Technische Universität München

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Pflanzenzüchtung

Genome-based prediction of testcross performance in maize (*Zea mays* L.)

Maria **Theresa Albrecht**

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. A. Tellier
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. C.-C. Schön
	2.	Univ.-Prof. Dr. H.-R. Fries

Die Dissertation wurde am 26.08.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 22.12.2014 angenommen.

Contents

List of Figures

List of Tables

List of Abbreviations

ASC	average silhouette coefficient
BLUE	best linear unbiased estimation
BLUP	best linear unbiased prediction
CS	calibration set
CV	cross-validation
CV-A	cross-validation with sampling across families
CV-aG	cross-validation across groups and tester
CV-R	cross-validation with random sampling
CV-W	cross-validation with sampling within families
DH	doubled haploid
dt	deciton
ES	estimation set
G	group
GBLUP	genome-based best linear unbiased prediction
GCA	general combining ability
GDC	grain dry matter content
GDY	grain dry matter yield
GP	genomic prediction
ha	hectare
LD	linkage disequilibrium
MAF	minor allele frequency
Mb	Megabases
PBC	point-biserial correlation
PBLUP	pedigree-based best linear unbiased prediction
PC	principal component
PCA	principal component analysis
QTL	quantitative trait loci
RRBLUP	ridge regression best linear unbiased prediction
SCA	specific combining ability
SNP	single nucleotide polymorphism
T	tester
TS	test set
UPGMA	unweighted pair group method with arithmetic mean
VC	VeraCode
VS	validation set

1 Introduction

Maize (*Zea maize* L.) originates from Mexico and was domesticated around 9000 years ago (Matsuoka et al. 2002). Today, maize is one of the most common crops grown in the world with a production of 850 Million tonnes for food, fodder, and bioenergy in 2010 (FAO 2010) demonstrating the achievements of maize breeding all over the world. In Germany, maize grain yields were on average 107.2 dt/ha and for silage production 476.1 dt/ha in 2011 (DMK 2011).

The revolution in maize breeding started with the first description of heterosis observed between a cross of two inbred lines by Shull (1908, 1909). Since the early 1950s, the first hybrid breeding programs started in Germany to exploit this heterosis effect and until today they have replaced traditional open pollinated varieties or landraces within breeding programs. For hybrid breeding, recurrent selection is conducted to genetically improve quantitatively inherited traits within the breeding population with including new material into every selection cycle (Hallauer and Miranda 1985). At the beginning of a breeding cycle, several crosses are generated from elite lines derived from the current breeding population. Out of these crosses, doubled haploid (DH) lines are produced by pollinating the S_0 plants with an inducer line (Röber et al. 2005). Each recurrent selection cycle includes multiple years where the selection candidates (DH lines) are crossed to several tester lines from the opposite pool and evaluated as testcrosses within multi-environment trials to assess their genetic potential and general combining ability (GCA). After several selection steps, hybrid performance is evaluated in factorial crosses to assess the specific combining ability (SCA) of possible hybrid partners. The selected lines will be used as parental lines for a new generation of DH lines and form the next recurrent selection cycle. Further improvements of the breeding methodology are now expected with the implementation of genome-based prediction of the performance of selection candidates, which has already been successfully implemented in animal breeding programs (Schaeffer 2006; Jannink et al. 2010).

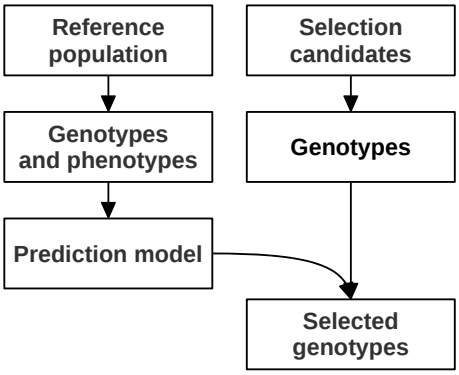


Figure 1: Concept of genome-based prediction.

1.1 Genome-based prediction

Basic concept of genome-based prediction

Historically, genome-based prediction (GP) was developed in animal breeding and led to improvements in assessing the potential of selection candidates (Goddard and Hayes 2007). First results were encouraging and genomic prediction was successfully implemented in cattle breeding programs (Hayes et al. 2009a; VanRaden et al. 2009). As illustrated in Figure 1, genomic prediction models are developed based on a large reference population for which genotypic and phenotypic data are available. The genetic value of non-phenotyped candidates can then be predicted and the candidates are selected based on their genomic information alone. In contrast to marker-assisted selection, GP does not aim at the selection of markers based on quantitative trait loci (QTL) mapping but uses all available markers for predictions. The assumptions behind this approach are based on the hypothesis that with a sufficiently high density of markers all available QTL are in linkage disequilibrium (LD) with random markers segregating in a population under study (Goddard and Hayes 2007). Whittaker et al. (2000) proposed a linear mixed model to estimate the effects of all available markers simultaneously. These estimates can be employed to predict breeding values of non-phenotyped animals (Meuwissen et al. 2001). In dairy breeding, this led to enormously shortened selection cycles, because the phenotyping of thousands of daughters of each bull took several years before the breeding value of a bull could be estimated.

In animal breeding, mixed models had already been implemented using the pedigree-

based kinship coefficients to estimate breeding values of selection candidates (Henderson 1984). Based on genome-wide marker information, the realized kinship matrix has now replaced the traditional kinship for best linear unbiased predictions (BLUPs) (Habier et al. 2007; VanRaden 2008). Due to the "kernel trick" (Schölkopf et al. 1998), this so called genome-based BLUP (GBLUP) model can be conveyed to the model where the markers are fitted directly (see Appendix) but is computationally more efficient, when the number of markers exceeds the number of individuals. In contrast to ridge regression, where all markers are assumed to have equal variance, further models have been developed taking unequal variances for each marker into account. For example, Bayesian prediction methods as BayesA and BayesB have been widely adopted in genomic prediction (Meuwissen et al. 2001), but their sensitivity to different hyper-parameter settings has been recently demonstrated by Lehermeier et al. (2013). Furthermore, none of the Bayesian models has outperformed GBLUP when applied to experimental data and traits with similar genetic architecture as the traits analyzed in this study (Wimmer et al. 2013). Improvements in prediction accuracy might be achieved with more complex models as compared to GBLUP (Ober et al. 2011).

Genome-based prediction in plant breeding

The selection cycle in maize breeding starts with several initial crosses of elite material to produce DH lines, which are fully homozygous. This breeding scheme leads to different family structures and higher LD as observed for animal breeding, where the degree of homozygosity is low. Furthermore, the DH lines are crossed to testers from the opposite heterotic pool and are evaluated in replicated multi-environmental trials to assess trait specific genotype by environment interactions. With GP, the genetic values of testcrosses can be predicted based on related lines for which genotypic and phenotypic data are available even in the absence of reliable pedigree data. Further advantages of GP compared to pedigree-based selection are the differentiation of lines with equal expected relatedness and the control of inbreeding rates during selection processes while selection intervals can be shortened (Heffner et al. 2009). Challenges for the implementation of GP within maize breeding schemes include unbalanced breeding designs and that extensive genetic substructures can occur in the breeding material when new material from related breeding populations is introgressed into the main germplasm of the cur-

rent breeding population. One further question for the implementation of GP into plant breeding programs is the optimal allocation of resources, i.e., number of lines, number of environments and the number of markers necessary to obtain reliable predictions.

The efficiency of GP compared to phenotypic selection can be derived from the "Breeder's equation" (Falconer and Mackay 1996), where the response to direct phenotypic selection (R_P) can be expressed as

$$R_P = i \cdot \sigma_g \cdot h, \tag{1}$$

where i is the selection intensity, h is the square-root of the trait heritability defined as the ratio of genetic to phenotypic variance and σ_g is the square-root of the genetic variance. Based on this equation, the relative efficiency of GP (R_M), which can be seen as indirect selection based on marker data, compared to direct phenotypic selection is:

$$\frac{R_M}{R_P} = \frac{i_M \cdot \sigma_g \cdot h_M \cdot r(\hat{g}, g)}{i \cdot \sigma_g \cdot h}, \tag{2}$$

where i_M is the selection intensity of indirect selection, h_M is the heritability of the marker, which is assumed to be $h_M = 1$, if the marker genotypes can be assessed without error, and $r(\hat{g}, g)$ is the correlation between predicted and true genetic values of the selection candidates (Technow et al. 2013). This correlation is also known as prediction accuracy and can be used for the comparison of different genome-based prediction models.

1.2 Assessing predictive ability

Different methods have been developed to assess the predictive ability or accuracy of prediction models. The need of useful validation procedures was proposed by Kurtz (1948) in the field of psychology, who criticized that models for the prediction of success of insurance sales managers based on a Rorschach test were not validated on an independent sample. However, in most studies only one sample is available and validation on an independent sample is time and cost intensive.

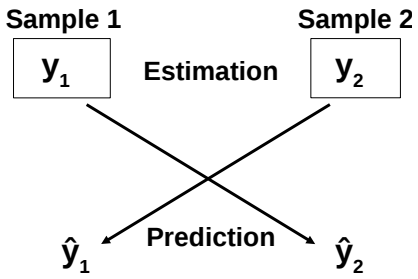


Figure 2: Basic scheme for cross-validation modified according to Mosier (1951).

Therefore, Mosier (1951) proposed a method to divide the data set into two disjoint samples (Figure 2). Model parameters are then estimated in one sample and used for prediction and cross-validation (CV) in the other sample. With this method one can validate the prediction models on disconnected subsets without the assessment of an additional data set. Advantages of CV compared to other model quality measures, e.g., Akaike information criterion, are to obtain out-of-sample estimates for the prediction accuracy even for non-nested models. On the other hand, dividing the data set into subsets causes a loss of information incurred in model development. Hence, optimizing the CV schemes to decrease the bias due to information loss without increasing the variance across subsets has been subject to previous studies (Breiman and Spector 1992; Utz et al. 2000). Furthermore, stratified CV schemes have been employed to account for distinct family structures (Legarra et al. 2008). Overall, CV can be applied to evaluate genome-based prediction models to obtain a direct estimate of the prediction accuracy achieved within different sets of breeding populations. However, first experimental studies in plant breeding have shown that predictive abilities obtained with CV in one year can only partly reflect the prediction of an independent sample evaluated in a different year or environment (Hofheinz et al. 2012; Utz et al. 2000). Therefore, the naive application of CV is not appropriate to obtain an adequate (unbiased) estimator of the predictive ability for a typical plant breeding scenario where the interest lies in predicting related crosses in a different year. For an overall evaluation of the potential of GP in maize breeding, results need to be validated on data derived from different breeding cycles and years.

1.3 Accounting for genetic structure

In contrast to animal breeding, populations in plant breeding are open for the introgression of new material during every recurrent selection cycle (Gordillo and Geiger 2008). Therefore, the training population might consist of individuals belonging to different genetic groups from the same heterotic pool, which can have a strong effect on the prediction of testcross values for different traits. As indicated by Windhausen et al. (2012), predictive abilities are highly affected by population structure, when the genetic groups differ in their mean performance. In addition, the choice of tester might be confounded with the maturity of the selection candidates such that early testers are crossed to late material of the breeding population and vice versa, but also with utilization aspects like silage or grain use. Furthermore, phenotyping of all possible tester combinations within early stages of a breeding cycle is usually not feasible, making predictions across genetic groups, testers, and years of interest for maize breeders.

In the context of GP, one main focus of studies dealing with experimental data has been the comparison of different GP models to improve prediction accuracies within plant populations, e.g., diversity panels (Crossa et al. 2010; Riedelsheimer et al. 2012; Rincent et al. 2012) or breeding populations (Zhao et al. 2012; Heslot et al. 2012). These plant populations differ mainly in their genetic substructure influencing predictive abilities. While diversity panels are designed to present a wide range of genetic variation, breeding populations comprise more closely related material derived from single or multiple crosses adapted to specific environmental conditions. The potential of GP in highly structured plant populations comprising different genetic groups or breeding cycles has been less investigated (Heslot et al. 2012; Hofheinz et al. 2012; Technow et al. 2012). However, these predictions are of high relevance in advanced cycle breeding populations.

In association studies, accounting for genetic substructure has already been discussed to correct for spurious associations due to relatedness (Astle and Balding 2009). When the genetic substructure is not known *a priori*, principal components and cluster analysis can be applied to marker data to detect subgroups in the material. A principal component analysis (PCA) represents the genetic relatedness between individuals in reduced dimensions with capturing the original variation. The principal components (PC) are linear combinations of the original data space derived from a single value decomposi-

tion (Schölkopf et al. 1998). The properties of PCs are that they are orthogonal and the first principal component explains most of the variability in the data set. To detect groups in data, different clustering methods are available, e.g., hierarchical clustering methods as UPGMA (unweighted pair group method with arithmetic mean; Sokal and Michener (1958)) and Ward's minimum variance (Ward 1963) method or k-means clustering (Hartigan and Wong 1979).

1.4 Objectives

The main goal of this thesis was to assess the potential of genome-based prediction models in maize breeding. For this purpose, genomic prediction of testcross values was analyzed in two experimental data sets representing typical maize breeding programs. The first experimental data set used here consists of 1377 DH lines genotyped with 1152 single nucleotide polymorphism (SNP) markers, pedigree data, and phenotypic data of two traits, grain dry matter yield (GDY) and grain dry matter content (GDC). A subset of this data set was additionally genotyped with a high density SNP array. The second data set consists of two calibration sets comprising 1073 and 857 DH lines evaluated as testcrosses in two different years, where two traits, GDY and GDC, were assessed. All lines were genotyped with a high density marker array. A selected set of DH lines from the first calibration set was additionally evaluated with several testers in the following year. The second data set can be further characterized according to the distinct genetic substructure and the different testers used for producing the testcrosses. Therefore, the objectives of this thesis were to

1. compare kinship coefficients between DH lines based on pedigree and genome-wide marker data in the context of prediction models,
2. assess the potential of prediction within biparental families or genetic groups versus population-wide prediction,
3. assess the potential of prediction across groups, testers, locations and years,
4. evaluate the impact of number of individuals, markers, and locations on predictive abilities, and
5. determine the impact of family and subpopulation structures on the predictive ability within two experimental data sets.

2 Materials and Methods

2.1 Plant materials

2.1.1 Maize 1

The first data set comprised a total of 1380 doubled haploid (DH) lines of maize (*Zea mays* L.). Thirty-six families were derived from crosses among 29 inbred lines and four single crosses all belonging to the dent heterotic group. Resulting S_0 plants were used for production of DH lines, which was performed with the *in vivo* haploid induction technology according to Röber et al. (2005). S_0 plants were pollinated with inducer line RWS and on average 38 DH lines per cross were produced. The smallest DH family comprised 14, the largest 60 lines. For all lines full pedigree information was available up to three generations (Appendix, Figure A1). The four largest biparental families, comprising 58-60 DH lines, were analyzed separately to assess prediction accuracy within individual families.

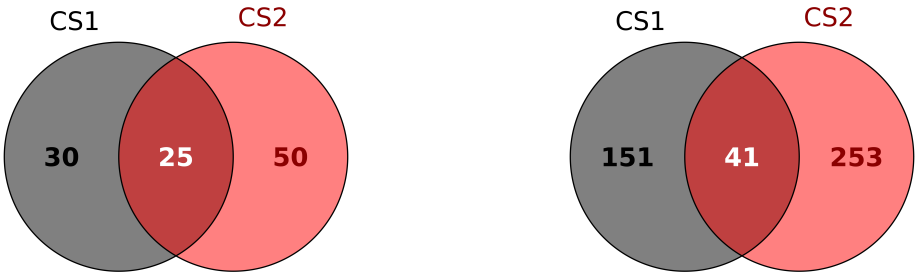
2.1.2 Maize 2

The second data set was derived from a different maize breeding program and consists of two sets of genetic material from the maize dent pool comprising 1073 and 857 DH lines. The two data sets represent two selection cycles of the same breeding program and form two calibration sets for GP.

In calibration set 1 (CS1), the 1073 DH lines were derived from 192 crosses among 55 parents (43 inbred lines and 12 single-crosses). In calibration set 2 (CS2), the 857 DH lines were obtained from 294 crosses among 75 parents (55 inbred lines and 20 single-crosses). The two calibration sets were connected by 25 parents (Figure 3). The number of DH lines derived from each cross was highly variable ranging from 1 to 63 DH lines with an average of 6 DH lines in CS1 and from 1 to 26 DH lines and an average of 3 DH lines in CS2. The number of progeny per parent ranged from 1 to 203 with an average of 39 DH lines in CS1 and from 1 to 130 with an average of 23 DH lines in CS2. Pedigree records included 479 ancestors representing a minimum of three generations. The DH

lines can be assigned to three *a priori* defined groups (G1, G2, and G3) of different genetic background, where G1 represents the core germplasm of both calibration sets (Appendix, Figure A2). Figure 4 shows the connection between groups G1, G2, and G3 based on parents used in each of the two calibration sets.

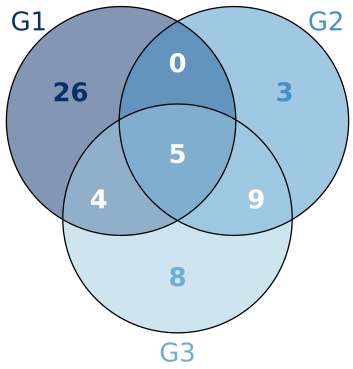
A subset of 97 DH lines from CS1 was selected based on their phenotypic performance to be evaluated in a second testcross cycle and forms a validation set (VS1) for CS1 and CS2. The DH lines belong to all three genetic groups ($N_{G1} = 67$, $N_{G2} = 8$ and $N_{G3} = 22$).



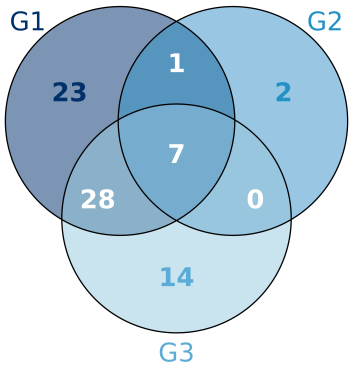
(a) Parental lines across CS1 and CS2

(b) Families across CS1 and CS2

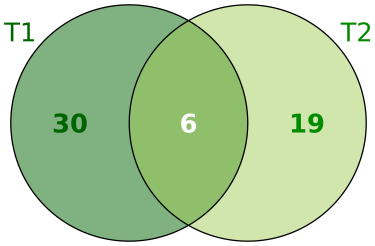
Figure 3: Venn-Diagram with parental lines and families across calibration sets CS1 and CS2 in Maize 2.



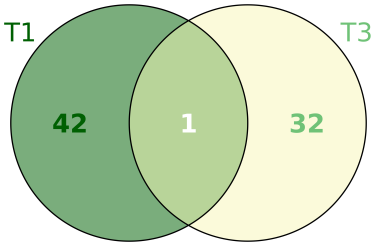
(a) Groups in CS1



(b) Groups in CS2



(c) Testers in CS1



(d) Testers in CS2

Figure 4: Venn-Diagram with parental lines across groups and testers for calibration set CS1 and CS2 of Maize 2.

2.2 Phenotypic analysis

2.2.1 Maize 1

All 1380 DH lines were evaluated as testcrosses with a single-cross tester in 2009 in seven European locations. Two-row plots were machine-planted and harvested as grain trials. Data were recorded for grain dry matter yield (GDY, dt/ha) and grain dry matter content (GDC, %). In each of the seven locations the experimental design consisted of 15 sets with 100 entries each. Trials were performed with one replication per location. Each set contained 92 DH lines and four checks each replicated twice. Outlying observations were removed from the data set based on extreme deviate standardized residuals according to Grubbs (1950). For each environment, trait values were adjusted for the effects of the sets based on the means of the replicated check varieties. Due to unreplicated field trials,

trait heritabilities were estimated with the pedigree-based prediction model (Eqn. 15).

2.2.2 Maize 2

Phenotypic performance of the DH lines of Maize 2 was evaluated in replicated field trials as testcrosses with a single-cross (T1) or double-cross tester (T2) for CS1 in 2010 and with one of two single-cross testers for CS2 in 2011 (T1 or T3). The three testers were derived from six parental lines (A, B, C, D, E, and F) from the opposite heterotic flint pool, with T1 being a cross of A × B, T2 was C.D × B.E, and T3 was B × F. For the VS1 in 2011, each DH line was tested with up to three testers, of which tester T1 was in common with both calibration sets. Each of the 192 crosses in CS1 and 294 crosses in CS2 was assigned to either of nine possible group/tester combinations. In CS1, four group/tester combinations were evaluated and in CS2 five combinations. A detailed summary of the distribution of DH lines across groups and testers in each calibration set is given in Table 1.

Field trials were conducted in six German locations in 2010 and 2011, where four locations were represented in both years. In 2010, entries were distributed across 16 trials. Each trial was conducted in four of the six locations resulting in 6-16 trials per location. In 2011, entries were distributed across 12 trials and each trial was conducted in three or four of the six locations, resulting in 4-12 trials per location. Phenotyping of the VS1 in 2011 was carried out in nine different locations consisting of 1-8 trials. Each trial was laid out in a 10 × 10 lattice design with two replications containing 94 entries and six

Table 1: Number of DH lines in different groups and tester subsets for (a) calibration set CS1 and (b) CS2.

(a) CS1

	G1	G2	G3	
T1	682	0	16	698
T2	0	145	230	375
	682	145	246	1073

(b) CS2

	G1	G2	G3	
T1	189	0	138	327
T3	393	15	122	530
	582	15	260	857

hybrid checks for CS1 in 2010 and VS1 in 2011 and 95 entries and five hybrid checks for CS2 in 2011. Entries comprised the genotyped DH lines and additional DH lines for which no genotypic data were available.

Measured phenotypic traits were grain dry matter yield (GDY, dt/ha) and grain dry matter content (GDC, %). Outliers were identified and removed based on maximum deviant residuals of a single stage model according to Grubbs (1950). Phenotypic analysis in each year was performed in a two-stage approach. In the first stage, adjusted means were calculated for each location with the following model formulated according to Piepho et al. (2003):

$$y_1 = g : trial/rep/block, \tag{3}$$

where y_1 are plot observations for each location, g is a variable for genotypes including checks and entries, and $trial$, rep and $block$ are the variables for trials, replications, and blocks, respectively. In this model notation, fixed effects are placed before the colon, random effects are placed after the colon.

In the second stage, location-specific adjusted means (y_2) of the first stage were passed to the following model:

$$y_2 = g : C \cdot loc + C \cdot D1 \cdot g \cdot loc + C \cdot D2 \cdot g \cdot loc, \tag{4}$$

with C being an indicator variable assigning genotypes to different factor levels such as checks or group/tester combinations and loc is a variable for random location effects which is assumed to be independent and identically normal distributed. Location specific residuals were assumed to follow a normal distribution with N(**0**,**R**), where **R** is a diagonal matrix containing squared standard errors of means from the first stage (Eqn. 3, Appendix Table A2 and A3) as weights (Möhring and Piepho 2009). The combination of two dummy variables ($D1$ and $D2$) separates the genotyped DH lines ($D1 = 1$, $D2 = 0$), the non-genotyped DH lines ($D1 = 0$, $D2 = 1$), and the checks ($D1 = 0$, $D2 = 0$). These dummy variables were introduced to estimate variance components for genotyped ($\sigma^2_{g_1}$) and non-genotyped entries ($\sigma^2_{g_2}$) and the corresponding genotype by location interac-

tions (Piepho et al. 2006). Estimates of heritabilities on an entry mean basis were calculated on genotyped DH lines only. Variance components were estimated with restricted maximum likelihood using the software ASReml 3.0 (Gilmour et al. 2009). Significance of variance components was tested according to Stram and Lee (1994) based on the following model with random genotype effects for estimating genotypic variances:

$$y_2 = C : C \cdot D1 \cdot g + C \cdot D2 \cdot g + C \cdot loc + C \cdot D1 \cdot g \cdot loc + C \cdot D2 \cdot g \cdot loc. \qquad (5)$$

Due to an unbalanced distribution of trials across locations used for the evaluation of testcrosses, trait heritabilities h^2 were approximated based on genotyped DH lines for both traits and both years in the second stage as follows (Holland et al. 2003):

$$h^2 = \frac{\sigma^2_{g_1}}{\sigma^2_{g_1} + \frac{\sigma^2_{g_1 \times loc}}{L} + \frac{\sigma^2_{e^*}}{L}}, \qquad (6)$$

where $\sigma^2_{g_1}$ is the genetic variance of genotyped DH lines, $\sigma^2_{g_1 \times loc}$ is the corresponding genotype by location interaction variance component, $\sigma^2_{e^*}$ is the residual variance calculated as mean of weights in $\mathbf{R}$ (Eqn. 4) and L is the harmonic mean of number of locations per genotype used in each year ($L = 4$ for CS1 in 2010, $L = 3.3$ for CS2 in 2011, and $L = 5.6$ for VS1 in 2011). Adjusted means from the second stage were submitted to the prediction models (Section 2.5).

2.3 Genotypic analysis

For Maize 1, marker analyses of the 1380 DH lines were performed with a VeraCode assay including 1152 biallelic SNP markers randomly distributed across the genome. For the majority of the markers, the positions in the maize genome were known from their alignment to the B73_RefGen_v1 sequence (Schnable et al. 2009). The average physical distance between adjacent markers was 2.9 Megabases (Mb). Markers with more than 10 % missing values or a minor allele frequency (MAF) < 0.01 were discarded, resulting in 732 useful SNPs in the population of DH lines. Three DH lines were discarded from the analysis due to low-quality marker data.

A subset of Maize 1 including 759 DH lines was additionally genotyped with 56110 SNPs using the Illumina® MaizeSNP50 BeadChip (Ganal et al. 2011). Only high-quality SNPs with a GenTrain-Score ≥ 0.7, a call frequency ≥ 0.9, a MAF ≥ 0.01, and non-redundant SNPs were used for further analysis, resulting in 20742 useful, polymorphic SNPs for 759 DH lines of Maize 1. The 20742 SNPs were equally distributed across the ten chromosomes of the maize genome and had an average distance of 0.11 Mb. For comparing the low-density and high-density SNP panels without an influence of the sample size N, the subset of 759 DH lines were also examined with 654 polymorphic SNPs from the VeraCode assay data.

For Maize 2, DH lines from both calibration sets were genotyped with the Illumina® MaizeSNP50 BeadChip (Ganal et al. 2011). Physical positions on the maize genome were known from an alignment to the B73 AGPv2 assembly (www.maizesequence.org). In both calibration sets, quality control of the DH lines (call rates ≥ 0.9) and SNP pruning based on the quality parameters (GenTrain-Score ≥ 0.7, call frequency ≥ 0.9, MAF ≥ 0.01, and discarding redundant SNPs) was performed, resulting in 15732 polymorphic SNPs for CS1 and 16846 SNPs for CS2. For the joint analysis of both calibration sets, the same SNP selection steps were performed resulting in 17734 polymorphic SNPs for the $N = 1930$ DH lines.

With fully homozygous inbred lines, only two genotypes can be distinguished. Marker genotypes were coded 0 or 2 depending on the number of copies of the minor allele. Missing marker genotypes were imputed based on family information for Maize 1 and based on family information and flanking markers using BEAGLE 3.3.1 (Browning and Browning 2009) and the R package `synbreed` (Wimmer et al. 2012) for Maize 2. If the cross from which the DH line was derived did not segregate at the SNP locus, the missing genotype was set to the genotype carried by its siblings. If the SNP marker did segregate in the respective cross, the genotype was substituted at random with one of the two possible genotypes at a probability of 0.5 (Maize 1) or with BEAGLE 3.3.1 (Maize 2).

Linkage disequilibrium (LD) between marker pairs was calculated as described by Hill and Robertson (1968) using the `synbreed` package (Wimmer et al. 2012) and the PLINK 1.07 software (Purcell et al. 2007). For the low density SNP panel, LD was measured between SNP pairs over the entire genome and for the high density SNP panels for

SNP pairs from the same chromosome. As a measure of LD, the squared correlation r^2 between alleles at two loci was used:

$$r^2 = \frac{(p_{vw} - p_v p_w)^2}{p_v(1-p_v)p_w(1-p_w)},$$

where p_{vw} , p_v, and p_w are the frequencies of the haplotype vw and alleles v at one locus and allele w at the other locus. For the SNPs from the VeraCode assay, significance of LD was further tested using a χ^2-test as suggested by Foulkes (2009). Differences in average MAF between both calibration sets were tested with a Mann–Whitney-Wilcoxon test.

