



Technische Universität München
Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Lehrstuhl für Pflanzenzüchtung

Investigation of genome-based prediction in differentially structured plant populations

Christina Lehermeier

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.

Vorsitzende: Univ.-Prof. D. P. Ankerst, Ph. D.

Prüfer der Dissertation:

1. Univ.-Prof. Dr. C.-C. Schön
2. Univ.-Prof. Dr. A. Tellier
3. Prof. D. Gianola, Ph. D.
University of Wisconsin – Madison/USA

Die Dissertation wurde am 07.05.2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 16.07.2015 angenommen.

The words 'true model' represent an oxymoron
(Burnham and Anderson, 2002)

Contents

	Page
List of Figures	V
List of Tables	VI
List of Abbreviations	VII
Publications included in this Thesis	IX
1 Introduction	1
1.1 Background	1
1.2 Outline	4
2 Material and Methods	7
2.1 Datasets	7
2.1.1 Simulated datasets	7
2.1.2 Experimental datasets	7
2.2 Genome-based prediction methods	12
2.3 Evaluation of genome-based prediction methods	21
2.3.1 Bayesian learning ability	21
2.3.2 Prediction performance	21
3 Discussion	24
3.1 Proposed methods for genome-based prediction	24
3.1.1 Comparison of method performance	24
3.1.2 Choice of model parameters	27
3.1.3 Bayesian learning ability	28
3.2 Population-specific factors affecting prediction performance	29
3.2.1 Linkage disequilibrium and linkage phases	30
3.2.2 Relatedness	32
3.2.3 Allele frequencies and the allele substitution effect	34
3.2.4 Heritability	35

3.3	Construction of the estimation set	36
3.3.1	Potential of multi-parental populations for genome-based prediction	37
3.3.2	Prediction across populations	39
3.4	Multivariate models for structured data	41
3.4.1	Prediction performance of the multivariate model	41
3.4.2	Multivariate models in animal breeding	44
3.4.3	Genomic correlations between sub-populations	45
3.4.4	Extending the multivariate model	46
3.5	Conclusions	49
4	Summary	50
5	Zusammenfassung	52
6	References	54
7	Appendix	74
7.1	Supporting Table	74
7.2	Publications	77
8	Acknowledgements	78
9	Curriculum Vitae	80

List of Figures

	Page
1 Number of markers and lines used in genome-based prediction studies in maize	2
2 Investigated plant species in publications on genomic selection.	3
3 Comparison of marginal prior densities	16
4 Investigated genome-based prediction methods in experimental plant data	25
5 Comparison of A-GBLUP, MG-GBLUP, and W-GBLUP	43
6 Estimated genomic correlations of Synbreed CS2 data	46
7 Predictive abilities of A-GBLUP, MG-GBLUP, and W-GBLUP in Synbreed CS2 data	47

List of Tables

	Page
2 Overview of datasets analyzed	8
A1 Overview of published experiments on genome-based prediction in plant species	74

List of Abbreviations

A-GBLUP	across-group genome-based best linear unbiased prediction
ANN	artificial neural networks
BLUP	best linear unbiased prediction
CV	cross-validation
CwC	cross-with-cross prediction
DH	doubled haploid
DMC	dry matter content
DMY	dry matter yield
DtSILK	days to silking
DtTAS	days to tasseling
FT	flowering time
GBLUP	genome-based best linear unbiased prediction
GDC	grain dry matter content
GDY	grain dry matter yield
GWAS	genome-wide association study/studies
LASSO	least absolute shrinkage and selection operator
LD	linkage disequilibrium
LOCO-CV	leave-one-cross-out cross-validation
MAF	minor allele frequency
MAS	marker assisted selection
MCMC	Markov chain Monte Carlo
MG-GBLUP	multi-group genome-based best linear unbiased prediction
NAM	nested association mapping
PBLUP	pedigree-based best linear unbiased prediction
PEV	prediction error variance
PH	plant height
PL	panicle length
QTL	quantitative trait locus/loci
R-CV	Random cross-validation
REML	restricted maximum likelihood
RF	random forest

RKHS	reproducing kernel Hilbert space (regression)
RR-BLUP	Ridge regression best linear unbiased prediction
SNP	single nucleotide polymorphism (marker)
SVM	support vector machines
TBV	true breeding value
W-GBLUP	within-group genome-based best linear unbiased prediction

Publications included in this Thesis

Lehermeier et al. (2013)

Lehermeier C, Wimmer V, Albrecht T, Auinger H-J, Gianola D, Schmid V J, Schön C-C (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. *Statistical Applications in Genetics and Molecular Biology* 12: 375–391

Abstract: Different statistical models have been proposed for maximizing prediction accuracy in genome-based prediction of breeding values in plant and animal breeding. However, little is known about the sensitivity of these models with respect to prior and hyperparameter specification, because comparisons of prediction performance are mainly based on a single set of hyperparameters. In this study, we focused on Bayesian prediction methods using a standard linear regression model with marker covariates coding additive effects at a large number of marker loci. By comparing different hyperparameter settings, we investigated the sensitivity of four methods frequently used in genome-based prediction (Bayesian Ridge, Bayesian Lasso, BayesA and BayesB) to specification of the prior distribution of marker effects. We used datasets simulated according to a typical maize breeding program differing in the number of markers and the number of simulated quantitative trait loci affecting the trait. Furthermore, we used an experimental maize dataset, comprising 698 doubled haploid lines, each genotyped with 56,110 single nucleotide polymorphism markers and phenotyped as testcrosses for the two quantitative traits grain dry matter yield and grain dry matter content. The predictive ability of the different models was assessed by five-fold cross-validation. The extent of Bayesian learning was quantified by calculation of the Hellinger distance between the prior and posterior densities of marker effects. Our results indicate that similar predictive abilities can be achieved with all methods, but with BayesA and BayesB hyperparameter settings had a stronger effect on prediction performance than with the other two methods. Prediction performance of BayesA and BayesB suffered substantially from a non-optimal choice of hyperparameters.

Candidate's contribution: Estimation of genome-based prediction models and calculation of Hellinger distances, discussion of results, creating figures and tables, writing the

first draft of the manuscript, and revision of the paper. Preliminary results of the study were in part included in the candidate's diploma thesis.

Lehermeier et al. (2014)

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

Abstract: The efficiency of marker-assisted prediction of phenotypes has been studied intensively for different types of plant breeding populations. However, a remaining question is how to incorporate and counterbalance information from bi-parental and multi-parental populations into model training for genome-wide prediction. To address this question, we evaluated testcross performance of 1,652 doubled-haploid maize (*Zea mays* L.) lines, genotyped with 56,110 single nucleotide polymorphism markers, and phenotyped for five agronomic traits in four to six European environments. Lines are arranged in two diverse half-sib panels representing two major European heterotic germplasm pools. The dataset contains ten related bi-parental dent families and eleven related bi-parental flint families generated from crosses of maize lines important for European maize breeding. With this new dataset we analyzed genome-based best linear unbiased prediction in different validation schemes and compositions of estimation and test sets. Further, we investigated theoretically and empirically marker linkage phases across multi-parental populations. In general, predictive abilities similar or higher than those within bi-parental families could be achieved by combining several half-sib families in the estimation set. For the majority of families, 375 half-sib lines in the estimation set were sufficient to reach the same predictive performance of biomass yield as an estimation set of 50 full-sib lines. In contrast, prediction across heterotic pools was not possible for most cases. Our findings have important impact on the experimental design in genome-based prediction as they provide guidelines for the genetic structure and required sample size of appropriate datasets used for model training.

Candidate's contribution: Method development and analyses of the data, discussion of results, creating figures and tables, writing the first draft of the manuscript, and revision of the paper.

Lehermeier et al. (2015)

Lehermeier C, Schön C-C, de los Campos G (2015) Assessment of genetic heterogeneity in structured plant breeding populations using multivariate whole-genome regression models. *Genetics*. doi:10.1534/genetics.115.177394

Abstract: Plant breeding populations exhibit varying levels of structure and admixture; these features are likely to induce heterogeneity of marker effects across sub-populations. Traditionally, structure has been dealt with as a potential confounder, and various methods exist to 'correct' for population stratification. However, these methods induce a mean-correction that does not account for heterogeneity of marker effects. The animal breeding literature offers a few recent studies that consider modeling genetic heterogeneity in multi-breed data using multivariate models. However, these methods have received little attention in plant breeding where population structure can have different forms. In this article we address the problem of analyzing data from heterogeneous plant breeding populations using three approaches: (a) a model that ignores population structure (A-GBLUP), (b) a stratified (i.e. within group) analysis (W-GBLUP), and (c) a multivariate approach that uses multi-group data and accounts for heterogeneity (MG-GBLUP). The performance of the three models was assessed on three different datasets: a diversity panel of rice (*Oryza sativa*), a maize (*Zea mays* L.) half-sib panel, and a wheat (*Triticum aestivum* L.) dataset that originated from plant breeding programs. The estimated genomic correlations between sub-populations varied from null to moderate depending on the genetic distance between sub-populations and traits. Our assessment of prediction accuracy features cases where ignoring population structure leads to a parsimonious more powerful model as well as others where the multivariate and stratified approach have higher predictive power. In general, the multivariate approach appeared slightly more robust than either the A- or W-GBLUPs.

Candidate's contribution: Contribution to method development, analyzing the data, discussion of results, creating figures and tables, writing the first draft of the manuscript, and revision of the paper.

1 Introduction

1.1 Background

More than a decade ago, Meuwissen et al. (2001) coined the term “genomic selection” for the prediction of breeding values in livestock and crops using DNA markers covering the whole genome. The idea of genomic selection can be summarized briefly: using a phenotyped and genotyped training population of individuals, the effects of all available markers are estimated simultaneously in a statistical model. Following this, the estimated marker effects are used to predict the phenotypic performance, or rather genotypic/breeding values, of a not yet phenotyped set of selection candidates. Based on these predicted breeding values, promising individuals are selected. The part of predicting breeding values based on DNA information in this thesis is called “genome-based prediction,” whereas a number of terms are used analogously in the literature, e.g. “genomic prediction” or “whole-genome prediction”. Especially when the trait of interest is quantitatively inherited, meaning several regions on the genome (quantitative trait loci, QTL) are influencing the trait, genomic selection is assumed to be superior to established marker-assisted selection (MAS) approaches (Jannink et al. 2010). Whereas in MAS significant marker-trait associations, or potential QTL, are identified by genome-wide association studies (GWAS) or QTL mapping approaches, and then used for selection, typically no pre-selection of significant markers is conducted in genomic selection. The potential advantage of the genomic selection approach is that by using all available markers in the model, preferably all QTL should be captured due to their linkage disequilibrium (LD) with markers, also those with small effects on the trait (Goddard and Hayes 2007). However, this approach also means that, usually, more marker effects need to be estimated than there are available phenotypic observations in the model, which requires the use of regularization methods, because ordinary least squares estimations of marker effects cannot be applied. Thus, Meuwissen et al. (2001) suggested Ridge regression best linear unbiased prediction (RR-BLUP) and two Bayesian methods (BayesA and BayesB) for estimating marker effects.

A few years before Meuwissen et al. (2001), Nejati-Javaremi et al. (1997) had already proposed the idea of estimating breeding values through molecular marker information. The suggested approach adapts the best linear unbiased prediction (BLUP) of breeding

values from Henderson (1975), using a relationship matrix constructed by marker data instead of pedigree information. Later, several authors showed the analog interpretation of BLUP by using a genome-based relationship matrix (GBLUP) or a pedigree-based relationship matrix (PBLUP) (VanRaden 2008; Habier et al. 2007) and also demonstrated that the GBLUP approach can be conveyed to the RR-BLUP approach from Meuwissen et al. (2001).

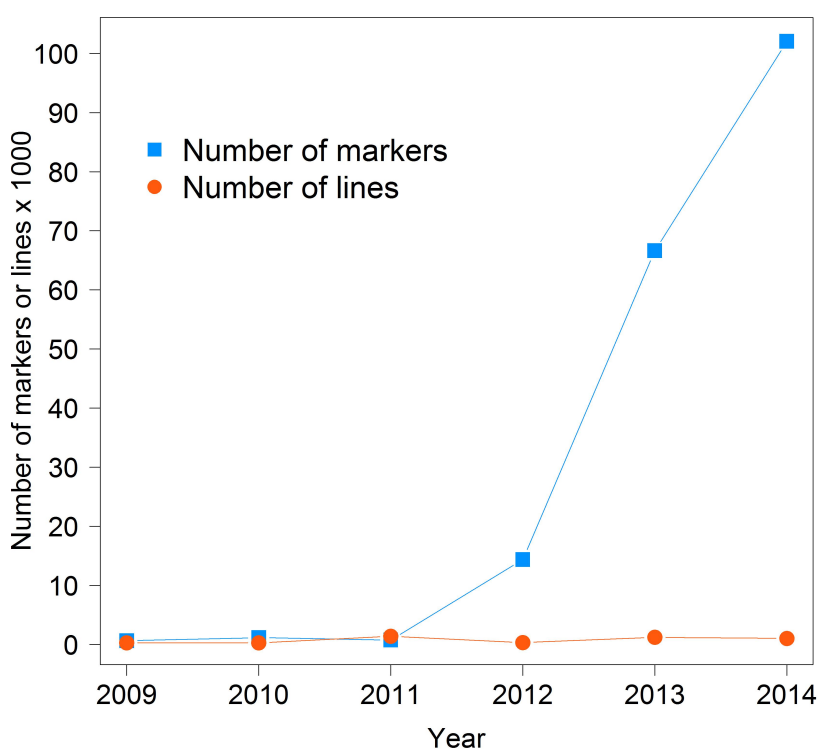


Figure 1: Average number of markers and lines used in published genome-based prediction studies in maize according to year of publication (graphic adapted from Gonzalez-Recio et al. (2014)).

The genomic selection approach quickly found its way into the animal breeding community, as here the traditional BLUP procedure using pedigree information was already well-established (Nakaya and Isobe 2012). In the meantime, technical advances on marker array developments in previous years made low-cost genotyping with several tens of thousands of single nucleotide polymorphism (SNP) markers possible for many species. Thus, genomic selection exhibited its practical impact, as genotyping costs de-

creased much more than phenotyping costs. Genomic selection was first applied in dairy cattle, where in 2008 a 50k SNP chip was released (Matukumalli et al. 2009) and the first experiments showed promising results (VanRaden et al. 2009). In the context of plant breeding, the first simulation studies were conducted by Bernardo and Yu (2007) to investigate the selection responses of genomic selection. The first empirical studies testing the efficiency of genomic selection compared to MAS were conducted in bi-parental families of maize, barley, and Arabidopsis using a few hundred polymorphic markers (Lorenzana and Bernardo 2009). Studies on breeding populations in wheat and maize followed (Crossa et al. 2010; Albrecht et al. 2011; Heffner et al. 2011a). Today, maize has been the most extensively studied crop in the context of genomic selection, where in 2011 a 50k SNP chip (Ganal et al. 2011) and just recently also a 600k SNP chip became available (Unterseer et al. 2014). Thus, similar to in livestock species (Gonzalez-Recio et al. 2014), in maize the number of markers has increased heavily during the last few years, while the number of available phenotypes has stayed relatively constant (see Figure 1).

In the last few years, genome-based prediction analyses have been conducted on a large variety of plant species, such as rice (Wimmer et al. 2013; Guo et al. 2014; Xu et al. 2014), sugar beet (Hofheinz et al. 2012; Würschum et al. 2013; Biscarini et al. 2014), cassava (de Oliveira et al. 2012; Ly et al. 2013), rye (Bernal-Vasquez et al. 2014; Wang et al. 2014), rapeseed (Würschum et al. 2014), sunflower (Reif et al. 2013), sugarcane (Gouy et al. 2013), soybean (Bao et al. 2014), oats (Asoro et al. 2011), and different trees (Kumar et al. 2012; Resende et al. 2012a,b,c; Munoz et al. 2014). Figure 2

presents an overview on absolute frequencies of the different species analyzed in genomic selection studies in experimental plant data.

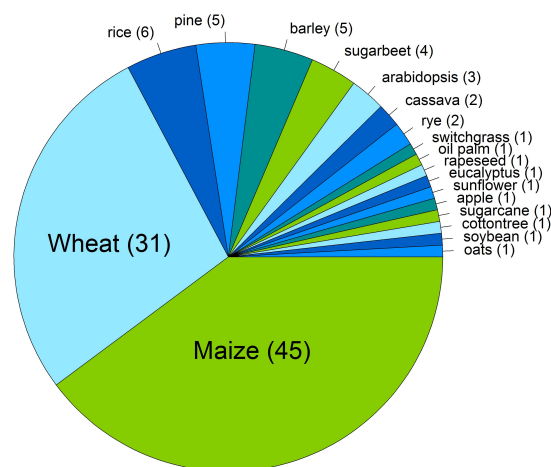


Figure 2: Investigated plant species in publications on genomic selection according to Table A1. Absolute number of studies are given in brackets.

1.2 Outline

A crucial step for genome-based prediction is finding a statistical method which can handle the huge amount of marker as input variables in conjunction with a limited number of phenotypic records (see also Figure 1). Many methods have been proposed and tested which induce different kinds of regularization and shrinkage on marker effects. In the context of Bayesian statistics, the amount of shrinkage on marker effects is specified by prior distributions. In the most frequently used method RR-BLUP (together with the analogous GBLUP) or its fully Bayesian version (Bayesian Ridge), the amount of shrinkage is marker-homogeneous. The underlying assumption here is that all markers explain the same amount of genotypic variation, which is quite stringent, as some markers might be in LD with QTL, while others are not. Thus, Meuwissen et al. (2001) proposed the BayesA and BayesB methods, which induce marker-heterogeneous shrinkage, albeit BayesB is the more general method which sets *a priori* a fraction π of marker effects to zero. Another broadly used method which induces heterogeneous shrinkage on marker effects is the Bayesian Lasso (Park and Casella 2008). This thesis investigates the performance of these four Bayesian methods (Bayesian Ridge, Bayesian Lasso, BayesA, and BayesB) for genome-based prediction, using two simulated and one experimental maize breeding datasets within the publication Lehermeier et al. (2013). In all Bayesian methods, hyperparameters need to be specified which adjust the strength of effect regularization. The sensitivity of the methods with respect to these hyperparameters is analyzed in Lehermeier et al. (2013). The goal of this study was not only to investigate the prediction performance of the methods, but also to assess their Bayesian learning ability, which is the ability of a method to learn from the data and to move away from the assigned prior distribution (Sorensen and Gianola 2002). The Bayesian learning ability of the methods was quantified by the Hellinger distance (Le Cam 1986) between marginal prior and posterior densities of the marker effects.

In addition to the application of the appropriate method, the composition of the training population is particularly important for good prediction performance in genome-based prediction. Using a multi-parental population of maize, consisting of two large half-sib panels representing two main heterotic germplasm pools for hybrid breeding (dent and flint), the optimum design of the training population was investigated in Lehermeier et al. (2014). In particular, the prediction power of a training population consisting of

progenies of crosses sharing both parents was compared to one consisting of progenies from crosses sharing one parent, also taking sample size into account. Furthermore, different complex traits were assessed, if prediction across heterotic maize pools was possible. While early studies on genome-based prediction assumed that performance was driven mainly by LD between markers and QTL, Habier et al. (2007) showed that prediction accuracy derives, to a large extent, from relatedness between the training population and selection candidates. However, LD patterns and relatedness between lines go hand in hand (Wientjes et al. 2012), and several studies have observed that linkage phases between markers differ more strongly for less related populations (Andrescu et al. 2007; de Roos et al. 2008; Toosi et al. 2009). Differences in linkage phases between the training population and selection candidates are problematic, as the consistency of the marker-QTL linkage phase is a crucial assumption for genome-based prediction as well as for MAS (Goddard and Hayes 2007; de Roos et al. 2009; Hayes et al. 2009). Thus, in Lehermeier et al. (2014), the influence of relatedness and linkage phase differences has been investigated and discussed within the specific situation present in large half-sib panels of maize. Additionally, it has been investigated how differences in trait heritability affect prediction performance.

Experimental plant data often exhibit strong population structures, due to geographic adaptations and natural and artificial selection. It was observed that if population structure is ignored, this can lead to false positive marker-trait associations in GWAS and to an over-optimistic assessment of accuracy in genome-based prediction, if mean differences in the different sub-populations are not accounted for (Windhausen et al. 2012; Albrecht et al. 2014; Guo et al. 2014). However, accounting for mean differences does not allow for marker effects that differ in different sub-populations. This is supposed to be problematic if the genetic model deviates from an additive model, as, for example, when epistatic or dominance effects exist. In such a case, marker effects may differ between populations, as the allele substitution effect is influenced by the allele frequency in the population (Falconer and Mackay 1996). Additionally, marker effects might be different due to differences in the LD patterns between markers and QTL in different populations. Marker effect differences in different sub-populations have so far been ignored in most genome-based prediction studies, for instance in Lehermeier et al. (2013, 2014), where the focus was on comparing different Bayesian methods and the applicable construction of the training population, respectively. Thus, Lehermeier et al. (2015)

investigate this issue using three differently structured experimental plant populations. A multivariate model is proposed to estimate population-specific marker effects, without losing information on the other sub-populations.

