Technische Universität München

Lehrstuhl für Pflanzenzüchtung

Efficiency of statistical methods for genome-based prediction

Valentin Wimmer

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Hans-Rudolf Fries
Prüfer der Dissertation:	1.	Univ.-Prof. Dr. Chris-Carolin Schön
	2.	Univ.-Prof. Dr. Aurélien Tellier
	3.	Univ.-Prof. Dr. Henner Simianer Georg-August-Universität Göttingen

Die Dissertation wurde am 24.02.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 23.06.2014 angenommen.

Summary

Plant and animal breeding programs are currently being revolutionized by technological developments in genomic research. With dense genome-wide marker data the genotypic value of an individual can be predicted based on its DNA profile. Statistical methods are required to derive marker effects from a training population comprising genotyped and phenotyped individuals. The proper choice of predictor variables and tuning of prediction methods is challenging in the face of high-dimensional marker data. Regularized regression and Bayesian methods are powerful techniques that cope with overfitting problems in high-dimensional marker data. Here, the performance of different statistical methods was compared with both simulated and experimental data sets taken from plant breeding populations. In this treatise, important determinants of prediction performance, such as choice of statistical method, trait heritability, marker density, and genetic trait architecture, were identified and then their effects were quantified.

First, the relative efficiency of genome-based prediction compared to pedigree-based prediction was investigated in an advanced cycle breeding population of maize (*Zea mays* L.). Marker data were incorporated into the genomic best linear unbiased prediction (GBLUP) method through a genome-based similarity matrix, and then predictive ability was estimated using cross-validation schemes. Next, alternative methods were explored by fitting marker effects within models allowing for variable selection and marker-specific prior distributions. A sensitivity analysis was performed to assess the influence of the prior specification on posterior inference and prediction performance. Furthermore, computer simulations were designed to investigate the accuracy of marker effects under different scenarios varying for the complexity of the true genetic model and the determinedness level of the data set. Real marker data from rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), and *Arabidopsis thaliana* (L.) populations were incorporated into the simulation schemes to obtain a realistic assessment of the effect of different levels of linkage disequilibrium on the accuracy of marker effects.

Results revealed a gain in predictive ability when using genome-based prediction methods compared to pedigree-based prediction. However, the performance of different methods can be severely derogated if inappropriate hyperparameters in the prior distribution were chosen. Given that hyperparameters were tuned properly, most methods performed similarly well. For complex traits, all methods captured primarily information on genetic relatedness, which emerge as major source of prediction accuracy. Variable selection enhanced predictive ability compared to methods retaining all markers in the model, but only if the sample size was much larger than the number of causal mutations underlying trait expression, and if only weak linkage disequilibrium among markers was present. Otherwise, all methods delivered genome-wide marker effects of low accuracy and were useful in predicting genotypic values but not in describing the genetic architecture of complex traits.

Applying genome-based predictions in plant breeding is still in its infancy and user-friendly software was lacking even though it is required to move genome-based prediction from theory into practice. To fill this gap, an open-source R package was developed providing a comprehensive analysis pipeline based on a unified data object, covering several statistical methods presented in this thesis.

In the near future, whole-genome sequence data will be available for genome-based prediction, and new computational as well as methodological challenges will arise when analyzing these data. The potential of pre-screening approaches to reduce data dimensionality was shown, though existing methods do not achieve a benefit in practice yet compared to using all available data. Further methodological developments are required to maximize performance when predicting complex traits based on whole-genome sequence data.

Zusammenfassung

Tier- und Pflanzenzüchtungsprogramme werden derzeit durch den technologischen Fortschritt in der Genomanalyse revolutioniert. Genomweite Markerdaten ermöglichen es den genotypischen Wert eines Individuums basierend auf seiner DNA vorherzusagen. Mit Regressionsmodellen werden die Markereffekte anhand eines Trainingsdatensatzes mit genotypisierten und phänotypisierten Individuen geschätzt. Aufgrund der hochdimensionalen Daten stellt die richtige Wahl der Einflussgrößen, der Methode und ihre Kalibrierung eine besondere Herausforderung dar. Regularisierte Regressionsverfahren und Bayesianische Methoden bieten die Möglichkeit, eine Überanpassung der Daten zu vermeiden. In dieser Arbeit wurde die Effizienz verschiedener Methoden anhand von experimentellen und simulierten Datensätzen verglichen. In diesem Zusammenhang wurden wichtige Einflussgrößen auf die Vorhersagefähigkeit von quantitativen Merkmalen wie die Wahl der Methode, die Heritabilität, die Markerdichte, sowie die genetische Merkmalsarchitektur untersucht und deren Effekt quantifiziert.

Zunächst wurde die relative Effizienz der genomweiten Vorhersage gegenüber verwandtschaftsbasierter Vorhersage in einem Mais (*Zea mays* L.) Züchtungsprogramm untersucht. Die Markerdaten wurden über eine genomische Ähnlichkeitsmatrix mittels der Methode der genomischen besten linearen unverzerrten Prädiktion (GBLUP) eingebunden und deren Vorhersagefähigkeit mit Kreuzvalidierung geschätzt. Anschließend wurden alternative Methoden, die die Verwendung marker-spezifischer Priori-Verteilungen und Variablenselektion ermöglichen, untersucht. Eine Sensitivitätsanalyse wurde durchgeführt, um den Einfluss der Parameter der Priori-Verteilung auf die Posteriori-Inferenz und die Vorhersagen zu erforschen. Mit Hilfe von Computersimulationen wurde die Genauigkeit der geschätzten Markereffekte untersucht. Dabei wurden verschiedene Szenarien simuliert, welche sich in der Komplexität des simulierten genetischen Modells und in dem Verhältnis zwischen der Anzahl der Beobachtungen und der Anzahl der Marker unterscheiden. Experimentelle genetische Markerdaten von Reis (*Oryza sativa* L.), Weizen (*Triticum aestivum* L.) und *Arabidopsis thaliana* (L.) wurden in die Simulationsroutinen integriert, um eine realistische Bewertung des Einflusses von verschiedenen Strukturen des Gametenphasenungleichgewichts auf die Genauigkeit der geschätzten Markereffekte zu erhalten.

Die Ergebnisse verdeutlichen, dass ein Anstieg an Vorhersagegenauigkeit mit genom-

basierten Methoden gegenüber verwandtschaftsbasierten Methoden erwartet werden kann. Allerdings kann die Vorhersagefähigkeit der Modelle deutlich beeinträchtigt werden, wenn ungeeignete Parameter in der Priori-Verteilung gewählt werden. Andernfalls liefern verschiedene Methoden ähnliche Vorhersagen, da alle Methoden die genetische Verwandtschaft zur Vorhersage nutzen und diese sich als primäre Quelle der Vorhersagefähigkeit herausstellte. Wenn die Anzahl der kausalen Mutationen, welche einem Merkmal zugrunde liegen, deutlich kleiner als die Stichprobengröße war, und kein starkes Gametenphasenungleichgewicht zwischen den Markern vorlag, konnte Variablenselektion die Vorhersagefähigkeit gegenüber Modellen, welche auf allen Markern basieren, erhöhen. In allen anderen Szenarien waren die genomweiten Markereffekte von geringer Genauigkeit und konnten zwar zur Vorhersage, nicht aber zur Beschreibung der genetischen Architektur komplexer Merkmale genutzt werden.

Die Anwendung genomweiter Vorhersage von genotypischen Werten in der Pflanzenzüchtung hat gerade erst begonnen und obwohl nutzerfreundliche Anwendersoftware der Schlüssel ist, um diesen Ansatz aus der Theorie in die Praxis zu bringen, war bislang noch keine solche Software vorhanden. Aus diesem Grund wurde ein umfassendes Erweiterungspaket für die Statistiksoftware R entwickelt, welches eine Analysepipeline basierend auf einem einheitlichen Datenobjekt beinhaltet. Dies ermöglicht die einfache Anwendung verschiedener statistischer Methoden aus dieser Arbeit.

In naher Zukunft werden genomweite Sequenzdaten zur Vorhersage von genotypischen Werten zur Verfügung stehen. Die Analyse dieser Datensätze erfordert sowohl neue rechentechnische als auch methodische Lösungen. Mit Computersimulationen wurde das Potential von Ansätzen, die eine Vorauswahl von Markern für die Vorhersage vornehmen, gezeigt. Allerdings liefert derzeit keine der untersuchten Methoden in der praktischen Anwendung eine Verbesserung der Vorhersagegenauigkeit gegenüber Modellen basierend auf allen Daten. Weitere methodische Entwicklungen werden benötigt, um die Vorhersagegenauigkeit von komplexen Merkmalen mit Hilfe von genomweiten Sequenzdaten zu maximieren.

Contents

Page

List of Figures

List of Tables

List of Abbreviations

BLUP	best linear unbiased prediction
BL	Bayesian Lasso
BMA	Bayesian model averaging
BRR	Bayesian Ridge regression
CRAN	comprehensive R archive network
CV	cross-validation
DH	doubled haploid
ES	estimation set
GBLUP	genome-based best linear unbiased prediction
GWAS	genome-wide association study
LASSO	least absolute shrinkage and selection operator
LD	linkage disequilibrium
MCMC	Markov chain Monte Carlo
MSE	mean-squared error
PBLUP	pedigree-based best linear unbiased prediction
PMSE	prediction mean-squared error
QTL	quantitative trait locus / loci
REML	restricted maximum likelihood estimation
RR-BLUP	Ridge regression best linear unbiased prediction
SIS	sure independence screening
SNP	single nucleotide polymorphism

Publications being part of the thesis

- Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzonova M, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, **123**: 339-350 [doi: 10.1007/s00122-011-1587-7]
- Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, **28**: 2086-2087 [doi: 10.1093/bioinformatics/bts335]
- Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. *Statistical Applications in Genetics and Molecular Biology*, **12**: 375-391 [doi: 10.1515/sagmb-2012-0042]
- Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Schön CC (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, **195**: 573-587 [doi: 10.1534/genetics.113.150078]

1 Introduction

1.1 Background

Genetic progress in crop and livestock breeding is required, in order to comply with the increasing demands for food and agricultural products. A key contribution is ongoing selection by plant breeders conducting field trials to score populations of selection candidates for agriculturally important traits. Individuals with the best genetic constitution are selected to become the founders of the next breeding cycle. With homozygous inbred lines at hand, the phenotype of a given genotype can be measured in multiple environments. With repeated measurements, the genotypic and environmental components of phenotypic variation can be separated (Falconer and Mackay 1996; Lynch and Walsh 1998). Thus, an important question in quantitative genetics is the relative contribution of the genotype to trait variation. For a given population, the heritable proportion of trait variation is quantified by the trait *heritability*, defined as the ratio of genotypic to phenotypic variance (Falconer and Mackay 1996). Another important factor is the *genetic trait architecture*, *i.e.*, the number of quantitative trait loci (QTL) controlling trait expression and their effect sizes. Most traits of agronomic importance follow a continuous distribution and are complex, *i.e.*, they are controlled by a large number of QTL (Fisher 1918; Schön *et al.* 2004; Hayes *et al.* 2010).

Plant and animal breeding programs are currently being revolutionized by technological developments in genomic research. For many species, marker arrays have been developed which determine the genotype of an individual at tens or hundreds of thousands of loci across the whole genome. Most markers are not causal, in that they do not affect phenotypic expression directly. Instead, it is assumed that the trait of interest is influenced by QTL which are not necessarily included in the marker panel. However, given that marker density is sufficiently high, QTL are likely to be in close proximity with at least one genetic marker. Neighboring loci on the genome tend to be inherited together and thus *linkage disequilibrium* (LD) can occur. LD is defined as non-random association of allele combinations of two or multiple loci, which arises from a shared history of mutation and recombination in a population (Hill and Robertson 1968). Consequently, the effects of most QTL are expected to be tagged by markers, and the marker effects are used as proxies for the QTL effects.

The availability of dense marker data for many species has led to a paradigm change in the breeding process. Selection based on the phenotypic value can be replaced by selection based on the genotypic value, which is determined by the sum of QTL effects carried by an individual. This approach, which is known as *genomic selection* after the seminal paper by Meuwissen *et al.* (2001), is based on a two-stage approach. First, a training set of individuals is both genotyped with markers and phenotyped for a specific trait. A statistical method is then used to estimate genome-wide marker effects, which are exploited in a second stage to predict the genotypic value of unphenotyped individuals based on their DNA marker profile. This approach is attractive from a practical point of view because it can accelerate breeding cycles (because the genotypic value is available as soon as a DNA sample is available) and can help to avoid cost-intensive phenotyping of a large number of selection candidates through performance trials (Schaeffer 2006; Jannink *et al.* 2010). Thus, the phenotype will no longer be the exclusive selection criterion but will be used to estimate marker or QTL effects in the training set (Lorenz *et al.* 2011). Genomic selection has been successfully implemented in dairy cattle breeding and has replaced traditional pedigree-based selection (Hayes *et al.* 2009). Until recently, predicting genotypic values has not been routinely implemented in plant breeding, although the integration of genomic selection in crop breeding programs shows good potential (Heffner *et al.* 2009; Lorenz *et al.* 2011; Wallace *et al.* 2014). However, it is unknown whether approaches applied to dairy cattle breeding can be integrated directly into crop breeding programs (Jonas and de Koning 2013).

An integral part of genomic selection is the statistical method used to estimate marker effects in the training set and because this approach does not yet involve a selection decision, this research field is called *genome-based* or *genomic prediction* rather than genomic selection. The major objective in genome-based prediction—and of this thesis—is to identify methods which predict genotypic values as accurately as possible. In addition to the choice of method, other factors that determine the prediction performance are not well understood. In this thesis, different statistical methods were evaluated for different experimental and simulated data sets to assess and quantify the influence of the choice of method and to measure the influence of factors such as trait heritability, marker density, and genetic trait architecture. In the following, the problem of genome-based prediction is presented from a statistical point of view and major challenges in the face of high-dimensional marker data are discussed.

1.2 Statistical models in genomic analyses

Today, the most frequently used genetic markers are single nucleotide polymorphisms (SNPs), which are positions in the DNA sequence where individuals differ with respect to the nucleotide (A, C, G, or T) they carry. Diploid individuals carry two alleles and so exhibit one out of three possible genotypes at each locus (*e.g.*, AA, AT, or TT). The coding of the SNP marker data depends on the design of the study and the model of inheritance. Here, additive effects are modeled and, thus, the genotype of an individual at each SNP is coded by the number of copies of a reference allele it carries, *i.e.*, 0, 1, or 2 for genotypes AA, AT, and TT if T is the reference allele. Genome-based prediction exploits a statistical model linking phenotypic variation to genetic variation at the DNA level. In this thesis, the phenotypic value y_i of an individual $i = 1, \dots, n$ is supposed to follow a Gaussian distribution, where the expectation is its genotypic value g_i. Thus,

$$y_i | g_i, \sigma^2 \sim \mathrm{N}(g_i, \sigma^2), \tag{1}$$

where the residual variance component σ^2 is due to non-genetic factors. Each individual is genotyped with p SNP markers and x_{ij} encodes the marker genotype of individual i at marker locus $j = 1, \dots, p$. All SNP markers can be used simultaneously as potential predictor variables in a regression model. In matrix notation, all phenotypic records are stacked in the n-dimensional vector $\mathbf{y} = (y_1, \dots, y_n)'$ and all genotypic data in the $n \times p$ dimensional matrix $\mathbf{X}$ of predictor variables, leading to the regression model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i \quad \text{for } i = 1, \dots, n \quad \text{or, in matrix notation,} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_j, \dots, \beta_p)'$ is the p-dimensional vector of marker effects to be estimated and $\mathbf{e} = (e_1, \dots, e_i, \dots, e_n)'$ is the n-dimensional vector of residual terms, with

$$\mathbf{e} | \sigma^2 \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_n \sigma^2), \tag{3}$$

where $\mathbf{I}_n$ is the n-dimensional identity matrix. In model (2), the response variable $\mathbf{y}$ is mean centered, in order to omit an intercept term without loss of generality. Suppose that the true model comprises a subset $p_0 \leq p$ of important predictor variables for the trait under study (*e.g.*, because they are causal mutations or tag QTL). Thus, the vector of true regression effects denoted as $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0p})'$ is expected to be *sparse*; *i.e.*, only a subset of predictor variables has a true nonzero regression coefficient. All other elements in $\boldsymbol{\beta}_0$ are true zero coefficients that can generate noise when predicting trait expression. Sparsity of the true model is a central concept when studying the behavior of different statistical methods, as illustrated in Wimmer *et al.* (2013) and Section 2.2.

Two major goals can be distinguished when searching for an appropriate model for the regression problem described previously: *variable screening* and *prediction*. For variable screening, variable selection can be used to pinpoint the positions of the true nonzero coefficients in $\boldsymbol{\beta}_0$. With genomic data, only those markers should be retained in the model that are considered to be causal mutations themselves or tag QTL for the trait under study. Thus, variable selection methods have the potential to describe the genetic architecture of complex traits, *i.e.*, the number of QTL and their effects. In Wimmer *et al.* (2013), as well as in Section 2.2, empirical results on the accuracy of estimated marker effects were presented and prospects and limitations of statistical methods for variable screening with high-dimensional marker data were discussed. For practical applications in breeding programs, prediction is of the utmost importance. In this treatise, it is crucial to obtain estimated regression coefficients $\hat{\boldsymbol{\beta}}$ that deliver accurate predictions of the genotypic value for a future observation using $\hat{g}_f = \mathrm{E}(y_f|\mathbf{x}_f) = \mathbf{x}'_f\hat{\boldsymbol{\beta}}$, where $\mathbf{x}_f$ is the p-dimensional vector of marker genotypes of the selection candidate.