2.4 Modeling the kinship between DH lines

In maize breeding, DH lines are evaluated as testcrosses in field trials to measure their general combining ability (GCA) by crossing them to a common tester from the opposite heterotic pool. In general, the genotypic variance (σ_c^2) and covariance ($\omega_{cc'}$) between relatives of a cross can be expressed as

$$\begin{aligned} \sigma_c^2 &= \sigma_{\alpha_1}^2 + \sigma_{\alpha_2}^2 + \sigma_{\delta_{12}}^2 \\ \omega_{cc'} &= \Phi_1\sigma_{\alpha_1}^2 + \Phi_2\sigma_{\alpha_2}^2 + \Phi_1\Phi_2\sigma_{\delta_{12}}^2 \end{aligned} \tag{7}$$

where $\sigma_{\alpha_1}^2$ and $\sigma_{\alpha_2}^2$ are the variances of GCA effects from parental populations 1 and 2, $\sigma_{\delta_{12}}^2$ is the interaction variance of specific combining ability (SCA) effects and Φ_1 and Φ_2 are the probabilities that alleles originating from populations 1 and 2 are identical by descent. Assuming that the DH lines are crossed to only one common tester, $\sigma_{\alpha_2}^2 = 0$ and $\Phi_2 = 1$ and the GCA and SCA cannot be estimated separately and the sum of both is therefore denoted as testcross variance. Therefore, the testcross variance (σ_t^2, see also Section 2.5) and the covariance ($\omega_{tt'}$) of the DH lines from one population is:

$$\begin{aligned} \sigma_t^2 &= \sigma_{\alpha_1}^2 + \sigma_{\delta_{12}}^2 \\ \omega_{tt'} &= \Phi_1(\sigma_{\alpha_1}^2 + \sigma_{\delta_{12}}^2) = \Phi_1\sigma_t^2 \end{aligned} \tag{8}$$

The probability that alleles are identical by descent can be estimated from pedigree and/or marker data as described in this Section and is denoted as kinship coefficients which are used to model the variance-covariance structure between DH lines.

2.4.1 Expected kinship coefficients

Expected kinship coefficients are based on pedigree data and give the expected probabilities that two alleles are identical by descent. In animal breeding, the numerator relationship coefficient, which is twice the kinship coefficient, was first applied by Henderson (1977) for best linear unbiased prediction of breeding values.

In both data sets, the matrix of expected kinship coefficients **K** was calculated according to Bernardo (2002) based on three generations of pedigree information. This method has been adopted for fully inbred lines in maize breeding and was implemented into the `synbreed` R-package (Wimmer et al. 2012).

2.4.2 Realized kinship coefficients

Instead of using pedigree-based estimates of kinship coefficients, genome-wide marker data make it possible to account for Mendelian sampling effects and increase the differentiation between equally related DH lines. Different methods have been proposed to estimate realized kinship coefficients (U_{ij}) between two individuals i and j from genome-wide marker data. Some coefficients are derived from the coefficient for alleles being identical by state (IBS), which are corrected by the average proportion of alleles being IBS between relatives in the base population. Other coefficients are derived from the correlation between alleles taken from gametes (Powell et al. 2010; Astle and Balding 2009).

Simple Matching

In plant breeding, the similarity between DH lines is frequently estimated with the simple matching coefficient U_{ij}^{SM} (Sneath and Sokal 1973). This method can be applied to

biallelic markers if the inbred lines are fully homozygous and one can form a 2×2 table with counting number of equal and unequal loci between individuals i and j. The pairwise coefficients U_{ij}^{SM} between DH line i and j was calculated from SNP genotypes as follows:

$$U_{ij}^{SM} = \frac{\sum_{m=1}^{M}[(w_{im}-1)(w_{jm}-1)]+M}{2M}, \tag{9}$$

with w_{im} and w_{jm} being the genotype scores for lines i and j coded as 0 or 2 for homozygous loci $m = 1, ..., M$, with M being the number of markers. The simple matching coefficient can be interpreted as average probabilities of alleles being IBS between two individuals (Astle and Balding 2009; Piepho 2009), which has to be corrected with the average proportion of alleles being IBS between relatives in the base population (Powell et al. 2010). Following Hayes and Goddard (2008), each element of the matrix was transformed with u_{min}, which is the minimum value of all pairwise similarity coefficients:

$$U_{ij}^{S} = \frac{U_{ij}^{SM} - u_{min}}{1-u_{min}}. \tag{10}$$

$$U_{ij}^{S} = \frac{\sum_{m=1}^{M}[(w_{im}-1)(w_{jm}-1)]+M+2Mu_{min}}{2M(1-u_{min})}. \tag{11}$$

This transformation leads to off-diagonal values between 0 and 1, while the diagonal elements are equal to $\frac{1}{2}(1+F)=1$, where F is the inbreeding coefficient of the DH lines in the population.

Trait specific kinship

The trait specific kinship is an extension of the previous kinship coefficient U_{ij}^{SM} according to Zhang et al. (2010). This kinship incorporates specific weights for each marker based on the estimates of marker effects for a trait in the reference population. The formula for the trait specific kinship coefficient $U_{ij}^{T'}$ is given by

$$U_{ij}^{T'} = \frac{\sum_{m=1}^{M}[(w_{im}-1)(w_{jm}-1)\cdot\omega_m] + M\cdot\sum_{m=1}^{M}\omega_m}{2M\cdot\sum_{m=1}^{M}\omega_m},$$

with ω_m being a trait specific weight for each marker m calculated as $2p_m(1-p_m)\cdot\hat{m}_m^2$, where $\hat{m}_m^2$ is the squared effect of marker m for a specific trait estimated within the reference population and p_m is the MAF of marker m in the population under study.

These kinship coefficients are again corrected by the minimum entry of the trait specific kinship matrix $\mathbf{U}_{T'}$ with

$$U_{ij}^{T} = \frac{U_{ij}^{T'} - min(\mathbf{U}_{T'})}{1 - min(\mathbf{U}_{T'})}$$

to obtain unbiased estimates of the kinship coefficients.

Kinship based on centered marker scores

The realized kinship coefficient according to Habier et al. (2007) and VanRaden (2008) is derived from the following formula

$$U_{ij} = \frac{1}{2}\frac{\sum_{m=1}^{M}[(w_{im}-2p_m)(w_{jm}-2p_m)]}{\sum_{m=1}^{M}2p_m(1-p_m)}. \tag{12}$$

The subtraction of the expected value $2p_m$ leads to mean-centered marker scores and the average of pairwise kinship coefficients between DH lines becomes zero. In contrast to kinship coefficients derived from the simple matching coefficient as described in the previous paragraph, this formula can lead to negative off-diagonal values. Hence, the direct interpretation as a probability of alleles being IBD does not hold, but an interpretation as a coefficient of correlation is straightforward (Powell et al. 2010). The diagonal elements also have a wider range than traditional kinship coefficients, but their expected values within the population is $\frac{1}{2}(1+F)$ with $F=1$ for DH lines.

Kinship based on centered and scaled marker scores

A similar approach was proposed by Astle and Balding (2009), but in contrast to VanRaden (2008), the marker scores are centered and scaled to have equal variances. Therefore, the formula for realized kinship coefficients is

$$U_{ij}^{AB} = \frac{1}{2M} \sum_{m=1}^{M} \frac{(w_{im} - 2p_m)(w_{jm} - 2p_m)}{2p_m(1-p_m)}. \tag{13}$$

Here, the formula puts more weight on markers with low allele frequencies, but problems can occur for markers having an allele frequency close to zero. This causes the undesirable property that the realized kinship coefficients between lines tend to infinity when they carry rare alleles with a MAF approaching to zero (Endelman and Jannink 2012).

2.4.3 Analysis of kinship structure

In Maize 2, relatedness of DH lines within and between the three genetic groups G1, G2, and G3 of both calibration sets was analyzed based on pedigree information. When information about genetic substructures in a data set is given *a priori*, the maximum kinship coefficient between a DH line i and all other lines from its own group ($k_{max,i}$ within) should be significantly higher than the maximum kinship coefficient between the DH line and all lines from the other groups ($k_{max,i}$ between)(Saatchi et al. 2011). For example, maximum kinship $k_{max,i}$ between individual i belonging to group G1 was derived from the following formula

$$\begin{aligned} k_{max,i,G1} &= max(k_{ij}) \quad \text{with} \quad i \in G1, j \in G1, j \neq i \\ k_{max,i,G2} &= max(k_{ij}) \quad \text{with} \quad i \in G1, j \in G2 \\ k_{max,i,G3} &= max(k_{ij}) \quad \text{with} \quad i \in G1, j \in G3 \end{aligned}$$

where k_{ij} is the expected kinship coefficient between individual i and j. Hence, for each DH line, three $k_{max,i}$ values were calculated. All $k_{max,i}$ $(i = 1, ..., N_{G1})$ values were

averaged over DH lines within group G1 and across G2 and G3 resulting in a mean $\overline{k}_{max}$ value within and between groups for G1. In addition, kinship coefficients between DH lines from the same group were averaged to result in mean kinship coefficients $\overline{k}$. Calculations were performed within and across all groups and tester subsets of both calibration sets and across calibration sets. In CS2, pairwise $\overline{k}$ and $\overline{k}_{max}$ values were also estimated between group/tester combinations.

As the range of the kinship coefficient derived from the simple matching coefficient (Eqn. 11) are equal to the pedigree-based kinship coefficients, the results from the kinship structure analysis described above can be compared with this realized kinship coefficient. Therefore, the kinship within and between groups and testers of Maize 2 were additionally analyzed with the modified simple matching coefficient.

Cluster analysis

For proper clustering results, decisions are required on an appropriate distance metric and clustering algorithm. To compare how different cluster methods can reveal genetic substructures in experimental data, three clustering methods, UPGMA, Ward, and k-means, were applied to the genotypic data of calibration sets CS1 and CS2 of Maize 2. Both hierarchical cluster analyses were performed with Rogers' distance as a genetic distance measure between homozygous DH lines based on biallelic SNP markers (Rogers 1972). For pairs of fully homozygous lines, Rogers' distance D_{ij} can derived from $1-U_{ij}^{SM}$ where U_{ij}^{SM} is the simple matching coefficient (Eqn. 9).

In UPGMA clustering, the distance between clusters i and j is calculated as average of all pairwise distances between all lines in cluster i and all lines in cluster j. In each step of the clustering process, the two clusters with the smallest average pairwise distance are joined. Ward's clustering method is based on an analysis of variance between the clusters. Two clusters are merged, if they yield the lowest error sum of squares, which is defined as the sum of squared distances between the line of the cluster and its centroid and is used as a measure of the tightness of a cluster (Kaufman and Rousseeuw 2005). For the third clustering approach, the k-means clustering, the number of clusters has to be predefined. In contrast to Ward's clustering, the observations are partitioned into a

specific number of clusters and the sums of squared distances between observed values in each cluster and the center of the cluster are minimized. For this study, the algorithm according to Hartigan and Wong (1979) which is implemented in R was applied directly to the genotype matrix.

The following two measures were used to determine the optimal number of clusters: point-biserial correlation (PBC) and average silhouette coefficient (ASC) (Odong et al. 2011). The PBC is the correlation between the original distance matrix and a reduced distance matrix with zeros and ones indicating lines in different and same clusters, respectivly. The ASC relates the distance between two lines in the same cluster with distances between lines not belonging to the same clusters. The PBC and ASC range between zero and one and the highest value indicates the optimum number of clusters.

To illustrate the genetic substructures which occurred in both calibration sets of Maize 2, a PCA was applied to the column-centered and scaled genotype matrix of both calibration sets separately.

2.5 Prediction models

For the prediction of testcross values in maize breeding, mixed effect models including best linear unbiased estimations (BLUEs) for the fixed effects and best linear unbiased prediction (BLUPs) for the random effects were used (Henderson 1984). In contrast to animal breeding, the variance-covariance structure of the random testcross effects is determined by the joint action of GCA of DH lines and the SCA of the testcrosses as described in Eqn. 8.

2.5.1 Pedigree-based best linear unbiased prediction

In the following model, the variance-covariance structure of testcross effects were modeled using pedigree-based estimates of kinship coefficients. The model is denoted as PBLUP and is described as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{t} + \mathbf{e}, \tag{14}$$

where $\mathbf{y}$ is a $N \times 1$ vector of adjusted means for N DH lines obtained from the phenotypic analysis (Eqn. 4); $\boldsymbol{\beta}$ is a vector of fixed effects, for Maize 1 $\boldsymbol{\beta}$ includes the overall mean and for Maize 2 $\boldsymbol{\beta}$ is a $c \times 1$ vector comprising $c = 4$ (CS1) and $c = 5$ (CS2) factor levels for group/tester combinations (see Table 1). The fixed effect was included to correct for genetic substructure within both calibration sets. Random testcross effects were modeled with the $N \times 1$ vector $\mathbf{t} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}\sigma_t^2)$, where $\mathbf{K}$ is a $N \times N$ matrix of pedigree-based kinship coefficients and σ_t^2 is the testcross variance as defined in Eqn. 8 pertaining to this model. The design matrices $\mathbf{X}$ and $\mathbf{Z}$ assign the adjusted means to the fixed and random effects, respectively. The $N \times 1$ residual vector $\mathbf{e}$ is assumed to be independent and normally distributed with $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_{e_P}^2)$, where $\mathbf{I}$ is an identity matrix and $\sigma_{e_P}^2$ is the residual variance.

A solution for the fixed and random effects is obtained by solving the mixed model equation according to Henderson (1977):

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\frac{\sigma_{\mathbf{e_P}}^2}{\sigma_{\mathbf{t}}^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Variances σ_t^2 and $\sigma_{e_P}^2$ were estimated using restricted maximum likelihood as implemented in ASReml 3.0. In contrast to Maize 2, where heritabilities were estimated in the phenotypic analysis across locations, trait heritabilities for Maize 1 were estimated based on the variance components derived from this PBLUP model using the standard formula:

$$\hat{h}^2 = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{\sigma}_{e_P}^2}. \tag{15}$$

2.5.2 Genome-based best linear unbiased prediction

In contrast to the PBLUP model, the variance-covariance structure of testcross effects can also be derived from genome-wide marker data as described in Section 2.4.2. The model is denoted as GBLUP and can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{16}$$

where the vectors $\mathbf{y}$, $\boldsymbol{\beta}$ and $\mathbf{e}$ and the design matrices $\mathbf{X}$ and $\mathbf{Z}$ are defined as in the PBLUP model with $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_{e_G})$ and $\sigma^2_{e_G}$ being the residual variance pertaining to the GBLUP model. The difference to PBLUP is the modeling of the random testcross effects $\mathbf{u}$, which are assumed to be normally distributed with $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{U}\sigma^2_u)$ in Maize 1 and $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{U_{AB}}\sigma^2_u)$ in Maize 2, where $\mathbf{U}$ and $\mathbf{U_{AB}}$ are $N \times N$ matrices of realized kinship coefficients based on marker data (Eqn. 12 and 13, respectively) and σ^2_u is the testcross variance pertaining to the GBLUP model.

Due to the "kernel trick", which is described in Schölkopf et al. (1998), the GBLUP model using the realized kinship based on the unscaled matrix of genotype scores ($\mathbf{W}$) is a transformation of the ridge regression BLUP (RRBLUP, Whittaker et al. (2000)), where the markers in $\mathbf{W}$ are fitted directly as BLUPs in the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{m} + \mathbf{e}, \tag{17}$$

with $\mathbf{m}$ being a vector of marker effects under the assumption that $\mathbf{m} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_m)$. The transformation of variance components and means is described in detail in the Appendix. In some cases, when the number of loci is smaller than the number of observations, it can be useful to apply RRBLUP to obtain marker effects directly. Based on GBLUP using the kinship coefficients $\mathbf{U}$ based on centered marker scores, marker effects can be calculated with the following transformation (Yang et al. 2011):

$$\hat{\mathbf{m}} = \mathbf{W}'\mathbf{U}^{-1}\hat{\mathbf{u}} / \sum_{m=1}^{M} 2p_m(1-p_m).$$

2.5.3 Pedigree- and genome-based best linear unbiased prediction

The expected and realized kinship coefficients can be combined in one model (Legarra et al. 2008). The testcross variance is decomposed into a component explained by the pedigree-based kinship coefficients and a component based on the marker data. Hence, the P+GBLUP model for testcross performance is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{t} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{18}$$

where the vectors $\mathbf{y}$, $\boldsymbol{\beta}$ and $\mathbf{e}$ and the design matrices $\mathbf{X}$ and $\mathbf{Z}$ are defined as in the PBLUP model with $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_{e_{PG}})$ and $\sigma^2_{e_{PG}}$ being the residual variance pertaining to the P+GBLUP model. Here, the vectors $\mathbf{t}$ and $\mathbf{u}$ comprise the random testcross values based on pedigree and marker data, respectively. Both vectors of testcross effects are assumed to be independent and normally distributed with $\mathbf{t} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}\sigma^2_{t_{PG}})$ and $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{U}\sigma^2_{u_{PG}})$, where $\mathbf{K}$ and $\mathbf{U}$ are the expected and realized kinship matrices from PBLUP and GBLUP, respectively.

2.6 Cross-validation

When only one data set is available to assess predictive abilities, cross-validation is a useful method to compare different models without the necessity of an additional validation sample. The basic concept of CV is to divide the data set into an estimation set (ES) for fitting the model and a test set (TS) for validating the estimated model parameters.

In this study, mean predictive abilities were constant for $4 < k \leq 20$, while the variability was lowest for $k = 5$ (Appendix, Figure A3). Therefore, 5-fold CV was applied to Maize 1 and Maize 2, which were randomly divided into five disjoint sets. Four sets were used for the ES and the remaining set as TS (CV-R, Figure 5). This procedure is repeated five times so that each subset was used once as TS, which ensures that each DH line was used once for validation. Assigning genotypes into five subsets was additionally repeated ten times to result in 50 CV runs. In Maize 1, average predictive abilities did not differ when the variance components were re-estimated in the ES or estimated in the data set and committed to the CV procedure, but the computational load was substantially

greater. Therefore, model parameters and variance components were estimated for both traits once in the complete Maize 1 data set ($N = 1377$) with ASReml 3.0 (Gilmour et al. 2009). For Maize 2, the breeding populations were less homogeneous compared to Maize 1 and variance components were re-estimated in each ES.

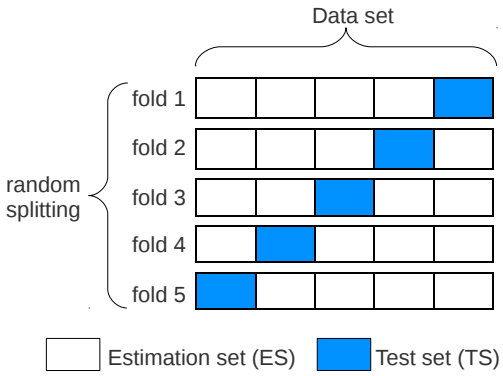


Figure 5: Scheme of 5-fold cross-validation with random sampling (CV-R) illustrating the five test sets (TS) and estimation sets (ES) of five runs within one replication.

Predictive abilities were calculated as the Pearson correlation coefficient between predicted and observed testcross performance or values of each TS. Based on vectors $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{t}}$, and $\hat{\mathbf{u}}$ estimated in the ES, predictive abilities of the different models were calculated in the corresponding TS as $r(\mathbf{y}_{TS}, \mathbf{X}_{TS}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{TS}\hat{\mathbf{t}})$ for PBLUP, $r(\mathbf{y}_{TS}, \mathbf{X}_{TS}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{TS}\hat{\mathbf{u}})$ for GBLUP and $r(\mathbf{y}_{TS}, \mathbf{X}_{TS}\hat{\boldsymbol{\beta}} + \mathbf{Z}_{TS}\hat{\mathbf{t}} + \mathbf{Z}_{TS}\hat{\mathbf{u}})$ for P+GBLUP. Here, the vector $\mathbf{y}_{TS}$ is a $N_{TS} \times 1$ vector of observations in the TS and $\mathbf{X}_{TS}$ is a $N_{TS} \times c$ and $\mathbf{Z}_{TS}$ is a $N_{TS} \times N$ design matrix for fixed and random effects, respectively. Due to the group structures in Maize 2, observed testcross performance was adjusted for the effect of its respective group/tester combination when the same combinations occurred in the ES and TS for CV-R. Hence, the predictive ability was $r(\mathbf{y}_{TS} - \mathbf{X}_{TS}\hat{\boldsymbol{\beta}}, \mathbf{Z}_{TS}\hat{\mathbf{t}})$ for PBLUP and $r(\mathbf{y}_{TS} - \mathbf{X}_{TS}\hat{\boldsymbol{\beta}}, \mathbf{Z}_{TS}\hat{\mathbf{u}})$ for GBLUP. The standard deviation of predictive abilities was calculated from means across folds within each replication.

The accuracy of a prediction model, which is the correlation between true and predicted testcross values, can be approximated by dividing the predictive ability by the square-root of the trait heritability h (Dekkers 2007; Legarra et al. 2008). However, the division by h does not influence the ranking of the different prediction models. Therefore, the focus in this study will be on predictive abilities for the comparison of models.

For Maize 1, another measure for the comparison of prediction models was applied because the selection of the top 10 % best DH lines within one testcross cycle is of interest for each plant breeder. Therefore, predicted testcross values of all five folds were merged within each replication and the best 10 % of DH lines were selected based on their pre-

dicted testcross performance. The observed testcross performance was averaged over the ten replications and compared between models.

The CV procedure described above was applied to Maize 1 and additionally to the four largest biparental families ($N = 58 - 60$) within this data set. In Maize 2, predictive abilities were assessed in both calibration sets and within genetic groups, testers or group/tester combinations of a given calibration set. The subset of group G3 crossed to tester T1 of CS1 and group G2 of CS2 were not analyzed because they comprised only 16 and 15 DH lines, respectively.

Effect of maturity on grain yield

Due to the genetic substructure in Maize 2, the impact of maturity on the prediction of grain yield within both calibration sets was determined. Using ES and TS sampled with 10×5-fold CV-R, predictive abilities for GDY were determined within each group/tester subset as the correlation between predicted testcross values of GDC and the observed testcross performance of GDY. Differences across group/tester subsets were visualized with elliptic contours. Each ellipse was shaped according to the 95 % confidence region of a bivariate normal distribution with mean and variance-covariance structure corresponding to the mean and variance-covariance matrix of the predicted testcross values and observed testcross performance within each group/tester combination.

Effect of sample size and marker density

To evaluate the effect of sample size on the accuracy of testcross prediction, the data set Maize 1 with $N = 1377$ DH lines was divided into 2, 4, and 8 subsets resulting in an array of subsets of size $N = 688$, 344, and 172, respectively. The procedure was repeated 16 times for $N = 688$, 8 times for $N = 344$, and 4 times for $N = 172$ to create 32 subsets for each sample size. Each of them was analyzed with PBLUP, GBLUP and P+GBLUP. Here, 10×5-fold CV with random sampling of the testcross progenies to the subsets was applied. In Maize 2, this procedure was applied to both calibration sets for comparing prediction within groups and tester subsets with a random sample of the complete data set. Therefore, CS1 with $N = 1073$ DH lines and CS2 with $N = 857$ DH lines were

divided into 2, 3, 4, 8, and 16 subsets of size $N = 536$, 358, 268, 134, and 67 for CS1 and $N = 428$, 286, 214, 107, and 54 for CS2, respectively. The sampling into subsets was repeated 24, 16, 12, 6, or 3 times to result in 48 subsets for each sample size.

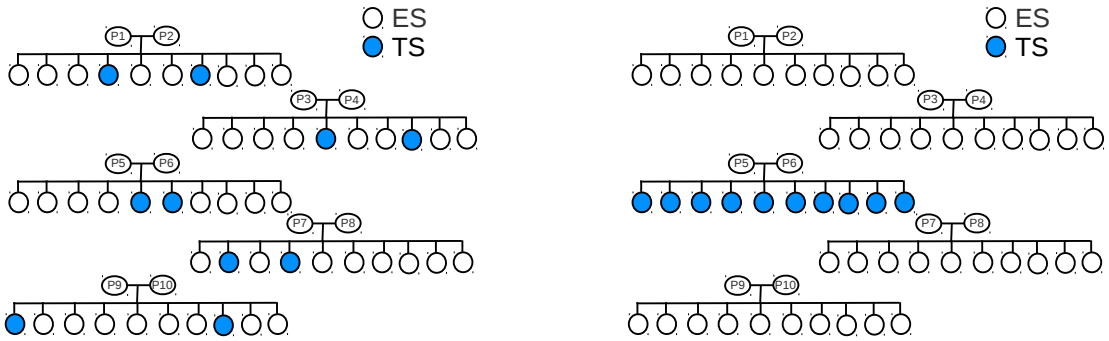
The effect of the marker density was evaluated within Maize 1. Here, two different marker arrays with high and low density were used to genotype a set of 759 DH lines.

2.6.1 Sampling within/across families

To account for the familiy structure in the data set, different sampling strategies can be used instead of random sampling within a data set. This implies sampling where the family structure of the complete data set is captured in each of the five subsets, i.e., stratified cross-validation (Kohavi 1995), or subsets reflect only a part of the family structure observed in the complete data set. These two sampling schemes, i.e., within family sampling (CV-W) and across family sampling (CV-A), are illustrated in Figure 6 (Legarra et al. 2008). With Maize 1, CV-W was performed for the entire data set (N =1377) and testcross progenies of each family were subdivided into five subsets. Four of them were assigned to the ES (N_{ES}=1093-1113), one to the TS. With CV-A, the 36 families of Maize 1 ($N = 1377$) were divided into four subsets of seven and one subset of eight families. Thus, the ES comprised 28 or 29 families, the TS contained the remaining families not included in the ES. Because the size of the families varied, the sample size of the ES also varied from N_{ES}=1002 to N_{ES}=1172. Due to the low number of progenies in each cross for both calibration sets in Maize 2, random sampling was used for CV instead of within and across family sampling.

2.6.2 Prediction across groups and testers

For Maize 2, the CV schemes were modified to assess predictive abilities across genetic groups and testers (CV-aG) conditional on a given TS. Here, the same TS as sampled from CV-R within each group or tester subset were used. The ES were sampled either from all remaining DH lines of a given calibration set ($ES_{G1,G2,G3}$) or from DH lines not belonging to the same group or tester represented in the TS. Sampling conditional on a TS from G3 in CS1 ($N_{TS} = 49$) is illustrated in Figure 7. Here, $ES_{G1,G2,G3}$ comprised



(a) Within family sampling

(b) Across family sampling

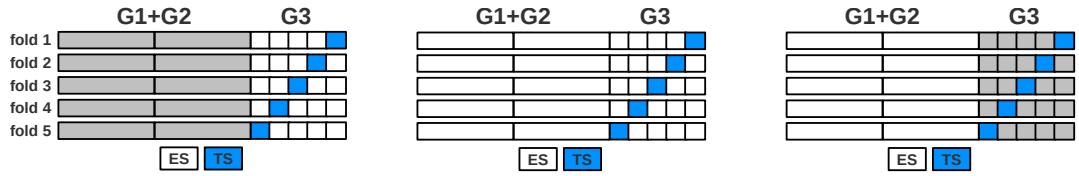
Figure 6: Example of stratified cross-validation with different estimation (ES) and test sets (TS) for sampling within (a) and across (b) five families derived from parental lines P1-P10.

the remaining 1024 DH lines from CS1, while $ES_{G1,G2}$ comprised only DH lines from the groups G1 and G2 ($N_{ES} = 827$). As groups G1 and G3 in CS2 were both crossed to testers T1 and T3, predictions were performed within and across group/tester subsets (further denoted as G1/T1, G1/T3, G3/T1, and G3/T3). Here, seven possible ES existed comprising the three ES described before, one ES comprising the remaining lines from the complete group, and three additional ES including the remaining group/tester subsets not belonging to the TS to assess predictive abilities within groups across tester and across groups within the same tester subset. When different group/tester combinations occurred in the ES and TS, a correction for the fixed group/tester effect was not possible and predictive abilities were calculated based on the unadjusted observed testcross performance $\mathbf{y}_{TS}$.

To account for the difference in sample size of the resulting ES for the prediction of group G2 in CS1, the size of the ES for the prediction across groups was kept constant at $N_{ES} = 116$. Sampling within each ES of the 50 CV runs was repeated 20 times resulting in 1000 ES. For the prediction across group/tester combinations in CS2, the size of the ES was reduced to $N_{ES} = 98$ which was sampled once from each of the 50 ES.