2 Material and Methods

2.1 Datasets

Different simulated and experimental plant datasets were analyzed in this study. An overview can be found in Table 2. In the following, a brief description of the datasets is given, and details can be found in the respective publications.

2.1.1 Simulated datasets

In Lehermeier et al. (2013) two simulated maize datasets, maizeA and maizeB, were used to assess the ability of different Bayesian methods to accurately predict true genotypic values (true breeding values, TBV). The simulation procedure mimics a typical maize breeding program with small effective population size and large LD as observed in experimental maize breeding populations. For both datasets 1250 fully homozygous recombinant inbred lines were simulated. QTL with equal additive effects on the phenotype were simulated and environmental errors were sampled from a normal distribution. Different marker densities and number of QTL were simulated for maizeA and maizeB. After the simulation of six breeding cycles, the maizeA dataset contained 1,117 polymorphic biallelic markers and 500 polymorphic QTL. With 7425 polymorphic markers and 369 polymorphic QTL maizeB contained more markers than phenotypes. Trait heritabilities, as correlation between phenotypic values and TBV, were 0.46 for maizeA and 0.64 for maizeB. The dataset maizeA has been made available within the `synbreedData` R package (Wimmer et al. 2012).

2.1.2 Experimental datasets

Synbreed CS1 maize dataset In addition to the two simulated maize datasets, an experimental maize dataset was used in Lehermeier et al. (2013), to investigate further the behavior of Bayesian methods on experimental data. The dataset was generated within the Synbreed project by KWS SAAT AG and represents a subset of calibration set 1 from Albrecht et al. (2014) comprising only the DH lines crossed to tester T1. In the following, the dataset is denoted as Synbreed CS1 maize dataset. It contains 698 genotyped and phenotyped doubled haploid (DH) maize (*Zea mays* L.) lines, which

Table 2: Overview of datasets analyzed in this thesis and corresponding publications

	n ¹	p ²	Traits	Used in publication
Simulated datasets				
maizeA	1,250	1,117	simulated Gaussian	Lehermeier et al. (2013)
maizeB	1,250	7,425	simulated Gaussian	Lehermeier et al. (2013)
Experimental datasets				
Synbreed CS1 maize	698	11,646	GDY (dt/ha) ³ , GDC (%) ⁴	Lehermeier et al. (2013)
Synbreed CS2 maize	842	16,847	GDY (dt/ha), GDC (%)	–
Cornfed maize	1652	34,116	DMY (dt/ha) ⁵ , DMC (%) ⁶ , PH (cm) ⁷ , Dt-TAS (d) ⁸ , DtSILK (d) ⁹	Lehermeier et al. (2014, 2015)
rice	337	36,861	FT (d) ¹⁰ , PH (cm), PL (cm) ¹¹	Lehermeier et al. (2015)
wheat	599	1,279	grain yield (four environments)	Lehermeier et al. (2015)

¹: number of lines; ²: number of markers; ³: grain dry matter yield; ⁴: grain dry matter content; ⁵: dry matter yield; ⁶: dry matter content; ⁷: plant height; ⁸: days to tasseling; ⁹: days to silking; ¹⁰: flowering time; ¹¹: panicle length

belong to the maize dent heterotic pool and were derived from 122 different crosses with, on average, six DH lines per cross.

DH lines were phenotyped as testcrosses in four German locations in the year 2010 for the traits grain dry matter yield (GDY; dt/ha) and grain dry matter content (GDC; %). Each location comprised eight trials with two replications, respectively. Trials were laid out in a 10 × 10 lattice design. Outlying observations were consecutively detected and removed based on maximum deviate residuals according to Grubbs (1950), with a significance value of 5%. In each location, phenotypic plot observations were adjusted in the first step for trial, replication, and block effects. In the second step, adjusted means over locations were calculated for each genotype and used for further analyses. Generalized heritabilities on an entry mean basis, estimated according to Cullis et al. (2006), were 0.74 for GDY and 0.94 for GDC.

Lines were genotyped with the Illumina MaizeSNP50 BeadChip[®] (Ganal et al. 2011) containing 56,110 single nucleotide polymorphism markers (SNPs). SNPs with a call frequency < 0.9, a GenTrainScore < 0.7, a minor allele frequency (MAF) < 0.01 as well as redundant SNPs were discarded. Missing values were imputed based on family information and on flanking markers using the software BEAGLE (Browning and Browning 2009) and R package synbreed (Wimmer et al. 2012). Finally, 11,646 high-quality polymorphic SNPs were used for further analyses.

Synbreed CS2 maize dataset In addition to the Synbreed CS1 dataset, this thesis includes analyses using calibration set 2 from Albrecht et al. (2014), which represents the selection cycle subsequent to calibration set 1 (Synbreed CS1), further denoted as the Synbreed CS2 dataset. According to pedigree information, the lines of the Synbreed CS2 dataset were separated into three genetic groups (G1, G2, and G3), including 582, 15, and 260 lines, respectively. For the analyses in this thesis, lines belonging to the smallest group G2 were excluded. Thus, 842 phenotyped and genotyped DH lines were used for analysis.

In 2011, the DH lines were phenotyped as testcrosses for GDY and GDC. Two different testers were used (T1 and T3). From group G1, 189 lines were crossed with tester T1 (G1_T1) and 393 with tester T3 (G1_T3). From group G3, 138 lines were crossed with tester T1 (G1_T1) and 122 lines with tester T3 (G1_T3). Field trials and phenotypic analyses were conducted as described for the year 2010, with details given by Albrecht

et al. (2014). Estimated heritabilities on a progeny mean basis were 0.71 and 0.95 for GDY and GDC, respectively. Lines were genotyped with the Illumina MaizeSNP50 BeadChip[®]. After conducting the same quality control steps as described above, 16,847 polymorphic high-quality SNPs were available for further analysis.

Cornfed maize dataset Two large maize panels were generated within the European PLANT-KBBE CornFed project and analyzed in Lehermeier et al. (2014). The generation and genetic structure of the panels were described in Bauer et al. (2013). The two panels represent two major European heterotic germplasm pools, namely dent and flint. In the dent panel, ten diverse dent founder lines were crossed to one common central dent line, F353. Similarly, in the flint panel, eleven diverse flint founder lines were crossed to one common central flint line, UH007. From each cross, fully homozygous DH lines were generated, resulting in 21 bi-parental families with an average of 79 DH lines per family. The design of each panel is similar to the US nested association mapping (NAM) panel (McMullen et al. 2009).

In 2011, the DH lines were phenotyped as testcrosses, whereby the common parent of the opposite pool was used as a tester. Field trials were conducted in four European locations for the dent panel and in six European locations for the flint panel. An augmented p-rep design, according to Williams et al. (2011), was used for the dent as well as the flint trials. For the dent panel, one-quarter of the entries were replicated at each location, and the trials were laid out in 120 incomplete blocks consisting of ten plots each. For the flint panel, one-third of the entries were replicated at each location, and the trials were laid out in 160 incomplete blocks consisting of eight plots each. The phenotypic traits dry matter yield (DMY, dt/ha), dry matter content (DMC, %), plant height (PH, cm), days to tasseling (DtTAS, d), and days to silking (DtSILK, d) were recorded. Phenotypic data analyses were conducted separately for the dent and flint panels. Outlying observations were detected and removed according to Grubbs (1950). For each trait and genotype, adjusted means were calculated over locations using a mixed model, fitted with the ASRem1-R package (Butler et al. 2009), adjusting for replication, block, and location effects. Family-specific genotypic, genotype-by-environment, and residual variance components, as well as their standard errors, were estimated by restricted maximum likelihood estimation (REML) using the ASRem1-R package. Using the resulting variance components, family-specific trait heritabilities on an entry mean basis were es-

estimated according to Hung et al. (2012). Standard errors of heritability estimates were derived using the delta method (Holland et al. 2003).

The DH lines and the parents of each cross were genotyped with the Illumina MaizeSNP50 BeadChip[®]. Markers with a call frequency < 0.9 , a GenTrainScore < 0.7 , or a MAF < 0.01 were discarded. Missing values were imputed based on family information and on flanking markers—as in the Synbreed maize datasets. After quality control, 34,116 high-quality polymorphic markers were available for the combined dent and flint datasets, 32,801 polymorphic markers for the dent panel, and 30,122 polymorphic markers for the flint panel. Overall, for the dent panel 841 and for the flint panel 811 phenotyped and genotyped lines were used for further analysis.

Rice dataset A publicly available rice dataset (<http://www.ricediversity.org/data/>) was used to investigate the performance of a multivariate genome-based prediction model for structured data (Lehermeier et al. 2015). The data represented a worldwide collected diversity panel of 413 inbred accessions of *Oryza sativa* (Zhao et al. 2011). The rice lines were classified into five sub-populations (indica, aus, temperate japonica, tropical japonica, and aromatic) by Zhao et al. (2011). A remaining admixed group of 62 lines could not be assigned to a specific group, and so these and the 14 lines of the smallest group “aromatic” were excluded from the original dataset in the study of Lehermeier et al. (2015), in order to obtain clear sub-populations with reasonable size. Three traits were chosen (flowering time at Aberdeen (FT, d), plant height (PH, cm), and panicle length (PL, cm)) for further analysis. These traits were also analyzed in Wimmer et al. (2013). FT was measured as days to heading in a greenhouse in Aberdeen (Scotland) across a nine-month period. Field trials for the other traits were conducted in Stuttgart (Arkansas) in the years 2006 and 2007, with two replications per year. Phenotypic means across replicates and years were calculated for each inbred line and used for further analysis.

Rice lines were genotyped with an Affymetrix 44K SNP array. Quality control and imputation of missing values were conducted by Zhao et al. (2011) and Wimmer et al. (2013). For the 312 lines used in the study of Lehermeier et al. (2015), 36,858 high-quality polymorphic markers were available.

Wheat dataset The wheat dataset originated from the CIMMYT Global Wheat Breeding program (Crossa et al. 2010) and is publicly available within the BGLR R package (Pérez and de los Campos 2014). Analyses using this dataset are included in Lehermeier et al. (2015). The wheat lines were phenotyped for grain yield in four environments, and they were genotyped with 1,447 binary Diversity Array Technology (DArT) markers (Wenzl et al. 2004). In the publicly available dataset, markers with a MAF < 0.05 had already been removed and missing genotypes imputed using Bernoulli samples (Crossa et al. 2010). In total, 1,279 markers were available. In the wheat data, two sub-populations were identified by the PSMix software (Wu et al. 2006) using genotypic data.

2.2 Genome-based prediction methods

Following classical quantitative genetic theory, phenotypic values y_i [$(i = 1, \dots, n)$ for n lines] are composed of the genotypic values g_i ($i = 1, \dots, n$) plus a number of environmental factors e_i ($i = 1, \dots, n$). Genome-based prediction aims at estimating unknown genotypic values g_i using genomic marker information. In the context of genome-based prediction, a variety of parametric and non-parametric methods have been proposed and investigated. This thesis concentrates on additive linear models. The adjusted means of the phenotypic traits described in section 2.1 enter the model as phenotypic response variable y_i and g_i is modeled by the sum of p marker effects: $\sum_{j=1}^p w_{ij}\beta_j$, with w_{ij} being the genotype of line i at marker locus j and β_j being the effect of the j th marker. Phenotypic values y_i ($i = 1, \dots, n$) were assumed to be normally distributed for all traits considered. The linear genome-based prediction model for n lines and p markers can be written in matrix notation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1)$$

where \mathbf{y} is the n -dimensional vector of phenotypes. Depending on the data, $\boldsymbol{\mu}$ is one-dimensional and models an overall mean (Lehermeier et al. 2013) or it is K -dimensional (K as the number of populations) and models population effects (Lehermeier et al. 2014; 2015). In the latter cases \mathbf{X} is an $n \times K$ -dimensional matrix which assigns phenotypes to $\boldsymbol{\mu}$. In Lehermeier et al. (2013) \mathbf{X} is an n -dimensional vector of ones. The $n \times p$ -

dimensional matrix \mathbf{W} includes the marker genotypes for the n lines, coded according to the number of a reference allele, and $\boldsymbol{\beta}$ is the p -dimensional vector of marker effects. The n -dimensional vector of residuals is assumed to be normally distributed with $\mathbf{N}(\mathbf{0}, \mathbf{I}_{n \times n} \sigma_\epsilon^2)$, where $\mathbf{I}_{n \times n}$ is the $n \times n$ -dimensional identity matrix and σ_ϵ^2 the residual variance.

With more marker covariates p than phenotypic observations n , which is typically the case in genome-based prediction, model (1) cannot be solved with ordinary least-squares estimation. Instead, marker effects $\boldsymbol{\beta}$ need to be penalized or shrunk towards zero. No penalization is employed on the effects $\boldsymbol{\mu}$; they are considered as so called fixed effects. In Ridge regression best linear unbiased prediction (RR-BLUP), suggested by Meuwissen et al. (2001) for genome-based prediction, the marker effects β_j ($j = 1, \dots, p$) are considered to be normally distributed with a common marker variance σ_β^2 :

$$\beta_j | \sigma_\beta^2 \sim \mathbf{N}(0, \sigma_\beta^2). \quad (2)$$

The RR-BLUP method estimates marker effects $\boldsymbol{\beta}$ by minimizing the residual sum of squares plus a penalty term for the marker effects, which is given by $\sigma_\epsilon^2 / \sigma_\beta^2 \sum_{j=1}^p \beta_j^2$. Here, $\lambda = \sigma_\epsilon^2 / \sigma_\beta^2$, also known as penalty or shrinkage parameter, controls the amount of shrinkage of marker effects β_j towards zero. The unknown variance components σ_β^2 and σ_ϵ^2 can be estimated with REML, or in a Bayesian setting prior distributions are assigned to the variance components. The latter approach has been used in Lehermeier et al. (2013) and Lehermeier et al. (2015). In Lehermeier et al. (2013) the Bayesian setting of the Ridge regression method, denoted as Bayesian Ridge, was applied. Here, inverse- χ^2 prior distributions are used as prior distributions for the variance components:

$$\sigma_\beta^2 | df_\beta, S_\beta \sim \chi^{-2}(df_\beta, S_\beta) \quad (3)$$

$$\sigma_\epsilon^2 | df_\epsilon, S_\epsilon \sim \chi^{-2}(df_\epsilon, S_\epsilon), \quad (4)$$

where df_β and df_ϵ are the degrees of freedom and S_β and S_ϵ are the scale parameters of the inverse- χ^2 distributions. Lehermeier et al. (2013) investigated the influence of the parameterization of these hyperparameters by varying S_β and keeping the other

parameters fixed. The scale parameter S_β influences the amount of shrinkage on marker effects, as it adjusts the expected value and mode of the inverse- χ^2 distribution, being:

$$\mathbb{E}(\sigma_\beta^2) = \frac{df_\beta}{df_\beta - 2} S_\beta \quad (5)$$

$$\text{mode}(\sigma_\beta^2) = \frac{df_\beta}{df_\beta + 2} S_\beta. \quad (6)$$

Thus, a larger scale S_β allows *a priori* greater variance for the marker effects σ_β^2 and yields less shrinkage.

Meuwissen et al. (2001) suggested the methods BayesA and BayesB for genome-based prediction, which were also investigated in Lehermeier et al. (2013). BayesA assumes *a priori* a heterogeneous variance of the marker effects. Thus, the prior setting of BayesA is:

$$\beta_j | \sigma_{\beta_j}^2 \sim \text{N}(0, \sigma_{\beta_j}^2) \quad (7)$$

$$\sigma_{\beta_j}^2 | df_\beta, S_\beta \sim \chi^{-2}(df_\beta, S_\beta). \quad (8)$$

BayesB is an extension of BayesA in as much that an additional parameter π is introduced, and the prior of the variance of the marker effects $\sigma_{\beta_j}^2$ is following mixture distribution:

$$\sigma_{\beta_j}^2 | \pi, df_\beta, S_\beta \sim \pi \delta_0(\cdot) + (1 - \pi) \chi^{-2}(df_\beta, S_\beta), \quad (9)$$

where $\delta_0(\cdot)$ denotes a point mass at zero. Thus, *a priori*, a portion π of marker effects is set to zero. With $\pi = 0$, BayesB reduces to BayesA. In Lehermeier et al. (2013), π of BayesB was set to 0.8 and the influence on model performance of S_β in BayesA and BayesB was investigated.

Additionally to the Bayesian Ridge, BayesA, and BayesB, the Bayesian Lasso (Park and Casella 2008) was explored for genome-based prediction. In the Bayesian Lasso, conditional Gaussian priors with mean zero are assigned to marker effects and, as in BayesA

and BayesB, the variance of a marker effect is specific to a marker locus. But instead of inverse- χ^2 prior distributions, exponential prior distributions are assigned to the variance parameters. The prior setting is as follows:

$$\beta_j | \tau_j, \sigma_\epsilon^2 \sim N(0, \sigma_\epsilon^2 \tau_j^2) \quad (10)$$

$$\tau_j^2 | \lambda \sim \text{Exp}(\lambda^2). \quad (11)$$

Parameter λ , which adjusts the amount of shrinkage on marker effects, can either be set to a fixed value or a prior distribution can be assigned to λ as well. Park and Casella (2008) suggested using a Gamma prior distribution for λ^2 : $\lambda^2 \sim \text{Gamma}(r, \delta)$, with rate parameter r and shape parameter δ . In Lehermeier et al. (2013), the influence of assigning different fixed λ values was investigated, as well as of assigning a Gamma prior distribution with varying values of rate parameter r . With $\text{Gamma}(\lambda^2 | r, \delta)$ as prior distribution of λ^2 , the prior distribution of λ is, according to the density transformation theorem, $2\lambda \text{Gamma}(\lambda | r, \delta)$. Thus, the prior mode of λ is $\sqrt{(2\delta - 1)/(2r)}$, and a smaller rate parameter r yields more shrinkage for the marker effects.

In all of the Bayesian methods presented so far, conditional Gaussian priors have been assigned to the marker effects. However, the marginal prior distributions for the marker effects differ according to each method. When a fixed value for σ_β^2 is set, as in RR-BLUP, the prior distribution for the marker effects is Gaussian with mean zero. When inverse- χ^2 prior distributions are used for the variances of the marker effects, as in the Bayesian Ridge and BayesA, the marginal distribution of the marker effects is a scaled Student-t distribution $t(0, df_\beta, S_\beta)$. For BayesB, the marginal prior distribution of the marker effects is a mixture distribution: $\beta_j | df_\beta, S_\beta \sim \pi \delta_0(\cdot) + (1 - \pi)t(0, df_\beta, S_\beta)$. The Bayesian Lasso, given a fixed λ , has a Laplace, also called a double-exponential, distribution as a marginal prior: $\beta_j | \lambda \sim \text{Lap}(\lambda)$. Figure 3 illustrates the densities of the different marginal prior distributions.

It is evident that the mass around zero increases when moving from the Gaussian distribution (RR-BLUP) to the t-distribution (Bayesian Ridge, BayesA), the Laplace distribution (Bayesian Lasso), and to the mixture distribution of BayesB, given $\pi = 0.8$ (de los Campos et al. 2013a). The higher the probability mass around zero, the stronger the small marker effects are shrunken towards zero.

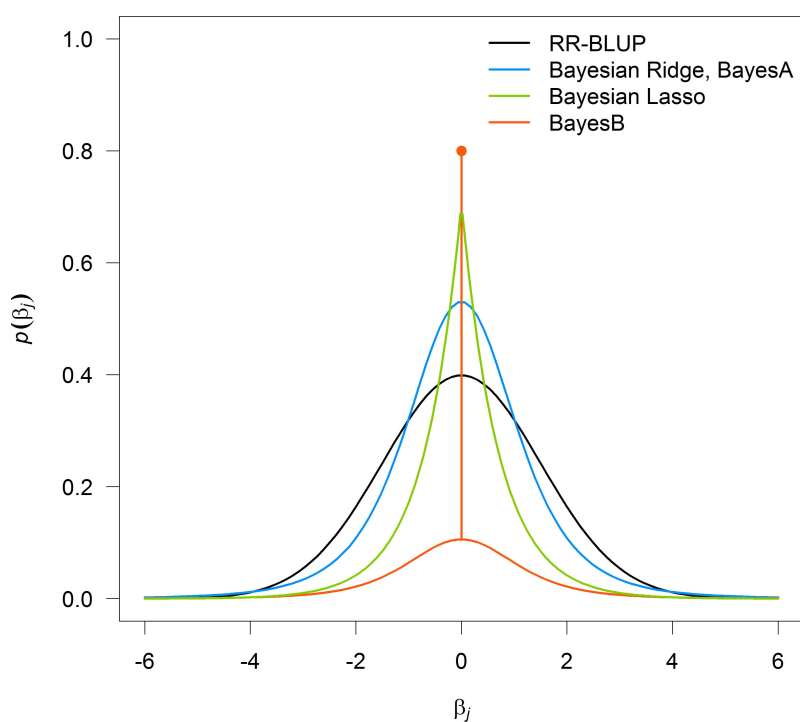


Figure 3: Marginal prior densities of the marker effects in RR-BLUP, Bayesian Ridge, BayesA, Bayesian Lasso, and BayesB. All prior densities are illustrated with zero mean and unit variance, and for the scaled-t density (BayesA, Bayesian Ridge) four degrees of freedom were chosen.

