MASTERARBEIT

Titel der Masterarbeit

A comparison of methods used in genomic selection

Verfasser

Dr.rer.nat. Ludwig Geroldinger

Wien, Februar 2015

Studienkennzahl lt. Studienblatt:	H 067 458
Studienrichtung lt. Studienblatt:	Individuelles Masterstudium, Ökologische Landwirtschaft
Betreuer:	Univ.-Prof. Dr. Johann Sölkner
Zweitbetreuer:	Dr. Gabor Meszaros

Acknowledgments

Foremost, I would like to thank my advisors, Professor Johann Sölkner and Dr. Gabor Meszaros, for the introduction to the exciting topic of genomic selection, as well as for the advice and liberty they gave me.

Having done my studies within the framework of an individual masters program, I want to thank Professor Werner Georg Nowak who guided the design of the program and provided essential support for the approval of the study proposal.

During my studies I have attended exciting lectures on recent developments in breeding technologies held by Professor Hermann Bürstmayr, Professor Clay Sneller and Professor Johann Sölkner. Further, I have gained from the seminars on theoretical population genetics held by Reinhard Bürger at the University of Vienna. I am grateful to all of them for the careful preparation of their classes and for stimulating discussions.

Finally, I would like to express my deep gratitude to my wife, my parents, and my friends who always believed in me and supported me.

Summary

Due to recent advances in sequencing technology, an increasing number of dense marker maps and fully sequenced genomes is becoming available for many populations in animal and plant breeding. In this thesis we study recently developed methods for the estimation of breeding values based on genomic data.

In the first part of the thesis, we present models which estimate genetic effects of markers, hence genetic breeding values, based on phenotypic records from single traits. These models differ in their assumptions on the genetic architecture of the trait, i.e., on the number of quantitative trait loci (QTLs) and on their effect-sizes. Although the differences in accuracy between the models are smaller than expected, we determine a strong model-dependent influence of the genetic architecture on the accuracy of breeding values. We show that Bayesian models usually perform better than linear mixed models if a few QTLs determine the trait, whereas the opposite may be true if many QTLs are underlying the trait. Further, we explore the influence of the density of markers, the heritability of the trait, and the number of phenotypic records on the performance of the methods.

In the second part of the thesis, we review multi-trait models. These models use the available data more efficiently than single-trait models by incorporating correlations between traits. The multi-trait models increase the accuracy of breeding values for low-heritability traits which are correlated to high heritability traits. Especially, if phenotypic records are missing for low-heritability traits, the use of multi-trait models is strongly recommended.

Zusammenfassung

Aufgrund der schnellen Weiterentwicklung biochemischer Methoden ist die Entschlüsselung von DNS Daten in den letzten Jahren sehr günstig geworden. Dadurch hat sich das Sequenzieren von Genomen als effizientes Hilfsmittel in der Pflanzen- und Tierzucht erwiesen. In dieser Arbeit besprechen wir Methoden zur Zuchtwertschätzung, welche auf hochauflösenden Marker-Arrays oder auf voll sequenzierten Genomen basieren.

Zuerst werden Modelle präsentiert, welche die Effekte von Genen, basierend auf phänotypischer Information an einem Merkmal, schätzen. Der Hauptunterschied zwischen den Modellen liegt in den unterschiedlichen Annahmen über die genetische Architektur des Merkmals, welche durch die Anzahl und durch die Effektgrößen der zugrundeliegenden Loci gegeben ist. Obwohl sich die Modelle weniger in deren Genauigkeit der Zuchtwertschätzung unterscheiden als man zunächst erwarten würde, konnte ein deutlicher Modell-abhängiger Effekt der genetischen Architektur auf die Genauigkeit der Zuchtwertschätzung festgestellt werden. Bayesianische Modelle treffen genauere Vorhersagen als gemischte Regressionsmodelle, wenn das Merkmal durch wenige Loci bestimmt wird. Falls allerdings das Merkmal von sehr vielen Loci beeinflusst wird, schätzen die Regressionsmodelle den Zuchtwert besser. Schließlich diskutieren wir den Einfluss der Anzahl an Markern, der Heritabilität und der Anzahl der phänotypischen Messungen auf die Genauigkeit der Modelle.

Im zweiten Teil der Arbeit, verallgemeinern wir die Modelle auf mehrere Merkmale. Diese Modelle nutzen die vorhandenen phänotypischen Informationen (meist) effektiver, da sie Korrelation zwischen den Merkmalen berücksichtigen. Die Genauigkeit der Zuchtwertvorhersagen profitiert hier besonders, wenn ein Merkmal mit kleiner Heritabilität zu einem Merkmal mit höherer Heritabilität korreliert ist. Auch wenn nur wenige phänotypische Messungen vorliegen, ist der Einsatz von Modellen mit mehreren Merkmalen gegenüber Modellen mit einem Merkmal vorteilhaft.

Contents

1 Introduction

Plant and animal breeding has a long tradition in human culture. Over the last few thousand years, domesticated plants and animals have been constantly selected to secure human nutrition. Artificial selection of breeding candidates is based on the evaluation of the genetic breeding value. Once breeding values are assigned to individuals, the individuals with highest breeding values are chosen for reproduction. The key task in breeding is the accurate assignment of breeding values which justifies the choice of selection candidates.

Over many centuries breeding values have been determined solely phenotypically, i.e., individuals with superior phenotypic values (e.g., yield, meat quality, disease resistance) were chosen for reproduction. This approach has two major problems. First, the phenotypic evaluation is often very time consuming and expensive. Second, low heritabilities of traits can lead to poor selective gains.

Recent developments in sequencing technology have made it possible to incorporate genetic information in the estimation of breeding values. The most prevalent sequencing technology is 'next-generation' sequencing which determines the exact order of nucleotide bases. Due to that sequencing technology, a large number of fully sequenced genomes has become available and the re-sequencing of complete genomes on the population scale is becoming increasingly popular. These developments make the estimation of genetic breeding values an achievable intent.

Fourteen years ago, Meuwissen et al. (2001) presented four statistical models to estimate breeding values from genetic data. These models are based on least-squares regression, best linear unbiased predictions (BLUP), and two similar Bayesian approaches. Due to their different assumptions on the genetic basis of the traits, the models have different statistical properties. Since real data usually does not satisfy all the assumptions of the models, there is still considerable uncertainty which of the methods (or their extensions in recent literature) is most appropriate. In general, this will depend on the genetic heritability and the genetic architecture of the trait under selection, as well as on the number and the quality of phenotypic and genotypic records.

In this thesis we present the models of Meuwissen et al. (2001) in a comprehensive way (Section 3) and relate them to classical animal breeding models which are based on pedigree information rather than on genetic data. In Section 4 we compare the accuracies of estimated breeding values between the models, and explore their dependence on factors such as the number of markers or the number of phenotypic records. We show that the differences in

accuracy are smaller than intuitively expected. However, the performance of the methods depends heavily on the genetic architecture, which influences BLUP and the Bayesian methods in opposite ways. Results from more recent literature which extend the methods of Meuwissen et al. (2001) are reviewed and the application of the models to real data is discussed. Whereas the methods in Section 3 and Section 4 are concerned with the estimation of breeding values reflecting fitness at a single trait, multi-trait models are presented in Section 5. These models perform more accurately, especially if traits are correlated and have low heritability.

Estimating breeding values from genomic information rather than pedigree information shortens the generation interval and is cheaper than phenotypic selection while maintaining high levels of accuracy (Meuwissen et al. 2001; Schaeffer 2006). Therefore, genomic selection is especially useful where phenotypic records are missing or are very expensive. Based on these advantages, we are confident that genomic selection will pave the way for a new era in plant and animal breeding. In the last few years genomic selection has already been successfully implemented in various cattle breeding programs.

Finally, we note that the methods presented in this theses are of special interest to organic agriculture. Organic farmers rely strongly on plants and animals which were selected to perform well without chemical fertilizers and antibiotics. Accounting for these aims in breeding programs, the term 'organic breeding value' has been shaped. The estimation of the 'organic breeding value' benefits from the methods presented below.

2 Basics

In this section we recapitulate the necessary terminology and concepts underlying the main parts of the thesis.

2.1 Genetic basics

Quantitative traits

Most traits under consideration in animal and plant breeding are quantitative, i.e., their phenotypic values can be measured on a continuous scale. Usually, these traits are determined by a large number of quantitative trait loci (QTLs). Despite of recent developments in sequencing technology, the QTLs influencing most traits are unknown and their determination is subject to ongoing research.

It is a statistical task to determine whether the variation in certain regions of the genome are associated with the phenotypic variation in the trait. If significant associations between a marker and trait variation are detected, the marker is linked to (at least) one QTL. A central concept for mapping QTLs is linkage disequilibrium.

Linkage disequilibrium

Linkage Disequilibrium (LD) is defined as the non-random association of alleles between two or more loci. In the literature different definitions of LD have been proposed to capture different aspects of that non-random association (see Slatkin (2008) for a review). In the following we give a classical definition of LD between two biallelic loci $\mathcal{A}$ and $\mathcal{B}$. For extensions to multiple loci with multiple alleles we refer to Bürger (2000). Let us denote the alleles at $\mathcal{A}$ by A and a, and those at $\mathcal{B}$ by B and b. The frequencies of the alleles A and B are denoted by p and q, respectively, and the frequencies of the four gametes, AB, Ab, aB, ab are designated x_1, x_2, x_3, x_4, respectively. One measure of LD is D, defined as

$$D = x_1x_4 - x_2x_3.$$

Therefore, D measures the difference of coupling genotypes AB/ab and repulsion genotypes Ab/aB. Since the measure D depends strongly on allele frequencies, many authors have investigated the squared correlation in allelic states

$$r^2 = \frac{D^2}{p(1-p)q(1-q)}, \tag{1}$$

which tends to exhibit 'better' statistical properties.