2.6.3 Effect of decreasing number of locations and sample size

The influence of different sets of locations on the predictive ability and the prediction of testcross values across locations was analyzed based on the subset of tester T1 from CS1



(a) CV-R: prediction of TS_{G3} with ES_{G3}

(b) CV-aG: prediction of TS_{G3} with $ES_{G1,G2,G3}$

(c) CV-aG: prediction of TS_{G3} with $ES_{G1,G2}$

Figure 7: Cross-validation schemes for the prediction within (CV-R) and across genetic groups (CV-aG). Scheme (a) illustrates the five test sets (TS_{G3}) for sampling within group G3 with the corresponding ES_{G3} and scheme (b) and (c) illustrate the two ES conditional on TS_{G3} containing all remaining DH lines of the calibration set and DH lines belonging to groups G1 and G2, respectively.

of Maize 2. All DH lines of this tester subset ($N = 698$) were evaluated in the same four locations and only negligible genetic substructure was expected as nearly all DH lines belong to group G1.

To analyze the influence of the number of locations, three, two, and one location out of four were selected and submitted to the second stage of phenotypic analysis to obtain adjusted means across locations for GBLUP (Schön et al. 2004). In each sampling step of locations, also subsets of the data set were selected to assess additionally the influence of the reference set size on predictive abilities. Therefore, the data set was divided into 2, 4, and 8 equally sized subsets containing $N = 349$, 175, and 87 DH lines, respectively. Prediction performance of these subsets was compared to prediction within the complete data set, as well as the evaluation within all four locations. A summary of the number of locations, DH lines and randomizations for each sampling step is given in Table 2.

2.6.4 Prediction across locations

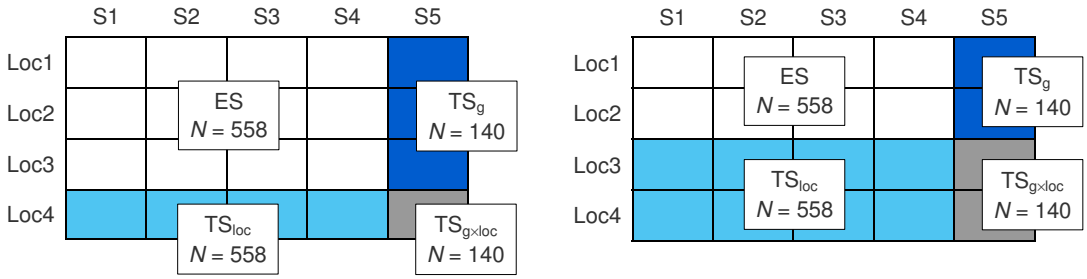
Prediction across locations was assessed to analyze the influence of genotype by location interactions. Therefore, the tester T1 data set was divided into five genotypic subsets (S) of CV-R each evaluated in the same four locations, resulting in 20 disconnected subsets (Figure 8). Assigning genotypes to subsets was repeated ten times as described for 10×5-fold CV-R.

The ES comprised adjusted means over two or three locations obtained from the phenotypic analysis. For each set of locations, heritabilities for both traits were assessed. The

Table 2: Number of location subsets (L) × number of subsets for a specific sample size (N) × number of randomizations for each possible location and sample size subsets of tester T1 in calibration set CS1 of Maize 2. Number of resulting data sets for each sampling step are in parenthesis.

	N			
L	698	349	175	87
4	$1 \times 1 \times 1$ (1)	$1 \times 2 \times 4$ (8)	$1 \times 4 \times 2$ (8)	$1 \times 8 \times 1$ (8)
3	$4 \times 1 \times 1$ (4)	$4 \times 2 \times 4$ (32)	$4 \times 4 \times 2$ (32)	$4 \times 8 \times 1$ (32)
2	$6 \times 1 \times 1$ (6)	$6 \times 2 \times 4$ (48)	$6 \times 4 \times 2$ (48)	$6 \times 8 \times 1$ (48)
1	$4 \times 1 \times 1$ (4)	$4 \times 2 \times 4$ (32)	$4 \times 4 \times 2$ (32)	$4 \times 8 \times 1$ (32)

phenotypic correlation was calculated between adjusted means obtained from locations in the ES and adjusted means obtained in each remaining location. Adjusted means obtained from each single locations were further analyzed with a cluster analysis based on the distance between locations according to Ouyang et al. (1995). Finally, each ES comprised three locations and four genotypic subsets resulting in $4 \times 10 \times 5 = 200$ possible ES and two locations and four genotypic subsets in $6 \times 10 \times 5 = 300$ ES. Following Utz et al. (2000), three different test sets (TS) were defined for each ES taking genotypic sampling (TS_g), sampling of locations (TS_{loc}) and both factors simultaneously ($TS_{g \times loc}$) into account. With TS_g, independent genotypes are predicted with the information from related DH lines in the same set of locations. In TS_{loc}, the same genotypes are predicted in an independent location. Both factors are combined in $TS_{g \times loc}$, where genotypes and locations are independent from the estimation set.



(a) Three locations **(b)** Two locations

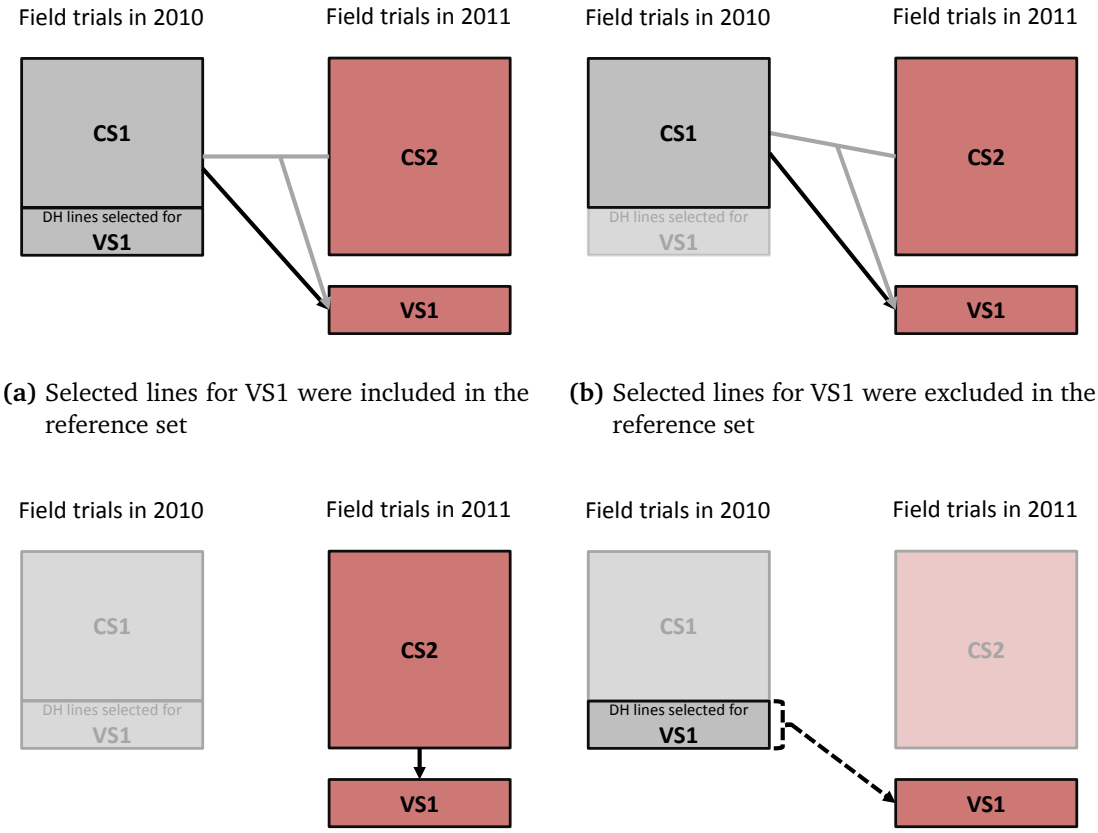
Figure 8: Sampling scheme for cross-validation with five genotypic subsets and four locations according to Utz et al. (2000). Colors are indicating possible ES including (a) three or (b) two locations and different TS for the predictions of independent genotypes (TS_g, dark blue), locations (TS_{loc}, light blue) or genotypes and locations ($TS_{g\times loc}$, grey).

2.6.5 Prediction across years

In Maize 2, two calibration sets evaluated in different years were available and predictions within each calibration set could be additionally validated by predicting testcross performance with the other calibration set. To investigate predictive abilities of GBLUP across years, data from both calibration sets were analyzed jointly ($N = 1930$) with a realized kinship matrix $\mathbf{U_{AB}}$ calculated from $M = 17734$ SNPs. The vector of fixed effects $\boldsymbol{\beta}$ in the joint model included $c = 9$ factor levels for each year/group/tester combination. To predict testcross performance of DH lines from CS2, phenotypic observations from CS2 were masked and predicted from data in CS1 and vice versa. Predictive abilities were measured as the correlation of predicted testcross values and observed testcross performance across all DH lines in each calibration set and for DH lines within each group/tester combination separately. To compare predictive abilities across locations and years, validation across years was also performed for the data subset of tester T1, which included DH lines from group G1 and G3 in both calibration sets.

Prediction across years was additionally assessed with the DH lines of CS1 which were selected in 2010 and evaluated again in 2011 as VS1. Therefore, different reference sets were used to predict the testcross values of the DH lines selected from CS1. The data set used for model training included observations from CS1, the observations from CS1 and CS2 (Figure 9a and 9b), or observations from CS2 (Figure 9c). For prediction of testcross performance of VS1 in 2011, the DH lines selected from CS1 could be in-

cluded (Figure 9a) or excluded from CS1 (Figure 9b). To compare the selection based on predicted testcross values with the selection based on observed testcross performance, correlations were calculated between observed testcross performance from 2010 and 2011 for DH lines included in VS1 (Figure 9d).



(a) Selected lines for VS1 were included in the reference set

(b) Selected lines for VS1 were excluded in the reference set

(c) CS2 in the reference set

(d) Correlation between years

Figure 9: Possible reference sets for predicting DH lines in the validation set VS1. The reference set included DH lines from calibration set CS1 (black arrow) or CS1 and CS2 (grey arrow), where the DH lines selected for VS1 were (a) included or (b) excluded from the reference set or (c) only observation from CS2. Scheme (d) illustrates the correlation between observations from 2010 and 2011 of DH lines in the VS1.

3 Results

3.1 Phenotypic analysis

3.1.1 Maize 1

Adjusted testcross values of the 1377 DH lines for GDY averaged across the seven locations ranged from 105.69 to 175.98 dt/ha and for GDC from 78.18 to 84.52 %. Family means varied for GDY from 132.38 to 159.52 dt/ha and for GDC from 80.34 to 83.09 %. Phenotypic correlations between the seven locations calculated from the testcross performance of the DH lines varied between 0.19 and 0.38 for GDY and between 0.32 and 0.64 for GDC. Estimates of the heritability based on variance components estimated with PBLUP were $\hat{h}^2_{GDY} = 0.84$ and $\hat{h}^2_{GDC} = 0.85$ (Appendix, Table A4).

3.1.2 Maize 2

Observed testcross performance of the DH lines for GDY and GDC are given for the two calibration sets and for each group/tester combination within CS1 and CS2 in Appendix Table A5. In CS1, adjusted means for GDY from phenotypic analysis ranged between 95.13 and 148.20 dt/ha with a mean of 126.71 dt/ha. Climatic conditions in 2011 were more favorable for maize production than in 2010 and GDY and GDC in CS2 were significantly ($p < 0.01$) higher than in CS1 with a range of 108.10 to 165.30 dt/ha and a mean of 144.31 dt/ha for GDY and a mean of 71.81 % for GDC (range 66.12-76.70 %. For the DH lines in the VS1 evaluated in 2011, trait values ranged between 126.77 and 157.58 dt/ha with a mean of 144.47 dt/ha for GDY and between 68.07 and 74.42 % with a mean of 71.44 % for GDC. In both calibration sets, phenotypic means differed significantly between the three genetic groups (G1-G3) and between groups of DH lines crossed to different testers. In CS2, these differences were reduced between group G1 and G3 for GDY and between group G2 and G3 for GDC.

Phenotypic correlations between GDY and GDC were -0.25 and -0.27 and genetic correlations were -0.45 and -0.37 in CS1 and CS2, respectively. In both calibration sets, genotypic and genotype by location interaction variances were highly significant for

both traits ($p < 0.01$). The resulting trait heritabilities on a progeny mean basis were $\hat{h}^2_{GDY} = 0.72$ and $\hat{h}^2_{GDC} = 0.94$ for CS1, $\hat{h}^2_{GDY} = 0.71$, $\hat{h}^2_{GDC} = 0.95$ for CS2 (Appendix, Table A6) and $\hat{h}^2_{GDY} = 0.77$ and $\hat{h}^2_{GDC} = 0.90$ for VS1.

3.2 Genotypic analysis

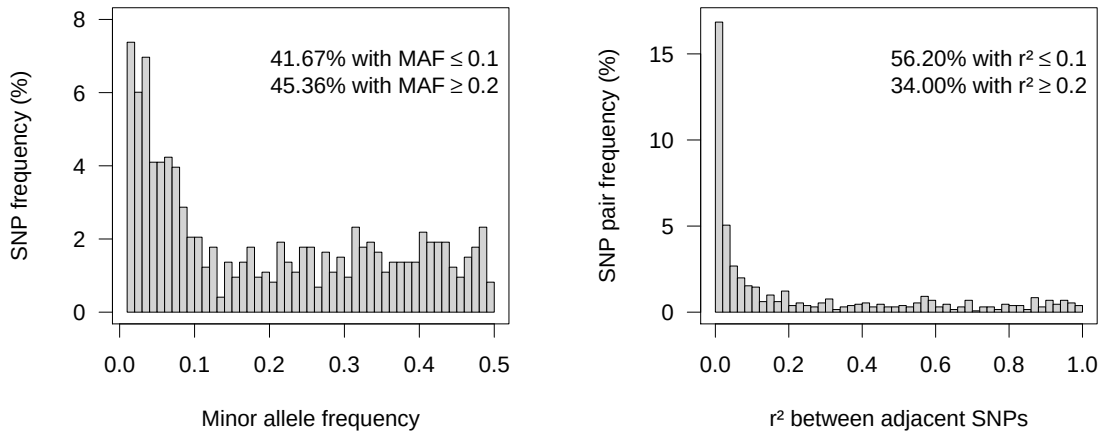
3.2.1 Maize 1

From the set of 732 polymorphic SNP markers 663 could be assigned to linkage groups, representing the ten chromosomes of maize. For sixty-nine markers the chromosomal position was unknown. The markers were evenly distributed across the genome with an average distance of 2.9 Mb. The number of SNPs per linkage group ranged from 48 on chromosome 10 to 96 on chromosome 1. The average MAF across markers was 0.19. The largest proportion of markers was observed for low MAF (Figure 10a), where 42 % of the markers had a MAF < 0.1. The number of polymorphic SNPs within the 36 crosses ranged from 78 to 600. Figure 10b illustrates the LD between adjacent SNPs. Only significant LD ($p < 0.05$) was included and average LD between adjacent SNPs was 0.23. As expected for an advanced cycle breeding population substantial long-range LD was detected (see Albrecht et al. (2011)).

For 18791 out of 20742 polymorphic SNPs, which were used for genotyping the subset of 759 DH lines, the map positions along the maize genome were known. The distribution of these SNPs is illustrated in Figure 11. The number of SNPs on each chromosome ranged between 1233 on chromosome 6 to 2884 on chromosome 1. Average distance between SNPs was 0.11 Mb and average LD between adjacent SNPs was high with 0.57.

3.2.2 Maize 2

For both calibration sets together, 17431 SNPs could be assigned to the physical map. The distance between neighboring markers along the genome was on average 0.12 Mb and ranged between 0 and 12.23 Mb. The number of SNPs on each chromosome ranged from 1014 on chromosome 10 to 2707 on chromosome 1.



(a) Minor allele frequency **(b)** Linkage disequilibrium

Figure 10: Histogram of (a) minor allele frequency (MAF) and (b) linkage disequilibrium (r^2) between adjacent markers ($M = 732$).

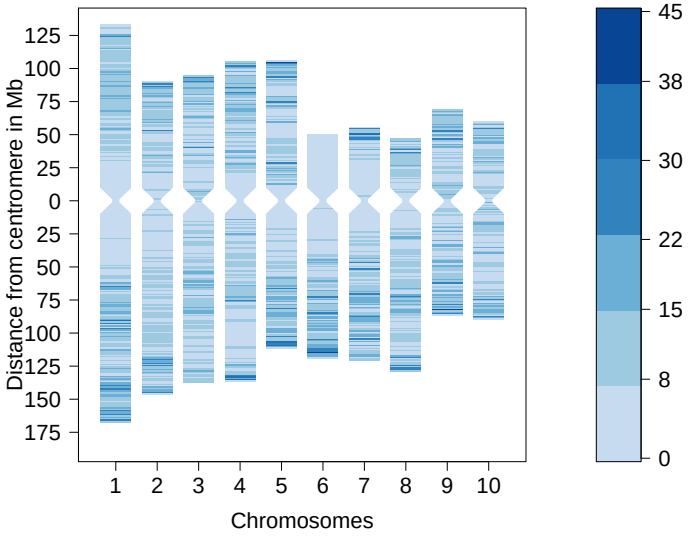
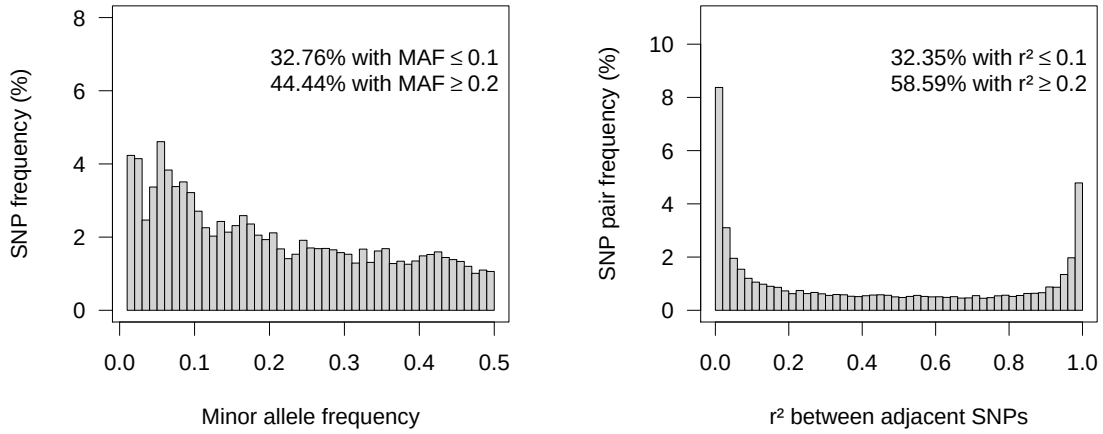


Figure 11: Distribution of $M = 18791$ mapped SNPs along chromosomes in the maize genome. Color-scale indicates the number of SNPs within 1 Mb.

For each calibration set, allele frequencies and LD were calculated for polymorphic SNPs and visualized in Figure 12 and 13, respectively. In CS1, the proportion of SNPs with MAF $\leq$ 0.1 was lower compared to CS2 leading to a significant decrease in mean MAF from 0.20 in CS1 to 0.19 in CS2 ($p < 0.01$). In addition, LD between adjacent SNPs was higher for DH lines from CS1 with a mean of 0.42 compared to DH lines from CS2 with a mean of 0.35. The percentage of adjacent SNPs with $r^2 \geq 0.2$ was 59 % in CS1

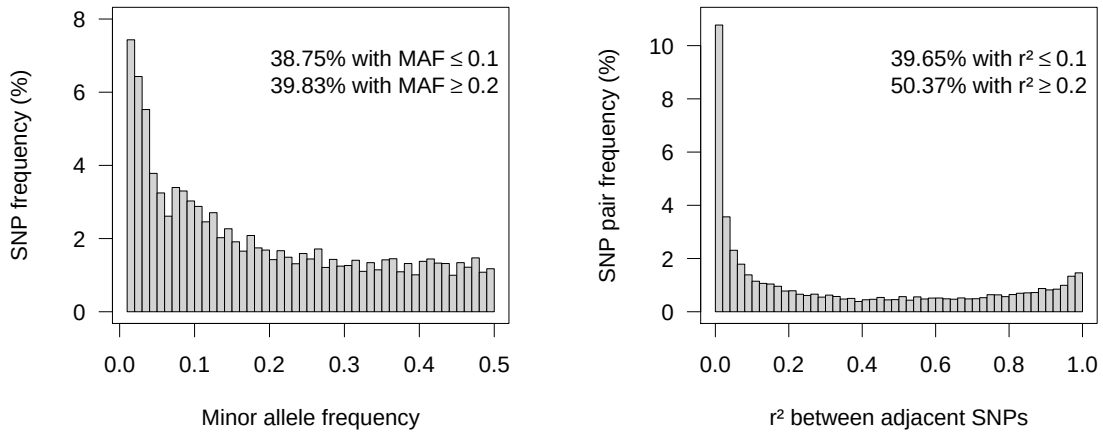
and 50 % in CS2. The observed differences in MAF and LD between the two calibration sets confirm differences in the distinct family structure in CS1 and CS2.



(a) Minor allele frequency

(b) Linkage disequilibrium

Figure 12: Histogram of (a) minor allele frequency (MAF) and (b) linkage disequilibrium (r^2) between adjacent markers ($M = 15732$) in calibration set CS1.



(a) Minor allele frequency

(b) Linkage disequilibrium

Figure 13: Histogram of (a) minor allele frequency (MAF) and (b) linkage disequilibrium (r^2) between adjacent markers ($M = 16846$) in calibration set CS2.

3.3 Substructure of genetic material in Maize 2

Table 3 shows the results from the kinship analysis according to Saatchi et al. (2011) for the DH lines from the two calibration sets CS1 and CS2 based on the expected (k_{ij}) and realized kinship coefficients (u^s_{ij}). Mean kinship coefficients within groups were different compared to the complete calibration set. Due to larger family sizes and a smaller number of parents, mean expected kinship coefficients in CS1 were higher than in CS2. In most cases, mean maximum kinship within groups was close to or exceeded 0.5 reflecting the high relatedness of DH lines derived from crosses of closely related parents. Maximum kinship within groups was always significantly higher than between groups confirming the prior assumption of genetic substructure (Appendix Figure A4). In CS2, the difference in mean maximum kinship within and between groups decreased compared to CS1.

Analogously as for the three genetic groups, the level of kinship can be analyzed for groups of DH lines crossed to the three different testers (Table 3). In CS1, the subset of tester T1 contained mainly DH lines from group G1, while the subset of tester T2 contained only lines from G2 and G3 (Table 1a). Thus, the kinship within and between groups of DH lines crossed to T1 and T2 reflected the kinship observed for the three genetic groups. In CS2, lines from G1 and G3 were crossed to both testers. However, average relatedness of DH lines crossed to T3 was higher than for lines crossed with T1, indicating a non-random assignment of crosses and DH lines to testers (see also Appendix Figure A5). This can also be inferred from the low number of parents connecting subsets of lines crossed with T1 and T3 (Figure 4d).

The different levels of mean kinship coefficients across groups can also be verified based on the realized kinship coefficients. Here, maximum realized kinship coefficients $\overline{u}^s_{max}$ within groups or tester subsets were also higher than between subsets. Lowest differences for expected and realized kinship occurred in both calibration sets in group G3.

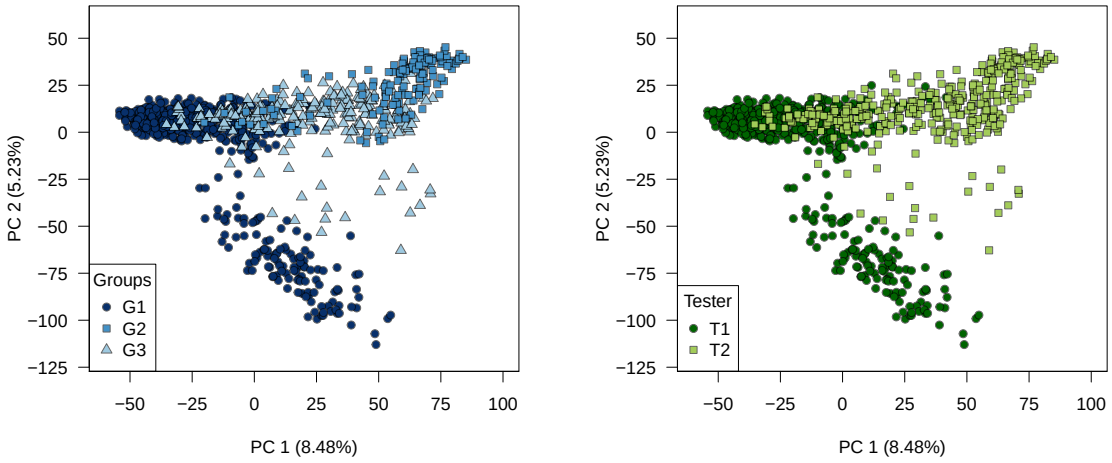
Table 3: Number of DH lines (N), number of markers (M), and mean ($\overline{k}$, $\overline{u}^s$) and maximum ($\overline{k}_{max}$, $\overline{u}^s_{max}$) expected and realized kinship within and between subsets for the complete data set, groups and tester subsets for calibration set CS1 and CS2. Realized kinship coefficients were estimated with Eqn.11 based on the simple matching coefficient of all DH lines in Maize 2.

		Group			Tester	
	CS1	G1	G2	G3	T1	T2
N	1073	682	145	246	698	375
No. crosses	192	118	19	55	122	70
M	15732	11038	4404	9703	11646	10394
$\overline{k}$	0.226 ± 0.121^a	0.298 ± 0.135	0.278 ± 0.141	0.220 ± 0.090	0.293 ± 0.135	0.192 ± 0.103
$\overline{k}_{max}$ within	0.566 ± 0.098^b	0.599 ± 0.102	0.550 ± 0.013	0.481 ± 0.052	0.597 ± 0.102	0.506 ± 0.054
$\overline{k}_{max}$ between	0.425 ± 0.103^b	0.335 ± 0.099	0.291 ± 0.088	0.323 ± 0.097	0.354 ± 0.101	0.306 ± 0.120
$\overline{u}^s$	0.455 ± 0.122^a	0.525 ± 0.124	0.550 ± 0.117	0.429 ± 0.095	0.521 ± 0.124	0.428 ± 0.107
$\overline{u}^s_{max}$ within	0.795 ± 0.076^b	0.812 ± 0.070	0.818 ± 0.059	0.727 ± 0.071	0.811 ± 0.070	0.762 ± 0.078
$\overline{u}^s_{max}$ between	0.700 ± 0.092^b	0.621 ± 0.095	0.581 ± 0.094	0.590 ± 0.105	0.641 ± 0.091	0.581 ± 0.104
	CS2	G1	G2	G3	T1	T3
N	857	582	15	260	327	530
No. crosses	294	181	7	106	141	153
M	16846	13001	727	14344	14052	9543
$\overline{k}$	0.207 ± 0.108^a	0.252 ± 0.118	0.279 ± 0.153	0.178 ± 0.079	0.156 ± 0.094	0.274 ± 0.117
$\overline{k}_{max}$ within	0.535 ± 0.104^b	0.576 ± 0.096	0.527 ± 0.044	0.440 ± 0.059	0.481 ± 0.083	0.568 ± 0.102
$\overline{k}_{max}$ between	0.406 ± 0.122^b	0.343 ± 0.091	0.376 ± 0.047	0.301 ± 0.112	0.286 ± 0.074	0.324 ± 0.066
$\overline{u}^s$	0.426 ± 0.118^a	0.472 ± 0.117	0.512 ± 0.114	0.375 ± 0.103	0.366 ± 0.107	0.482 ± 0.120
$\overline{u}^s_{max}$ within	0.750 ± 0.108^b	0.784 ± 0.104	0.726 ± 0.060	0.657 ± 0.080	0.680 ± 0.100	0.790 ± 0.095
$\overline{u}^s_{max}$ between	0.672 ± 0.110^b	0.613 ± 0.098	0.634 ± 0.072	0.552 ± 0.117	0.590 ± 0.090	0.633 ± 0.084

[a]: standard deviation attached; [b]: within and across calibration sets

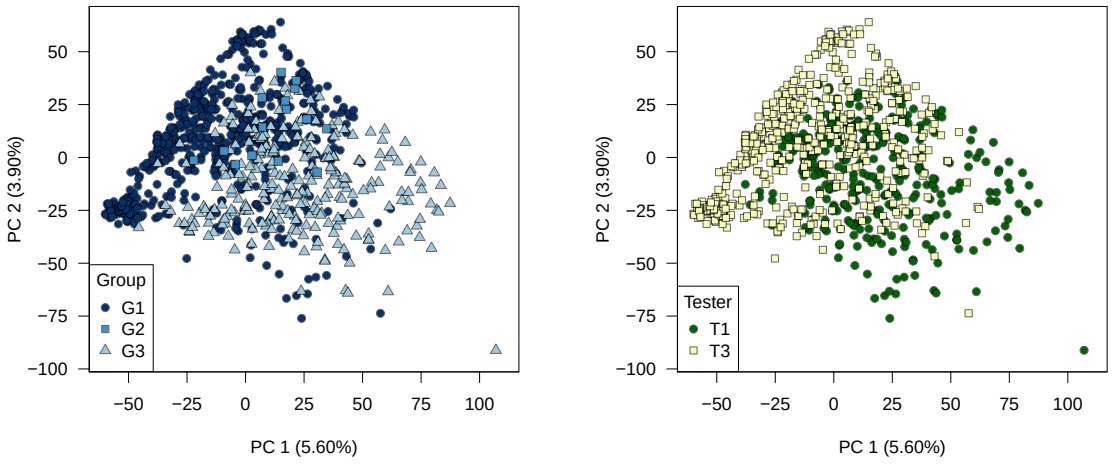
Principal component analysis

The different extent of genetic substructure in both calibration sets CS1 and CS2 is illustrated based on the first two PCs of the genome-wide marker data (Figure 14 and 15). In CS1, there was a clear separation of the predefined groups G1, G2 and G3 along the first and second PC, which explained 8.5 and 5.2 % of the total variation in marker data, respectively (Figure 14a). This separation was also represented between tester subsets. In CS2, the three groups were less separated than in CS1 within the space of the first two PCs, which explained only 5.6 and 3.9 % of the total variation (Figure 15a).



