Aus dem Institut für Pflanzenbau und Pflanzenzüchtung II
der Justus-Liebig-Universität Gießen
Professur für Biometrie und Populationsgenetik
Prof. Dr. Matthias Frisch

Implementation of genome-wide prediction methods in applied plant breeding programs

Dissertation zur Erlangung des akademischen Grades eines

Doktors der Agrarwissenschaften

- Dr. agr. -

im Fachbereich
Agrarwissenschaften, Ökotrophologie und Umweltmanagement
der Justus-Liebig-Universität Gießen

vorgelegt von

Nina Hofheinz
aus Ewersbach, Hessen

Gießen, im Juni 2014

Contents

[1] Hofheinz, N., D. Borchardt, K. Weissleder and M. Frisch (2012) Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theoretical and Applied Genetics* **125**: 1639-1645.

[2] Hofheinz, N., and M. Frisch (2014) Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3: Genes | Genomes | Genetics* **4**: 539-546.

Abbreviations

ANOVA	analysis of variance
BLUP	best linear unbiased prediction
cM	centimorgan
CMS	cytoplasmic male sterility
GBS	genotyping-by-sequencing
GWP	genome-wide prediction
HEM	heteroscedastic effects model
LD	linkage disequilibrium
MAS	marker-assisted selection
Mb	megabases
QTL	quantitative trait loci
REML	restricted maximum likelihood
RIR	ridge regression employing preliminary estimates of the heritability
RMLV	modification of the restricted maximum likelihood procedure that yields heteroscedastic variances
RMLA	estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to analysis of variance components
RR-BLUP	ridge regression BLUP
RRWA	ridge regression with weighing factors according to analysis of variance components
SNP	single nucleotide polymorphism

Chapter 1

General introduction

Plant breeding aims at improving crop cultivars for particular traits in order to fulfill human needs and to meet environmental requirements. Every breeding program consists of the three main stages: (1) Generate genetic variability; (2) select potential parents for cultivars; and (3) test the experimental cultivars. My work mainly focused on selection of the best performing individuals from the created genetic variation in the second stage. With the invention of molecular markers, marker-assisted selection (MAS) for improving quantitative traits emerged in the 1990s (*cf.* Xu and Crouch 2008) and facilitated the effective selection of breeding candidates based on their genotype. In a first step, quantitative trait loci (QTL) are mapped with molecular markers and in a second step, marker effects are estimated (*cf.* Lande and Thompson 1990; Collard and Mackill 2008). MAS has successfully been implemented in various crops for monogenic and oligogenic traits like resistances (Collard and Mackill 2008). The main drawback of MAS is that only markers with significant effects are considered and therefore it fails to capture the complete genetic variance (Goddard and Hayes 2007). This results in an overestimation of QTL effects and has been referred to as the Beavis effect (*cf.* Utz and Melchinger 1994; Xu 2003). By avoiding the selection of markers with significant effects, Meuwissen *et al.* (2001) introduced genome-wide prediction (GWP) to the animal breeding community. The focus of GWP lies on improving complex traits, for which MAS was not successful (Heffner *et al.* 2009). With GWP, breeding candidates are selected based on molecular markers covering the whole genome. This avoids the critical step of selecting significant markers and is expected to improve quantitative traits, which are influenced by many genes with small effects on the trait. Widely used synonymous terms for GWP are genomic selection, genomic prediction and genome-based prediction.

Implementation of GWP in plant breeding programs

High-throughput marker systems enabled cheap genotyping and triggered the increasing research of GWP in animal and plant breeding. A commonly used marker type for genotyping of plants are single nucleotide polymorphisms (SNPs), which are highly abundant (*cf.* Gupta *et al.* 2001). Due to variations of single nucleotide bases in the genome, SNPs are able to differentiate between individuals on the molecular level. As phenotyping is still cost- and time-intensive, one potential of GWP lies in shifting the focus from field testing of parental lines to the prediction of their performance from molecular data (*cf.* Nakaya and Isobe 2012; Heffner *et al.* 2010). Much research has been done in the field of GWP, its practical implementation in applied plant breeding programs is, however, still progressing rather slowly.

The process of GWP is illustrated in Figure 1.1. In a first step, genetic effects of markers are estimated with a statistical method for a set of individuals named “estimation set”. Individuals in the estimation set are genotyped and phenotyped for the traits of interest. The genetic effects can be used in order to predict genotypic values of individuals in the “prediction set”. Individuals in the prediction set are only genotyped with molecular markers across the whole genome. Estimated genetic effects from the estimation set and genotypic data from the individuals in the prediction set are both required for the prediction of genotypic values.

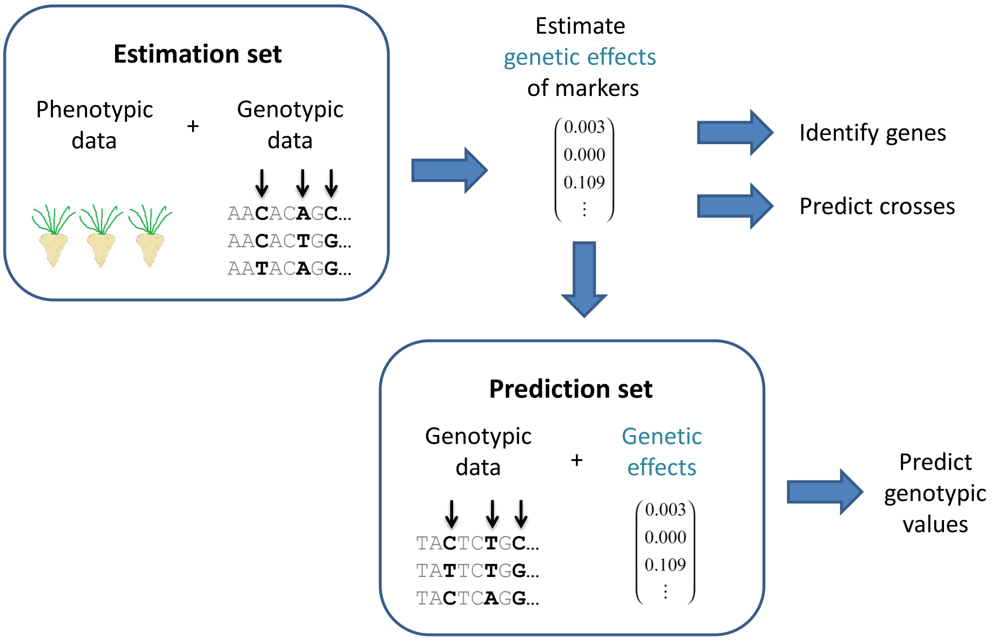

Figure 1.1. Illustration of GWP in applied plant breeding programs. Individuals in the estimation set are genotyped and phenotyped, whereas individuals in the prediction set are only genotyped. Genetic effects of markers are estimated with a statistical method for GWP. Three exemplary SNP markers are indicated by black arrows. The genetic effects can be used to identify genes, predict crosses or predict genotypic values for individuals in the prediction set.

The obtained genetic effects can also be used to identify functional genes, which can further be used for marker-assisted introgression programs (Figure 1.1). However, accurate localization and effect size estimation of markers controlling the traits of interest are prerequisites for this application. Besides the identification of functional genes, plant breeders have high expectations for the prediction of crosses. The process of selecting favorable parental combinations for crosses has a very high impact on the success of a breeding program. Crosses can be predicted by estimating expectation and variance of the performance of a population derived from crossing two parental genotypes. The success of cross prediction with GWP strongly depends on accurate marker effect estimates in order to model the recombination of parental alleles.

GWP and sugar beet hybrid breeding

Throughout my work, the main focus was on the prediction of test cross performance of sugar beet (*Beta vulgaris* L.) inbred lines in an applied hybrid breeding program. The discovery of cytoplasmic male sterility (CMS) by Owen (1945) enabled the commercial hybrid production of sugar beet. Historically, the crop has faced an extreme genetic bottleneck in the 1960s when the system of CMS was developed with the simultaneous introduction of the highly desired monogermic seed character (Biancardi *et al.* 2010). Ever since, hybrid sugar beet seed is produced by crossing monogermic, cytoplasmic male-sterile plants with multigermic, fertile pollinators. In sugar beet hybrid breeding programs, the number of subpopulations of each heterotic pool equals the number of years required for a recurrent selection cycle. This ensures one completed breeding cycle each year and therefore the permanent availability of improved breeding material. The progeny of each breeding cycle is evaluated in field trials for its test cross performance.

After the invention of molecular markers, the introgression of the major gene *RZ1* into pollinator lines for resistance against rhizomania became a popular example for an effective application of MAS in plant breeding (De Biaggi *et al.* 2010). Recently, we demonstrated the successful application of GWP in sugar beet breeding (Hofheinz *et al.* 2012) and thereafter, concordant results were presented by Würschum *et al.* (2013). The genome sequence of sugar beet has been published most recently by Dohm *et al.* (2014) and will facilitate the availability of cheap genome-wide dense molecular marker maps.

Statistical methods for GWP

Genome-wide dense marker maps lead to an overparameterization of GWP methods by fitting more marker data (p) than individuals (n). In order to overcome this problem, ridge regression or variable selection methods can be employed. The latter is utilized by the broad variety of Bayesian methods (*cf.* Meuwissen *et al.* 2001; Kärkkäinen and Sillanpää 2012). Here, specified prior distributions result in heteroscedastic marker variances. These marker-specific variances are expected to meet genetic requirements of monogenic and oligogenic traits, for example resistance traits, which are only influenced by one or a small number of genes. The main drawbacks of the Bayesian methods are their tremendous computational demand and difficulties concerning the definition of prior distributions and hyperparameters.

Throughout my work, I focused on methods employing shrinkage factors using ridge regression. Here, all markers are included in the model, but each estimated genetic effect is shrunk with a shrinkage factor λ. If the degree of shrinkage is determined by dividing the error variance (σ_e) by the genetic variance (σ_g), the estimates are equivalent to best linear unbiased predictions (BLUP) of genetic effects (*cf.* Piepho 2009). The required restricted maximum likelihood (REML) estimates of the variance components can be obtained by using an expectation-maximization algorithm. BLUP estimates of genetic effects are obtained by shrinking each genetic effect to the same extent, no matter how much influence the marker has on the trait. A constant shrinkage factor results in homoscedastic genetic variances for all markers. Homogeneous shrinkage proved to work well for the prediction of quantitative traits, for example yield-related traits, which are influenced by many genes, each having a small effect on the trait (*cf.* Crossa *et al.* 2010; Hofheinz *et al.* 2012).

The assumption of homogeneous shrinkage will, however, circumvent the applications of GWP methods in which accurate marker effect estimates are needed. Therefore, heteroscedastic marker variances are a prerequisite for applications like prediction of the performance of crosses or identification of genes in applied plant breeding programs. Due to the rapidly growing dimension of data sets, solutions which address the computational obstacles that arise with the Bayesian methods are required. Our goal was to develop novel, computationally efficient ridge regression methods employing heteroscedastic marker variances (Hofheinz and Frisch 2014). Here, each genetic effect is shrunk with a marker-specific shrinkage factor. Simultaneously, Shen *et al.* (2013) introduced the heteroscedastic effects model (HEM) as a generalized ridge regression approach employing marker-specific shrinkage in a non-Bayesian framework.

Implementation of GWP in different crops

In the beginning of the 21st century, the pioneering publication of Meuwissen *et al.* (2001) rapidly revolutionized the field of animal breeding. A few years later, many simulation studies which focused on plant breeding followed (*e.g.*, Bernardo and Yu 2007; Wong and Bernardo 2008; Bernardo 2009; Zhong *et al.* 2009). The main goal of these studies was to show the theoretical potential of GWP to overcome the shortcomings of MAS and QTL mapping. Recently, GWP studies employing experimental data sets from different crops such as maize (*cf.* Piepho 2009; Crossa *et al.* 2010; Albrecht *et al.* 2011), wheat (*cf.* Heffner *et al.* 2011; Zhao *et al.* 2013), barley (*cf.* Lorenz *et al.* 2012), sugar beet (Hofheinz *et al.* 2012; Würschum *et al.* 2013), rapeseed (Würschum *et al.* 2014) and trees such as apple (Kumar *et al.* 2012) or loblolly pine (Resende *et al.* 2012) were published.

The focus of the above-mentioned studies was on the validation of GWP methods for their prediction accuracy. All individuals in the data sets of these studies were genotyped and phenotyped and cross validation was used to assess the prediction accuracy. Cross validation randomly divides the data set into two parts: The estimation set is used for estimating the marker effects and the prediction set is used for validating the predicted genotypic values. As a measure of prediction accuracy from repeated cross validations, the Pearson correlation coefficient between observed phenotypic values and predicted genotypic values is usually calculated. Additionally, the estimated correlation can be divided by the square root of the heritability (*cf.* Daetwyler *et al.* 2013). This measure is useful when the investigated data sets basically consist of individuals belonging to the same breeding cycle, as it is the case in the above-mentioned GWP studies. However, plant breeders desire the prediction of lines with marker effects estimated in a previous breeding cycle. For such cases, cross validation might not be sufficient for assessing the accuracy of prediction and therefore, independent validation is required. Further research about the transferability of marker effect estimates to subsequent breeding cycles needs to be investigated.

The common conclusion from most empirical studies comparing GWP methods employing homo- and heteroscedastic marker variances was that differences in prediction accuracies were negligible regardless of the trait architecture (*cf.* Heslot *et al.* 2012; Wimmer *et al.* 2013). However, less effort has been invested so far in the comparison of GWP methods with respect to their accuracy of marker effect estimation. Sizes of marker effect estimates obtained with different GWP methods have been compared in few studies (Lorenz *et al.* 2012; Kumar *et al.* 2012; Shen *et al.* 2013) and remarkable differences were demonstrated. Kumar *et al.* (2012) described a strong shrinkage of small effect markers and little shrinkage of markers with greater effects for a Bayesian GWP method. In the simulation study of Wimmer *et al.* (2013), accuracies of estimated marker effects were compared for ridge regression BLUP (RR-BLUP) and several variable selection methods. It was shown that variable selection outperformed RR-BLUP for certain combinations of model complexity and determinedness level. These results highlight the importance of further investigations for accuracy of marker effect estimates from GWP methods. High accuracies are a prerequisite for applications of GWP like the identification and functional analysis of genes for introgression or for the prediction of cross performance. Moreover, besides accuracy of marker effect estimates, other criteria like computational efficiency, user-friendliness, etc. will become important for the evaluation of GWP methods in prospective studies.