Linkage disequilibrium between loci is reduced by recombination. Let $D(t)$ denote the measure D at generation t, and let c denote the recombination rate between two loci. Then we have

$$D(t+1) = (1-c)D(t). \tag{2}$$

Factors which cause LD include selection, migration, and random genetic drift. In general, it is difficult to determine which of these factors was dominating in the generation of observed LD. In breeding however, random genetic drift may be the most important determinant of the average LD in the genome since population sizes are rather small. Migration can often be ruled out by the breeding design, and LD due to selection is mostly localized around specific genes.

Whereas in infinite panmictic populations, neutral loci will eventually be in linkage equilibrium (LE) (see eq. (2)), LD is maintained in finite populations. Sved (1971) showed that the expectation of r^2 in a finite population with an effective population size N is given by

$$\mathrm{E}[r^2] = \frac{1}{1+4Nc}. \tag{3}$$

In breeding, the population size N is mostly known and $\mathrm{E}[r^2]$ is usually observed. Therefore, one could infer the recombination rate c from (3). If the parameters $N, \mathrm{E}[r^2]$, and c do not satisfy (3), selection will most likely have shaped the genotype frequencies.

Phenotypic variances

In quantitative genetics, phenotypic values are usually decomposed into genetic and environmental components. Similarly, the phenotypic variance σ_p^2 of a trait is decomposed into its genetic component σ_g^2, its environmental component σ_e^2, and the interaction of both $\sigma_{g\times e}^2$, i.e.,

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + \sigma_{g\times e}^2 \tag{4}$$

For simplicity, we neglect the genotype environment interaction in this manuscript, i.e., we always assume $\sigma_{g\times e}^2 = 0$.

The genetic variance σ_g^2 can further be partitioned into a component which is due to the additive effects of the loci σ_a^2, a component emerging from dominance σ_d^2, and a component due to all kind of genetic interactions σ_i^2 (e.g., epistasis). Then, we have

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 + \sigma_i^2. \tag{5}$$

For rigorous definitions of the variance components we refer to Bürger (2000) or Lynch and Walsh (1998).

Further, we recall the commonly used definition of the (narrow-sense) heritability of a trait

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}, \tag{6}$$

which measures the proportion of the total phenotypic variance that can be explained by additive effects.

If multiple traits are under consideration, their genetic variances are often correlated. If the correlation between two traits is high, it is very likely that these traits are influenced by the same set of genes. Then selection on one of the traits also influences the second trait. Most commonly, genetic correlations between traits are summarized in the matrix $\mathbf{\Sigma}_g = (c^{(ij)})$ (often referred to as the **G**-matrix), where $c^{(ii)}$ is the additive genetic variance of trait i and $c^{(ij)}$ is the additive genetic covariance between trait i and trait j. Similarly to considering additive genetic covariances, one defines the environmental covariances $\mathbf{\Sigma}_e$ between traits (see Section 5).

2.2 Linear models

In the following we introduce general linear models and mixed models which are common tools from applied statistics for the estimation of breeding values. The presentation follows Lynch and Walsh (1998, Chapters 8 and 26), where further examples and applications can be found. Let p, q, and n be natural numbers. We denote the set of $p \times q$ matrices with real entries by $\mathbb{M}^{p,q}$ and the identity matrix by $\mathbf{I}_p \in \mathbb{M}^{p,p}$.

General linear models

We assume $n > p$. A general linear model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{7}$$

where

$\mathbf{y} \in \mathbb{M}^{n,1}$ is a vector of observations with mean $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$,

$\boldsymbol{\beta} \in \mathbb{M}^{p,1}$ is a vector of fixed effects which shall be inferred,

$\mathbf{X} \in \mathbb{M}^{n,p}$ is an incidence matrix, relating the observations $\mathbf{y}$ to $\boldsymbol{\beta}$,

$\mathbf{e} \in \mathbb{M}^{n,1}$ is a vector of random error terms with mean $\mathrm{E}(\mathbf{e}) = 0$ and the (invertible) covariance matrix $\mathbf{R} = \mathrm{Cov}(\mathbf{e}) \in \mathbb{M}^{n,n}$.

Since the underlying linear system in (7) is overdetermined, it usually has no solution in $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ has to be estimated. A common estimation-method for $\boldsymbol{\beta}$ is the method of least squares. The simplest version of it assumes that the residual effects are uncorrelated with the same variance, i.e., $\mathbf{R} = \rho \mathbf{I}_n$, where $\rho = \sigma_e^2$ will be the environmental variance (4) in all applications considered here. Then, the minimization of the (unweighted) sum of squared differences between the data values $\mathbf{y}$ and their corresponding modeled values is an unbiased estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2. \tag{8}$$

It can be easily shown that $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{9}$$

where for any matrix $\mathbf{U}$, $\mathbf{U}^T$ denotes the transposed matrix.

Example 1. Suppose that three sires are chosen randomly from a population and are mated to three randomly chosen damns. Two, one, and three offspring are evaluated from these matings, respectively. Let y_{ij} denote the phenotypic value of the jth offspring from the ith sire, μ denote the phenotypic mean of the population, s_i denote the effect of the ith sire, and e_{ij} the residual effects. We apply the general linear model (7) with $p = 4$, $n = 6$ to estimate the average effect of each sire

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \mu \\ s_1 \\ s_2 \\ s_3 \end{pmatrix}, \ \mathbf{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{31} \\ e_{32} \\ e_{33} \end{pmatrix}.$$

The estimator $\hat{\boldsymbol{\beta}}$ is then obtained from (9). Numerical values for the y_{ij} may be milk yield or indicators of meat quality.

Mixed models

A mixed model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{10}$$

where $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{e}$ are as in (7) and

- $\mathbf{u} \in \mathbb{M}^{q,1}$ is a vector of random effects which is independent of $\mathbf{e}$, with mean $\mathrm{E}(\mathbf{u}) = 0$ and with (invertible) covariance matrix $\mathrm{Cov}(\mathbf{u}) = \mathbf{G} \in \mathbb{M}^{q,q}$,

$\mathbf{Z} \in \mathbb{M}^{n,q}$ is an incidence matrix, relating the observations $\mathbf{y}$ to $\mathbf{u}$.

The mixed model (10) partitions the residual effects from (7), $\mathbf{e}^*$, into two components, $\mathbf{e}^* = \mathbf{Zu} + \mathbf{e}$. Both models yield the same estimations of $\boldsymbol{\beta}$, however, the mixed model incorporates estimations of random effects in the vector $\mathbf{u}$. The variables $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$ are known and the variables $\boldsymbol{\beta}$ and $\mathbf{u}$ have to be inferred. Here, we assume that the covariance matrices $\mathbf{R}$ and $\mathbf{G}$ are known. However, usually these have to be inferred in a separate step; see Lynch and Walsh (1998, Chapter 27), by methods such as ANOVA or REML. In the context of breeding, the vector $\boldsymbol{\beta}$ typically contains the population mean, elements of population structure, gender and treatment, whereas the vector $\mathbf{u}$ usually accounts for genetic effects. The matrices $\mathbf{X}$ and $\mathbf{Z}$, also called design matrices, often contain only 0 and 1 indicating whether the corresponding effect is contributing to the phenotype.

For historical reasons, we refer to inferences of fixed effects as estimates, and to inferences of random effects as predictions. A variety of methods has been developed to infer the estimates and predictions in (10). We present and use the approach of Henderson, who found the best linear unbiased estimator (BLUE) $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and the best linear unbiased prediction (BLUP) $\hat{\mathbf{u}}$ for $\mathbf{u}$ for (10). Assuming that $\boldsymbol{\beta}$ and $\mathbf{u}$ are independent, we infer $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ as the solutions of the following linear system

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \tag{11}$$

Alternatively, one can show that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \tag{12a}$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{12b}$$

where

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} \tag{12c}$$

is the covariance matrix of $\mathbf{y}$. Although equation (12) looks nicer than (11), the computation of $\mathbf{V}^{-1}$ is usually computationally more demanding than the computation of $\mathbf{G}^{-1}$ and $\mathbf{R}^{-1}$.

Example 2. Suppose that three randomly chosen sires are mated with randomly chosen dams. Two offspring from each mating are evaluated. However, in contrast to Example 1, some offspring experience different treatments than others. Let y_{ijl} denote the phenotypic value of the lth offspring from the ith sire with treatment j. In order to compensate for different treatments we apply the general mixed linear model (10) to estimate the effect of the

sires on their offspring. Let us assume that the first offspring of the first and the third sire was assigned to treatment 1, whereas the second offspring of the first and the third sire was assigned to treatment 2. Further, both offspring of the second sire were assigned to treatment 1. This yields the following vectors of phenotypic observations, incidence matrices of fixed and random effects, and the vectors of these effects, respectively

$$\mathbf{y} = \begin{pmatrix} y_{111} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{311} \\ y_{322} \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \ \mathbf{u} = \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix}.$$

The BLUE values $(\hat{\beta}_1, \hat{\beta}_2)$ and the BLUP values for the sire effects $(\hat{s}_1, \hat{s}_2, \hat{s}_3)$ can be obtained from equations (11) or (12). A numerical example is given by

$$\mathbf{y}^T = (9, 12, 11, 6, 7, 14), \ \mathbf{G} = 2\mathbf{I}_6, \ \mathbf{R} = 6\mathbf{I}_6,$$

wherefore

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 8.22 \\ 13.06 \end{pmatrix} \text{ and } \hat{\mathbf{u}} = \begin{pmatrix} -0.056 \\ 0.111 \\ -0.056. \end{pmatrix}.$$

2.3 Bayesian statistics

In Section 3 we will use the following theorem of Bayes,

$$P(x|y) = P(y|x)\frac{P(x)}{P(y)}, \tag{13}$$

where

x is a hypothesis which is effected by data y,

$P(x)$ is the probability of x before y is observed, also called the prior probability,

$P(y)$ is the probability of y independently of the hypothesis x,

$P(x|y)$ is the probability of x given y, also called the posterior probability,

$P(y|x)$ is the probability of y given x, also known as the likelihood.