(a) Groups **(b)** Testers

Figure 14: First two principal components of the marker data from calibration set CS1 colored according to (a) groups and (b) testers.



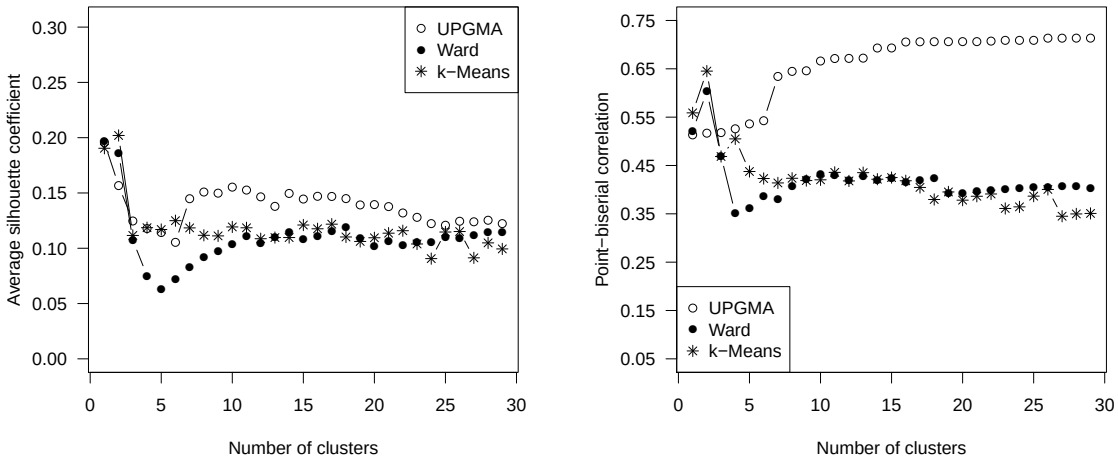
(a) Groups **(b)** Testers

Figure 15: First two principal components of the marker data from calibration set CS2 colored according to (a) groups and (b) testers.

Cluster analysis

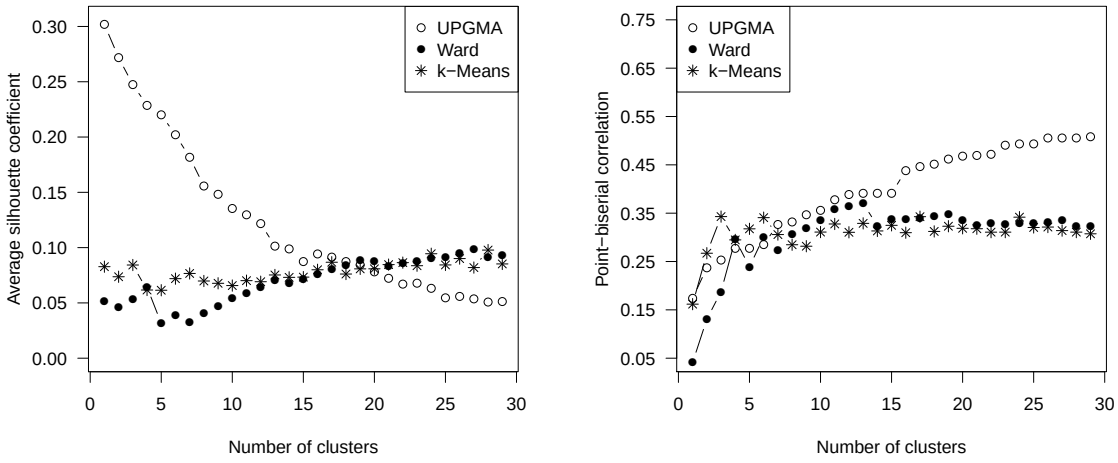
The results from the cluster analysis are illustrated in Figure 16, 17, and 18. Based on the maximum ASC, the optimum number of clusters in CS1 was two for UPGMA and Ward's clustering method, but three for k-means clustering, which is in accordance with the number of predefined genetic groups in CS1. With PBC, the maximum correlation was reached with three clusters for Ward and k-means, while for UPGMA, the optimum number of clusters was much higher than for the other clustering methods, which was also observed by Odong et al. (2011). In CS2, the optimum number of clusters varied strongly across clustering methods irrespective of the coefficient applied indicating that the optimum number of clusters could not be determined. These results confirm the marginal separation of the genetic groups in the space of the first two PCs.

In Figure 18, two and three clusters obtained with UPGMA, Ward's, and k-means clustering are illustrated within the space of the first two PCs. For the first two clusters, the data set was divided into similar subsets along the first PC with Ward's method and UPGMA clustering. The first two clusters were also in accordance with the *a priori* known groups (Figure 14), where the first cluster represented group G1 which is the core germplasm of the breeding population. The other two groups G2 and G3 clustered into the same cluster. The cluster subsets differed when more than three clusters were considered, where the third cluster based on UPGMA included only three DH lines. With Ward's method, the first three clusters split the data set along the first and second PC. Subsets derived from k-means clustering are illustrated in Figure 18c. The subsets for three clusters were similar to these derived from Ward's method as both methods minimize the squared sums between observations and the mean within a cluster.



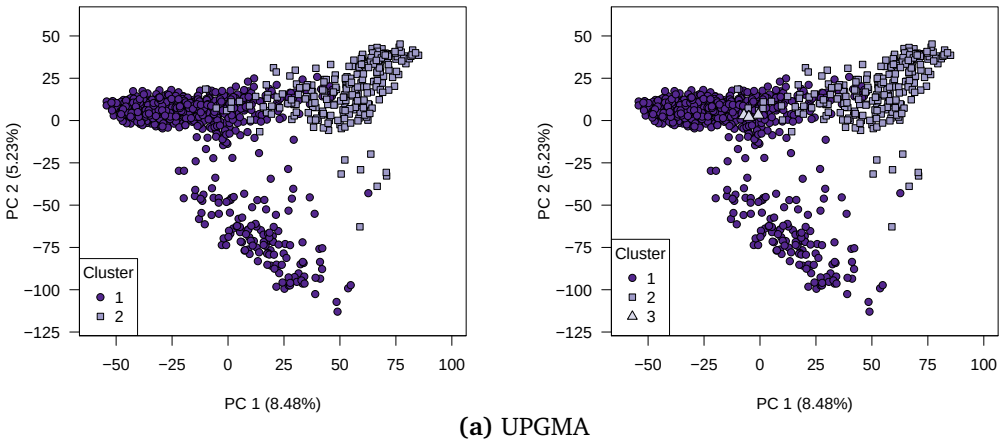
(a) Average silhouette coefficient **(b)** Point-biserial correlation

Figure 16: Optimum number of clusters obtained with (a) average silhouette coefficient and (b) point-biserial correlation for an increasing number of clusters in calibration set CS1.

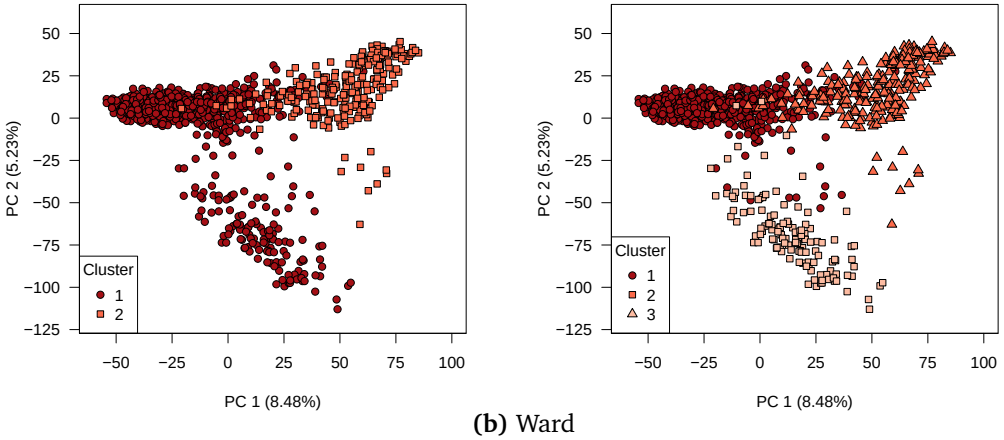


(a) Average silhouette coefficient **(b)** Point-biserial correlation

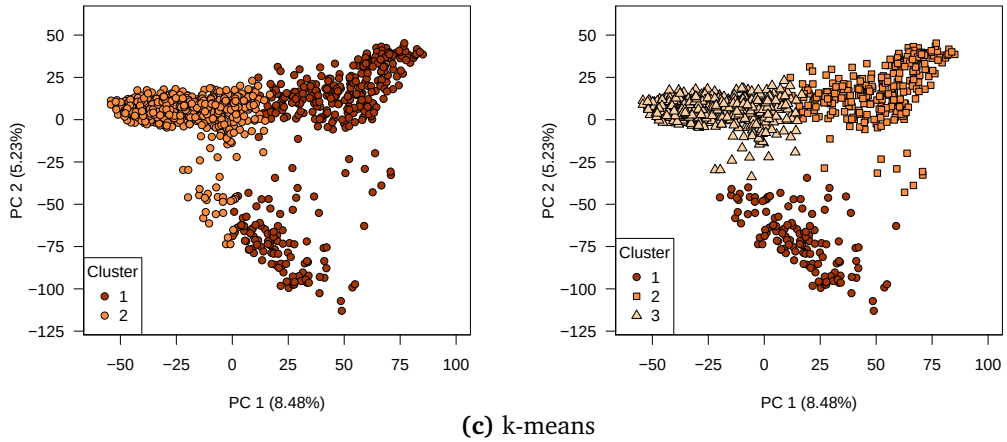
Figure 17: Optimum number of clusters obtained with (a) average silhouette coefficient and (b) point-biserial correlation for an increasing number of clusters in calibration set CS2.



(a) UPGMA



(b) Ward



(c) k-means

Figure 18: Two and three clusters obtained with (a) UPGMA, (b) Ward, and (c) k-means cluster analysis plotted within the space of the first two principal components of the marker data from calibration set CS1.

3.4 Predictive abilities obtained with different cross-validation schemes

3.4.1 Within/across family prediction

Mean predictive abilities of Maize 1 obtained with 10 × 5-fold CV-W, i.e., within family sampling, and CV-A, i.e., across family sampling, are shown for PBLUP, GBLUP and P+GBLUP for both traits in Table 4. For GDY and CV-W, the mean predictive ability of GBLUP and P+GBLUP over the 50 cross-validation runs was $r_{TS} = 0.66$ and 0.68, respectively, which was substantially higher than for PBLUP. P+GBLUP performed better than GBLUP but predicted testcross values were highly correlated ($r = 0.90$). Mean predictive abilities of CV-A were markedly reduced and much more variable compared to CV-W. For GDY, the mean predictive ability for PBLUP was reduced to $r_{TS} = 0.11$ and for GBLUP to $r_{TS} = 0.44$. Mean predictive abilities for GDC were generally higher than for GDY in both CV schemes.

Comparing the models on the basis of mean phenotypic performance of 10 % best lines selected based on their predicted testcross performance, resulted in the same ranking of models as obtained from comparing predictive abilities. Lines selected based on predictions from GBLUP and P+GBLUP performed substantially better than those selected based on PBLUP with both CV schemes, but P+GBLUP did not show an advantage over GBLUP.

Within each of the four large biparental families in Maize 1, the number of polymorphic markers varied from 116 to 212 (Figure 19). Mean predictive abilities differed between families with a range of $r_{TS} = 0.24$ to 0.54 for GDY and from $r_{TS} = 0.43$ to 0.78 for GDC and were highly variable across CV runs. Except for GDC in biparental family 3, predictive abilities were smaller than those obtained with the complete data set.

Table 4: Mean predictive abilities and observed testcross performance of 10 % best predicted DH lines with their standard deviations derived from 10 × 5-fold cross-validation with sampling within (CV-W) and across (CV-A) families for PBLUP, GBLUP and P+GBLUP for traits grain yield (GDY) and grain dry matter content (GDC) estimated in Maize 1 ($N = 1377$ and $M = 732$).

Model	Predictive abilities		10 % best predicted	
	GDY	GDC	GDY	GDC
CV-W				
PBLUP	0.509 ± 0.004	0.498 ± 0.004	157.40 ± 0.17	82.41 ± 0.01
GBLUP	0.664 ± 0.006	0.719 ± 0.003	159.00 ± 0.21	82.70 ± 0.02
P+GBLUP	0.679 ± 0.006	0.724 ± 0.003	159.04 ± 0.16	82.70 ± 0.02
CV-A				
PBLUP	0.113 ± 0.114	0.308 ± 0.048	154.79 ± 1.70	81.87 ± 0.28
GBLUP	0.440 ± 0.035	0.594 ± 0.037	157.30 ± 0.37	82.57 ± 0.05
P+GBLUP	0.426 ± 0.048	0.593 ± 0.029	156.90 ± 0.67	82.55 ± 0.05

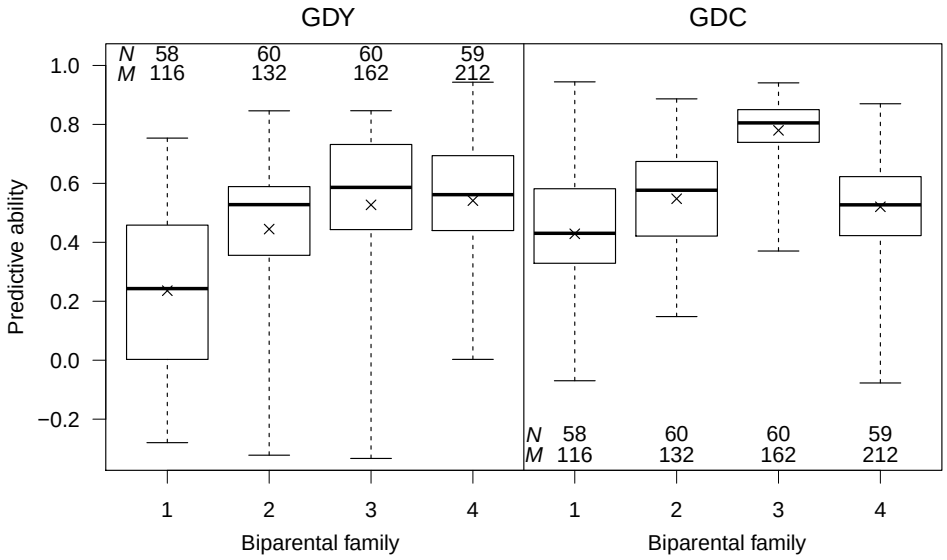
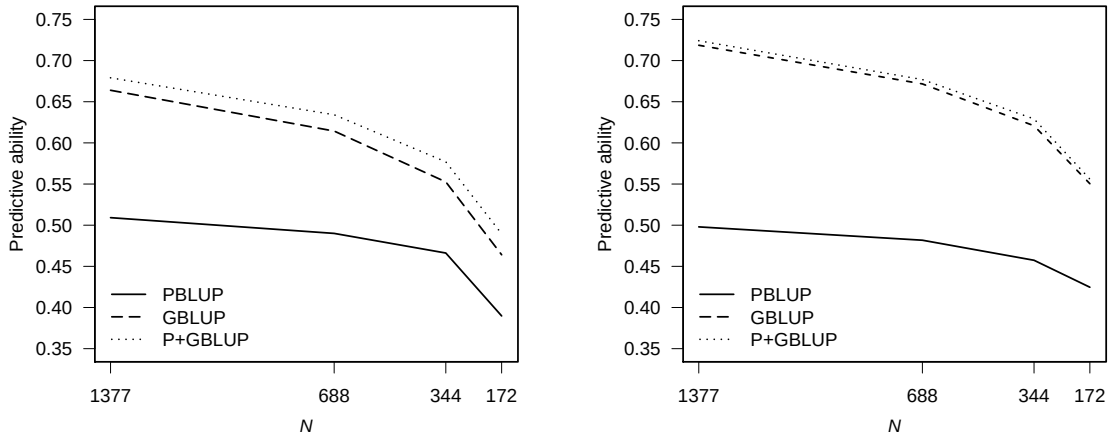


Figure 19: Predictive abilities within four biparental families (BF1-4) obtained from 10×5-fold cross-validation with random sampling within each family for the traits grain dry matter yield (GDY) and grain dry matter content (GDC). The boxplots show the range, median (bar) and mean (×) of 50 CV runs for both traits grain dry matter yield (GDY) and grain dry matter content (GDC). Numbers above and below boxplots indicate number of DH lines (N) and number of polymorphic markers (M) in each family.

3.4.2 Influence of sample size and marker density on predictions

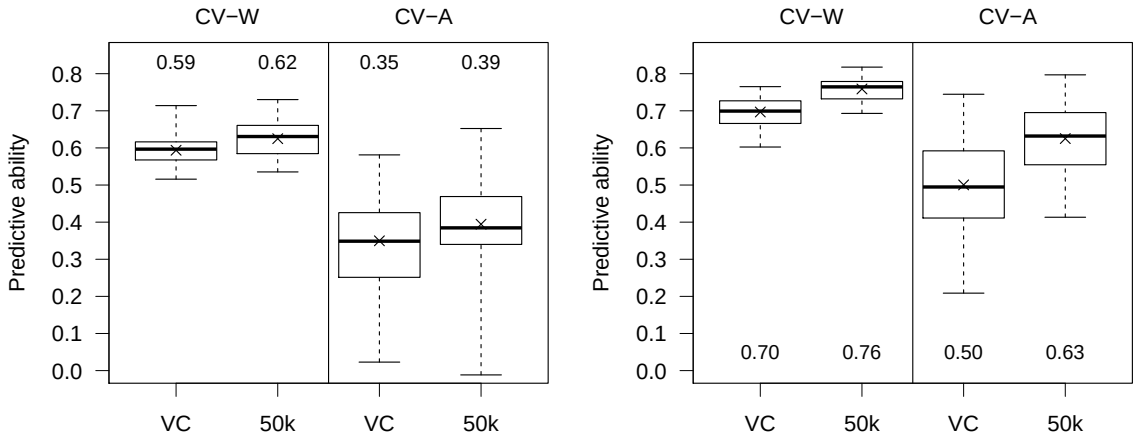
The connection between sample size of the reference population and predictive ability is illustrated in Figure 20 for Maize 1. The size of the data set had a strong effect on the predictive ability of all models for predicting testcross performance for GDY and GDC. The reduction in predictive ability was more pronounced for GBLUP and P+GBLUP than for PBLUP but performance of the models with genomic information was still better. A reduction of the data set from 1377 to 688 or 344 DH lines led only to a slight decrease in predictive abilities. A further reduction from 344 to 172 DH lines led to strong decrease in predictive abilities for all models and both traits.

An increase of the marker density from 654 to 20742 SNPs for the subset of $N = 759$ DH lines of Maize 1 led to substantially higher predictive abilities for both traits (Figure 21). For GDY, predictive abilities increased from 0.59 to 0.62 with CV-W and from 0.35 to 0.39 for CV-A. Differences in predictive abilities between CV-W and CV-A were reduced with higher marker densities. But, predictive abilities for GBLUP with high marker densities were lower than for the complete data of Maize 1 with $N = 1377$ DH lines and $M = 732$ SNPs.



(a) Grain dry matter yield **(b)** Grain dry matter content

Figure 20: Average predictive ability of PBLUP, GBLUP and P+GBLUP with decreasing sample size for (a) grain dry matter yield and (b) grain dry matter content measured in Maize 1 ($M = 732$).



(a) Grain dry matter yield **(b)** Grain dry matter content

Figure 21: Comparison of predictive abilities of GBLUP with different SNP densities, VeraCode (VC) with $M = 654$ and 50k SNP chip with $M = 20742$ SNPs, determined from 10×5-fold cross-validation with sampling within (CV-W) and across families (CV-A) for (a) grain yield and (b) grain dry matter content. The boxplots show the range, median (bar) and mean (×) of 50 CV runs for both traits grain dry matter yield (GDY) and grain dry matter content (GDC). Numbers above or below boxplots indicate average predictive abilities.

3.4.3 Prediction within calibration sets and genetic groups

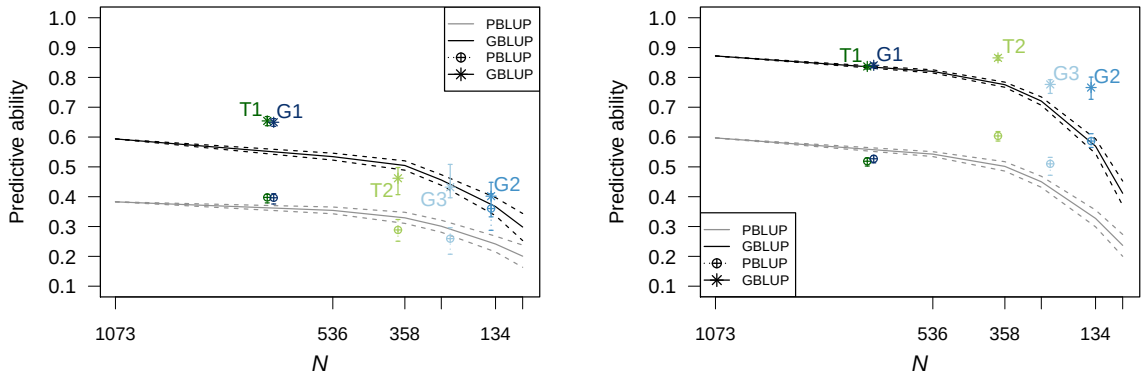
In both calibration sets of Maize 2, GBLUP consistently outperformed PBLUP irrespective if cross-validation procedure CV-R was performed in the entire data set, within genetic groups or within groups of lines crossed to the same tester (Figure 22 and 23). Predictive abilities obtained with PBLUP were of similar magnitude when compared across calibration sets, although the size of the data set decreased in CS2. Mean predictive abilities obtained with GBLUP were higher in CS1 than in CS2 for both traits.

In CS1, average predictive abilities were $r_{TS} = 0.59$ for GDY in the complete set of lines and ranged between $r_{TS} = 0.40$ for the smallest group G2 ($N_{G2} = 145$) and $r_{TS} = 0.65$ for the largest group G1 ($N_{G1} = 682$) with the highest mean expected kinship coefficient. Although the size of the ES was increased when CV-R was performed in the complete calibration set ($N = 1073$), there was no gain in predictive ability for GDY compared to a prediction within group G1. Mean predictive ability for GDY of tester T1 ($N_{T1} = 698$) was also substantially higher than within CS1, while predictive abilities within tester T2 were low ($r_{TS} = 0.46$). For GDC however, mean predictive ability was $r_{TS} = 0.87$ for

CS1 and higher than within groups. Although group G2 and G3 were small ($N_{G2} = 145$, $N_{G3} = 246$), predictive abilities for GDC were still high ($r_{TS} = 0.77$ and $r_{TS} = 0.78$, respectively). Predictions for GDC within tester T2 ($r_{TS} = 0.87$) performed better than within tester T1 ($r_{TS} = 0.84$).

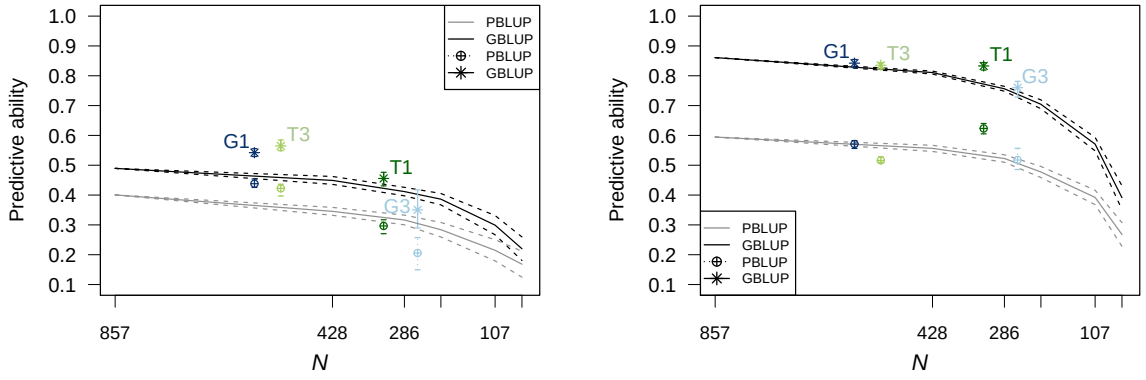
For CS2, mean predictive ability for GDY was $r_{TS} = 0.49$ for the complete data set and ranged between $r_{TS} = 0.35$ for group G3 ($N_{G3} = 260$) to $r_{TS} = 0.56$ for T3 ($N_{T3} = 530$). Predictive abilities within group G1 ($r_{TS} = 0.54$) were again higher than within the complete calibration set ($N = 857$). Highest predictive abilities for GDY were observed for tester T3, which was crossed to DH lines from all groups. For GDC, mean predictive ability was $r_{TS} = 0.86$ within CS2 and decreased when the predictions were performed within subsets of smaller sample sizes.

Results for the prediction of testcross values for GDY based on effects from GDC obtained from CV-R of the complete calibration sets are illustrated in Figure 24. Within CS1, prediction of GDY based on GDC resulted in a mean predictive ability of $|r_{TS}| = 0.22$ to 0.44 measured within each group/tester combination, but was lower than mean predictive ability for GDY obtained in CS1 with CV-R ($r_{TS} = 0.59$). Within CS2, the correlation between predicted testcross values of GDC and observed testcross values of GDY decreased substantially within the subsets of G1 and G3 tested to T1 ($|r_{TS}| = 0.01$ and 0.04, respectively). For the subsets of groups belonging to tester T3, higher correlations were observed.