1.3 Challenges in statistical modeling

With marker data at hand, the phenotypic value of an individual can be regressed on the marker loci. The standard technique employed to obtain regression coefficients in linear models is least-squares estimation. However, this method is limited to the case where more individuals than markers are available and extensions are required when analyzing high-dimensional marker data, as discussed in the following.

The theoretical properties of regression models are well-described in the framework of problems with a small, fixed number p of well-chosen predictor variables and large numbers of observations n. Nowadays, the number of markers is continuing to increase while the number of individuals with phenotypic observations remains limited because phenotyping is the bottleneck especially in plant breeding programs. Thus, high-dimensional data sets emerge whereby $n \ll p$, *i.e.*, the number of predictor variables exceeds the number of observations by far, known as *large p, small n problem.* From a statistical point of view, the regression problem in (2) is underdetermined and a simultaneous fit of all markers using least-squares estimation is not possible.

The concept of genome-based prediction relies on LD among markers and QTL (Meuwissen *et al.* 2001). A sufficient coverage of the genome with markers is required to maximize the probability that at least one marker is in close proximity with a QTL. On the other hand, in dense marker maps, LD among markers leads to multicollinearity in the matrix of the predictor variables. Multicollinearity describes a situation whereby multiple linear dependencies are present among predictor variables (Myers 1994). Under this scenario, spurious correlations of unimportant predictor variables and the response variable can emerge, leading to a severe bias for estimated regression coefficients, wrong standard errors, and even misleading scientific conclusions (Miller 2002; Hastie *et al.* 2009).

1.4 Addressing dimensionality through regularization

Methods to cope with the large p, small n problem in high-dimensional data sets are of paramount interest for genome-based prediction. Promising approaches are based on statistical methods applying *regularization* through constraints in the objective function, by variable selection, or by introducing a prior distribution for the unknown parameters in a Bayesian framework. First, it is described how regularization influences bias and variance of the estimated marker effects and why regularization in regression models can enhance prediction performance.

1.4.1 The bias-variance tradeoff in regression analyses

By using regularization, an unbiased estimator for a marker effect such as the least-squares estimator is turned into a biased estimator with smaller variance. Suppose $\hat{\theta}_n$ is an unbiased estimator of the true parameter θ obtained from n observations. If this estimator is multiplied by a regularization parameter $a \in [0, 1)$, the estimator $\hat{\theta}^*_n$ for θ is biased because it is *shrunken* such that $\mathrm{E}(\hat{\theta}^*_n) = a\theta < \theta$. However, $\mathrm{Var}(\hat{\theta}^*_n) = a^2 \,\mathrm{Var}(\hat{\theta}_n) < \mathrm{Var}(\hat{\theta}_n)$ and, thus, the variance is reduced at the expense of the estimation bias (de los Campos *et al.* 2013a). To measure the precision and accuracy of an estimator, the mean-squared error (MSE) is suitable as it comprises both the squared bias and the variance of the estimator (Hastie *et al.* 2009):

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}_n) &= \mathrm{E}\left\{\left[\hat{\theta}_n - \theta\right]^2\right\} \\
&= \mathrm{E}\left\{\left[\hat{\theta}_n - \mathrm{E}(\hat{\theta}_n)\right]^2\right\} + \left[\mathrm{E}(\hat{\theta}_n) - \theta\right]^2 \\
&= \mathrm{Var}(\hat{\theta}_n) + \left[\mathrm{Bias}(\hat{\theta}_n)\right]^2 . \qquad (4)
\end{aligned}
$$

The MSE can be minimized by an appropriate choice of a, leading to a bias-variance tradeoff. This concept is fundamental to illustrate how regularization can enhance the performance of genome-based prediction methods. The extent of regularization in a regression model will be measured by the *number of effective parameters* defined by the trace of the *hat matrix* that projects the vector of observed values to the vector of fitted values following Tibshirani (1996) and Gianola (2013). A method that fits the training data extremely accurately has a hat matrix that is close to the diagonal matrix, and, thus, a large number of effective parameters. Such a method will exhibit a small bias but has potentially a large prediction variance. Hence, the ability to predict independent test data is reduced because the method exaggerates minor fluctuations in the data and suffers from overfitting. On the other hand, with too much regularization, the method experiences a lack of fit, leading to increased errors in both training and test data (*i.e.*, the method suffers from underfitting).

For genomic prediction based on dense marker data the crucial task is to optimize this bias-variance tradeoff. The “best” estimate with respect to the MSE typically trades a small bias for a large reduction in prediction variance (Figure 1). In many cases, the variance term will decrease with sample size n, while the bias is independent of n. Thus, larger samples tend to support a higher number of effective parameters because the relative disadvantage of a large variance is reduced (Miller 2002). Furthermore, a different extent of regularization can be advantageous if the major goal is variable screening instead of prediction. This question will be investigated in Section 2.3.1. When considering the question of an appropriate method for genome-based prediction, the concept shows why different methods exhibit different properties and can deliver different prediction performances. Thus, different statistical methods were investigated in this thesis and their efficiency was assessed through computer simulations and experimental data sets. In particular the choice of regularization parameters is crucial when optimizing the bias-variance tradeoff, as illustrated in Section 2.3. Next, the specific statistical methods employed in this thesis will be described.

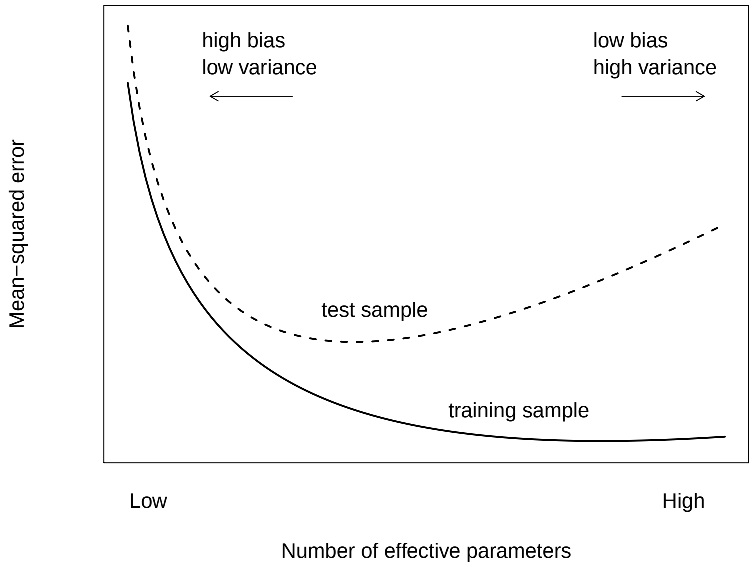

Figure 1: Typical behavior of the mean-squared error in a training sample and a test sample (dashed) as a function of the number of effective parameters (figure adopted from Hastie *et al.* (2009, p.38)).

1.4.2 Penalized least-squares estimates

In penalized regression models, a penalty function augments the objective function of least-squares estimation in order to constrain the size of the regression coefficients and to induce regularization. In this thesis, three approaches based on penalized least-squares estimators were employed where the estimated marker effects for (2) are obtained as

$$\hat{\boldsymbol{\beta}}(\lambda) = \text{argmin}_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \cdot \text{Pen}(\boldsymbol{\beta}) \right\} \tag{5}$$

where $\mathbf{y}$ denotes the n-dimensional vector of phenotypic observations, $\mathbf{X}$ the $n \times p$ matrix of marker genotypes, and $\boldsymbol{\beta}$ the p-dimensional vector of marker effects. In (5) $||\mathbf{x}||_2^2$ denotes the squared L_2 norm of a vector $\mathbf{x} = (x_1, ..., x_p)'$, defined as $||\mathbf{x}||_2^2 = \sum_{j=1}^{p} x_j^2$. The choice of penalty function $\text{Pen}(\boldsymbol{\beta})$ is crucial for the properties of the resulting estimator and it was demonstrated empirically that different penalty functions were preferable for different scenarios (Wimmer *et al.* 2013). The regularization parameter $\lambda \geq 0$ is the multiplier of the penalty function that controls the extent of regularization while notation $\hat{\boldsymbol{\beta}}(\lambda)$ is used to stress the dependency of the solution on the choice of λ. The larger the value of λ, the smaller the number of effective parameters. In Section 2.3.1 the influence of the choice of λ on the properties of the resulting estimator $\hat{\boldsymbol{\beta}}(\lambda)$ will be investigated with marker data.

One statistical method used for genomic prediction in this thesis is *Ridge regression* (Hoerl and Kennard 1970) where the penalty function is defined by the L_2 norm of the regression coefficients, *i.e.*, $\text{Pen}(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2^2$. The estimated marker effects are obtained as

$$\hat{\boldsymbol{\beta}}(\lambda)^{\text{Ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}, \tag{6}$$

where $\mathbf{I}_p$ is the p-dimensional identity matrix. Thus, the value λ is added to the diagonal values of $\mathbf{X}'\mathbf{X}$, which assures a unique inverse and stabilizes the solution when $\mathbf{X}'\mathbf{X}$ is ill-conditioned due to multicollinearity (Hoerl and Kennard 1970). The bias of the estimator in (6) increases but the variance decreases monotonically with λ and the MSE can be improved compared to the least-squares estimator (Hoerl and Kennard 1970). In this thesis, Ridge regression was implemented according to the best linear unbiased prediction (BLUP) approach, where the value for λ was derived from the noise-to-signal ratio in the data, *i.e.*, the ratio of environmental to marker variance components, as estimated through restricted maximum likelihood estimation (REML) (see Wimmer *et al.* (2013) for more details). This method is known as “Ridge regression BLUP” (RR-BLUP) in the context of genome-based prediction. It is worth mentioning that RR-BLUP performs no variable selection and retains all markers in the model.

From a computational perspective in the $n \ll p$ case, it is more convenient to replace the RR-BLUP method with the genomic BLUP (GBLUP) method where marker data are used to construct a genome-based similarity matrix among individuals (Habier *et al.* 2007). Consequently, the GBLUP method requires the inverse of a $n \times n$ instead of a $p \times p$ matrix as in RR-BLUP (see Equation 6), but both methods deliver the same predicted genotypic values (Goddard *et al.* 2009; Hayes *et al.* 2009) and the estimated variance components are transferable between models (Albrecht *et al.* 2011). The GBLUP method is closely related to the pedigree-based BLUP (PBLUP) method that has been used for livestock improvement for decades. In PBLUP, the expected relationship between relatives based on pedigree is exploited to predict genotypic values. A pedigree-based relationship matrix is constructed to account for the expected covariance among the genotypic values of the individuals being random effects in a linear mixed model (Henderson 1984).

A special case of regularization is *variable selection*, where predictor variables are effectively removed from the model by setting their regression coefficients to zero. Variable selection is of great interest in genetic analyses when it is likely that only a fraction of markers will be important as predictor variables, *e.g.*, because they tag QTL effects (Meuwissen *et al.* 2001). With the L_1 norm penalty function $\text{Pen}(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$ the estimator in (5) is called the *least absolute shrinkage and selection operator* (LASSO, Tibshirani 1996). This penalty function is continuous, but the first derivative (*i.e.*, the signum function) is not continuous in zero. As a consequence, the LASSO features variable selection such that $\hat{\beta}_j(\lambda) = 0$ for many predictor variables and LASSO can select at most n nonzero coefficients for all values of λ (Hastie *et al.* 2009). In this thesis, suitable values for λ were determined using cross-validation (CV) as described in Section 1.5 and Wimmer *et al.* (2013). The resulting sparse model can be used to pinpoint QTL. However, it is challenging to distinguish these from noise in the face of multicollinearity caused by LD among markers. In Wimmer *et al.* (2013), scenarios where variable selection was successful were identified when analyzing genome-based prediction data.

Several LASSO and Ridge regression extensions have been proposed in the literature. Zou and Hastie (2005) introduced the *elastic net* for the analysis of high-dimensional data sets with correlated predictor variables. The penalty function for the elastic net is defined as a compromise between the L_1 and L_2 norm penalty functions, *i.e.*, $\text{Pen}(\boldsymbol{\beta}) = \alpha||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||_2^2$ (Friedman *et al.* 2010). Both LASSO and Ridge regression can be considered as special cases of the elastic net with $\alpha = 1$ and $\alpha = 0$, respectively. However, it is not straightforward to see theoretically when the elastic net outperforms LASSO (Bühlmann and Mandozzi 2013). Thus, the performance of both methods was compared empirically in Wimmer *et al.* (2013).

1.4.3 Bayesian methods for genome-based prediction

Besides the choice of penalty function, the choice of appropriate regularization parameters affects the performance of penalized regression methods. The regularization parameter λ in (5) can be defined by CV or the BLUP approach. An alternative view on regularization is the Bayesian framework, where the regularization parameter is considered to be an additional random variable in the model. This offers the possibility to estimate the regularization parameter along with the regression coefficients from the data. Here, regression coefficients are assumed to be unknown while a predefined prior distribution conveys regularization. Inferences are based on the posterior distribution which is obtained from the prior distribution that is updated with the likelihood function of the data. Thus, the prior distribution will affect posterior inference, but the influence is expected to vanish for large sample sizes (Bernardo and Smith 1994).

The prior distribution conveys information about the unknown parameters as the penalty function introduces constraints into the optimization problem described in (5). In the regression framework of (2), the prior distributions may differ between the regression coefficients and typically involve a second level in the model hierarchy with hyperprior distributions based on hyperparameters that must be specified in advance. These hyperparameters control the shape of the prior distribution, and, hence, the bias-variance tradeoff. An important question is the extent to which the prior distribution influences posterior inference (Gianola 2013). In this context, *sensitivity analysis* refers to the process of modifying the prior specification in order to investigate its impact on posterior inference (Gelman *et al.* 2004). In Lehermeier *et al.* (2013), a sensitivity analysis was performed to assess the influence of hyperparameters in the prior distribution for different Bayesian genome-based prediction methods. The specific Bayesian methods applied for genome-based prediction in this thesis will be described briefly in the following while additional information can be found in de los Campos *et al.* (2013a).

In Bayesian Ridge regression, the same prior distribution is employed for all marker effects such that

$$\beta_j|\sigma^2_\beta \sim \mathrm{N}(0, \sigma^2_\beta) \text{ for } j = 1, \ldots, p, \tag{7}$$

where the variance parameter σ^2_β is considered as a random variable, and a scaled inverse-χ^2 distribution is assigned as the hyperprior distribution, *i.e.*,

$$\sigma^2_\beta|\nu, S \sim \chi^{-2}(\nu, S). \tag{8}$$

The scaled inverse-χ^2 distribution has two hyperparameters, the scale parameter $S > 0$ and $\nu > 0$ degrees of freedom. A large value for S in (8) will lead to a large value for $\mathrm{E}(\sigma^2_\beta) = \nu S^2/(\nu - 2)$ (for $\nu > 2$) and thus a diffuse prior distribution for β_j in (7), indicating a small extent of regularization. The common variance component in Bayesian Ridge regression reflects that marker effects along the genome originate from the same distribution (de los Campos *et al.* 2013a), which does not imply that the same amount of regularization is applied to all markers, because shrinkage is allele frequency-dependent (Gianola 2013). When the value for σ^2_β is derived by BLUP using estimated variance components instead of modeling the prior distribution as in (8), the prior distribution in (7) describes the prior distribution pertaining to RR-BLUP.

The next method employed for genomic prediction was Bayesian Lasso where the prior distribution for the marker effects is described by a conditional mixture of Gaussian distributions (Park and Casella 2008; de los Campos *et al.* 2009). Consequently, the assumption of the same variance parameter across all markers is relaxed and marker-specific prior distributions are modeled within the following model hierarchy

$$\begin{aligned} \beta_j|\sigma^2, \tau_j^2 &\sim \mathrm{N}(0, \sigma^2\tau_j^2), & (9)\\ \tau_j^2|\lambda &\sim \mathrm{Exp}(\lambda^2), & (10)\\ \lambda^2|r, \delta &\sim \mathrm{Ga}(r, \delta), & (11) \end{aligned}$$

where (11) specifies a Gamma distribution for the regularization parameter λ with shape parameter δ and scale parameter r. The hyperprior distribution in (11) can be omitted when a fixed value for λ is supplied in (10). In Lehermeier *et al.* (2013) as well as Section 2.3.2, the advantage of using a Gamma distribution compared to a fixed value was investigated. In general, a large λ value will lead to a sharp prior distribution for β_j, *i.e.*, more shrinkage toward zero. It is expected that shrinkage of the Bayesian Lasso is stronger compared to the Gaussian prior distribution in Bayesian Ridge regression or RR-BLUP (Gianola 2013), a question that will be approached empirically in Section 2.3.2.

Finally, the Bayesian methods BayesA and BayesB were investigated which were proposed by Meuwissen *et al.* (2001) in the context of genome-based prediction. The prior distribution for the marker effects $j = 1, \ldots, p$ in BayesA and BayesB can be expressed through the following model hierarchy:

$$\begin{aligned} \beta_j|\sigma_{\beta_j}^2 &\sim \mathrm{N}(0, \sigma_{\beta_j}^2), & (12)\\ \sigma_{\beta_j}^2|\pi, \nu, S &\sim \pi\delta_0(\cdot) + (1-\pi)\chi^{-2}(\nu, S), & (13) \end{aligned}$$

where $\delta_0(\cdot)$ denotes a point mass at zero and $\pi \in [0, 1]$ controls the fraction of marker effects included in the model. Note that zero variance for a regression coefficient indicates complete certainty about its effect size in the Bayesian framework (Gianola *et al.* 2009). The BayesB method was implemented such that zero variance leads to a zero effect to induce variable selection while BayesB reduces to BayesA if $\pi = 0$. Again, both methods relax the assumption of an equal variance component for all marker effects and assign a variance that is specific for each marker. Thus, Bayesian Lasso, BayesA, and BayesB apply differential shrinkage to each marker, which is expected to be advantageous compared to Bayesian Ridge regression for traits with QTL of sizeable effect. BayesB does additional variable selection compared to BayesA and Bayesian Lasso which is considered to be advantageous if the true model has a sparse representation. However, with an inappropriate choice of π, either too many or too few markers will be selected and both scenarios can affect prediction performance adversely. The efficiency of variable selection compared to methods retaining all predictor variables in the model was investigated in Wimmer *et al.* (2013).