Objectives

The main goal of my thesis research was to investigate GWP with newly developed ridge regression methods in applied plant breeding programs with a focus on sugar beet. Specifically, my objectives were to:

(1) Compare cross validation with independent validation using sugar beet lines from a subsequent breeding cycle. High and low heritable traits were investigated to analyze whether marker effects estimated in one breeding cycle can be used for the prediction of test cross performance in the subsequent breeding cycle.

(2) Propose a ridge regression approach that approximates BLUP estimates of genetic effects (RIR) to reduce the required computing time.

(3) Suggest novel heteroscedastic ridge regression approaches, where shrinkage factors are obtained by estimating single-marker variance components (RMLA, RRWA) or by modifying the restricted maximum

likelihood procedure with fixed residual variances in variance component estimation (RMLV) as alternatives to the Bayesian GWP methods.

(4) Investigate the properties of the novel ridge regression approaches with respect to prediction accuracy, computational efficiency and accuracy of effect estimates by analyzing simulated data and data sets from applied breeding programs of maize, wheat and sugar beet.

Chapter 2

Genome-based prediction of test cross performance in two subsequent breeding cycles [1]

[1] Hofheinz, N., D. Borchardt, K. Weissleder and M. Frisch (2012) Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theoretical and Applied Genetics* **125**: 1639-1645.

Theor Appl Genet (2012) 125:1639–1645
DOI 10.1007/s00122-012-1940-5

ORIGINAL PAPER

Genome-based prediction of test cross performance in two subsequent breeding cycles

Nina Hofheinz · Dietrich Borchardt ·
Knuth Weissleder · Matthias Frisch

Received: 26 June 2012 / Accepted: 29 June 2012 / Published online: 20 July 2012

Abstract Genome-based prediction of genetic values is expected to overcome shortcomings that limit the application of QTL mapping and marker-assisted selection in plant breeding. Our goal was to study the genome-based prediction of test cross performance with genetic effects that were estimated using genotypes from the preceding breeding cycle. In particular, our objectives were to employ a ridge regression approach that approximates best linear unbiased prediction of genetic effects, compare cross validation with validation using genetic material of the subsequent breeding cycle, and investigate the prospects of genome-based prediction in sugar beet breeding. We focused on the traits sugar content and standard molasses loss (ML) and used a set of 310 sugar beet lines to estimate genetic effects at 384 SNP markers. In cross validation, correlations >0.8 between observed and predicted test cross performance were observed for both traits. However, in validation with 56 lines from the next breeding cycle, a correlation of 0.8 could only be observed for sugar content, for standard ML the correlation reduced to 0.4. We found that ridge regression based on preliminary estimates of the heritability provided a very good approximation of best linear unbiased prediction and was not accompanied with a loss in prediction accuracy. We conclude that prediction accuracy assessed with cross validation within one cycle of a breeding program can not be used as an indicator for the accuracy of predicting lines of the next cycle. Prediction of lines of the next cycle seems promising for traits with high heritabilities.

Communicated by M. Sillanpää.

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

D. Borchardt · K. Weissleder
KWS Saat AG, 37555 Einbeck, Germany

Introduction

Prediction of genetic values with genome-wide dense marker maps was proposed in an animal breeding context by Meuwissen et al. (2001). Simulation studies (Bernardo and Yu 2007; Bernardo 2009; Wong and Bernardo 2008; Xu 2003; Zhong et al. 2009) suggested that it can overcome shortcomings limiting the application of QTL mapping and marker assisted selection in plant breeding.

In a study with maize, test cross performance for kernel dry weight of 208 doubled haploid lines was assessed in five locations (Piepho 2009). The lines were genotyped with 136 SNP and SSR markers and the model fit of various ridge regression models was assessed. It was suggested that genotype × environment interactions and genetic effects not captured by markers should be included in genome-based prediction models. Parametric and semi-parametric models for genome-based prediction were compared in a study using phenotypic data of 599 wheat lines grown in four environments and 300 maize lines grown under two different conditions (Crossa et al. 2010). 1,447 markers were used for the genotyping of the wheat lines and 1,148 markers for the maize lines. In cross validation, correlations between observed and predicted performance in the range of 0.4–0.5 were observed for grain yield and up to 0.79 for flowering time. Genome-based prediction with mixed linear models was investigated in a study with 1,380 doubled haploid maize lines grown in seven environments and phenotyped for the traits grain dry matter yield and grain dry matter content (Albrecht et al. 2011). The lines were genotyped with 1,152 SNP markers. In cross validation, correlations between predicted and observed test cross performance up to 0.74 were observed. While Piepho (2009) used the model fit in the estimation set as a measure to compare alternative models, Crossa et al. (2010) and Albrecht et al. (2011) used cross validation in the estimation set to assess prediction accuracy. However, no results are available investigating the accuracy of genome-based prediction when the set of lines to be predicted belongs to the breeding cycle that follows the breeding cycle to which the estimation set belongs.

The goal of our study was to assess the accuracy of genome-based prediction of test cross performance for sugar content (SC) and standard molasses loss (ML) in sugar beet (*Beta vulgaris* L.) by using data from two subsequent cycles of a breeding program. In particular our objectives were to (1) compare ridge regression employing preliminary estimates of the heritability (RIR) with best linear unbiased prediction (BLUP) for predicting marker effects, (2) compare cross validation for assessing prediction accuracy of genome-based prediction with validation using data from a subsequent breeding cycle, (3) draw conclusions on the potential of genome-based prediction in sugar beet breeding.

Methods

Plant material

The estimation set consisted of 310 inbred lines randomly derived from 34 crosses among 9 diploid sugar beet lines. The number of progenies from each cross ranged from two to seven. The 56 lines of the validation set were derived from 8 crosses among 6 lines of the estimation set. The number of progenies from each cross ranged from 3 to 11. The line development included selection between crossing parents as well as selection between lines. The lines were selected for high performance and to maintain the genetic diversity within the breeding pool.

Field data

Test cross performance of the lines of the estimation set was evaluated for SC (%) and ML (%) in field trials at six European locations with one tester. The lines were a subset of a larger trial that was set up in 10 × 10 lattices with two replications. The lines of the validation set were evaluated as part of a larger trial that employed alpha lattices with block size 10 at six European locations. A two-replicate design was employed. The first replicate was assigned to the first of two testers and the second replicate to the second. Four standard genotypes were included. The field trials were analysed with a two-stage analysis. In the first stage the adjusted entry means were calculated for each environment. These were combined in an analysis of series of experiments. The error variance for the analysis of the series was obtained by pooling the individual error variances. Following the practice of commercial sugar beet breeding, relative values were calculated that refer to the average of the standard lines. The relative values were calculated from the means across environments. The use of relative values might not be totally consistent with assumptions implicitly made by our further analyses; however, in the present study, our focus is on practical applicability.

Marker data

Genotyping was carried out with the same marker set of 384 SNPs in the estimation set and in the validation set. The nine chromosomes of sugar beet had lengths of about 1 M and the total map length was 10.25 M. Hence, the average map distance between two adjacent markers was 3.6 cM. Markers with more than two alleles, more than 20 % missing values, or a low degree of polymorphism ($1 - \sum f_i^2 < 0.1$, where f_1, f_2 are the allele frequencies at a marker) were discarded. This resulted in 300 SNPs for the estimation set and 198 SNPs for the validation set that were used for the calculations.

The marker data were used to investigate the relatedness of the material (Fig. 1) and the decay of linkage disequilibrium depending between pairs of loci depending on their map distance (Fig. 2).

Linear model

For estimating the genetic effects of the SNPs we used the linear model

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad (1)$$

where $\mathbf{y}$ is the vector of N phenotypic values, β_0 a fixed intercept, $\mathbf{Z}$ the design matrix relating the marker data to genotypes, $\mathbf{u}$ the vector of genetic effects, and $\mathbf{e}$ the vector of residuals. The genetic effects $u_l (l = 1\ldots m)$ at the m SNPs were assumed to follow a normal distribution with expectation 0 and variance σ_u^2. The residuals were assumed to follow a normal distribution with expectation 0 and variance σ_e^2. It was assumed that $\mathrm{cov}(u_i,u_j) = 0\ (i \neq j)$ and $\mathrm{cov}(e_k,e_l) = 0\ (k \neq l)$.

We assume that the possible allele effects at each locus follow a distribution with a common variance. An alternative model takes the allele frequencies at the individual loci into account and assumes that in the estimation set each locus contributes equally to the genetic variance (Crossa et al. 2010). For predicting the genetic values in a

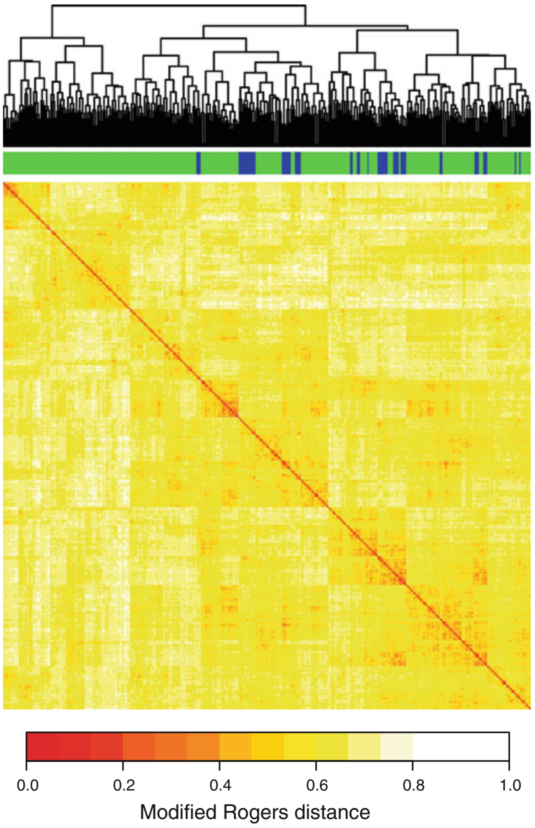

Fig. 1 Relatedness of the employed inbred lines based on the Modified Rogers distance determined from the SNP marker data. Average linkage clustering was used for ordering the distance matrix. Lines of cycle *n* are marked in *green* and lines of cycle $n + 1$ in *blue* (color figure online)

new validation set, our approach seems more suitable, because allele frequencies in the estimation and validation sets are most likely different.

The assumption of independent residuals in Eq. 1 is simplifying, because adjusted means are neither uncorrelated nor necessarily homoscedastic. It remains open to further research, whether more advanced linear models, that combine the analysis of the field design and the modeling of marker effects are able to increase the accuracy of prediction of genetic values.

Best linear unbiased prediction

We used an expectation-maximization (EM) algorithm to obtain restricted maximum likelihood (REML) estimates of the variance components σ_u^2 and σ_e^2 (Searle et al. 1992, p. 303). The EM algorithm is known to be slow in convergence and commercial software implements more sophisticated numerical approaches. However, it showed good performance for our data set. Convergence was reached with less than 10 iterations and computing times less than one second were required when using starting values determined on basis of Eq. 7. The algorithm showed high numerical stability and similar performance for other data sets from sugar beet and maize breeding programs.

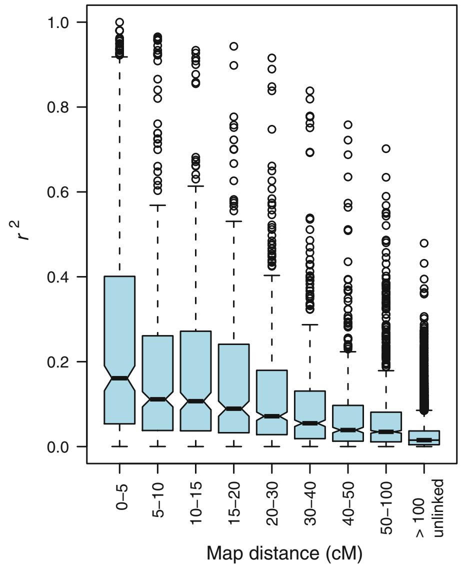

Fig. 2 Distribution of the pairwise linkage disequilibrium measure r^2 depending on the map distance between SNPs

To obtain best linear unbiased predictions (BLUP) of the genetic effects we solved (Searle 1987, p. 509)

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \lambda^2\mathbf{I} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix} \quad (2)$$

for **u** where

$$\lambda^2 = \sigma_e^2/\sigma_u^2 \quad (3)$$

An LU decomposition with back substitution (Press et al. 1992, p. 44) was used for solving Eq. 2.

With respect to terminology, we follow the literature on linear models (Searle 1987, Searle et al. 1992) and Meuwissen et al. (2001), and use the abbreviation BLUP for the best linear unbiased prediction of the elements of the **u** vector. Albrecht et al. (2011) employed the term random regression (Model RR) for a similar model.

Prediction with ridge regression

RIR was carried out by solving the mixed model equations (Eq. 2) with a fixed shrinkage parameter λ^2. As a starting point, we used the convenient but incorrect assumption (Bernardo and Yu 2007), that the variance due to each marker can be approximated by dividing an estimate of the genotypic variance by the number of markers. As pointed out by Piepho (2009), estimates of the genotypic variance are usually obtained from models assuming independent genotype effects. This is in contrast with the marker-based ridge regression model that implies correlation among genotypic effects. Due to this fundamental difference in the models, we do not claim mathematical rigour for the RIR approach suggested in the following.