The most commonly used method in genome-based prediction is the RR-BLUP method. When there are more markers to estimate than there are phenotypes in the model ($p > n$), it is computationally more efficient to convey the RR-BLUP method to the so called genome-based best linear unbiased prediction (GBLUP) method:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon}, \quad (12)$$

where \mathbf{g} is the n -dimensional vector of genotypic effects, or, when testcrosses are considered, as in the analyzed maize data, the vector of testcross effects and \mathbf{Z} is an $n \times n$ -dimensional matrix assigning phenotypes to genotypic effects. All other variables are defined as in model (1). The vector of genotypic effects \mathbf{g} is assumed to follow: $\mathbf{g} | \sigma_g^2 \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$, where \mathbf{G} is a genomic relationship matrix and σ_g^2 is the genotypic variance or the testcross variance when testcrosses are considered. The genomic relationship matrix \mathbf{G} is calculated using marker data \mathbf{W} by: $1/V(\mathbf{W}-\mathbf{P})(\mathbf{W}-\mathbf{P})'$, where V is a scaling factor and the $n \times p$ -dimensional matrix \mathbf{P} centers marker genotypes to zero mean (Habier et al. 2007). Habier et al. (2007) and VanRaden (2008) suggested to use $V = \sum_{j=1}^p 2p_j(1-p_j)$ as a scaling factor, where p_j is the frequency of the reference allele of marker j . Using scaled marker genotypes with unit variance, the scaling factor V reduces to the number of markers p (Aistle and Balding 2009). Further, V is modified depending on the specific data situation.

When sub-populations are present in the dataset, it is questionable whether one should assume that marker effects are equal across sub-populations. Following classical quantitative genetics theory, the additive effect (also called allele substitution effect) at a marker locus depends on the allele frequencies in a population, if dominance and/or epistasis are present (Falconer and Mackay 1996). In addition, it is a rather strict assumption that the genotypic variance σ_g^2 and the residual variance σ_ϵ^2 are constant over all sub-populations. Thus, when sub-populations are present in the data, a multivariate model for genome-based prediction might be more suitable (Lehermeier et al. 2015). The model estimates population-specific marker effects using:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} \mu_1 \\ \mathbf{1}_{n_2} \mu_2 \\ \vdots \\ \mathbf{1}_{n_K} \mu_K \end{pmatrix} + \begin{pmatrix} \mathbf{W}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_K \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_K \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_K \end{pmatrix}, \quad (13)$$

where \mathbf{y}_k ($k = 1, \dots, K$) is the n_k -dimensional vector of phenotypic values of sub-population k , μ_k is an intercept specific for sub-population k with $\mathbf{1}_{n_k}$ as n_k -dimensional vector of ones, \mathbf{W}_k ($k = 1, \dots, K$) is the $(n_k \times p)$ -dimensional marker matrix of sub-population k and $\boldsymbol{\beta}_k$ ($k = 1, \dots, K$) is the p -dimensional vector of marker effects specific for sub-population k . The complete vector of marker effects $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \dots, \boldsymbol{\beta}'_K)'$ is assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix $\mathbf{B} \otimes \mathbf{I}_{p \times p}$: $\boldsymbol{\beta} \sim \text{MVN}_{K \cdot p \times K \cdot p}(\mathbf{0}, \mathbf{B} \otimes \mathbf{I}_{p \times p})$. Here, \otimes denotes the Kronecker product, $\mathbf{I}_{p \times p}$ is the $p \times p$ -dimensional identity matrix, and \mathbf{B} is the variance-covariance matrix of the marker effects among sub-populations:

$$\mathbf{B} = \begin{pmatrix} \sigma_{\beta_1}^2 & \sigma_{\beta_{12}} & \cdots & \sigma_{\beta_{1K}} \\ \sigma_{\beta_{21}} & \sigma_{\beta_2}^2 & \cdots & \sigma_{\beta_{2K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\beta_{K1}} & \sigma_{\beta_{K2}} & \cdots & \sigma_{\beta_K}^2 \end{pmatrix}. \quad (14)$$

As for model (1), model (13) can also be reformed to an analogous GBLUP model similar to model (12). Hence, an augmented vector of genotypic values needs to be formed which contains the genotypic value of each line in each sub-population: $\mathbf{g}^* = (\mathbf{g}'_1, \mathbf{g}'_2, \dots, \mathbf{g}'_K)'$, with $\mathbf{g}_k^* = \mathbf{W}_k \boldsymbol{\beta}_k$, where $\mathbf{W} = (\mathbf{W}'_1, \mathbf{W}'_2, \dots, \mathbf{W}'_K)'$ is the full marker matrix of all $n = \sum_{k=1}^K n_k$ lines in the sample. Thus, the full augmented vector \mathbf{g}^* can be expressed as $\mathbf{g}^* = (\mathbf{I}_{K \times K} \otimes \mathbf{W}) \boldsymbol{\beta}$. As the vector of marker effects $\boldsymbol{\beta}$ follows a multivariate normal distribution, \mathbf{g}^* , as linear combination of $\boldsymbol{\beta}$, is multivariate normal as well. The mean and covariance of \mathbf{g}^* can be derived as:

$$\mathbb{E}(\mathbf{g}^*) = (\mathbf{I}_{K \times K} \otimes \mathbf{W}) \mathbb{E}(\boldsymbol{\beta}) = \mathbf{0} \quad (15)$$

$$\text{Cov}(\mathbf{g}^*) = (\mathbf{I}_{K \times K} \otimes \mathbf{W}) (\mathbf{B} \otimes \mathbf{I}_{p \times p}) (\mathbf{I}_{K \times K} \otimes \mathbf{W}') \quad (16)$$

$$= (\mathbf{B} \otimes \mathbf{W}) (\mathbf{I}_{K \times K} \otimes \mathbf{W}') \quad (17)$$

$$= (\boldsymbol{\Sigma}_g \otimes \mathbf{G}). \quad (18)$$

Thus, $\mathbf{g}^* | \boldsymbol{\Sigma}_g \sim \text{MVN}_{(n \cdot k) \times (n \cdot k)}(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \mathbf{G})$, where $\mathbf{G} = 1/V(\mathbf{W} - \mathbf{P})(\mathbf{W} - \mathbf{P})'$ is the genomic relationship matrix and $\boldsymbol{\Sigma}_g = V \cdot \mathbf{B}$ is the genomic variance-covariance matrix among sub-populations:

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \sigma_{g_1}^2 & \sigma_{g_{12}} & \cdots & \sigma_{g_{1K}} \\ \sigma_{g_{21}} & \sigma_{g_2}^2 & \cdots & \sigma_{g_{2K}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{g_{K1}} & \sigma_{g_{K2}} & \cdots & \sigma_{g_K}^2 \end{pmatrix} \quad (19)$$

Only some of the entries of \mathbf{g}^* are really linked to phenotypes, and so the multivariate GBLUP model is:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} \mu_1 \\ \mathbf{1}_{n_2} \mu_2 \\ \vdots \\ \mathbf{1}_{n_K} \mu_K \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 \mathbf{g}_1^* \\ \mathbf{Z}_2 \mathbf{g}_2^* \\ \vdots \\ \mathbf{Z}_K \mathbf{g}_K^* \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_K \end{pmatrix}, \quad (20)$$

where \mathbf{Z}_k is an $n_k \times n$ -dimensional matrix linking phenotypes with the entries of \mathbf{g}_k^* . In a Bayesian setting, the genotypic variance-covariance matrix $\boldsymbol{\Sigma}_g$ is assigned an inverse-Wishart prior distribution: $\boldsymbol{\Sigma}_g \sim W^{-1}(\Psi, \nu)$. The inverse-Wishart distribution is the multivariate analogon of the inverse- χ^2 distribution. Thus, the model is a multivariate version of the Bayesian Ridge or Bayesian GBLUP method.

The population-specific residuals $\boldsymbol{\epsilon}_k$ ($k = 1, \dots, K$) are assumed to be uncorrelated and to follow a normal distribution with population-specific residual variances:

$\boldsymbol{\epsilon}_k \sim \text{N}(\mathbf{0}, \mathbf{I}_{n_k \times n_k} \sigma_{\epsilon_k}^2)$. Identical inverse- χ^2 prior distributions are assigned to the residual

variances for all $k = 1, \dots, K$: $\sigma_{\epsilon_k}^2 \sim \chi^{-2}(df_0, S_0)$, leading to the following joint prior distribution for the complete vector of residuals $\epsilon = (\epsilon'_1, \dots, \epsilon'_K)'$:

$$p(\epsilon, \sigma_{\epsilon_1}^2, \dots, \sigma_{\epsilon_K}^2) = \prod_{k=1}^K \prod_{i=1}^{n_k} N(\epsilon_{k_i} | \mathbf{0}, \sigma_{\epsilon_k}^2) \chi^{-2}(\sigma_{\epsilon_k}^2 | df_0, S_0). \quad (21)$$

In Bayesian analysis, estimates for all unknown parameters Ω are obtained using the joint posterior distribution of Ω , given the data \mathbf{y} . Using Bayes' theorem, the joint posterior distribution is proportional to the product of the likelihood of the data and the prior distribution:

$$p(\Omega | \mathbf{y}) \propto \prod_{i=1}^n N\left(y_i | \sum_{k=1}^K x_{ik} \mu_k + \sum_{j=1}^p w_{ij} \beta_j, \sigma_{\epsilon}^2\right) \times p(\Omega) \quad (22)$$

With unknown variance components, this posterior distribution does not have a closed form; however, samples can be obtained using Markov chain Monte Carlo (MCMC) algorithms (Gelman et al. 2004). When the full conditional posterior distributions of each unknown parameter are known, a Gibbs sampler can be used to obtain samples from the posterior distribution. Here, it is in turn sampled from the full conditional posterior distribution of each unknown parameter. After a convergence phase (burn in), these samples can be seen as samples from the joint posterior distribution. To make sure that the algorithm converged in this study, sample paths were checked visually and the Geweke diagnostic (Bernardo and Smith 2009) was used. Point estimates of the unknown parameters were obtained by forming the mean of the post burn in samples. In the RR-BLUP or GBLUP method with fixed variance components, the solution of the covariate effects has a closed form and can be derived analytically.

2.3 Evaluation of genome-based prediction methods

2.3.1 Bayesian learning ability

To evaluate the Bayesian learning ability (Sorensen and Gianola 2002) of different genome-based prediction methods, the Hellinger distance (Le Cam 1986) can be used, as it measures the distance between two densities f and l by:

$$H(f, l) = \sqrt{\frac{1}{2} \int_{-\infty}^{\infty} (\sqrt{f(u)} - \sqrt{l(u)})^2 du}. \quad (23)$$

Marginal posterior densities of the marker effects were estimated using the post burn in MCMC samples via kernel density estimation. The Hellinger distances between marginal prior (f) and posterior (l) densities of the marker effects—estimated by Bayesian Ridge, Bayesian Lasso, BayesA, and BayesB—were calculated numerically. The integral in H was approximated by the trapezoidal rule. The values of $H(f, l)$ can take values from 0 to 1, where a value of 0 is taken when $f = l$, indicating that no Bayesian learning took place as the posterior density did not move away from the prior. Alternatively, this would also be the case when the prior density already perfectly explained the data.

2.3.2 Prediction performance

To evaluate the prediction performance of the different methods, cross-validation was used. In cross-validation, the dataset is split into different subsets, using one subset as a test set and the other subsets as an estimation set. The data in the estimation set are used to estimate all unknown parameters, and the estimated values are then used to predict the genotypic values of the test set. Different validation schemes were used for addressing the questions in this thesis. To compare the prediction performance of different methods, five-fold cross-validation (5-fold CV) was conducted within the complete Synbreed CS1 dataset (Lehermeier et al. 2013) and within every Cornfed family separately (Lehermeier et al. 2014). Here, the dataset was split randomly into five subsets, where one subset formed the test set and the other subsets the estimation set. Every subset formed the test set once, so that model estimation was conducted five times. To compare the prediction performance of the different Cornfed families, when the estima-

tion set size is constant, 50 lines of one family were used randomly as an estimation set, and the rest of the lines from the same family were then deployed as a test set (R-CV). For every family, this procedure was randomly repeated 50 times, and estimated prediction performance was averaged over the 50 replications. In this prediction setting within families, highly related full-sib lines are always used in the estimation and test sets. To assess if lines of one bi-parental family can be predicted by using the progeny from the other crosses as an estimation set, leave-one-cross-out cross-validation (LOCO-CV) was used (Lehermeier et al. 2014). Here, one family formed the test set and all other families from the same half-sib panel (dent or flint) formed the estimation set. The influence of estimation set size was investigated by randomly sampling 25 to 675 half-sib lines in increments of 25 in the estimation set. For every estimation set size, sampling was randomly repeated 100 times and prediction performance was averaged over replications. To evaluate if adding half-sib lines to an estimation set of full-sib lines could further increase prediction performance by increasing estimation set size, all available half-sib lines were additionally included in the estimation sets of R-CV. The prediction performances of these estimation sets were compared to the prediction performances of R-CV, where only full-sib lines were used in the estimation set. For investigating prediction performance across heterotic pools, all Cornfed dent lines were used as an estimation set and all flint lines as a test set, and vice versa. Additionally, cross-with-cross (CwC) prediction was performed (Lehermeier et al. 2014). Here, every Cornfed family was once used as an estimation set and once as a test set. In order to attain comparable estimation and test set sizes, 50 lines from each family were randomly sampled, to form the estimation and test sets. Sampling was randomly repeated 100 times.

To investigate the prediction performance of a multivariate model for genome-based prediction in data with a sub-population structure, half of the lines from every sub-population were randomly sampled in the estimation set, and the rest of the lines formed the test set (Lehermeier et al. 2015). Prediction performance was evaluated within each sub-population separately. Sampling was randomly repeated 100 times, and estimated prediction performance was averaged over the 100 replications.

In all datasets and validation settings, prediction performance was estimated as Pearson correlation between observed phenotypic values (\mathbf{y}_{TS}) and predicted genotypic values of the test set ($\hat{\mathbf{g}}_{TS}$): $\text{cor}(\mathbf{y}_{TS}, \hat{\mathbf{g}}_{TS})$. In the context of genome-based prediction, this estimate is frequently denoted as predictive ability. Generally, one is interested in how close the

predicted genotypic value is to the true genotypic value (\mathbf{g}_{TS}), thus, on $\text{cor}(\mathbf{g}_{TS}, \hat{\mathbf{g}}_{TS})$. This estimate is frequently denoted as prediction accuracy. Only in simulated datasets, where the true genotypic values are known, as in the simulated datasets in Lehermeier et al. (2013), can this correlation be calculated directly. In experimental data, true genotypic values are unknown and the prediction accuracy needs to be approximated, according to Dekkers (2007), by:

$$\text{cor}(\mathbf{g}_{TS}, \hat{\mathbf{g}}_{TS}) = \frac{\text{cor}(\mathbf{y}_{TS}, \hat{\mathbf{g}}_{TS})}{\sqrt{h_{TS}^2}}, \quad (24)$$

where h_{TS}^2 is the heritability in the test set.

3 Discussion

This chapter combines a discussion of the results from the three publications underlying this thesis (Lehermeier et al. 2013, 2014, and 2015) and includes further explanations and results.

3.1 Proposed methods for genome-based prediction

The statistical method used for genome-based prediction has to cope with a large number of parameters which need to be estimated and generally by far exceed the number of observations. Many methods have been proposed for this purpose. One large group of methods, employed in this thesis, is formed by parametric models which assume *a priori* a linear and additive effect of the markers on the phenotype (reviewed in de los Campos et al. (2013a)). This group of parametric models can be further separated into penalized frequentist regression models, as Ridge regression (RR-BLUP), GBLUP, LASSO (Tibshirani 1996), and elastic net (Zou and Hastie 2005); and Bayesian shrinkage models, discussed in Gianola (2013) as the Bayesian alphabet. Another group of methods is represented by machine learning methods which do not assume *a priori* a linear and additive effect of the markers on the phenotypes. Reproducing kernel Hilbert space regression (RKHS), support vector machines (SVM), boosting, random forest (RF), and artificial neural networks (ANN) belong to these machine learning methods which have been applied in genome-based prediction (Gonzalez-Recio et al. 2014).

Figure 4 shows the frequency of the different methods used for genome-based prediction in plant breeding, according to their application to experimental data in peer-reviewed publications. Here, the most frequently used method is GBLUP, or the equivalent Ridge regression BLUP (RR-BLUP).

3.1.1 Comparison of method performance

In Lehermeier et al. (2013) the prediction performance of four Bayesian shrinkage methods—Bayesian Ridge, Bayesian Lasso, BayesA, and BayesB—has been compared. The Bayesian Ridge method, which assigns *a priori* equal shrinkage to all markers (i.e. marker-homogeneous shrinkage), yielded in all datasets equal or higher predictive abilities than the other three methods. Irrespective of the number of markers and observations,

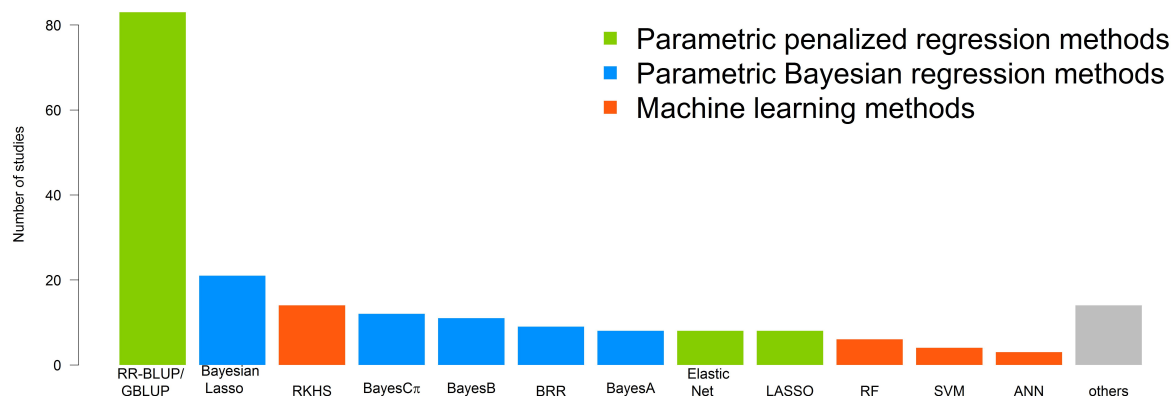


Figure 4: Investigated genome-based prediction methods as used in published studies on different experimental plant datasets according to Table A1. RKHS: reproducing kernel Hilbert space regression; BRR: Bayesian Ridge; RF: random forest; SVM: support vector machines; ANN: artificial neural networks.

marker-specific shrinkage did not outperform marker-homogeneous shrinkage in this study. Performance in the experimental dataset was similar to the simulated datasets for both traits. One reason for the good performance of the Bayesian Ridge might be the large number of QTL affecting the target traits. The two quantitative traits in the experimental data are assumed to be affected by many QTL (Schön et al. 2004). Both simulated datasets comprised more than 300 segregating QTL, leading to a ratio relating to the number of QTL per number of lines in the estimation set (n_{QTL}/n_{ES}) larger than 0.3. Wimmer et al. (2013) showed, based on published simulation experiments (Zhong et al. 2009; Meuwissen and Goddard 2010; Daetwyler et al. 2010; Zhang et al. 2010), the clear influence of n_{QTL}/n_{ES} on the relative performance of BayesB (performing marker-heterogeneous shrinkage) compared to RR-BLUP (performing marker-homogeneous), with a BayesB superiority only for n_{QTL}/n_{ES} ratios < 0.5 .

Furthermore, long range LD among markers is pervasive in maize breeding populations. This was also observed in the experimental maize data investigated in Lehermeier et al. (2013). If there is strong LD, many SNPs are expected to be in LD with at least one QTL. It is conjectured that not only the large number of QTL, but also the strong correlation between markers arising from LD are reasons for the superiority of the Bayesian Ridge method over marker-specific shrinkage methods in terms of predictive abilities. Results

from Wimmer et al. (2013) also suggest that large LD does not permit marker-specific shrinkage methods like BayesB to perform efficient variable selection. If interest lies mainly in the prediction of phenotypic traits, there may be little difference if a small effect is assigned to a group of highly correlated markers, as happens in the Bayesian Ridge, or a larger effect is assigned to only one of them, as in BayesB.