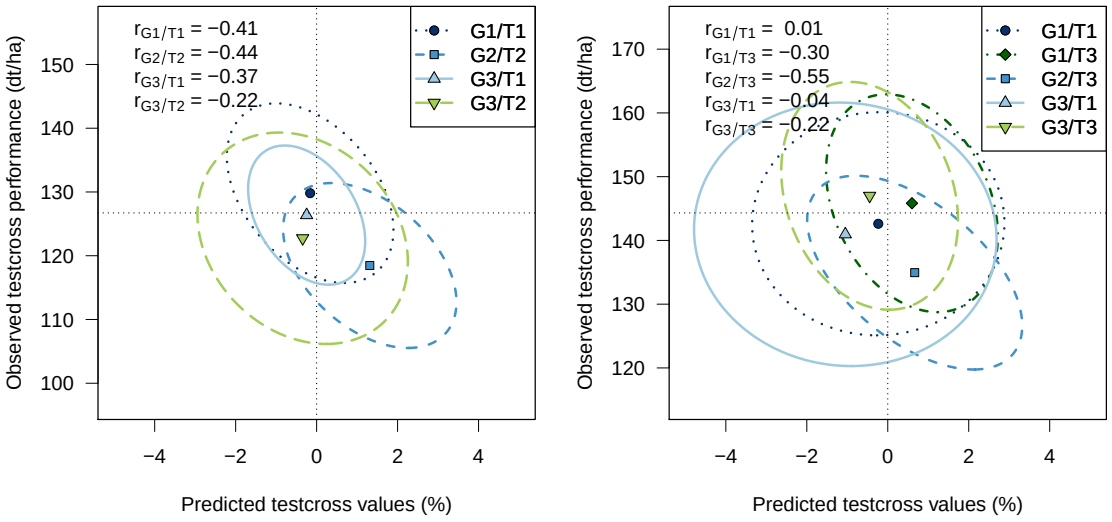
(a) Grain dry matter yield **(b)** Grain dry matter content

Figure 22: Predictive abilities with decreasing number of observations (N) for PBLUP and GBLUP assessed with 10×5-fold cross-validation within random subsets of the complete data set and within groups and tester subsets of calibration set CS1 for (a) grain yield and (b) grain dry matter content. Stars and circles with whiskers indicate average predictive abilities and range of 10 replications averaged across five folds. Dashed lines indicate 95 % confidence intervals for predictive abilities across random subsets.



(a) Grain dry matter yield **(b)** Grain dry matter content

Figure 23: Predictive abilities with decreasing number of observations (N) for PBLUP and GBLUP assessed with 10×5-fold cross-validation within random subsets of the complete data set and within groups and tester subsets of calibration set CS2 for (a) grain yield and (b) grain dry matter content. Stars and circles with whiskers indicate average predictive abilities and range of 10 replications averaged across five folds. Dashed lines indicate 95 % confidence intervals for predictive abilities across random subsets.



(a) CS1 **(b)** CS2

Figure 24: Correlation of observed testcross performance for grain dry matter yield with predicted testcross values of grain dry matter content obtained with GBLUP and CV-R within (a) calibration set CS1 and (b) CS2 visualized as elliptical contours representing 95% confidence intervals of means for individual group/tester subsets. The respective group/tester subsets are indicated by the symbol in the center of the ellipse and different types of lines. Correlations (r) are given for individual group/tester subsets.

3.4.4 Prediction across groups and testers

Results for the prediction across genetic groups and testers obtained with procedure CV-aG are given for both calibration sets in Table 5 and 6.

In CS1, tester T1 was crossed mainly to DH lines from group G1 (see Table 1a) and T2 was crossed only to lines of groups G2 and G3. Thus, the effects of genetic substructure and SCA cannot clearly be separated in CS1. The small subset of the combination from group G3 and tester T1 can be neglected as no large effect on predictive abilities within the subset of T1 compared to group G1 has been observed. For GDY, predictive abilities in group G1 decreased when the ES was augmented with DH lines from G2 and G3 even though sample size of the ES was almost doubled (Table 5). The same effect was observed for group G3 even though the increase in sample size of the ES was more dramatic than for group G1. In contrast for G2, there was an increase in predictive ability from $r_{TS} = 0.40$ to 0.44 when the ES included also lines from groups G1 and G3. The increase in sample size in this case was almost tenfold with $N_{ES} = 1044$ for $\mathrm{ES}_{\mathrm{G1,G2,G3}}$ as compared to $N_{ES} = 116$ for $\mathrm{ES}_{\mathrm{G2}}$. To account for the effect of sample size, testcross values of DH lines from G2 were predicted with randomly sampled 116 DH lines from the 1044 DH lines of $\mathrm{ES}_{\mathrm{G1,G2,G3}}$ which lead to a decrease of average predictive abilities from $r_{TS} = 0.44$ to 0.23. For all groups in CS1, predictive abilities decreased substantially ($r_{TS} = 0.26 - 0.30$), when the ES consisted only of DH lines from genetic groups that were not included in the TS. For GDC, augmenting the ES with lines from the other genetic groups led to markedly higher predictive abilities for all groups. Prediction of genetic values of DH lines from groups G2 and G3 worked well ($r_{TS} = 0.69$ and 0.74, respectively) even when the ES comprised only DH lines from the other genetic groups as long as the sample size of the ES was large enough. For the TS of G1, predictive abilities decreased ($r_{TS} = 0.57$) when only the groups G2 and G3 were included in the ES ($N_{ES} = 391$). As expected in CS1, very similar results could be observed when prediction was performed across the two groups of lines crossed to T1 and T2 as they were largely reflecting differences in genetic groups (Table 6). The comparison of predictive abilities for T1 based on $\mathrm{ES}_{\mathrm{T2}}$ ($N_{ES} = 375$) and T2 based on $\mathrm{ES}_{\mathrm{T1}}$ ($N_{ES} = 698$) showed that the sample size of the respective ES had a strong effect on prediction across testers for both traits.

In CS2, genetic groups G1, G2, and G3 were not as clearly differentiated based on marker or pedigree data as in CS1 (Figure 15 and Table 3). In addition, subsets of genetic groups G1 and G3 were crossed to both testers (T1 and T3). The mean kinship within the subsets of DH lines crossed to tester T1 were lower than compared to the subsets crossed to T3. Therefore, results for the prediction across groups in CS2 are illustrated separately for the four group/tester subsets (Table 5 and 6). For GDY, the prediction of G1/T1 and G1/T3 with $ES_{G2,G3}$ and G3/T1 with $ES_{G1,G2}$, respectively, resulted in a much smaller difference in predictive ability between within and across group predictions as compared to differences in CS1 (Table 5) reflecting the higher connectedness between groups in CS2. For G3/T3, predictive abilities increased with $ES_{G1,G2}$ as compared to the prediction within the group/tester subset, probably driven by the larger sample size of the ES. For GDC, predictive abilities increased for group/tester combinations when DH lines from all groups comprised the ES. For G1/T1 and G3/T3, predictions within the group/tester subsets were reduced compared to the prediction within the complete group G1 and G3 (Figure 23), respectively, and prediction based on all groups lead to an increase in predictive ability.

In addition for CS2, pairwise predictions across each group/tester combination (G1/T1, G1/T3, G3/T1, and G3/T3) were performed separately (Table 7). Results were very similar for both traits. The best predictive ability was obtained within each group/tester combination except for G3/T3 where an increase in predictive ability could be achieved with ES from G1/T3 having a substantially larger sample size ($N_{ES} = 393$) than the ES for G3/T3 ($N_{ES} = 98$). Predictive abilities decreased more when the tester changed as compared to the genetic group, which might have been the result of relatively small genetic diversity between G1 and G3 in CS2 or the non-random assignment of DH lines from the respective groups to testers T1 and T3. With keeping the size of the ES constant at $N_{ES} = 98$ (Figure 25), predictive abilities of GBLUP were highly correlated with the mean and maximum kinship coefficients between DH lines of the group/tester combinations (Figure 26). For expected kinship coefficients, predictive abilities across group/tester subsets were significantly ($p < 0.01$) correlated with $\overline{k}_{max}$. For $\overline{k}$, correlations were only significant for GDY ($p < 0.01$).

Table 5: Predictive abilities of the prediction within and across genetic groups with specific estimation sets (ES) for the same test set (TS) of calibration set CS1 and CS2 in Maize 2. Size of each TS (N_{TS}) and corresponding ES (N_{ES}) are given in the table.

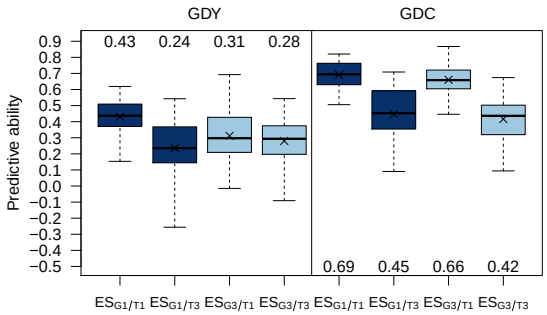
	TS		ES		Predictive ability ± standard deviation	
	Group	N_{TS}	Group	N_{ES}	GDY	GDC
CS1	G1	136	G1	546	0.650 ± 0.007	0.841 ± 0.005
			G1,G2,G3	937	0.631 ± 0.007	0.857 ± 0.004
			G2,G3	391	0.292 ± 0.004	0.568 ± 0.003
	G2	29	G2	116	0.402 ± 0.035	0.766 ± 0.022
			G1,G2,G3	1044	0.441 ± 0.021	0.817 ± 0.014
			G1,G3	928	0.303 ± 0.018	0.694 ± 0.012
	G3	49	G3	197	0.432 ± 0.034	0.776 ± 0.013
			G1,G2,G3	1024	0.385 ± 0.024	0.826 ± 0.006
			G1,G2	827	0.260 ± 0.007	0.737 ± 0.006
CS2	G1/T1	38	G1/T1	151	0.442 ± 0.029	0.724 ± 0.029
			G1,G2,G3	799	0.464 ± 0.017	0.839 ± 0.013
			G2,G3	275	0.424 ± 0.013	0.753 ± 0.007
	G1/T3	78	G1/T3	315	0.592 ± 0.009	0.832 ± 0.007
			G1,G2,G3	779	0.567 ± 0.010	0.848 ± 0.006
			G2,G3	275	0.401 ± 0.009	0.688 ± 0.002
	G3/T1	28	G3/T1	110	0.331 ± 0.048	0.804 ± 0.016
			G1,G2,G3	829	0.349 ± 0.036	0.836 ± 0.009
			G1,G2	597	0.292 ± 0.020	0.740 ± 0.012
	G3/T3	24	G3/T3	98	0.298 ± 0.036	0.536 ± 0.025
			G1,G2,G3	833	0.409 ± 0.020	0.768 ± 0.019
			G1,G2	597	0.442 ± 0.013	0.556 ± 0.013

Table 6: Predictive abilities of the prediction within and across tester subsets with specific estimation sets (ES) for the same test set (TS) of calibration set CS1 and CS2 in Maize 2. Size of each TS (N_{TS}) and corresponding ES (N_{ES}) are given in the table.

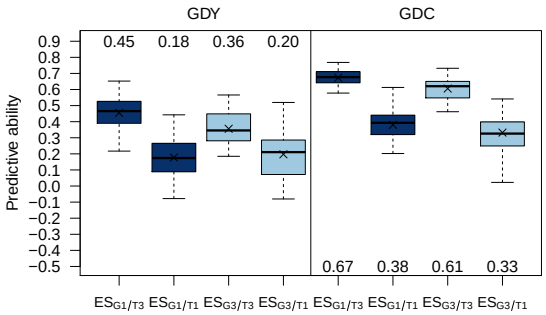
	TS		ES		Predictive ability ± standard deviation	
	Tester	N_{TS}	Tester	N_{ES}	GDY	GDC
CS1	T1	140	T1	558	0.654 ± 0.012	0.837 ± 0.005
			T1,T2	933	0.636 ± 0.009	0.850 ± 0.006
			T2	375	0.288 ± 0.004	0.567 ± 0.004
	T2	75	T2	300	0.462 ± 0.029	0.865 ± 0.005
			T1,T2	998	0.500 ± 0.019	0.885 ± 0.004
			T1	698	0.405 ± 0.007	0.785 ± 0.003
CS2	G1/T1	38	G1/T1	151	0.442 ± 0.029	0.724 ± 0.029
			T1/T3	799	0.464 ± 0.017	0.839 ± 0.013
			T3	530	0.344 ± 0.006	0.680 ± 0.010
	G3/T1	28	G3/T1	110	0.331 ± 0.048	0.804 ± 0.016
			T1,T3	829	0.349 ± 0.036	0.836 ± 0.009
			T3	530	0.249 ± 0.018	0.497 ± 0.028
	G1/T3	78	G1/T3	315	0.592 ± 0.009	0.832 ± 0.007
			T1,T3	779	0.567 ± 0.009	0.848 ± 0.006
			T1	327	0.204 ± 0.009	0.527 ± 0.006
	G3/T3	24	G3/T3	98	0.298 ± 0.036	0.536 ± 0.025
			T1,T3	833	0.409 ± 0.018	0.768 ± 0.019
			T1	327	0.208 ± 0.028	0.490 ± 0.015

Table 7: Predictive abilities of the prediction within and across group/tester subsets with specific estimation sets (ES) for the same test set (TS) in calibration set CS2. Size of each TS (N_{TS}) and corresponding ES (N_{ES}) and the mean ($\overline{k}$) and maximum ($\overline{k}_{max}$) expected kisnhip coefficients between group/tester combinations are given in the table.

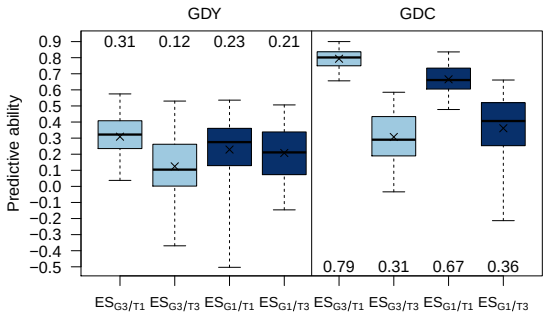
TS		ES		$\overline{k}$	$\overline{k}_{max}$	Predictive ability ± standard deviation	
Subset	N_{TS}	Subset	N_{ES}			GDY	GDC
G1/T1	38	G1/T1	151	0.20	0.52	0.442 ± 0.029	0.724 ± 0.029
		G1/T3	393	0.20	0.31	0.326 ± 0.008	0.651 ± 0.013
		G3/T1	138	0.13	0.33	0.338 ± 0.011	0.715 ± 0.007
		G3/T3	122	0.15	0.24	0.288 ± 0.016	0.462 ± 0.016
G1/T3	78	G1/T3	315	0.32	0.60	0.592 ± 0.009	0.832 ± 0.007
		G1/T1	189	0.20	0.34	0.172 ± 0.010	0.482 ± 0.006
		G3/T3	122	0.22	0.40	0.374 ± 0.008	0.647 ± 0.002
		G3/T1	138	0.14	0.28	0.205 ± 0.011	0.377 ± 0.005
G3/T1	28	G3/T1	110	0.16	0.42	0.331 ± 0.048	0.804 ± 0.016
		G3/T3	122	0.16	0.24	0.120 ± 0.014	0.322 ± 0.019
		G1/T1	189	0.13	0.35	0.288 ± 0.017	0.753 ± 0.012
		G1/T3	393	0.14	0.21	0.230 ± 0.020	0.466 ± 0.032
G3/T3	24	G3/T3	98	0.25	0.46	0.298 ± 0.036	0.536 ± 0.025
		G3/T1	138	0.16	0.27	0.194 ± 0.020	0.325 ± 0.020
		G1/T3	393	0.22	0.42	0.446 ± 0.009	0.564 ± 0.013
		G1/T1	189	0.15	0.26	0.208 ± 0.014	0.400 ± 0.020



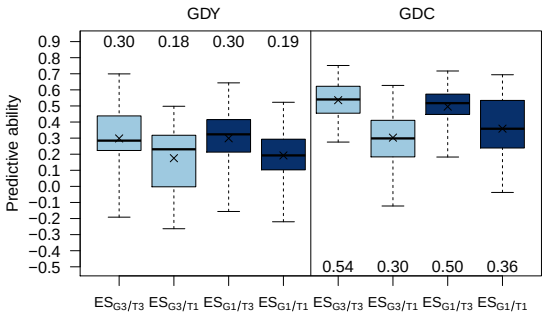
(a) Specific ES for G1/T1 ($N_{TS} = 38$)



(b) Specific ES for G1/T3 ($N_{TS} = 28$)

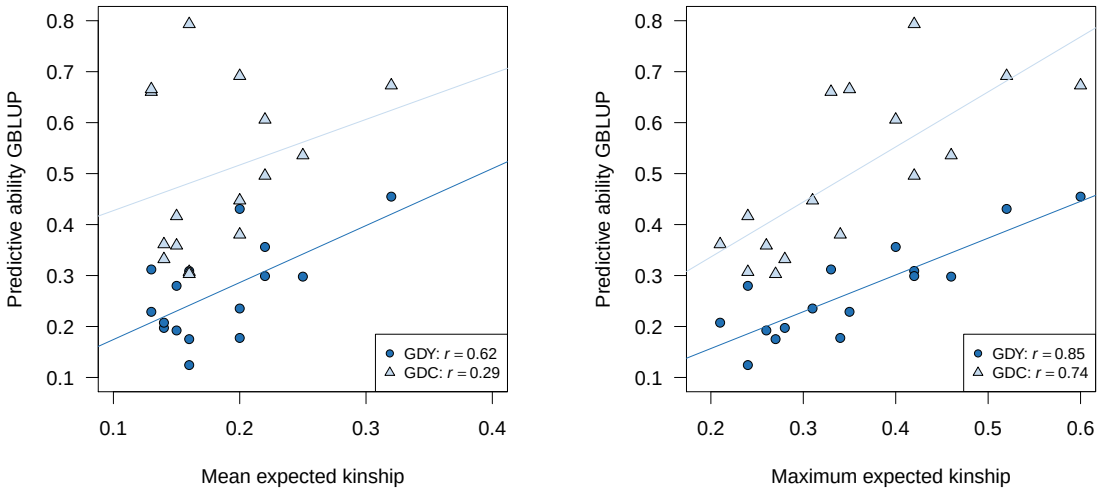


(c) Specific ES for G3/T1 ($N_{TS} = 78$)



(d) Specific ES for G3/T3 ($N_{TS} = 24$)

Figure 25: Predictive abilities obtained with 10×5-fold cross-validation across group/tester subsets of calibration set CS2. Specific estimation sets (ES) were sampled with constant size ($N_{ES} = 98$) for test sets (TS) of (a) G1/T1, (b) G1/T3, (c) G3/T1, and (d) G3/T3 containing the same group/tester subset or the remaining group/tester subsets. The boxplots show the range, median (bar) and mean (×) of 50 CV runs for both traits grain dry matter yield (GDY) and grain dry matter content (GDC). Numbers above and below boxplots indicate average predictive abilities.



(a) Mean expected kinship coefficients

(b) Maximum expected kinship coefficients

Figure 26: Mean predictive ability of GBLUP for traits GDY and GDC obtained for the prediction within and across group/tester combinations (G1/T1, G1/T3, G3/T1, and G3/T3) in calibration set CS2 plotted against (a) mean ($\overline{k}$) and (b) maximum kinship ($\overline{k}_{max}$) within and across group/tester combinations. The sample size of the estimation sets was fixed to $N_{ES} = 98$ for each of the 16 possible combinations. Correlations (r) between predictive abilities and mean and maximum expected kinship coefficients, respectively, are given for both traits in the legend.

3.4.5 Predictive abilities based on different kinship coefficients between testcrosses

A pairwise comparison between different kinship matrices for the Tester T1 subset of CS1 in Maize 2 is visualized in Figure 27. The Pearson's correlation coefficient between the different kinship coefficients derived from pedigree or genome-wide marker data ranged between $r(\mathbf{K}, \mathbf{U_{AB}}) = 0.55$ to $r(\mathbf{U}, \mathbf{U_{AB}})$=0.97. As both realized kinship matrices $\mathbf{U}$ and $\mathbf{U_{AB}}$ are based on a marker matrix with column-centered genotype scores, this high accordance was expected. The realized kinship coefficients $\mathbf{U_S}$, $\mathbf{U_{T_{GDY}}}$ and $\mathbf{U_{T_{GDC}}}$ were also highly correlated ($r = 0.88 - 0.94$) as both were derived from the simple matching coefficient. The lowest correlations were observed between the expected and realized kinship coefficients ($r = 0.55 - 0.76$), where $\mathbf{U_S}$ showed the highest accordance with $\mathbf{K}$.

Average predictive abilities derived from PBLUP and GBLUP for the tester T1 subset of CS1 based on these kinship coefficients are shown in Table 8. Confirming the results from Section 3.4.3, lowest predictive abilities were observed for the pedigree-based kinship. For GDY, predictive abilities obtained with different realized kinship coefficients showed no large differences. Larger differences occurred between GBLUP with different kinship coefficients for the trait GDC. Here, highest predictive abilities ($r_{TS} = 0.84$) were obtained with $\mathbf{U_S}$ and $\mathbf{U}$. The predicted testcross values derived from the GBLUP model with these two realized kinship coefficients were correlated with $r = 1$ but shifted by a constant.

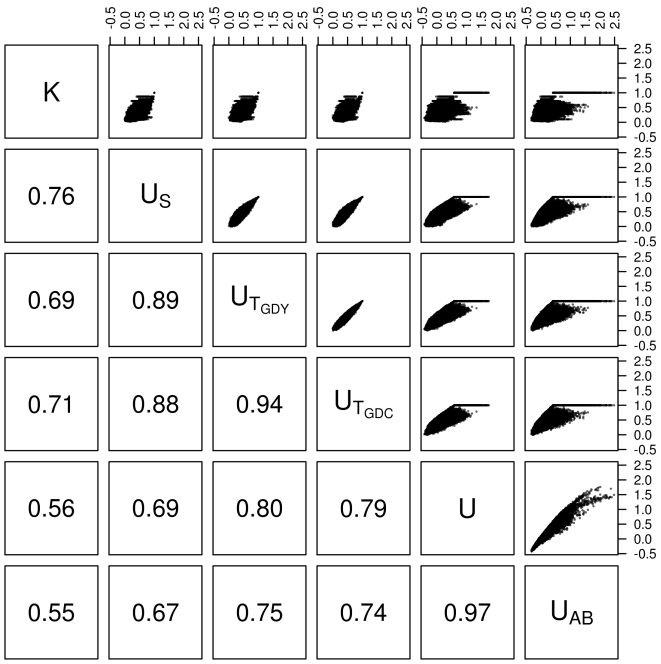


Figure 27: Scatterplot of pairwise comparisons between kinship coefficients. Lower triangle shows the correlation coefficients between kinship matrices for DH lines from subset of tester T1 of calibration set CS1 in Maize 2.

Table 8: Mean predictive abilities and attached standard errors for grain yield (GDY) and grain dry matter content (GDC) obtained with 10×5-fold cross-validation with random sampling of PBLUP and GBLUP with different realized kinship coefficients for tester T1 subset of calibration set CS1 in Maize 2.

Kinship	Predictive ability ± standard deviation	
coefficients	GDY	GDC
K	0.398 ± 0.007	0.519 ± 0.008
$\mathbf{U_S}$	0.654 ± 0.011	0.844 ± 0.004
$\mathbf{U_T}$	0.649 ± 0.012	0.831 ± 0.004
U	0.654 ± 0.011	0.844 ± 0.004
$\mathbf{U_{AB}}$	0.654 ± 0.012	0.837 ± 0.005

3.4.6 Predictive abilities with decreasing number of locations and sample size

The dependency of predictive abilities on the number of locations (L) and number of individuals (N) within the data set of tester T1 in CS1 of Maize 1 is displayed for both traits in Table 9. Predictive abilities decreased with decreasing number of locations, which were used for the phenotypic evaluation of testcrosses. Moreover, within a constant set of locations, predictive abilities decreased when the number of individuals decreased. For GDY, predictive abilities were reduced from $r_{TS} = 0.65$ within the set of four locations and with all 698 DH lines to $r_{TS} = 0.32$ when only 87 DH lines and one location were used for testcross evaluation. For GDC, predictive abilities were reduced from $r_{TS} = 0.84$ to 0.50.

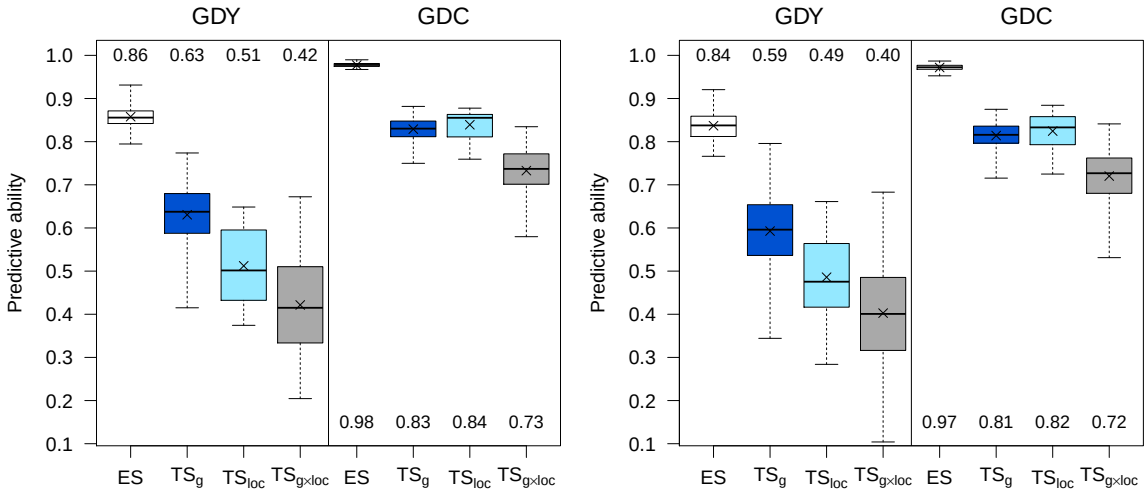
Table 9: Predictive abilities and standard deviations within subsets of specific size (N) and number of locations (L) obtained from 10×5-fold cross-validation for grain yield (GDY) and grain dry matter content (GDC). Predictive abilities were averaged across randomizations for each combination of $N \times L$ subsets as stated in Table 2.

		N			
Trait	L	698	349	175	87
GDY	4	0.654	0.592 ± 0.058	0.530 ± 0.071	0.432 ± 0.103
	3	0.631 ± 0.038	0.568 ± 0.068	0.505 ± 0.085	0.407 ± 0.119
	2	0.593 ± 0.061	0.531 ± 0.086	0.468 ± 0.111	0.375 ± 0.142
	1	0.523 ± 0.097	0.466 ± 0.106	0.407 ± 0.139	0.322 ± 0.157
GDC	4	0.837	0.766 ± 0.013	0.685 ± 0.045	0.534 ± 0.084
	3	0.829 ± 0.009	0.758 ± 0.017	0.679 ± 0.045	0.531 ± 0.082
	2	0.814 ± 0.016	0.744 ± 0.023	0.667 ± 0.049	0.522 ± 0.088
	1	0.777 ± 0.025	0.708 ± 0.031	0.636 ± 0.057	0.500 ± 0.098

3.4.7 Prediction across locations

Figure 28 illustrates the predictive abilities with sampling of genotypic subsets, locations, and both factors simultaneously based on the possible combination of three and two locations in the ES as described in Figure 8. With genotypic sampling, predictive abilities ranged between $r_{TS} = 0.41$ and 0.77 with an average of $r_{TS} = 0.63$ for GDY. Accounting for sampling of locations resulted in lower mean predictive abilities (TS_{loc}, $r_{TS} = 0.51$). Predictive abilities were reduced even more, when the sampling accounted for genotypes and locations simultaneously ($TS_{g\times loc}$; $r_{TS} = 0.42$). Predictive abilities were reduced for all sampling schemes, when only two locations were sampled for the ES. Predictive abilities for grain dry matter content were higher than for grain yield for all CV schemes. For GDC, accounting for sampling of locations did not lead to reduced predictive abilities compared to genotypic sampling.

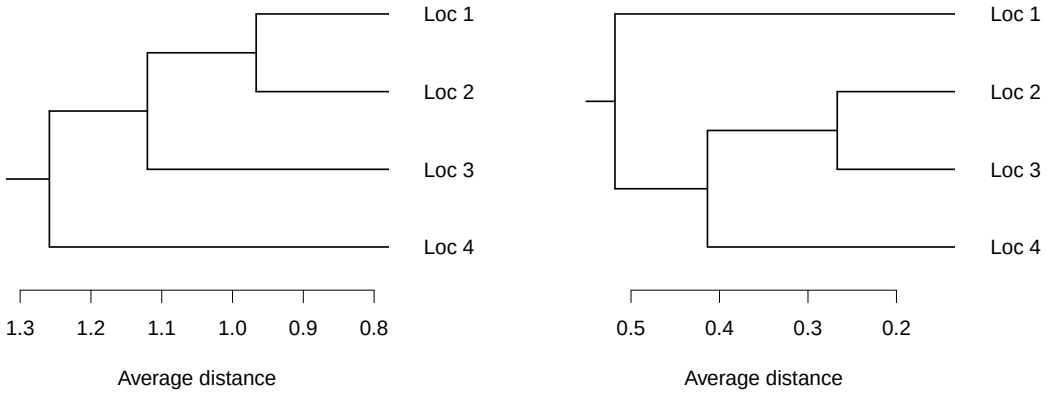
In Figure 29, the correlation between locations is illustrated with a dendrogram. When highly correlated locations were sampled for the ES, estimated heritabilities were high. In Figure 30, mean predictive abilities are illustrated for each possible combination of locations in the TS when the ES included three or two environments against the heritability in each ES. Figure 31 illustates predictive abilities against the phenotypic correlation between ES and TS. With genotypic sampling (TS_g), predictive abilities were significantly ($p < 0.01$) dependent on trait heritabilities for both traits obtained from a specific set of locations in the ES. Mean predictive abilities across locations (TS_{loc}) were significantly ($p < 0.001$) reflected in the phenotypic correlations between adjusted means of locations in the ES and TS. Best predictions of testcross values were observed when heritabilities of the ES were high and the adjusted means from the TS and ES were highly correlated.