To determine λ^2 we used preliminary estimates of the heritability h_{p}^2 that are typically available for the traits under selection in a breeding program. In a simplified model, these can be interpreted as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \tag{4}$$

where σ_u^2 is the genetic variance and σ_e^2 the residual variance. This approximation is a second point where we do not claim mathematical rigour for our approach: The masking variance used to obtain the heritability estimate typically includes not only the residual variance but further variance components. These are totally ignored in our interpretation of the heritability. This is expected to result in an inflated value for the error variance, resulting in a stronger shrinkage of the genotypic effects. However, if we make this simplification, we can write

$$\frac{\sigma_e^2}{\sigma_g^2} = \frac{1}{h^2} - 1, \tag{5}$$

and together with the assumption of equal variances of the marker effects, we can use the approximation

$$\sigma_u^2 \approx \frac{1}{m}\sigma_g^2 \tag{6}$$

to define

$$\lambda^2 = \frac{\sigma_e^2}{\sigma_u^2} = m\left(\frac{1}{h_{\mathrm{p}}^2} - 1\right). \tag{7}$$

Using a shrinkage factor as defined in Eq. 7 can be regarded as an approximation of the BLUP approach. The difference between RIR and BLUP is that with RIR the shrinkage factor is determined from genetic and residual variances that were approximated from results on preliminary estimates of heritability, while in the BLUP approach these variances are estimated from the data. Hence, if the variance components correspond to the marker data, as is the case in the simulation example of Shepherd et al. (2010), then Eq. 7 results in BLUP. If preliminary estimates for the heritability are used, then it approximates BLUP. To determine λ^2 for our experimental data, we used preliminary estimates of the heritabilities of $h_{\mathrm{p}}^2 = 0.9$ and 0.4 for the traits SC and ML. These values are not estimated for the particular set of material under consideration, nor approaches were employed to obtain the most precise heritability estimates possible for unbalanced data (Piepho and Möhring 2007). The appeal of the method lies in the fact that it employees rule-of-thumb estimates of the heritability that are easily available in breeding programs.

Validation

For assessing the prediction accuracy we carried out (a) cross validation within one breeding cycle and (b) validation with lines of the next breeding cycle. In each of 100 cross validation runs, the lines of the first breeding cycle were divided randomly to two parts, 254 lines were used to estimate marker effects and 56 lines to validate the effects. The correlations between observed and predicted test cross performance for RIR and BLUP were averaged over the 100 runs. For validation with lines from the next breeding cycle, we estimated the marker effects with the lines from the first breeding cycle and predicted the test cross performance of the lines of the subsequent breeding cycle. Then we assessed the correlation between the predicted and observed test cross performance.

Results

For SC the correlation between observed and predicted test cross values in the estimation set was $r = 0.94$ with RIR (employing a $h_p^2 = 0.9$) and $r = 0.93$ for BLUP. In cross validation, correlations of on average 0.82 were observed for both prediction models. Prediction of the test cross values of lines of the next breeding cycle resulted in correlations $r = 0.79$ (RIR) and 0.80 (BLUP, Fig. 3).

For ML the correlation between predicted and observed test cross values in the estimation set was slightly greater for BLUP ($r = 0.94$) than for RIR ($r = 0.90$). However, in cross validation similar average correlations of $r = 0.85$ (RIR) and 0.86 (BLUP) were observed for both prediction models. Despite these high correlations in cross validation, that were even greater than those observed for SC, the transferability of the effect estimates to the next breeding cycle was low. A correlation of $r = 0.41$ was observed for RIR (employing a $h_{\mathrm{p}}^2 = 0.4$) and $r = 0.39$ was observed for BLUP.

Discussion

Accuracy of prediction methods

Bayesian methods provided better prediction accuracy than BLUP in the study that initially suggested genome-based

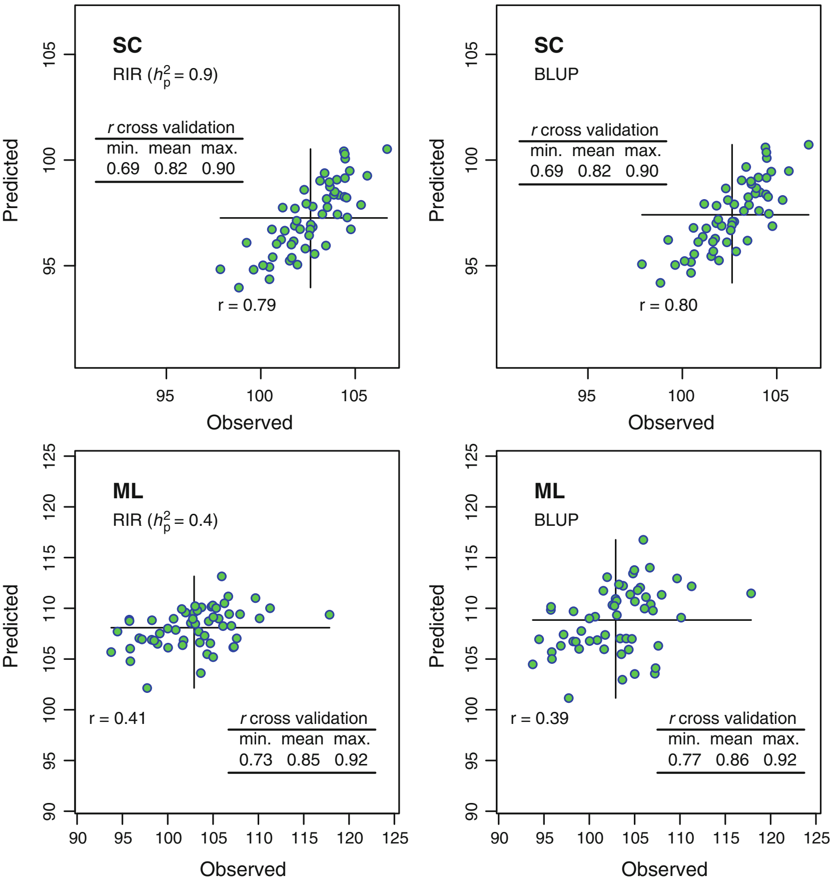

Fig. 3 Prediction of test cross performance for SC and ML. Observed versus predicted test cross performance in the validation set for prediction with RIR and BLUP. h^2_p is the preliminary estimate of the heritability employed in RIR and *r* the correlation between predicted and observed values. In the tables, the minimum, the mean, and the maximum of the correlations between predicted and observed test cross performance in the cross validation runs with the estimation set are presented

prediction of genetic values (Meuwissen et al. 2001). Since then much effort was invested in Bayesian estimation methods (Gianola and van Kaam 2008; Park and Casella 2008; Gianola et al. 2006) that allow for distributions of the genetic effects with unequal variances, because it was expected that they provide improved prediction accuracy. However, as pointed out by Piepho (2009) and Bernardo and Yu (2007), the fact that all genetic effects are modelled as realizations of random variables with the same variance does not imply that all loci contribute equally to the genetic value. It was suggested by Piepho (2009) and Goddard and Hayes (2007) that the advantage of Bayesian estimation over BLUP observed by Meuwissen et al. (2001) might be a consequence of the effect distributions in the employed simulation model. Bernardo and Yu (2007) concluded that for plant models, Bayesian methods would provide little, if any, advantage and Zhong et al. (2009) found BLUP to outperform Bayesian estimation. For grain yield in maize Albrecht et al. (2011) and Crossa et al. (2010) found that prediction accuracy of BLUP was similar to that of Bayesian estimation with varying variances. However, for flowering time superiority of Bayesian estimation was observed (Crossa et al. 2010). In accordance with Daetwyler et al. (2010), a possible conclusion from these studies is that approaches with variable variances might be superior for traits that are controlled by a few major genes. In contrast, for polygenic traits that follow closely the infinitesimal model of quantitative genetics, models with constant variances might be more appropriate. Schneider et al. (2002) detected five QTLs for SC on five chromosomes. They also found several QTLs for potassium, sodium, and alpha-amino nitrogen, which account for the trait ML. These results suggest that, due to the polygenic inheritance of SC and ML, BLUP is an appropriate method for genome-wide prediction in our data set.

Average correlations between predicted and observed test cross performance from the cross validation of BLUP were 0.82 for SC and 0.86 for ML (Fig. 3). These values confirm the results of Albrecht et al. (2011) and Piepho (2009) that BLUP can provide precise predictions, and support the hypothesis that for polygenic traits BLUP with constant variances is a suitable prediction method. RIR based on preliminary estimates of the heritability provided the same prediction accuracy as BLUP for both traits. With the present data set consisting of roughly 300 lines and 300 markers, obtaining BLUPs was not technically challenging. However, with large data sets convergence problems could occur. For such data sets an RIR approach might prove useful. In conclusion, BLUP provided genome-based predictions of high accuracy, and approximating BLUP on basis of preliminary estimates of heritabilities with RIR is a computationally simple alternative that was not accompanied with losses in prediction accuracy.

Cross validation and validation with the subsequent breeding cycle

The average correlations between predicted and observed test cross performance in cross validation were 0.82 (SC) and 0.86 (ML). Compared with results from maize and wheat (Crossa et al. 2010; Albrecht et al. 2011) these values are high. An explanation for the high correlations might be the homogeneity of the material in the investigated breeding pool. With an average distance between two adjacent markers of ≈ 3 cM, prediction of genetic values still relies on gametic disequilibrium between marker and QTL alleles. If the breeding material in a pool is homogeneous, then the linkage phase of marker and QTL alleles is expected to be the same for large parts of the material, resulting in high prediction accuracy. In more diverse breeding material, however, more dense marker maps, ideally to the point that each gene underlying a trait can be directly traced by a SNP, are expected to improve prediction accuracy.

The correlation between observed and predicted values in cross validation was smaller for SC ($h_\mathrm{p}^2 = 0.9$) than for ML ($h_\mathrm{p}^2 = 0.4$). This result indicates that even for traits with low heritabilities, good correlations between observed and predicted performance can be obtained in cross validation. The relatedness of the genotypes within a breeding pool can be a reason for such high correlations. The following example illustrates the problem. Assume several full sib lines that share common marker alleles at several loci not underlying the trait under consideration. In addition, they share a high performance. Some lines are part of the estimation set in a cross validation run and others are part of the validation set. As a consequence, high effect estimates are assigned to the common marker alleles, and these effects are validated by the sister lines in the validation set. An important conclusion from these results is that cross validation in breeding pools of related material does not necessarily correct prediction models for over-fitting. In consequence, high correlations between predicted and observed performance in cross validation do not guarantee a good transferability of the estimated effects to a different set of breeding material.

In contrast to cross validation, where the correlations between predicted and observed performance were high for both traits traits, in independent validation large differences were observed. While for SC correlations amounted to 0.8, only correlations of 0.4 were observed for ML. These correlations correspond well to the preliminary estimates of the heritability $h_\mathrm{p}^2 = 0.9$ (SC) and 0.4 (ML). This indicates that cross validation can only provide limited information on the accuracy of predicting line performance with effects estimated from a previous breeding cycle. In particular it remains open to further research whether results comparing the accuracy of different prediction models are robust with respect to the difference between cross validation and independent validation.

Application in breeding programs

Test cross performance of lines in hybrid breeding can be predicted either with effects estimated from related lines of the same breeding cycle or with effects estimated in a previous breeding cycle. Prediction of untested lines with an estimation set from the same breeding cycle can be implemented by generating more candidate lines than will be evaluated in field trials. After having evaluated a portion of the lines in field trials, the performance of the second portion of lines is predicted, and the lines with the best predictions were included in the second stage of line testing. Employing genome-based selection in such a scenario is conceptually similar to the assessment of prediction accuracy with cross validation. Due to the relatedness of the breeding material, even random associations between markers and phenotypes can be exploited by genome-based prediction. The high correlations in cross validation suggest that a considerable gain in response to selection can be realized with such applications.

Prediction of lines with an estimation set from the previous breeding cycle can be implemented as follows. More candidate lines are generated than will be evaluated in the field trials. All of these are genotyped and those with the best predicted test cross values were evaluated in the field. This can be regarded as indirect selection where the correlation ρ between the trait under selection and the trait to be improved is the correlation between the gene effects in the estimation set and the gene effects in the validation set (which could be called in this context more appropriately prediction set). The upper bound of this correlation is limited by a measure for the heritability, that takes into account not only the variance components of the field trial, but in addition the genetic change through recombination. We conclude that for assessing the accuracy of genome-based prediction with effects estimated in previous breeding cycles, cross validation within one cycle is not sufficient, but independent validation is required. Our results suggest that such predictions are only promising for traits with high heritabilities.

Acknowledgments We thank Gregory Mahone for proof reading the manuscript. We thank two anonymous reviewers for their helpful comments.

References

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

Bernardo R (2009) Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Sci 49:419–425

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

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

Daetwyler HD, Pong-Wong R, Villanueva B, Wooliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031

Gianola D, van Kaam JBCHM (2008) Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776

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

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

Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103:681–686

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

Piepho HP, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

Schneider K, Schäfer-Pregl R, Borchardt DC, Salamini F (2002) Mapping QTLs for sucrose content, yield and quality in a sugar beet population fingerprinted by EST-related markers. Theor Appl Genet 104:1107–1113

Searle SR (1987) Linear models for unbalanced data. Wiley, New York

Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York

Shepherd RK, Meuwissen THE, Woolliams JA (2010) Genomic selection and complex trait prediction using a fast EM algorithm applied to genomewide-markers. BMC Bioinformatics 11:529

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

Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163:789–801

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

Chapter 3

Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation [1]

[1] Hofheinz, N., and M. Frisch (2014) Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3: Genes | Genomes | Genetics* **4**: 539-546.

