original marker data, similar or higher efficiencies for HP prediction can be achieved.

Marker-based HP prediction of inter-group crosses between lines, which were marker genotyped but not testcross evaluated, was not investigated hitherto. Heterosis, which considerably contributes to maize grain yield, was so far not incorporated into marker-based HP prediction. Combined analyses of field trials from multiple experiments of a breeding program provide valuable data for HP prediction. With a mixed linear model analysis of such unbalanced data from nine factorial mating experiments, best linear unbiased prediction (BLUP) values for HP, GCA, SCA, line *per se* performance, and heterosis of 400 hybrids were obtained in the third study. The prediction efficiency was assessed in cross-validation for prediction of hybrids, of which none (Type 0) or one (Type 1) parental inbred was testcross evaluated. An extension of the established HP prediction method based on BLUP of GCA and SCA, but not using marker data, resulted in prediction efficiency intermediate for Type 1 and very low for Type 0 hybrids. Combining line *per se* with marker-based heterosis estimates (TEAM-LM) mostly resulted in the highest prediction efficiencies of grain yield and grain dry matter content for both Type 0 and Type 1 hybrids. For the heterotic

trait grain yield, the highest prediction efficiencies were generally obtained with marker-based TEAM approaches.

In conclusion, this thesis research provided methods for the marker-based prediction of HP. The experimental results suggested that marker-based HP prediction is an efficient tool which supports the selection of superior hybrids and has great potential to accelerate commercial hybrid breeding programs in a very cost-effective manner. The significance of marker-based HP prediction is further enhanced by recent advances in production of doubled haploid lines and high-throughput technologies for rapid and inexpensive marker assays.

# Chapter 7
# Zusammenfassung

In kommerziellen Maiszüchtungsprogrammen werden zur Entwicklung von ertragreichen Hybridsorten sehr viele Inzuchtlinien erzeugt. Aus der Vielzahl möglicher Kreuzungen kann jedoch in Feldversuchen nur ein geringer Teil auf Ertragsleistung hin geprüft werden. Die Vorhersage der Hybridleistung (HL) durch die Nutzung von Prüfergebnissen verwandter Kreuzungen ermöglicht das Auffinden aussichtsreicher, aber bislang ungeprüfter Hybriden. Ziel dieser Forschungsarbeit ist die Entwicklung von Methoden zur Nutzung molekularer Marker für die Vorhersage der HL auf der Grundlage unbalancierter Daten aus kommerziellen Maiszüchtungsprogrammen.

Ein Erfolg versprechender Ansatz zur Vorhersage der HL nutzt die Summe der Markereffekte von Genloci eines quantitativen Merkmals (quantitative trait loci, QTL); ein Vergleich mit gängigen Vorhersagemethoden, die auf allgemeiner Kombinationseignung (general combining ability, GCA) basieren, fehlt jedoch. Auch eine Vorhersage der spezifischen Kombinationseignung (specific combining ability, SCA) ist mit diesem Ansatz möglich, wurde bislang jedoch nicht für die Vorhersage der HL genutzt. Ziele der ersten Studie waren die Identifikation von QTL für Kornertrag und Korntrockenmassegehalt, die Kombination von GCA mit marker-basierten SCA-Schätzwerten zur HL-Vorhersage und ein Vergleich der marker-basierten Vorhersage mit gängigen Methoden. Hybriden aus vier faktoriellen Dent × Flint Kreuzungsexperimenten wurden in Feldversuchen geprüft und ihre Elternlinien mit molekularen Markern genotypisiert. Durch Kreuzvalidierung mit Einzelwerten wurde die Güte der Vorhersage für Hybriden untersucht, bei denen beide Eltern bereits in Testkreuzungen geprüft worden waren (Typ 2). Dabei wurde mit der gängigen GCA-basierten Methode eine bessere Vorhersage der HL erreicht als mit ausschließlich marker-basierten Methoden. Bei größerer Bedeutung der SCA war die Kombination von GCA mit marker-basierter SCA jedoch dem einfachen GCA-basierten Ansatz überlegen.