In general, the solution to these Bayesian methods cannot be computed analytically. Instead, Markov chain Monte Carlo (MCMC) techniques are used to sequentially generate samples from the full conditional posterior distributions of the unknown parameters. Together, the samples approximate the joint posterior distribution. Monitoring convergence, *i.e.*, checking whether the distribution converges toward a unique stationary posterior distribution, is crucial for valid statistical inference (Gelman *et al.* 2004). The algorithms involve a *burn-in phase*, where early samples are not used for posterior inference. In this thesis, the Markov chains were inspected visually to assess their convergence status and the required burn-in phase.

To summarize, regularized regression and Bayesian methods are powerful techniques for coping with overfitting problems in high-dimensional marker data to achieve an optimal bias-variance tradeoff. In Table 1, an overview of the models applied in this thesis is presented. Those methods with a variable selection feature are highlighted and a list of the unknown parameters for each method and the technique to tune them (CV, BLUP, or MCMC) is presented. Those methods that are available through the `synbreed` R package (Wimmer *et al.* 2012) are indicated.

1.5 Model assessment

When comparing statistical methods for genome-based prediction, it is important to assess their prediction performance in an independent validation set, because performance in the training data can be a poor description of the performance in test data (Figure 1). In this thesis, K-fold CV was used as an assumption-free method to assess the prediction

Table 1: Overview of statistical methods based on marker data used and compared in this thesis.

Method	Variable selection	Unknown parameters for marker effects	Tuning technique	Used in publication	Available through the `synbreed` package
RR-BLUP	no	λ	BLUP	Wimmer *et al.* (2013)	yes
GBLUP	no	λ	BLUP	Albrecht *et al.* (2011)	yes
LASSO	yes	λ	CV	Wimmer *et al.* (2013)	no
Elastic net	yes	λ, α	CV	Wimmer *et al.* (2013)	no
BRR[1]	no	σ^2_β	MCMC	Lehermeier *et al.* (2013)	yes
BL[2]	no	λ or r and δ	MCMC	Lehermeier *et al.* (2013)	yes
BayesA	no	ν, S	MCMC	Lehermeier *et al.* (2013)	no
BayesB	yes	π, ν, S	MCMC	Lehermeier *et al.* (2013) Wimmer *et al.* (2013)	no

[1]: BRR = Bayesian Ridge regression; [2]: BL = Bayesian Lasso

performance of different statistical methods in experimental and simulated data sets. In brief, the procedure utilized to compare different methods using K-fold CV is as follows:

1. Divide the training data set into K mutually exclusive sets $D_1, \dots, D_K$ of (almost) equal size. A typical choice is $K = 5$.

2. Define for each $k = 1, \dots, K$ the estimation set (ES) as $D_{\text{ES}} = \{D_i : i \neq k, i = 1, \dots, K\}$ and the test set (TS) as $D_{\text{TS}} = D_k$. By n_{ES} and n_{TS} the number of individuals in the ES and TS are denoted. A genome-based prediction method is fitted based on observations in the ES. If the method involves the tuning of a regularization parameter λ, an additional CV layer is used to select λ based on a grid search by repeating steps 1 to 4 in the ES (*i.e.*, for LASSO).

3. Compute the predicted genotypic values $\hat{\mathbf{g}}_{\text{TS}}$ for observations in the TS.

4. Evaluate prediction performance with a criteria $C(\hat{\mathbf{g}}_{\text{TS}}, \mathbf{y}_{\text{TS}})$ for the discrepancy of predicted genotypic values ($\hat{\mathbf{g}}_{TS}$) and observed phenotypic values ($\mathbf{y}_{TS}$) in the TS.

In most cases, different methods were compared based on their *predictive ability*, defined as Pearson's correlation coefficient $r_{\hat{g}y} = r(\hat{\mathbf{g}}_{TS}, \mathbf{y}_{TS})$ between predicted genotypic values and observed phenotypic values in the TS. Ideally, the methods should be compared based on their prediction accuracy defined as $r_{\hat{g}g} = r(\hat{\mathbf{g}}_{TS}, \mathbf{g}_{TS})$, *i.e.*, the correlation between predicted and true genotypic values, as the aim of genome-based prediction is to predict genotypic and not phenotypic values. However, prediction accuracy can be assessed only in computer simulations where true genotypic values are known; otherwise, predictive ability is influenced by trait heritability. To compare different traits with different heritabilities in experimental data sets, prediction accuracy was approximated by $r_{\hat{g}g} \approx r_{\hat{g}y}/h$ where h is the square root of trait heritability (Legarra *et al.* 2008). In addition, the prediction performance was measured by the prediction mean-squared error (PMSE), defined as

$$\text{PMSE}(\hat{\mathbf{g}}_{\text{TS}}, \mathbf{y}_{\text{TS}}) = \frac{1}{n}||\hat{\mathbf{g}}_{\text{TS}} - \mathbf{y}_{\text{TS}}||_2^2.$$

Besides prediction performance, an interesting question is the accuracy of genome-wide marker effects. An evaluation of the accuracy of estimated marker effects is not possible with experimental data sets where true marker effects are unknown; instead, computer simulations were employed in Wimmer *et al.* (2013) to compare the accuracy of estimated marker effects between different methods *in silico* (see Section 1.6.2). In Wimmer *et al.* (2013), the main measure for the accuracy of estimated marker effects was the normalized L_2 error between the vector of estimated ($\hat{\boldsymbol{\beta}}$) and true marker effects ($\boldsymbol{\beta}_0$) known from the simulation routine, which is defined as

$$L_2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \frac{||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2}{||\boldsymbol{\beta}_0||_2}. \tag{14}$$

This criterion describes the accuracy of both true zero and nonzero coefficients across scenarios and methods. The ability of the different variable selection methods to pinpoint causal mutations was investigated based on the number of true positive nonzero coefficients (TP), false positive true zero coefficients (FP), true negative true zero coefficients

(TN), and false negative true nonzero coefficients (FN), defined as

$$\mathrm{TP}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \sum_{j=1}^{p} \mathbb{1}(\hat{\beta}_j \neq 0 | \beta_{0j} \neq 0),$$

$$\mathrm{FP}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \sum_{j=1}^{p} \mathbb{1}(\hat{\beta}_j \neq 0 | \beta_{0j} = 0),$$

$$\mathrm{TN}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \sum_{j=1}^{p} \mathbb{1}(\hat{\beta}_j = 0 | \beta_{0j} = 0),$$

$$\mathrm{FN}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \sum_{j=1}^{p} \mathbb{1}(\hat{\beta}_j = 0 | \beta_{0j} \neq 0),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. From these measures one computes the sensitivity and specificity of the vector of estimated regression coefficients, defined as

$$\begin{aligned} \mathrm{Sens}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) &= \frac{\mathrm{TP}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)}{\mathrm{FN}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) + \mathrm{TP}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)}, \\ \mathrm{Spec}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) &= \frac{\mathrm{TN}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)}{\mathrm{FP}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) + \mathrm{TN}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)}, \end{aligned}$$

respectively. Both measures range between 0 and 1, and high values indicate good performance. In Wimmer *et al.* (2013), variable selection was considered to be successful when the sensitivity was above a predefined threshold of 0.8.

1.6 Data sets

Besides method validation in an independent data set, it is crucial to compare different methods using different data sets in order to make reliable inferences about the general performance of different genome-based prediction methods. In this section, a brief summary of the experimental and simulated data sets analyzed in this study will be presented. Further information can be found in the original publications (Albrecht *et al.* 2011; Lehermeier *et al.* 2013; Wimmer *et al.* 2013).

1.6.1 Experimental data sets

In Albrecht *et al.* (2011), an experimental data set of 1377 doubled haploid (DH) lines of maize (*Zea mays* L.) derived from 36 crosses was analyzed. Each family contributed 14 to 60 DH lines to the final data set. All lines were phenotyped as testcrosses with a single tester for two quantitative traits (grain dry matter yield and content) and genotyped using 732 polymorphic SNP markers.

The experimental data set in Lehermeier *et al.* (2013) consisted of 698 DH lines of maize derived from 122 crosses. On average, six DH lines were derived from each family with a minimum of 1 and a maximum of 63. All DH lines were phenotyped as testcrosses with a single tester for two quantitative traits (grain dry matter yield and content) and genotyped for 11646 polymorphic SNPs.

In Wimmer *et al.* (2013), three publicly available data sets of rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), and *Arabidopsis thaliana* (L.) were analyzed. For the rice data, phenotypic and genoptypic data for 413 rice varieties and 36901 SNPs were used from the original publication (Zhao *et al.* 2011). Out of 34 available traits, four traits with contrasting genetic architecture were selected based on the genome-wide association study (GWAS) results. The wheat data set (Poland *et al.* 2012) consisted of 254 breeding lines phenotyped for yield, thousand-kernel weight, and days to heading. Genotypic data on 2056 polymorphic SNPs were available. The *Arabidopsis* data set (Atwell *et al.* 2010) consisted of 199 accessions genotyped with 215908 SNP markers. Four out of the 107 available traits with contrasting genetic architecture were selected based on the GWAS results in Atwell *et al.* (2010).

1.6.2 Simulated data sets

In Lehermeier *et al.* (2013) and Wimmer *et al.* (2013), different simulation procedures were employed to investigate the influence of sample size, heritability, LD structure, and genetic trait architecture on the efficiency of different statistical methods *in silico*. In all cases, a predefined number p_0 of true nonzero coefficients (or QTL) were simulated and the genotypic values of n individuals were obtained as

$$g_i = \sum_{j=1}^{p_0} q_{ij}, \; i = 1, ..., n$$

where q_{ij} is the effect of the j-th QTL allele for individual i. Phenotypic values were simulated by adding random environmental residuals

$$y_i = g_i + e_i, \; i = 1, ..., n$$

where $e_i \sim \mathrm{N}(0, \sigma^2)$ and σ^2 was calibrated to achieve a predefined heritability. In Lehermeier *et al.* (2013), data was simulated following standard procedures, as in Meuwissen *et al.* (2001), but considering the specifics of a typical commercial maize breeding program. The objective of this simulation scheme was to generate data sets with an LD and a family structure such as observed in the experimental data set in Albrecht *et al.* (2011). The loci that were assigned to be QTL were not included in the marker panel, resembling the latent QTL assumption. These simulated data sets were used to compare the prediction performance of different Bayesian methods in Lehermeier *et al.* (2013).

In Wimmer *et al.* (2013), a different approach was used to control various parameters that are expected to influence the efficiency of statistical methods for genome-based prediction. A simulation scheme with varying numbers of markers, individuals, and trait heritabilities was employed. The set of true nonzero coefficients was included in the set of predictor variables to validate the accuracy of estimated marker effects based on both true zero and nonzero coefficients directly. In the first simulation procedure, the influence of determinedness level, defined as the ratio n/p, true model complexity level, defined as the ratio p_0/n, and the trait heritability was investigated in order to investigate their marginal influence on the efficiency of different statistical methods. In a second simulation procedure, real marker data was incorporated from the rice, wheat, and *Arabidopsis* data sets to obtain a more realistic picture of what can be expected in real data. This approach encapsulates the real LD structure (but not the genetic relatedness structure between individuals) but is limited to the actual observed sample sizes. Thus, a third simulation procedure was employed where the LD structure of the real data set was conveyed to simulated data sets of varying sample size.

It is worth mentioning that even though no LD structure was specified in procedure 1, the finite sample size will generate spurious correlations (Fan and Lv 2008). Moreover, it is important to be aware of similarities and differences in the simulated and experimental data sets. An important difference between all procedures in Wimmer *et al.* (2013) and the experimental data sets is that no genetic relatedness between individuals was simulated (see discussion in Sections 2.1 and 2.6).

1.7 Outline of the thesis

Using the statistical methods as well as experimental and simulated data sets described previously, the following research questions were investigated in the publications being part of this thesis (Albrecht *et al.* 2011; Wimmer *et al.* 2012; Lehermeier *et al.* 2013; Wimmer *et al.* 2013). First of all, the general usefulness of genome-based prediction in plant breeding was investigated in Albrecht *et al.* (2011). With marker data at hand, pedigree-based relationship coefficients were replaced by genome-based similarity coefficients measuring the realized proportion of shared alleles between pairs of individuals. The latter approach should provide more accurate predictions of the genotypic values because genome-based similarity coefficients account for Mendelian sampling (Hayes *et al.* 2009; Goddard 2009). This hypothesis was investigated with an experimental data set of maize testcross values and two complex traits. With Albrecht *et al.* (2011), empirical results at the population level including multiple families of maize have become available, describing the target data structure of advanced cycle breeding programs. Marker data were incorporated through the GBLUP method and different CV schemes were conducted to assess predictive abilities within and across families. The specific objectives in Albrecht *et al.* (2011) were to

- compare the accuracy of genome-based predictions with pedigree-based predictions in an advanced cycle breeding population of maize,
- assess the influence of the sample size on predictive ability, and,
- evaluate prediction performance within and across families through CV.

Results presented by Albrecht *et al.* (2011) are complemented by investigating the influence of increased marker density in Section 2.1.2. With increasing marker density, the probability of tagging QTL through markers is increased. Thus, an interesting question is the number of markers required for genome-based prediction, the answer to which is presented in Section 2.1.2 and Wimmer *et al.* (2013) for different plant populations. Large differences in magnitude of prediction performance were observed for different data sets and traits. Thus, in Section 2.1.3, important determinants of prediction accuracy such as sample size and trait heritability as well as theoretical formulas to predict prediction accuracy from these factors will be discussed.

In Section 1.4 it was shown that different statistical methods employ different approaches to optimize the bias-variance tradeoff. Thus, an important question is the choice of an appropriate genome-based prediction method. The GBLUP method is well-established as a standard method and is the most prevalent method used in plant and animal breeding (de los Campos *et al.* 2013b). The underlying assumption of this method is that all markers contribute according to the same Gaussian distribution for all marker effects (Hayes *et al.* 2009). It was envisaged from computer simulations that Bayesian methods allowing for marker-specific variances in the prior distribution are superior for traits controlled by a small number of QTL (Meuwissen *et al.* 2001; Daetwyler *et al.* 2010). Consequently, choice of prediction method can affect prediction performance and, thus, different Bayesian methods (Bayesian Ridge Regression, Bayesian Lasso, BayesA, and BayesB) were compared in Lehermeier *et al.* (2013) using experimental and simulated data sets. These methods require the specification of hyperparameters (see Table 1) but their impact on posterior inferences and prediction performance was largely unknown because most studies considered only a single set of hyperparameters. Thus, the objectives in Lehermeier *et al.* (2013) were to perform a sensitivity analysis in order to investigate

- the influence of hyperparameters in the prior distribution on the posterior distribution of marker effects and predictive ability of different Bayesian methods, and,
- whether models allowing for marker-specific prior variances instead of modeling the same prior distributions across all markers enhance prediction performance.

The influence of the hyperparameters and marker-specific prior variances was demonstrated in Lehermeier *et al.* (2013) while different approaches to select hyperparameters will be discussed in Section 2.3.

Variable selection constitutes a special case of regularization with marker effects being effectively removed from the model. Theoretically, methods with a built-in variable selection feature can outperform methods using all available markers when the vector of true nonzero coefficients is sufficiently sparse (Bühlmann and van de Geer 2011). The influence of the true model complexity was demonstrated with computer simulations in Donoho and Stodden (2006) for different variable selection methods where recovery of true nonzero coefficients with a high probability was only possible if $p_0 \ll n$. In addition, the authors highlighted the role of the determinedness level for statistical inference in high dimensions. With a constant true model complexity level, the performance of LASSO was derogated with respect to the recovery of true nonzero coefficients if the determinedness level was reduced. However, the simulations in Donoho and Stodden (2006) did not account for the properties of marker data, such as the discrete nature of predictor variables and LD among SNPs. Thus, the objective in Wimmer *et al.* (2013) was to identify scenarios where LASSO can recover causal mutations in high-dimensional marker data using different simulation procedures (Section 1.6.2). Besides sparsity, collinearity in the matrix of predictor variables confines prospects of variable selection (Bühlmann and van de Geer 2011). Thus, experimental data from three different plant species where incorporated into the study to assess the influence of different degrees of collinearity caused by LD. A total of 11 quantitative traits with presumably different genetic architecture were analyzed and the following questions were addressed in Wimmer *et al.* (2013):

- Under which conditions can it be expected that variable selection in addition to shrinkage enhances prediction performance for quantitative traits, and what are the factors that impact the ability of a method to recover true nonzero coefficients?
- What are upper bounds for the number of true nonzero coefficients which can be identified under different scenarios of sample size, trait heritability, and extent of LD?
- How accurate are marker effects estimated by different statistical methods and can genome-wide marker effects describe the genetic architecture of complex traits?