Since $P(y)$ is independent of x and therefore often considered to be constant, one also writes

$$P(x|y) \propto P(y|x)P(x).$$

In statistics, (13) is the foundation of a field called Bayesian inference which is often applied in the analysis of genomic data. The key point of (13) is that it is often easier (but still difficult) to calculate $P(y|x)$ than $P(x|y)$.

Example 3. We consider a set of coins, where 99% are fair coins and 1% are two-headed coins. One flips a coin three times and obtains three heads. What is the probability that the coin was two-headed?

Let y denote the observation that each flip results in a head and let x be the probability that the coin was two-headed. Then we get $P(x) = 0.01$, $P(y|x) = 1$. Let further x^* be the probability that the coin was fair, $P(x^*) = 1 - P(x) = 0.99$ and $P(y|x^*) = 0.5^3 = 0.125$. Because of $P(y) = P(x)P(y|x) + P(x^*)P(y|x^*)$, we obtain

$$P(x|y) = \frac{1 * 0.01}{(1 * 0.01) + (0.99 * 0.125)} = 0.075.$$

In practice, one often often applies Bayes' rule iteratively. It is convenient to make an educated guess for the prior, i.e., the distribution of $P(x)$ and, subsequently, perform an 'experiment' to determine the posterior distribution, i.e., the distribution of $P(x|y)$. This posterior distribution may serve as a prior for the next step. In general, determining the posterior distribution is difficult, since the likelihood is unknown. For some priors, e.g., for conjugate priors, analytical results are available. Conjugate prior distributions have the property that their posterior distributions belong to the same family of probability distributions if the likelihood function fulfills certain properties. For example, normal distributed priors yield normal distributed posteriors if the likelihood function is Gaussian.

A common numerical tool for inferring the posterior are Markov chain Monte Carlo algorithms. In the following we describe Gibbs sampling, a simple, commonly used Markov chain Monte Carlo algorithm. Gibbs sampling generates a sequence of observations from a multivariate probability distribution f where direct sampling is difficult.

Assume that $f(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$ is known $(1 \leq j \leq n)$ but $f(x_1, \ldots, x_n)$ is unknown. In order to obtain l samples from $f(x_1, \ldots, x_n)$, we choose an initial value $(x_1^{(0)}, \ldots, x_n^{(0)})$ and proceed iteratively. Let $(x_1^{(i)}, \ldots, x_n^{(i)})$ denote the ith sample. The variable $x_j^{(i)}$ is sampled from the conditional distribution

$$f(x_j|x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, , x_n^{(i-1)}), \tag{14}$$

i.e., we sample each variable from its distribution conditioned on all other variables. This recurrently updates all variables.

In order to obtain l independent samples $(x_1^{(i)}, \ldots, x_n^{(i)})$, $1 \leq i \leq l$, it is common practise to ignore a number of samples at the beginning and then consider only every ith sample. The reason for this is that (i) it takes some iterations to reach the stationary distribution of the underlying Markov chain and (ii) subsequent samples are not independent but exhibit correlations.

2.4 Genome wide association studies

Genome wide association studies (GWAS) examine the correlation of genetic variability with phenotypic variance. These studies compare two groups of individuals with different phenotypes, e.g., individuals infected by a disease with healthy individuals. Both groups are either fully sequenced or genotyped using arrays of single-nucleotide polymorphisms (SNPs). If a marker is more frequent in one of the groups, one may conclude that this marker is associated with a QTL influencing the trait which is differing between the groups.

For some traits, e.g., human height, GWAS have found hundreds of SNPs that are significantly associated with the trait (Yang et al. 2010). However, in most cases the detected SNPs only explain a small proportion of the heritability of the trait (missing heritability problem).

Hypothesis that could potentially resolve the missing heritability problem are subject to ongoing investigation and debate. Two main hypotheses, which are opposed in Gibson (2012), are:

A The SNPs used in GWAS explain all genetic variance, but the effect of each single SNP is too small to be detected.

B The QTLs are not in perfect LD with any of the SNPs and are therefore undetected.

Hypothesis A suggests that an increase in sample size would resolve the missing heritability problem, whereas hypotheses B demands an increase in the number of markers. A further hypothesis is:

C Substantial amounts of quantitative genetic variation are due to epistasis (non-additive gene-interaction), compromising the wide-spread assumption of additivity.

If hypothesis C was true, we would have to estimate the effects of haplotypes including SNPs which are not necessarily neighbored. This would be much more demanding than the estimation of effects at single SNPs.

In the following we present a simple method of testing for associations between markers and phenotypes. We consider a sample of size n from a population. If m markers are available,

the association of each marker i with a phenotype can be tested by using the linear models

$$\mathbf{y} = \mu \mathbf{1}_n + \mathbf{X}_i \mathbf{g}_i + \mathbf{e}, \quad 1 \leq i \leq m, \tag{15}$$

where $\mu = \frac{1}{n} \sum_l y_l$ is the mean phenotypic value and $\mathbf{1}_n$ denotes a column vector with n 1s. The $\mathbf{g}_i = (g_{i1}, \ldots, g_{ik_i})^T$ is a vector representing the k_i genetic effects at marker i. Considering a biallelic SNP with alleles A and a, we may estimate the effects of the three genotypes AA, Aa, and aa, hence $k_i = 3$ (or $k_i = 2$ with appropriate scaling). If we are estimating a single effect only, the $\mathbf{g}_i = g_i$ would be a scalar. With $\mathbf{y}^* = \mathbf{y} - \mu \mathbf{1}_n$ and $\boldsymbol{\beta}_i = \mathbf{g}_i$, the model in (15) is equivalent to the model in (7).

The null-hypotheses is that the marker is not associated with the trait. If the p-value obtained from the estimator $\hat{\mathbf{g}}_i$ is less than the predetermined significance level α we reject the null-hypothesis and conclude that the marker is associated with the trait. The power of this test depends on

(i) the number of genotyped individuals n,

(ii) the magnitude of LD between the marker loci and QTLs,

(iii) the effect sizes of the QTLs on the trait,

(iv) the allele frequency distribution at the marker locus,

(v) the significance level α.

Testing the association of a single marker with the trait for large amounts of markers (up to one million) with the model in (15) is problematic due to the multiple-testing problem. Conservative corrections for the multiple-testing problem, such as the Bonferoni correction do not take into account that the pairwise tests are not independent, since a pair of markers is usually not in LE. Therefore, testing the association of haplotypes (instead of SNPs) with the phenotype, or fitting of all markers simultaneously is usually a superior method. The latter method has been applied successfully by Yang et al. (2010), who explained 45% of the variance in human height. We will consider this method in more detail in the next chapter. It has been proposed for genomic selection by Meuwissen et al. (2001) and has been successfully applied ever since.

Finally, we note that missing information on population structure is a further severe problem in GWAS. If not accounted for, population structure yields a significant increase in false positives. If one has knowledge on the structure of the population, one can use mixed models instead of linear models to account for it (compare Example 1 and Example 2).

3 Genomic selection for single traits

A traditional way to use marker information for breeding is to perform GWAS (with the model in (15)) and use the markers linked to beneficial genes for marker assistant selection. As indicated in the previous section, a superior method would be to include all marker information in one statistic. This idea was first explored by Meuwissen et al. (2001) and led to the concept of genomic selection which we present in this section. Recent reviews on further developments can be found in Lorenz et al. (2011), Hayes and Daetwyler (2013), Heslot et al. (2012), Hayes et al. (2009a), and Mrode (2014).

The procedure of genomic selection requires a dense marker coverage of the whole genome. Then, the genome is divided into chromosome segments to which a genomic value can be assigned. Conceptually, genomic selection is performed in two steps. First, the genomic values of chromosome segments are determined in a reference population. Second, genomic estimated breeding values for selection candidates not in the reference population will be determined. Here, we outline the methods for the first step by estimating effects of chromosome segments based on phenotypic information from a single trait. Aspects of the second step are discussed in Section 4.

Once the genomic values ($\mathbf{g}_i$) of chromosome segments have been estimated, the genomic estimated breeding value (GEBV) of an individual is given by

$$\text{GEBV} = \sum_{i=1}^{m} \mathbf{X}_i \mathbf{g}_i, \tag{16}$$

where m denotes the number of markers, $\mathbf{g}_i = (g_{i1}, \ldots, g_{ik_i})^T$ is the vector of genotypic effects at marker i and $\mathbf{X}_i \in \mathbb{M}^{1,k_i}$ are entries of a design matrix which assigns markers to individuals.

The difficulty of estimating the $\mathbf{g}_i$ in a reference population emerges from the fact that, usually, we have many more markers than individuals. In the following we present four different approaches to estimate the $\mathbf{g}_i$.

3.1 Least-squares estimation

In the least-squares (LS) estimation, marker effects are treated as fixed effects (cf. Section 2.2). The method proceeds in two steps.

First, we repeat the single marker regression from (15) in a group of n individuals. Based on this analysis, we select the m most significant markers (out of the $\tilde{m}$ initial markers), where $m < n$. This could be done by considering the log-likelihood of every marker, i.e., we take

the markers that satisfy

$$\log(L(\mathbf{y}|\hat{\mathbf{e}})) = -\frac{1}{2}\left(n\ \log(\hat{\sigma}_e^2) + \frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{\hat{\sigma}_e^2}\right) > \alpha, \quad 1 \leq i \leq \tilde{m},$$

where $\log(.)$ denotes the natural logarithm, $\hat{\mathbf{e}}$ is an estimate for the vector of environmental effects, $\hat{\sigma}_e^2 = \hat{\mathbf{e}}^T\hat{\mathbf{e}}/(n - \text{rank}([\mathbf{1}_n\mathbf{X}_i]))$ is an estimate of the error variance, and α is a sufficiently stringent significance level so that $m < n$.