Gametenphasenungleichgewicht zwischen Markern kann zur Minderung der Vorhersagegüte führen. Daher wurde in der zweiten Studie eine multiple lineare Regression (MLR) für die HL-Vorhersage genutzt. Darüber hinaus wurden benachbarte Markerloci mit starkem Gametenphasenungleichgewicht zu Haploblöcken zusammengefasst. Ein Ansatz zur HL-Vorhersage auf der Grundlage der Gesamteffekte assoziierter Marker (total effects of associated markers, TEAM) wurde für multiallelische Haploblockdaten entwickelt. Die genomweite Suche nach signifikanten QTL bedingt ein multiples Testen vieler Markerloci und erhöht das Risiko falsch-positiver Prüfergebnisse. Daher wurde für den TEAM-Ansatz die Einhaltung der sog. „false discovery rate" berücksichtigt. Ein beträchtlicher Informationsverlust wird durch das Fehlen weniger Markerdatenpunkte verursacht, wenn die Vorhersagemethode vollständige Daten erfordert. Der TEAM-Ansatz wurde deshalb so weiterentwickelt, dass auch Marker mit fehlenden Werten nutzbar sind. In der Kreuzvalidierung wurde berücksichtigt, dass innerhalb faktorieller Kreuzungsschemata häufig nur ein Teil der Linien einer heterotischen Gruppe mit allen Linien der anderen Gruppe gekreuzt werden. Die Güte der Vorhersagemethoden wurde mit denselben Daten wie in der vorherigen Studie geprüft. Die Ergebnisse zeigten, dass mit Haploblockdaten anstelle der ursprünglichen Markerdaten eine ähnliche oder höhere Vorhersagegüte für HL erzielt werden kann.

Die marker-basierte Leistungsvorhersage von Kreuzungen zwischen Linien, die zwar mit Markern genotypisiert, nicht aber in Testkreuzungen geprüft wurden, war noch nicht erforscht. Ebenso wurde Heterosis, die wesentlich zum Kornertrag von Maishybriden beiträgt, bislang bei der marker-basierten HL-Vorhersage nicht berücksichtigt. Mit einem gemischten linearen Modell wurden in der dritten Studie unbalancierte Daten aus neun faktoriellen Kreuzungsexperimenten zur Ermittlung von „best linear unbiased prediction" (BLUP) Werten für HL, GCA, SCA, Linieneigenleistung und Heterosis von 400 Hybriden analysiert. Durch Kreuzvalidierung wurde die Vorhersagegüte für Kreuzungen zweier Linien untersucht, wovon keine (Typ 0) oder nur eine (Typ 1) in Testkreuzungen geprüft worden war. Die gängige Vorhersagemethode, basierend auf BLUP Werten für GCA und SCA, erzielte für

Typ 1-Hybriden eine mittlere und für Typ 0-Hybriden eine sehr geringe Vorhersagegüte. Die Kombination von Linieneigenleistung mit marker-basierter Heterosis (TEAM-LM) erreichte meist die höchste Vorhersagegüte für Kornertrag und Korntrockenmassegehalt sowohl für Typ 1- als auch Typ 0-Hybriden. Für Kornertrag wurde mit den marker-basierten TEAM Ansätzen generell die höchste Vorhersagegüte erzielt.

In der vorliegenden Arbeit wurden Methoden zur marker-basierten Vorhersage der HL entwickelt und bewertet. Nach diesen Ergebnissen ist die marker-basierte Vorhersage der HL ein effizientes Werkzeug zur Selektion überlegener Hybriden und ermöglicht die Beschleunigung kommerzieller Hybridzuchtprogramme in sehr kosteneffizienter Weise. Insbesondere haben Fortschritte bei (1) der Herstellung von doppelhaploiden Linien und (2) der schnellen und kostengünstigen Markeranalyse mittels Hochdurchsatz-technologien die Voraussetzungen geschaffen, um die in dieser Studie als aussichtsreich eingestuften Vorhersageverfahren künftig erfolgreich in praktischen Maiszüchtungsprogrammen einzusetzen.