Computer simulations have revealed that methods performing variable selection or allowing for a non-Gaussian distribution of marker effects work well for traits influenced by a small number of QTL (Zhong *et al.* 2009; Daetwyler *et al.* 2010). However, different results have been obtained with experimental data sets. Here, a common finding among many empirical studies is that most methods perform similarly and no single best method across all traits and data sets emerged. This leads to the following additional question:

- Why are some genome-based prediction methods advantageous in computer simulations but not in studies based on experimental data?

Results in Wimmer *et al.* (2013) confirmed that variable selection can only enhance prediction performance when the number of true nonzero coefficients was small compared to the sample size, when strong LD was absent, and trait heritability was large. This is typically not the case in plant breeding populations showing strong LD, as discussed in

Wimmer *et al.* (2013) and sparsity might not be a reasonable assumption for the true model when analyzing high-dimensional marker data (see discussion in Section 2.5.2).

With whole-genome sequence data, the causal mutations are expected to be included among the predictor variables (Meuwissen and Goddard 2010). Thus, sparsity of the true model might become a reasonable assumption. However, analyzing whole-genome sequence data is challenging given the large number of potential predictor variables, leading to both statistical and computational challenges. Thus, methods performing a pre-selection of predictor variables to reduce dimensionality are of great interest and their prospects and limitations are discussed in Section 2.6 based on computer simulation results. Important topics related to the use of computer simulations for model assessment are discussed in Section 2.7.

Integration of genome-based prediction in breeding for crop improvement is still in its infancy. At the initiation of this thesis in 2010, no comprehensive software package was available to derive predictions from large-scale genomic data. However, both research and practical applications will be advanced by the availability of open-source software integrating analysis procedures for genotypic, phenotypic, and pedigree data (Heffner *et al.* 2009). Thus, Wimmer *et al.* (2012) provides

- an open-source software package within the R environment (R Development Core Team 2012) which implements genome-based prediction and covers special cases that occur in plant breeding populations with DH lines and repeated measurements, and,
- a unified data object to integrate phenotypic, genotypic, and pedigree data in various formats and from different species.

The publicly available R package `synbreed` described in Wimmer *et al.* (2012) provides a versatile analysis pipeline for genomic prediction and was released on the comprehensive R archive network (CRAN, `http://cran.r-project.org/web/packages/synbreed/`). In Section 2.8 it is shown how the package implements a unified analysis pipeline for genome-based prediction including several statistical methods described in this thesis.

The general discussion in Section 2 concatenates the topics of all publications and presents possible directions of future research, including additional investigations and results on several topics.

2 General discussion

The central research questions of this thesis were to investigate the relative efficiency of genome-based prediction in maize, to explore the performance of statistical methods which might improve the prediction performance compared to the GBLUP method, and to provide user-friendly software for the application of genome-based prediction. In this section, essential conclusions for genome-based prediction are discussed which emerge from findings in Albrecht *et al.* (2011), Wimmer *et al.* (2012), Lehermeier *et al.* (2013), and Wimmer *et al.* (2013) as well as from further unpublished results. Finally, possible directions of future research when analyzing whole-genome sequence data are discussed and the most important take-home messages of this work are summarized.

2.1 Genome-based prediction in maize and factors affecting prediction performance

2.1.1 Prediction with marker data instead of pedigree data

In Albrecht *et al.* (2011), testcross values were predicted with different models, and predictive ability was assessed using different CV schemes. In the first linear mixed model, the expected variance-covariance structure of the random effects for the genotypic values pertaining to the DH lines was inferred from three generations of pedigree data (PBLUP method). For the second linear mixed model, marker data were incorporated using different genome-based similarity coefficients including the GBLUP method. Finally, a linear mixed model including both pedigree and marker data was fitted. Predictive ability and accuracy were estimated using CV with a random allocation of individuals as well as CV within and across families. To investigate the influence of the sample size on predictive ability, the size of the training data set was reduced by retaining only fractions (1/2, 1/4, and 1/8) of DH lines chosen at random from the complete data set. The main result was that in all scenarios prediction accuracies were maximized with models incorporating marker data. For grain yield, an average gain of 29% within and 300% across families with respect to prediction accuracy was observed relative to a model using pedigree data alone. The increase of predictive ability in CV with GBLUP was confirmed with the data used in Lehermeier *et al.* (2013) where a gain of 76% in predictive ability compared to the PBLUP method was observed for grain yield (Figure 2).

The results in Albrecht *et al.* (2011) and Lehermeier *et al.* (2013) are promising for the implementation of genome-based prediction in maize breeding, because an increase in accuracy can be expected if predictions are based on marker data instead of pedigree data. In some scenarios, a further small, but significant increase in predictive ability was observed for models using marker data and pedigree data simultaneously. This was expected where marker coverage was low (Carré *et al.* 2013) and it was confirmed in experimental studies for wheat (de los Campos *et al.* 2009; Crossa *et al.* 2014). However, the additional advantage of pedigree data is expected to vanish if marker density is sufficiently high (de los Campos *et al.* 2010; Lorenz *et al.* 2011).

2.1.2 Benefit of increasing marker density

The studies in Albrecht *et al.* (2011) and Lehermeier *et al.* (2013) vary with respect to the number of markers used (732 and 11646 polymorphic SNPs, respectively). As more markers increase the probability of tagging QTL, an important question that needs to be addressed is the minimum number of markers required to obtain sufficient genome coverage for prediction. To investigate the influence of the number of markers on predictive ability, low density marker panels were generated *in silico* through a random masking of subsets of the available SNPs. These subsets were used to fit the GBLUP method and the predictive ability was evaluated using CV without accounting for family structure. The procedure was repeated 10 times for each number of markers, in order to account for randomness when sampling these marker subsets. This procedure was conducted for the data sets of Albrecht *et al.* (2011) and Lehermeier *et al.* (2013) and trait grain yield.

A plateau for the predictive ability was reached with approximately 2000 SNPs from the data in Lehermeier *et al.* (2013) (Figure 2). Conversely, predictive ability for the data set in Albrecht *et al.* (2011) did not yet plateau with the maximum number of available markers (732 SNPs). Thus, the full potential of accuracy with this data could not be exploited with the given number of markers. Indeed, an increase in predictive ability was observed when a subset of the DH lines used in Albrecht *et al.* (2011) was evaluated with CV using the same high-density marker panel as in Lehermeier *et al.* (2013) compared to the original 732 markers (results not shown). For the data in Albrecht *et al.* (2011), 100 SNPs were sufficient on average to predict testcross values for yield as accurately as using pedigree data, while for the data in Lehermeier *et al.* (2013), predictions based on 30 SNPs already outperformed PBLUP on average (Figure 2). The different performance of the PBLUP method in the two studies can be explained by their family structure. The data in Albrecht *et al.* (2011) comprised several large biparental families while the data in Lehermeier *et al.* (2013) consisted of many families with a small number of progenies. PBLUP was not very efficient for the latter design and the relative advantage of models using marker data increased.

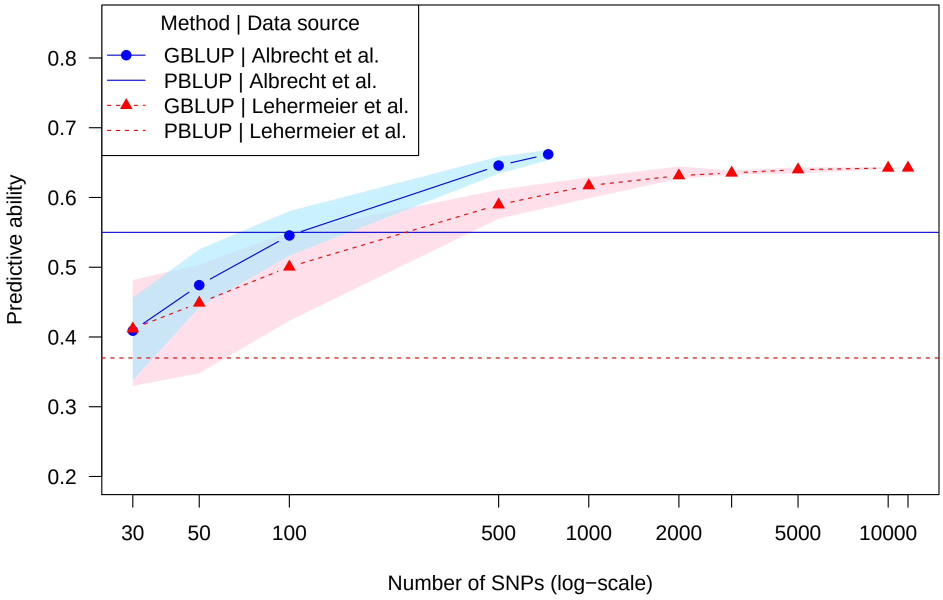

Figure 2: Predictive ability for marker subsets of the data in Albrecht *et al.* (2011) and Lehermeier *et al.* (2013). Predictive ability for grain yield was estimated with GBLUP and fivefold cross-validation with random sampling of individuals for different numbers of randomly selected SNP markers. Shaded areas indicate the range of 10 replications for each number of SNPs. Horizontal lines indicate the predictive ability of the PBLUP method.

The number of SNPs required to reach a plateau as in Figure 2 is expected to increase with increasing genome size and decreasing extent of LD (Jannink *et al.* 2010). With LD, neighboring markers tend to be ambiguous and subsets of SNPs can deliver similar predictive abilities. Interestingly, with roughly 2000 SNPs a plateau in predictive ability when using the RR-BLUP method was observed for the rice, wheat, and *Arabidopsis* data sets in Wimmer *et al.* (2013) even though they show very different marker densities and genome sizes. In de los Campos *et al.* (2010), more SNPs were required to reach the plateau when predicting the US-Holstein net merit index with dairy cattle data, and to outperform predictions based on the parent average (*i.e.*, based on their pedigree). This can be explained by the lower extent of LD in dairy cattle compared to the plant breeding populations. However, the sample size in de los Campos *et al.* (2010) was also larger than in the plant populations considered in Wimmer *et al.* (2013). One hypothesis is that statistical methods cope better with larger numbers of markers as sample size increases (Section 1.4.1). The influence of sample size on the number of markers required to reach the plateau has not been investigated but warrants further research using large experimental data sets and resampling with subsets of both individuals and markers.

2.1.3 Expected prediction accuracies

An important question for the practical application of genome-based prediction in breeding programs is the expected prediction accuracy for different traits and experimental settings. In Albrecht *et al.* (2011) and Wimmer *et al.* (2013), it was observed empirically that predictive ability was influenced by both the sample size n_{ES} and trait heritability h^2, respectively. Thus, an interesting question is the prediction accuracy that can be expected theoretically for these scenarios. Daetwyler *et al.* (2010) proposed a formula for the expected prediction accuracy based on these two factors as well as the number of independent chromosome segments M_e:

$$r_{\hat{g}g} = \sqrt{\frac{n_{\mathrm{ES}}h^2}{n_{\mathrm{ES}}h^2 + M_e}}. \tag{15}$$

The value for M_e increases with increasing genome size or increasing effective population size N_e (Goddard *et al.* 2011). For fixed h^2 and M_e, a diminishing return of prediction accuracy is expected when the sample size is increased (Figure 3). In Albrecht *et al.* (2011), this hypothesis was corroborated and a nonlinear relationship between $r_{\hat{g}g}$ and n with a diminishing return for large n was observed. An interesting question is whether the observed prediction accuracies match the expectation from Equation (15). However, this requires profound knowledge of the value of M_e, which was not available here, and is difficult to determine in experimental populations (Erbe *et al.* 2013).

For practical applications, knowledge about the sample size required to reach a plateau for the prediction accuracy will be valuable to determine optimal sample sizes when allocating resources in a breeding program. However, nominal sample size n_{ES} can be a poor description of the amount of information conveyed by the training data set, as demonstrated in the following. In Wimmer *et al.* (2013), the prediction performance evaluated with CV and the RR-BLUP method did not reach a plateau for the predictive ability with the largest size of the training set ($n_{\mathrm{ES}} = 1600$) in computer simulations (Figure 3). It produced a similar shape compared to the expected curve with $M_e = 1000$. Differences to the study in Albrecht *et al.* (2011) can be explained by the different relatedness structures within these data sets. In Wimmer *et al.* (2013), the training data set consisted of unrelated individuals, while in Albrecht *et al.* (2011) individuals were related through their family structure. Thus, the data sets were likely to differ with respect to their effective populations size and in turn their M_e. It is expected that adding an individual from one family does not add much information to the training data set if already other members of the same family were included. Indeed, this diminishing return of the sample size was much more pronounced in Albrecht *et al.* (2011) compared to Wimmer *et al.* (2013) where no family structure was present. Hence, increasing sample size of the training data set does not necessarily increase prediction performance, because the genetic structure of the population must also be considered. When considering training data sets of a given size, those consisting of genetically independent individuals are expected to deliver marker effects maximizing predictive ability achieved in an independent test data set (de los Campos *et al.* 2013b).

In addition to the composition of the training data set, genetic relatedness between individuals in the training and test data set contributes to prediction performance. There is increasing evidence that relatedness among individuals captured by markers is the major source of accuracy in genome-based prediction (Habier *et al.* 2010; Pérez-Cabal *et al.* 2012; Clark *et al.* 2012; de los Campos *et al.* 2013b). Partly for this reason, low-density marker panels provide already good predictive abilities as observed in Figure 2 (Weigel *et al.* 2009; Vazquez *et al.* 2010), because marker effects mainly reflect the resemblance of relatives. Under this hypothesis, it is not necessary to have markers in close proximity to all QTL in order to achieve high prediction accuracy for selection candidates related to the individuals in the training data set. Thus, to compare the prospects of genome-based prediction across scenarios, it is important to compare the efficiency of genome-based methods with pedigree-based methods, in order to assess the additional benefit when exploiting marker data for prediction.

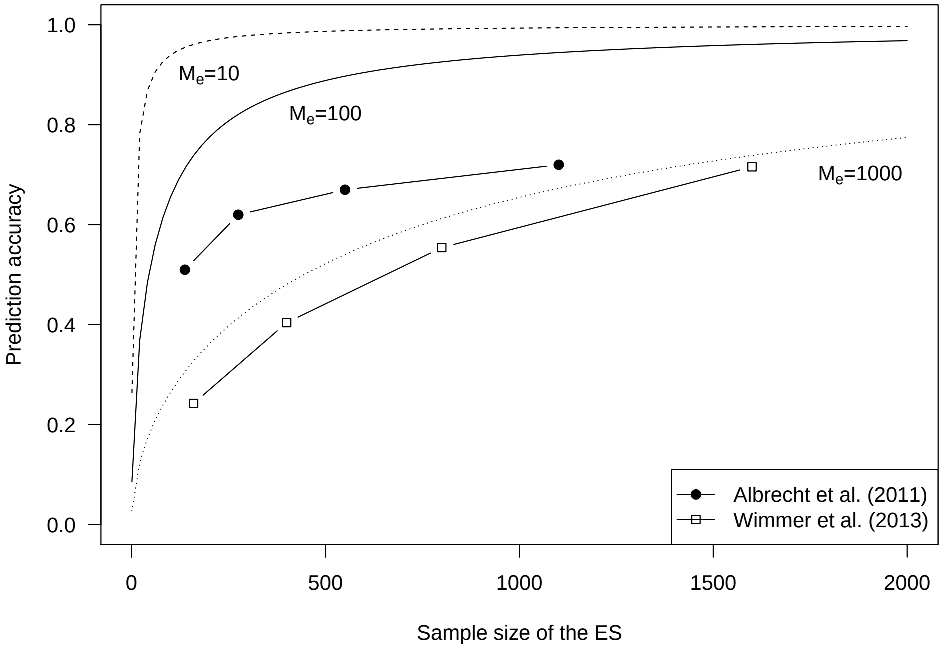

Figure 3: Expected prediction accuracy $r_{\hat{g}g}$ for different sample sizes evaluated based on the deterministic formula in Daetwyler *et al.* (2010) using $h^2 = 0.75$ and different values for the number of independent chromosome segments ($M_e = 10$, 100, and 1000). Superimposed are estimated average prediction accuracies for different sample sizes derived in Albrecht *et al.* (2011, Table 3) with the G-BLUP method and approximated accuracies using $r_{\hat{g}g} = r_{\hat{g}y}/h$ from Wimmer *et al.* (2013, Table 2) with the RR-BLUP method in simulations using $p_0/n = 0.25$, $p = 2000$, and $h^2 = 0.75$.

In Albrecht *et al.* (2011), predictive abilities achieved with genome-based prediction were significantly reduced when prediction was conducted across families compared to within families due to the lower extent of relatedness between training and testing data sets. Thus, observed predictive abilities in experimental studies strongly depend on the specific population structure and the training-testing design, *i.e.*, the relatedness of individuals in the training and testing data set. Using deterministic formulas to predict genome-based prediction accuracies remains challenging and warrants further research to obtain formulas which will account for both the structure in the training set and the training-testing relationships. Currently, deterministic formulas cannot replace empirical evaluations based on model assessment techniques such as CV to estimate prediction accuracies. As a consequence, estimated marker effects must be updated constantly because otherwise prediction performance will decrease over generations as selection candidates become less related to those in the training data set (Meuwissen *et al.* 2001; Habier *et al.* 2010).