Heteroscedastic Ridge Regression Approaches for Genome-Wide Prediction With a Focus on Computational Efficiency and Accurate Effect Estimation

Nina Hofheinz and Matthias Frisch[1]
Institute of Agronomy and Plant Breeding II, Justus Liebig University, 35392 Giessen, Germany

ABSTRACT Ridge regression with heteroscedastic marker variances provides an alternative to Bayesian genome-wide prediction methods. Our objectives were to suggest new methods to determine marker-specific shrinkage factors for heteroscedastic ridge regression and to investigate their properties with respect to computational efficiency and accuracy of estimated effects. We analyzed published data sets of maize, wheat, and sugar beet as well as simulated data with the new methods. Ridge regression with shrinkage factors that were proportional to single-marker analysis of variance estimates of variance components (*i.e.*, RRWA) was the fastest method. It required computation times of less than 1 sec for medium-sized data sets, which have dimensions that are common in plant breeding. A modification of the expectation-maximization algorithm that yields heteroscedastic marker variances (*i.e.*, RMLV) resulted in the most accurate marker effect estimates. It outperformed the homoscedastic ridge regression approach for best linear unbiased prediction in particular for situations with high marker density and strong linkage disequilibrium along the chromosomes, a situation that occurs often in plant breeding populations. We conclude that the RRWA and RMLV approaches provide alternatives to the commonly used Bayesian methods, in particular for applications in which computational feasibility or accuracy of effect estimates are important, such as detection or functional analysis of genes or planning crosses.

KEYWORDS
genome-wide prediction
ridge regression
heteroscedastic marker variances
linkage disequilibrium
plant breeding populations
GenPred
Shared data resources

Best linear unbiased prediction (BLUP) and Bayesian approaches were suggested by Meuwissen *et al.* (2001) for predicting genotypic values with DNA markers. These genome-wide prediction (GWP) approaches have proven to be useful in plant breeding populations (*cf.* Crossa *et al.* 2010; Albrecht *et al.* 2011; Hofheinz *et al.* 2012). To overcome the problem of overparameterization triggered by more available marker data (p) than number of observations (n), shrinkage factors (ridge regression; BLUP) or variable selection (Bayesian approaches) can be used. Shrinkage factors can be constant for all markers or marker-specific with the use of homo- or heteroscedastic genetic variances.

Homoscedastic genetic variances at all markers in the linear model are regarded as a major shortcoming of the BLUP approach because many traits are assumed to be controlled by only a subset of the genes of an individual, not by all of them. This shortcoming motivated the development of Bayesian approaches that allow for heteroscedastic marker variances but at the expense of being computationally demanding (*cf.* Meuwissen *et al.* 2001, Shepherd *et al.* 2010, Kärkkäinen and Sillanpää 2012). To avoid the computational demands of Bayesian approaches, a linear model approach that uses heteroscedastic marker variances for data sets with more genotypes than markers was proposed by Piepho (2009). The generalized ridge regression (heteroscedastic effects model, or HEM) of Shen *et al.* (2013) also allows marker-specific shrinkage for overparameterized situations. These authors emphasized the need for computationally efficient GWP approaches with heteroscedastic marker variances.

The accuracy of the predicted genotypic values for GWP approaches with homoscedastic and heteroscedastic marker variances was compared, *e.g.*, for fruit traits in apple (Kumar *et al.* 2012),

doi: 10.1534/g3.113.010025
Manuscript received October 15, 2013; accepted for publication January 19, 2014; published Early Online January 21, 2014.

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.113.010025/-/DC1
[1] Corresponding author: Institute of Agronomy and Plant Breeding II, Justus Liebig University, 35392 Giessen, Germany. E-mail: matthias.frisch@uni-giessen.de

Fusarium head blight resistance in barley (Lorenz *et al.* 2012), 13 traits important in wheat breeding (Heffner *et al.* 2011), and for eight data sets in wheat, barley, *Arabidopsis*, and maize data sets (Heslot *et al.* 2012). The common conclusion was that in most instances the accuracy of predicting genotypic values was comparable for the investigated approaches. In particular, (1) none of the approaches was clearly superior under a broad range of applications; and (2) the BLUP approaches proved to provide good prediction accuracies, even for traits that are not supposed to follow closely the infinitesimal model of quantitative genetics, such as resistances. This finding was confirmed in a simulation study by Wimmer *et al.* (2013), who recommend the use of BLUP in plant breeding populations with large linkage disequilibrium (LD) extent, small sample sizes, and medium trait heritabilities.

The focus of the aforementioned studies was on the prediction of genotypic values of the individuals of a prediction set, and high prediction accuracies were observed when the individuals of the training and the prediction set were related (*cf.* Hofheinz *et al.* 2012). If training and prediction sets are a finite population of related individuals, then long chromosome stretches are expected to be in LD. In such populations, it is sufficient for a high prediction accuracy of genotypic values that the effects of chromosome stretches in LD are estimated with high accuracy. A high accuracy of estimating the effects of single markers is not necessary. Even if the estimated effects of single markers might be different for the different GWP approaches, the sum of the effects on a chromosome stretch in LD might be of similar size. This can be regarded as an explanation why different GWP approaches with homoscedastic and heteroscedastic variances result in a prediction of gentoypic values of similar accuracy.

The focus of this research lies on the accuracy of GWP approaches with respect to estimating the effects of single markers. This accuracy is important for the identification and functional analysis of genes, for the identification of target genes for marker-assisted gene introgression programs, and for the prediction of the performance of crosses. Predicting crosses, *i.e.*, estimating expectation and variance of the performance of a population derived from a cross of two parental genotypes, is an application of GWP in which plant breeders have high expectations, but no reports of successful implementations have been published. Predicting crosses builds on modeling the breaking up of existing LD and the recombination of favorable alleles originating from the two parents of a cross. Both the accurate localization of markers linked to the investigated trait and the accurate estimation of the effects via a GWP approach are of central importance for the success of such a prediction.

Our objectives were (1) to present novel heteroscedastic ridge regression approaches that improve existing approaches with respect to computational efficiency or accuracy of effect estimates and (2) to demonstrate their properties with computer simulations and with data sets of maize, wheat, and sugar beet.

METHODS

Linear model

For estimating the genetic effects of m biallelic single-nucleotide polymorphism (SNP) markers, a linear model, as follows, can be used:

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1}$$

$\mathbf{y}$ is the vector of N phenotypic values, β_0 is a fixed intercept, $\mathbf{Z}$ is the design matrix relating the marker data to genotypes, $\mathbf{u}$ is the vector of genetic effects, and $\mathbf{e}$ is the vector of residuals. The elements of $\mathbf{Z}$ are coded as linear regression on the number of one of the two alleles, *i.e.*, as 0,1,2. The genetic effects u_l (l = 1. . .m) and the residuals are normally distributed with $u_l \sim \mathrm{N}(0, \sigma_l^2)$ and $e_k \sim \mathrm{N}(0, \sigma_e^2)$ (k = 0. . .N). Furthermore, cov(u_i, u_j) = 0 ($i \neq j$) and cov(e_k, e_l) = 0 ($k \neq l$).

In ridge regression, the genetic effects u_l are predicted by solving the following mixed-model equations

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \boldsymbol{\Lambda}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}, \tag{2}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix that defines the amount of shrinkage. If its elements λ_l (l = 1. . .m) are defined as $\lambda_l = \sigma_e^2/\sigma_l^2$ and $\sigma_l^2 = \sigma_k^2$ for all l, $k \in \{1...m\}$, then the predictions u_l are the BLUPs (*cf.* Piepho 2009). This approach uses typically variance components σ_g^2 and σ_e^2 estimated from the data set under investigation.

In an approximative approach, preliminary rule of thumb estimates of the heritability h_p^2 can be used to define $\lambda_l = (1/h_\mathrm{p}^2 - 1)m$ (ridge regression employing preliminary estimates of the heritability (RIR), Hofheinz *et al.* 2012). In the following, we suggest approaches to determine marker-specific shrinkage parameters λ_l for ridge regression.

Shrinkage by single-marker variance component estimates

A moment estimator of the variance component for each marker can be obtained from a random single-factor analysis of variance (ANOVA) as follows:

$$\hat{\sigma}_l^{2*} = \frac{\mathrm{MQM}_l - \mathrm{MQE}_l}{\frac{1}{2}\left(N - \sum_i n_i^2/N\right)}. \tag{3}$$

MQM$_l$ and MQE$_l$ are the mean squares due to the marker and the error in the ANOVA for the l-th marker, N is the total number of individuals, and n_i (i = 1,2,3) are the numbers of individuals in the three marker classes.

The $\hat{\sigma}_l^{2*}$ are not independent and, therefore, they do not sum up to the genetic variance, which means that they cannot be used directly to determine the shrinkage factor. However, they can be used to partition the total genetic variance to the individual markers:

$$\hat{\sigma}_l^2 = \hat{\sigma}_g^2 \frac{\hat{\sigma}_l^{2*}}{\sum\limits_{l'=1}^{m} \hat{\sigma}_{l'}^{2*}}. \tag{4}$$

Here, the proportion of the genetic variance that is assigned to a marker l is proportional to the contribution of the single-marker ANOVA variance component of marker l to the sum of the single marker variance components of all markers. This results in shrinkage factors

$$\lambda_l = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_g^2} \frac{\sum_{l'} \hat{\sigma}_{l'}^{2*}}{\hat{\sigma}_l^{2*}}. \tag{5}$$

The approach used to determine the shrinkage factors in Equation 5 is abbreviated as RMLA (*i.e.*, estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components).

The estimation of the genetic and error variance components from the data set under consideration can be replaced by using preliminary estimates of the heritability h_p^2 as suggested by Hofheinz *et al.* (2012). This results in shrinkage factors

$$\lambda_l = \left(1/h_\mathrm{p}^2 - 1\right) m \, \frac{\sum_l \sigma_l^{2*}}{\sigma_l^{2*}}. \tag{6}$$

We abbreviate this procedure RRWA (*i.e.*, ridge regression with weighing factors according to ANOVA variance components).

Shrinkage by fixing the residual variance in variance component estimation

BLUPs of **u** in a linear model as defined by Equation 1 can be obtained with an iterative procedure on basis of the expectation-maximization algorithm (Searle *et al.* 1992) that consists of solving the mixed-model equations in Equation 2 for the parameter vector and then solving the following,

$$\begin{aligned} \hat{\sigma}_e^{\,2} &= \left(\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'\mathbf{Z}'\mathbf{y}\right)/(N-1) \\ \hat{\sigma}_l^2 &= \left(\hat{\mathbf{u}}_l'\hat{\mathbf{u}}_l - \hat{\sigma}_e^2 \mathrm{tr}\mathbf{C}_{ll}\right)/q_l \end{aligned} \tag{7}$$

for the variance components until convergence is reached (Misztal and Schaeffer 1986). Here q_l is the number columns of the design matrix **Z** that correspond to the variance component σ_l^2 and tr$\mathbf{C}_{ll}$ is the trace of the inverse of the coefficient matrix of Equation 2 that corresponds to the variance component.

If $\sigma_l^2 = \sigma_k^2$ (l, $k \in \{1 \ldots m\}$) is the constant variance of marker effects, $\mathbf{C}_{ll}$ is the complete coefficient matrix, and q_l the number of columns of **Z** (assuming full column rank), then the procedure can be used to obtain the variance components that yield the BLUPs.

A modification can be used to determine marker-specific shrinkage factors for ridge regression. First, $\hat{\sigma}_e^2$ is estimated as with BLUP. Then, the iterative procedure is repeated, but with two modifications: (1) The residual error σ_e^2 is not updated in each iteration round but instead the residual variance is held fixed for the value estimated in the first round. (2) For each marker, a different $\hat{\sigma}_l^2$ is estimated. This results in m values for $\hat{\sigma}_l^2$ and those are used to define the shrinkage factor for ridge regression as $\lambda_l = \hat{\sigma}_e^2/\hat{\sigma}_l^2$. We abbreviate this procedure RMLV (*i.e.*, modification of the restricted maximum likelihood procedure that yields heteroscedastic variances).

Software

We implemented the RIR, RMLA, RRWA, and RMLV approaches in our software SelectionTools (www.uni-giessen.de/population-genetics/downloads), which was also used for computer simulations. To perform reparametrized BLUP we used the R package rrBlupMethod6 (Piepho *et al.* 2012). The package BLR (Pérez *et al.* 2010) was applied for performing the Bayesian LASSO (BL). We used 1500 iterations and discarded the first 500 iterations as burn-in. The R package bigRR (Shen *et al.* 2013) was used for the HEM approach. A summary of all approaches used in the present study is given in Table 1. The code for all calculations is available in the Supporting Information, File S1, File S2, File S3, and File S4.

Experimental data sets

Three experimental data sets were used to investigate the prediction accuracy, size of effect estimates, and computing time of GWP approaches. The first data set consisted of 300 tropical maize lines from the International Maize and Wheat Improvement Center (CIMMYT), which were genotyped with 1148 SNP markers (Crossa *et al.* 2010). The traits grain yield (GY), female flowering, male flowering, and anthesis-silking interval were analyzed. Each trait was evaluated under severe drought stress and well-watered conditions.

The second data set consisted of 306 elite wheat lines from CIMMYT, which were genotyped with 1717 diversity array technology markers (Pérez-Rodríguez *et al.* 2012). The averages of all employed environments for the traits GY and days to heading were analyzed. The maize and the wheat data sets are available as an online supplement to the publications. The third data set consisted of 310 inbred lines from a commercial sugar beet breeding program, which were genotyped with 300 SNP markers (Hofheinz *et al.* 2012). The traits sugar content and molasses loss were analyzed. Genotypic and phenotypic data for both traits are available in the File S4.

To assess the accuracy of predicting genotypic values, we used repeated random subsampling to divide the data for cross validation. The first subset was used to estimate the marker effects and contained 80% of the data. The second subset contained 20% of the data and was used to validate the effects. The correlations between observed and predicted values were averaged over 100 cross validation runs.

Simulations

Computer simulations were used to investigate prediction accuracy of GWP approaches with respect to map position and effect size. To investigate the effect of high and low LD, we simulated random intermating of a large F1 population for either three or 19 generations (ngen = 3, 19). From the last intermating generation, 600 random doubled haploid lines were developed. We simulated 10 chromosomes, each of 1.6 M length, which were evenly covered with markers. To

Table 1 Summary of GWP approaches organized by the assumption of marker variances in the present study

Approach	Marker Variances		Reference/R Package
	Homoscedastic	Heteroscedastic	
BLUP	x		Meuwissen *et al.* (2001)
rrBlupM6	x		Piepho *et al.* (2012)
RIR	x		Hofheinz *et al.* (2012)
BL		x	Pérez *et al.* (2010)
HEM		x	Shen *et al.* (2013)
RMLA		x	New approach
RMLV		x	New approach
RRWA		x	New approach

GWP, genome-wide prediction; BLUP, best linear unbiased prediction; RIR, ridge regression employing preliminary estimates of the heritability; BL, Bayesian LASSO ; HEM, heteroscedastic effects model; RMLA, estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to analysis of variance components; RMLV, modification of the restricted maximum likelihood procedure that yields heteroscedastic variances; RRWA, ridge regression with weighing factors according to analysis of variance components.