In the second step, the effects of the non-significant markers are set to zero and a multiple regression over all significant markers is performed to infer the $\mathbf{g}_i = (g_{i1}, \ldots, g_{ik_i})^T \in \mathbb{M}^{k_i,1}$:

$$\mathbf{y} = \mu\mathbf{1}_n + \sum_{i=1}^{m} \mathbf{X}_i\mathbf{g}_i + \mathbf{e}, \tag{17}$$

where $\mathbf{X}_i = (X_{iuv}) \in \mathbb{M}^{n,k_i}$. The model (17) is equivalent to a linear model of the form in (7), i.e.,

$$\mathbf{y} - \mu\mathbf{1}_n = \mathbf{X}\mathbf{g} + \mathbf{e}, \tag{18}$$

where

$$\mathbf{X} = \begin{pmatrix} X_{111}, \ldots, X_{11k_1}, \ldots, X_{i11}, \ldots, X_{i1k_i}, \ldots X_{m11}, \ldots X_{m1k_m} \\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ X_{1n1}, \ldots, X_{1nk_1}, \ldots, X_{in1}, \ldots, X_{ink_i}, \ldots X_{mn1}, \ldots X_{mnk_m} \end{pmatrix} \in \mathbb{M}^{n,k}, \tag{19a}$$

$$\mathbf{g} = (g_{11}, \ldots, g_{1k_1}, \ldots, g_{i1}, \ldots, g_{ik_i}, \ldots g_{m1}, \ldots g_{mk_m})^T \in \mathbb{M}^{k,1}, \tag{19b}$$

and $k = \sum_{i=1}^{m} k_i$. Therefore, the results derived for (7) can be applied. Some authors use the notation of (18), whereas we stick to (17).

The procedure of LS exhibits two major problems. First, the choice of the significance level α is problematic. It has to be stringent, since the the single marker regression requires $m < n$. Otherwise the approach of LS cannot be performed. Further, due to multiple testing, the effect sizes will usually be overestimated.

3.2 BLUP

Here, we treat the marker effects as random effects (cf. Section 2.2) and estimate them by incorporating the mixed model (10) with $\mathbf{X} = \mathbf{1}_n$, and $\boldsymbol{\beta} = (\mu, \ldots, \mu)$, wherefore $\mathbf{X}\boldsymbol{\beta} = \mu\mathbf{1}_n$ and

$$\mathbf{y} = \mu\mathbf{1}_n + \sum_{i=1}^{m} \mathbf{Z}_i\mathbf{g}_i + \mathbf{e} \tag{20a}$$

$$= \mu\mathbf{1}_n + \mathbf{Z}\mathbf{g} + \mathbf{e}, \tag{20b}$$

where $\mathbf{Z}$ or $\mathbf{g}$ are taken by merging the $\mathbf{Z}_i$ or $\mathbf{g}_i$ analogously to (19). We assume that the marker effects are independently and identically normally distributed with variance $\sigma_g^2 = \sigma_{\mathbf{g}_i}^2$ $(1 \leq i \leq m)$, i.e., $\mathbf{G} = \sigma_g^2 \mathbf{I}_m$. Further, we assume that the residuals are independently and identically normally distributed with variance σ_e^2, i.e., $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. We set $\lambda = \sigma_e^2/\sigma_g^2$ and obtain from (11) that

$$\begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_n^T \mathbf{1}_n & \mathbf{1}_n^T \mathbf{Z} \\ \mathbf{Z}^T \mathbf{1}_n & \mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{I}_n \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_n^T \mathbf{y} \\ \mathbf{Z}^T \mathbf{y} \end{pmatrix}. \tag{21}$$

In the following we give a short example which we adapted from Hayes and Daetwyler (2013).

Example 4. Let us consider five individuals and each individual has been genotyped for ten biallelic markers with alleles A and a. The genotypes AA, Aa, and aa are encoded with 0, 1, and 2 at each marker, respectively. We set $k_i = 1$ and only estimate the (additive) effect of allele a. The entry y_i (component of $\mathbf{y}$) gives the phenotypic value of the ith individual, g_j (component of $\mathbf{g}$) denotes the effect of allele a at the jth marker, and z_{ij} (entry of $\mathbf{Z}$) gives the value of the jth marker at the ith individual. We choose the following numerical values

$$\mathbf{y} = \begin{pmatrix} 0.19 \\ 1.23 \\ 0.86 \\ 1.23 \\ 0.45 \end{pmatrix}, \ \mathbf{Z} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 0 & 2 \\ 1 & 0 & 0 & 1 & 1 & 1 & 2 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 1 \end{pmatrix}, \ \lambda = \sigma_e^2/\sigma_g^2 = 1,$$

and obtain from (21) that

$$\hat{\mu} = 0.47, \ \ \hat{\mathbf{g}} = (0.29, -0.05, -0.05, 0.08, -0.02, 0.13, 0.13, -0.08, 0.11, -0.08)^T. \tag{22}$$

Therefore, the estimated effect of one copy of allele a at the first marker is 0.29, and so on. In the following we use this information to predict the breeding value for two selection candidates. We set $\tilde{\mathbf{Z}} = (\tilde{z}_{ij})$, where $\tilde{z}_{ij}$ connects the jth marker and the ith selection candidate. Let us assume that

$$\tilde{\mathbf{Z}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 2 & 1 & 0 & 1 \end{pmatrix}.$$

Then, equations (16) and (22) yield the GEBVs of the two selection candidates $\mathrm{GEBV}_1 = 0.47$ and $\mathrm{GEBV}_2 = 0.57$.

As mentioned in Section 2.2, σ_e^2 and σ_g^2 have to be inferred by independent methods; see Lynch and Walsh (1998). Usually, the BLUP method performs better than LS, since multiple testing and significance levels do not yield complications; see Section 4.

Variations of the BLUP method have been proposed and are concerned with incorporating different forms of regularizations to prevent overfitting. The most commonly used regularization methods are the Tikhonov regularization (or ridge regression), the Lasso method, and the elastic net method (see Ogutu et al. (2012) for a comparison). These methods yield different estimators for $\hat{\mu}$ and $\hat{\mathbf{g}}$ by posing different penalizations on the marker effects. We note that there exists a wide variety of possible regularizations, e.g., Schulz-Streeck and Piepho (2010) proposed geostatistical methods where genetic distances are treated analogously to spatial distances in geostatistics.

3.3 Bayesian method A

The following method shall relax the assumption on the distribution of marker effects made by the BLUP method. There it was assumed that the effects of the markers are identically (normally) distributed. For many traits there is considerable evidence that this is not true, and that traits are often determined by a few loci with large effects and many with small effects. The following Bayesian approach is accounting for these genetic architectures and allows us to incorporate assumptions on the underlying genetics of the trait.

In the following we present a hierarchical model at two levels. We create one model at the level of marker effects and a second model at the level of variances at each marker. The first step is similar to BLUP. We assume that the marker effects at position j are independently normally distributed with variance $\sigma^2_{\mathbf{g}_j}$ and infer μ and $\mathbf{g}$ from the model

$$\mathbf{y} = \mu \mathbf{1}_n + \mathbf{Z}\mathbf{g} + \mathbf{e}. \tag{23}$$

In contrast to Section 3.2 the variances for each segment j may differ, i.e., we do not assume $\sigma^2_{\mathbf{g}_i} = \sigma^2_{\mathbf{g}_j}$ for $i \neq j$. In order to estimate $\hat{\mathbf{g}}_i$, we need some further information on these variances. In contrast to the BLUP method where the $\sigma^2_{\mathbf{g}_i}$ were assumed to be known, here, we infer the $\sigma^2_{\mathbf{g}_i}$ and the $\mathbf{g}_i$ simultaneously.

In the second step, the distribution of the variances of marker effects is modeled by a Bayesian approach (see Section 2.3). Following Meuwissen et al. (2001), we assume a scaled inverted chi-square distribution as a prior for the distribution of variances of marker effects,

$$\text{prior}(\sigma^2_{\mathbf{g}_i}) = \chi^{-2}(\nu, S), \tag{24}$$

where S is a scaling parameter and ν denotes the degrees of freedom. One can show that this is a conjugate prior, i.e., that the resulting posterior distribution is also an inverted chi-square

distribution if the data is normally distributed. It is given by

$$\text{post}(\sigma^2_{\mathbf{g}_i}|\mathbf{g}_i) = \chi^{-2}(\nu + k_i, S + \mathbf{g}_i^T\mathbf{g}_i), \tag{25}$$

where k_i is the number of possible effects at marker i (if a single effect is estimated, we have $k_i = 1$).

The choice of the prior is in accordance with the results of Hayes and Goddard (2001), who studied the distribution of marker effects. Meuwissen et al. (2001) used the parameters $v = 4.012$ and $S = 0.002$, which were determined numerically to fit the mean and variance of the prior to the simulated distribution of marker effects under mutation-drift balance. As mentioned in Section 2.3, there is no such thing as a true prior but many choices could be suitable; see Xu (2003) for the choice of a different prior.

We cannot use (25) directly to infer $\sigma^2_{\mathbf{g}_i}$, since the $\sigma^2_{\mathbf{g}_i}$ depend on the unknown effects $\mathbf{g}_i$ and vice versa. Therefore, Meuwissen et al. (2001) used Gibbs sampling to estimate $\mathbf{g}_i$ and $\sigma^2_{\mathbf{g}_i}$ from (23) and (25). In order to apply the algorithm (see Section 2.3), we further assume that the prior distribution of the error variance (assuming $\sigma^2_e = \sigma^2_{\mathbf{e}_i}$ for $1 \leq i \leq n$) is given by

$$\text{prior}(\sigma^2_e) = \chi^{-2}(-2, 0), \tag{26}$$

which yields the posterior

$$\text{post}(\sigma^2_e|\mathbf{e}_i) = \chi^{-2}(n - 2, \mathbf{e}_i^T\mathbf{e}_i). \tag{27}$$

The algorithm proceeds as follows:

1. We choose initial values for the $\mathbf{g}_i = (g_{i1}, \ldots, g_{ik_i})^T$ and μ.

2. We sample the $\sigma^2_{\mathbf{g}_i}$ from (25).

3. Given $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)^T$ and μ, we calculate $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{g} - \mathbf{1}_n$ and sample σ^2_e from (27).