2.2 Accuracy of estimated marker effects and the prospects of variable selection methods

In the context of genome-based prediction, it is expected that variable selection methods can lead to more persistent predictions across generations compared to GBLUP, because these methods do not rely on relatedness only, but instead tag QTL effects (Zhong *et al.* 2009; Habier *et al.* 2010; Lorenz *et al.* 2011). Moreover, the promise of variable selection methods is that they enhance prediction performance for traits controlled by a few major QTL and can reveal the underlying mechanisms of genetic trait architecture. Thus, a crucial question is do variable selection methods applied in plant breeding populations have the potential to provide more accurate marker effects by retaining only true nonzero coefficients and, if so, under what scenarios? Different methods can deliver very different marker effects as illustrated for LASSO and RR-BLUP with flowering time in the rice data analyzed in Wimmer *et al.* (2013). For this trait, biological prior knowledge about the *Hd1* gene with a large effect on chromosome 6 was available (Zhao *et al.* 2011). With LASSO a small number of markers with large effects were selected in this region while these effects were distributed across many SNPs in RR-BLUP (Figure 4).

In Wimmer *et al.* (2013), both the accuracy of estimated marker effects and predictive ability were investigated with the variable selection methods LASSO, the elastic net, and BayesB in comparison to RR-BLUP with different experimental data sets and a vast simulation study including more than 1000 scenarios from three simulation procedures with true models of varying complexity (see Section 1.6.2). The following major factors determining the efficiency of variable selection and the accuracy of estimated marker effects were identified and quantified:

1. Trait heritability: The accuracy of estimated marker effects increased with increasing trait heritability.
2. Extent of LD among markers: Accurate marker effects were only observed in scenarios without a large extent of LD among the markers. Otherwise, LD can create ambiguity, because true nonzero coefficients act as proxies for true zero coefficients (Gianola 2013), thus leading to non-accurate estimates of individual marker effects.

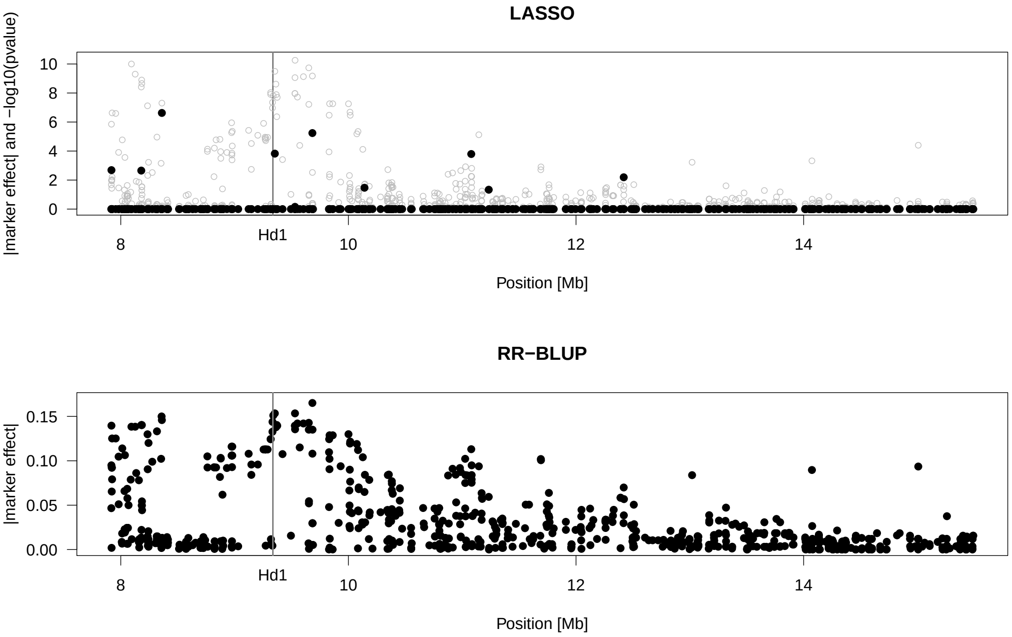

Figure 4: Marker effects of LASSO and RR-BLUP for trait flowering time in the rice data. Genome-wide marker effects were estimated and those near the *Hd1* gene on chromosome 6 are visualized. Top panel: Filled circles represent absolute values of marker effects obtained by LASSO, while gray open circles indicate the $-\log_{10}(p\text{-value})$ for each marker as obtained from the genome-wide association study in Zhao *et al.* (2011). Bottom panel: Filled circles represent the absolute values of marker effects obtained by RR-BLUP.

3. Complexity of the true model: Similar to the study in Donoho and Stodden (2006), variable selection was only successful if the number of true nonzero coefficients was much smaller than the sample size.

The major finding in Wimmer *et al.* (2013) was that variable selection methods outperformed RR-BLUP with respect to both the accuracy of marker effects and predictive ability only if assumptions 1 to 3 were fulfilled simultaneously. Otherwise, prediction performance was similar across different methods because all methods cannot reveal the set of true nonzero coefficients. This explains why many experimental studies reported only small differences for the predictive ability of methods with and without variable selection feature (Heslot *et al.* 2012; Riedelsheimer *et al.* 2012; Wimmer *et al.* 2013) although considerable advantages of variable selection methods were observed in computer simulations (Meuwissen *et al.* 2001; Daetwyler *et al.* 2010). The reason is, that the assumptions for successful variable selection can be fulfilled *in silico* but are unlikely to be fulfilled for many experimental settings given small sample sizes and long-range LD among markers.

These results shed light on the ongoing discussion about the choice of method to predict complex traits and the role of variable selection. Not only the genetic trait architecture but also the sample size of the training data set limits the efficiency of variable selection for genome-based prediction. The well-known requirement that more observations than predictor variables are required to estimate regression coefficients accurately in linear models is violated with data structures commonly employed for genome-based prediction or GWAS. For example, Huber (1981, p.197) proposed $n/p \geq 5$ as a rule of thumb for meaningful coefficients in regression analyses, and this cannot be circumvented by means of regularization or prior distributions, unless the vector of true regression coefficients is sufficiently sparse (Gianola 2013; Wimmer *et al.* 2013). Theory and empirical results such as in Donoho and Stodden (2006) and Wimmer *et al.* (2013) point to the conclusion that the $n/p \geq 5$ rule can be replaced by $n/p_0 \geq 5$ to obtain meaningful estimates for regression coefficients when $n < p$. Thus, variable selection can circumvent the curse of dimensionality, but only when the true model is sufficiently sparse. In addition, absence of multicollinearity caused by correlations among true zero and nonzero coefficient is expected to be crucial (Bühlmann and van de Geer 2011). This was confirmed in Wimmer *et al.* (2013) where the necessary sample size for successful recovery increased tremendously with increasing extent of LD.

It is important to keep in mind that statistical methods can be useful for genome-based prediction, even though they deliver marker effects of low accuracy (Gianola 2013). The reason is that a high-dimensional data set does not contain enough information for estimating the effects of p markers when $n < p$, but the n genotypic values are likelihood-identifiable (Gianola 2013). An example is RR-BLUP which is expected to provide marker effects of low accuracy in many scenarios but still can predict accurately genetic values because it can exploit genetic relatedness for prediction.

The accuracy of individual marker effects was measured by the normalized L_2 error in Wimmer *et al.* (2013). This measure does not account for cases where a true nonzero coefficient was not recovered, but the effect was assigned to a flanking true zero coefficient in high LD. Gianola *et al.* (2009) measured accuracy at the level of genomic regions by considering a window of four flanking markers around each true nonzero coefficient. Then accuracy was evaluated as the fraction of retrieved genomic regions. Using this kind of measure might be more appropriate if the goal is to describe accuracy at the level of genomic regions instead of individual marker effects. Based on simulations in Gianola *et al.* (2009), it is expected that the drop in accuracy from scenarios with low LD to scenarios with high LD will not be as severe as observed in Wimmer *et al.* (2013) based on the normalized L_2 error.

2.3 Choice of regularization and hyperparameters

Besides the choice of an appropriate prediction method, the proper choice of regularization parameters or hyperparameters is crucial to optimize the bias-variance tradeoff for a specific setting and to maximize prediction performance. Regularization parameters can be specified based on prior knowledge or derived from the training data using resampling strategies such as CV to tune them by a grid search. The advantage of CV is that this approach is assumption-free, but this comes at the expense of high computational costs. Strategies such as BLUP evaluate a single fit to derive the regularization parameter. Different aspects associated with the choice of regularization parameters in LASSO, RR-BLUP, and the Bayesian Lasso are discussed in the next section.

2.3.1 Variable screening and prediction

The key parameter to control the behavior of LASSO is the regularization parameter λ. If λ increases, the extent of regularization increases and fewer variables will be selected. In Wimmer *et al.* (2013), LASSO was tuned for prediction using CV; however, tuning for variable screening may require more severe regularization to obtain sparser solutions and fewer false positives (Bühlmann and van de Geer 2011).

For LASSO, it can be illustrated how different values of the regularization parameter affect the prediction performance and accuracy of estimated marker effects. A sequence of 100 λ values as obtained by the `glmnet` R package (Friedman *et al.* 2010) was generated and the corresponding set of estimated nonzero coefficients was computed. In order to identify the best choice of λ, the PMSE (evaluated with fivefold CV), sensitivity, specificity, and normalized L_2 error of the estimated marker effects were displayed for the grid of λ values. Different scenarios for true model complexity level were generated with simulation procedure 1, according to Wimmer *et al.* (2013), using $n = 500$ and $n = 1000$ as well as $p_0 = 250$ and $p_0 = 125$, respectively. The number of markers and trait heritability were constant with $p = 2000$ and $h^2 = 0.75$.

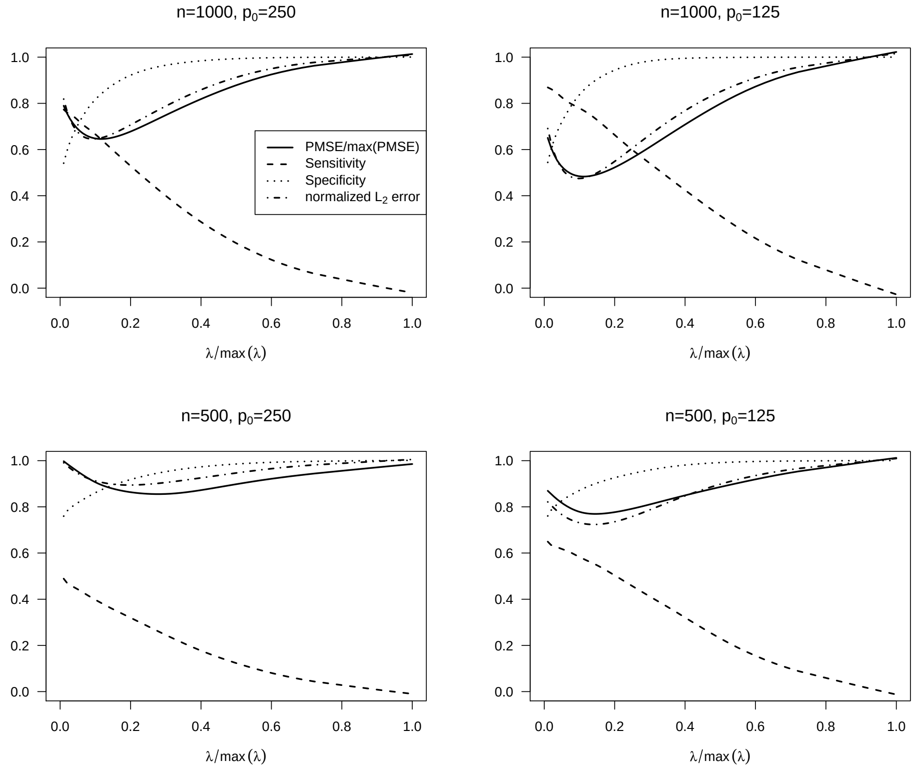

Figure 5: Performance of LASSO for different values of λ, n, and p_0, with data simulated according to procedure 1 in Wimmer *et al.* (2013) using $p = 2000$ and $h^2 = 0.75$. Displayed are the results for the four performances measures averaged across 10 replications for each scenario.

As expected, the choice of regularization parameter λ had a significant influence on the different performance measures (Figure 5). The results revealed that prediction performance measured by the PMSE was maximized at an intermediate value of λ, while optimal sensitivity and specificity were observed as expected for small and large values of λ, respectively. Interestingly, although the normalized L_2 error measured the accuracy at the level of marker effects, the curve in Figure 5 followed closely the curve for the PMSE. The PMSE measures the squared bias and variance of the predicted genotypic values, while the normalized L_2 error measures the squared bias and variance of the estimated marker effects. Presumably, for both measures, the tradeoff between bias and variance was optimized by similar λ values. Thus, tuning LASSO for the PMSE delivered an (almost) optimal normalized L_2 error, but not the optimal sensitivity or specificity.

The absolute values of the sensitivity and the normalized L_2 error depended on the combination of n and p_0, while best performance was observed as expected for $n = 1000$ and $p_0 = 125$. When n was halved to $n = 500$ ($p_0 = 125$), the normalized L_2 error increased more than when the number of true nonzero coefficients was doubled to $p_0 = 250$ (both scenarios had the same true model complexity level of $p_0/n = 0.25$). The reason is the lower determinedness level n/p in the first scenario. Specificity was influenced mainly by sample size, which determines the upper bound for the number of selected predictor variables in LASSO and thus the upper bound for the number of false positives. The curve around the optimum value for λ with respect to the PMSE was sharp when $n = 1000$ and $p_0 = 125$ but flat for $n = 500$ and $p_0 = 250$. Thus, it appears that the choice of regularization parameter for LASSO was of particular importance in scenarios where variable selection is expected to work well (*i.e.*, when the ratio p_0/n is small) in order to select the optimum number of markers into the model. It is important to make sure that the grid of λ values that is investigated within CV is sufficiently dense and covers the global minimum. In general, the default values generated by the `glmnet` R package, following Friedman *et al.* (2010), worked well for the data structures presented in this thesis.

2.3.2 Sensitivity analysis of Bayesian genome-based prediction methods

With Bayesian methods, the choice of regularization parameters using CV, as in LASSO, is computationally prohibitive when using MCMC algorithms. Alternatively, the Bayesian framework allows to incorporate prior distributions for the regularization parameters. Through Bayesian learning, suitable regularization parameters are derived from the data. However, an additional layer of hyperprior distributions must be specified, which requires additional hyperparameters. These hyperparameters can affect posterior inference, but most studies comparing different Bayesian prediction methods report results only for a single set of hyperparameters, and limited effort has been made to assess their influence on prediction performance. A notable exception is Gianola *et al.* (2009), where the influence of hyperparameters on posterior inference was investigated for the BayesA method with computer simulations. Here, the practical consequences of the choice of hyperparameters on the bias-variance tradeoff will be demonstrated with the example of the Bayesian Lasso and it will be discussed how methods should be tuned for prediction in practice.

In Lehermeier *et al.* (2013), a sensitivity analysis was conducted with four frequently used genome-based prediction methods (Bayesian Lasso, Bayesian Ridge, BayesA, and BayesB) using simulated and experimental maize data. The influence of the prior specification on posterior inference was investigated with different scenarios where one hyperparameter controlling the extent of regularization was altered while the other hyperparameters were kept constant across scenarios. For each method, one scenario was defined using "optimal" hyperparameters according to Pérez *et al.* (2010). These *ad-hoc* formulas relate the hyperparameters to trait heritability, and thus the prior belief about signal-to-noise ratio in the data. For Bayesian Lasso, three scenarios named BL1-3 were defined with a fixed parameter λ in Equation (10), while in scenarios BL4-6 three different Gamma distributions in Equation (11) were assigned to λ. In the latter case, the prior for the regularization parameter λ should be updated through Bayesian learning. The scenarios were defined such that for both fixed and random λ one scenario was close to the optimal signal-to-noise ratio (BL1 and BL4), while two other scenarios were expected to generate underfitting (BL3 and BL6) or overfitting (BL2 and BL5), respectively. Scale parameter r of the Gamma distribution in the random scenario was selected such that the mode of the prior distribution for λ in one random scenario matches the value of one fixed scenario. To detect overfitting or underfitting, the model fit was evaluated by the correlation of observed and predicted testcross values in the ES as well as predictive ability in the TS using CV for each scenario.

Results in Table 2 demonstrate the huge influence of the choice of hyperparameters on the number of effective parameters and the correlation of predicted genotypic values and observed phenotypic values in the ES and TS. Bayesian Lasso scenario BL2 with small fixed regularization parameter λ generated a high number of effective parameters leading to overfitting, *i.e.*, a good fit in the ES but poor prediction performance in the TS. Vice versa, scenario BL3 produced as expected a low number of effective parameters and underfitting. Both overfitting and underfitting were avoided when a Gamma prior distribution was modeled for λ. A similar predictive ability was obtained in scenario BL1 where λ was selected according to the heritability. For comparison, predictions were contrasted with the RR-BLUP method, where the regularization parameter was tuned by the noise-to-signal ratio in the ES (using REML estimates of the marker and residual variance

Table 2: Number of effective parameters ('No. eff. par.') calculated by the trace of the hat matrix following Gianola (2013) as well as the average correlation of predicted genotypic values and observed phenotypic values in the TS ('Corr TS') and ES ('Corr ES') across the five cross-validation folds and methods Bayesian Lasso with six scenarios following Lehermeier *et al.* (2013) as well as method RR-BLUP.

Scenario	No. eff. par.	Corr TS	Corr ES
BL1	224.5	0.53	0.76
BL2	632.8	0.42	0.92
BL3	13.77	0.35	0.44
BL4	189.3	0.52	0.73
BL5	194.7	0.52	0.74
BL6	189.4	0.52	0.73
RR-BLUP	172.2	0.52	0.73

components). Interestingly, RR-BLUP had a similar number of effective parameters and predictive ability as compared with the BL4-6 and BL1 models even though it employs the same prior variance parameter for all markers instead of variances specific for each marker as in Bayesian Lasso. Due to the shape of the Laplace distribution employed by Bayesian Lasso (Park and Casella 2008), a smaller number of effective parameters was expected for the Bayesian Lasso compared to RR-BLUP which employs a Gaussian distribution (Gianola 2013). However, the results will be considerably influenced by the specific choice of hyperparameters (D. Gianola, personal communication).