Table 2 Computing time (sec) required for the estimation of marker effects with different GWP approaches

	Homoscedastic Marker Variances			Heteroscedastic Marker Variances				
	RIR	BLUP	rrBLUPM6	RMLV	RRWA	RMLA	BL	HEM
Simulated data, 500 individuals								
330 markers	0.03	0.16	0.91	5.07	0.05	0.16	5.14	39.92
810 markers	0.05	3.18	1.55	50.30	0.13	3.38	7.99	49.56
1610 markers	0.23	32.11	1.68	330.60	0.30	28.22	11.77	63.65
Crossa *et al.* (2010), 264 maize lines								
1135 SNP markers	0.10	9.08	0.37	118.20	0.14	9.17	11.10	8.79
Pérez-Rodríguez *et al.* (2012), 306 wheat lines								
1717 DArT markers	0.23	61.8	0.62	405.60	0.37	60.60	8.96	12.49
Hofheinz *et al.* (2012), 310 sugar beet lines								
300 SNP markers	0.01	0.12	0.35	3.72	0.04	0.11	5.51	3.69

For the maize data set, the trait GY-WW was investigated, for the wheat data set the trait GY, and for the sugar beet data set the trait SC. GWP, genome-wide prediction; RIR, ridge regression employing preliminary estimates of the heritability; BLUP, best linear unbiased prediction; RMLV, modification of the restricted maximum likelihood procedure that yields heteroscedastic variances; RRWA, ridge regression with weighing factors according to analysis of variance components; RMLA, estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to analysis of variance components; BL, Bayesian LASSO; HEM, heteroscedastic effects model; SNP, single-nucleotide polymorphism; DArT, diversity array technology; GY, grain yield; WW, well-watered; SC, sugar content.

investigate the effect of high, medium, and low marker density, we considered distances between two adjacent markers of 1 cM, 2 cM, or 5 cM (md = 1, 2, 5). Two genes affected the trait on each chromosome; they were 0.401 M and 1.201 M distant from the telomere. Each had a positive effect of 2.5 on the trait. Both favorable alleles originated from the same parental line of the F1 population. To obtain phenotypic values, for each of the 600 doubled haploid lines, a random normally distributed residual was added to the genotypic value. The residual effect was chosen such that the heritability of the trait was $h^2 = 0.5$ or $h^2 = 0.8$. Estimation of marker effects in the simulated data set was replicated 50 times for each GWP approach and the estimated effects sizes for each marker were averaged over the replications.

RESULTS

Computational efficiency

The computing time required to estimate marker effects with the simulated and experimental data sets was compared with a Linux workstation with 8 GB RAM and an Intel Core Quad 2.80 GHz processor. Among the approaches with homoscedastic marker variances, RIR was the fastest, and among those with heteroscedastic marker variances, RRWA was the fastest (Table 2). With both approaches, marker effect estimation took less than a second for all investigated data sets. RMLV was the slowest approach; in particular, for large data sets, the required computing time was considerable greater than that required for the other approaches.

Prediction accuracy of GWP approaches

For the approaches BLUP, RRWA, RMLA, BL, and HEM, the correlation between predicted and observed phenotypic values ranged between 0.31 for flowering time in the maize data set and 0.86 for molasses loss in the sugar beet data set (Table 3). The differences in prediction accuracy between the data sets were pronounced; however, a clear trend with respect to differences between the GWP approaches was not observable. Prediction accuracies were nearly identical for the approaches BLUP, RIR, and RRBlupM6; therefore, only the results for

Table 3 Correlation between observed and predicted phenotypic values determined with cross validation for different traits in the maize, wheat, and sugar beet data sets

Trait-Environment		Heteroscedastic Marker Variances				
	BLUP	RMLV	RRWA (h_p^2)	RMLA	BL	HEM
Crossa *et al.* (2010), 284 maize lines (264 lines, GY)						
MFL-WW	0.36	0.28	0.35 (0.8)	0.38	0.36	0.35
MFL-SS	0.45	0.28	0.38 (0.8)	0.39	0.45	0.44
FFL-WW	0.31	0.27	0.32 (0.8)	0.31	0.31	0.32
FFL-SS	0.51	0.35	0.46 (0.8)	0.47	0.48	0.50
ASI-WW	0.51	0.35	0.50 (0.8)	0.52	0.51	0.47
ASI-SS	0.51	0.35	0.44 (0.8)	0.46	0.50	0.45
GY-WW	0.54	0.36	0.46 (0.9)	0.50	0.54	0.52
GY-SS	0.43	0.19	0.34 (0.9)	0.37	0.43	0.35
Pérez-Rodríguez *et al.* (2012), 306 wheat lines						
GY-average	0.65	0.54	0.66 (0.8)	0.66	0.63	0.63
DTH-average	0.59	0.41	0.57 (0.9)	0.60	0.58	0.55
Hofheinz *et al.* (2012), 310 sugar beet lines						
SC	0.83	0.78	0.80 (0.9)	0.80	0.83	0.82
ML	0.85	0.82	0.84 (0.4)	0.86	0.86	0.85

For the RRWA approach, the preliminary heritability estimates h_p^2 are given in brackets. BLUP, best linear unbiased prediction; RMLV, modification of the restricted maximum likelihood procedure that yields heteroscedastic variances; RRWA, ridge regression with weighing factors according to analysis of variance components; RMLA, estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to analysis of variance components; BL, Bayesian LASSO; HEM, heteroscedastic effects model; GY, grain yield; MFL, male flowering; WW, well-watered; SS, severe drought stress; FFL, female flowering; ASI, anthesis-silking interval; DTH, days to heading; SC, sugar content; ML, molasses loss.

BLUP are presented. The RMLV approach showed considerable lower prediction accuracies than the other approaches, ranging from r = 0.19 to 0.82. Similar trends were observed with the simulated data (data not shown).

Size of effect estimates in the wheat data set

In the wheat data set for the trait GY, markers for which the effects estimated with the BLUP approach were high had even greater effects with the RMLA approach (Figure 1). With RMLV, the differences in size between small and large effects were even greater. Most marker effects were shrunken to zero, and only a subset of markers had remarkably high effect estimates. The approaches RRWA and RMLA estimated marker effects of identical effect sizes. HEM and RRWA estimated marker effects of comparable magnitude. Both shrank many marker effects toward zero and estimated greater effects for the remaining markers. However, the marker effects shrunken near zero were not the same for both approaches.

Simulation study on accuracy of marker effect estimates

For all combinations of marker distance (md = 1, 2, 5) and LD (ngen = 3, 19) the BLUP approach estimated the true marker effects with the least accuracy and the RMLV approach with the greatest accuracy (Figure 2). The BL, HEM, and RMLA approaches reached greater accuracies than the BLUP approach but still were outperformed considerably by the RMLV approach.

The accuracy of the BLUP approach was in particular low for the combination of small marker distances (md = 1) and high LD (ngen = 3). Here only RMLV resulted in usable effect estimates. With decreasing marker distances and decreasing LD the accuracy of the effect estimates obtained by the BLUP approach increased. However, the other approaches still provided effect estimates with considerable greater accuracy.

The greatest accuracy of effect estimates was achieved for large marker distances (md = 5) and low LD (ngen = 19), but still the BLUP showed a considerable underestimation of the true effects.

In addition to the simulations with a heritability of h^2 = 0.8 (Figure 2), we performed the same set of simulations with a heritability of h^2 = 0.5. The accuracy of effect estimates was lower but showed the same trends as with h^2 = 0.8 (File S5).

DISCUSSION

Heteroscedastic marker variances

For highly polygenic traits that follow closely the infinitesimal model of quantitative genetics, like yield, GWP approaches assuming homoscedastic marker variances are expected to be efficient for

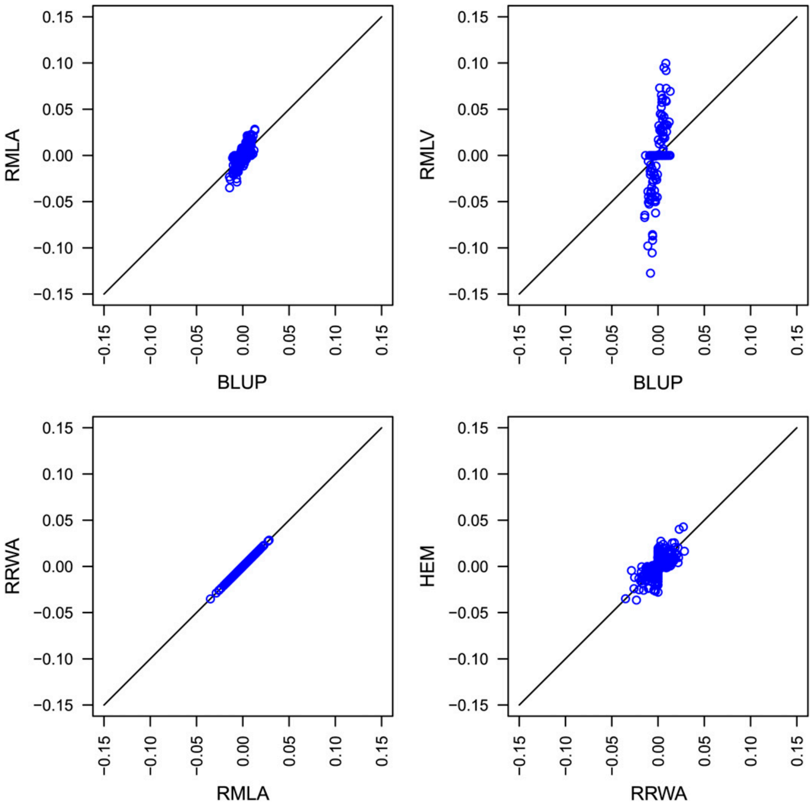

Figure 1 Comparison of the estimated marker effects for grain yield (GY) in the wheat data set for the best linear unbiased prediction (BLUP), ridge regression with weighing factors according to analysis of variance components (RRWA), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to analysis of variance components (RMLA), modification of the restricted maximum likelihood procedure that yields heteroscedastic variances (RMLV), and heteroscedastic effects model (HEM) approaches.

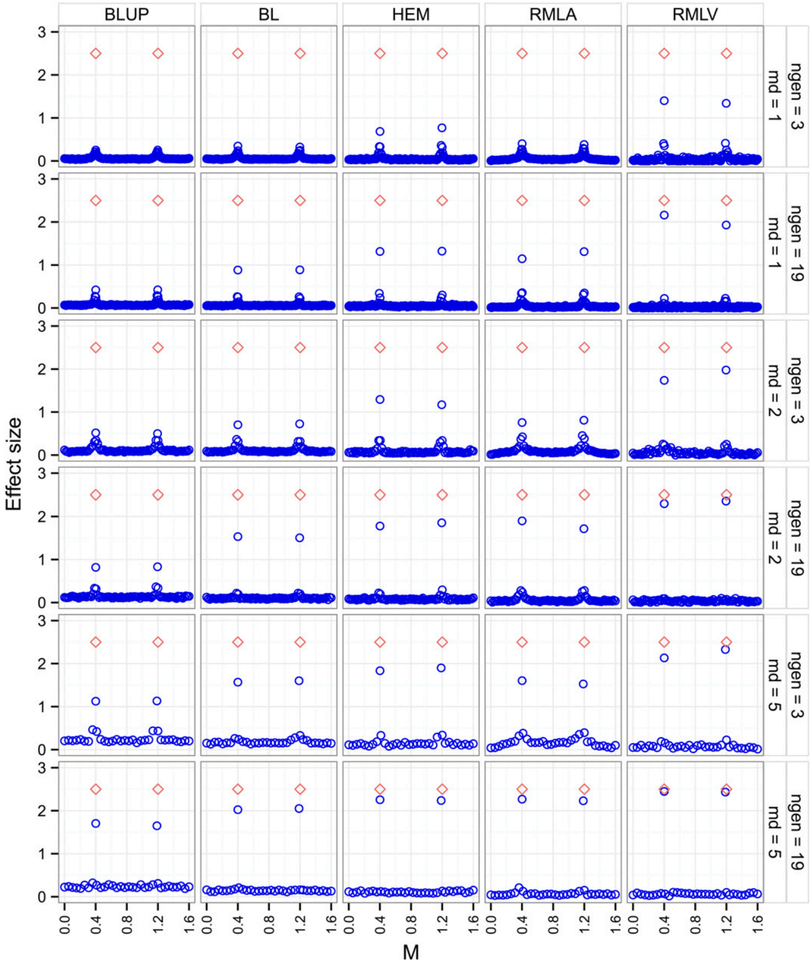

Figure 2 Marker effects (blue circles) estimated with different GWP approaches in the simulated data set plotted against marker locations [M] for the first chromosome. The positions of the simulated quantitative trait loci are symbolized by open red diamonds, ngen is the number of random intermating generations, and md is the marker distance [cM] of two adjacent markers.

predicting genotypic values. However, GWP approaches with heteroscedastic marker variances model better the genetic basis of traits when the number of markers is substantially greater than number of genes underlying the trait. This is the case for SNP maps with high marker densities or for traits that are controlled by only a few genes. Bayesian models were the first heteroscedastic GWP approaches. Their two main drawbacks are that choosing a suitable prior is required and that they are computationally very demanding. Dense marker maps have become state of the art and aggravate the problem of high computing times required for Bayesian approaches. Hence, fast and efficient heteroscedastic GWP approaches are necessary (*cf.* Shen *et al.* 2013).