However, it was suggested in the literature that parametric variable selection methods, for example BayesB, BayesC π , or LASSO, can be used in a broader sense and are superior to the marker-homogeneous shrinkage methods GBLUP or RR-BLUP for predictions across families, populations, and/or generations (Meuwissen 2009; Daetwyler et al. 2013). The assumption here is that variable selection methods rely more on information from LD whereas GBLUP mainly uses information from relatedness (Habier et al. 2007, 2010). Thus, two variable selection methods have been tested and compared with GBLUP for prediction across families and pools with the Cornfed data (Lehermeier et al. 2014). As a frequentist approach, LASSO was used, which shrinks effects directly to zero (Tibshirani 1996); and as a Bayesian approach, BayesC π was used, which is a generalization of the Bayesian Ridge and includes a random prior probability π of setting a marker effect to zero (Habier et al. 2011). The prediction performance of BayesC π was highly similar to GBLUP for all cases. LASSO showed more differences compared to GBLUP and BayesC π , but in general no superiority compared to GBLUP was observed. The reason why no superiority of variable selection methods was observed for prediction across populations might be that variable selection methods are successful if the trait is only affected by a small number of QTL, and if LD among markers is low (Wimmer et al. 2013), which was not given for the Cornfed data. Furthermore, the advantage of GBLUP compared to variable selection methods for predictions across populations may be that it does not lose variables that could be useful for a distantly related test set. Also, other studies showed an inferior BayesC π performance compared to GBLUP for genome-based prediction (Heslot et al. 2012), especially for predictions across families, as in a study with oats by Asoro et al. (2011).

It has also been conjectured that the performance of Ridge regression-type methods may change compared to that of marker-heterogeneous shrinkage methods when marker coverage is more dense (de los Campos et al. 2009). However, with denser marker coverage, the ratio p/n will increase further. This exemplifies the importance of choosing appropriate model parameters and good Bayesian learning ability of the methods, which

will be discussed in the next two chapters.

3.1.2 Choice of model parameters

In Lehermeier et al. (2013), the influence of the choice of hyperparameters of Bayesian Ridge, Bayesian Lasso, BayesA, and BayesB on prediction performance was investigated. Hyperparameters which are chosen for the prior densities control the amount of shrinkage on the marker effects. The parameter choice showed a high impact on prediction performance for some methods. If the prior parameter setting was appropriate, the predictive abilities and accuracies of BayesA and BayesB were high and equal to those of the Bayesian Ridge and Bayesian Lasso with random λ . However, finding “optimal” parameters is not straightforward. The proposed guidelines for finding an “optimal” scale S_β and λ , according to Pérez et al. (2010), did not always yield the best parameter setting in terms of predictive ability. Alternative formulas have been proposed for finding an “optimal” scale parameter S_β , e.g. in Habier et al. (2010, 2011), but both approaches are based on strong assumptions, e.g. the independence of marker effects, which may be inadequate if strong LD among markers translates into a joint dependence of their effects.

In experimental data, there is additional uncertainty in finding “optimal” parameters, because the genotypic and residual variance components are unknown. In practical applications of genome-based prediction, variance components can only be estimated based on the training dataset, and not on phenotypic values of the test dataset. Thus, there may be additional uncertainty, as the data distribution may change from the training to the test set, especially when training and test data come from different populations or generations. In Lehermeier et al. (2013), hyperparameters were chosen based on the mode of the prior distributions. An option would be to choose hyperparameters based on the mean of the prior densities, which would change the formulas for finding hyperparameters and gives additional uncertainty for the hyperparameter choice. An alternative to using an ad hoc formula would be to find hyperparameters iteratively via cross-validation within the estimation set, but this would mean much higher computational demands (de los Campos et al. 2013a). Hence, Bayesian methods that are robust with respect to the choice of hyperparameters are highly desirable. For practical applications it is recommended to choose a less informative parameter setting and rather weak prior distributions, and to investigate the influence of the prior choice. Therefore, it is

especially important to conduct some kind of prior robustness diagnostic (Roos et al. 2015).

3.1.3 Bayesian learning ability

In Lehermeier et al. (2013), the strong influence of the choice of hyperparameters was observed. These findings indicate a lack of Bayesian learning ability, which was also discussed by Gianola et al. (2009). To quantify the ability of Bayesian learning for the respective methods, the Hellinger distance between the marginal prior and the posterior densities of marker effects for the simulated dataset maizeA was calculated in Lehermeier et al. (2013). The Hellinger distance is related to the Kullback Leibler distance, although opposed to the Kullback Leibler distance it is a symmetric measure and is defined for the whole range of the two densities which are compared (Shemyakin 2014). A greater distance indicates that the posterior density has moved away from the prior density, and that Bayesian learning has taken place. For the BayesA and BayesB methods, quite small distances were observed, whereas in the Bayesian Ridge and the Bayesian Lasso the distance between prior and posterior density was much larger. A small distance between prior and posterior density can of course also be the result of a perfect prior density which was assigned. However, this happens with probability close to zero, if prior knowledge is limited. In combination with partly reduced predictive ability, and the fact that all BayesA and BayesB scenarios yielded a small Hellinger distance between prior and posterior density, irrespective of the hyperparameter setting, this is very unlikely to be the reason for the small distances found in Lehermeier et al. (2013). From the Hellinger distances it can be seen that BayesA and BayesB have less Bayesian learning ability than the Bayesian Ridge and the Bayesian Lasso, and so the influence of the choice of hyperparameters on prediction is significant. In BayesA and BayesB, the degrees of freedom of the fully conditional posterior distribution of $\sigma_{\beta_j}^2$ are $df_{\beta} + 1$, and thus only one degree of freedom higher than the degrees of freedom of the prior distribution of $\sigma_{\beta_j}^2$, independently of the number of observations (n) or markers (p) in the model (Gianola et al. 2009). In contrast, in the Bayesian Ridge, the degrees of freedom of σ_{β}^2 increases with the number of markers in the model.

In genomic datasets, Bayesian learning is limited due to the $p > n$ situation. In real life, as with next generation sequencing data, p will get even larger than in the simulated and experimental data investigated here, and it is expected to increase much more than

n , which is also seen in Figure 1. Thus, methods with a strong Bayesian learning ability, such as the Bayesian Ridge and Bayesian Lasso, are required for genome-based prediction. Only a few studies have investigated the sensitivity of Bayesian methods regarding prior specification in the context of genome-based prediction (e.g. Knürr et al. (2013), Lehermeier et al. (2013), Yang et al. (2015)), and more emphasis should be placed on sensitivity analyses, especially if the Bayesian alphabet (Gianola 2013) is extended.

3.2 Population-specific factors affecting prediction performance

For the allocation of resources available in a breeding program, it would be highly desirable if one could predict the accuracy of a genome-based prediction experiment in advance, based on known factors. Different deterministic formulas and variations thereof have been proposed for this purpose which take into account the number of individuals in the estimation set, the trait heritability, the number of independently segregating chromosome segments M_e (depending on effective sample size), and partly also the number of markers (Daetwyler et al. 2008; Goddard 2009; Daetwyler et al. 2010; Goddard et al. 2011; Erbe et al. 2013). Using these deterministic formulas and selection theory, Riedelsheimer and Melchinger (2013) derived how to optimize analytically genomic selection strategies within one breeding cycle and bi-parental family. However, uncertainty about the correct choice of M_e exists in these deterministic formulas. Additionally, they assume that the estimation and test set are composed of individuals from the same population without structure, as would be the case, for example, for predictions within full-sib families. Following Henderson (1963), VanRaden (2008) proposed a formula based on the genomic relationship matrix, in order to assess how reliably the genomic values for each individual in the test set will be predicted. De los Campos et al. (2013b) adopt this formula to derive upper bounds of prediction accuracy, under perfect and imperfect LD between markers and QTL. Wientjes et al. (2012) investigate how the deterministic formulas of Daetwyler et al. (2008) and VanRaden (2008) are affected by LD differences and complex family structure. Recently, Wientjes et al. (2015) extended the formula of VanRaden (2008) for prediction across populations.

Although various formulas have been proposed, assessing the accuracy of prediction still remains difficult, and it is not yet fully understood in what manner prediction accuracy is affected by population-specific parameters. In the following, population-specific factors

that are supposed to affect genome-based prediction accuracy are examined further.

3.2.1 Linkage disequilibrium and linkage phases

Unless high-quality whole-genome sequencing data are used for genome-based prediction, functional polymorphisms might not be directly included as predictors in the model. However, markers are included which are assumed to be in linkage disequilibrium (LD) with the QTL affecting the trait. LD is defined as non-random association of alleles at two loci on the same gamete, while common measures of LD are based on the difference of the frequency of the haplotype AB (p_{AB}) and the frequency which would be expected under linkage equilibrium: $D_{AB} = p_{AB} - p_A p_B$, where p_A is the allele frequency of allele A at the first locus, and p_B is the allele frequency of allele B at the second locus. A crucial assumption for efficient genome-based prediction is that the LD between the marker and QTL given in the estimation set is also given in the test set. And even more importantly, linkage phases need to be consistent between the estimation set and selection candidates, meaning the sign of D_{AB} between marker and QTL need to be equal for both populations (Goddard and Hayes 2007; de Roos et al. 2009; Hayes et al. 2009).

In the Cornfed data, which comprises bi-parental and multi-parental populations, different factors contribute to the creation of LD. Within a single bi-parental DH family, population structure is absent and LD is generated purely by linkage. For DH lines derived from the same cross of two homozygous parents, the LD between two polymorphic loci can be calculated based on the recombination frequency r as $\pm 0.25(1 - 2r)$, with the sign depending on the linkage phase of the parents (same linkage phase of both parental gametes returns a positive value). Thus, D_{AB} can take values 0.25 or -0.25, if no recombination takes place ($r = 0$). If at least one locus is monomorphic, or if recombination frequency is at its maximum value of 0.5, $D_{AB} = 0$. Two bi-parental families with one common parent are expected to have equal linkage phases, because only gametes with the same linkage phase will show polymorphism with the common parent at both loci. In Lehermeier et al. (2014), this was shown in the pairwise comparison of linkage phases between families from the same heterotic pool, where, as expected, all SNP pairs had equal linkage phases. If two bi-parental families are combined, the arising LD is calculated as the weighted mean of the LD within each family plus the LD that arises due to admixture because of differences in allele frequencies (Nei and Li 1973):

$$D_{AB} = \pi_1 D_{AB_1} + \pi_2 D_{AB_2} + \pi_1 \pi_2 (p_{A_1} - p_{A_2})(p_{B_1} - p_{B_2}), \quad (25)$$

where π_1 (π_2) is the frequency of family 1 (2) in the admixed population, D_{AB_1} (D_{AB_2}) is the value of LD in family 1 (2), p_{A_1} (p_{A_2}) and p_{B_1} (p_{B_2}) are the frequencies of alleles A and B in family 1 (2), respectively. Accordingly, if two half-sib families are combined, the sign of D_{AB} can only change compared to a third family if allele frequencies differ between the two families, because D_{AB_1} and D_{AB_2} have equal signs. Allele frequencies within biparental families are expected to be 0.5 for segregating loci, but differences emerge if one family segregates whereas the other family does not segregate for specific loci. If multiple families (say K families) are combined, as in the estimation set of LOCO-CV in Lehermeier et al. (2014), with frequencies π_k ($k = 1, \dots, K; \sum_{k=1}^K \pi_k = 1$), the above formula for the LD in the combined population can be extended according to Charcosset and Essioux (1994) to:

$$D_{AB} = \sum_{k=1}^K \pi_k D_{AB_k} + \sum_{k=1}^K \pi_k \left(p_{A_k} - \sum_{k=1}^K \pi_k p_{A_k} \right) \left(p_{B_k} - \sum_{k=1}^K \pi_k p_{B_k} \right). \quad (26)$$

Here, $\sum_{k=1}^K \pi_k p_{A_k}$ and $\sum_{k=1}^K \pi_k p_{B_k}$ are the allele frequencies p_A and p_B in the combined population. If A and B denote the alleles at locus 1 and locus 2 of the common parent of K half-sib families, all families have LD values of $D_{AB_k} \geq 0$ ($k = 1, \dots, K$), and the allele frequencies p_A and p_B are expected to be equal to or larger than 0.5. Thus, for loci in linkage equilibrium within each family ($D_{AB_k} = 0$, for all $k = 1, \dots, K$) or with low LD, D_{AB} can become negative if at least one family does not segregate for one locus ($p_{A_k} = 1$ and $p_{B_k} = 0.5$ or vice versa). This leads to differences in allele frequencies for the different families and consequently to the creation of negative LD in a combination of multiple half-sib families, due to admixture. However, if there is strong LD within families, the probability that linkage phases change from one full-sib family compared to a combination of families connected through a common parent is low. By investigating the linkage phases of each Cornfed dent (flint) family, compared to the combination of all other Cornfed dent (flint) families, it was observed that the concordance of linkage phases was, in general, very high for SNP pairs less than five mega base pairs apart. However,

variations were observed for different families. This variation of linkage phases was related to the genetic similarity of a founder line (the parental line which is specific for a family) with all other founder lines. The average simple matching coefficient (Sneath and Sokal 1973) of one dent (flint) founder line with the other dent (flint) founder lines was highly correlated with the concordance of linkage phases, thereby reflecting the relationship between the variation in linkage phases and SNP allele frequencies in the admixed estimation sets of LOCO-CV. As expected from the high correlation of the calculated mean simple matching coefficients between one founder line and the other founder lines and the fraction of equal linkage phases, both factors were similarly associated with predictive abilities from LOCO-CV.

3.2.2 Relatedness

The impact of the relatedness of the founder lines contributing to the estimation and test set was investigated further using results from CwC prediction within both half-sib panels. Here, linkage phases are identical between estimation and test set, and differences in predictive abilities are expected to emerge only due to differences in relatedness. As a general trend, it was observed that high relatedness between founder lines tended to result in high prediction performance and low relatedness to yielding low prediction performance. However, the association between relatedness and predictive ability was generally quite weak. This confirms results from Daetwyler et al. (2013), who found associations between relatedness and accuracy at the “macro level” but not at the “micro level”. Previous results from animal breeding showed high correlations between relatedness and accuracy (Clark et al. 2012; Pszczola et al. 2012). However, in these studies accuracy was derived from prediction error variance (PEV) and not from cross-validated correlations. Several studies have shown the equality between PEV and accuracy (e.g. Clark et al. (2012); Pszczola et al. (2012)). However, this equality just holds if data distribution does not change from the estimation set to the test set with equal genotypic and residual variance components of the GBLUP method for both populations. This is quite unlikely if the test set and the estimation set are formed by distantly related populations. In the Cornfed data, large differences in genotypic and residual variance components were observed for different bi-parental families, even if they were half-sibs and belonged to the same heterotic pool. Consequently, the results derived from PEV are restricted to special cases, as PEV does not take into account the future prediction

errors of independent lines that were not included in the estimation model. Thus, cross-validated results provide more realistic findings for prediction scenarios compared to investigations of PEV. For this purpose, Gianola et al. (2014) suggested using bagging to derive cross-validated prediction errors for each individual in the test set, and observed little concordance between the theoretical model-based PEV and the empirically derived PEV.

In Lehermeier et al. (2014), it was also investigated how predictive ability is affected when additional unrelated lines are included in the estimation set consisting of related half-sib lines. Here including additional lines from the opposite pool (i.e., unrelated crosses with potentially different linkage phases) to the estimation set of LOCO-CV had no impact on predictive abilities. Thus, including completely unrelated crosses in the estimation set was neither a benefit nor a drawback for prediction performance. This is relatively surprising, as one might expect that including unrelated crosses in the estimation set would lead to a decrease in prediction performance, as different linkage phases between markers and QTL might exist which lead to incorrectly estimated marker effects. Different linkage phases between markers and QTL across the dent and flint pool in the Cornfed data are most likely considering the observed opposite linkage phases between markers. However, as the relationship between dent and flint lines observed by markers was very small, the GBLUP method does not borrow a lot of information from the other pool when predicting lines from the same pool. Accordingly, it was also stated by de los Campos et al. (2013b) that data from unrelated individuals contribute little to prediction performance. In contrast, including additional half-sib families in the estimation set for prediction within bi-parental families, which show relatedness with the estimation and test set, considerably increased predictive abilities.

Interestingly, in the CwC procedure, conducted in Lehermeier et al. (2014), it was observed that prediction performance was relatively high across families derived from lines originating from the same breeding program, although the lines belonged to different heterotic pools, and only small values of relatedness could be observed in the marker data. This could be due to similar selection strategies and adaptation to similar environmental conditions within one breeding program. Here, relatedness might be higher at causal loci than observed at marker data.

3.2.3 Allele frequencies and the allele substitution effect

Considering a biallelic context, in a random mating population the allele substitution effect of a QTL for a given trait corresponds to the least squares estimate of the QTL effect on the phenotype and depends on half the difference between the genotypic values of both homozygotes at the QTL a , the dominance deviation d , and the allele frequency at the QTL p_A in the population (Falconer and Mackay 1996): $\alpha = a + d(1 - 2p_A)$, under the assumption that epistatic interactions are absent. Thus, if dominance is present, the allele substitution effect α depends on the allele frequency in the population. If no dominance exists, but interaction of the QTL with another locus (say locus 2 with alleles BB/bb), the allele substitution effect becomes: $\alpha = a + 2p_B i$, where p_B is the allele frequency at locus 2 in the population, and i is the interaction deviation between both loci. Therefore, also without dominance, the allele substitution effect is allele frequency-dependent if epistatic interactions exist with other loci.

These considerations hold for random mating populations, but they need to be modified for a set of inbred lines which are generally considered in hybrid plant breeding. If a population of DH lines is considered, no dominance deviation is observed, as the heterozygotes are missing and the allele substitution effect α reduces to a , as long as no epistasis is present. The maize populations considered in this thesis were phenotyped as testcrosses with a tester line, in which case the allele substitution effect depends also on the genotype of the tester (Melchinger et al. 1998b).

Thus, the allele substitution effect of a QTL measured on testcrosses, and thus of markers that are in LD with the QTL, depends on the tester genotype if dominance is present and also on the allele frequencies of the DH population when epistatic interactions exist.

Due to these considerations, QTL, and therefore marker effects, are likely to differ in distinct populations when they vary in their allele frequencies and/or when testcrosses are produced with different tester lines. This hampers prediction across populations and the accurate estimation of QTL and marker effects in datasets with a population structure. Studies in animal, plant, and human genetics report that the genotypic variance of quantitative traits mainly consists of additive genetic variance and only little of dominance variation and variance due to epistatic interaction. However, one cannot infer from such observations that non-additive effects are absent, because dominance and epistasis also contribute to additive genetic variance (Hill et al. 2008; Maki-Tanila and Hill 2014).

3.2.4 Heritability

In quantitative genetics, heritability in the broad sense is defined as the fraction of phenotypic variance which can be explained by all genetic factors (Falconer and Mackay 1996). While in animal breeding and human genetics, heritability is commonly estimated by using data from related individuals, in plant breeding heritability can be estimated using replicated samples. Typically in plant breeding single phenotypic values of lines are not considered but (adjusted) means of phenotypic observations from several replications are, partly also from different locations and years. Thus, the so-called operative heritability of the recorded phenotypic value of a line can be increased by increasing the number of replications.

In Lehermeier et al. (2014), the heritability on a line mean basis was estimated for each family separately. While within the flint pool heritability was higher than in the dent pool, due to more locations, differences in heritability within each heterotic pool of the Cornfed data mainly originated from differences in genotypic variance and genotype by environment interactions for the different crosses, as all crosses were tested in the same locations and with identical number of replications. Thus, adjusted means are expected to have equal precision for different crosses. These differences in heritability were mainly driven by differences in genotypic variances, which was also suggested by a high correlation between heritability and the genotypic variance within families within each Cornfed pool.

In Lehermeier et al. (2014), a strong association between family-specific heritability and predictive abilities obtained through prediction within families was observed. Here, heritability is equal for the estimation and the test sets, and higher prediction performance comes from a higher signal-to-noise ratio within both of these sets. Guo et al. (2014) also showed that prediction performance is affected by genomic heritability in the estimation set, as well as by heritability within the test set. To account for the signal-to-noise ratio within the test set, a commonly used approach is to correct predictive ability by dividing it by the square root of the heritability within the test set (Dekkers 2007; Legarra et al. 2008; Daetwyler et al. 2013). This procedure should approximate the correlation of the estimated with the true genetic values of the test set, often also called prediction accuracy. As heritability might not be estimated with the highest precision within (small) families, this approach was not followed in Lehermeier et al. (2014), in order to

avoid introducing additional bias into estimating prediction performance, and further, to avoid introducing autocorrelation in the comparison between heritability and accuracy. Greater heritability in the estimation set yields more accurately estimated effects. A higher heritability goes along with a higher Mendelian sampling term within families, which is the main source of genomic prediction capacity within bi-parental families. In Lehermeier et al. (2014), a positive correlation of predictive ability and heritability in the estimation set was also observed for prediction across families, which suggests the desirability of using estimation sets with high heritability and genotypic variation. Recently, this was also claimed by Guo et al. (2014); however, it was observed in Lehermeier et al. (2014) that genotypic variances in the progenies could not be predicted by marker-based similarities of the parental lines. This is in accordance with observations made by Hung et al. (2012), based on the US NAM population and numerous other studies investigating the relationship between genetic similarity and variance in various crops (Helms et al. 1997; Burkhamer et al. 1998; Melchinger et al. 1998a; Gumber et al. 1999). Thus, it is a difficult task to select appropriate parental lines for new crosses which would lead to a large genotypic variance within progeny for building a reliable estimation set for prediction.