# Acknowledgements

I am very grateful to my academic supervisor Prof. Dr. A.E. Melchinger for his advise, suggestions and continuous support during this thesis research. Thanks to Prof. Dr. H.-P. Piepho and Prof. Dr. R. Blaich for serving on my graduate committee. Sincere thanks to Prof. Dr. M. Frisch for many discussions and his persistent advice and support.

Many thanks to Prof. Dr. B.S. Dhillon, Prof. Dr. M. Frisch, Dr. H.P. Maurer, Prof. Dr. A.E. Melchinger, J. Möhring, Dr. J. Peleman, Prof. Dr. H.-P. Piepho, and Dr. A.P. Sørensen for being co-authors of the publications. Special thanks to Prof. Dr. H.F. Utz for his valuable comments and suggestions and to Dr. J. Muminović for the editorial contribution to publications.

I highly appreciate the great help in organisational matters from H. Beck, B. Boesig, B. Devezi-Savula, and S. Meyer, in field experiments from F. Mauch, H. Pöschel, Dr. W. Schipprack, and R. Schoch, in laboratory work from C. Beuter, E. Kokai-Kota, and Dr. B. Kusterer, and in computer services from T. Schmidt.

Many thanks to my office mates Dr. H.P. Maurer and PD Dr. J.C. Reif, and also to S.U. Bhosale, C. Bolduan, H. Burger, Dr. S. Dreisigacker, Dr. K.C. Falke, S. Fischer, Dr. C. Flachenecker, N. Friedl, Dr. G.A. Gordillo, Dr. M. Heckenberger, C. Knopf, V. Kühr, Dr. F. Longin, Dr. Z. Micic, Dr. V. Mirdita, Dr. J.M. Montes, Dr. J. Muminović, Dr. H. Parzies, V. Prigge, Dr. B. Stich, S.I. Strube, Dr. Z. Sušić, H.H. Voß, T. Wegenast, Dr. K. Wilde, and all unmentioned members of the institute for creating a pleasant work environment.

I dedicate my special gratitude to my dearest Mareike and my family.

This research was supported by the Deutsche Forschungsgemeinschaft (DFG).

# Curriculum vitae

| | |
|---|---|
| Name | Tobias Alexander Schrag |
| Date and Place of Birth | 31/05/1974 in Mutlangen/Ostalbkreis |
| Nationality | German |
| Marital Status | married |
| School Education | 1980-1984, elementary school (Grundschule in Heiningen)<br>1984-1993, high school (Freihof-Gymnasium in Göppingen), Abitur 05/1993 |
| Civil Service | 10/1993-12/1994, Diakonie in Stetten i.R., Texdat in Weinheim/Bergstr. |
| Professional Education | 08/1995-08/1997, Horticulture apprenticeship, Pressmar in Kuchen/Fils |
| University Education | 10/1997-09/1999, Horticulture sciences, University of Hannover (Vordiplom)<br>09/1999-06/2000, Horticulture sciences, SAC Auchincruive and Strathclyde University in Glasgow (UK)<br>08/2000-08/2003, Horticulture sciences, University of Hannover (Diplom-Agraringenieur)<br>since 09/2003, Doctorate candidate in Applied Genetics and Plant Breeding (Prof. Dr. A.E. Melchinger) at University of Hohenheim, Stuttgart |
| Agricultural Experiences | 07/2000, Hild Samen/Nunhems in Marbach/ Neckar |
| Employment Record | since 05/2007, Research & development scientist, KWS SAAT AG in Einbeck |