Our RMLA approach, as well as the HEM approach of Shen *et al.* (2013), provides computational efficient alternatives to Bayesian approaches. The core of both approaches is to determine an individual shrinkage factor for each marker and then apply these shrinkage factors in ridge regression. The shrinkage factors for HEM are determined on basis of a BLUP estimate of the marker effects u_i, whereas RMLA uses a single-marker ANOVA. From a computational point of view, obtaining the BLUP estimates requires iterative procedures, whereas RMLA requires only the calculation of sums of squares. Consequently, determining shrinkage factors for RMLA is simpler and faster than for HEM. A second property that distinguishes RMLA from HEM is that the shrinkage factors for HEM are based on a first approximation, which uses homoscedastic marker variances; in contrast, the shrinkage factors for RMLA are based on a first approximation using heteroscedastic marker variances.

The computational efficiency of HEM was similar to that of RMLA for the data set of Crossa *et al.* (2010), but HEM was faster than RMLA for the data set of Pérez-Rodríguez *et al.* (2012) (Table 2). This advantage can be attributed to the optimized fitting algorithm of HEM, which makes its running time proportional to the number of individuals and not to the number of markers, as is the case for RMLA. Adopting a similar approach for RMLA might provide increased performance for dense marker maps. We chose to implement a different strategy for obtaining better performance. Approximating RMLA with RRWA uses preliminary estimates of the heritability instead of estimating the genetic and the residual variance from the data set under investigation. This results in a heteroscedastic ridge regression approach that does not need iterative procedures at all. RMLA and its approximation RRWA yielded the same effect estimates (Figure 1), and estimating the marker effects with RRWA took less than 1 sec for medium-sized data sets (Table 2). RRWA outperformed the other investigated approaches by factors between 10 and 100.

Prediction of genotypic values and size of estimated effects

The accuracy of predicting genotypic values was comparable for homo- and heteroscedastic genetic GWP approaches (*cf.* Heffner *et al.* 2011). Our results confirm that in general no advantage of a particular approach can be observed with respect to predicting genotypic values (Table 3). The size of effect estimates, however, was clearly different in the wheat data set of Pérez-Rodríguez *et al.* (2012) for the five investigated GWP approaches (Figure 1). The estimated effects for grain yield were greater for RMLA and HEM than for BLUP. RMLV resulted in the greatest effects and the most effects shrunken near zero. Hence, the similarity of GWP approaches with respect to predicting genotypic values is not caused by similar estimated marker effects.

We conclude that a high accuracy of estimated marker effects is not a prerequisite for high prediction accuracies of genotypic values, as long as marker alleles that were in positive LD in the estimation set are still to a large extent in positive LD in the individuals for which the genetic values were predicted. However, because there are considerable differences in the estimated marker effects between the GWP approaches, the choice of the GWP approach is expected to have an impact on the success of such applications of GWP that rely on the accuracy of estimates of single marker effects.

Importance of accurate effect estimates

Identification of known candidate genes in *Arabidopsis* (Shen *et al.* 2013) and apple (Kumar *et al.* 2012) was possible with effect estimates obtained by heteroscedastic GWP approaches. In contrast, no successful identification of genes was reported with results from homoscedastic BLUP estimates. This can be regarded as an indication that the greater effects obtained by the heteroscedastic approaches (Figure 1) are modeling the genetic basis of traits controlled by few genes better than homoscedastic BLUP and that accurate marker effect estimates are a prerequisite for the identification and fine mapping of functional genes.

An application of GWP that is most anticipated by plant breeders is planning crosses. In planning crosses, the probability distribution of the genotypic values of a population is investigated, which was derived from the cross of two parents with known phenotype and marker genotype. This distribution depends on the recombination between loci in the two parents of the cross, which breaks up the LD present in the parents. Here, it is not sufficient that the sum of genotypic values on a chromosome stretch in LD is correctly estimated. Instead, the effect of each single marker needs to be estimated with high accuracy. These two applications demonstrate that there is a need for GWP that provide accurate effect estimates for single markers.

With experimental data sets, the differences between GWP approaches with respect to effects sizes can be investigated (Figure 1), but it is not possible to evaluate which of the different effects at a marker is in fact the better estimate of the true (but unknown) effect. The importance of the two aforementioned applications and the fact that with experimental data the true effects are unknown motivated our simulation study.

Accuracy of effect estimates depending on the GWP approach

In breeding populations of crop species, the level of LD is typically high. Li *et al.* (2011) observed 20.6 cM for sugar type inbreds of sugar beet, and Stich *et al.* (2005) found an average LD length of 33 cM in European elite maize germplasm. The simulations with high marker density (md = 1) and high LD (ngen = 3) represent such a genetic situation (Figure 2). Here BLUP estimates of the genetic effects of traits controlled by two genes are underestimated. The underestimation is so severe that a useful application of the BLUP effect estimates seems unrealistic. Although there is still a considerable underestimation of RMLV in this scenario, this approach was the only that provided an effect estimate useful for applications like prediction of crosses and identification of functional genes.

In conclusion, our results confirm the results of previous studies that the BLUP can provide accurate predictions of genotypic values. However, for dense markers and strong LD, the effect estimates of BLUP are very imprecise. For applications of GWP that rely on accurate effect estimations, heteroscedastic approaches are superior. In particular, the RMLV approach is a promising approach for providing accurate GWP effect estimates.

LITERATURE CITED

Albrecht, T., V. Wimmer, H. J. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123: 339–350.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

Heffner, E. L., J.-L. Jannink, and M. E. Sorrells, 2011 Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome J. 4: 65–75.

Heslot, N., H.-P. Yang, M. E. Sorrells, and J. L. Jannink, 2012 Genomic selection in plant breeding: a comparison of models. Crop Sci. 52: 146–160.

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

Kärkkäinen, H. P., and M. J. Sillanpää, 2012 Back to basics for Bayesian model building in genomic selection. Genetics 191: 969–987.

Kumar, S., D. Chagné, M. C. A. M. Bink, R. K. Volz, C. Withworth *et al.*, 2012 Genomic selection for fruit trait quality in apple (*Malus* x *domestica* Borkh.). PLoS ONE 7: e36674.

Li, J., A. K. Lühmann, K. Weißleder, and B. Stich, 2011 Genome-wide distribution of genetic diversity and linkage disequilibrium in elite sugar beet germplasm. BMC Genomics 12: 484.

Lorenz, A. J., K. P. Smith, and J.-L. Jannink, 2012 Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. Crop Sci. 52: 1609–1621.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Misztal, I., and L. R. Schaeffer, 1986 Nonlinear model for describing convergence of iterative methods of variance component estimation. J. Dairy Sci. 69: 2209–2213.

Pérez, P., G. de los Campos, J. Crossa, and D. Gianola, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression Package in R. Plant Genome 3: 106–116.

Pérez-Rodríguez, P., D. Gianola, J. M. González-Camacho, J. Crossa, Y. Manés *et al.*, 2012 Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2: 1595–1605.

Piepho, H.-P., 2009 Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49: 1165–1176.

Piepho, H.-P., J. O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo *et al.*, 2012 Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Theor. Appl. Genet. 52: 1093–1104.

Searle, S. R., G. Casella, and C. E. McCulloch, 1992 *Variance Components*. Wiley, New York.

Shen, X., M. Alam, F. Fikse, and L. Rönnegård, 2013 A novel generalized ridge regression method for quantitative genetics. Genetics 193: 1255–1268.

Shepherd, R. K., T. H. E. Meuwissen, and J. A. Wooliams, 2010 Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. BMC Bioinformatics 11: 529.

Stich, B., A. E. Melchinger, M. Frisch, H. P. Maurer, M. Heckenberger *et al.*, 2005 Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor. Appl. Genet. 111: 723–730.

Wimmer, V., C. Lehermeier, T. Albrecht, H. J. Auinger, Y. Wang *et al.*, 2013 Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195: 573–587.

Communicating editor: D.-J. De Koning

Chapter 4

General discussion

Prediction of genotypic values

Cross validation and independent validation

The main focus of GWP applications in plant breeding has been on the prediction of genotypic values. In order to assess the prediction accuracies of GWP methods, many studies used cross validation methods on experimental data sets. Further, the correlations between observed and predicted performances were calculated. In a study with sugar beet inbred lines, we assessed correlations between observed and predicted test cross performances for 310 inbred lines (Hofheinz *et al.* 2012). The lines were genotyped with 384 SNPs and BLUP yielded in cross validation prediction accuracies of 0.86 and 0.82 for the low- and highly heritable traits molasses loss and sugar content, respectively. Similar high prediction accuracies between 0.72 and 0.80 were reported in a study employing a diversity set of 924 sugar beet inbred lines (Würschum *et al.* 2013). The inbred lines were genotyped with 677 SNPs and phenotyped for yield- and quality-related traits. Using mixed linear models, Albrecht *et al.* (2011) observed prediction accuracies between 0.72 and 0.74 for grain yield in a data set of 1,380 doubled haploid maize lines. They were genotyped with 1,152 SNPs. Such high prediction accuracies suggest that

GWP can be implemented in practical hybrid breeding programs for the prediction of test cross performance of lines from the same breeding cycle with only a moderate number of SNPs.

In the above-mentioned studies, average distances between two adjacent markers of 1 cM (Würschum *et al.* 2013), 3 cM (Hofheinz *et al.* 2012) and 2.9 Mb (Albrecht *et al.* 2011) resulted in prediction accuracies which were based on gametic disequilibrium between marker and QTL alleles. The reported high prediction accuracies were most likely caused by the homogeneity of the breeding material. Homogeneity results in identical linkage phases of marker and QTL alleles for large parts of the breeding material. Here, a division into estimation and prediction set for cross validation leads to a situation in which closely related lines with high performance and the same marker alleles at loci not influencing the trait under study are assigned to both sets. Hence, high effect estimates are assigned to these marker alleles even though the loci do not account for the trait of interest. Consequently, high prediction accuracies are the result of such a validation. We assessed a prediction accuracy of 0.86 for the trait molasses loss, which typically reaches heritabilities of $h_p^2 = 0.4$ in sugar beet breeding programs (Hofheinz *et al.* 2012). This result gives clear evidence for the phenomenon of overfitting due to relatedness in cross validations.

Accounting for relatedness between inbred lines, Albrecht *et al.* (2011) predicted maize test cross performance with respect to within and across family sampling. Here, the high degree of relatedness between estimation and validation set and high linkage disequilibrium (LD) between markers and QTL lead to higher prediction accuracies within families. Moreover, Würschum *et al.* (2013) investigated prediction within families when the marker effects were estimated in a diverse set of sugar beet lines. Prediction accuracies were consequently lower than those observed when the diversity set was used for both estimation and prediction. Despite those findings, a certain degree of relationship between individuals of estimation and prediction set is essential for the prediction of genotypic values. However, the degree of relation highly influences the amount of prediction accuracy. The proposed approach of Daetwyler *et al.* (2013) to account for relatedness between individuals of estimation and prediction set by adding a measure for relationship might improve the comparability of studies.

The prediction of genotypic values with marker effects estimated in a previous breeding cycle cannot be evaluated with cross validation. Instead, independent validation with data sets from two breeding cycles is required. Most recently, Albrecht *et al.* (2014) reported a small decrease in predictive ability with independent validation compared to cross validation in a study employing two highly heritable traits in maize. These results support our hypothesis that a transferability of effects estimated in one breeding cycle for predicting genotypes of a subsequent breeding cycle is only promising for highly heritable traits like sugar content in sugar beet (Hofheinz *et al.* 2012).

Similarity of prediction accuracies between GWP methods

The broad variety of GWP methods employing homo- and heteroscedastic marker variances achieved similar prediction accuracies in many studies employing experimental plant breeding data. Heslot *et al.* (2012) compared several GWP methods for their prediction accuracies in data sets of *Arabidopsis*, barley, maize and wheat and found no superior method. Four GWP methods showed similar prediction accuracies for 13 traits in a wheat data set (Heffner *et al.* 2011). Comparing GWP methods for traits with different genetic architectures in maize inbred lines, Riedelsheimer *et al.* (2012) found similar performance of all methods for both highly polygenic traits and metabolites with one known major QTL. In a study with data sets of maize, wheat and sugar beet, we compared six GWP methods and found no method to be advantageous (Hofheinz and Frisch 2014). Moreover, we demonstrated similar prediction accuracies reached by different GWP methods.

The main reason for small differences of prediction accuracies can be seen in the fact that long chromosome stretches are in LD in practical plant breeding populations. Observed average LD of sugar beet sugar type inbreds was 20.6 cM (Li *et al.* 2011) and European elite maize germplasm revealed an average LD of 33 cM (Stich *et al.* 2005). Average extents of intra-chromosomal LD in spring and winter wheat were 20.8 cM and 19.2 cM, respectively (Chao *et al.* 2010). Hence, tracing an individual QTL effect with marker-specific shrinkage as realized by heteroscedastic marker variances will reach prediction accuracies similar to those obtained by spreading the QTL effect along a LD stretch with homoscedastic marker variances. Wimmer *et al.* (2013) concluded that aside from a large LD extent, small sample sizes and medium trait heritabilities in applied plant breeding populations are reasons why GWP methods with heteroscedastic marker variances do not yield higher prediction accuracies. Besides, the authors showed that only the combination of large effective population sizes (N_e) and traits influenced by few genes can lead to a superiority of GWP methods employing heteroscedastic marker variances for the prediction of genotypic values. However, effective population sizes are typically small in plant breeding populations, *e.g.*, $N_e = 21.2$ in a study with sugar beet yield types (Li *et al.* 2011). In applied plant breeding populations, GWP methods employing homoscedastic marker variances such as BLUP are therefore highly recommendable for the prediction of genotypic values. Besides easy implementation, their main advantage lies in computationally efficient performance.

Computationally efficient ridge regression approaches

High-throughput marker systems strongly increase marker densities and require GWP methods, which are able to fit the growing parameters in computationally efficient ways. Rapid advances have recently been made in developing genome-wide dense molecular markers with a low per-sample cost using genotyping-by-sequencing (GBS). In wheat breeding, 34,749 SNPs have been identified with GBS and successfully been used for GWP employing BLUP (Poland *et al.* 2012). Up to 235,265 SNPs were identified with GBS in a maize data set and used in a cross validation (Crossa *et al.* 2013). To overcome the increasing computing time needed for marker effect estimation in such data sets, especially when cross validation runs are involved, computationally efficient GWP methods are required.