4. We sample the mean μ from the normal distribution

$$\mathcal{N}\left(\frac{\mathbf{1}_n^T\mathbf{y} - \mathbf{1}_n^T\mathbf{X}\mathbf{g}}{n}; \frac{\sigma^2_e}{n}\right), \tag{28}$$

   which includes the information on σ^2_e from step three.

5. Given the values for μ, σ^2_e, and $\sigma^2_{\mathbf{g}_i}$, we sample the effects $\mathbf{g}_i$ from the normal distribution

$$\mathcal{N}\left(\frac{\mathbf{X}_{ij}^T\mathbf{y} - \mathbf{X}_{ij}^T\mathbf{X}\tilde{\mathbf{g}}_{ij} - \mathbf{X}_{ij}^T\mathbf{1}_n\mu}{\mathbf{X}_{ij}^T\mathbf{X}_{ij} + \lambda_i}; \frac{\sigma^2_e}{\mathbf{X}_{ij}^T\mathbf{X}_{ij} + \lambda_i}\right), \tag{29}$$

where $\mathbf{X}_{ij}$ is the column of $\mathbf{X}$ corresponding the jth effect at marker i, $\tilde{\mathbf{g}}_{ij}$ is equal to $\mathbf{g}$ except that the jth effect at marker i is set to zero, and $\lambda_i = \sigma_e^2/\sigma_{\mathbf{g}_i}^2$. We note that the distributions (28) and (29) are of the form (14).

One repeats these steps for a large number of cycles until the sampled distributions of $\mathbf{g}_i$ and $\sigma_{\mathbf{g}_i}^2$ have converged. In Meuwissen et al. (2001) the algorithm ran for 10000 cycles, and the first 1000 cycles were discarded as a burn in.

3.4 Bayesian method B

The following adaptation of Bayesian method A considers the fact that in reality there are many markers with no genetic variance. The prior (24) in Bayesian method A, does not have a peak at $\sigma_{\mathbf{g}_i}^2 = 0$, but instead is infinitesimally small. We adapt the Bayesian method A by using the following prior:

$$\sigma_{\mathbf{g}_i}^2 = 0, \qquad \text{with probability } \pi, \tag{30a}$$

$$\sigma_{\mathbf{g}_i}^2 = \chi^{-2}(\nu, S), \qquad \text{with probability } 1-\pi. \tag{30b}$$

The Gibbs sampling algorithm as described in Section 3.3 can not be used with the prior (30), since the sampling of $\sigma_{\mathbf{g}_i}^2 = 0$ from (25) is impossible if $\mathbf{g}_i^T \mathbf{g}_i > 0$. Therefore, the Metropolis-Hasting algorithm is proposed for sampling $\sigma_{\mathbf{g}_i}^2$ and $\mathbf{g}_i$ simultaneously; see Meuwissen et al. (2001) for details.

Meuwissen et al. (2001) used the parameters $v = 4.234$ and $S = 0.0429$ in (30), which were determined numerically to fit the mean and variance of the simulated distribution of marker effects conditioned on $\sigma_{\mathbf{g}_i}^2 > 0$. Assuming mutation-drift balance at all markers, $1-\pi$ can be determined as the expected proportion of segregating QTLs, using classical results from theoretical population genetics (Appendix of Meuwissen et al. 2001).

Further developments

For practical applications it may not be very accurate to infer the parameters ν, S, and most importantly π from neutrality, since background selection or demographic bottlenecks may have influenced these parameters. Recent studies have taken that into account and improved various aspects of Bayesian method A and B.

Jia and Jannink (2012) introduced an version of the Metropolis-Hasting algorithm which re-estimates ν and S from the data rather than considering it to be fixed. Verbyla et al. (2009) combined Bayesian method B with stochastic search variable selection (SSVS) which

is an efficient algorithm for identifying SNPs with positive effects and, therefore, implicitly estimates π. A different approach was chosen by Habier et al. (2011) whose models, Bayesian method Cπ and Bayesian method Dπ, estimate π together with the variance of segregating markers. These models perform slightly better than the Bayesian method B, but their main advantage is the increase in computational performance.

3.5 Relation to standard animal models

In the following we elucidate the connection between the above models and the classical animal model

$$\mathbf{y} = \mu \mathbf{1}_n + \tilde{\mathbf{Z}}\mathbf{a} + \mathbf{e}, \tag{31}$$

where $\mathbf{a} = (a_i)$ and a_i is the additive genetic effect of individual i. Similarly to (20), the model (31) is a special case of (10), but now the vector of random effects consists of phenotypic additive effects instead of marker effects. Incorporating the standard assumption on the variance of residual effects $\mathrm{Cov}(\mathbf{e}) = \mathbf{R} = \sigma_e^2 \mathbf{I}_n$ and recalling (12), the solution of (31) depends only on the covariance matrix $\mathrm{Cov}(\mathbf{a})$ of additive effects.

It can be easily shown that the additive genetic covariance between individuals i and j is given by $2\theta_{ij}\sigma_a^2$, where θ_{ij} is the probability that a particular allele in individual i is identical by descent to a particular allele in individual j. Therefore, the covariance of additive effects in the animal model (31) is given by

$$\mathrm{Cov}(\mathbf{a}) = \sigma_a^2 \mathbf{A}, \text{ where } \mathbf{A} = (2\theta_{ij}).$$

The matrix $\mathbf{A}$ denotes the classical relationship matrix which can be determined from pedigree information.

Following Habier et al. (2007) we show that the coefficients of genetic relationships (given by $\mathbf{A}$) are closely related to marker information (given by the incidence matrix $\mathbf{X}$ in (18) or $\mathbf{Z}$ in (20) and (23)). Let $\mathbf{Z}$ denote the incidence matrix and $\mathbf{z}_i^T$ the ith row of $\mathbf{Z}$. Then the element ij of $\mathbf{Z}\mathbf{Z}^T$ is given by $\mathbf{z}_i^T \mathbf{z}_j$. Treating $\mathbf{z}_i^T \mathbf{z}_j$ as a random variable, its expected value is given by

$$\mathrm{E}[\mathbf{z}_i^T \mathbf{z}_j] = \sum_{l=1}^{m} \mathrm{E}[z_{il} z_{jl}] = \sum_{l=1}^{m} \mathrm{Cov}[z_{il} z_{jl}] + \mathrm{E}[z_{il}]\,\mathrm{E}[z_{jl}],$$

where m is the number of markers. Defining the $z_{il} \in \{0, 1, 2\}$ as in Example 4, we obtain that $\mathrm{E}[z_{il}] = 2p_l$, where p_l is the allele frequency at the lth locus. Further, we have $\mathrm{Cov}[z_{il} z_{jl}] =$

$4\theta_{ij}p_l(1-p_l)$ and obtain

$$\mathrm{E}[\mathbf{z}_i^T\mathbf{z}_j] = 4\theta_{ij}\sum_{l=1}^{m} p_l(1-p_l) + 4\sum_{l=1}^{m} p_l^2.$$

Consequently,

$$\mathrm{E}[\mathbf{Z}\mathbf{Z}^T] = \mathbf{A}\left(2\sum_{l=1}^{m} p_l(1-p_l)\right) + 4\ \mathbf{1}_n\mathbf{1}_n^T\sum_{l=1}^{m} p_l^2,$$

hence $\mathrm{E}[\mathbf{Z}\mathbf{Z}^T]$ is proportional to $\mathbf{A}$ modulo a constant.

Since the matrix $\mathrm{E}[\mathbf{Z}\mathbf{Z}^T]$ is proportional to the covariance matrix of marker effects $\mathbf{G} = \mathrm{Cov}(\mathbf{g})$, the matrix $\mathbf{G}$ is proportional to $\mathbf{A}$ modulo a constant. If the number of markers approaches infinity, $\mathbf{Z}\mathbf{Z}^T$ converges to $\mathrm{E}[\mathbf{Z}\mathbf{Z}^T]$. Therefore, for infinitely many markers the BLUP-model (20) is equivalent to the classical animal model (31) (the constant is incorporated into the mean μ).

This result holds independently of the LD between markers and QTLs. If markers are in LD with QTLs, the matrix $\mathbf{Z}\mathbf{Z}^T/\sum_{l=1}^{m} 2p_l(1-p_l)$ includes variation in relationships (e.g., between full sibs) and, therefore, provides more information about the covariance between relatives than $\mathbf{A}$. Also for the least squares method and the Bayesian methods, the incidence matrix $\mathbf{Z}$ reflects genetic relationships, wherefore the accuracy of GEBVs is positive even if all markers are in LE with the QTLs.

4 Accuracy of methods in genomic selection

Here, we present and review various investigations which compare the accuracy of the methods presented in Section 3 under different scenarios. Except for the discussion in Section 4.4, the results assume dense marker arrays rather than full sequence data.

4.1 Results from Meuwissen et al. (2001)

In Meuwissen et al. (2001) the four methods presented in Section 3 were compared by applying them to simulated data. The simulation model was set up as follows. A finite neutrally evolving population with effective population size $N = 100$ evolved until it reached mutation-drift balance. The genome had a length of 1000 centi-Morgan (cM) and a marker each cM. In the middle of each pair of markers a QTL was located. The mutation rates at the QTLs and at the marker loci were $u_1 = 2.5 * 10^{-5}$ and $u_2 = 2.5 * 10^{-3}$, respectively. Mutational effects were drawn from a gamma distribution with shape parameter $\beta = 0.4$ and scale parameter 1.66. The heritability of the trait was fixed to $h^2 = 0.5$.