3.3 Construction of the estimation set

A crucial question in genome-based prediction is how to construct an estimation set appropriately. Two different situations can be considered here. The first option has a set of genotyped selection candidates, but only limited resources to phenotype them. Here, the question is which lines should be phenotyped so that a best possible estimation set can be constructed to predict all lines accurately. This question is considered in Rincent et al. (2012) by using optimization algorithms based on the prediction error variance from the GBLUP method. Alternatively, the other situation is that one has a set of genotyped selection candidates and a (potentially large) population of lines that are phenotyped and genotyped. Here, the question is which lines should be chosen to build the estimation set for those specific selection candidates? The main difference in the two situations is that in the latter case the set of selection candidates is fixed, while it is dependent on the construction of the estimation set in the first case. The latter situation is considered and investigated in Riedelsheimer et al. (2013), Jacobson et al. (2014),

and Lehermeier et al. (2014) using multi-parental populations of maize.

3.3.1 Potential of multi-parental populations for genome-based prediction

Several studies have investigated the accuracy of genome-based prediction within bi-parental families of maize (Lorenzana and Bernardo 2009; Albrecht et al. 2011; Guo et al. 2012; Zhao et al. 2012a; Massman et al. 2013). Here, the estimation set for genome-based prediction is composed of lines from the same bi-parental family as the lines in the test set. Thus, the estimation and test sets have the same genetic background and are highly related, LD within bi-parental families is high and based purely on linkage, and population structure is absent. These are all factors which are advantageous for genome-based prediction, and so prediction accuracy within bi-parental families is supposed to be the maximum that can be achieved with a given sample size (Cossa et al. 2014). However, this approach does not allow for predicting newly generated crosses; instead, a part of the newly generated cross needs to be phenotyped first. Diversity panels have also been considered for genome-based prediction, because they sample maximum allelic diversity, e.g. in maize described by Rincent et al. (2012) and Riedelsheimer et al. (2012a). The advantages of both approaches for genome-based prediction are assumed to be combined when multi-parental populations are considered for forming estimation sets. First, the genome-based prediction approach mainly excels within families, as it can capture the Mendelian sampling term and is thus far better than pedigree-based prediction. Second, when several bi-parental families are combined to form the estimation set, the probability of QTL segregating that are relevant to the family in the test set is higher.

In Lehermeier et al. (2014) the prediction performance of multi-parental populations, compared to predictions within bi-parental populations, was investigated. The results showed that, for a given family, similar predictive abilities could be obtained by combining several half-sib families in the estimation set as compared to predicting within the bi-parental family, albeit only with an increase in the estimation set size. Predictive abilities over all traits and families were, on average, 0.50 when the testcross performance of progeny from a given family was predicted using all available half-sib lines in the estimation set (LOCO-CV). In contrast, within bi-parental families (R-CV), predictive abilities were, on average, 0.54. Prediction performance, with 50 full-sib lines forming the estimation set, could be increased to an average value of 0.63 by adding all additional

available half-sib families to the estimation set. This was also observed in Lehermeier et al. (2015) for the Cornfed dent data, where for all families except progeny derived from crosses with UH304 prediction performance was higher when all remaining lines from the same heterotic pool were used as the estimation set to predict the testcross values of half of the lines from one family compared to when only the remaining 50% full-sib lines were used as an estimation set.

One possible reason why prediction with multiple half-sib families worked well given a sufficient estimation set size is that the probability of changing linkage phases between the estimation and test sets is low, if families connected by a common parent are combined into one estimation set. Additionally, allele frequencies are expected to be 0.5 within families for loci where both parental lines have different alleles. Thus, allele substitution effects are expected to be similar for the different families. For practical applications it is appealing that those results show that new crosses can be predicted with high accuracy, if multiple half-sib families are used in the estimation set as long as they share one common parent and thus linkage phases between markers and QTL are largely consistent. For the Cornfed data, predictive abilities did not significantly change if four, six, or nine half-sib families constituted the estimation set of size $n = 200$.

Within a given heterotic pool, all lines from the Cornfed data share one common parent. Thus, no unrelated or unconnected families that come from the same heterotic pool are available, and we cannot investigate how accurate the prediction is when unconnected families from the same pool form the estimation set. Furthermore, no progeny from crosses sharing one of the founder lines was available. These situations were considered in Riedelsheimer et al. (2013) and Jacobson et al. (2014). Riedelsheimer et al. (2013) investigated prediction performance with five interconnected bi-parental maize families generated from crosses with four parental lines from the flint heterotic pool. As expected, they observed the highest prediction performance with full-sib lines and low prediction performance with unrelated lines. When half-sib lines were used as the estimation set, it was better when they were derived from crosses where both parental lines of the full-sib lines in the test set were represented, compared to the situation where they only shared one parent. Jacobson et al. (2014) compared the prediction performance of differently constructed estimation sets, using 970 bi-parental families of maize. Similarly, as observed in Riedelsheimer et al. (2013), genome-based prediction was most efficient when full-sib lines were used as the estimation set, and half-sib families yielded

better prediction performance compared to progeny from unrelated parents.

In Lehermeier et al. (2014) it was observed for the Cornfed dent panel that the prediction performance of LOCO-CV relative to R-CV differed for the investigated traits. For DMY, predictive abilities obtained with R-CV were around 0.4, and the same prediction performance could be reached with half-sib lines in LOCO-CV. For DMC, however, predictive abilities within families were much higher (0.48-0.77) and could generally not be reached with the full set of half-sib lines in LOCO-CV. The substantially decreased prediction performance of LOCO-CV, compared to predictions within families for DMC, might indicate that the importance of genetic factors contributing to variations in DMC within each bi-parental family is increased compared to DMY.

In Lehermeier et al. (2014, 2015), an outlier family within the Cornfed dent data was found with the one derived from UH304 for the trait DMY, but not for the other traits. This family's lines could not be predicted based on other dent families for DMY in the LOCO-CV scenario. As all dent families are equally connected through one common parent, no relevant differences in linkage phases were observed between the UH304 derived family and the other dent families. Thus, differences in linkage phases, as suggested in other outlying cases of low prediction performance (Riedelsheimer et al. 2013; Würschum et al. 2013), cannot be a reason for the poor prediction performance seen here. UH304 was the only Iodent founder line within the dent pool, and it is a more recently developed line compared to the other founder lines and may have experienced different selection pressure, especially for DMY. DH lines of this family had higher genetic similarities amongst each other compared to the other dent families, as the two parental lines UH304 and the central line F353 have both an Iodent background and showed high relatedness. However, despite the low genotypic variation observed from marker data, family UH304 showed an intermediate genotypic variance for all traits.

3.3.2 Prediction across populations

In animal breeding, prediction across populations was investigated using Holstein and Jersey breeds of dairy cattle, where prediction performance was close to zero when a Holstein estimation set was used to predict a Jersey test set, and vice versa (Hayes et al. 2009; Erbe et al. 2012). In maize, one of the most extreme settings is prediction across heterotic pools established in hybrid breeding. In Central Europe, an important heterotic pattern is dent and flint crosses. Generally, predictions across the two maize

pools of the Cornfed data failed, with predictive abilities close to zero when dent lines were predicted based on information from flint lines, and vice versa. This can be readily explained by the missing relatedness and inconsistent linkage phases between the pools. Technow et al. (2013) investigated predictive abilities across dent and flint lines for the line *per se* performance of the trait for Northern corn leaf blight resistance. They obtained slightly higher predictive abilities than were observed for DMY and DMC in the Cornfed data, with a predictive ability up to 0.25 for the prediction of line *per se* performance across flint and dent heterotic pools. Their findings are in line with the observation of slightly increased prediction performance across pools for flowering traits, compared to DMY and DMC in the Cornfed data. In breeding, line selection is mainly based on their testcross performance in relation to yield, and heterosis is maximized if opposite alleles are fixed in each pool. Thus, dent and flint lines are assumed to differ mainly at QTL affecting yield related traits. A companion study, performing QTL mapping with the same data (Giraud et al. 2014), detected less than 15% common QTL for the two pools across the five investigated traits and no overlapping QTL in dent and flint for DMY, which, according to Giraud et al. (2014), might indicate that complementary alleles have been fixed in both groups, due to selection. An additional obstacle is that not line *per se* but testcross performance was predicted. Although the amount of specific combining ability compared to general combining ability is relatively small (Bernardo 2010), especially when heterotic groups are distinct such as in maize (Reif et al. 2007; Albrecht et al. 2014), the allele substitution effect might differ when tester alleles are different and dominance is present (see chapter 3.2.3).

To date, predicting completely distinct genetic material has not been successful. Further research is therefore necessary to develop new methods adapted to such prediction scenarios. It has been assumed that Bayesian and other methods applying variable selection rely more on information from LD, whereas GBLUP mainly uses information from relatedness (Habier et al. 2007). Thus, it was assumed in the literature that sparse models can be used in a broader context and are superior to GBLUP for across-family prediction or prediction across generations (Meuwissen et al. 2009; Daetwyler et al. 2013). Here, variable selection methods did not improve prediction performance across pools in the Cornfed data, and results from BayesC π were similar compared to results from GBLUP.

3.4 Multivariate models for structured data

As indicated in the previous section and section 3.1.1 for prediction across breeds or populations, it was suggested in the literature that variable selection methods might be advantageous. However, these methods assume that marker effects are identical in different populations. This is a strict assumption, as the allele substitution effect can be allele frequency-dependent if dominance and/or epistasis exist, and additionally the marker-QTL LD might differ between populations (see section 3.2.1 and 3.2.3). For genome-based prediction in multi-parental maize populations, Schulz-Streeck et al. (2012a) addressed this issue by including interaction effects between markers and families into the RR-BLUP method. In doing so they observed very similar predictive abilities for the classical RR-BLUP method and the model including interaction effects. Although this approach yields family-specific marker effects, the covariance structure between sub-populations (or families) is restricted to being constant when more than two sub-populations are considered. Thus, Lehermeier et al. (2015) investigated a multivariate GBLUP method (MG-GBLUP) which estimates population-specific marker effects but at the same time uses all available data information for the estimation.

3.4.1 Prediction performance of the multivariate model

In Lehermeier et al. (2015), a multivariate genome-based prediction method (MG-GBLUP) for analyzing structured plant populations was introduced. This method was compared with a GBLUP method, ignoring sub-structure in the data and estimating common marker effects across all sub-populations (A-GBLUP) and analyses within each sub-population separately (W-GBLUP), therefore allowing for marker effects to differ across sub-populations but not to borrow information between sub-populations. Methods A-GBLUP and W-GBLUP can be considered as special cases pertinent to the multivariate method. With perfect genomic correlations between sub-populations, the MG-GBLUP method arrives at A-GBLUP, and with genomic correlations at zero MG-GBLUP arrives at W-GBLUP. Lehermeier et al. (2015) analyzed three differently structured plant datasets: a rice diversity panel with four sub-populations, the Cornfed dent data representing a large half-sib maize panel comprising ten bi-parental families, and wheat data from a breeding program which could be clustered into two sub-populations.

The rice data represent the most heterogeneous material analyzed in this study, with

very clear and distant sub-populations. This was also indicated by a high amount of variance explained by the first two eigenvalues of the marker-derived relationship matrix. Thus, estimated genomic correlations between sub-populations from method MG-GBLUP were close to zero. Consequently, the MG-GBLUP method approached the W-GBLUP method, and predicted genomic values from both methods were highly correlated. As expected, both methods also yielded very similar predictive abilities. With different allele frequencies, and thus different allele substitution effects, in the four different rice sub-populations, one would expect A-GBLUP to perform worse than the other two methods. Estimated sub-population-specific marker effects differed highly for the four sub-populations. Nevertheless, A-GBLUP, which forces marker effects to be identical over sub-populations, yielded prediction performance quite similar to MG-GBLUP and W-GBLUP with sub-population-specific marker effects. An explanation for the relatively good prediction performance of A-GBLUP might be that different alleles segregate in different rice sub-populations. Thus, a lot of SNPs are only polymorphic within a single sub-population, and SNP effect estimation is only marginally affected by the other sub-populations.

In contrast to the situation in the rice data, MG-GBLUP performed more akin to A-GBLUP than to W-GBLUP in the Cornfed dent data. Predicted genomic values from the multivariate model were highly correlated with the A-GBLUP method when ignoring sub-population structure. Estimated genomic correlations among maize families were rather weak but substantially higher than among the rice sub-populations, as all Cornfed dent lines share one common parent. With respect to predictive ability, on average the A-GBLUP method performed best and the multivariate method yielded similar, albeit slightly decreased, predictive abilities, while W-GBLUP yielded the lowest predictive abilities.

In the wheat data, predicted genomic values from the multivariate method (MG-GBLUP) were highly similar to the predicted genomic values from the within cluster GBLUP method (W-GBLUP), as observed in the rice data. Comparing the results obtained from the different environments, it was evident that, as expected, the lower the genomic correlation among clusters, the higher the correlation among predicted genomic values obtained with W-GBLUP and MG-GBLUP. Thus, the lower the genomic correlation between sub-populations, the less information is borrowed from the other sub-population within the multivariate method. Considering predictive abilities, the ranking of the methods

varied over the different environments. Averaging the results over all environments, the multivariate method outperformed the other two methods.

The scheme in Figure 5 visualizes the similarities between the A-GBLUP, MG-GBLUP, and W-GBLUP analyses in the rice, the Cornfed dent, and wheat datasets. Here, the edges of the triangles represent the Euclidean distance of the predicted genomic values from the different methods averaged over multiple traits or environments. In rice and wheat, the MG-GBLUP method approaches the W-GBLUP method. Furthermore, it is clearly evident that the difference between the three methods is relatively small in rice and wheat, which is due to less of a relationship across sub-populations in these datasets. Thus, A-GBLUP and MG-GBLUP borrow little information across sub-populations and are closer to the W-GBLUP analysis. In contrast, in the Cornfed dent data, the distance between A-GBLUP and W-GBLUP is much larger, as here more relatedness across sub-populations exists. Furthermore, the MG-GBLUP method converges to the A-GBLUP method in this case.

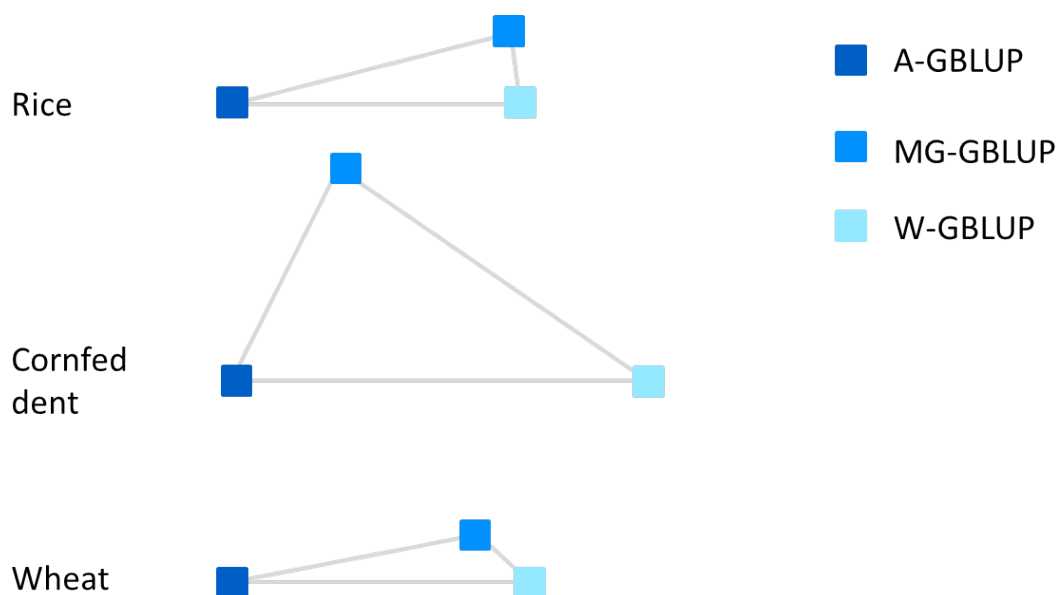


Figure 5: Visualization of the distance between the A-GBLUP, MG-GBLUP, and W-GBLUP analyses in the rice, the Cornfed dent, and the wheat data. The lengths of the triangle edges correspond to the Euclidean distances of the predicted genomic values taken from the different methods, averaged over different traits/environments.

Higher estimated genomic heritabilities were observed for MG-GBLUP compared to W-GBLUP for all three datasets. In addition, correlations of observed and predicted genomic values within the estimation set were higher for MG-GBLUP than for W-GBLUP

and A-GBLUP. This, together with partly reduced predictive abilities of MG-GBLUP compared to A-GBLUP or W-GBLUP, indicates some kind of overfitting behavior in relation to MG-GBLUP. Although MG-GBLUP seems to better fit the data than the other two methods, it partly showed reduced prediction performance compared to A-GBLUP and W-GBLUP. MG-GBLUP is more complex and has more parameters to estimate than the other two methods. Typically, in line with increasing model complexity, the estimation error in the training data decreases, but it might adapt too strongly to the training data and lead to poor predictions for independent test data (Hastie et al. 2009). It seems that although MG-GBLUP is the better fitting model and leads to less biased estimates, it does in some cases result in reduced prediction performance compared to the less complex models, due to greater prediction variance (bias-variance trade-off).

Both special cases of MG-GBLUP, namely the A-GBLUP and W-GBLUP, have advantages and disadvantages. A-GBLUP makes use of all available data to estimate marker effects; however, it imposes the homogeneity of marker effects and variance components on sub-populations. Conversely, W-GBLUP analyses allow for the heterogeneity of marker effects and variance components but do not allow for borrowing information between sub-populations. The MG-GBLUP method lies in between A-GBLUP and W-GBLUP. Borrowing information between sub-populations is most important when the within sub-population sample size is small. If that is the case, W-GBLUP analyses are inferior to A-GBLUP and MG-GBLUP, as especially seen in the Cornfed data. When sample size is small, it might also occur that a more sparse model with less parameters to estimate is advantageous. This might explain the good prediction performance of A-GBLUP, even when large differences between sub-populations are observed.

3.4.2 Multivariate models in animal breeding

Multivariate models have also been suggested for genome-based predictions combining different breeds in animal breeding. Olson et al. (2012) used a multivariate GBLUP method for the joint analysis of Holstein, Jersey, and Brown Swiss cattle, where they assigned different fixed values of genomic correlations between breeds. They found a slight increase in prediction performance when using the multivariate method with an assumed correlation of 0.3 between breeds, compared to a within-breed or an across-breed analysis. Makgahlela et al. (2013) investigated different Nordic Red breeds by applying a multivariate approach, but they could not estimate covariance between breeds

with REML estimation, due to convergence problems, and so they assigned a covariance of 0, which reduced the model to an analysis within breeds comparable to W-GBLUP in Lehermeier et al. (2015). Karoui et al. (2012) used a multivariate model for genome-based prediction in a combined dataset of three major French dairy cattle breeds, where the variance-covariance structure between breeds was estimated via a Gibbs sampler similar to that employed in this thesis. They observed mostly a slight increase in prediction performance using the multivariate model compared to analysis within breeds (corresponding to W-GBLUP in the study of Lehermeier et al. (2015)). Nonetheless, the results were very similar to assuming a very high correlation of 0.95 among breeds, which nearly corresponds to an overall GBLUP method similar to A-GBLUP. Recently, Zhou et al. (2014) investigated for Nordic Holstein and Nordic Red a multivariate GBLUP method using differently constructed realized relationship matrices. They observed only a small increase in prediction performance using the multivariate method over within-breed GBLUP, and they did not compare results to an overall GBLUP approach ignoring sub-population structure.

3.4.3 Genomic correlations between sub-populations

The multivariate method provides useful estimates of the genomic correlations between sub-populations. Estimated genomic correlations based on the variance-covariance matrix Σ_g from the multivariate GBLUP model (20) are equivalent to the marker correlations based on the variance-covariance matrix of the marker effects \mathbf{B} from model (13). Those correlations between sub-populations are trait-specific, as traits are assumed to be affected by different QTL in line with different contributions of epistasis and dominance. Thus, genomic correlations provide a measure of the similarity in relation to QTL effects for a specific trait between sub-populations and marker-QTL LD consistency within sub-populations (Karoui et al. 2012). For the Cornfed dent data, genomic correlations were higher between families derived from D06, D09, and UH250 than between the other families. The founder lines of these families originated from the breeding program initiated by the University of Hohenheim and were more related than the other founder lines. As those founder lines are related, their progenies show more likely segregation at identical loci and QTL. Thus, markers are assumed to have similar effects on a specific trait. For DMY, the family derived from UH304 showed a correlation close to zero with the other maize families. In Lehermeier et al. (2014), it was shown that the

yield records from the lines of this family could not be predicted based on lines from other families. The founder line UH304 is related to the common parent of all maize lines, namely F353. Thus, many QTL which explain genotypic variations for DMY within the other maize families might be monomorphic within family derived from UH304 (Giraud et al. 2014). In contrast, higher genomic correlations between this family and the other families were estimated for DMC, while higher prediction performance was also observed for this trait in Lehermeier et al. (2014). Generally, genomic correlations were lower for DMC than for DMY, indicating that functional polymorphisms affecting DMC are more family-specific than those affecting yield.