It was observed that different methods can deliver a similar number of effective parameters, even though they employ different prior distributions for the marker effects. These results have a strong impact on the ongoing discussion about the best choice of method for genome-based prediction. One can conjecture that it is less important which specific method is used for prediction; rather, the choice of regularization parameters influences prediction performance. Results from Lehermeier *et al.* (2013) as well as Table 2 indicate that it was better to assign a Gamma distribution to λ instead of using a fixed value because the regularization parameter can be updated by Bayesian learning then. Recall that in BL4-6 the regularization parameter was estimated through the hierarchical model described in Equations 9 to 11, while in model BL1 and RR-BLUP the regularization parameter was tuned by the signal-to-noise ratio in the ES. All models delivered a similar predictive ability and a similar number of effective parameters. Thus, it can be suggested that it is sufficient to tune the regularization parameters based on this value, given that a reasonable estimate of the signal-to-noise ratio is available. Otherwise, the regularization parameter should be treated as a random variable in the Bayesian framework (as in BL4-6), or CV should be used to tune the regularization parameter (as in LASSO). The number of effective parameters defined by the trace of the hat matrix was a valuable tool to compare different methods or scenarios. An interesting question is whether one can identify the optimal extent of regularization in advance by using data properties such as sample size, marker density, genetic trait architecture, or trait heritability which warrants further research.

2.4 Efficiency of Bayesian methods

2.4.1 Modeling marker-specific prior variances

Computer simulations and studies from animal breeding suggest that in some cases Bayesian methods can be more efficient for genome-based prediction compared to GBLUP. Recall that the underlying assumption behind GBLUP (or RR-BLUP) is that all marker effects originate from the same prior distribution (this does not imply that all marker effects are equal). Bayesian methods such as BayesA are expected to outperform RR-BLUP by allowing for prior distributions specific to each marker. This approach should overcome shortcomings of methods using a common variance component for traits influenced by a few QTL of sizeable effects (Meuwissen *et al.* 2013). However, Gianola (2013) pointed out that all marker effects possess the same marginal distribution in BayesA and any differences to Bayesian Ridge regression or RR-BLUP occur because shrinkage is marker-effect specific in BayesA such that small effects receive more regularization. This additional flexibility comes at the expense of additional hyperparameters in the prior distribution confining Bayesian learning abilities (Gianola *et al.* 2009; Gianola 2013). Empirical results in Lehermeier *et al.* (2013) suggest that marker-effect specific shrinkage in BayesA, BayesB, or Bayesian Lasso did not enhance prediction performance compared to Bayesian Ridge regression.

2.4.2 Influence of the prior distribution on posterior inference

Different Bayesian methods are expected to vary with respect to their Bayesian learning abilities. Theoretical results in Gianola *et al.* (2009) illustrated that Bayesian learning was limited for BayesA and BayesB, and the scale parameter S in (13) was influential, as it controls the extent of regularization. Empirical results in Lehermeier *et al.* (2013) corroborated this hypothesis, and the distance between prior and posterior distribution was quantified with their Hellinger distance (Le Cam 1986). Bayesian Lasso and Bayesian Ridge regression were less influenced by the prior distribution compared to BayesA and BayesB. Unexpectedly, the influence of the prior distribution was more pronounced in BayesA than in BayesB, although BayesB has the drawback of a fixed hyperparameter value for π (Habier *et al.* 2011; Gianola 2013). Meuwissen *et al.* (2001) selected the value for π based on knowledge taken from the simulation procedure. Nonetheless, this approach was not feasible with experimental data, and approaches to estimate π from data are of interest.

2.4.3 Estimating the variable selection intensity from data

In BayesB, the fraction of markers π assigned zero variance in (13) is not subject to Bayesian learning. The method BayesCπ is expected to overcome the shortcoming of a fixed value for π in BayesB, because π is treated as an additional random variable which is inferred from the data (Habier *et al.* 2011). However, convergence problems in the Markov chains were observed for the parameter π when applying this method to the data sets in Wimmer *et al.* (2013) (data not shown). A similar observation was made by Wolc *et al.* (2011) and Colombani *et al.* (2012), and this might be explained by the large extent of LD in these data sets leading to ambiguity with respect to the information conveyed by the SNP markers. This was confirmed by the posterior inclusion probabilities for BayesB within the experimental data sets of Wimmer *et al.* (2013). Most markers had an equal probability of being included in the model, confirming that a large number of marker subsets with equal predictive power exists (results not shown). In all scenarios with BayesB, the value $\pi = 0.8$ was used in Lehermeier *et al.* (2013), but through a grid search with different values for π it was observed that other values delivered similar predictive abilities in the maizeA data set (data not shown). Thus, BayesB was partly indifferent to whether the effects were assigned to a smaller number of loci obtained by variable selection or if the QTL effects were distributed across many markers. In scenarios showing ambiguity among many markers, there will be no advantage of BayesCπ compared to BayesB.

2.4.4 The Bayesian elastic net

For many penalized least-squares estimation methods, such as LASSO, Ridge regression, or the elastic net, Bayesian counterparts are available. For the standard elastic net (Zou and Hastie 2005), a two-dimensional grid search was required to optimize prediction performance in Waldron *et al.* (2011). Tuning multiple regularization parameters is computationally demanding when using a multidimensional grid search in conjunction with CV. The promise of the Bayesian elastic net (Li and Lin 2010) is that it will assess the optimal allocation of the regularization parameters to the L_1 and L_2 norm penalty functions through Bayesian learning. It is expected that for traits with a few major QTL, more weight should be given to the L_1 norm part in order to emphasize the variable selection feature of LASSO, while for complex traits, the L_2 norm part is expected to become more important. Stuckart (2012) explored the efficiency of the Bayesian elastic net compared to the standard elastic net as well as LASSO, Ridge regression, and their Bayesian counterparts using experimental data for *Arabidopsis* ($n = 426$, $p = 1260$) and four quantitative traits obtained from Kover *et al.* (2009). No relevant differences between the methods with respect to their predictive ability were observed. For the Bayesian elastic net, convergence problems for the regularization parameters were identified within the MCMC algorithm, and posterior inference was influenced considerably by the hyperparameters in the prior distribution adopted (Stuckart 2012). Thus, the algorithm was ambiguous regarding whether to emphasize variable selection through the L_1 norm penalty function or distributing marker effects across the genome through the L_2 norm penalty function. This confirmed that different marker subsets are likely to have the same predictive power when there is strong LD among markers, and variable selection does not improve prediction for complex traits in this case.

2.4.5 Application of Bayesian methods for genome-based prediction

To summarize, the implementation of Bayesian methods requires specifying hyperparameters that will affect posterior inference to a certain degree. Those methods that allow for strong Bayesian learning of the regularization parameters are preferable because they can circumvent overfitting more effectively. As discussed in Section 2.2, Bayesian methods can be efficient for prediction but not necessarily to estimate marker effects accurately in high-dimensional marker data. Computation times of Bayesian methods are considerably higher compared to penalized least-squares techniques. A grid search to select hyperparameters with CV is not feasible with the Bayesian methods, indicating the need for Bayesian methods that are robust with respect to the choice of hyperparameters.

With experimental data, no advantage of Bayesian methods was observed compared to GBLUP with respect to their predictive ability (Lehermeier *et al.* 2013). An emerging approach is *Bayesian model averaging* (BMA) to utilize an ensemble of prediction models in order to account for uncertainty associated with the choice of a single model (Raftery *et al.* 1997). In the context of genome-based prediction, different models can capture different aspects of the underlying genetic trait architecture, and it emerges that a unified model might be superior for predicting genotypic values (Jannink *et al.* 2010). The predictive performance of the average model is expected to be superior compared to the best single model (Sorensen and Gianola 2002) but further research is required to explore the possibilities of BMA for genome-based prediction.

2.5 Genetic trait architecture and complexity of the true model

2.5.1 Estimating the number of QTL

The proper choice of a statistical method for genome-based prediction will be advanced by knowledge about the genetic architecture of the trait under study (Zhong *et al.* 2009; Daetwyler *et al.* 2010; Coster *et al.* 2010). If the number of QTL (N_{QTL}) are known for a specific trait, the results in Wimmer *et al.* (2013) could guide the choice of an appropriate method. For traits with $N_{QTL} \ll n$, methods with a variable selection feature are expected to enhance prediction performance. Thus, estimates for N_{QTL} will be valuable in practice and therefore some researchers have tried to estimate N_{QTL} from the data; for example, Daetwyler *et al.* (2010) presented a deterministic formula to obtain $\hat{N}_{QTL}$ from the observed prediction accuracies of BayesB.

This formula was validated using the BayesB derived predictive abilities for different scenarios from Table 4 in Wimmer *et al.* (2013). For these data structures a discrepancy between the actual simulated number of QTL and the estimated values based on the formula in Daetwyler *et al.* (2010) was observed. This might be explained by the fact that the formula does not account for the LD and relatedness structures when estimating N_{QTL}, while these factors where major determinants of the prediction performance for the data sets analyzed in Wimmer *et al.* (2013). Alternatively, Habier *et al.* (2011) used the method BayesCπ to estimate $\hat{N}_{QTL}$ in computer simulations based on the posterior estimate of π. The idea is that the method will set a large fraction π of marker effects to zero when the trait has a sparse representation. They found that $\hat{N}_{QTL} = p \cdot (1 - \hat{\pi})$ was a poor description of the simulated value of N_{QTL} in many scenarios, probably because of the difficulties in finding a proper estimate for π using BayesCπ as described in Section 2.4.3. Thus, estimating N_{QTL} remains challenging, but there is evidence that many quantitative traits are influenced by a large number of QTL with small effects (Schön *et al.* 2004; Buckler *et al.* 2009), with few exceptions such as kernel carotenoid content in maize (Wallace *et al.* 2014).

2.5.2 Definition of the true model

In the context of genome-based prediction, it was assumed in Wimmer *et al.* (2013) that the true model is defined by a subset of the observed loci with additive effects on the trait under study. This concept was powerful when illustrating the joint influence of the complexity level of the true model, the determinedness level, and the trait heritability on the efficiency of different statistical methods. However, is sparsity a reasonable assumption for the true model underlying complex traits? With marker data, it is generally assumed that the QTL will be latent and marker effects are expected to tag QTL effects. Thus, the existence of a single true model involving a single subset of important predictor variables is implausible, and the definition of true nonzero coefficients is not straightforward when LD is present and multiple markers tag QTL (Li and Sillanpää 2012) or single markers tag multiple QTL. Partly for this reason, many alternative models can deliver the same predictive power (see Section 2.4), but a paradigm change is expected with whole genome-sequence data where causal mutations are included in the data (Meuwissen and Goddard 2010). Here, sparsity can become a reasonable assumption and the true model is defined unambiguously by the set of causal mutations, as discussed in the next section.

2.6 Toward the analysis of whole-genome sequence data

With the progress in whole-genome resequencing technologies, it will soon become feasible to derive the full genome sequence of a large number of individuals. These data are valuable for genome-based prediction, and benefits of whole-genome sequence data compared to dense marker data are expected for the following reasons:

- Causal polymorphisms are expected to be included in whole-genome sequence data while they are unlikely to be included in marker panels. In the latter case, markers are in LD with the mutations, but this association breaks down over generations due to recombination. It is desirable to have the causal mutations directly in the data, in order to achieve more persistent predictions over generations based directly on the estimated effects of the causal mutations. Computer simulations revealed a substantial increase in predictive ability with sequence data compared to predictions based on dense marker data, given that a method with a variable selection feature was used (Meuwissen and Goddard 2010).
- The proportion of genetic variance captured by whole-genome sequence data is expected to increase compared to current marker panels because accuracy is no longer bounded by the extent of LD between causal mutations and markers (Druet *et al.* 2014; Wray *et al.* 2013).
- Whole-genome sequence data are expected to tag structural genetic variations such as insertions, deletions, or copy number variants in addition to SNPs (Daetwyler *et al.* 2013). These can be included as additional predictor variables in genome-based prediction models.

Given these advantages, whole genome-sequence data are likely to replace current SNP arrays for genome-based predictions, once sufficiently large training data sets are available (Ober *et al.* 2012; Meuwissen *et al.* 2013). On the downside, sequencing data will increase the number of potential predictor variables by a factor of 10^2 to 10^3 and new computational challenges will emerge especially for Bayesian methods given the computation times for

MCMC algorithms that scale with the number of predictor variables. Furthermore, it was demonstrated *in silico* that even for traits with a low true model complexity level a loss in efficiency of different statistical methods can be expected when the determinedness level (*i.e.*, the ratio of n to p) is much smaller than one (see Wimmer *et al.* (2013) as well as Figure 6 where $p_0/n = 0.1$). Dimension reduction of whole-genome sequence data by pre-screening predictor variables emerges as a strategy to reduce the number of predictor variables and circumvent dimensionality, as indicated in Wimmer *et al.* (2013). In the following, the potential of methods that perform an educated selection of predictor variables is discussed.

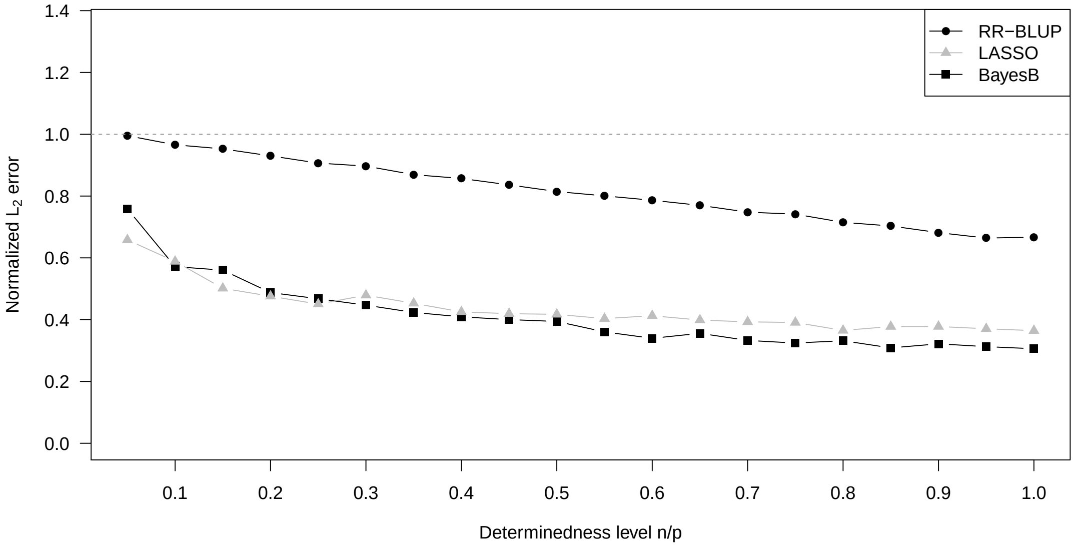

Figure 6: Performance of RR-BLUP, LASSO, and BayesB for different values of n/p using simulation procedure 1 in Wimmer *et al.* (2013) with $p = 2000$, $p_0/n = 0.1$, and $h^2 = 0.75$.

Fan and Lv (2008) proposed sure independence screening (SIS) to reduce the number of predictor variables from a very high to a moderate scale, for example below the sample size. This method was previously used in the context of genome-based prediction with SNP marker data (Kärkkäinen and Sillanpää 2013) and is computationally feasible even with millions of predictor variables. The procedure works as follows. First, the predictor variables are ordered by their marginal correlation with the response variable, and only those that rank among the top d predictor variables are retained for further analysis and, thus, the subset of predictor variables identified by SIS is given by

$$\hat{S}_{\mathrm{SIS}}(d) = \left\{ j : |\rho_j| \geq |\rho_{(d)}|,\ j = 1, \ldots, p \right\},$$

where $|\rho_j|$ denotes the absolute value of the marginal correlation of predictor variable $j = 1, \ldots, p$ with the response variable and $|\rho_{(d)}|$ denotes the absolute value of the d-th largest correlation. Thus, the underlying procedure in SIS is initially similar to the approaches applied in a GWAS. The main difference is that the subset $\hat{S}_{\mathrm{SIS}}(d)$ is not the final result, but instead an additional variable selection method will be applied within $\hat{S}_{\mathrm{SIS}}(d)$ to further reduce the number of nonzero coefficients. Here, LASSO was applied after SIS (SIS-LASSO), and the power of LASSO in this subset is expected to be higher compared to the original data set because the determinedness level n/p increases (Ishwaran and Rao 2011). The performance was compared in computer simulations for whole-genome sequence data using $p = 250000$, $n = 200$, $h^2 = 0.75$, and varying p_0 and d within simulation procedure 1 of Wimmer *et al.* (2013). Each scenario was replicated 10 times by simulating new data, and the results were averaged over replications.