The population was simulated for 1000 generations, after which it reached mutation-drift balance. In generations 1001 and 1002 the population size was expanded to 2000, and genotypic and phenotypic values were recorded. The individuals from generation 1003 were supposed to be juveniles (not having phenotypic records) whose breeding values were estimated with the methods from Section 3. The true breeding value (TBV) of each individual was recorded in the simulation. Table 1 shows the correlation r_{BV} of the TBV and the GEBV as well as the regression b_{BV} of the TBV on the GEBV. Both, r_{BV} and b_{BV}, are measures for the accuracy of genomic selection. The regression b_{BV} would be equals 1 if the estimations were unbiased.

	r_{BV}	b_{BV}
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
Bayesian method A	0.798	0.827
Bayesian method B	0.848 ± 0.012	0.946 ± 0.018

Table 1: The correlation r_{BV} of the TBV and the GEBV and the regression b_{BV} of the TBV on the GEBV for the different methods presented in Section 3. The means and standard deviations were taken over five replicates (which is rather low). For BLUP, the value $\sigma^2_{\mathbf{g}_i} = 0.0028$ was used, whereas $1 - \pi = 0.053$ was used in Bayesian method B. (The values are taken from Table 2 in Meuwissen et al. (2001).)

The LS method performs most poorly, which is probably due to the poor detection of QTLs (cf. Figure 1 in Meuwissen et al. 2001) and the high false positive rate. BLUP and the Bayesian methods resulted in a rather high accuracy but the Bayesian methods performed slightly better.

We recall that BLUP assumes the same variance of the distribution of marker effects at all markers, whereas the Bayesian methods incorporate different variances at each marker. Therefore, it seems surprising that in despite of these very different assumptions, BLUP, Bayesian method A and Bayesian method B exhibit rather similar performance.

In some sense, the results in Table 1 give the optimal accuracy of the methods, because parameters such as the mutation rate, the variance of genetic and environmental effects could be obtained from the simulation design, whereas in practical applications they have to be estimated.

4.2 The methods with real data

For real data, even smaller differences between the accuracy of BLUP and the accuracy of the Bayesian methods than in Table 1 are often observed. For example, Verbyla et al. (2009) compared BLUP, Bayesian method A and a variant of Bayesian method B (SSVS) for the relative and absolute protein and fat content of meat in cattle. Approximately 1500 bulls were genotyped for 39000 markers and their phenotype was the average performance of their offspring for the trait. The GEBVs of the bulls were compared with the phenotypically evaluated breeding values (PEBVs). The results are displayed in Table 2.

	Protein kg		Fat kg		Protein %		Fat %	
	r_{BV}	b_{BV}	r_{BV}	b_{BV}	r_{BV}	b_{BV}	r_{BV}	b_{BV}
BLUP	0.60	1.06	0.56	0.99	0.66	0.89	0.65	0.93
Bayesian method A	0.57	1	0.53	0.86	0.64	1	0.72	0.86
Bayesian method B (SSVS)	0.58	0.99	0.56	0.9	0.67	0.97	0.74	0.87

Table 2: The correlation r_{BV} of the PEBV and the GEBV and the regression b_{BV} of the PEBV on the GEBV for the different methods for real data from 1800 bulls. (The values are taken from Table 3 in Verbyla et al. (2009).)

The relative fat content is the trait with the highest deviations between different methods. This is in accordance with evidence that this trait is determined mainly by one mutation with large effect, which contradicts the assumption of BLUP that all marker effects are identically distributed. Therefore, BLUP shrinks the effect of this mutation which results in a loss of

accuracy. The Bayesian methods are more flexible and can capture the concentrated genetic architecture more accurately.

4.3 Factors determining the accuracy

Above we have seen that the methods presented in Section 3 perform rather similarly, both in simulations and with real data. In this section we will explore the underlying reason, but also discuss many general aspects determining the performance of all methods. The accuracy of genomic selection may be defined in several ways. As in Section 3, here, we measure accuracy by the correlation of the true to the estimated breeding value r_{BV}.

Although we will not give further details on other measures of accuracy, we note that in praxis the true breeding value is unknown (it can only be determined in simulation studies) and, therefore, our definition of r_{BV} may be of limited use. Then, the definition of accuracy depends on the chosen predicant which may be individual phenotypic values or average offspring performance, or something else.

Next, we discuss the influence of the following factors on the accuracy of genomic selection:

(i) the level of LD between markers and QTLs,

(ii) the number of markers capturing genetic relationships,

(iii) the number of phenotyped individuals in the reference population and the heritability of the trait,

(iv) the genetic architecture of the trait,

(v) the divergence between breeding and reference population, and

(vi) the time between estimation and application of marker effects.

(i) Linkage disequilibrium between markers and QTLs

In order to obtain accurate genomic breeding values at least one marker shall be in LD with each QTL. The higher the average LD along the genome, the less markers are required.

For the results shown in Table 1, one marker was positioned at each cM (see text above Table 1). Table 3 shows the change in accuracy if there would have been marker distances of 2cM and 4cM. For BLUP and Bayesian method B, as expected, the accuracy declines as the number of markers decreases. In the LS method the accuracy slightly increases with less

markers, which is probably due to the fact that less effects have to be estimated with less markers, which reduces the problem of overfitting.

Table 3 and Table 5 do not show results for Bayesian method A, since Bayesian method A has high computational costs and is supposed to perform worse than Bayesian method B in any case.

	Marker Spacing (cM)		
	1	2	4
LS	0.318	0.354	0.363
BLUP	0.732	0.708	0.668
Bayesian method B	0.848	0.810	0.737

Table 3: The accuracy of genomic selection (measured by r_{BV}) for different marker spacings in the simulation study of Meuwissen et al. (2001) which was described in Section 4.1. For BLUP with the marker spacings of 1cM, 2cM, and 4cM, the values $\sigma_g^2 = 0.0028$, $\sigma_g^2 = 0.0056$, and $\sigma_g^2 = 0.0112$ were used for the genetic variances, respectively. For Bayesian method B, $1 - \pi = 0.053$, $1 - \pi = 0.106$, and $1 - \pi = 0.212$ were used for the different marker spacings. (The values are taken from Table 4 in Meuwissen et al. (2001).)

Table 3 shows that high accuracies can be obtained if a marker is positioned every cM, i.e., two neighbored markers recombine at rate $c = 0.01$. Then, the expected r^2 (see eq. (3)) between two neighbored markers is $\mathrm{E}[r^2] = 0.2$ (assuming an effective population size of $N = 100$ as in Meuwissen et al. (2001); see Section 4.1). This is in accordance with the findings of Calus et al. (2008), who showed that an average value of $r^2 \approx 0.2$ is sufficient to predict genomic breeding values accurately.

In Meuwissen et al. (2001) the prediction individuals were offspring of the training individuals. Meuwissen (2009) showed that that if the prediction individuals and the training individuals are unrelated, the number of required markers for accurate genomic predictions is much higher. It is given by $10NL$, where L is the length of the genome in Morgans. Therefore, with $N = 100$ and $L = 10$ (as assumed in Meuwissen et al. 2001) ten markers per cM would be necessary if the prediction individuals and the training individuals were unrelated.

In Holstein Friesian cattle the effective population size is $N \approx 100$ and the length of the genome is $L \approx 30$. Following Meuwissen (2009), $10NL \approx 30000$ markers should be sufficient to predict genomic breeding values accurately. Complementary, we could aim for $\mathrm{E}[r^2] = 0.2$ between neighbored markers, following the result of Calus et al. (2008). In Holstein Friesian cattle a r^2 of 0.2 occurs every 100kb (kilobases). Since the genome has an approximate length of 3000Mb (megabases), we also conclude that 30000 markers are required for accurate

genomic predictions.

(ii) The number of markers and genetic relationships

As pointed out in Section 3.5, a higher number of markers does not only increase the probability of associations with QTLs, but also captures the genetic relationships more accurately.

The different methods differ in their capability to capture these relationships, since the fitted number of markers in LE varies between the methods. Habier et al. (2007) analyzed the average number of markers in LE which are captured by the models (Table 4). The least squares method fits only a few markers and captures genetic relationships the least. Bayesian method B includes more markers in LE, but much less than the BLUP method which captures all markers. We note that the results for Bayesian method B depend on the choice of the parameter $1 - \pi$, i.e., the proportion of markers which are assumed to have positive variance.

	Number of markers in LE		
	100	1000	2000
LS	1.9 ± 1.0	7.4 ± 2.8	11.0 ± 3.0
BLUP	100	1000	2000
Bayesian method B	12.6 ± 2.0	20.3 ± 3.1	21.4 ± 2.3

Table 4: The average number of markers in LE captured by the methods LS, BLUP, and Bayesian method B. The standard deviation was obtained from 96 replicates. For Bayesian method B the parameter $1 - \pi$ was set to the number of QTLs (ten QTLs were assumed) divided by the number of markers in LE, i.e., $1 - \pi = 0.1$, $1 - \pi = 0.01$, and $1 - \pi = 0.02$ in the left, middle, and right column, respectively. (The values are taken from Table 2 in Habier et al. (2007).)

Whereas the BLUP method reflects the relationships in the population most accurately (Table 4), the overall accuracy is often highest with the Bayesian methods (Table 3), since the Bayesian methods utilize information on LD better. Zhong et al. (2009) found that there is usually a trade-off between the capability of models in capturing the LD between markers and QTLs, and their capability of capturing the genetic relationships.