3.4.4 Extending the multivariate model

It is conceivable that the MG-GBLUP method might be useful for the analysis of testcross performance when different testers are used. This situation is present in the Synbreed CS2 data analyzed in Albrecht et al. (2014). This dataset contains two large genetic groups containing 582 (G1) and 260 (G3) lines, respectively. About one-third of the lines in group G1 and one half of G3 were crossed to tester T1 and to tester T3, respectively. Thus, in the data, four sub-populations can be considered: (i) G1_T1 with lines of group G1 crossed to tester T1, (ii) G1_T3 with lines of group G1 crossed to tester T3, (iii) G3_T1 with lines from group G3 crossed to tester T1, and (iv) G3_T3 with lines from group G3 crossed to tester T3.

G1_T1		0.22 ± 0.17	0.30 ± 0.16	0.28 ± 0.17
G3_T1	0.64 ± 0.17		0.27 ± 0.17	0.26 ± 0.18
G1_T3	0.67 ± 0.16	0.59 ± 0.17		0.51 ± 0.14
G3_T3	0.53 ± 0.17	0.49 ± 0.18	0.59 ± 0.14	
	G1_T1	G3_T1	G1_T3	G3_T3

Figure 6: Estimated genomic correlations (\pm posterior standard deviations) of traits GDY (upper diagonal) and GDC (lower diagonal) between the sub-populations from the Synbreed CS2 data.

Figure 6 shows genomic correlations between these four sub-populations for the traits GDY and GDC estimated by the MG-GBLUP method. Genomic correlations for lines belonging to the same genetic group, albeit crossed to different tester lines, were 0.30 and 0.26 for grain yield and 0.67 and 0.49 for grain dry matter content. Under the assumption that the lines tested with different testers represent a random sample from the same population, one would expect a correlation of one, if both testers had identical alleles at all QTL or if no dominance and epistasis existed (Melchinger et al. 1998b). Generally, genomic correlations were lower for GDY than for GDC. This is in line with the study of Melchinger et al. (1998b), where the QTL mapping results were more consistent across testers for grain moisture than for grain yield, which can be explained by significant contributions of dominance affecting grain yield.

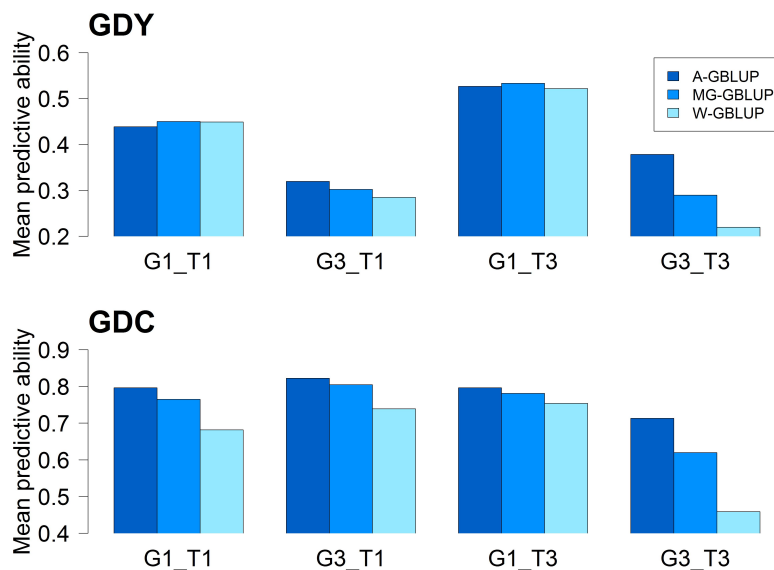


Figure 7: Mean predictive abilities for traits GDY and GDC of A-GBLUP, MG-GBLUP, and W-GBLUP evaluated within sub-populations from the Synbreed CS2 data.

Figure 7 shows the prediction performance of the MG-GBLUP method in comparison to the special cases A-GBLUP and W-GBLUP. For most sub-populations, the A-GBLUP method yielded the highest predictive abilities, as was also observed for the Cornfed dent data. Due to a smaller sample size of G3 compared to G1, predictive abilities in G3 were lower than in G1, while W-GBLUP, without borrowing information across groups, performed poorly in G3, especially for GDY. Exceptions were observed for G1 (G1_T1 and G1_T3) for the trait GDY, where the MG-GBLUP method performed best. For ge-

netic group G1 and trait GDY, it was already observed by Albrecht (2015) that predictive ability decreased when additional lines from the other groups constituted the estimation set, which corresponds to the comparison of the A-GBLUP and W-GBLUP analyses herein. Here, the MG-GBLUP was superior, as it accounted for differences between genetic groups and applied testers but at the same time did not ignore valuable information from the other groups.

A rather strict assumption in MG-GBLUP is that marker effects among sub-populations are homogeneously correlated along the genome. One option would be to extend MG-GBLUP to a marker-heterogeneous shrinkage method such as BayesA or BayesB (Meuwissen et al. 2001), where the correlation structure is specific to each marker. However, this substantially increases the number of parameters which need to be estimated, possibly leading to a highly under-determined model. Another option which would be more restrictive, but where less parameters would need to be estimated, would be to build clusters of markers and to assume different correlation structures for each one. This approach would be similar in principle to the one suggested by Akdemir and Jannink (2015), and it would allow for the notion that some markers are highly correlated among sub-populations, for example in regions where the marker-QTL LD is constant across sub-populations. Another group of markers might be very specific for each sub-population, for example due to LD with different QTL. However, as MG-GBLUP already seemed to suffer from overfitting, it is likely that extending the model to marker-heterogeneous correlation structures would not increase prediction performance, as long as sample size was not increased in parallel.

The methods discussed in Lehermeier et al. (2015) assume that individuals cluster in homogeneous, clearly separable groups; however, genetic variations are sometimes better described by a continuum in which some individuals belong to homogeneous sub-populations and others are admixed. Clearly, the methods discussed in Lehermeier et al. (2015) do not accommodate admixed groups adequately, and so further research is needed to develop methods that can deal simultaneously with admixed lines and distinct structures.

3.5 Conclusions

This thesis investigated different plant populations for genome-based prediction. The main conclusions from this work are:

- Marker-heterogeneous shrinkage methods were not superior to marker-homogeneous shrinkage methods such as Bayesian Ridge and GBLUP in experimental data with high LD and a presumably large number of small effect QTL.
- Bayesian Ridge and Bayesian Lasso showed a higher Bayesian learning ability and less sensitivity regarding hyperparameter specification than BayesA and BayesB.
- Through a number of theoretical considerations, it was shown that linkage phases between markers and QTL can change only slightly if families genetically connected by a common parent are combined. Thus, besides full-sib lines, half-sib lines represent an appropriate basis on which to construct an estimation set. This has the advantage that larger estimation sets can be constructed and newly generated crosses can also be predicted.
- Prediction across heterotic pools was not possible, due to a lack of relatedness. However, the results indicated that prediction performance is higher among lines originating from the same breeding program, even if no relatedness can be observed with marker data.
- It was shown that multivariate models can be used to infer the similarity of marker effects among sub-populations in a heterogeneous population. Thus, genetic heterogeneity between sub-populations can be characterized. Not only were genomic correlations between sub-populations dependent on the genetic distance of the sub-populations, but they also showed large differences among traits.
- Substructure in the data complicates genome-based prediction tasks. The choice of the best analysis method depends on the genetic heterogeneity in the data and on the sample size. While in highly heterogeneous populations the W-GBLUP and MG-GBLUP methods tend to perform better, A-GBLUP is superior in populations with more closely related sub-populations and with small sample sizes per sub-population.

4 Summary

Technical advances in the last few years have led to reduced costs in terms of time and money for genotyping in comparison to phenotyping plants. This has led to the rise and steady development of genomic selection in several species. The idea of genomic selection is to select lines based on their genome-based predicted breeding values instead of on their phenotypes. For this purpose, the effects of genetic markers need to be estimated in a first step based on a genotyped and phenotyped dataset (training population). A number of frequentist and Bayesian methods was suggested in the literature for genomic selection, which can cope with the much higher number of markers in comparison to the number of phenotypes. This thesis investigated, with simulated and empirical maize breeding data, different Bayesian methods (Bayesian Ridge, Bayesian Lasso, BayesA, and BayesB) with respect to their prediction performance as well as their sensitivity to prior parameter settings. Using the Hellinger distance between marginal prior and posterior density, a reduced Bayesian learning ability was quantified for BayesA and BayesB in comparison to Bayesian Ridge and Bayesian Lasso.

Besides the choice of an appropriate method for genome-based prediction, the construction of a suitable estimation set for a specific population of selection candidates is crucial. The optimum design of an estimation set was investigated using a multi-parental population of maize lines. This population comprised 1,652 genotyped and phenotyped double haploid (DH) maize lines arranged in two large half-sib families. These two half-sib families represent two main heterotic germplasm pools in Europe: dent and flint. Ten bi-parental dent families were generated from crosses of ten diverse dent maize lines with one common dent parental line (F353), and eleven bi-parental flint families were generated from crosses of eleven diverse flint maize lines with one common flint parental line (UH007). Using different estimation and test set compositions, the efficiency of genome-based prediction was investigated. It was observed that an estimation set consisting of several half-sib families yielded similar or even higher predictive abilities than an estimation set of 50 full-sib lines. Largely consistent marker linkage phases in half-sib families, which were investigated theoretically and empirically, are one reason why half-sib families are a good basis for the construction of an estimation set. Prediction across heterotic pools was not possible in most cases, and it was shown that including additional unrelated lines in the estimation set had no impact on prediction

performance.

Plant breeding populations often exhibit enormous substructures. Generally, these are corrected for mean differences in different sub-populations when structured data are analyzed for genome-based prediction. Very rarely differences in marker effects which likely occur due to, for example, differences in allele frequencies and linkage phases are accounted for. Three options to analyze structured data in the context of genome-based prediction were investigated. The first method ignored the population structure in the data and estimated common marker effects for all sub-populations (A-GBLUP). The second method estimated marker effects within each sub-population separately (W-GBLUP). Whereas W-GBLUP accounts fully for heterogeneity in the data in contrast to A-GBLUP, it has the disadvantage that sample size is reduced enormously, which in turn hampers the precise estimation of marker effects. Thus, a multivariate method (MG-GBLUP) was proposed which estimates population-specific marker effects, though marker effects are allowed to be correlated among sub-populations. Differently structured datasets were used to investigate the three methods. Depending on the diversity of the data, A-GBLUP or W-GBLUP yielded better prediction performance. In most cases, the multivariate method, which is a generalization of A-GBLUP and W-GBLUP, converged towards the better performing method and showed similar or better prediction performance. The genomic correlations which were estimated using the multivariate method provided information about trait-specific heterogeneity between sub-populations.

5 Zusammenfassung

Biotechnologische Fortschritte der letzten Jahre führten dazu, dass in der Pflanzenzüchtung die Genotypisierung von Linien deutlich günstiger und schneller durchzuführen ist als die Phänotypisierung. Dies führte zu einem Aufschwung und zu einer stetigen Weiterentwicklung der genomischen Selektion in verschiedenen Spezies. Die Idee bei der genomischen Selektion ist Linien nicht anhand ihrer Phänotypen sondern basierend auf ihren vorhergesagten genomischen Zuchtwerten zu selektieren. Dazu werden die Effekte der genetischen Marker mittels eines genotypisierten und phänotypisierten Datensatzes (Referenzpopulation) geschätzt. Eine Vielzahl frequentistischer und Bayesianischer Methoden wurde in der Literatur für die genomische Selektion vorgeschlagen, welche in der Lage sind die in der Regel deutlich höhere Anzahl an Markern im Vergleich zu den beobachteten Phänotypen zu bewältigen. In dieser Arbeit wurden mittels simulierter Daten und einer Mais Züchtungspopulation verschiedene Bayesianische Methoden (Bayesian Ridge, Bayesian Lasso, BayesA und BayesB) bezüglich ihrer Vorhersagegenauigkeit und Sensitivität gegenüber der Priori Parameterwahl untersucht. Durch die Berechnung der Hellinger Distanz zwischen marginaler Priori- und Posteriori-Dichte wurde eine reduzierte Fähigkeit zu Bayesianischem Lernen bei den Methoden BayesA und BayesB im Vergleich zu dem Bayesianischen Ridge und Bayesianischen Lasso quantifiziert.

Neben der Wahl einer geeigneten statistischen Methode für die genomische Vorhersage ist vor allem auch entscheidend, wie eine geeignete Referenzpopulation für bestimmte Populationen an Selektionskandidaten zusammengestellt werden soll. Diese Frage wurde mittels einer multiparentalen Maispopulation untersucht. Diese Population setzt sich zusammen aus 1652 genotypisierten und phänotypisierten Doppelhaploiden (DH) Maislinien welche in zwei großen Halbgeschwisterfamilien angeordnet sind. Diese beiden Halbgeschwisterfamilien repräsentieren dabei zwei wichtige heterotische Gruppen in Europa: Dent und Flint. Aus Kreuzungen von zehn diversen Dent-Maislinien mit einer gemeinsamen Dent-Elternlinie (F353) wurden zehn biparentale Dent-Familien generiert und aus Kreuzungen von elf diversen Flint-Maislinien mit einer gemeinsamen Flint-Elternlinie (UH007) wurden elf biparentale Flint-Familien generiert. Die Effizienz der genomischen Vorhersage mit verschiedenen Zusammensetzungen von Schätz- und Testsets wurde anhand dieses Datensatzes untersucht. Hierbei wurde beobachtet, dass

durch die Kombination mehrerer Halbgeschwisterfamilien im Schätzset eine ähnliche oder sogar höhere Vorhersagegenauigkeit erreicht werden kann als wenn 50 Vollgeschwisterlinien verwendet werden. Konsistente Marker Kopplungsphasen bei Halbgeschwisterfamilien, welche theoretisch und empirisch untersucht wurden, sind hierbei ein Aspekt weshalb mehrere Halbgeschwisterfamilien eine geeignete Referenzpopulation bilden. Die Vorhersage über heterotische Gruppen hinweg war in den meisten Fällen nicht möglich und es wurde gezeigt, dass das Hinzufügen unverwandter Linien zu dem Schätzset keinen Einfluss auf die Vorhersagegenauigkeit hat.

Pflanzenzüchtungspopulationen weisen häufig eine starke genetische Struktur auf. Bei der Analyse von Daten mit Substrukturen wurde bisher häufig nur für Mittelwertsunterschiede in den verschiedenen Subpopulationen korrigiert. Selten wurde dabei berücksichtigt, dass sich auch die Markereffekte in den verschiedenen Subpopulationen mit hoher Wahrscheinlichkeit unterscheiden, etwa aufgrund unterschiedlicher Allelfrequenzen oder Kopplungsphasen. Drei verschiedene Optionen wurden für die genomische Vorhersage strukturierter Daten untersucht. Einmal wurde die Populationsstruktur vernachlässigt und es wurden für alle Subpopulationen gleiche Markereffekte geschätzt (A-GBLUP). Desweiteren wurde eine stratifizierte Analyse durchgeführt in der Markereffekte innerhalb jeder Subpopulation einzeln geschätzt wurden (W-GBLUP). Während W-GBLUP anders als A-GBLUP die Heterogenität in den Daten berücksichtigt, hat W-GBLUP den Nachteil, dass die Stichprobengröße zur präzisen Schätzung der Markereffekte stark reduziert wird. Aus diesem Grund wurde eine multivariate Analyse (MG-GBLUP) vorgeschlagen, mit der populationsspezifische Markereffekte geschätzt werden können und Korrelationsstruktur zwischen den Markereffekten der Subpopulationen zugelassen wird. Datensätze mit sehr unterschiedlicher Populationsstruktur wurden verwendet, um die drei Analysemethoden zu vergleichen. Es zeigte sich, dass abhängig von der Diversität der Datensätze A-GBLUP oder W-GBLUP die bessere Vorhersagegenauigkeit lieferte. Das multivariate Modell, welches eine Verallgemeinerung von A-GBLUP und W-GBLUP darstellt, näherte sich in den meisten Fällen der besser abschneidenden Methode an und zeigte eine ähnliche oder höhere Vorhersagegenauigkeit. Die mit der multivariaten Analyse geschätzten genomischen Korrelationen zwischen den Subpopulationen gaben Aufschluss über die merkmals-spezifische Heterogenität zwischen den Subpopulationen.

6 References

- Akdemir D, and Jannink J L (2015), Locally epistatic genomic relationship matrices for genomic association and prediction. *Genetics* 199(3):857–871
- Albrecht T (2015), Genome-based prediction of testcross performance in maize (*Zea mays* L.). Dissertation, Technische Universität München, Freising
- Albrecht T, Auinger H J, Wimmer V, Ogutu J O, Knaak C, Ouzunova M, Piepho H P and Schön C C (2014), Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theor. Appl. Genet.* 127(6):1375–1386
- Albrecht T, Wimmer V, Auinger H J, Erbe M, Knaak C, Ouzunova M, Simianer H, and Schön C C (2011), Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.* 123(2):339–350
- Andreescu C, Avendano S, Brown S R, Hassen A, Lamont S J, and Dekkers J C M (2007), Linkage disequilibrium in related breeding lines of chickens. *Genetics* 177(4):2161–2169
- Asoro F G, Newell M A, Beavis W D, Scott M P and Jannink J L (2011), Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* 4(2):132–144
- Astle W, and Balding D J (2009), Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* 24(4):451–471
- Bao Y, Vuong T, Meinhardt C, Tiffin P, Denny R, Chen S, Nguyen H T, Orf J H, and Young N D (2014), Potential of association mapping and genomic selection to explore PI 88788 derived soybean cyst nematode resistance. *Plant Genome* 7(3)
- Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincint R, Schipprack W, Altmann T, Flament P, Melchinger A E, Menz M, Moreno-González J, Ouzunova M, Revilla P, Charcosset A, Martin O C, and Schön C C (2013), Intraspecific variation of recombination rate in maize. *Genome Biol.* 14(9):R103

-
- Bentley A R, Scutari M, Gosman N, Faure S, Bedford F, Howell P, Cockram J, Rose G A, Barber T, Irigoyen J, Horsnell R, Pumfrey C, Winnie E, Schacht J, Beauchêne K, Praud S, Greenland A, Balding D, and Mackay I J (2014), Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. *Theor. Appl. Genet.* 127(12):2619–2633
- Bernal-Vasquez A M, Möhring J, Schmidt M, Schönleben M, Schön C C, and Piepho H P (2014), The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye. *BMC Genomics* 15(1):646
- Bernardo J M, and Smith A F M (2009), *Bayesian theory*. Wiley, Chichester
- Bernardo R (2010), *Breeding for quantitative traits in plants*. Stemma Press, Woodbury, MN, 2nd edition
- Bernardo R, and Yu J (2007), Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47(3):1082–1090
- Biscarini F, Stevanato P, Broccanello C, Stella A, and Saccomani M (2014), Genome-enabled predictions for binomial traits in sugar beet populations. *BMC Genetics* 15(1):87
- Browning B L, and Browning S R (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(2):210–223
- Burgueño J, de los Campos G, Weigel K, and Crossa J (2012), Genomic prediction of breeding values when modeling genotype \times environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52(2):707–719
- Burkhamer R L, Lanning S P, Martens R J, Martin J M, and Talbert L E (1998), Predicting progeny variance from parental divergence in hard red spring wheat. *Crop Sci.* 38(1):243–248
- Butler D G, Cullis B R, Gilmour A B, and Gogel B J (2009), *ASReml-R reference manual*. Department of Primary Industries and Fisheries, Brisbane, Australia

-
- de los Campos G, Hickey J M, Pong-Wong R, Daetwyler H D, and Calus M P L (2013a), Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193(2):327–345
- de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, and Cotes J M (2009), Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182(1):375–385
- de los Campos G, Vazquez A I, Fernando R, Klimentidis Y C, and Sorensen D (2013b), Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9(7):e1003608
- Charcosset A, and Essioux L (1994), The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. *Theor. Appl. Genet.* 89(2-3):336–343
- Charmet G, Storlie E, Oury F X, Laurent V, Beghin D, Chevarin L, Lapiere A, Perretant M R, Rolland B, Heumez E, Duchalais L, Goudemand E, Bordes J, and Robert O (2014), Genome-wide prediction of three important traits in bread wheat. *Mol. Breeding* 34(4):1843–1852
- Clark S A, Hickey J M, Daetwyler H D, and van der Werf J H (2012), The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44(4):1–9
- Combs E, and Bernardo R (2013), Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* 6(1):1–7
- Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A, Pomiès V, Riou V, Suryana E, and Bouvet J M (2015), Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.* 128(3):397–410
- Crossa J, Beyene Y, Kassa S, Perez P, Hickey J M, Chen C, de los Campos G, Burgueno J, Windhausen V S, Buckler E, Jannink J L, Lopez Cruz M A, and Babu R (2013),