(a) Three locations in the ES

(b) Two locations in the ES

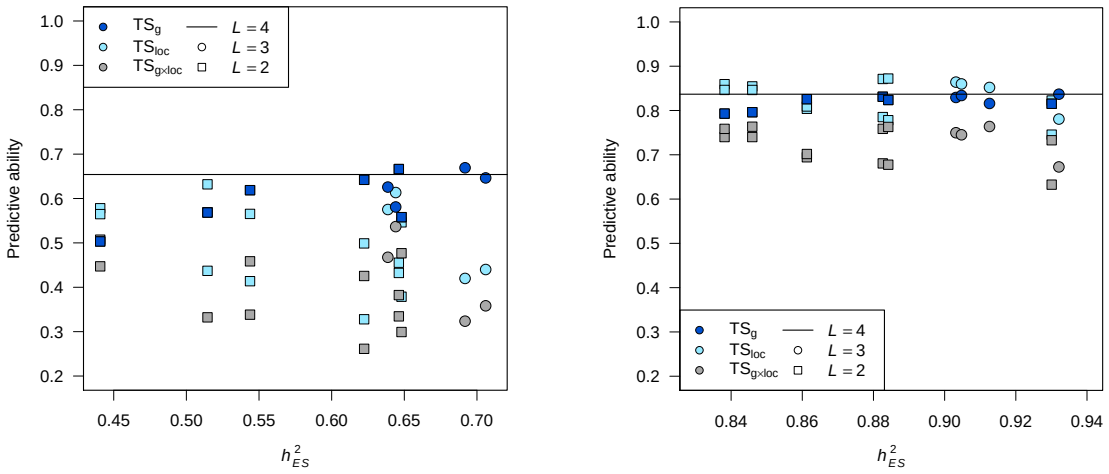
Figure 28: Predictive abilities of different TS derived from (a) 300 possible ES of three locations and (b) 200 possible ES of two locations from genotypic and environmental sampling in cross-validation for both traits grain yield (GDY) and grain dry matter content (GDC). Numbers above and below boxplots show average predictive abilities.



(a) Grain dry matter yield

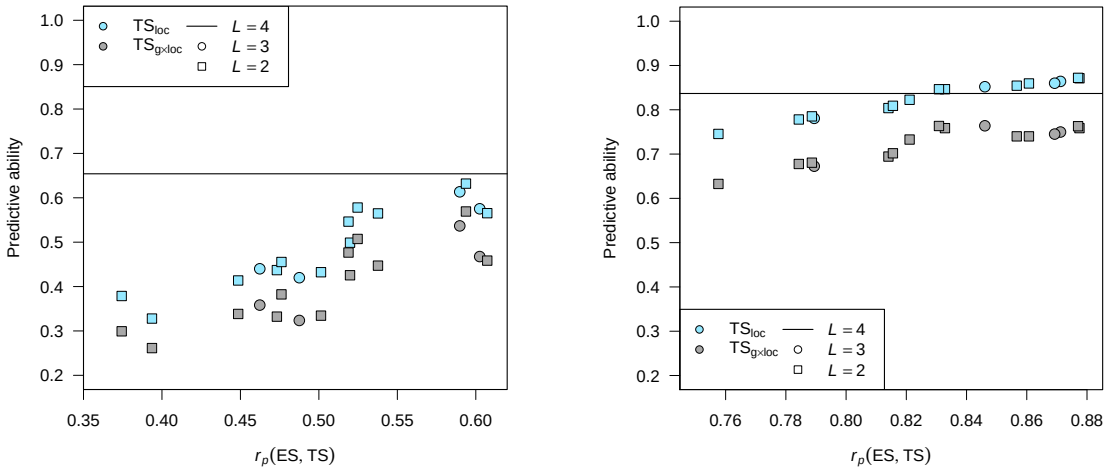
(b) Grain dry matter content

Figure 29: Dendrogram of locations estimated with average linkage clustering. Distance between locations was calculated based on (a) grain yield and (b) grain dry matter content.



(a) Grain dry matter yield **(b)** Grain dry matter content

Figure 30: Mean predictive abilities within three specific test sets (TS_g, TS_{loc},$TS_{g\times loc}$) for different subsets of locations in the estimation set including two or three locations against trait heritabilities in the estimation set (h^2_{ES}) for (a) grain yield and (b) grain dry matter content.



(a) Grain dry matter yield **(b)** Grain dry matter content

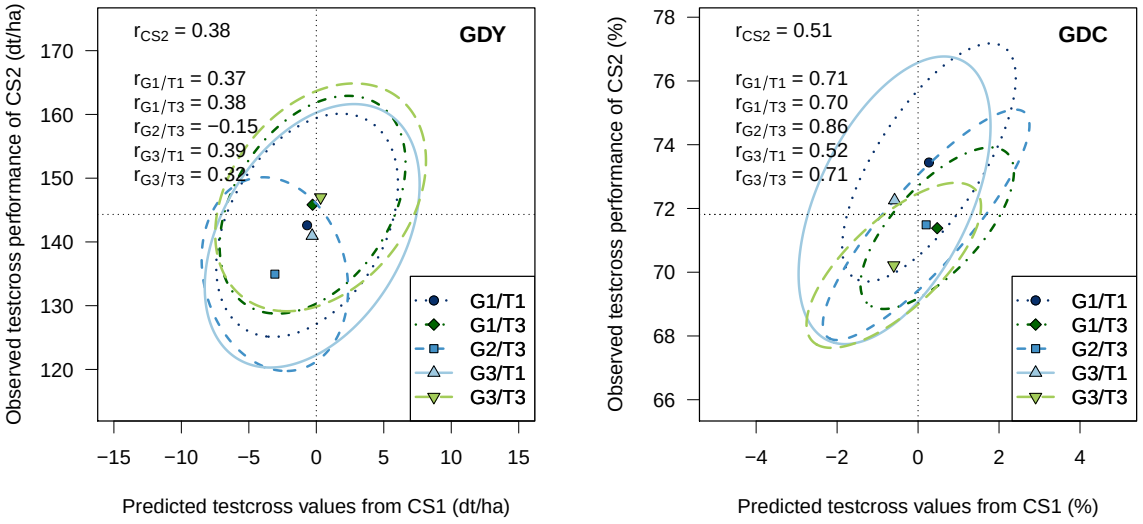
Figure 31: Mean predictive abilities within specific test sets (TS_{loc},$TS_{g\times loc}$) for different subsets of locations in the estimation set including two or three locations against phenotypic correlations between adjusted means of locations in the estimation and test set ($r_p(ES, TS)$) for (a) grain dry matter yield and (b) grain dry matter content.

3.4.8 Prediction across years

In both calibration sets, a subset of DH lines was evaluated with tester T1. Therefore, the testcross values of DH lines from the group/tester combination G1/T1 in 2011 (CS2) were predicted based on the data from the tester T1 subset in 2010 (CS1). The correlations between predicted and observed testcross values of G1/T1 in CS2 were $r = 0.36$ for GDY and $r = 0.61$ for GDC (Appendix, Table A8). The prediction across years resulted in even more reduced predictive abilities as compared to the prediction of an independent set of DH lines in an independent location within the same year (see $TS_{g\times loc}$).

The prediction across the two complete calibration sets was also investigated. Figure 32 shows correlations between observed testcross performance of DH lines in one year and their predicted testcross values derived from model training with the entire calibration set of the other year separated for the respective group/tester combination. For GDY, predictive abilities across years ranged between 0.30 and 0.44 for all group/tester subsets except for DH lines from group G2 that could not be predicted accurately ($r_{TS} = 0.28$ in CS1 and -0.15 in CS2). For GDC, predictive abilities across years ranged between 0.50 and 0.74. Within the respective group/tester combinations, predictions across years were of similar magnitude irrespective of whether model training was performed on CS1 and predictions on CS2 or vice versa. However, when calculated across all subsets, predictive abilities were higher in CS1 ($r_{CS1} = 0.58$ for GDY and $r_{CS1} = 0.74$ for GDC) than in CS2 ($r_{CS1} = 0.38$ for GDY and $r_{CS1} = 0.51$ for GDC). Predictions across years for GDY were in a similar range as predictions obtained for the prediction across locations when sampling independent genotypes (Figure 28, $TS_{g\times loc}$).

For predicting testcross values of DH lines of VS1 evaluated in 2011, different reference sets were used for model training. The results from these predictions are shown for both traits in Table 10. For GDY, the highest correlation ($r = 0.60$) was obtained when the observations from CS1 were used for predicting the testcross performance in 2011. For GDC, predictive abilities were higher than the phenotypic correlation with CS1 when the testcross values were predicted based on genotypic information from CS1. An additional increase was observed when the reference set included also the DH lines from CS2, which was evaluated in the same year as the VS1. For both traits, predictive abilities dropped when the selected DH lines were excluded from CS1 in the reference set.



(a) Prediction of GDY from CS1

(b) Prediction of GDC from CS1

(c) Prediction of GDY from CS2

(d) Prediction of GDC from CS2

Figure 32: Observed against predicted testcross values of calibration set CS1 and CS2 predicted with the other calibration set for traits grain yield (GDY) and grain dry matter content (GDC) as elliptical contours representing 95 % confidence intervals for each group/tester combination. Predictive abilities as correlation (r) between observed and predicted testcross values are stated within the figures for each calibration set and within group/tester combinations.

Table 10: Correlation of observed testcross performance for grain dry matter yield and grain dry matter content of DH lines in the validation set VS1 with observed or predicted testcross values from calibration set CS1 and/or CS2. The DH lines selected for the VS1 were a subset of the DH lines from CS1 and their observations from 2010 were included or excluded for model training. Confidence intervals for $\alpha = 0.05$ are in parentheses.

Correlation of observed testcross performance of VS1 in 2011 with	Grain dry matter yield		Grain dry matter content	
Observed testcross values of VS1 in 2010	0.599	[0.454, 0.713]	0.814	[0.734, 0.872]
Predicted testcross values based on model training with				
Observed values of VS1 from 2010 included				
CS1	0.555	[0.400, 0.680]	0.846	[0.778, 0.895]
CS1 + CS2	0.582	[0.433, 0.700]	0.850	[0.783, 0.897]
Observed values of VS1 from 2010 excluded				
CS1	0.519	[0.357, 0.651]	0.772	[0.676, 0.842]
CS1 + CS2	0.548	[0.391, 0.674]	0.798	[0.712, 0.861]
CS2	0.498	[0.332, 0.635]	0.743	[0.639, 0.821]

4 Discussion

4.1 Modeling the kinship between DH lines

The application of pedigree-based kinship coefficients has been widely adopted into plant breeding programs (Bernardo 2002). In association studies, kinship coefficients have been applied to correct for spurious associations which rely only on relationships between individuals instead of true marker trait associations (Astle and Balding 2009). When pedigree data is not available, marker information can be used to model these kinships. One of the first suggestions for marker-based kinship estimates have been reported in Fernando and Grossman (1989). For genomic prediction, the estimated kinship coefficients are used to model the variance-covariance structure of random effects to predict the genetic value of related individuals. For traits regulated by a large number of genes with small effects, e.g., grain yield and grain dry matter content (Melchinger et al. 1998; Schön et al. 2004), and populations with strong long range LD, these mixed effects models have been shown to perform well with respect to prediction accuracies (Lorenzana and Bernardo 2009; Zhong et al. 2009; Piepho 2009; Crossa et al. 2010; Wimmer et al. 2013).

4.1.1 Predictive abilities with pedigree- and genome-wide marker data

Irrespective of the CV scheme employed, genome-enabled predictions performed substantially better than the model based on pedigree data alone. So far, PBLUP has been a standard procedure for predicting the performance of selection candidates within a breeding population. The relative performance of GBLUP compared to PBLUP indicates an advantage of marker information in contrast to pedigree data only (Daetwyler et al. 2013). As pointed out by Goddard (2008), for predicting the magnitude of this increase in accuracy the effective number of segregating loci in the population under study is most relevant. In advanced cycle breeding populations with small effective sample size and doubled haploid lines generated from F_1 plants, extensive LD can be expected and was shown for the experimental material under study. Consequently, the variation in realized genetic relationship among DH lines sharing the same expected relatedness, i.e.,

within full-sib families, is high (Appendix, Figure A1 and A2), leading to an increase in prediction accuracy for models using genomic data. In Maize 1 for 264 DH lines of a randomly chosen test set from CV-W, Figure 33 illustrates predicted testcross values for grain yield derived from PBLUP and GBLUP relative to their respective family mean calculated from adjusted means of the full data set. While in PBLUP testcross performance of all DH lines derived from the same cross obtain the same predicted value, variation of testcross values predicted with GBLUP is large within each of the 36 families leading to higher predictive abilities.

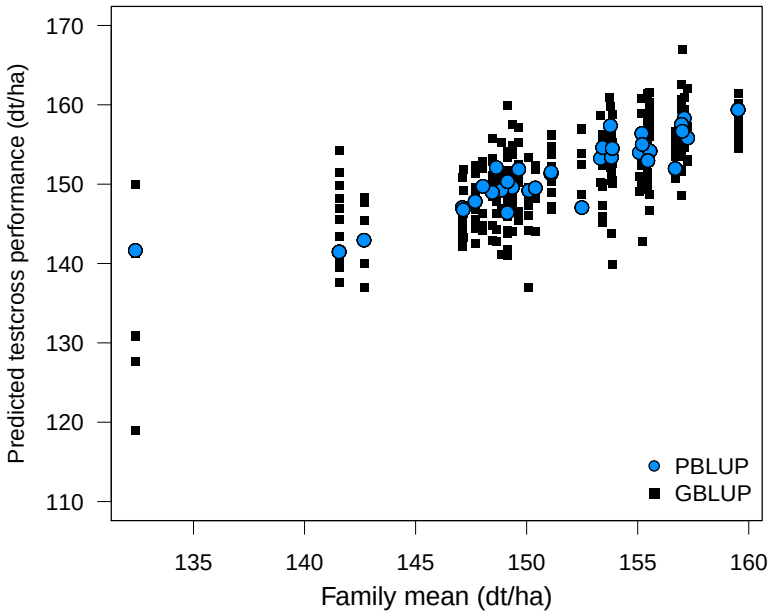


Figure 33: Predicted testcross performance with PBLUP and GBLUP obtained from one test set of within family sampling in Maize 1 plotted against the respective family means calculated from adjusted means for grain yield.

Including pedigree information in addition to genome-wide marker data in the model (P+GBLUP) improved prediction of testcross performance but only to a small extent and only with CV-W (Table 4). A modest improvement of predictive abilities of models including pedigree information in addition to marker data was also observed by Crossa et al. (2010). Goddard (2008) pointed out that including a polygenic term in the model might be beneficial for capturing the effects of alleles with low frequency. However, the relative performance of P+GBLUP was equal compared to GBLUP when predictions were calculated for distantly related DH lines in CV-A (Table 4). As observed in this study,

even with high-quality pedigree and genome-wide marker data for all individuals in model training and prediction, P+GBLUP did not outperform GBLUP with mixed effects models due to the redundancy of the data.

The conclusions on the relative performance of the three models were further confirmed when comparing the mean phenotypic testcross performance of the 10 % best lines selected based on their predicted testcross performance for grain yield. Lines selected based on predictions from GBLUP and P+GBLUP performed markedly better than those selected based on PBLUP with CV-W and CV-A, but P+GBLUP did not have an advantage over GBLUP.

In Maize 2, the relative performance of PBLUP to GBLUP in both calibration sets confirmed the differences in family substructure (see also Section 4.2.3). Due to the smaller family sizes, prediction performance of PBLUP for GDY was high in CS2 (Figure 23). However, realized genetic kinship coefficients capture the variation within large crosses and the advantage of GBLUP over PBLUP was higher in CS1 than in CS2.

4.1.2 Predictive abilities with different genome-based kinship coefficients

The estimation of genomic relatedness was one of the first applications of molecular markers in plant breeding and has been successfully used for a wide range of applications such as management of heterotic pools, prediction of heterosis, and diversity analyses. Many different measures of relatedness have been proposed for quantifying the kinship between pairs of individuals (Reif et al. 2005). In maize breeding, estimates of kinship between fully homozygous inbred lines are frequently calculated from the proportion of shared marker alleles corrected for the average proportion of alleles alike in state between unrelated individuals in the population under study (Bernardo 1993). A similar approach is the application of $\mathbf{U_S}$, but instead of correcting with the proportion of alleles alike in state estimated from unrelated individuals, the minimum value of alleles shared between DH lines (s_{min}) as suggested by Hayes and Goddard (2008) was taken as a correction term. As shown in the Appendix, the choice of the correction factor for $\mathbf{U_S}$ in GBLUP affects the estimated variance components in a predictable form and predicted testcross values from GBLUP will be shifted in scale but ranked identically when

compared to GBLUP with **U**. Predictive abilities obtained in CV were not influenced by the choice of one of these two kinship coefficients (Table 8).

The kinship between DH lines can be estimated from genome-wide covariances of allele counts which can be interpreted as deviations of allele sharing from that expected for unrelated individuals. In this study, allele frequencies were estimated directly from the data to obtain the elements of the matrix **U** for GBLUP. As pointed out by Habier et al. (2007), correction of allele counts (elements w_{im}) with their expected values ($2p_m$) subtracts the same constant for all individuals in the population under study and thus only leads to a scale-shift. Analogously as for GBLUP, testcross values predicted with RRBLUP are shifted in scale but ranked identically when compared to GBLUP. The transformation of RRBLUP to GBLUP is only possible, when the final matrix of kinship coefficients is not further modified. Although negative kinship coefficients can occur, the interpretation as a correlation coefficient is straightforward and negative values should not be set to zero which was recently applied in association studies (Bernardo 1993; Stich et al. 2008).

Comparing the kinship coefficient based on centered or centered and scaled marker scores (VanRaden 2008; Astle and Balding 2009), only the second one puts more weights on the rare alleles, while the first results in identical estimates as RRBLUP. Endelman and Jannink (2012) discussed that different shrinkage of the kinship matrix can influence predictive abilities when $N > M$ and the phenotypic accuracies are low. With multi-environmental trials for maize breeding populations which are genotyped with a high density panel of SNPs, it was expected that there might be no large difference of predictive abilities as observed for the data sets of this study (Table 8). However, Scutari et al. (2013) concluded that the kinship based on column-scaled marker scores might be superior to the kinship according to VanRaden (2008).

4.2 Implications from cross-validation and validation

Cross-validation is a nonparametric method for the selection and evaluation of prediction models. The main feature of this method is to split the data set into two subsets to obtain estimates of the predictive ability and accuracy without a second trial. Different sampling strategies were applied to analyze factors influencing predictive ability and to

obtain an estimate of predictive ability for the different models. Dividing the data set into subsets is always a loss of information in estimating the effects of a model, because one subset of observations is discarded in model fitting. But with decreasing test set size, the variance of predictive abilities increases. Breiman and Spector (1992) proposed that 5- and 10-fold CV are suited for model selection and Kohavi (1995) obtained an optimum for predictive abilities with 10- and 20-fold CV. In this study, an optimal choice was $k = 5$ (Appendix, Figure A3). Similar sampling strategies were applied in a QTL study with 5-fold CV by Utz et al. (2000).

4.2.1 Stratified cross-validation

In stratified CV, each subset has the same distribution as the original data set. Kohavi (1995) proposed to use this sampling strategy to obtain less biased estimates of the predictive ability. This can be achieved by sampling equally within crosses (Table 4). In Maize 1, CV-W yielded high average accuracies for both traits, when the genetic relationship between DH lines was modeled based on genomic data ($\overline{r}_{TS} \geq 0.66$). These high values are in accordance with analytical and computer simulation results presented by Hayes et al. (2009b) who also showed high prediction accuracies within families as compared to random mating populations. Comparing CV-W with splitting the data set randomly into subsets for 5-fold CV resulted in similar accuracies. Random sampling was able to capture the family distribution of the whole data set, because each family was sufficiently represented in the complete data set since even the smallest full-sib family contained at least 14 DH lines. In addition, a high degree of relatedness between the DH lines in the ES and TS as well as long range haplotype blocks within families lead to high LD between markers and QTL causing these high predictive abilities obtained with CV-W.

In contrast to CV-W, across family sampling (CV-A) was applied as described by Legarra et al. (2008). As expected, average predictive abilities with using the information of distant relatives were lower compared to using information from close relatives. The low average accuracies obtained in CV-A with PBLUP ($\bar{r}_{TS} \leq 0.31$) indicated that in the population under study, families derived from different crosses were only distantly related by pedigree. For GBLUP and P+GBLUP, predictive abilities decreased not as

severely as with PBLUP, indicating that substantial LD between the markers and QTL was captured by the markers. Furthermore, predictive abilities obtained with CV-A varied considerably more compared to CV-W, which might be a result of the highly variable degree of relatedness between lines in the ES and the corresponding TS. As pointed out in Habier et al. (2010), the relatedness between lines in the ES and TS must be known to provide reliable accuracies for the prediction of testcross performances. On the other hand, due to sampling whole families of different size, the size of the ES varied over the 50 CV runs ($1002 \leq N_{ES} \leq 1172$). But, the possible minimum size of ES within the CV-A procedures could have been 902 which can still result in reliable predictive abilities as shown in Figure 20.

Predictive abilities obtained within full-sib families are considered to be higher than those for random mating populations because allele effects are estimated more accurately and the effective number of independently segregating loci controlling the phenotype is reduced (Hayes et al. 2009b). In addition, when model training is performed within biparental families with relatively low marker densities, high LD and the lack of population structure will increase predictive abilities for GP (Crossa et al. 2014). Therefore, predictive abilities of GBLUP were assessed within the four largest biparental families ($58 \leq N \leq 60$) of Maize 1. Predictive abilities obtained for the four families varied strongly from $r_{TS} = 0.26$ to 0.59 for GDY and from $r_{TS} = 0.47$ to 0.85 for GDC (Figure 19), probably the result of different genetic relatedness between the parents of the crosses or differing heritabilities within families (Lehermeier et al. 2014). Based on pedigree information alone, the expected kinship between progenies of all crosses showed the same degree of relatedness of 0.5. The number of polymorphic markers differed substantially. In the family with the lowest accuracy, only 116 of the 732 markers were polymorphic as compared to 212 in the family with the highest accuracy. Due to the small family sizes, predictive abilities varied highly across the 50 CV runs. Considering the substantially higher and much less variable predictive abilities obtained when using the full data set and CV-W, higher predictive abilities can be achieved with taking into account information from related families or population wide LD and it does not seem appropriate to perform model training within individual families (Jannink et al. 2010).

4.2.2 Allocation of ressources

The main factors influencing predictive ability are the resources available for developing the training population. The influence of the size of the training set on predictive abilities was analyzed in both breeding populations. For this purpose, the complete data set of Maize 1 was reduced to $N = 688$, 344, and 172. The values for the predictive abilities were still high up to $N = 344$, below this limit the predictive abilities dropped substantially for both traits (Figure 20). A similar slope was observed for the two calibration sets in Maize 2 (Figure 22 and 23). As shown by Lorenzana and Bernardo (2009) with biparental plant populations, the limits of the population size were about 100-200 for highly heritable traits in maize. Results from this study with multiple crosses indicate that the critical limit for the size of the training population is in the order of 400 DH lines.

The optimal marker density used for genomic prediction depends on the sample size and genetic structure observed in a breeding population. The resulting structure and extent of LD influences the efficiency and success of genome-wide prediction. Observed LD within a European elite maize breeding population is mainly caused by relatedness, population stratification, and genetic drift (Stich et al. 2005). In Maize 1, containing 1377 DH lines out of an advanced cycle breeding population, the level of LD was high even between marker pairs not located on the same chromosome. This might be an effect of admixture of different allele frequencies coming from different populations and the small effective population sizes expected for a maize breeding population. With this high level of LD, a low marker density as used in Maize 1 seems to already capture the necessary information required for genomic selection and only a small additional gain of predictive ability was expected with a higher marker density. The results obtained with the high density marker panel confirmed this assumption (Figure 21), as increasing the marker density led only to a small increase in predictive abilities for CV-W. In contrast, higher marker densities were more efficient than the lower marker panel for the prediction of less related material in CV-A. These findings are corroborated by the study of Habier et al. (2010) who observed that accuracies between unrelated individuals are decreasing, if LD is only based on selection rather than historic mutations and linkage.

The allocation of field resources to optimize breeding schemes is another crucial point

for the implementation of genomic prediction. Therefore, the influence of the number of locations on predictive ability was additionally analyzed. In general, Burgueño et al. (2011) and Guo et al. (2013) observed a gain in predictive ability if a multi-environment analysis was applied instead of a single-environment analysis for the prediction of newly developed genotypes. Similar results could be observed from this study, using half of the available number of locations reduced predictive abilities for grain yield to the same amount compared to using only half of the DH lines evaluated in all locations for model training (Table 9). In contrast for GDC, reducing the sample size had a larger effect on predictive abilities than reducing the number of location. Therefore, when the number of plots available for the field trials is limited, it might be advisable to reduce the number of locations before discarding DH lines for model training.

4.2.3 Optimizing the population for model training

In commercial maize breeding, new improved elite material is recombined in successive cycles in recurrent selection schemes (Gordillo and Geiger 2008). Unrelated genetic material from other breeding populations is introgressed into the core germplasm to maintain the genetic variability within the breeding pool. At the beginning of a breeding cycle, genetic groups will occur where the relatedness within groups is higher than between groups, which was observed for Maize 2 in this study. As pointed out by Habier et al. (2010), it is useful to select DH lines, which are closely related to the non-phenotyped testcrosses to form the training population. Therefore, an estimation of testcross effects should include the most recent field data from a related training population. On the other hand, to maintain long term selection gain, Rincent et al. (2012) recommended capturing high genetic variation within the training population by including a diverse set of parents for model training. Hence, the optimization of the training population for GP in maize breeding schemes is still an open question.

In Maize 2, predictive abilities within and across genetic groups were influenced by the kinship between groups and the size of the ES. The most important factor influencing predictive abilities across groups was the average and maximum kinship between these groups, which has also been described by Habier et al. (2010) and Saatchi et al. (2011) for cattle breeding populations. Predictive abilities for GBLUP across group/tester sub-

sets in CS2 of Maize 2 were highly dependent on the maximum kinship captured between estimation and test set when fixing the sample size of the ES in CV (Figure 26). However, when new genetic material is integrated into an existing breeding population, there is generally a strong imbalance in the number of lines derived from adapted and new genetic material and inferences on the association of the degree of relatedness and predictive ability are not as straightforward.

While adding progenies of unrelated material does not change the maximum kinship between ES and TS, a reduction of the predictive ability of GBLUP for GDY was observed for some scenarios. Especially for group G1, which represents the core germplasm of the calibration sets in Maize 2, predictive abilities did not increase although material of the other groups was added to the ES. These results confirm similar conclusions by Riedelsheimer et al. (2013), where the prediction accuracy in multi-parental crosses of maize could not be improved with including unrelated families to the training population. Thus, when new genetic material is introgressed into the breeding population, predictions within the main genetic group might be advisable until a higher connectivity between groups is reached by recombination.

However, results obtained in Maize 2 indicate that increasing the ES size by adding unrelated material to an existing ES of small sample size improves predictive abilities of GBLUP, which was observed for G2 in CS1 and in CS2 (Table 5). Also for GDC, predictive abilities slightly increased or were not affected when the ES included lines from all genetic groups ($ES_{G1,G2,G3}$). Similar results were observed in animal breeding (Erbe et al. 2012), where the highest prediction accuracies were observed, if the reference set pooled multiple breeds to predict a TS including Jersey bulls only. Hayes et al. (2009a) argued that SNPs capturing effects across multiple breeds must be adjacent to the potential QTL, because they are in high LD across all breeds. Therefore, predictions might be more persistent over generations when multiple groups are included in the reference population.