(iii) The number of phenotyped individuals and the heritability of the trait

How does the number of phenotypic records in the initial experiment influence the accuracy of GEBV? The results shown in Table 1 assume 2000 phenotypic records. Table 5 shows that the accuracy reduces by 67%, 21%, and 17% for the methods LS, BLUP, and Bayesian method B, respectively, if the number of records decreases from 2000 to 500.

	Phenotypic Records		
	500	1000	2000
LS	0.124	0.204	0.318
BLUP	0.579	0.659	0.732
Bayesian method B	0.708	0.787	0.848

Table 5: The accuracy of genomic selection (measured by r_{BV}) for different numbers of phenotypic records in the simulation study of Meuwissen et al. (2001). We recall from Section 4.1 that $h^2 = 0.5$. (The values are taken from Table 3 in Meuwissen et al. (2001).)

In Meuwissen et al. (2001) and Table 5 the heritability of the trait was set to $h^2 = 0.5$. In a more recent study, Goddard (2009) derived analytical results determining the influence of the heritability on the accuracy in dependence on the number of phenotypic records. Assuming that the QTL effects are normally distributed and that the number of QTLs Q is equal to the number of independent chromosome segments M_e, he derived that with BLUP

$$r_{\mathrm{BV}} = \sqrt{1 - \frac{\lambda}{2n\sqrt{a}} \frac{\log(1 + a + 2\sqrt{a})}{1 + a - 2\sqrt{a}}}, \tag{32a}$$

where

$$a = 1 + \frac{2\lambda}{n}, \quad \lambda = \frac{M_e k}{h^2}, \quad k = \frac{1}{\log(2N)}, \quad M_e = Q = 2NL, \tag{32b}$$

h is the heritability of the trait (6), n is the number of phenotyped individuals in the reference population, N is the effective population size in the breeding population, and L is the length of the genome in Morgans.

Equation (32) shows that large reference populations are necessary for high accuracy if the heritability of the trait is small; see Figure 1. Although (32) is based on the assumptions of normally distributed QTL effects, (32) predicts the accuracy of genomic selection well, both with simulated data (Hayes et al. 2009a) and with real data (Hayes et al. 2009b).

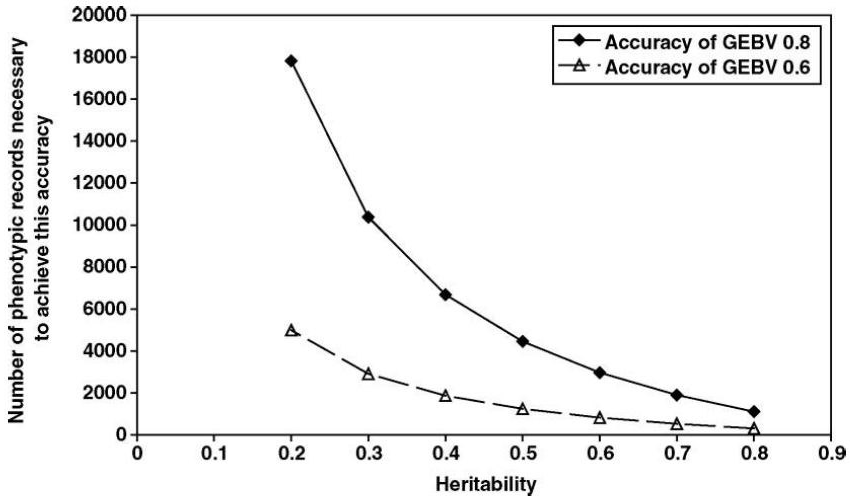

Figure 1: The number of phenotypic records necessary to achieve an accuracy of 0.6 (△) or 0.8 (⧫) for the GEBV. The plot is based on equation (32), which assumes that the QTL effects are normally distributed with the same variance at all positions, i.e., the BLUP method. The population size was $N = 1000$ and $L = 29$. (The Figure is taken from Hayes et al. (2009a).)

(iv) The genetic architecture of the trait

Since different assumptions on the genetic architecture of the trait are made in BLUP and the Bayesian models, the genetic architecture will influence the relative accuracy of the methods. In the BLUP model we assume that the effects at all markers are normally distributed with the same variance (Section 3.2), whereas the Bayesian methods are more flexible by incorporating different distributions of effect sizes at markers (Sections 3.3 and 3.4). In the following we elucidate the impact of these assumptions.

Daetwyler et al. (2010) used an approach similar to Hayes et al. (2009a), for determining the accuracy of BLUP and Bayesian method B. For the BLUP model they found that

$$r_{\mathrm{BV}} = \sqrt{\frac{nh^2}{nh^2 + M_e}}, \tag{33}$$

where M_e is the number of independent chromosome segments. In practical applications M_e is hard to determine. Incorporating different assumptions on the coalescent process, Goddard (2009) and Hayes et al. (2009b) suggested that $\frac{2NL}{\log(4NL)} \leq M_e \leq 2NL$. The numerical study of Daetwyler et al. (2010) suggests that $\frac{2NL}{\log(4NL)}$ is the better approximation (Table 5 in Daetwyler et al. 2010). If $M_e \approx 2NL$, equation (33) approximates (32). An important conclusion from (33) is that the accuracy in BLUP does not depend on the number of QTLs Q (Figure 2). In contrast, it was assumed in (32) that $Q = M_e$.

Daetwyler et al. (2010) conjectured that if Q is sufficiently large, the accuracy of Bayesian method B should converge to the accuracy of BLUP. Indeed, they could show that for Bayesian method B

$$r_{\mathrm{BV}} = \sqrt{\frac{nh^2}{nh^2 + \min(Q, M_e)}}. \tag{34}$$

Therefore, the Bayesian method performs better than BLUP if and only if if the number of QTLs is smaller than the number of independent chromosome segments, i.e., $Q < M_e$ (Figure 2).

The validity of the analytical approximations (33) and (34) was tested with simulations. It was confirmed that the accuracy of BLUP is independent of the number of QTLs Q (Figure 2). For Bayesian method B the accuracy is highest with low Q and decreases as Q increases. For large Q the accuracy with Bayesian method B reaches a plateau which is a bit lower than the accuracy of BLUP. (Equations (33) and (34) predict the same accuracy for $Q \geq M_e$.) All lines in Figure 2 increase with the number of phenotypic records n and the heritability h^2 (results not shown). The critical number of QTLs where the accuracy of BLUP and Bayesian method B are identical (at the intersections of solid and dashed lines in Figure 2) also increases with n, N, and h^2. If Q is very large ($Q \gg M_e$), the difference in accuracy between the two methods decreases to zero as n goes to infinity.

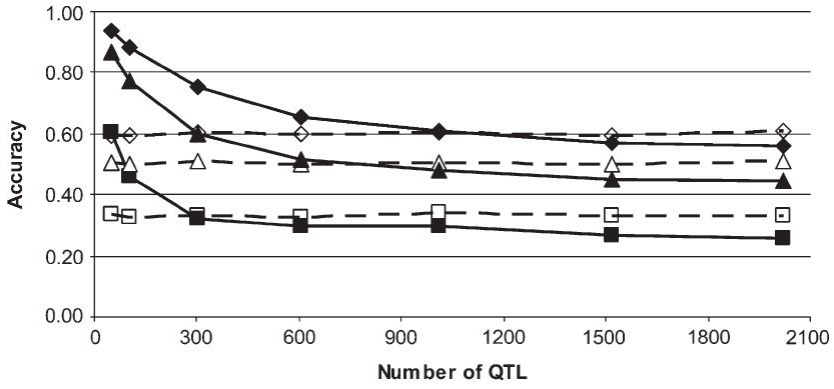

Figure 2: The accuracy of genomic selection r_{BV} for the BLUP model (empty symbols) and Bayesian method B (full symbols). Heritabilities are $h^2 = 0.1$ ($\square$), $h^2 = 0.3$ ($\triangle$), and $h^2 = 0.5$ ($\Diamond$). The number of independent chromosome segments was $M_e = 1887$. The parameter π for Bayesian method B was simply inferred from the simulation. (The Figure is taken from Daetwyler et al. (2010).)

(v) Genomic selection across breeds

Genomic selection is always applied to populations different to the reference population in which the marker effects were estimated. Often, the breeding population even origins from different lines or breeds than the reference population. Furthermore, the age and the environment may differ between the two populations.

The accuracy of genomic selection relies on the LD between the markers and the QTLs. Increasing divergence between the two populations reduces the probability that predictions on LD persist. One possibility to measure the divergence between populations is the average correlation of LD between breeds, which will increase with an increasing number of markers. De Roos et al. (2008) investigated the minimum number of markers between cattle breeds to obtain an average correlations of 0.9 between r^2 values. For example, they found that between Jersey and Angus cattle 300 000 markers are necessary. We note that this is a factor ten higher than the estimation we presented in (i).

A further problem is that the effect sizes of QTLs may differ between breeds. Recently, a series of examples has been documented where the number of QTLs and their effects differed between Holstein and Jersey cattle (e.g., Hayes et al. 2009a). Compensating for that, the idea was brought up to use multi-breed reference populations in breeding programs (Pryce et al. 2011; Erbe et al. 2012). However, these studies found very limited increase in the accuracy of genomic predictions when multi-breed reference populations were used instead of single-breed reference populations.

Finally, we note that the presence of genotype by environment interactions may limit genomic predictions across breeds.

(vi) How often shall the breeding value be re-estimated?

Linkage disequilibrium between the markers and the actual QTLs decays over time due to recombination; see equation (2). Therefore, it is necessary to re-estimate the effect of each marker after some generations. If a marker would coincide with a QTL this would not be necessary for short evolutionary timescales. With an increasing number of markers, the time period between re-estimations can be reduced.

In the study of Meuwissen et al. (2001) it was shown that the accuracy of estimated breeding values was strongly reduced after five generations. It was suggested to re-estimate breeding values every third generation (see Table 6). Whereas Table 6 shows the decay in accuracy only for Bayesian method B, it was shown in Habier et al. (2007) that the decay in accuracy is faster with BLUP than with the Bayesian methods.