Approximating BLUP estimates of genetic effects

Since BLUP yields accurate predictions of genotypic values, efforts are being made towards improving its computational efficiency. We provided RIR as a ridge regression approach employing preliminary estimates of the heritability (Hofheinz *et al.* 2012). Variance components are not estimated directly from the data as with BLUP. Instead, a rapid approximation of genetic and residual variance components from results of preliminary estimates of the heritability is used (Table 4.1). These estimates are only rule-of-thumb values for the traits under study and are typically available in applied plant breeding programs. Therefore the application of the RIR is straightforward. Neither these preliminary estimates of the heritability nor the assumption that dividing an estimate of the genotypic variance by the total number of markers approximates the variance due to each marker are mathematically rigorous. However, RIR resulted in identical prediction accuracies as obtained with BLUP in the sugar beet data set (Hofheinz *et al.* 2012).

An alternative computation of BLUP termed “rrBlupMethod6” was developed to address the need for computationally efficient GWP methods (Piepho *et al.* 2012). The method allows for a fixed residual variance, as the authors emphasized that the residual variance is readily available in applied plant breeding programs from obtaining adjusted entry means of the phenotypes. Therefore, computational efficiency of rrBlupMethod6 is achieved by omitting the re-estimation of the residual variance. However, we showed that

RIR clearly outperformed conventional BLUP and was faster than rrBlupMethod6 (Hofheinz and Frisch 2014). RIR took 0.23 sec for the estimation of marker effects in the wheat data set of Pérez-Rodríguez *et al.* (2012), whereas rrBlupMethod6 and BLUP were slower with computing times of 0.62 sec and 61.8 sec, respectively. RIR took less than 0.3 seconds for the estimation of marker effects in simulated data and experimental data sets of maize, wheat, and sugar beet. Therefore, RIR proved to be a computationally efficient GWP method with high prediction accuracies.

Heteroscedastic ridge regression approaches

Plant breeders have high expectations for applications of GWP such as the identification of functional genes and prediction of the performance of crosses. Accurate marker effect estimates are a prerequisite for these applications and therefore homogeneous shrinkage of all marker effects will not suffice. Since the Bayesian methods require high computing times, computationally efficient alternatives to the Bayesian methods are required. We proposed RMLA, RRWA and RMLV (Table 4.1) as novel heteroscedastic ridge regression approaches, which allow marker-specific shrinkage by employing heteroscedastic marker variances (Hofheinz and Frisch 2014).

Table 4.1. Newly developed ridge regression methods in my thesis.

	Marker variances		
Method	Homoscedastic	Heteroscedastic	Calculation of the shrinkage factor λ
RIR	x		Approximation of BLUP using preliminary rule-of-thumb estimates of the heritability
RMLA		x	Estimation of single-marker variance components
RRWA		x	Approximation of RMLA using preliminary rule-of-thumb estimates of the heritability
RMLV		x	Fixation of the first residual variance component estimate

The RMLV approach is characterized by a modification of the iterative procedure of variance component estimation. Heteroscedastic marker variances are obtained by fixing the first estimate of the residual variance for further iterations. From a computational point of view, RMLV is a rather slow approach. However, the benefit of RMLV lies in most accurate effect estimation, which is a prerequisite for the above-mentioned applications. It remains to be investigated whether RMLV can be optimized in order to increase its computational efficiency.

Ridge regression with shrinkage factors that are proportional to random single-marker analysis of variance (ANOVA) estimates of variance components is employed with the methods RMLA and RRWA. For both methods, a moment estimator of the variance component for each marker is obtained by performing a random single-factor ANOVA. These estimates are used to partition the total genetic variance to each marker. The difference between both methods is that with RMLA the variance components are estimated directly from the data set under study. In contrast to RMLA, preliminary estimates of the heritability are used for a rapid approximation of variance components with RRWA, as is the case with RIR (Hofheinz *et al.* 2012).

Shen *et al.* (2013) proposed HEM as a computationally efficient generalized ridge regression method in which running time is proportional to the number of individuals. HEM is therefore especially efficient when the number of markers exceeds the number of individuals ($p > n$). This advantage of HEM became apparent in our study for the wheat data set (Hofheinz and Frisch 2014). For the remaining data sets, HEM showed computing times similar to those of RMLA. The main difference to RMLA is that HEM bases shrinkage factors for each marker on a first approximation of BLUP. Therefore, homoscedastic marker variances are involved in the first approximation with HEM. Another distinction to RMLA is that iterations are involved for obtaining the BLUPs of each marker with HEM, whereas RMLA requires only the calculation of sums of squares. RMLA and its approximation RRWA resulted in identical effect estimates in our study (Hofheinz and Frisch 2014). Moreover, RRWA took less than one second to estimate the marker effects and outperformed all investigated heteroscedastic approaches in experimental data sets of maize, wheat and sugar beet with respect to computational efficiency. It can be concluded that RRWA is the most rapid GWP method employing heteroscedastic marker variances.

Accuracy of single marker effect estimates

Accuracy of estimated effects for single markers is a neglected, but very important criterion for the evaluation of GWP methods. Several studies compared sizes of marker effect estimates between GWP methods in data sets of *Arabidopsis*, apple, barley and wheat (Shen *et al.* 2013; Kumar *et al.* 2012; Lorenz *et al.* 2012; Hofheinz and Frisch 2014). Here, BLUP shrank each estimated marker effect to the same extend by employing homoscedastic marker variances. The studies revealed that heteroscedastic GWP methods estimated large effects for a few markers while most marker effects were shrunk close to zero. Moreover, we found remarkable differences between GWP methods employing heteroscedastic marker variances (Hofheinz and Frisch 2014). The methods HEM, RMLA and RRWA estimated greater effects for markers that already had comparatively high effect estimates with BLUP. Greatest effect sizes were estimated for some markers with RMLV, while remaining markers were shrunk strongly. Based on the reported differences between GWP methods in estimated marker effect sizes, heteroscedastic GWP methods are expected to model the genetic architecture of traits controlled by only a few genes with large effects in a proper way. The comparisons showed that the earlier discussed similarity in prediction accuracies of the GWP methods is not induced by similar effect estimates. Accurate marker effect estimation has no benefit on the prediction accuracy when the same marker alleles are in positive LD in estimation and prediction set. However, accuracy of marker effect estimates has a major impact on further applications of GWP like the identification of functional genes and prediction of crosses.

In order to compare GWP methods for their accuracy of marker effect estimates in different genetic scenarios, we conducted a simulation study

(Hofheinz and Frisch 2014). A trait controlled by two genes on each of the ten chromosomes, for example a polygenic resistance, was simulated with known positions and effect sizes. Considering the presence of dense marker maps and high LD levels in applied plant breeding programs, the simulation study showed that BLUP greatly underestimates the genetic effects of the trait and its application seems not appropriate. Throughout the simulation study, RMLV proved to be most useful for GWP applications that require accurate marker effect estimates. It outperformed each of the GWP methods employing heteroscedastic marker variances, even though it also showed a considerable underestimation when high LD was present. In situations in which the distance of two adjacent markers was 5 cM and therefore in the range of the low LD level that is present after 19 generations of random mating, almost precise marker effect estimates were obtained with RMLV. These results indicate that with the presence of short-distance LD, high effects will only be estimated for markers close to the gene. However, effect estimates will be spread over those markers that are located in the same LD stretch as the gene, when longer LD stretches are present. Similar results were found by van den Berg *et al.* (2013), who implemented Bayes C(π) for QTL fine mapping in a simulation study. It can be concluded that RMLV provides most accurate single marker effect estimates in each genetic scenario.

Practical implementations of GWP in plant breeding programs

Prediction of selection candidates in hybrid breeding programs

Since my work mainly focused on sugar beet data sets, the following demonstrations are about possible implementations of my results in applied hybrid breeding programs. In hybrid breeding, the most promising parental inbred lines are typically identified by comparing their test cross performances, *i.e.* crossing them to carefully chosen testers to estimate their general combining ability. Test cross performances are evaluated in field trials, which are extremely resource intensive and limit the number of candidate lines in plant breeding programs. For resource efficiency, GWP can be implemented for the prediction of untested lines. Here, estimation and prediction sets belong to the same breeding cycle. More candidate lines are generated and each of them is genotyped, whereas only a part of them is evaluated in field trials. Test cross performances of lines which were not evaluated are then predicted by performing GWP and the lines with the highest predictions enter the second stage of line testing.

Another implementation of GWP in hybrid breeding programs is the prediction of highly heritable traits with an estimation set from the previous breeding cycle. More candidate lines are generated and they are all genotyped. Marker effects estimated in the previous breeding cycle are used for the prediction of test cross performances of all generated candidate lines. Hence, instead of evaluating each generated candidate line, only those candidate lines with the highest predicted test cross performances are evaluated in field trials. The usage of RIR can be recommended for the prediction of genotypes, because it provides high prediction accuracies and is computationally efficient.

Potential for genome-wide prediction of crosses

Having identified the most promising parental inbred lines, the next stage in a hybrid breeding program is the selection of parental combinations. Therefore, parental inbred lines are crossed and the progeny is evaluated for maximum expression of the desired characteristics. Plant breeders anticipate the prediction of the performance of crosses in order to support the critical step of identifying promising crosses. Cross prediction can be implemented in the breeding program by crossing only those parental inbred lines with best cross predictions and evaluating their progeny in field trials. Prediction of crosses has been a theoretical approach of implementing GWP until now and therefore little knowledge about its practical application in breeding programs is known. Typically, genome-wide marker data and phenotypic values of parental genotypes are available in breeding programs. This data can be used to predict expectation and variance of the performance of a population derived from crossing two parental genotypes. When planning crosses with GWP, the probability distribution of genotypic values in the derived population needs to be investigated. This distribution depends on the recombination between loci of the two parents. In a cross, recombination breaks up existing LD in the parents. Hence, summing up genotypic values for chromosome stretches, which works well for obtaining high prediction accuracies, is not sufficient for the prediction of crosses.

Accurate marker effect estimates are essential for the prediction of the performance of a cross. Iwata *et al.* (2013) recently predicted the simulated segregation patterns of two fruit-related traits in a progeny population of Japanese pear (*Pyrus pyrifolia*) in order to identify promising crosses. However, only the GWP methods BayesA and RR-BLUP were compared. As expected, RR-BLUP resulted in lower accuracies between observed and predicted segregation. We demonstrated in our simulation study that GWP methods differ substantially in their accuracy of marker effect estimation (Hofheinz and Frisch 2014). It remains open to further research whether the most accurate single marker effect estimates obtained with RMLV can be used for reliable cross predictions in applied plant breeding programs.

Identification of functional genes

The identification of functional genes is an application of GWP which has recently been investigated. The known gene *RPM1* of the monogenic trait AvrRpm1 was identified in a data set of *Arabidopsis* employing the heteroscedastic method HEM (Shen *et al.* 2013). In another study, LASSO and elastic net precisely identified major metabolite QTLs (Riedelsheimer *et al.* 2012). Homogeneous shrinkage of marker effects with RR-BLUP diluted the effects of the candidate gene and major QTLs in both studies. Therefore, BLUP estimates are not appropriate for the identification of genes. Identified candidate genes could be used for marker-assisted gene introgression in applied plant breeding programs. However, there is no evidence in the above-mentioned studies that the estimated marker effects represent the true genetic effects.

If experimental data sets are employed for GWP, the true genetic effects are not known. Therefore, simulated data sets with known positions and sizes of genes influencing the trait of interest are required for the evaluation of GWP methods concerning accuracy of effect estimates. In our simulation study, accuracy of effect estimates was higher when a high trait heritability was simulated (Hofheinz and Frisch 2014). Similar results were found by van den Berg *et al.* (2013), who conclude that high trait heritabilities and large data sets are a prerequisite for QTL mapping. RMLV can be recommended for the identification of genes, because it reached greatest marker effect accuracies in all combinations of LD and marker distances in our simulation study.

The presented results suggest that GWP is a promising tool with numerous applications in plant breeding programs such as the prediction of genotypic values, identification of functional genes or prediction of crosses. Our proposed ridge regression methods are an important contribution to the existing variety of GWP methods: Besides high prediction accuracies, the ridge regression methods in this thesis research have the potential to be computationally efficient and to estimate accurate marker effects.

Chapter 5

Summary

Genome-wide prediction (GWP) was suggested in order to overcome the shortcomings of quantitative trait loci mapping and marker-assisted selection. The latter failed to improve quantitative traits which are influenced by many genes with small effects, because only markers with significant effects are considered. GWP, in contrast, is based on molecular markers covering the whole genome. Genetic effects of markers are simultaneously estimated with a statistical GWP method in an estimation set consisting of genotyped and phenotyped individuals. Research has mainly focused on using the estimated genetic effects for the prediction of genotypic values for individuals in a prediction set which are only genotyped. However, plant breeders anticipate the usage of estimated genetic effects for the identification of functional genes for gene introgression or for the prediction of the performance of crosses. The objective of the present study was therefore the development of novel ridge regression methods that improve existing GWP methods with respect to accuracy of predicted genotypic values, accuracy of marker effect estimates and computational efficiency. For this purpose, their properties were compared in simulated data and data sets from applied plant breeding programs of maize, wheat and sugar beet.

The accuracy of predicted genotypic values is usually assessed using cross validation within related individuals of the same breeding cycle. The obtained accuracies are typically high. In the present study, prediction of test cross performance was investigated for the first time with an independent validation of a data set originating from two subsequent breeding cycles of an applied sugar beet breeding program. It was demonstrated that genetic effects which were estimated in a certain cycle of a breeding program can be used for prediction of genotypic values in the subsequent breeding cycle, if the trait under consideration has a high heritability, as for example sugar content.