-
- Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3: Genes, Genomes, Genetics* 3(11):1903–1926
- Crossa J, Campos G d l, Perez P, Gianola D, Burgueno J, Araus J L, Makumbi D, Singh R P, Dreisigacker S, Yan J, Arief V, Banziger M, and Braun H J (2010), Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186(2):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
- Cuevas J, Perez-Elizalde S, Soberanis V, Perez-Rodriguez P, Gianola D, and Crossa J (2014), Bayesian genomic-enabled prediction as an inverse problem. *G3: Genes, Genomes, Genetics* 4(10):1991–2001
- Cullis B R, Smith A B, and Coombes N E (2006), On the design of early generation variety trials with correlated data. *J. Agric. Biol. Envir. S.* 11(4):381–393
- Daetwyler H D, Bansal U K, Bariana H S, Hayden M J, and Hayes B J (2014), Genomic prediction for rust resistance in diverse wheat landraces. *Theor. Appl. Genet.* 127(8):1795–1803
- Daetwyler H D, Calus M P L, Pong-Wong R, de los Campos G, and Hickey J M (2013), Genomic prediction in animals and plants: simulation of data, validation, reporting and benchmarking. *Genetics* 193(2):347–365
- Daetwyler H D, Pong-Wong R, Villanueva B, and Woolliams J A (2010), The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185(3):1021–1031
- Daetwyler H D, Villanueva B, and Woolliams J A (2008), Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3(10):e3395
- Dawson J C, Endelman J B, Heslot N, Crossa J, Poland J, Dreisigacker S, Manès Y, Sorrells M E, and Jannink J L (2013), The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Res.* 154:12–22

-
- Dekkers J C M (2007), Marker-assisted selection for commercial crossbred performance. *J. Anim. Sci.* 85(9):2104–2114
- El-Kassaby Y A, Klapste J, and Guy R D (2012), Breeding without breeding: selection using the genomic best linear unbiased predictor method (GBLUP). *New Forest* 43(5-6):631–637
- Erbe M, Gredler B, Seefried F R, Bapst B, and Simianer H (2013), A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS ONE* 8(12):e81046
- 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(7):4114–4129
- Falconer D S, and Mackay T F C (1996), *Introduction to quantitative genetics*. Longman, Essex, England
- Fodor A, Segura V, Denis M, Neuenschwander S, Fournier-Level A, Chatelet P, Homa F A A, Lacombe T, This P, and Le Cunff L (2014), Genome-wide prediction methods in highly diverse and heterozygous species: proof-of-concept through simulation in grapevine. *PLoS ONE* 9(11):e110436
- Fritsche-Neto R, DoVale J C, Lanes d C M d, Resende M D V d, and Miranda G V (2012), Genome-wide selection for tropical maize root traits under conditions of nitrogen and phosphorus stress. *Acta Sci.-Agron.* 34(4)
- Ganal M W, Durstewitz G, Polley A, Bérard A, Buckler E S, Charcosset A, Clarke J D, Graner E M, Hansen M, Joets J, Le Paslier M C, McMullen M D, Montalent P, Rose M, Schön C C, Sun Q, Walter H, Martin O C, 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(12):e28334
- Gelman A, Carlin J B, Stern H S, and Rubin D B (2004), *Bayesian data analysis*. Chapman & Hall/CRC

-
- Gianola D (2013), Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics* 194(3):573–596
- Gianola D, de los Campos G, Hill W G, Manfredi E, and Fernando R (2009), Additive genetic variability and the Bayesian alphabet. *Genetics* 183(1):347–363
- Gianola D, Weigel K A, Krämer N, Stella A, and Schön C C (2014), Enhancing genome-enabled prediction by bagging genomic BLUP. *PLoS ONE* 9(4):e91693
- Giraud H, Lehermeier C, Bauer E, Falque M, Segura V, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Schipprack W, Flament P, Melchinger A E, Menz M, Moreno-Gonzalez J, Ouzunova M, Charcosset A, Schön C C, and Moreau L (2014), Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. *Genetics* 198(4):1717–1734
- Goddard M (2009), Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136(2):245–257
- Goddard M, and Hayes B (2007), Genomic selection. *J. Anim. Breed. Genet.* 124(6):323–330
- Goddard M, Hayes B, and Meuwissen T (2011), Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* 128(6):409–421
- Gonzalez-Recio O, Rosa G, and Gianola D (2014), Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Sci.* 166:217–231
- González-Camacho J M, de los Campos G, Pérez P, Gianola D, Cairns J E, Mahuku G, Babu R, and Crossa J (2012), Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125(4):759–771
- Gouy M, Rousselle Y, Bastianelli D, Lecomte P, Bonnal L, Roques D, Efile J C, Rocher S, Daugrois J, Toubi L, Nabeneza S, Hervouet C, Telismart H, Denis M, Thong-Chane A, Glaszmann J C, Hoarau J Y, Nibouche S, and Costet L (2013), Experimental assessment of the accuracy of genomic selection in sugarcane. *Theor. Appl. Genet.* 126(10):2575–2586

-
- Gowda M, Zhao Y, Würschum T, Longin C F, Miedaner T, Ebmeyer E, Schachschneider R, Kazman E, Schacht J, Martinant J P, Mette M F, and Reif J C (2013), Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. *Heredity* 112:552–561
- Grubbs F E (1950), Sample criteria for testing outlying observations. *Ann. Math. Stat.* 21(1):27–58
- Gumber R K, Schill B, Link W, v Kittlitz E, and Melchinger A E (1999), Mean, genetic variance, and usefulness of selfing progenies from intra- and inter-pool crosses in faba beans (*Vicia faba* L.) and their prediction from parental parameters. *Theor. Appl. Genet.* 98(3-4):569–580
- Guo T, Li H, Yan J, Tang J, Li J, Zhang Z, Zhang L, and Wang J (2013a), Performance prediction of F1 hybrids between recombinant inbred lines derived from two elite maize inbred lines. *Theor. Appl. Genet.* 126(1):189–201
- Guo Z, Tucker D M, Basten C J, 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(3):749–762
- Guo Z, Tucker D M, Lu J, Kishore V, and Gay G (2012), Evaluation of genome-wide selection efficiency in maize nested association mapping populations. *Theor. Appl. Genet.* 124(2):261–275
- Guo Z, Tucker D M, Wang D, Basten C J, Ersoz E, Briggs W H, Lu J, Li M, and Gay G (2013b), Accuracy of across-environment genome-wide prediction in maize nested association mapping populations. *G3: Genes, Genomes, Genet.* 3(2):263–272
- Habier D, Fernando R L, and Dekkers J C M (2007), The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177(4):2389–2397
- Habier D, Fernando R L, Kizilkaya K, and Garrick D J (2011), Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12(186)
- Habier D, Tetens J, Seefried F R, 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

-
- Hastie T, Tibshirani R, and Friedman J (2009), *The elements of statistical learning: data mining, inference, and prediction*. Springer, Stanford, California, USA
- Hayes B J, Bowman P J, Chamberlain A C, Verbyla K, and Goddard M E (2009), Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41(1):51
- Heffner E L, Jannink J L, Iwata H, Souza E, and Sorrells M E (2011a), Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51(6):2597–2606
- Heffner E L, Jannink J L, and Sorrells M E (2011b), Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4(1):65–75
- Helms T, Vallad G, McClean P, and Orf J (1997), Genetic variance, coefficient of parentage, and genetic distance of six soybean populations. *Theor. Appl. Genet.* 94(1):20–26
- Henderson C R (1963), *Selection index and expected genetic advance*. Natl. Acad. Sci., Washington, DC
- Henderson C R (1975), Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31(2):423
- Heslot N, Akdemir D, Sorrells M E, and Jannink J L (2014), Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127(2):463–480
- Heslot N, Jannink J L, and Sorrells M E (2013), Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. *Crop Sci.* 53(3):921–933
- Heslot N, Yang H P, Sorrells M E, and Jannink J L (2012), Genomic selection in plant breeding: A comparison of models. *Crop Sci.* 52(1):146–160
- Hill W G, Goddard M E, and Visscher P M (2008), Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* 4(2):e1000008

-
- 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(8):1639–1645
- Hofheinz N, and Frisch M (2014), Heteroscedastic Ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3: Genes, Genomes, Genet.* 4(3):539–546
- Holland J B, Nyquist W E, and Cervantes-Martínez C T (2003), Estimating and interpreting heritability for plant breeding: An update. *Plant Breed. Rev.* 22:9–112
- Hung H Y, Browne C, Guill K, Coles N, Eller M, Garcia A, Lepak N, Melia-Hancock S, Oropeza-Rosas M, Salvo S, Upadyayula N, Buckler E S, Flint-Garcia S, McMullen M D, Rocheford T R, and Holland J B (2012), The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity* 108(5):490–499
- Isidro J, Jannink J L, Akdemir D, Poland J, Heslot N, and Sorrells M E (2014), Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 128(1):145–158
- Jacobson A, Lian L, Zhong S, and Bernardo R (2014), General combining ability model for genomewide selection in a biparental cross. *Crop Sci.* 54(3):895
- Jannink J L, Lorenz A J, and Iwata H (2010), Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics and Proteomics* 9(2):166–177
- Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, and de los Campos G (2014a), A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127(3):595–607
- Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, and Lorenz A (2014b), Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15(1):740
- Jiang Y, Zhao Y, Rodemann B, Plieske J, Kollers S, Korzun V, Ebmeyer E, Argillier O, Hinze M, Ling J, Röder M S, Ganai M W, Mette M F, and Reif J C (2014), Potential and limits

-
- to unravel the genetic architecture and predict the variation of Fusarium head blight resistance in European winter wheat (*Triticum aestivum* L.). *Heredity* 114:318–326
- Karoui S, Carabaño M J, Díaz C, and Legarra A (2012), Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genet. Sel. Evol.* 44(1):39
- Knürr T, Läärä E, and Sillanpää M J (2013), Impact of prior specifications in a shrinkage-inducing Bayesian model for quantitative trait mapping and genomic prediction. *Genet. Sel. Evol.* 45(1):24
- Kumar S, Chagné D, Bink M C A M, Volz R K, Whitworth C, and Carlisle C (2012), Genomic selection for fruit quality traits in apple (*Malus × domestica* Borkh.). *PLoS ONE* 7(5):e36674
- Lado B, Matus I, Rodriguez A, Inostroza L, Poland J, Belzile F, del Pozo A, Quincke M, Castro M, and von Zitzewitz J (2013), Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes, Genomes, Genet.* 3(12):2105–2114
- Le Cam L (1986), *Asymptotic methods in statistical decision theory*. Springer Series in Statistics, Springer, New York
- Legarra A, Robert-Granie C, Manfredi E, and Elsen J M (2008), Performance of genomic selection in mice. *Genetics* 180(1):611–618
- Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, Flament P, Melchinger A E, 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 C C (2014), Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. *Genetics* 198(1):3–16
- Lehermeier C, Wimmer V, Albrecht T, Auinger H J, Gianola D, Schmid V J, and Schön C C (2013), Sensitivity to prior specification in Bayesian genome-based prediction models. *Stat. Appl. Genet. Mol. Biol.* 12(3):375–391
- Li Z, and Sillanpää M J (2012), Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor. Appl. Genet.* 125(3):419–435

-
- Lipka A E, Lu F, Cherney J H, Buckler E S, Casler M D, and Costich D E (2014), Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. PLoS ONE 9(11):e112227
- Lorenz A J, Smith K, and Jannink J L (2012), Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52(4):1609–1621
- Lorenzana R E, and Bernardo R (2009), Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120(1):151–161
- Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch H G, Okechukwu R, Dixon A G, Kulakow P, and Jannink J L (2013), Relatedness and genotype \times environment interaction affect prediction accuracies in genomic selection: A study in Cassava. Crop Sci. 53(4):1312–1325
- Makgahlela M, Mäntysaari E, Strandén I, Koivula M, Nielsen U, Sillanpää M, and Juga J (2013), Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J. Anim. Breed. Genet. 130(1):10–19
- Maki-Tanila A, and Hill W G (2014), Influence of gene interaction on complex trait variation with multilocus models. Genetics 198(1):355–367
- Massman J M, Gordillo A, Lorenzana R E, and Bernardo R (2013), Genomewide predictions from maize single-cross data. Theor. Appl. Genet. 126(1):13–22
- Massman J M, Jung H J G, and Bernardo R (2012), Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. 53(1):58–66
- Matukumalli L K, Lawley C T, Schnabel R D, Taylor J F, Allan M F, Heaton M P, O'Connell J, Moore S S, Smith T P L, Sonstegard T S, and Van Tassell C P (2009), Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4(4):e5350
- McMullen M D, Kresovich S, Villeda H S, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon

-
- D, Lepak N, Mitchell S E, Peterson B, Pressoir G, Romero S, Rosas M O, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz J C, Goodman M, Ware D, Holland J B, and Buckler E S (2009), Genetic properties of the maize nested association mapping population. *Science* 325(5941):737–740
- Melchinger A E, Gumber R K, Leipert R B, Vuylsteke M, and Kuiper M (1998a), Prediction of testcross means and variances among F3 progenies of F1 crosses from testcross means and genetic distances of their parents in maize. *Theor. Appl. Genet.* 96(3-4):503–512
- Melchinger A E, Utz H F, and Schön C C (1998b), 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(1):383–403
- Meuwissen T, and Goddard M (2010), Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185(2):623–631
- Meuwissen T H (2009), Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* 41(35)
- Meuwissen T H, Solberg T R, Shepherd R, and Woolliams J A (2009), A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.* 41:2
- Meuwissen T H E, Hayes B J, and Goddard M E (2001), Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4):1819–1829
- Morota G, Koyama M, M Rosa G J, Weigel K A, and Gianola D (2013), Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. *Genet. Sel. Evol.* 45(1):17
- Munoz P R, Resende M F R, Huber D A, Quesada T, Resende M D V, Neale D B, Wegrzyn J L, Kirst M, and Peter G F (2014), Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. *Crop Sci.* 54(3):1115–1123

-
- Nakaya A, and Isobe S N (2012), Will genomic selection be a practical method for plant breeding? *Ann. Botany* 110(6):1303–1316
- Nei M, and Li W H (1973), Linkage disequilibrium in subdivided populations. *Genetics* 75(1):213–219
- Nejati-Javaremi A, Smith C, and Gibson J P (1997), Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* 75(7):1738–1745
- de Oliveira E J, de Resende M D V, da Silva Santos V, Ferreira C F, Oliveira G A F, da Silva M S, de Oliveira L A, and Aguilar-Vildoso C I (2012), Genome-wide selection in cassava. *Euphytica* 187(2):263–276
- Olson K, VanRaden P, and Tooker M (2012), Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci.* 95(9):5378–5383
- Onogi A, Ideta O, Inoshita Y, Ebana K, Yoshioka T, Yamasaki M, and Iwata H (2015), Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). *Theoretical and Applied Genetics* 128(1):41–53
- Ornella L, Pérez P, Tapia E, González-Camacho J M, Burgueño J, Zhang X, Singh S, Vicente F S, Bonnett D, Dreisigacker S, Singh R, Long N, and Crossa J (2014), Genomic-enabled prediction with classification algorithms. *Heredity* 112:616–626
- Ornella L, Singh S, Perez P, Burgueño J, Singh R, Tapia E, Bhavani S, Dreisigacker S, Braun H J, Mathews K, and Crossa J (2012), Genomic prediction of genetic values for resistance to wheat rusts. *Plant Genome* 5(3):136–148
- Ould Estagvirou S, Ogutu J O, Schulz-Streeck T, Knaak C, Ouzunova M, Gordillo A, and Piepho H P (2013), Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC Genomics* 14(1):860
- Ould Estagvirou S B, Ogutu J O, and Piepho H P (2014), Influence of outliers on accuracy estimation in genomic prediction in plant breeding. *G3: Genes, Genomes, Genet.* 4(12):2317–2328
- Owens B F, Lipka A E, Magallanes-Lundback M, Tiede T, Diepenbrock C H, Kandianis C B, Kim E, Cepela J, Mateos-Hernandez M, Buell C R, Buckler E S, DellaPenna D, Gore

-
- M A, and Rocheford T R (2014), A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. *Genetics* 198(4):1699–1716
- Park T, and Casella G (2008), The Bayesian Lasso. *J. Am. Stat. Assoc.* 103(482):681–686
- Peiffer J A, Flint-Garcia S A, De Leon N, McMullen M D, Kaeppler S M, and Buckler E S (2013), The genetic architecture of maize stalk strength. *PLoS ONE* 8(6):e67066
- Peiffer J A, Romay M C, Gore M A, Flint-Garcia S A, Zhang Z, Millard M J, Gardner C A C, McMullen M D, Holland J B, Bradbury P J, and Buckler E S (2014), The genetic architecture of maize height. *Genetics* 196(4):1337–1356
- Piepho H P (2009), Ridge regression and extensions for genomewide selection in maize. *Crop Sci.* 49(4):1165–1176
- Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, and Jannink J L (2012), Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5(3):103–113
- Pérez P, and de los Campos G (2014), Genome-wide regression & prediction with the BGLR statistical package. *Genetics* 198(2):483–495
- Pérez P, de los Campos G, Crossa J, and Gianola D (2010), Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *Plant Genome* 3:106
- Pérez-Rodríguez P, Gianola D, Gonzalez-Camacho J M, Crossa J, Manes Y, and Dreisigacker S (2013), Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes, Genomes, Genet.* 2(12):1595–1605
- Pszczola M, Strabel T, Mulder H, and Calus M (2012), Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95(1):389–400
- Reif J C, Gumpert F M, Fischer S, and Melchinger A E (2007), Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics* 176(3):1931–1934

-
- Reif J C, Zhao Y, Würschum T, Gowda M, and Hahn V (2013), Genomic prediction of sunflower hybrid performance. *Plant Breed.* 132(1):107–114
- Resende M D V, Resende M F R, Sansaloni C P, Petroli C D, Missiaggia A A, Aguiar A M, Abad J M, Takahashi E K, Rosado A M, Faria D A, Pappas G J, Kilian A, and Grattapaglia D (2012a), Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol.* 194(1):116–128
- Resende M F R, Munoz P, Resende M D V, Garrick D J, Fernando R L, Davis J M, Jokela E J, Martin T A, Peter G F, and Kirst M (2012b), Accuracy of genomic selection methods in a standard data set of Loblolly pine (*Pinus taeda* L.). *Genetics* 190(4):1503–1510
- Resende M F R, Muñoz P, Acosta J J, Peter G F, Davis J M, Grattapaglia D, Resende M D V, and Kirst M (2012c), Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol.* 193(3):617–624
- Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, and Melchinger A E (2012a), Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* 44(2):217–220
- Riedelsheimer C, Endelman J B, Stange M, Sorrells M E, Jannink J L, and Melchinger A E (2013), Genomic predictability of interconnected bi-parental maize populations. *Genetics* 194(2):493–503
- Riedelsheimer C, and Melchinger A E (2013), Optimizing the allocation of resources for genomic selection in one breeding cycle. *Theor. Appl. Genet.* 126(11):2835–2848
- Riedelsheimer C, Technow F, and Melchinger A E (2012b), Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics* 13(1):452
- Rincent R, Laloe D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodriguez V M, Moreno-Gonzalez J, Melchinger A, Bauer E, Schön C C, 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(2):715–728
- de Roos A P W, Hayes B J, and Goddard M E (2009), Reliability of genomic predictions across multiple populations. *Genetics* 183(4):1545–1553
- de Roos A P W, Hayes B J, Spelman R J, and Goddard M E (2008), Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179(3):1503–1512
- Roos M, Martins T G, Held L, and Rue H (2015), Sensitivity analysis for Bayesian hierarchical models. *Bayesian Anal.* 10(2):321–349
- Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink J L, and Sorrells M (2012), Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. *Plant Genome J.* 5(2):51–61
- Schön C C, Utz H F, Groh S, Truberg B, Openshaw S, and Melchinger A E (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(1):485–498
- Schulz-Streeck T, Ogutu J O, Gordillo A, Karaman Z, Knaak C, and Piepho H P (2013), Genomic selection allowing for marker-by-environment interaction. *Plant Breed.* 132(6):532–538
- Schulz-Streeck T, Ogutu J O, Karaman Z, Knaak C, and Piepho H P (2012a), Genomic selection using multiple populations. *Crop Sci.* 52(6):2453–2461
- Schulz-Streeck T, Ogutu J O, and Piepho H P (2012b), Comparisons of single-stage and two-stage approaches to genomic selection. *Theor. Appl. Genet.* 126(1):69–82
- Shemyakin A (2014), Hellinger distance and non-informative priors. *Bayesian Anal.* 9(4):923–938
- e Silva F F, Viana J M S, Faria V R, and de Resende M D V (2013), Bayesian inference of mixed models in quantitative genetics of crop species. *Theor. Appl. Genet.* 126(7):1749–1761