4.2.4 Accounting for genetic substructures

In Maize 2, CS1 is characteristic for an early cycle from a maize breeding program showing extensive genetic substructure, CS2 represents a cycle where genetic groups can not be distinguished. In both calibration sets, the core set of germplasm in the breeding populations is formed by genetic group G1 with the highest performance level and high average kinship between DH lines.

Accounting for population structure has mainly been discussed in the context of genome-wide association studies to correct for spurious associations due to admixture (Astle and Balding 2009). In the context of GP, Windhausen et al. (2012) observed that prediction accuracies are overestimated due to different phenotypic performance levels across groups. A correction based on principal components was recently proposed by Guo et al. (2014) and illustrated that predictive abilities in structured breeding populations are biased. The same effect was observed in Maize 2. When the substructure was not included as fixed effects in the prediction model within CS1, predictive abilities increased from r_{TS}=0.59 to 0.73 for GDY (Appendix, Table A7), because markers captured the variation across groups in addition to the variation within groups. To predict the genetic potential for grain yield of an untested DH line, the focus of the selection is on the genetic variation within groups and the estimation of testcross effects should be independent from the variation across genetic groups.

Results from this study emphasize the importance of taking substructure into account. If the substructure within a data set is unknown, marker data can be used to assess population structures, e.g., with cluster analyses (Odong et al. 2011; Saatchi et al. 2011; Heslot et al. 2012). The information from a principal component analysis can be used to illustrate genetic groups and correct for their effects when estimating predictive abilities (Guo et al. 2014). In Maize 2, both calibration sets were analyzed with a cluster and principal component analysis. All clustering methods applied to CS1 were able detect main substructures in this data set. Main clusters detected in CS1 with the different cluster analyses could be distinguished in the space of the first two PCs. However, a prediction within clusters derived from the different clustering methods resulted only in slightly better predictive abilities as a random sample of CS1 (Appendix, Figure A6). Furthermore, the prediction within group G1 outperformed predictions within the largest

clusters derived from all clustering methods. In CS2, less variability was explained by the PCs and the optimum number of clusters could not be clearly identified confirming that the substructure in CS2 was reduced compared to CS1.

Results obtained within each calibration set of Maize 2 can only partly reflect the effect of genetic substructure in prediction of an independent sample. Only few studies so far have validated their results from CV in an independent sample. As observed by Hofheinz et al. (2012) for sugar beet and Windhausen et al. (2012) for maize, results obtained with CV within a population can be overestimated compared to a validation on lines from another related population. Both concluded that high predictive abilities evaluated within the same population do not necessarily lead to good predictions in a different validation set as LD and linkage phases might change across populations. Therefore, results obtained from CV-R in CS1 of Maize 2 were validated on CS2 and vice versa (Figure 32). As expected, predictive abilities of the validation decreased compared to the results obtained with CV-R. Due to the higher extent of genetic substructure in CS1, correlations within CS1 obtained with CS2 as training set were higher than predictions obtained with CS1 as training set. However, the overall correlations were substantially inflated ($r_{CS1} = 0.58$ for GDY) or deflated ($r_{CS2} = 0.51$ for GDC) because the across group variation was not captured in the prediction across years. The correlations between observed and predicted testcross values within the respective group/tester subsets captured by the within group variation were of the same magnitude for both calibration sets. Improvements could be observed, when the VS1 in 2011 was predicted with the training sets capturing the within group variation of CS1 and the year interaction from CS2.

To avoid biased predictive abilities due to differences in mean performance across genetic groups, the observed testcross performance should be adjusted in advance by using estimates of fixed effects. However, if the fixed effects in the validation sample cannot be estimated within the training population, for example due to different group/tester combinations in both calibration sets as observed for Maize 2, these differences need to be known *a priori*. Otherwise, results should be handled with caution and might be difficult to interpret (de los Campos et al. 2013).

4.2.5 Multi-trait predictions

In computer simulations, multi-trait GP models are very promising to be advantageous if traits differ substantially in heritability or if data on one trait are incomplete (Jia and Jannink 2012). In European maize breeding, the two most important traits are yield and maturity and their negative correlation is undesirable. Therefore, it is important to establish to which extent predictive abilities in GDY can be influenced by differences in GDC. In Maize 2, results show that depending on the genetic group and the tester, marker effects predicted for GDC could effectively predict GDY, making it difficult to break up the negative correlation between the two traits in selection (Figure 24). On the other hand, the tight association between the two traits could be used in multi-trait prediction models to enhance the prediction of GDY due to the higher and more stable prediction accuracies obtained for GDC. However, results from this study show that genetic substructure plays an important role for GP and that the genetic correlation between GDY and GDC strongly varied for the different group/tester combinations. In addition, it is hard to rule out the possibility that the different weather conditions experienced in the two years of evaluation affected the association between GDY and GDC in different ways. Thus, modeling the genetic variance-covariance between traits is exceedingly challenging for real life experimental data. It remains to be seen if employing a multi-trait prediction approach can simultaneously improve the predictive accuracies of both traits.

4.2.6 Prediction across testers

Another important question for breeders is the comparison of predictive abilities across testers, as multiple testers are often used for producing testcrosses. The genetic correlation (r_g) between testers can be used to estimate the accordance between testers. In Maize 2, DH lines were crossed to only one of two testers in each calibration set and the genetic correlation r_g could not be directly assessed. The extent of r_g depends on the general and specific combining ability, which cannot be separated when testcrosses were produced with only one inbred or single-cross tester (see Section 2.4). However, correlations between testers are expected to be medium to high as in hybrid maize breeding, heterotic pools emerged a long time ago which reduced the influence of specific compared to general combining ability (Reif et al. 2007). From the literature, r_g between

testers of greater than 0.6 for GDY and GDC have been observed for testcrosses derived from biparental families with two different testers (Melchinger et al. 1998). Concerning predictive abilities obtained from CV-R within testers in CS1 and the fact that the testers were mainly crossed to different genetic groups, the observed predictive abilities obtained from CV across testers were higher than expected with $r_g = 0.6$ as assumed by Windhausen et al. (2012) (Table 6). In CS2, predictive abilities across testers were generally lower as expected although both single-cross testers T1 and T3 shared a common parent. However in this data set, the DH lines were not randomly assigned to the tester subsets in both calibration sets and this effect cannot be separated from the specific combining ability of each tester. Thus, the choice of the tester can have a strong effect on predictive abilities and the exploitation of specific combining ability for GP needs further attention.

4.2.7 Prediction across locations and years

The prediction of untested DH lines in a different set of locations and years is of importance for any breeder. The evaluation of DH lines in multiple years is time and cost consuming. Therefore, DH lines are generally evaluated in multiple locations representing the target environment to capture genotype by environment interactions within one year only. Using the subset of tester T1 from CS1 of Maize 2, predictive abilities across locations were compared to validation in the next year to assess whether prediction across locations can reflect genotype by environment interactions necessary for the prediction across years. All DH lines belonging to this subset were evaluated in the same four locations and tester T1 was crossed to subsets of groups G1 and G3 in both calibration sets.

The results for the prediction across locations were highly dependent on the selected locations for the ES. While the prediction of untested DH lines within a set of locations was dependent on the heritability obtained from the selected locations (Figure 30), the prediction of testcross performance in a different location was dependent on the correlation between these locations (Figure 31). Best predictions of genotypes within the TS_{loc} and $TS_{g\times loc}$ were obtained when the ES included locations which represented the whole range of the target environment and were therefore highly correlated with the en-

vironment in the TS. Similar results were observed by Burgueño et al. (2011), i.e., predictive abilities improved when not only highly correlated environments were selected for the ES. Hence, environments for calibration of prediction models should be highly heritable but still covering a wide range of genotype by location interactions to obtain a good prediction across locations and genotypes. Comparing the predictions across locations with predictions across years, large differences could be observed between the two traits. For GDY, mean predictive abilities obtained in $TS_{g\times loc}$ were of the same magnitude as the correlation with predicted testcross values from CS1 and observations from CS2 ($r_{G1/T1} = 0.36$). In contrast for GDC, predictions across locations and genotypes within one year overestimated predictive abilities compared to validation across years ($r_{G1/T1} = 0.61$). This might be an effect of the different weather conditions in both years. In 2011, weather conditions were more favorable for maize production than in 2010 influencing the level of maturity reached at harvest. In addition, these results are comparable with previous QTL studies. As observed by Utz et al. (2000) for traits like GDY and GDC, the proportion of genetic variance explained by the QTL was always higher for the validation in $TS_{g\times loc}$ than in an independent validation sample, although the power of QTL detection was reduced due to the lower number of observations in the ES compared to analyses with the complete data set. However, predictions of GDC for G1/T1 in CS2 using the complete CS1 data set yielded markedly higher predictive abilities ($r_{G1/T1} = 0.71$, Figure 32) which were comparable to those obtained with the predictions across locations. To obtain a more complete picture on the effects of genotype by environment interactions on predictive abilities and the usefulness of accounting for them in GP will require more than two calibration sets. Connections between years by common check cultivars might increase the potential to fully capture genotype by environment interactions. Further improvements for the predictions across environments might also be achieved with modeling the genotype by environment interactions within the GP model as suggested by Burgueño et al. (2011) and Guo et al. (2013).

5 Conclusions

High predictive abilities obtained with different cross-validation and validation scenarios are promising for the implementation of genomic prediction into maize breeding programs. The data sets presented in this study exhibited different degrees of genetic substructure and it was demonstrated that due to the complexity of the data sets, the implementation of GP is not straightforward.

Prediction models should account for data heterogeneity, different testers and genotype by environment interactions. In general, predictive abilities were highly dependent on the relatedness between estimation and test sets. When data heterogeneity is high and the connection between genetic groups is low, a prediction within groups can outperform a prediction including multiple groups. However, when across group variation is reduced by recombinations, information for the predictions can be gained from related testcrosses even from different genetic groups or families. Observed correlations between testers were lower than expected. To increase the information gain for the calibration of prediction models, the connectivity between different testers should be increased by common check cultivars crossed to all available testers used for producing the testcrosses. Predictions across years were highly influenced by the across group variation and predictive abilities were over- as well as underestimated. For prediction across years, information will be gained from related lines evaluated in several years to capture genotype by environment interactions. By including correlated environments representing the full spectrum of the target environment and modeling genotype by environment interactions in the prediction model, predictive abilities could be increased.

This study gives valuable insights into the re-allocation of resources for maize breeding to effectively implement GP and shows that not only the size of the population for model training but also the genetic variation captured by the population is of utmost importance.

6 Summary

Genomic prediction is a newly developed method to predict the genetic potential of untested individuals based on their genotypic profile. This method has been successfully implemented into dairy breeding, where phenotypic performance of a bull cannot be directly assessed and the evaluation of daughters is time and cost consuming. With decreasing genotyping costs, genomic prediction has also become of interest for plant breeders, where multi-environmental trials are necessary to assess the performance of newly developed lines. First results on the genome-based prediction of testcross performance from this study are encouraging for the implementation of genomic prediction into maize breeding programs. However, the implementation of genomic prediction in plant breeding is not straightforward and more research is needed for the prediction within structured breeding material or the prediction across testers and years.

The experimental data sets used in this study were derived from two commercial maize breeding programs. The first data set comprised 1377 doubled haploid (DH) lines evaluated as testcrosses with one tester in seven locations. All DH lines were genotyped with a low-density array including 1152 biallelic single nucleotide polymorphism (SNP) markers, a subset of the DH lines was additionally genotyped with a high-density SNP array comprising 56110 markers. The second data set was composed of two calibration sets derived from subsequent breeding cycles comprising 1073 and 857 DH lines. Both calibration sets included germplasm from three genetic groups and were evaluated as testcrosses with three different testers in four locations and two consecutive years. Genotyping was performed with the high-density SNP array (56110 markers). Selected DH lines of the first calibration set were additionally evaluated in the second year. In all data sets, the traits grain dry matter yield and content were assessed.

All data sets were characteristic for maize breeding populations exhibiting different degrees of genetic substructure and distinct family and tester composition. To obtain an estimate of the relative performance of genome-based compared to pedigree-based prediction of testcross performance, models employed for the prediction of testcross performance differed in how the kinship between DH lines was modeled based on pedigree and marker data. Different cross-validation and validation procedures were applied to the data sets to assess the genome-based prediction performance within and across families,

genetic groups, testers, locations, and years.

It was demonstrated that genome-based predictions outperformed pedigree-based predictions in plant breeding populations. Different methods for estimating genome-based kinship coefficients between DH lines resulted in similar or even identical predictive abilities. The largest influence on prediction performance was observed for genetic substructure in the data sets. In most cases, mean predictive abilities across families or subsets of genetic groups and testers were lower than predictive abilities within families or subsets, due to the reduced relatedness among DH lines in the estimation and test sets. Genome-based prediction across years achieved promising results but predictive abilities were reduced compared to the prediction within a year. The optimal calibration population should represent the full genetic variation of the breeding population. Locations selected for evaluating this population should represent the complete target environment. However, accounting for data heterogeneity, different testers, and genotype by environment interactions should be improved with increasing the connectedness between calibration and validation population by including common test units.

7 Zusammenfassung

In der Pflanzenzüchtung beruht die Evaluierung von ungeprüften Selektionskandidaten bisher auf ressourcenintesiven, mehrortigen Feldversuchen. Die Vorhersage des genetischen Wertes einer Linie basierend auf ihrem DNA Profil ist hingegen eine neue und vielversprechende Methode. In der Rinderzüchtung wurde diese Methode bereits erfolgreich etabliert, da die bisherige Zuchtwertschätzung von Bullen eine teure und langwierige Prüfung der Bullentöchter beinhaltete. Mit der stetigen Abnahme der Genotypisierungskosten ist die genombasierte Vorhersage auch für die Pflanzenzüchtung in greifbare Nähe gerückt. Erste Ergebnisse aus dieser Studie sind vielversprechend für die Implementierung der genomischen Vorhersage in Maiszuchtprogrammen. Jedoch ist die genaue Kenntnis des Einflusses von Populationsstrukturen auf die Vorhersage über Populationen sowie über Tester und über Jahre notwendig.

Die Datensätze für diese Studie stammen aus zwei Zuchtprogrammen. Der erste Datensatz bestand aus 1377 Doppelhaploiden (DH)-Linien, welche als Testkreuzungen mit einem gemeinsamen Tester an sieben Standorten geprüft wurden. Alle DH-Linien wurden mit 1152 biallelen Einzelnukleotid-Polymorphismus (*engl.* single nucleotide polymorphism, SNP)-Markern genotypisiert, ein Teil der Linien wurde zusätzlich noch auf einer Hochdurchsatzplattform mit 56110 SNP-Markern genotypisiert. Der zweite Datensatz enthielt zwei Kalibrierungspopulationen mit 1073 und 857 DH-Linien aus zwei aufeinanderfolgenden Zyklen desselben Zuchtprogramms. Beide Populationen wurden als Testkreuzungen mit drei unterschiedlichen Testern an vier Standorten in 2010 beziehungsweise 2011 geprüft. Die Genotypisierung erfolgte ebenfalls mit der 50k SNP-Plattform. Aus der ersten Kalibrierungspopulation wurden einige DH-Linien selektiert und als Validierungspopulation erneut im zweiten Jahr geprüft. In allen Datensätzen wurden die Merkmale Korntrockenmasseertrag und -gehalt bestimmt.

Alle Datensätze waren charakteristisch für europäische Maiszuchtprogramme und spiegelten unterschiedliche Grade an Familien-, Tester- und Populationsstrukturen wider. Für die Vorhersage der Testkreuzungsleistung wurden verschiedene statistische Modelle etabliert, welche sich in der Modellierung der Verwandtschaft zwischen den Linien, basierend auf Marker- und Abstammungsdaten, unterschieden, um die relative Effizienz der genomischen zur stammbaumbasierten Vorhersage abschätzen zu können. Um die Vorher-

sageleistung innerhalb und zwischen Familien, Gruppen, Testern, Orten und Jahren ermitteln zu können wurden unterschiedliche Kreuzvalidierungs- und Validierungsszenarien auf die Datensätze angewendet.

Wie erwartet war die genomische Vorhersage der Testkreuzungsleistung dem Modell basierend auf den Abstammungsdaten überlegen, während die verschiedenen Methoden zur Schätzung der genomischen Verwandtschaft zwischen den Linien nur zu geringen Unterschieden in den Vorhersagegenauigkeit führten. Den größten Einfluss auf die Vorhersageleistung hatte die genetische Struktur innerhalb der Kalibrierungspopulationen. Die Vorhersage innerhalb der Familien, Gruppen und Testerteildatensätze funktionierte meist deutlich besser als die Vorhersage zwischen Familien und Gruppen, da die Verwandtschaft zwischen den DH-Linien im Schätz- und Testset im zweiten Fall deutlich geringer war. Obwohl die Vorhersage über Jahre stark von der Substruktur beeinflusst wurde, waren die Vorhersagegenauigkeiten hoch, aber geringer als innerhalb der Jahre. Aus den Ergebnissen lässt sich schließen, dass die optimale Kalibrierungspopulation das gesamte genetische Spektrum des Zuchtprogramms widerspiegeln sollte. Des Weiteren sollten die ausgewählten Umwelten zur Phänotypisierung der Kalibrierungspopulation möglichst gut der Zielumwelt entsprechen. Um die Heterogenität in den Datensätzen zu minimieren und somit den Einfluss der Gruppen und Tester, sowie der Genotyp-Umwelt-Interaktion zu reduzieren, sollte anhand von gemeinsamen Prüfgliedern zwischen den Testern und Jahren die Verbindung zwischen Kalibierungs- und Validierungspopulation optimiert werden.

8 References

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

Astle W, and Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451–471

Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet 85:1055–1062

Bernardo R (2002) Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN, USA

Breiman L, and Spector P (1992) Submodel selection and evaluation in regression. The X-random case. Int Stat Rev 60:291–319

Browning BL, and Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for Large Data Sets of Trios and unrelated individuals. Am J Hum Genet 84:210–223

Burgueño J, Crossa J, Cotes JM, Vicente FS, and Das B (2011) Prediction assessment of linear mixed models for multienvironment trials. Crop Sci 51:944–954

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, and Calus MPL (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193:327–345

Crossa J, Campos Gdl, Pérez P, Gianola D, Burgeno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, and Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

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

Daetwyler HD, Calus MPL, Pong-Wong R, de Los Campos G, and Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124:331–341

DMK (2011) Fakten - Statistik - Bedeutung des Maisanbaues in Deutschland - Flächenerträge von Körnermais und Silomais in Deutschland - Deutsches Maiskomitee e.V. (DMK). http://www.maiskomitee.de, accessed: January 27th, 2013

Endelman JB, and Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3 2:1405–1413

Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, Mason B, and Goddard M (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95:4114–4129

Falconer DS, and Mackay TFC (1996) Introduction to quantitative genetics. Longman, Harlow, UK

FAO (2010) Food and Agriculture Organization of the United Nations - FAOSTAT. http://faostat3.fao.org, accessed: January 27th, 2013

Fernando R, and Grossman M (1989) Marker assisted selection using best linear unbiased prediction. Genet Sel Evol 21:467

Foulkes AS (2009) Applied Statistical Genetics with R: For Population-based Association Studies. Springer, New York, USA

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schön CC, Sun Q, Walter H, Martin OC, and Falque M (2011) A large maize (*Zea mays* L.) SNP

genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS One 6:e28334

Gilmour AR, Gogel BJ, Cullis BR, Thompson R, Butler D, Cherry M, Collins D, Dutkowski G, Harding SA, and Haskard K (2009) ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK

Goddard ME, and Hayes B (2007) Genomic selection. J Anim Breed Genet 124:323–330

Goddard ME (2008) Genomic selection: prediction of accuracy and maximization of long term response. Genetica 136:245–257

Gordillo GA, and Geiger HH (2008) Alternative recurrent selection strategies using doubled haploid lines in hybrid maize breeding. Crop Sci 48:911–922

Grubbs FE (1950) Sample criteria for testing outlying observations. Annal Math Stat 21:27–58

Guo Z, Tucker DM, Wang D, Basten CJ, Ersoz E, Briggs WH, Lu J, Li M, and Gay G (2013) Accuracy of across-environment genome-wide prediction in maize nested association mapping populations. G3 3:263–272

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, and Gay G (2014) The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127:749–762

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

Habier D, Tetens J, Seefried FR, Lichtner P, and Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:5

Hallauer AR, and Miranda JB (1985) Quantitative genetics in maize breeding. Iowa State Univ. Press, Ames, IA, USA

Hartigan JA, and Wong MA (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Appl Stat 28:100–108

Hayes BJ, and Goddard ME (2008) Technical note: prediction of breeding values using marker-derived relationship matrices. J Anim Sci 86:2089–2092

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, and Goddard ME (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51

Hayes BJ, Visscher PM, and Goddard ME (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91:47–60

Heffner EL, Sorrells ME, and Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Henderson CR (1977) Best linear unbiased prediction of breeding values not in the model for records. J Dairy Sci 60:783–787

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph

Heslot N, Yang HP, Sorrells ME, and Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160

Hill WG, and Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

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

Holland JB, Nyquist WE, and Cervantes-Martínez CT (2003) Estimating and interpreting heritability for plant breeding: an update. In Plant breeding reviews page 9–111 John Wiley & Sons, Inc., Oxford, UK

Jannink JL, Lorenz AJ, and Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166–177

Jia Y, and Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513–1522

Kaufman L, and Rousseeuw PJ (2005) Finding groups in data: an introduction to cluster analysis. Wiley, Hoboken, NJ, USA

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Joint Con Artif 14:1137–1145

Kurtz AK (1948) A research test of the Rorschach test. Pers Psych 1:41–51

Legarra A, Robert-Granie C, Manfredi E, and Elsen JM (2008) Performance of genomic selection in mice. Genetics 180:611–618

Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, Flament P, Melchinger AE, Menz M, Meyer N, Moreau L, Moreno-Gonzalez J, Ouzunova M, Pausch H, Ranc N, Schipprack W, Schönleben M, Walter H, Charcosset A, and Schön CC (2014) Usefulness of multi-parental populations of maize (*Zea mays* L.) for genome-based prediction of testcross performance. Genetics 198:3–16

Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, and Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. Stat Appl Genet Mol Biol 12:3750–391

Lorenzana RE, and Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, and Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci USA 99:6080–6084

Melchinger AE, Utz HF, and Schön CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics 149:383–403

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

Mosier CI (1951) I. Problems and designs of cross-validation. Educ Psychol Measurement 11:5–11

Möhring J, and Piepho HP (2009) Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci 49:1977–1988

Ober U, Erbe M, Long N, Porcu E, Schlather M, and Simianer H (2011) Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 188:695–708

Odong TL, Heerwaarden J, Jansen J, Hintum TJL, and Eeuwijk FA (2011) Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet 123:195–205

Ouyang Z, Mowers RP, Jensen A, Wang S, and Zheng S (1995) Cluster analysis for genotype × environment interaction with unbalanced data. Crop Sci 35:1300–1305

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165–1176

Piepho HP, Büchse A, and Emrich K (2003) A Hitchhiker's guide to mixed models for randomized experiments. J Agron Crop Sci 189:310–322

Piepho HP, Williams ER, and Fleck M (2006) A note on the analysis of designed experiments with complex treatment structure. Hort Sci 41:446–452

Powell JE, Visscher PM, and Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet 11:800–805

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81:559–575

Reif JC, Melchinger AE, and 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, Gumpert FM, Fischer S, and Melchinger AE (2007) Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176:1931–1934

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

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, and Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503

Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schön CC, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, and Moreau L (2012) Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (*Zea mays* L.). Genetics 192:715–728

Rogers J (1972) Measures of genetic similarity and genetic distance. In Studies in genetics VII p. 145–153, Univ. of Texas, Austin, TX, USA

Röber F, Gordillo G, and Geiger H (2005) In vivo haploid induction in maize – Performance of new inducers and significance of doubled haploid lines in hybrid breeding. Maydica 50:275–284

Saatchi M, McClure MC, McKay SD, Rolf MM, Kim JW, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, et al. (2011) Accuracies of genomic breeding values in American Angus beef cattle using k-means clustering for cross-validation. Genet Sel Evol 43:40

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123:218–223

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schölkopf B, Smola A, and Müller KR (1998) Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Comput 10:1299–1319

Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, and 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

Scutari M, Mackay I, and Balding D (2013) Improving the efficiency of genomic selection. Stat Appl Genet Mol Biol 12:517–527

Shull GH (1908) The composition of a field of maize. J Hered 4:296–301

Shull GH (1909) A pure-line method in corn breeding. J Hered 5:51–58

Sneath PHA, and Sokal RR (1973) Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman, San Francisco, CA, USA

Sokal RR, and Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409–1438

Stich B, Melchinger AE, Frisch M, Maurer HP, Heckenberger M, and Reif JC (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111:723–730

Stich B, Mohring J, Piepho HP, Heckenberger M, Buckler ES, and Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754

Stram DO, and Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171–1177

Technow F, Riedelsheimer C, Schrag TA, and Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:1181–1194

Technow F, Bürger A, and Melchinger AE (2013) Genomic prediction of Northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 3:197–203

Utz HF, Melchinger AE, and Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from

experimental data in maize using cross validation and validation with independent samples. Genetics 154:1839–1849

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, and Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Ass 58:236–244

Whittaker JC, Thompson R, and Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252

Wimmer V, Albrecht T, Auinger HJ, and Schön CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086–2087

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

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer C, and Melchinger AE (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2:1427–1436

Yang J, Lee SH, Goddard ME, and Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82

Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, and Zhang Q (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5:e12648

Zhao Y, Gowda M, Liu W, Würschum T, Maurer H, Longin F, Ranc N, and Reif J (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776

Zhong S, Dekkers JCM, Fernando RL, and Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364

9 Appendix

Transformation of variance components

As described in Section 2.5.2, the GBLUP model with the realized kinship according to Habier et al. (2007) and VanRaden (2008) can be transformed to the RRBLUP model. The transformation of variance components can be derived from the following equations under the assumption that the variance-covariance structure of **y** is equal in both models.

The variance components for GDY were estimated with both models in tester T1 subset of Maize 2 (Table A1). The realized kinship coefficients were calculated from the genotype matrix **W** with the following formula in matrix notation:

$$\mathbf{U} = \frac{(\mathbf{W} - \mathbf{P})(\mathbf{W} - \mathbf{P})'}{2 \cdot \sum_{m=1}^{M} 2p_m(1 - p_m)}$$

Table A1: Variance components and mean for GDY estimated with GBLUP and RRBLUP in the subset of tester T1 in calibration set CS1 of Maize 2.

	$\hat{\sigma}^2_u$	$\hat{\sigma}^2_{u_s}$	$\hat{\sigma}^2_m$	$\hat{\sigma}^2$	μ
GBLUP($\mathbf{U}$)	21.73			12.96	129.74
GBLUP($\mathbf{U_S}$)		35.74		12.96	124.46
RRBLUP			0.00329	12.96	131.89

$$
\begin{aligned}
Var(y) = \mathbf{V_{GBLUP}} &= \mathbf{V_{RRBLUP}} \\
\mathbf{ZUZ}'\hat{\sigma}_u^2 + \mathbf{I}\hat{\sigma}^2 &= \mathbf{WIW}'\hat{\sigma}_m^2 + \mathbf{I}\hat{\sigma}^2 \\
\mathbf{ZUZ}'\hat{\sigma}_u^2 &= \mathbf{WIW}'\hat{\sigma}_m^2 \\
\mathbf{Z}\frac{(\mathbf{W}-\mathbf{P})(\mathbf{W}-\mathbf{P})'}{2\cdot\sum_{m=1}^{M}2p_m(1-p_m)}\mathbf{Z}'\hat{\sigma}_u^2 &= \mathbf{WW}'\hat{\sigma}_m^2 \quad \text{with} \quad \mathbf{Z}=\mathbf{I} \\
\frac{(\mathbf{W}-\mathbf{P})(\mathbf{W}-\mathbf{P})'}{2\cdot\sum_{m=1}^{M}2p_m(1-p_m)}\hat{\sigma}_u^2 &= \mathbf{WW}'\hat{\sigma}_m^2 \\
\hat{\sigma}_u^2 &= 2\cdot\sum_{m=1}^{M}2p_m(1-p_m)\cdot\hat{\sigma}_m^2 \\
21.73 &\approx 0.00329\cdot 6608.51 = 21.74
\end{aligned}
$$

The variance components are unaffected by **P**, since the same column vector **p** is subtracted from each vector of marker genotypes **m** and the slope of the regression of the phenotype on each SNP is not changed. The term **P** is constant and captured by the intercept as described in the next section. The same transformation can be derived for the comparison of variance components of GBLUP with $\mathbf{U_S}$ and RRBLUP, where $\hat{\sigma}^2_{u_s} = 2M(1-s_{min})\cdot\hat{\sigma}^2_m \approx 10948\cdot 0.00329$.