First, the potential of pre-screening predictor variables with respect to prediction performance was explored under the (unrealistic) assumption that one had a pre-screening procedure at hand which can remove a subset of the true zero coefficients in advance. In Figure 7, the predictive ability of LASSO is displayed for different numbers of causal mutations in the simulation scheme when fractions of superfluous true zero coefficients were removed from the data prior to the analysis. Across all scenarios, predictive ability was largest when only four causal mutations were underlying trait expression, and the efficiency of LASSO could not be further improved by reducing the number of superfluous SNPs in advance. When more mutations were simulated, predictive ability was considerably increased when superfluous markers were removed by pre-screening. With $p_0 = 40$, almost no predictive ability was observed when all available markers were used for prediction, but when only 250 SNPs were retained (*i.e.*, more than 99% of the superfluous SNPs were removed), a predictive ability $r_{\hat{g}y} > 0.60$ was observed. These results do not only demonstrate the potential of pre-screening methods with whole-genome sequence data but also indicate that LASSO, with its built-in variable selection feature through the L_1 norm penalty function, was already very efficient for traits controlled by a small number of causal mutations and cannot be improved significantly with pre-screening methods. This confirmed theoretical results on the high efficiency of LASSO in scenarios where the number of true nonzero coefficients was small (Eldar and Kutyniok 2012).

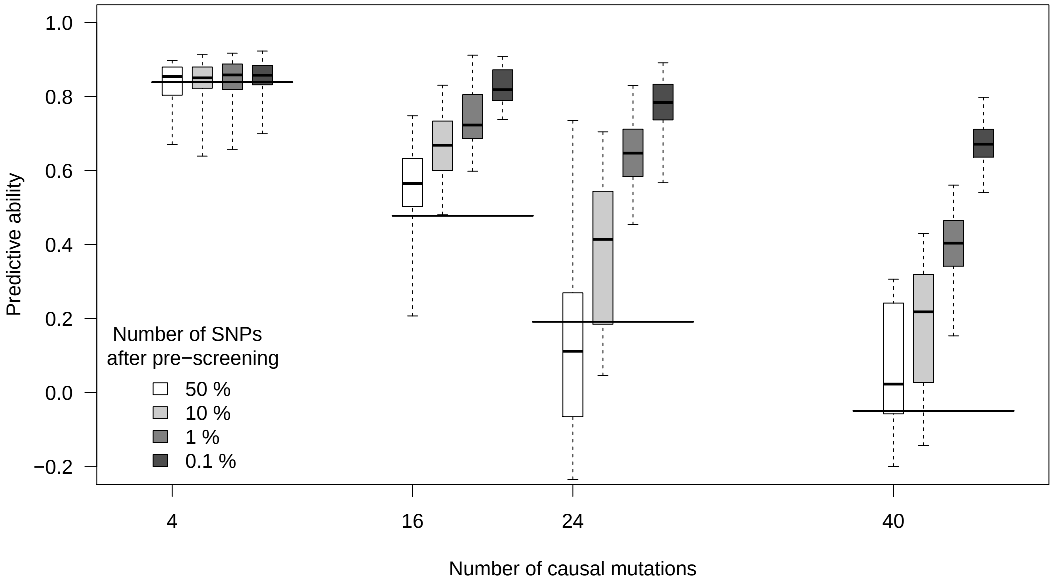

Figure 7: Prospects of pre-screening predictor variables with high-dimensional marker data. Data were simulated using $p = 250000$, $n = 200$, $h^2 = 0.75$, and varying p_0 and 10 replications per scenario using simulation procedure 1 of Wimmer *et al.* (2013). Boxplots display predictive ability estimated with fivefold cross-validation for scenarios where all true causal mutations were retained in the data but the number of non-causal loci was subsequently reduced in four steps from 50% to 0.10% in advance. Horizontal lines indicate the average predictive ability of LASSO without pre-screening predictor variables (see data in Table 3).

Next, the performance of SIS-LASSO was investigated under a scenario where it was not known in advance which subset of predictor variables comprised true zero coefficients. The sensitivity with respect to recovery of causal mutations was investigated for different values of p_0 and d for LASSO and SIS-LASSO. When $p_0 \leq 12$, the highest sensitivity was observed using LASSO without pre-screening (Figure 8). With more causal mutations, the performance of SIS-LASSO with $d = 2500$ was slightly increased, but at the expense of markedly more false positives (data not shown). An interesting measure was the minimum number of markers which need to be included such that all causal mutations were retained after pre-screening with SIS. With $p_0 = 4$, on average more than 90000 markers (36.5% of 250000) must be selected according to their marginal correlation on average before all four causal mutations were included (range: $104 - 238597$ SNPs). These numbers increased with increasing p_0.

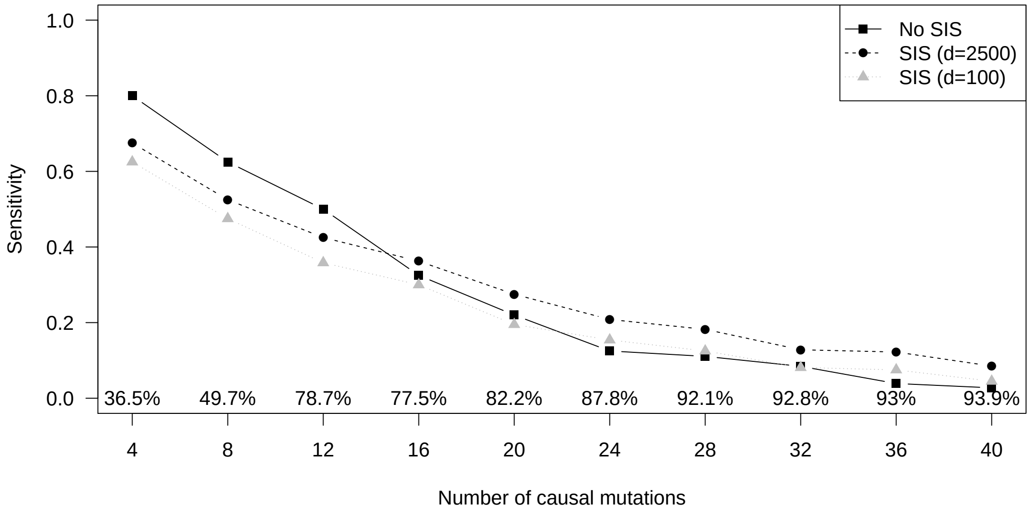

Figure 8: Sensitivity of LASSO and sure independence screening (SIS) followed by LASSO (using $d = 2500$ and $d = 100$). Numbers at the bottom indicate the minimum proportion of markers with largest marginal correlations which must be included such that all causal mutations were retained after pre-screening with SIS given the number of causal mutations. Data were simulated according to procedure 1 in Wimmer *et al.* (2013) using $p = 250000$, $n = 200$, and $h^2 = 0.75$.

For prediction, no advantage was observed when the number of predictor variables was reduced with SIS in the ES and then followed by an application of LASSO (see Table 3). With $p_0 = 4$ or $p_0 = 16$, a significant advantage of LASSO without pre-screening was observed. Differences vanished with increasing number of causal mutations but predictive ability also approached zero. It is worth mentioning that this decrease in predictive ability with increasing p_0 will not be as severe with experimental data compared to Table 3, because a certain level of predictive ability will be retained due to relatedness structures between training and testing data sets (see also Section 2.1.3).

The results in Figure 7 clearly illustrate the prospects of pre-screening predictor variables for prediction when analyzing whole-genome sequence data. Unfortunately, this potential cannot be retrieved by statistical procedures such as SIS because several pitfalls are associated with dimension reduction approaches for high-dimensional marker data. First, the

Table 3: Average predictive ability ± standard error from 10 replications with fivefold cross-validation for LASSO and sure independence screening followed by LASSO (SIS-LASSO); p_0: number of causal mutations; data were simulated according to procedure 1 in Wimmer *et al.* (2013) using $p = 250000$, $n = 200$, and $h^2 = 0.75$.

p_0	SIS-LASSO ($d = 100$)	SIS-LASSO ($d = 2500$)	LASSO
4	0.60 ± 0.01	0.77 ± 0.02	0.84 ± 0.01
16	0.26 ± 0.03	0.37 ± 0.02	0.48 ± 0.02
24	0.08 ± 0.02	0.14 ± 0.02	0.19 ± 0.05
40	-0.01 ± 0.04	0.02 ± 0.04	-0.05 ± 0.07

use of marginal correlations can be misleading when conditional correlations are important (Guyon and Elisseeff 2003). In particular, some predictor variables might be useful for prediction but only in connection with other predictor variables, for example in epistatic networks, while SIS considers only marginal correlations. Second, dimension reduction requires a sparse true model and will not be successful for complex traits because true nonzero coefficients are likely to be removed through any dimension reduction technique. This was confirmed by the results on the minimum number of predictor variables that must be included such that all causal mutations were retained after pre-screening with SIS.

In the literature, different pre-screening techniques have been employed. Scutari *et al.* (2013) used Markov blankets for pre-selecting predictor variables in conjunction with Ridge regression, LASSO, and the elastic net, but the authors did not observe an advantage compared to scenarios without pre-screening. For LASSO, there are approaches available that can pre-screen predictor variables based on their marginal association with the response variable, ensuring that almost certainly no true nonzero coefficient is removed (Tibshirani *et al.* 2012). In a study for human height, de los Campos *et al.* (2013b) used a similar approach compared to SIS and ranked markers according to their p-values from a GWAS. They observed a slight advantage when using only the highest-ranking SNPs for prediction. Thus, exploring pre-screening techniques to enhance the efficiency of genome-based prediction using whole-genome sequence data warrants further research.

2.7 Design of computer simulations

2.7.1 Encapsulating real marker data in computer simulations

Computer simulations are powerful tools for studying the efficiency of different statistical methods for genome-based prediction (Daetwyler *et al.* 2013). In Wimmer *et al.* (2013), a versatile simulation framework to investigate the efficiency of different statistical methods *in silico* was presented. Experimental data sets from three different plant species (rice, wheat, and *Arabidopsis*) were incorporated into the simulation scheme to encapsulate their LD structure and obtain realistic scenarios reflecting the LD structure of these experimental data sets. This provided the unique opportunity to control important parameters such as trait heritability or the number of QTL, as well as to explore scenarios relevant for real-life applications. Method comparisons based on this kind of simulation scheme will also be valuable in acquiring a more detailed picture of what can be expected when analyzing whole-genome sequence data. An interesting approach will be to integrate experimental whole-genome sequencing data into simulation procedure 3 in Wimmer *et al.* (2013), for example, in order to perform power calculations for the sample size required to identify a given number of causal mutations for a given trait heritability and determinedness level.

Multicollinearity is limiting the ability of statistical methods to recover true nonzero coefficients within high-dimensional marker data and computer simulations were a viable tool for assessing the influence of LD on the accuracy of estimated marker effects *in silico* (Wimmer *et al.* 2013). Different approaches were explored compared to the simulation scheme in Lehermeier *et al.* (2013) in order to generate LD, and, hence, correlations among markers. In Lehermeier *et al.* (2013), new genotypes were simulated, while in Wimmer *et al.* (2013) resampling of real genotypes was used to encapsulate the actual LD structure in the marker data of rice, wheat, and *Arabidopsis.* The former strategy was more flexible but it must be validated whether it mimics real data sets (Daetwyler *et al.* 2013). Simulations based upon real genotypes are limited to the specific data structures under study. The shortcoming of using a fixed sample size was circumvented by simulation procedure 3 in Wimmer *et al.* (2013) employing a Cholesky decomposition of the correlation structure of the SNP marker data to convey the LD structure of the real data to simulated data sets of arbitrary sample size. A similar approach was taken by

Wientjes *et al.* (2013) to investigate the influence of LD on the reliability of genome-based predictions. Hoerl *et al.* (1986) proposed a technique to arrive at predefined collinearity levels. Their approach will be useful when investigating the influence of LD in a more universal framework that is not restricted to specific experimental data sets but this is left for future studies.

2.7.2 Allele frequencies and effect distributions

Besides LD structure and the number of QTL, the distribution of QTL effects is expected to influence the efficiency of genome-based prediction methods (Coster *et al.* 2010). In Wimmer *et al.* (2013) and Lehermeier *et al.* (2013), QTL effects were assigned equal values across QTL or were sampled from a uniform distribution. Other simulation studies used Gamma (Meuwissen *et al.* 2001), Gaussian, or Laplace (Daetwyler *et al.* 2010) distributions to simulate QTL effects, and the choice of distribution might affect the results, although the difference between a Gaussian and a Laplace distribution was small in Daetwyler *et al.* (2010). Causal mutations in Wimmer *et al.* (2013) were sampled from the given marker loci in simulation procedure 2, but in nature QTL might have a different allele frequency spectrum than markers. In particular, some marker panels are known to exhibit an ascertainment bias of SNPs, with a tendency toward a uniform minor allele frequency distribution (Daetwyler *et al.* 2013). Simulations in Druet *et al.* (2014) revealed high accuracies of genome-based prediction only when the QTL had the same allele frequency spectrum as the SNPs. Modifications of the computer simulations in Wimmer *et al.* (2013) with different allele frequencies for the QTL are interesting topics for future research. Moreover, the independent assignment of QTL to marker loci can be extended toward assigning true nonzero coefficients within known pathways, in order to simulate epistatic effects.

2.8 Software implementation

As shown in Section 2.1, there is a great potential for applying genome-based predictions in plant breeding, and in Sections 2.3 and 2.4 the need for model assessment and exploring alternative methods and hyperparameter settings was demonstrated. Thus, a user-friendly and well-documented software package is crucial to facilitate the application and validation of different methods, in order to bring genome-based prediction from theory into practice. Such software was lacking and the `synbreed` R package (Wimmer *et al.* 2012) was the first open-source software offering a comprehensive analysis pipeline with the whole functionality to implement genome-based prediction in breeding programs within one software. The `synbreed` package covers the processing and coding of raw marker data, the estimation of genome-based similarity and pedigree-based relationship coefficients, the application and validation of different prediction methods, and the visualization of results (Figure 9). Such a pipeline is crucial for comparing effectively a large number of different settings and model specifications to maximize prediction performance. The package provides all the tools required to fit the GBLUP method, which is the benchmark method for many method comparisons. Moreover, implementations of Bayesian Lasso and Bayesian Ridge regression in the `BLR` package (Pérez *et al.* 2010) were embedded within the `synbreed` package. This allows researchers to conduct standard analyses, for example using linear mixed models or Bayesian methods as described in this thesis. The prediction performance of different methods can be compared through CV schemes. Data flow is streamlined by the special class of `gpData` objects ('genomic prediction `Data`') developed to facilitate genomic prediction analyses. This class resembles a generic data structure which is suitable for a wide range of statistical methods employing genotypic and phenotypic data such as genome-based prediction, GWAS, or QTL mapping. Once an object of class `gpData` is created, it can be efficiently stored and all further analysis steps rely on its structure in order to gear the different functions. This approach was innovative within software for genome-based prediction and enhances the reproducibility of results. Moreover, it is a step toward customized high-throughput analysis pipelines, which are required in large-scale applications of real breeding programs.

To summarize, the `synbreed` package provides a valuable tool within the plant and animal genetics researcher's software toolbox, and it is now an established tool in the plant breeding community after the release on CRAN in March 2012 (see Figure 10). Where necessary, the package provides gateways to other software packages to broaden the type of possible applications. This includes the `R/qtl` package (Broman *et al.* 2003) for QTL mapping, `Plink` (Purcell *et al.* 2007) for a GWAS, or `Beagle` (Browning and Browning 2009) for imputing missing values in the marker matrix. With the unified data object and several transformation tools, conversion and transfer effort between software packages is tremen-

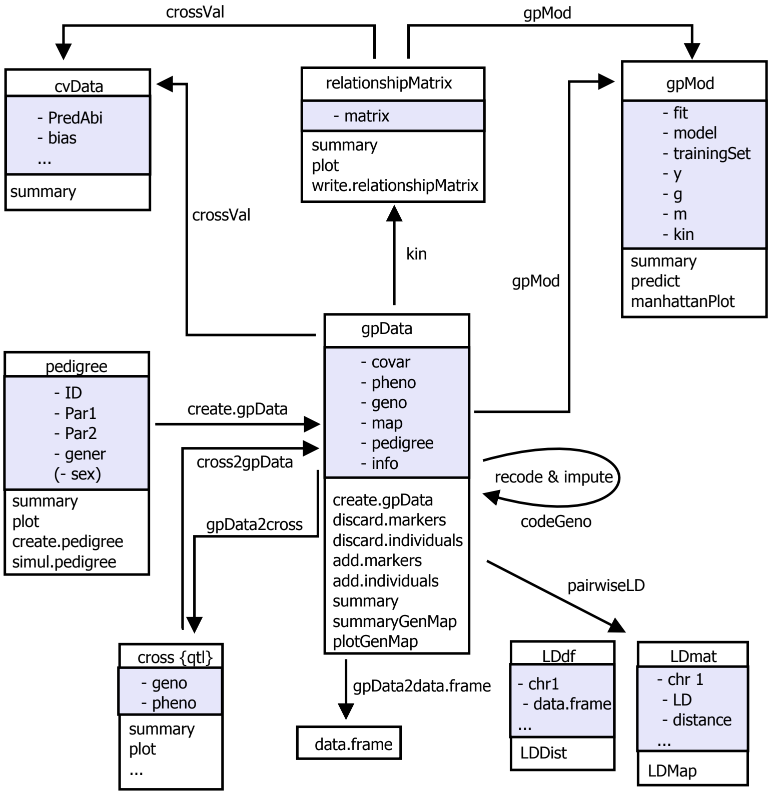

Figure 9: Overview of object classes, methods, and functions within the `synbreed` package (Wimmer *et al.* 2012). Each box indicates a class, together with class name, elements, and available functions and methods. The arrows indicate data flow while the origin indicates the input argument and the head is the return value of the function.

dously reduced and it is straightforward to have access to the functions implemented in these software packages. The design of the `synbreed` package is very flexible and covers special cases in plant breeding, such as repeated measurements of DH lines. This flexibility contributed to the fact that the package has been applied successfully in the public and private sectors with widespread applications and data from different crop, tree, and livestock species. Based on the download statistics of CRAN (`cran.r-project.org`), 2290 downloads and users from more than 100 different countries have been identified (see Figure 10). All code was implemented within the R language, therefore permitting users to customize the methods to their specific needs. The package has already been extended by the `impute.R` package to include genotype imputation and phasing using a reference panel of haplotypes (Y. Badke, personal communication; see `https://www.msu.edu/~steibelj/JP_files/vignette_impute1025.pdf`). The package was released together with a package vignette (available from `http://cran.r-project.org/web/packages/synbreed/vignettes/IntroSyn.pdf`), providing hands-on tutorials with example data sets (from the accompanying package `synbreedData`, `http://cran.r-project.org/web/packages/synbreedData/`). Thus, the package was able to fill the gap in the availability of user-friendly software for next-generation genetics research and application in plant breeding programs.