-
- Sneath P H, and Sokal R R (1973), Numerical taxonomy: the principles and practice of numerical classification. Freeman, San Francisco, CA
- Sorensen D, and Gianola D (2002), Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York
- Sun X, Ma P, and Mumm R H (2012), Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. PLoS ONE 7(11):e50604
- Technow F, Bürger A, and Melchinger A E (2013), Genomic prediction of Northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3: Genes, Genomes, Genet. 3(2):197–203
- Technow F, and Melchinger A E (2013), Genomic prediction of dichotomous traits with Bayesian logistic models. Theor. Appl. Genet. 126(4):1133–1143
- Technow F, Schrag T A, Schipprack W, Bauer E, Simianer H, and Melchinger A E (2014), Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197(4):1343–1355
- Tibshirani R (1996), Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. B 58(1):267–288
- Toosi A, Fernando R L, and Dekkers J C M (2009), Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88(1):32–46
- Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom T M, Fries R, Pausch H, Bertani C, Davassi A, Mayer K F, and Schön C C (2014), A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. BMC Genomics 15(1):823
- VanRaden P (2008), Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11):4414–4423
- VanRaden P, Van Tassell C, Wiggans G, Sonstegard T, Schnabel R, Taylor J, and Schenkel F (2009), Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92(1):16–24

-
- Wang D, Salah El-Basyoni I, Stephen Baenziger P, Crossa J, Eskridge K M, and Dweikat I (2012), Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity* 109(5):313–319
- Wang Y, Mette M, Miedaner T, Gottwald M, Wilde P, Reif J C, and Zhao Y (2014), The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. *BMC Genomics* 15(1):556
- Wen W, Guo T, Tovar V H C, Li H, Yan J, and Taba S (2012), The strategy and potential utilization of temperate germplasm for tropical germplasm improvement: a case study of maize (*Zea mays* L.). *Mol. Breeding* 29(4):951–962
- Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, and Kilian A (2004), Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *P. Natl. Acad. Sci.* 101(26):9915–9920
- Wientjes Y C, Veerkamp R F, Bijma P, Bovenhuis H, Schrooten C, and Calus M P (2015), Empirical and deterministic accuracies of across-population genomic prediction. *Genet. Sel. Evol.* 47(1)
- Wientjes Y C J, Veerkamp R F, and Calus M P L (2012), The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193(2):621–631
- Williams E, Piepho H P, and Whitaker D (2011), Augmented p-rep designs. *Biometrical J.* 53(1):19–27
- Wimmer V, Albrecht T, Auinger H J, and Schön C C (2012), synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28(15):2086–2087
- Wimmer V, Lehermeier C, Albrecht T, Auinger H J, Wang Y, and Schön C C (2013), Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195(2):573–587
- Windhausen V S, Atlin G N, Hickey J M, Crossa J, Jannink J L, Sorrells M E, Raman B, Cairns J E, Tarekegne A, Semagn K, Beyene Y, Grudloyma P, Technow F, Riedelsheimer

-
- C, and Melchinger A E (2012), Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3: Genes, Genomes, Genet.* 2(11):1427–1436
- Würschum T, Abel S, and Zhao Y (2014), Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. *Plant Breed.* 133(1):45–51
- Würschum T, Reif J C, Kraft T, Janssen G, and Zhao Y (2013), Genomic selection in sugar beet breeding populations. *BMC Genet.* 14(1):85
- Wu B, Nianjun L, and Zhao H (2006), PSMIX: an R package for population structure inference via maximum likelihood method. *BMC Bioinformatics* 7:317
- Xu S (2013), Genetic mapping and genomic selection using recombination breakpoint data. *Genetics* 195(3):1103–1115
- Xu S, Zhu D, and Zhang Q (2014), Predicting hybrid performance in rice using genomic best linear unbiased prediction. *P. Natl. Acad. Sci.* 111(34):12456–12461
- Yang W, Chen C, and Tempelman R J (2015), Improving the computational efficiency of fully Bayes inference and assessing the effect of misspecification of hyperparameters in whole-genome prediction models. *Genet. Sel. Evol.* 47(1)
- Zapata-Valenzuela J, Whetten R W, Neale D, McKeand S, and Isik F (2013), Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. *G3: Genes, Genomes, Genet.* 3(5):909–916
- Zhang X, Pérez-Rodríguez P, Semagn K, Beyene Y, Babu R, López-Cruz M A, San Vicente F, Olsen M, Buckler E, Jannink J L, Prasanna B M, and Crossa J (2015), Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity* 114:291–299
- Zhang Z, Liu J, Ding X, Bijma P, de Koning D J, and Zhang Q (2010), Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5(9):e12648
- Zhao K, Tung C W, Eizenga G C, Wright M H, Ali M L, Price A H, Norton G J, Islam M R, Reynolds A, Mezey J, McClung A M, Bustamante C D, and McCouch S R (2011),

Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2:467

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

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

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

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

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

Zhong S, Dekkers J C M, Fernando R L, and Jannink J L (2009), Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics* 182(1):355–364

Zhou L, Lund M, Wang Y, and Su G (2014), Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *J. Anim. Breed. Genet.* 131(4):249–257

Zou H, and Hastie T (2005), Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B* 67(2):301–320

7 Appendix

7.1 Supporting Table

Table A1: Overview of used genome-based prediction methods and analyzed species in published genomic selection studies, using different experimental plant datasets (up to the end of 2014).

Publication	Methods	Species
Albrecht et al. (2011)	GBLUP	maize
Albrecht et al. (2014)	GBLUP	maize
Asoro et al. (2011)	RR-BLUP, BayesC π	oats
Bao et al. (2014)	RR-BLUP, BayesL, BayesC π , SVM, RF	soybean
Bentley et al. (2014)	RR-BLUP, LASSO, Elastic Net	wheat
Bernal-Vasquez et al. (2014)	RR-BLUP	rye
Biscarini et al. (2014)	GBLUP	sugar beet
Burgueño et al. (2012)	GBLUP	wheat
Charmet et al. (2014)	GBLUP, BRR, BayesL, RKHS, RF	wheat
Combs and Bernardo (2013)	RR-BLUP	maize
Cros et al. (2015)	GBLUP, BayesL, BRR, BayesC π , BayesD π	oil palm
Crossa et al. (2010)	BayesL, RR-BLUP, RKHS	wheat, maize
Crossa et al. (2013)	RKHS, Bayesian GBLUP	maize
Cuevas et al. (2014)	BRR, BayesA (variations)	wheat, maize
Daetwyler et al. (2014)	GBLUP, BayesR	wheat
Dawson et al. (2013)	GBLUP	wheat
de los Campos et al. (2009)	BayesL	wheat
de Oliveira et al. (2012)	RR-BLUP	cassava
El-Kassaby et al. (2012)	GBLUP	cotton tree
e Silva et al. (2013)	Bayesian GBLUP	maize
Fodor et al. (2014)	RR-BLUP, BayesL	loblolly pine
Fritsche-Neto et al. (2012)	GBLUP	maize
Gianola et al. (2014)	GBLUP, Bagging GBLUP	wheat
González-Camacho et al. (2012)	ANN, RKHS, BayesL	maize
Gouy et al. (2013)	RR-BLUP, BayesL, RKHS, PLSR	sugar cane
Gowda et al. (2013)	RR-BLUP	wheat
Guo et al. (2012)	RR-BLUP, BayesA, BayesB	maize
Guo et al. (2013a)	GBLUP	maize
Guo et al. (2013b)	RR-BLUP	maize
Guo et al. (2014)	GBLUP	rice, maize
Heffner et al. (2011a)	RR-BLUP, BayesA, BayesB, BayesC π	wheat
Heffner et al. (2011b)	RR-BLUP, BayesC π	wheat
Heslot et al. (2012)	RR-BLUP, BayesL, Elastic Net, wBSR, BayesC π , E-Bayes, RKHS, SVM, RF, ANN	arabidopsis, wheat, barley, maize
Heslot et al. (2013)	BayesL	barley
Heslot et al. (2014)	GBLUP, sparse group LASSO	wheat
Hofheinz et al. (2012)	RR-BLUP	sugar beet
Hofheinz and Frisch (2014)	RR-BLUP, BayesL	maize, wheat, sugar beet
Isidro et al. (2014)	RR-BLUP	wheat, rice

Continued on next page

Table A1 – *Continued from previous page*

Publication	Methods	Species
Jacobson et al. (2014)	GBLUP	maize
Jarquín et al. (2014a)	Bayesian GBLUP	wheat
Jarquín et al. (2014b)	GBLUP	wheat
Jiang et al. (2014)	RR-BLUP, RKHS, BayesC π	wheat
Kumar et al. (2012)	RR-BLUP, BayesL	apple
Lado et al. (2013)	GBLUP	wheat
Lehermeier et al. (2013)	BRR, BayesL, BayesA, BayesB	maize
Lehermeier et al. (2014)	GBLUP, BayesC π	maize
Li and Sillanpää (2012)	LASSO, Elastic Net, BayesL, adapt. Lasso	barley
Lipka et al. (2014)	RR-BLUP, LASSO, Elastic Net	switchgrass
Lorenz et al. (2012)	RR-BLUP, BayesL, BayesC π	barley
Lorenzana and Bernardo (2009)	RR-BLUP, E-Bayes	maize, arabidopsis, barley
Ly et al. (2013)	GBLUP	cassava
Massman et al. (2012)	GBLUP	maize
Massman et al. (2013)	GBLUP, RR-BLUP	maize
Morota et al. (2013)	RKHS	wheat
Munoz et al. (2014)	GBLUP	loblolly pine
Onogi et al. (2015)	GBLUP, RKHS, LASSO, Elastic Net, RF, BayesL, extended BayesL, WBSR	rice
Ornella et al. (2012)	GBLUP, BayesL, SVM	wheat
Ornella et al. (2014)	RR-BLUP, BayesL, RKHS, RF, SVM	maize, wheat
Ould Estaghevrou et al. (2013)	GBLUP	maize
Ould Estaghevrou et al. (2014)	GBLUP	maize
Owens et al. (2014)	RR-BLUP, LASSO, Elastic Net	maize
Peiffer et al. (2013)	GBLUP	maize
Peiffer et al. (2014)	GBLUP	maize
Pérez-Rodríguez et al. (2013)	BRR, BayesA, BayesB, BayesL, ANN, RKHS	wheat
Piepho (2009)	RR-BLUP	maize
Poland et al. (2012)	GBLUP	wheat
Reif et al. (2013)	GBLUP	sunflower
Resende et al. (2012a)	GBLUP	eucalyptus
Resende et al. (2012b)	RR-BLUP, BayesA, BayesL, BayesC π	loblolly pine
Resende et al. (2012c)	RR-BLUP	loblolly pine
Riedelsheimer et al. (2012a)	GBLUP	maize
Riedelsheimer et al. (2012b)	RR-BLUP, LASSO, Elastic Net, RKHS, BayesB	maize
Riedelsheimer et al. (2013)	GBLUP	maize
Rincet et al. (2012)	GBLUP	maize
Rutkoski et al. (2012)	RR-BLUP, BayesL, RKHS, RF	wheat
Schulz-Streeck et al. (2012a)	RR-BLUP, Elastic Net, LASSO	maize
Schulz-Streeck et al. (2012b)	GBLUP	maize
Schulz-Streeck et al. (2013)	RR-BLUP	maize
Sun et al. (2012)	RR-BLUP, BayesA, BayesB, RKHS	maize
Technow and Melchinger (2013)	Bayesian GBLUP, BayesB	wheat, maize
Technow et al. (2013)	Bayesian GBLUP	maize

Continued on next page

7 APPENDIX

Table A1 – *Continued from previous page*

Publication	Methods	Species
Technow et al. (2014)	GBLUP, BayesB	maize
Wang et al. (2012)	adaptive LASSO	wheat
Wang et al. (2014)	RR-BLUP	rye
Wen et al. (2012)	GBLUP	maize
Wimmer et al. (2013)	GBLUP, BayesB, Elastic Net, LASSO	arabidopsis, wheat, rice
Windhausen et al. (2012)	GBLUP	maize
Würschum et al. (2013)	RR-BLUP	sugar beet
Würschum et al. (2014)	RR-BLUP, BayesB	rapeseed
Xu (2013)	LASSO	rice
Xu et al. (2014)	GBLUP	rice
Zhang et al. (2015)	GBLUP	maize
Zapata-Valenzuela et al. (2013)	GBLUP	loblolly pine
Zhao et al. (2012a)	RR-BLUP	maize
Zhao et al. (2012b)	RR-BLUP	maize
Zhao et al. (2013a)	RR-BLUP	maize
Zhao et al. (2013b)	RR-BLUP, BayesA, BayesB, BayesC, BayesC π	wheat
Zhao et al. (2014)	RR-BLUP, BayesC π , W-BLUP (weighted BLUP)	wheat

7.2 Publications

The publications underlying this thesis, including supporting material, can be accessed via following links:

Lehermeier et al. 2013 <http://www.degruyter.com/view/j/sagmb.2013.12.issue-3/sagmb-2012-0042/sagmb-2012-0042.xml>

Lehermeier et al. 2014 <http://www.genetics.org/content/198/1/3>

Lehermeier et al. 2015 <http://www.genetics.org/content/early/2015/06/29/genetics.115.177394>

8 Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Chris-Carolin Schön who gave me the opportunity to do this PhD. Thank you for the constant advice, support and encouragement.

Thanks a lot also to Prof. Gustavo de los Campos (Ph. D.) for giving me the opportunity to stay at UAB for three months as well as for making a second visit possible. Thank you for sharing your ideas with me, your support and the good collaboration on the third paper of my thesis.

I also like to thank Prof. Daniel Gianola (Ph. D.) for his inspiring ideas and discussions during his visits at TUM and for being part of my graduate committee. Thanks a lot also to Prof. Donna Ankerst (Ph. D.) and Prof. Dr. Aurélien Tellier for serving on my graduate committee.

I would like to thank all co-authors and project partners for the good collaboration and for many interesting discussions.

A big note of thanks to the complete TUM plant breeding group and all members of the last years for creating a very nice working environment and for various help and assistance during this thesis. Especially I want to mention: Dr. Eva Bauer and Dr. Nicole Krämer for their support on the second paper of my thesis; Dr. Theresa Albrecht for her help during my start at the chair of plant breeding and for ongoing support and discussions as well as for proof reading this thesis; Dr. Valentin Wimmer and Hans-Jürgen Auinger for various help and the very good collaboration especially during the work on the first paper of this thesis and ongoing scientific discussions; Manfred Schönleben for his help especially with the Cornfed maize data, for introducing me to the most important maize lines, and his enthusiasm for my results; my office mates and remaining PhD fellows Wiltrud Erath, Sandra Unterseer, Sebastian Steinemann, Christos Dadousis, Flavio Foiada, Dr. Sebastian Gresset, and Katrin Töpner for their help with any problems and always having encouraging words.

This research was made possible through financial support from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Synbreed - Synergistic plant and animal breeding” (FKZ 0315528A) and through the PLANT-KBBE Initiative CornFed, with funding from the BMBF (Germany), Agence Nationale de la Recherche (ANR, France), and Ministry of Science and Innovation (MICINN, Spain).

8 ACKNOWLEDGEMENTS

Zu guter letzt: Vielen Dank an meine Familie, vor allem meinen Eltern und Mathias für ihre beständige Unterstützung, ihre Geduld und dafür, dass sie immer für genügend Essen für mich sorgten.

9 Curriculum Vitae

Personal Information

Christina Lehermeier

Date of birth: February, 23 1987

Place of birth: Straubing, Germany

Nationality: German

Education and Position

- | | |
|-------------------|--|
| 10-2011 – current | Research Assistant, Chair of Plant Breeding, Technische Universität München, Germany |
| 10-2011 – current | Ph.D. student, Chair of Plant Breeding, Technische Universität München, Germany
Supervisor: Prof. Dr. Chris-Carolin Schön |
| 03-2014 – 05-2014 | Visiting Scholar, Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, USA
Supervisor: Prof. Gustavo de los Campos, PhD |
| 10-2006 – 08-2011 | Master of Science (Diplom) in Statistics, Department of Statistics, Ludwig-Maximilians-Universität München, Germany
Thesis: Bayesianische Prädiktion in der Pflanzenzüchtung mittels molekularer Marker
Supervisor: Prof. Dr. Volker Schmid, Prof. Dr. Chris-Carolin Schön |
| 09-1997 – 06-2006 | Veit-Höser-Gymnasium Bogen (secondary school)
Degree: Allgemeine Hochschulreife (Abitur) |
| 09-1993 – 07-1997 | Grundschule Oberalteich (elementary school), Germany |

Publications

- de los Campos G., Veturi Y., Vazquez A. I., **Lehermeier C.**, Pérez-Rodríguez P. (2015) Incorporating genetic heterogeneity in whole-genome regressions using interactions. *Journal of Agricultural, Biological, and Environmental Statistics* (submitted)
- Giraud H., **Lehermeier C.**, Bauer E., Falque M., Segura V., Bauland C., Camisan C., Campo L., Meyer N., Ranc N., Schipprack W., Flament P., Melchinger A.E., Menz M., Moreno-González J., Ouzunova M., Charcosset A., Schön C.-C., Moreau L. (2014) Linkage disequilibrium with linkage analysis of multi-line crosses reveals different multi-allelic QTL for hybrid performance in the flint and dent heterotic groups of maize. *Genetics* 198:1717-1734
- Lehermeier C.**, Schön C.-C., de los Campos G. (2015) Assessment of genetic heterogeneity in structured plant breeding populations using multivariate whole-genome regression models. *Genetics*. doi:10.1534/genetics.115.177394
- Lehermeier C.**, Krämer N., Bauer E., Bauland C., Camisan C., Campo L., Flament P., Melchinger A.E., Menz M., Meyer N., Moreau L., Moreno-González J., Ouzunova M., Pausch H., Ranc N., Schipprack W., Schönleben M., Walter H., Charcosset A., Schön C.-C. (2014) Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. *Genetics* 198:3-16
- Lehermeier C.**, Wimmer V., Albrecht T., Auinger H.-J., Gianola D., Schmid V. J., Schön C.-C. (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. *Statistical Applications in Genetics and Molecular Biology* 12: 375–391
- Wimmer V., **Lehermeier C.**, Albrecht T., Auinger H.-J., Wang Y., Schön C.-C. (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* 195:573–587

Contributions to scientific conferences

- Albrecht T., Wimmer V., Auinger H.-J., **Lehermeier C.**, Knaak C., Ouzunova M., Schön C.-C. (2012) Prediction of maize testcross performance across environments. 4th

International Conference of Quantitative Genetics, Edinburgh

Giraud H., **Lehermeier C.**, Bauer E., Falque M., Segura V., Bauland C., Camisan C., Campo L., Meyer N., Ranc N., Schipprack W., Flament P., Melchinger A.E., Menz M., Moreno-González J., Ouzunova M., Charcosset A., Schön C.-C., Moreau L. (2015) Joint analysis of European nested association mapping populations reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. 57th Annual Maize Genetics Conference, St. Charles

Lehermeier C., Wimmer V., Albrecht T., Auinger H.-J., Schön C.-C., Schmid V.J. (2012) Sensitivity to prior specification in Bayesian models for genomic prediction in maize. 4th International Conference of Quantitative Genetics, Edinburgh (own poster presentation)

Lehermeier C., Bauer E., Schönleben M., Walter H., Bauland C., Camisan C., Campo L., Meyer N., Ranc N., Schipprack W., Altmann T., Flament P., Melchinger A. E., Menz M., Moreno-González J., Ouzunova M., Charcosset A., Schön C.-C. (2012) Genomic prediction of testcross performance in multi-line cross of maize (*Zea mays* L.). XVth EUCARPIA Biometrics in Plant Breeding Conference, Hohenheim (own poster presentation)

Lehermeier C., Krämer N., Bauer E., Bauland C., Camisan C., Campo L., Flament P., Melchinger A.E., Menz M., Meyer N., Moreau L., Moreno-González J., Ouzunova M., Pausch H., Ranc N., Schipprack W., Schönleben M., Walter H., Charcosset A., Schön C.-C. (2014) Usefulness of multiparental populations of maize for genome-based prediction. GPZ congress, Kiel (own talk)

Lehermeier C., Schön C.-C., de los Campos G. (2015) Assessment of genetic heterogeneity in structured plant breeding populations using multivariate whole-genome regression models. Quantitative Genetics and Genomics Research Conference, Lucca (own poster presentation)

Lehermeier C., Knaak C., Ouzunova M., de los Campos G., Schön C.-C. (2015) Assessment of genetic heterogeneity in a maize breeding population using multivariate whole-genome regression models. Synbreed Colloquium, Freising (own poster presentation)

- Schön C.-C., Wimmer V, **Lehermeier C.** (2014) Efficiency of variable selection in genome-wide prediction for traits of different genetic architecture. Proceedings, 10th World Congress of Genetics Applied to Livestock Production (WCGALP), Vancouver
- Wimmer V, Albrecht T, **Lehermeier C.**, Auinger H.-J., Wang Y., Knaak C., Ouzunova M., Schön C.-C. (2012) Inferences about the genetic architecture of complex traits from genome-based prediction models. XVth EUCARPIA Biometrics in Plant Breeding Conference, Hohenheim
- Wimmer V, **Lehermeier C.**, Albrecht T., Auinger H.-J., Wang Y., Knaak C., Ouzunova M., Schön C.-C. (2013) Efficiency of variable selection in genome-wide prediction for traits of different genetic architecture. Quantitative Genetics and Genomics Research Conference, Galveston