Transformation of the intercept

The following equations describe how the intercept in the RRBLUP model is shifted by a constant, when the term **P** is subtracted from the genotype matrix **W**. If RRBLUP is estimated with a column-centered genotype matrix, i.e., $\mathbf{W} - \mathbf{P}$, the estimated intercept equals that of GBLUP, while the marker variances are not changed as described in the previous section. Therefore, the modified RRBLUP can be rewritten as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + (\mathbf{W} - \mathbf{P})\mathbf{m} + \mathbf{e} \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{m} - \mathbf{P}\mathbf{m} + \mathbf{e}. \end{aligned}$$

As **Pm** is constant for each predicted testcross value, the term captures the difference of the intercepts, i.e., $\mathbf{X}\boldsymbol{\beta}$ in that case, and can be expressed as

$$\begin{aligned} \mu_{GBLUP} - \mu_{RRBLUP} &= \sum_{m=1}^{M} (\mathbf{P} \cdot \hat{\mathbf{m}}) \\ 129.74 - 131.89 &\approx -2.155 \end{aligned}$$

Additional Tables and Figures

Table A2: Variance components and repeatability (rep^2) from the first stage of phenotypic analysis for grain yield (GDY) and grain dry matter content (GDC) for each location in 2010 of Maize 2.

		Variance components					
Trait	Location	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_{trial}$	$\hat{\sigma}^2_{rep}$	$\hat{\sigma}^2_{block}$	$\hat{\sigma}^2_e$	$\hat{rep}^2$
GDY	Loc 1	44.93	9.68	3.54	1.61	45.02	0.50
	Loc 2	24.13	5.24	2.40	8.90	32.09	0.43
	Loc 3	35.42	36.53	7.12	4.63	35.39	0.50
	Loc 4	34.44	7.03	3.49	5.31	49.44	0.41
	Loc 5	75.03	2.22	0.68	0.00	48.62	0.61
	Loc 6	23.31	15.93	6.38	10.64	62.43	0.27
GDC	Loc 1	1.48	0.38	0.11	0.10	0.24	0.86
	Loc 2	0.96	0.05	0.06	0.14	0.24	0.80
	Loc 3	0.92	0.27	0.01	0.02	0.17	0.84
	Loc 4	0.99	0.25	0.04	0.08	0.22	0.82
	Loc 5	1.44	0.56	0.00	0.03	0.16	0.90
	Loc 6	1.57	0.02	0.03	0.06	0.38	0.81

Table A3: Variance components and repeatability (rep^2) from the first stage of phenotypic analysis for grain yield (GDY) and grain dry matter content (GDC) for each location in 2011 of Maize 2.

		Variance components					
Trait	Location	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_{trial}$	$\hat{\sigma}^2_{rep}$	$\hat{\sigma}^2_{block}$	$\hat{\sigma}^2_e$	$\hat{rep}^2$
GDY	Loc 1	58.52	40.61	3.54	6.29	58.73	0.50
	Loc 2	56.63	2.83	2.82	3.96	36.81	0.61
	Loc 3	60.02	2.99	0.04	0.07	41.22	0.59
	Loc 4	54.64	0.00	22.43	8.19	45.05	0.55
	Loc 7	79.44	66.86	16.16	33.92	69.51	0.53
	Loc 8	85.01	18.87	0.00	6.27	44.04	0.66
GDC	Loc 1	1.90	3.16	0.07	0.08	0.30	0.86
	Loc 2	1.02	0.09	0.08	0.08	0.21	0.83
	Loc 3	1.04	0.44	0.01	0.05	0.15	0.88
	Loc 4	2.54	0.33	0.29	0.13	0.37	0.87
	Loc 7	2.15	0.43	0.31	0.29	0.43	0.83
	Loc 8	2.29	0.90	0.05	0.10	0.17	0.93

Table A4: Variance components and heritability (h^2) from the prediction models PBLUP, GBLUP and P+GBLUP for grain yield (GDY) and grain dry matter content (GDC) of Maize 1.

Trait	Model	$\hat{\sigma}^2_t$	$\hat{\sigma}^2_u$	$\hat{\sigma}^2_e$	$\hat{h}^2$	LogL	AIC
GDY	PBLUP	79.36 ± 19.49		14.86 ± 10.3	0.84	-3499.69	7003
	GBLUP		44.22 ± 5.27	34.92 ± 1.56		-3343.46	6691
	P+GBLUP	36.92 ± 12.14	35.20 ± 4.73	14.37 ± 6.45		-3318.45	6643
GDC	PBLUP	0.862 ± 0.22		0.152 ± 0.12	0.85	-376.53	757
	GBLUP		0.690 ± 0.075	0.284 ± 0.01		-123.12	250
	P+GBLUP	0.213 ± 0.09	0.641 ± 0.072	0.167 ± 0.05		-113.39	233

Table A5: Averages of adjusted entry means of grain yield (GDY) and grain dry matter content (GDC) within tester subsets and groups of calibration sets CS1 and CS2 of Maize 2.

	Trait	Tester	Group			
			G1	G2	G3	Mean
CS1	GDY	T1	129.81 ± 0.22[a]		126.37 ± 1.12	129.74 ± 0.22
		T2		118.47 ± 0.44	122.74 ± 0.47	121.09 ± 0.32
		Mean	129.81 ± 0.22	118.47 ± 0.44	122.98 ± 0.42	126.71 ± 0.18
	GDC	T1	69.00 ± 0.04		68.92 ± 0.17	68.99 ± 0.04
		T2		69.73 ± 0.09	67.83 ± 0.08	68.57 ± 0.06
		Mean	69.00 ± 0.04	69.73 ± 0.09	67.90 ± 0.08	68.84 ± 0.03
CS2	GDY	T1	142.61 ± 0.52		140.96 ± 0.72	141.91 ± 0.43
		T3	145.83 ± 0.35	134.95 ± 1.60	147.00 ± 0.66	145.79 ± 0.30
		Mean	144.79 ± 0.29	134.95 ± 1.60	143.79 ± 0.49	144.31 ± 0.25
	GDC	T1	73.44 ± 0.11		72.26 ± 0.16	72.94 ± 0.09
		T3	71.38 ± 0.05	71.49 ± 0.38	70.22 ± 0.10	71.12 ± 0.05
		Mean	72.05 ± 0.05	71.49 ± 0.38	71.30 ± 0.09	71.81 ± 0.05

[a] Standard errors centered for each group/tester combination attached

Table A6: Variance components and heritability (h^2) from the second stage of phenotypic analysis for grain yield (GDY) and grain dry matter content (GDC) in calibration sets CS1 and CS2 of Maize 2.

	Trait	$\hat{\sigma}^2_{g_1}$	$\hat{\sigma}^2_{g_1 \times loc}$	$\hat{\sigma}^2_{e^*}$	$\hat{h}^2$
CS1	GDY	25.05 ± 1.52	11.91 ± 0.95	27.31	0.72
	GDC	1.106 ± 0.05	0.128 ± 0.01	0.170	0.94
CS2	GDY	39.01 ± 2.69	19.46 ± 1.64	33.88	0.71
	GDC	1.658 ± 0.08	0.050 ± 0.01	0.268	0.95

Table A7: Average predictive abilities of PBLUP and GBLUP with and without correction for group/tester effects obtained with 10×5-fold cross-validation of grain yield (GDY) and grain dry matter content (GDC) in calibration set CS1 and CS2 of Maize 2.

	Model	Predictive ability ± standard deviation			
		with correction		without correction	
		GDY	GDC	GDY	GDC
CS1	PBLUP	0.383 ± 0.004	0.597 ± 0.004	0.641 ± 0.002	0.646 ± 0.003
	GBLUP	0.594 ± 0.006	0.872 ± 0.003	0.726 ± 0.003	0.868 ± 0.003
CS2	PBLUP	0.400 ± 0.011	0.594 ± 0.005	0.467 ± 0.010	0.748 ± 0.004
	GBLUP	0.489 ± 0.012	0.861 ± 0.004	0.535 ± 0.010	0.842 ± 0.006

Table A8: Correlation (r_{TS}) between predicted and observed testcross values for grain yield (GDY) and grain dry matter content (GDC) from the prediction across years within groups, tester subsets and group/tester subsets from calibration set CS1 and CS2 of Maize 2.

CS1 → CS2		r_{TS}		CS2 → CS1		r_{TS}	
		GDY	GDC			GDY	GDC
G1	G1	0.393	0.391	G1	G1	0.429	0.682
	G1/T1	0.341	0.620				
	G1/T3	0.451	0.653				
G2	G2	0.141	0.830	G2	G2	-0.070	0.485
G3	G3	0.284	0.446	G3	G3	0.272	0.600
	G3/T1	0.384	0.411		G3/T1	0.298	0.396
	G3/T3	0.224	0.648		G3/T2	0.304	0.627
T1	T1	0.302	0.557	T1	T1	0.341	0.589
	G1/T1	0.358	0.610		G1/T1	0.322	0.595
	G3/T1	0.235	0.403		G3/T1	0.338	0.531

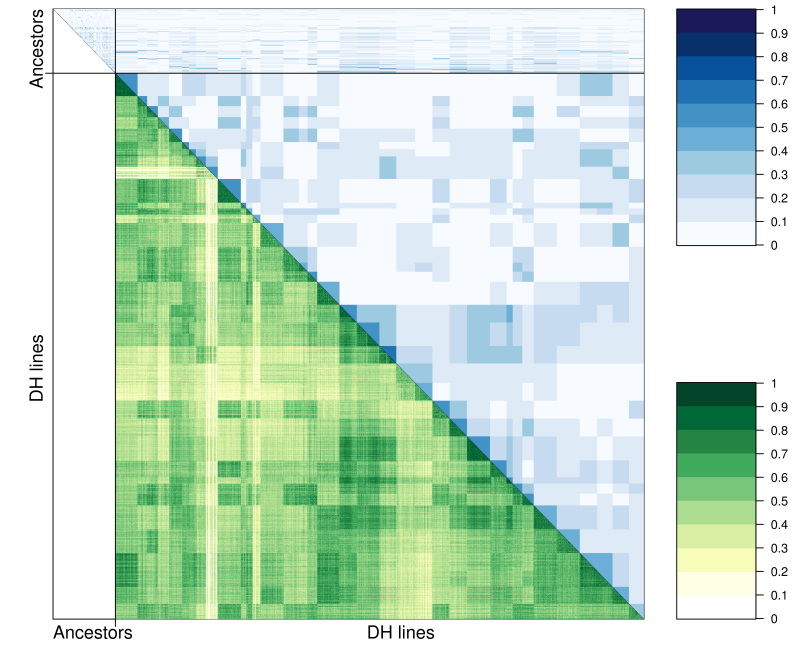


Figure A1: Expected (upper triangle and diagonal, blue) and realized kinship coefficients (lower triangle, green) based on the modified simple matching coefficient in Maize 1.

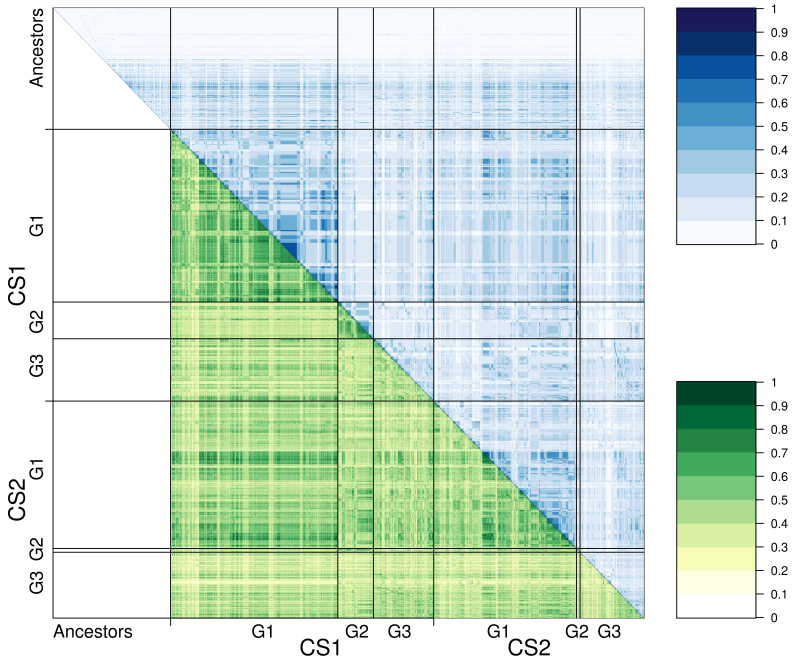


Figure A2: Expected (upper triangle and diagonal, blue) and realized kinship coefficients (lower triangle, green) based on the modified simple matching coefficient of calibration set CS1 and CS2 in Maize 2.

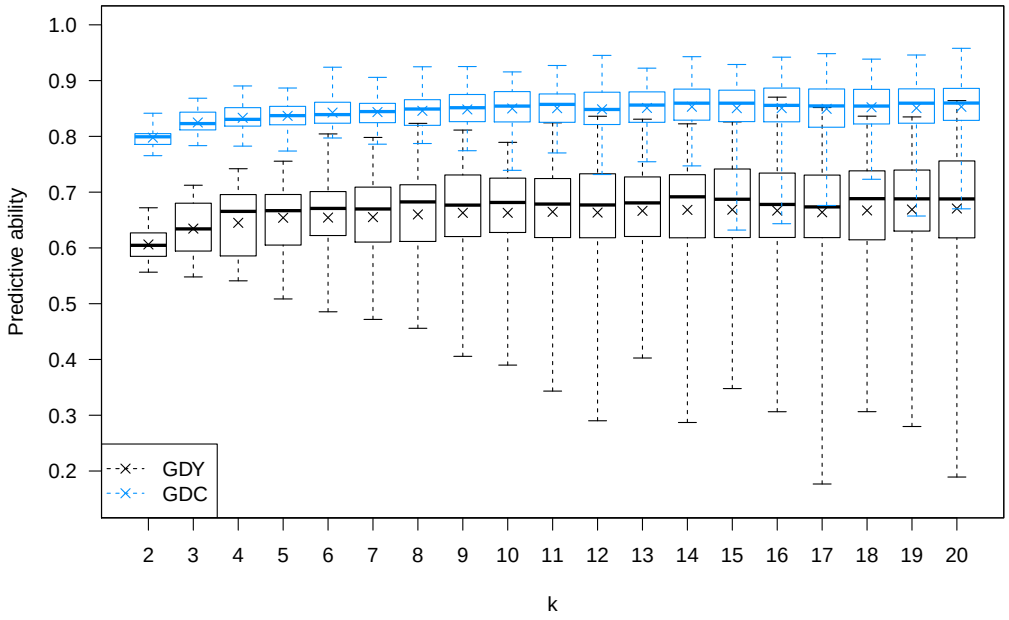
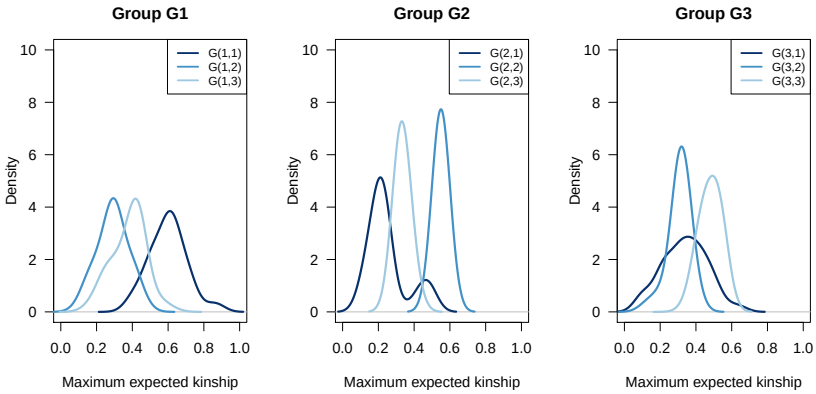
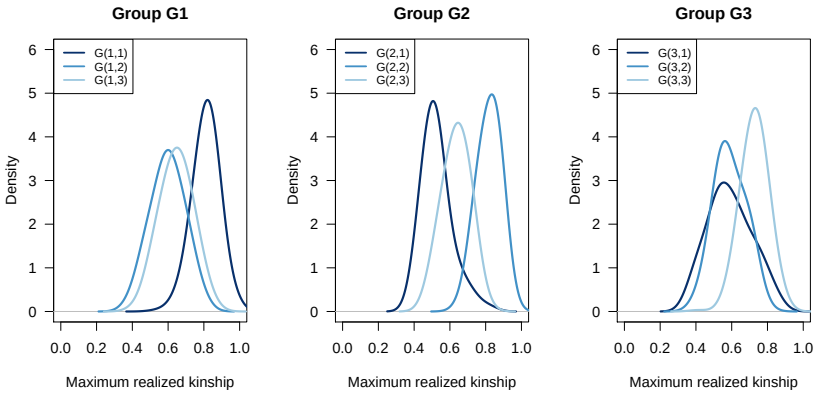


Figure A3: Predictive abilities of $10 \times k$-fold cross-validation with increasing number of folds for grain yield (GDY) and grain dry matter content (GDC) within the subset of tester T1 of calibration set CS1 of Maize 2 ($N = 698$, $M = 11646$).

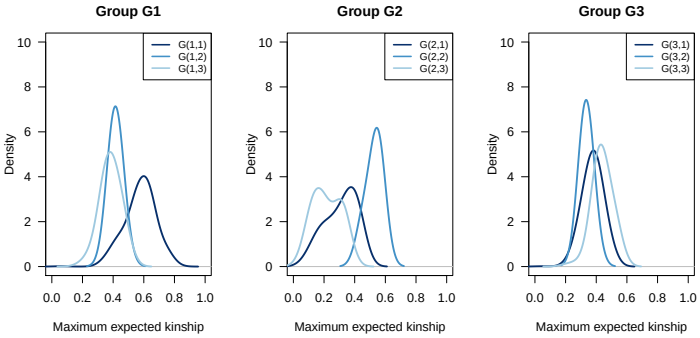


(a) Maximum expected kinship across groups

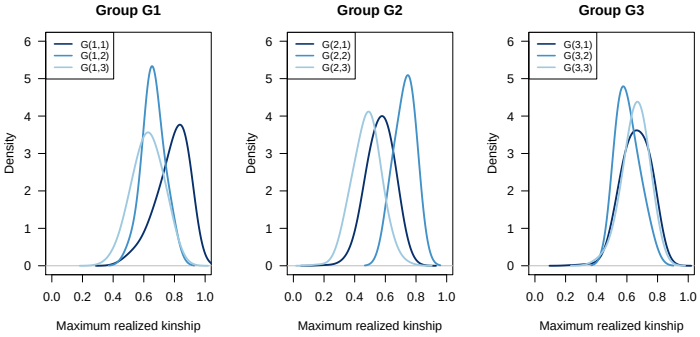


(b) Maximum realized kinship across groups

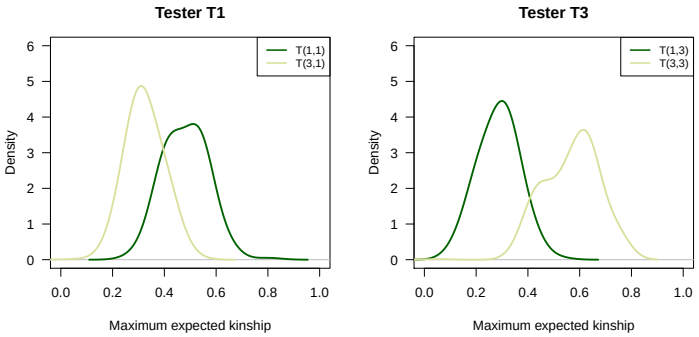
Figure A4: Density plots of maximum expected and realized kinship based on the modified simple matching coefficient across different subsets of groups and clusters in calibration set CS1 of Maize 2.



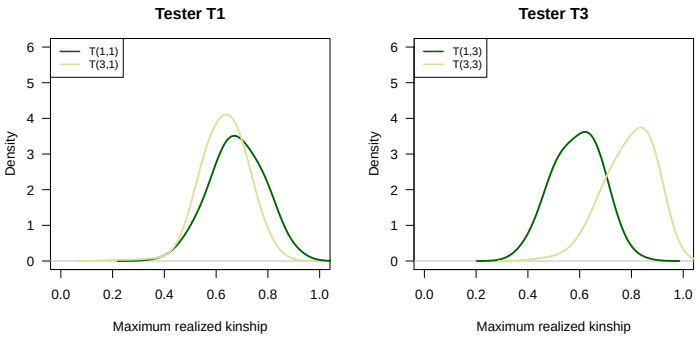
(a) Maximum expected kinship across groups



(b) Maximum realized kinship across groups

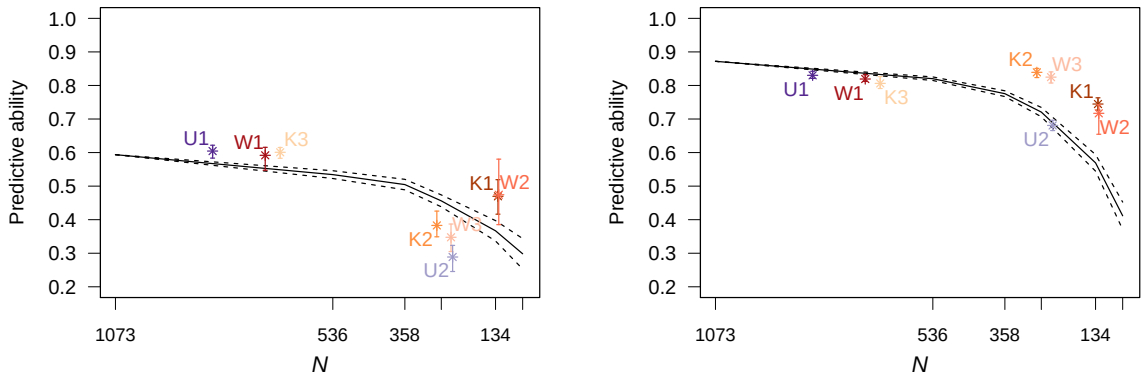


(c) Maximum expected kinship across tester



(d) Maximum realized kinship across tester

Figure A5: Density plots of maximum expected and realized kinship based on the modified simple matching coefficient across different subsets of groups and clusters in calibration set CS2 of Maize 2.



(a) Grain yield

(b) Grain dry matter content

Figure A6: Predictive abilities with decreasing number of observations (N) for GBLUP assessed with CV-R within random subsets of the complete data set and within clusters derived from UPGMA (U1, U2), Ward's (W1, W2, W3), and k-means (K1, K2, K3) clustering in calibration set CS1 of Maize 2 for (a) grain yield and (b) grain dry matter content. Stars and circles with whiskers indicate average predictive abilities and range of 10 replications averaged across five folds. Dashed lines indicate 95 % confidence intervals for predictive abilities across random subsets.

10 Publications out of this thesis

The following papers have been published in advance out of this thesis:

Albrecht T, Wimmer V, Auinger HJ, Erbe M, Ouzunova M, Knaak C, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350

The final publication is available at
`http://link.springer.com/article/10.1007/s00122-011-1587-7`.

The following Sections include parts of this paper: 2.1.1, 2.2.1, 2.4.1, 2.5, 2.6.1, 3.1.1, 3.2.1, 3.4.1, 3.4.2, 4.1.1, and 4.2.1

Candidate's contribution: analyzing data, discussion of results, composing graphs and tables, writing parts of the manuscript and revision of the paper. The overall contribution was 40 %.

Albrecht T, Auinger HJ, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho HP, Schön CC (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor Appl Genet 127:1375–1386

The final publication is available at
`http://link.springer.com/article/10.1007/s00122-014-2305-z`.

The following Sections include parts of this paper: 2.1.2, 2.2.2, 2.6.2, 2.6.3, 2.6.4, 3.1.2, 3.2.2, 3.3, 3.4.3, 3.4.4, 3.4.7, 3.4.8, 4.2.3, 4.2.4, 4.2.6, and 4.2.7

Candidate's contribution: development of validation schemes, writing functions and analyzing data, discussion of results, composing graphs and tables, writing the first draft of the manuscript and revision of the paper. The overall contribution was 75 %.

11 Acknowledgements

First, I want to thank my supervisor Prof. Dr. Chris-Carolin Schön for her advice, interest, and support throughout this thesis.

Many thanks to Dr. Milena Ouzunova, Dr. Carsten Knaak, and Dr. Sofia daSilva from KWS SAAT AG for providing me the experimental data sets from commercial maize breeding programs and their helpful discussions.

I am grateful to Prof. Dr. Ruedi Fries, Dr. Hubert Pausch, and Hildrun Walter for processing DNA samples and the high-density SNP genotyping arrays.

A lot of thanks to Prof. Dr. Hans-Peter Piepho, Prof. Dr. Daniel Gianola, Prof. Dr. Donna Ankerst, Dr. Joseph O. Ogutu, Hans-Jürgen Auinger, and Valentin Wimmer for their suggestions on the statistical analysis.

Special thanks to Christina Lehermeier, Manfred Schönleben, and Sebastian Gresset for the discussions about maize breeding and statistics during coffee breaks or in the train between Munich and Freising.

Many thanks goes also to Dr. Eva Bauer, Dr. Nicole Krämer, Wiltrud Erath, Flavio Foiada, Sebastian Steinemann, Dr. Yu Wang, and the complete team of the Chair of Plant Breeding at the Technische Universität München for their help and assistance during this thesis.

Many thanks also to Dr. Lorenz Hartl, Prof. Dr. Volker Mohler, and my new working group at the Bavarian State Research Center for Agriculture for their patience and support while finishing this thesis.

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” (FKZ: 0315528A).

12 Curriculum Vitae

Personal Information

Theresa Albrecht

Birth: September 7th 1984 in Halle (Saale), Germany

Work and Education

since 01/2013 — Research Scientist, Bavarian State Research Center for Agriculture, Institute for Crop Science and Plant Breeding

07/2009 – 07/2012 — Ph.D. student, Technische Universität München, Chair of Plant Breeding

Thesis "Genome-based prediction of testcross performance in maize (*Zea mays* L.)." within the project "Synbreed – Synergistic Plant and Animal Breeding"

Supervisor: Prof. Dr. Chris-Carolin Schön

09/2004 – 05/2009 — Diplom in Agricultural Biology, Universität Hohenheim, Germany

Thesis "Phenotypic evaluation of a finger millet [*Eleusine coracana* L.] core collection under different field conditions in India."

Supervisor: Prof. Dr. Albrecht E. Melchinger

05/2004 — Abitur, Feodor-Lynen-Gymnasium, Planegg, Germany

Publications

Albrecht T, Auinger HJ, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho HP, Schön CC (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor Appl Genet 127:1375–1386

Mohler V, **Albrecht T**, Mrva K, Hartl L (2014) Genetic analysis of falling number in three bi-parental common winter wheat populations. Plant Breed 133:448–453

Mohler V, Diethelm M, Castell A, **Albrecht T**, Friedlhuber R, Livaja M, Hartl L (2014) CORNET Efficient Wheat: The influence of *Rht-D1* on agronomic performance and quality traits in common winter wheat. In: Tagungsband der 64. Jahrestagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs. Mutagenese und Hochdurchsatz-Screening. Höhere Bundeslehr- und Forschungsanstalt für Landwirtschaft Raumberg-Gumpenstein, A-Irdning, pp.31–32

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

Lehermeier C, Wimmer V, **Albrecht T**, Auinger HJ, Gianola D, Schmid V, Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. Stat Appl Genet Mol Bio 12:375–391

Albrecht T, Schön CC (2012) Genom-basierte Vorhersage der Testkreuzungsleistung bei Mais. In: Tagungsband der 62. Jahrestagung der Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs. Von markergestützter Selektion zu genomischer Selektion in der Pflanzenzüchtung. Höhere Bundeslehr- und Forschungsanstalt für Landwirtschaft Raumberg-Gumpenstein, A-Irdning, pp.3–6

Wimmer V, **Albrecht T**, Auinger HJ, Schön CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086-2087

Albrecht T, Wimmer V, Auinger HJ, Erbe M, Ouzunova M, Knaak C, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350

Upadhyaya HD, Sarma NDRK, Ravishankar CR, **Albrecht T**, Narasimhudu Y, Singh SK, Varshney SK, Reddy VG, Singh S, Dwivedi SL, Wanyera N, Oduori COA, Mgonja MA, Kisandu DB, Parzies HK, Gowda CLL (2010) Developing a mini-core collection in finger millet using multilocation data. Crop Sci 50:1924–1931