In plant breeding populations linkage disequilibrium (LD) stretches cover substantial parts of a chromosome. Thus, accurate marker effect estimation most likely has no benefit on the accuracy of predicting genotypic values. Negligible differences of prediction accuracies were consequently observed between GWP methods employing homoscedastic and heteroscedastic marker variances. For the prediction of genotypic values, ridge regression employing preliminary estimates of the heritability (RIR) was the fastest GWP method among those employing homoscedastic marker variances.

The development of high-throughput marker systems facilitated the availability of low-cost, dense marker maps. Thus, computationally efficient ridge regression methods are required for GWP, especially when heteroscedastic marker variances are employed. Accurate estimation of the true genetic effects for each marker is an important criterion for heteroscedastic GWP methods, if they are used for the identification of functional genes for gene introgression or the prediction of the performance of crosses. A modification of the expectation-maximization algorithm that yields heteroscedastic marker variances (RMLV) and ridge regression with weighing factors according to analysis of variance components (RRWA) provide alternative solutions to the computationally demanding Bayesian methods. RRWA outperformed all of the investigated GWP methods employing heteroscedastic marker variances by factors between 10 and 100 in terms of computational efficiency. Most accurate marker effects in a simulated data set were estimated using RMLV, especially in situations with long LD stretches along the chromosomes and high marker densities, which often occur in plant breeding programs.

It can be concluded that the proposed novel ridge regression methods are promising for providing accurate predictions of genotypic values, accurate marker effect estimates and computational efficiency as was shown in a simulation study and data sets of applied breeding programs of maize, wheat and sugar beet.

Chapter 6

Zusammenfassung

Genomweite Vorhersage wurde entwickelt, um die Mängel der QTL-Kartierung und der markergestützten Selektion zu überwinden. Letztere verzeichnete keinen Erfolg im Hinblick auf die Verbesserung quantitativer Merkmale, da diese von vielen Genen mit jeweils kleinen Effekten beeinflusst werden. Die Begründung dafür ist, dass ausschließlich Marker mit signifikanten Effekten berücksichtigt werden. Im Vergleich dazu basiert die genomweite Vorhersage auf molekularen Markern, die das gesamte Genom abdecken. Die genetischen Effekte dieser Marker werden mittels einer statistischen Methode gleichzeitig in einem Schätzdatensatz, welcher genotypisierte und phänotypisierte Individuen beinhaltet, geschätzt. Die Forschung konzentrierte sich hauptsächlich auf die Verwendung der geschätzten genetischen Effekte für die Vorhersage von genotypischen Werten der Individuen eines Vorhersagedatensatzes, welche nur genotypisiert sind. Pflanzenzüchter erhoffen sich jedoch vor allem die Nutzung der geschätzten genetischen Effekte für die Identifikation funktioneller Gene für Genintrogression oder für die Vorhersage von Kreuzungsleistungen. Das Ziel der vorliegenden Arbeit war demnach die Entwicklung neuartiger Ridge-Regressions Methoden, welche die bisherigen Methoden zur genomweiten Vorhersage in Bezug auf Genauigkeit der geschätzten genotypischen Werte, Genauigkeit der geschätzten Markereffekte und rechnerische Effizienz verbessern. Zu diesem Zweck wurden die Eigenschaften der Methoden verglichen, und zwar anhand eines simulierten Datensatzes und solchen aus angewandten Pflanzenzüchtungsprogrammen von Mais, Weizen und Zuckerrübe.

Die Genauigkeit der vorhergesagten genotypischen Werte wird gewöhnlich mit einer Kreuzvalidierung innerhalb verwandter Individuen des gleichen Zuchtzyklus bestimmt. Somit sind die erhaltenen Genauigkeiten üblicherweise hoch. In der vorliegenden Arbeit wurde die Vorhersage von Testkreuzleistungen zusätzlich mit einer unabhängigen Validierung eines Datensatzes, der aus zwei nachfolgenden Zuchtzyklen eines Zuckerrübenzuchtprogramms besteht, untersucht. Eine Übertragbarkeit der genetischen Effekte, die in einem Zuchtzyklus geschätzt wurden und zur Vorhersage der genotypischen Werte im nachfolgenden Zuchtzyklus verwendet werden, ist für hoch heritable Merkmale, wie z.B. Zuckergehalt, Erfolg versprechend.

In Pflanzenzüchtungspopulationen weisen lange Chromosomenabschnitte Gametenphasenungleichgewicht auf. Daher begünstigt eine präzise Schätzung der Markereffekte höchst wahrscheinlich nicht die Genauigkeit der Vorhersage von genotypischen Werten. Geringfügige Unterschiede in den Vorhersagegenauigkeiten wurden zwischen den Vorhersagemethoden mit homoskedastischen und heteroskedastischen Markervarianzen beobachtet. Für die Vorhersage von genotypischen Werten war die Ridge-Regression, welche vorläufige Schätzer der Heritabilität verwendet (RIR), die schnellste Methode unter den Methoden, die homoskedastische Markervarianzen annehmen.

Die Entwicklung von Hochdurchsatzmarkersystemen ermöglichte die Verfügbarkeit von kostengünstigen, dichten Markerkarten. Folglich sind rechnerisch effiziente Ridge-Regressionsmethoden für die genomweite Vorhersage nötig, insbesondere bei heteroskedastischen Markervarianzen. Präzise Schätzungen des wahren genetischen Effekts von jedem Marker sind ein wichtiges Kriterium für heteroskedastische genomweite Vorhersagemodelle, wenn sie für die Identifikation funktioneller Gene für Genintrogression oder für die Vorhersage von Kreuzungsleistungen verwendet werden. Eine Modifikation des Expectation-Maximization-Algorithmus, welche zu heteroskedastischen Markervarianzen führt (RMLV), und Ridge-Regression mit

Gewichtungsfaktoren in Abhängigkeit von der Varianzkomponentenanalyse (RRWA) sind Alternativen zu den Bayes'schen Methoden. Denn letztere zeichnen sich durch einen hohen Rechenaufwand aus. RRWA hat die übrigen untersuchten Methoden zur genomweiten Vorhersage mit heteroskedastischen Markervarianzen um Faktoren zwischen 10 und 100 in Bezug auf rechnerische Effizienz übertroffen. Die präzisesten Markereffekte in einem simulierten Datensatz wurden mit RMLV geschätzt, insbesondere wenn sich lange Abschnitte entlang der Chromosomen im Gametenphasenungleichgewicht befanden und die Markerdichte hoch war. Dies kommt sehr häufig in Pflanzenzüchtungsprogrammen vor.

Wie in Datensätzen aus angewandten Zuchtprogrammen von Mais, Weizen und Zuckerrübe gezeigt wurde, sind die vorgeschlagenen neuartigen Ridge-Regressions Methoden vielversprechend, sie erzielen präzise Vorhersagen von genotypischen Werten und präzisere Markereffektschätzer und sind zudem rechnerisch effizient.

Chapter 7

Literature

Albrecht, T., H.-J. Auinger, V. Wimmer, J.O. Ogutu, C. Knaak *et al.* (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theor. Appl. Genet.* **127**: 1375-1386.

Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak *et al.* (2011) Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.* **123**: 339-350.

Bernardo, R. (2009) Genomewide selection for rapid introgression of exotic germplasm in maize. *Crop Sci.* **49**: 419-425.

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

Biancardi, E., J.M. McGrath, L.W. Panella, R.T. Lewellen, and P. Stevanato (2010) Sugar Beet. In J.E. Bradshaw (ed.) Root and Tuber Crops. Handbook of Plant Breeding, Volume 7. Springer Science+Bussiness Media, LLC, New York, NY USA. p. 173-219.

Chao, S., J. Dubcovsky, J. Dvorak, M.-C. Luo, S.P. Baenziger *et al.* (2010) Population- and genome-specific patterns of linkage disequilibrium and

SNP variation in spring and winter wheat (*Triticum aestivum* L.) *BMC Genomics* **11**: 727.

Collard, B.C.Y., and D.J. Mackill (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Phil. Trans. R. Soc. B* **363**: 557-572.

Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J.M. Hickey *et al.* (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 (Bethesda)* **3**: 1903-1926.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño *et al.* (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**: 713-724.

Daetwyler, H.D., M.P.L. Calus, R. Pong-Wong, G. de los Campos, and J.M. Hickey (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* **193**: 347-365.

De Biaggi, M., P. Stevanato, D. Trebbi, M. Saccomani, and E. Biancardi (2010) Sugar beet resistance to rhizomania: State of the art and perspectives. *Sugar Tech* **12**(3-4): 238-242.

Dohm, J.C., A.E. Minoche, D. Holtgräwe, S. Capella-Gutiérrez, F. Zakrzewski *et al.* (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*) *Nature* **505**: 546-549.

Goddard, M.E., B.J. Hayes (2007) Genomic selection. *J. Anim. Breed. Genet.* **124**: 323-330.

Gupta, P.K., J.K. Roy, and M. Prasad (2001) Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr. Sci.* **80**: 524–535.

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome.* **4**: 65-75.

Heffner, E.L., A.J. Lorenz, J.-L. Jannink, and M.E. Sorrells (2010) Plant breeding with genomic selection: Gain per unit time and cost. *Crop Sci.* **50**: 1681-1690.

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink (2009) Genomic selection for crop improvement. *Crop Sci.* **49**: 1-12.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink (2012) Genomic selection in plant breeding: A comparison of models. *Crop Sci.* **52**: 146-160.

Hofheinz, N., and M. Frisch (2014) Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3 (Bethesda)* **4**: 539-546.

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

Iwata, H., T. Hayashi, S. Terakami, N. Takada, T. Saito *et al.* (2013) Genomic prediction of trait segregation in a progeny population: a case study of Japanese pear (*Pyrus pyrifolia*) *BMC Genetics* **14**: 81.

Kärkkäinen, H.P., and M.J. Sillanpää (2012) Back to bascis for bayesian model building in genomic selection. *Genetics* **191**: 969-987.

Kumar, S., D. Chagné, M.C.A.M. Bink, R.K. Volz, C. Withworth *et al.* (2012) Genomic selection for fruit quality traits in apple (*Malus* x *domestica* Borkh.). *PLoS ONE* **7**(5): e36674.

Lande, R., and R. Thompson (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**: 743-756.

Li, J., A.-K. Lühmann, K. Weißleder, and B. Stich (2011) Genome-wide distribution of genetic diversity and linkage disequilibrium in elite sugar beet germplasm. *BMC Genomics* **12**: 484.

Lorenz, A.J., K.P. Smith, and J.-L. Jannink (2012) Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. *Crop Sci.* **52**: 1609-1621.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.

Nakaya, A., and S.N. Isobe (2012) Will genomic selection be a practical method for plant breeding? *Ann. Bot.* **110**(6): 1303-1316.

Owen F.V. (1945) Cytoplasmically inherited male-sterility in sugar beets *Jour. Ag. Res.* **71**(10): 423-440.

Pérez-Rodríguez, P., D. Gianola, J.M. González-Camacho, J. Crossa, Y. Manès *et al.* (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 (Bethesda)* **2**(12): 1595-1605.

Piepho, H.P., J. O. Ogutu, T. Schulz-Streeck, B. Estaghvirou, A. Gordillo *et al.* (2012) Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. *Crop Sci.* **52**: 1093-1104.

Piepho, H.P. (2009) Ridge regression and extensions for genomewide selection in maize. *Crop Sci.* **49**: 1165-1176.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu *et al.* (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome* **5**: 103-113.

Resende, M.F.R. Jr., P. Muñoz, M.D.V. Resende, D.J. Garrick, R.L. Fernando *et al.* (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* **190**: 1503-1510.

Riedelsheimer, C., F. Technow, and A.E. Melchinger (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics* **13**: 452.

Shen, X., M. Alam, F. Fikse, and L. Rönnegård (2013) A novel generalized ridge regression method for quantitative genetics. *Genetics* **193**: 1255-1268.

Stich, B., A.E. Melchinger, M. Frisch, H.P. Maurer, M. Heckenberger *et al.* (2005) Linkage disequilibrium in european elite maize germplasm investigated with SSRs. *Theor. Appl. Genet.* **111**: 723-730.

Utz, H.F., and A.E. Melchinger (1994) Comparison of different approaches to interval mapping of quantitative trait loci. In J.W. van Ooijen, and J. Jansen (eds.) Biometrics in Plant Breeding: Applications of Molecular Markers. Wageningen, p. 195-204.

van den Berg, I., S. Fritz, and D. Boichard (2013) QTL fine mapping with Bayes C(π): a simulation study. *Genet. Sel. Evol.* **45**: 19.

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

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

Würschum, T., S. Abel, and Y. Zhao (2014) Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. *Plant Breeding* **133**: 45-51.

Würschum, T., J.C. Reif, T. Kraft, G. Janssen, and Y. Zhao (2013) Genomic selection in sugar beet breeding populations. *BMC Genetics* **14**: 85.

Xu, Y., and J.H. Crouch (2008) Marker-assisted selection in plant breeding: From publications to practice. *Crop Sci.* **48**: 391-407.

Xu, S. (2003) Theoretical basis of the beavis effect. *Genetics* **165**: 2259-2268.

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

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

Acknowledgments

I want to express my gratitude to my academic supervisor Prof. Dr. Matthias Frisch for his continuous support, many suggestions and his advise during my thesis work.

Many thanks to Prof. Dr. Dr. h.c. Wolfgang Friedt for being my second supervisor.

Sincere thanks to Dr. Dietrich Borchardt, Dr. Knuth Weißleder and their colleagues for the good collaboration, many meetings and helpful discussions.

I thank all my colleagues at the institute of biometry and population genetics at the Justus Liebig University for the nice working atmosphere. I very much appreciate the endless organisational assistance of Mrs. Renate Schmidt. Thanks to my office mate Eva Herzog for good times and surviving in Ghana, Siberia and the dessert. I want to thank Carola Zenke-Philippi and Gregory Mahone for proof-reading.

Finally, I want to thank my family, Julia Brennecke, Marie Moos and all other friends for endless encouragement and cheerful times.

Eidesstattliche Erklärung

Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe.

Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.

Gießen, 15. Oktober 2014

Nina Hofheinz