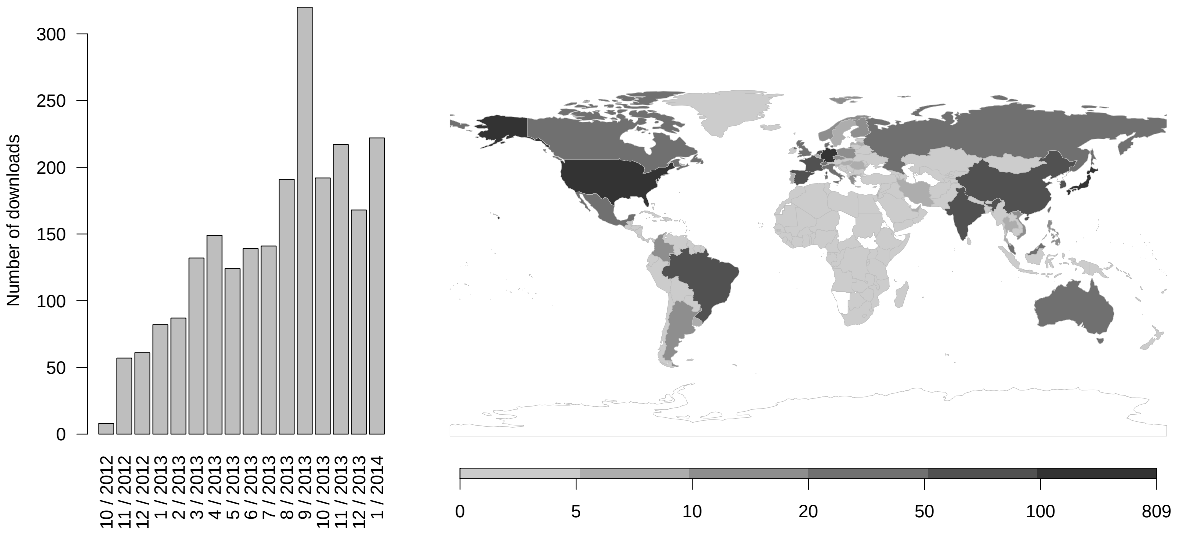

Figure 10: Left-hand side: number of downloads of the `synbreed` package from CRAN per month; right-hand side: number of downloads per country (last updated on 2014 - 01 - 31).

2.9 Conclusions

A series of novel results for genome-based prediction in plant breeding and new insights into the properties of statistical methods with high-dimensional marker data have been achieved by this thesis. The findings are valuable from both a practical and theoretical point of view. Here, the most important inferences from this work are summarized:

- Genome-based prediction can deliver accurate predictions of genotypic values for different quantitative traits and plant species. In particular, *predictive ability was higher compared to models based on pedigree data only.* These results are encouraging for the implementation of genome-based prediction into breeding programs.
- An important distinction has to be made between the tasks of variable screening and prediction of genotypic values. The information content in high-dimensional marker data was sufficient to predict genotypic values but mostly not sufficient to accurately estimate individual marker effects. These results suggest a paradigm change in genome-based prediction. It was originally envisaged that predictability stems from marker effects tagging QTL effects, but given the low accuracy of individual marker effects, *relatedness among individuals emerges as a major source of predictive ability.*
- *Under restrictive assumptions, variable selection methods can circumvent dimensionality* and successfully identify true nonzero coefficients. The most important assumptions are the existence of a sparse true model, the absence of strong correlations among markers, and a high trait heritability.
- *LASSO can be very efficient to pinpoint causal mutations within whole-genome sequence data*, but only when the number of mutations is considerably smaller than the sample size.
- Regularized regression and Bayesian methods are powerful prediction techniques that cope with overfitting problems in high-dimensional marker data through a bias-variance tradeoff. The amount of regularization is controlled by one or more tuning parameters. *Their influence on the results can be remarkable.* Bayesian Lasso and Bayesian Ridge regression were less influenced by the prior distribution compared to BayesA and BayesB.
- *GBLUP is a viable method with competitive predictive abilities in several experimental data sets.* The efficiency of this method was not affected by the genetic

architecture of the trait under study. For complex traits, Bayesian methods allowing for marker-specific prior variance components do not outperform methods assuming an equal contribution of all loci *a priori*.

- The `synbreed` R package *established a versatile analysis pipeline for genome-based prediction*, covering several methods presented in this thesis.

No general recommendation for a specific genome-based prediction method can be given, because results demonstrate that no method was consistently superior for all purposes. In general, a method that contributes satisfactory answers to one question might not be appropriate for answering another question. Consequently, the proper choice of method becomes an empirical question that must be answered case-by-case using model assessment techniques for experimental data such as CV. Nevertheless, the body of experimental studies, together with the computer simulations presented in this thesis, provided several guidelines which can be applied to appraise the efficiency of different statistical methods under various scenarios. For all methods, the proper choice and constant tuning of regularization parameters is crucial before employing these methods within routine applications for genome-based prediction.

3 References

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

Atwell, S., Huang, Y., Vilhjálmsson, B., Willems, G., Horton, M., *et al.* (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**: 627 – 631.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, New York, USA.

Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **7**: 889 – 890.

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, **846**: 210 – 223.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., *et al.* (2009). The genetic architecture of maize flowering time. *Science*, **325**: 714 – 718.

Bühlmann, P. and Mandozzi, J. (2013). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comp Stat*, doi:10.1007/s00180–013–0436–3.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer, Berlin/Heidelberg, Germany.

Carré, C., Gamboa, F., Cros, D., Hickey, J., Gorjanc, G., *et al.* (2013). Genetic prediction of complex traits: integrating infinitesimal and marked genetic effects. *Genetica*, **141**: 239 – 246.

Clark, S., Hickey, J., Daetwyler, H., and van der Werf, J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet Sel Evol*, **44**: doi:10.1186/1297–9686–44–4.

Colombani, C., Legarra, A., Fritz, S., Guillaume, F., Croiseau, P., *et al.* (2012). Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. *J Dairy Sci*, **96**: 575 – 591.

Coster, A., Bastiaansen, J. W. M., Calus, M. P. L., van Arendonk, J. A. M., and Bovenhuis, H. (2010). Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genet Sel Evol*, **42**: doi:10.1186/1297–9686–42–9.

Crossa, J., Pérez, P., Hickey, J., Burgueno, J., Ornella, L., *et al.* (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, **112**: 48 – 60.

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

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, **185**: 1021 – 1031.

de los Campos, G., Gianola, D., and Allison, D. B. (2010). Predicting genetic predisposition in humans: The promise of whole-genome markers. *Nat Rev Genet*, **11**: 880 – 886.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013a). Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics*, **193**: 327 – 345.

de los Campos, G., Naya, H., Gianola, D., José Crossa, A. L., Manfredi, E., *et al.* (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, **182**: 375 – 385.

de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y., and Sorensen, D. (2013b). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet*, **9**: e1003608.

Donoho, D. L. and Stodden, V. (2006). Breakdown point of model selection when the number of variables exceeds the number of observations. In *Proceedings of the International Joint Conference on Neural Networks*, 1916 – 1921.

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, **112**: 39 – 47.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed Sensing: Theory and Applications*. Cambridge University Press, Cambridge, UK.

Erbe, M., Gredler, B., Seefried, F. R., Bapst, B., and Simianer, H. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS ONE*, **8**: e81046.

Falconer, D. and Mackay, T. (1996). *Introduction to Quantitative Genetics*. Longman Technical, Harlow, Essex, UK.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J Roy Stat Soc B*, **70**: 849–911.

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans Roy Soc Edinb*, **52**: 399–433.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, **30**: 1 – 22.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, Second Edition*. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, **194**: 573 – 596.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**: 347 – 363.

Goddard, M., Hayes, B., and Meuwissen, T. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J Anim Breed Genet*, **128**: 409 – 421.

Goddard, M. E. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, **136**: 245 – 257.

Goddard, M. E., Wray, N. R., Verbyla, K., and Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. *Stat Sci*, **24**: 517 – 529.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J Mach Learn Res*, **3**: 1157 – 1182.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, **177**: 2389 – 2397.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**: doi:10.1186/1471–2105–12–186.

Habier, D., Tetens, J., Seefried, F., Lichtner, P., and Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in german Holstein cattle. *Genet Sel Evol*, **42**: doi:10.1186/1297–9686–42–5.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction, Second Edition.* Springer Series in Statistics. Stanford, California, USA.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci*, **92**: 433 – 443.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet*, **6**: e1001139.

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased accuracy of artifical selection by using the realized relationship matrix. *Genetics Research Cambridge*, **91**: 47 – 60.

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

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Canada.

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

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet*, **6**: 226 – 231.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**: 55 – 67.

Hoerl, R. W., Schuenemeyer, J. H., and Hoerl, A. E. (1986). A simulation of biased estimation and subset selection regression techniques. *Technometrics*, **28**: 369–380.

Huber, P. (1981). *Robust Statistics*. Wiley, New York, USA.

Ishwaran, H. and Rao, J. S. (2011). Generalized ridge regression: geometry and computational solutions when p is larger than n. Technical report, Cleveland Clinic and University of Miami, Miami, Florida, USA.

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, **9**: 166 – 177.

Jonas, E. and de Koning, D.-J. (2013). Does genomic selection have a future in plant breeding? *Trends in Biotechnology*, **31**: 497 – 504.

Kärkkäinen, H. P. and Sillanpää, M. J. (2013). Fast genomic predictions via Bayesian G-BLUP and multilocus models of threshold traits including censored Gaussian data. *G3*, **3**: 1511 – 1523.

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., *et al.* (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet*, **5**: e1000551.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer, New York, USA.

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

Lehermeier, C., Wimmer, V., Albrecht, T., Auinger, H.-J., Gianola, D., *et al.* (2013). Sensitivity to prior specification in Bayesian genome-based prediction models. *Stat Appl Genet Mol Biol*, **12**: 375 – 391.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, **5**: 151 – 170.

Li, Z. and Sillanpää, M. J. (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor Appl Genet*, **125**: 419 – 435.

Lorenz, A., Chao, S., Asoro, F., Heffner, E., Hayashi, T., *et al.* (2011). Genomic selection in plant breeding: Knowledge and prospects. *Advances in Agronomy*, **110**: 77 – 123.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, Massachusetts, USA.

Meuwissen, T., Hayes, B., and Goddard, M. (2013). Accelerating improvement of livestock with genomic selection. *Annual Review of Animal Biosciences*, **1**: 221 – 237.

Meuwissen, T. H. E. and Goddard, M. E. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*, **185**: 623 – 631.

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

Miller, A. (2002). *Subset selection in regression. Second Edition*. Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Myers, R. H. (1994). *Classical and Modern Regression with Applications*. PWS-Kent, Belmont, California, USA.

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., *et al.* (2012). Using whole genome-sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet*, **8**: e1002685.

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

Pérez, P., de los Campos, G., Cross, J., and Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the BLR package in R. *The Plant Genome*, **3**: 106 – 116.

Pérez-Cabal, M., Vazquez, A., Gianola, D., Rosa, G., and Weigel, K. (2012). Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Front Genet*, **3**: doi: 10.3389/fgene.2012.00027.

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., *et al.* (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**: 559 – 575.

R Development Core Team (2012). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J Am Stat Assoc*, **92**: 179 – 191.

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., *et al.* (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet*, **44**: 217 – 220.

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

Schön, C.-C., Utz, H. F., Groh, S., Truberg, B., Openshaw, S., *et al.* (2004). Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics*, **167**: 485 – 498.

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

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer, New York, USA.

Stuckart, C. (2012). *Methoden des Elastic Net zur sparsamen Variablenselektion und deren Anwendung in der Genetik*. Master's thesis, Department of Statistics, Ludwig-Maximilians Universtität München, Munich, Germany.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J Roy Stat Soc B*, **58**: 267 – 288.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., *et al.* (2012). Strong rules for discarding predictors in lasso-type problems. *J Roy Stat Soc B*, **74**: 245 – 266.

Vazquez, A. I., Rosa, G. J. M., Weigel, K. A., de los Campos, G., Gianola, D., *et al.* (2010). Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. *J Dairy Sci*, **93**: 5942 – 5949.

Waldron, L., Pintilie, M., Tsao, M.-S. S., Shepherd, F. A., Huttenhower, C., *et al.* (2011). Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, **27**: 3399–3406.

Wallace, J. G., Larsson, S. J., and Buckler, E. S. (2014). Entering the second century of maize quantitative genetics. *Heredity*, **112**: 30–38.

Weigel, K., de los Campos, G., González-Recio, O., Naya, H., Wu, X., *et al.* (2009). Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci*, **92**: 5248 – 5257.

Wientjes, Y. C. J., Veerkamp, R. F., and Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, **193**: 621 – 631.

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

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

Wolc, A., Arango, J., Settar, P., Fulton, J., O'Sullivan, N., *et al.* (2011). Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet Sel Evol*, **43**: doi:10.1186/1297–9686–43–23.

Wray, N. R., Yang, J., j. Hayes, B., Pricse, A. L., Goddard, M. E., *et al.* (2013). Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*, **14**: 507 – 515.

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., *et al.* (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa. Nature communications*, **467**: doi: 10.1038/ncomms1467.

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

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

4 Appendix

4.1 Publications

Many results obtained in this dissertation were already published in international, peer-reviewed journals. These publications including online supporting material are available from the following websites:

1. Albrecht *et al.* (2011): `http://link.springer.com/article/10.1007%2Fs00122-011-1587-7#page-1`
2. Wimmer *et al.* (2012): `http://bioinformatics.oxfordjournals.org/content/28/15/2086`
3. Lehermeier *et al.* (2013): `http://www.degruyter.com/view/j/sagmb.2013.12.issue-3/sagmb-2012-0042/sagmb-2012-0042.xml`
4. Wimmer *et al.* (2013): `http://www.genetics.org/content/195/2/573`

5 Acknowledgements

First of all, I would like to thank my academic supervisor Prof. Dr. Chris-Carolin Schön for her constant support, encouragement, and many inspiring ideas. Thank you for giving me the opportunity to work on this project.

Sincere thanks to Prof. Dr. Aurélien Tellier, Prof. Dr. Henner Simianer, and Prof. Dr. Hans-Rudolf Fries for serving on my graduate committee.

I would like to thank all co-authors, project partners, and in particular the whole plant breeding group of TUM for many interesting discussions and the enjoyable joint work. Many thanks to Nicole Krämer, Christina Lehermeier, Sandra Unterseer, and Raphael Wimmer for proofreading and suggestions to this thesis.

This research was made possible through financial support from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed - Synergistic plant and animal breeding" (FKZ 0315528A).

Finally, I would like to thank my family who always supported me in the last years.

6 Curriculum Vitae

Personal Information

Valentin Wimmer

Date of birth: January 7, 1985

Place of birth: Munich, Germany

Education

03/2010 – 10/2013: Doctorate in plant breeding, Chair of plant breeding, Technische Universität München, Germany

10/2007 – 11/2009: M.Sc. in Statistics, Department of Statistics, Ludwig-Maximilians-Universität München, Germany

10/2004 – 09/2007: B.Sc. in Statistics, Department of Statistics, Ludwig-Maximilians-Universität München, Germany

09/1995 – 06/2004: Secondary school, Deutsch-Herren-Gymnasium Aichach, Germany

Publications

Schön CC, Wimmer V (2014) Statistical models for the prediction of genetic values. Chapter in *Risk - A Multidisciplinary Introduction*, Springer (in print)

Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Schön CC (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, **195**: 573-587 [doi: 10.1534/genetics.113.150078]

Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. *Statistical Applications in Genetics* *and Molecular Biology*, **12**: 375-391 [doi: 10.1515/sagmb-2012-0042]

Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, **28**: 2086-2087 [doi: 10.1093/bioinformatics/bts335]

Albrecht T*, Wimmer V*, Auinger HJ, Erbe M, Knaak C, Ouzonova M, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. *Theoretical and Applied Genetics*, **123**: 339-350 [doi: 10.1007/s00122-011-1587-7]

Pyrka P*, Wimmer V*, Fenske N, Fahrmeir L, Schwirtz A (2011) Factor analysis in performance diagnostic data of competitive Ski-Jumpers and Nordic Combined athletes. *Journal of Quantitative Analysis in Sports*: 7, Iss. 3, Article 8. [doi: 10.2202/1559-0410.1300]

Wimmer V, Fenske N, Pyrka P, Fahrmeir L (2011) Exploring competition performance in decathlon using semi-parametric latent variable models. *Journal of Quantitative Analysis in Sports*: 7, Iss. 4, Article 6. [doi: 10.2202/1559-0410.1307]

* = equal contribution