Generation	r_{BV}
1003	0.848
1004	0.804
1005	0.768
1006	0.758
1007	0.734
1008	0.718

Table 6: Accuracy (measured by r_{BV}) for successive generations after the estimation of breeding values in generation 1002. Results are based on the simulation study described in Section 4.1 for Bayesian method B. (The values are taken from Table 5 in Meuwissen et al. (2001).)

In De Roos et al. (2008) this issue was investigated for real data from Holstein bulls with 20000 markers. The authors found that after two generations, accuracy is reduced by a factor of 0.1, which is consistent with Table 6. Muir (2007) showed that the accuracy persists longer if the reference population is genotyped for more generations rather than for a single generation.

4.4 Full sequence data vs. marker data

We have seen that the accuracy of genomic selection increases with the number of markers, because an increasing number of markers captures genetic relationships and the LD more accurately. However, it is not immediately clear, whether full sequence data will further increase the accuracy compared to SNP arrays which already contain sufficiently many markers so that several SNPs are in high LD with each QTL.

Meuwissen and Goddard (2010) tackled this question with simulations and revealed the following aspects which encourage the use of full sequence data compared to high density SNP arrays. First, they found that accuracy increases approximately linearly with $\log(m)$, i.e., with the log of the number of markers. It is a pity that this conclusion was based on simulations for $m \leq 33000$, because recent SNP arrays contain up to 10^6 markers. Therefore, further investigations on the dependence of accuracy on m would be desirable if m is very large.

Second, with full sequence data it is much more likely than for SNP arrays that causative mutations are included in the data set. If a causative mutation was included in the analyzed data, accuracy increased by 5% (Meuwissen and Goddard 2010). More importantly, the inclusion of the causative mutation prevents the rapid decline in accuracy over the generations after the estimation of marker effects (cf. Table 6).

In the simulations of Meuwissen and Goddard (2010), who assumed 3 or 30 QTLs, Bayesian

method B performed almost twice as good as BLUP for full sequence data. Although we already know from above (e.g., Figure 2) that Bayesian method B outperforms BLUP if the number of QTLs is low, the relative difference between Bayesian method B and BLUP increases as the marker density increases. The reason is that with an increasing number of markers, the proportion of markers accounting for no genetic variance decreases. By modifying the parameter π, Bayesian method B is able to adapt to an increasing number of markers, whereas BLUP cannot.

5 Genomic selection for multiple traits

In Section 3 we presented methods for the estimation of GEBVs in single traits. Applying these methods separately to a set of traits one could apply classical index selection to select for multiple traits. However, many traits are genetically correlated and it may be more efficient to determine a GEBV for multiple traits simultaneously. Multiple-trait models for this purpose have been investigated only very recently by Jia and Jannink (2012), Guo et al. (2014), Calus and Veerkamp (2011), and Hayashi and Iwata (2013). (Also in the context of GWAS similar approaches have been used lately; see Korte et al. (2012) and Segura et al. (2012)).

5.1 Models

We introduce a multivariate version of BLUP (cf. Section 3.2) and a multivariate version of Bayesian method A (cf. Section 3.3). For notational simplicity, we assume that each marker i has only a single effect, i.e., $k_i = 1$.

Multivariate BLUP

Let n be the number of phenotypic records per trait (we assume that n is independent of the trait) and let o be the number of traits. Furthermore, let $\mathbf{y}^{(j)}$ be the vector of observations at trait j and let $\mathbf{y} = (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(o)})^T$ be the full vector of observations. Similarly to (20), we assume that each trait j follows the model

$$\mathbf{y}^{(j)} = \mu^{(j)}\mathbf{1}_n + \mathbf{Z}^{(j)}\mathbf{g}^{(j)} + \mathbf{e}^{(j)}, \quad 1 \leq j \leq o, \tag{35}$$

wherefore

$$\begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(o)} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_n & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1}_n \end{pmatrix} \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \\ \vdots \\ \mu^{(o)} \end{pmatrix} + \begin{pmatrix} \mathbf{Z}^{(1)} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}^{(2)} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{Z}^{(o)} \end{pmatrix} \begin{pmatrix} \mathbf{g}^{(1)} \\ \mathbf{g}^{(2)} \\ \vdots \\ \mathbf{g}^{(o)} \end{pmatrix} + \begin{pmatrix} \mathbf{e}^{(1)} \\ \mathbf{e}^{(2)} \\ \vdots \\ \mathbf{e}^{(o)} \end{pmatrix}. \tag{36}$$

In the following, we analyze the covariance matrices $\mathbf{R} \in \mathbb{M}^{no,no}$ and $\mathbf{G} \in \mathbb{M}^{mo,mo}$ of environmental and genetic effects, respectively, by incorporating commonly used assumptions of BLUP (m denotes the number of markers).

We assume that the covariances of environmental effects between individuals are independent of the trait and are given by the matrix $\tilde{\mathbf{R}} \in \mathbb{M}^{n,n}$. Let $e^{(ij)}$ denote the environmental

covariance between trait i and j within an individual and $\mathbf{\Sigma}_e = (e^{(ij)}) \in \mathbb{M}^{o,o}$. Then

$$\mathbf{R} = \tilde{\mathbf{R}} \otimes \mathbf{\Sigma}_e = \begin{pmatrix} \tilde{\mathbf{R}}e^{(11)} & \tilde{\mathbf{R}}e^{(12)} & \dots & \tilde{\mathbf{R}}e^{(1o)} \\ \tilde{\mathbf{R}}e^{(21)} & \tilde{\mathbf{R}}e^{(22)} & \dots & \tilde{\mathbf{R}}e^{(2o)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{R}}e^{(o1)} & \tilde{\mathbf{R}}e^{(o2)} & \dots & \tilde{\mathbf{R}}e^{(oo)} \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product.

Similarly, we assume that the covariances of marker effects (relationships) between individuals are independent of the trait and are given by the matrix $\tilde{\mathbf{G}} \in \mathbb{M}^{m,m}$. With $c^{(ij)}$ we denote the covariance of breeding values between trait i and j within an individual, $\mathbf{\Sigma}_g = (c^{(ij)}) \in \mathbb{M}^{o,o}$ (cf. Section 2.1). Then,

$$\mathbf{G} = \tilde{\mathbf{G}} \otimes \mathbf{\Sigma}_g = \begin{pmatrix} \tilde{\mathbf{G}}c^{(11)} & \tilde{\mathbf{G}}c^{(12)} & \dots & \tilde{\mathbf{G}}c^{(1o)} \\ \tilde{\mathbf{G}}c^{(21)} & \tilde{\mathbf{G}}c^{(22)} & \dots & \tilde{\mathbf{G}}c^{(2o)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{G}}c^{(o1)} & \tilde{\mathbf{G}}c^{(o2)} & \dots & \tilde{\mathbf{G}}c^{(oo)} \end{pmatrix}. \tag{37}$$

With these considerations and taking into account that $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$, the solution of (36) is easily obtained from (11). As in the univariate case, it is usually assumed for (36) that the environmental and genetic effects are independently and identically normally distributed, i.e., $\tilde{\mathbf{R}} = \sigma_e^2 \mathbf{I}_n$ and $\tilde{\mathbf{G}} = \sigma_g^2 \mathbf{I}_m$.

Multivariate Bayesian method A

Here, we outline the multivariate Bayesian method A following the presentation of Jia and Jannink (2012). The model is similar to the univariate case, wherefore we proceed in close analogy to Section 3.3.

Instead of (23) we use the multivariate model (36). We denote the vector of effects of marker i on the o traits with $\mathbf{g}_i = (g_i^{(1)}, \dots, g_i^{(o)}) \in \mathbb{M}^{1,o}$ and assume that the marker effects at position i are independently (multivariate) normally distributed with covariance matrix $\mathbf{\Sigma}_{\mathbf{g}_i} \in \mathbb{M}^{o,o}$. As in the univariate Bayesian approach, we allow for $\mathbf{\Sigma}_{\mathbf{g}_i} \neq \mathbf{\Sigma}_{\mathbf{g}_j}$ if $i \neq j$ which is ignored in BLUP.

The prior distribution of the covariance matrix of marker substitution effects $\mathbf{\Sigma}_{\mathbf{g}_i}$ is supposed to be the inverse of the multivariate version of the χ^2-distribution, i.e., the inverse of the Wishart distribution. Thus, instead of (24), we have

$$\text{prior}(\mathbf{\Sigma}_{\mathbf{g}_i}) = \text{inv-Wis}(\nu, \mathbf{S}), \tag{38}$$

where ν denotes the degrees of freedom and $\mathbf{S} \in \mathbb{M}^{o,o}$ is a positive definite scale matrix. The multivariate version of the conditional distribution (25) is given by

$$\text{post}(\mathbf{\Sigma}_{\mathbf{g}_i}|\mathbf{g}_i) = \text{inv-Wis}(\nu + 1, \mathbf{S} + \mathbf{g}_i^T \mathbf{g}_i),$$

where we assumed that each marker i has only a single effect ($k_i = 1$). Concerning the environmental effects, (26) generalizes to

$$\text{prior}(\mathbf{\Sigma}_e) = \text{inv-Wis}(-2, \mathbf{0}),$$

where $\mathbf{0} \in \mathbb{M}^{o,o}$ in the zero matrix. This yields the conditional distribution

$$\text{post}(\mathbf{\Sigma}_e|\mathbf{e}) = \text{inv-Wis}(n - 2, \mathbf{e}^T \mathbf{e}),$$

where $\mathbf{e} = (\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(o)}) \in \mathbb{M}^{n,o}$.

A detailed description of the multivariate sampling algorithm is given in the Appendix of Jia and Jannink (2012). Note that the algorithm in Jia and Jannink (2012) also samples ν and $\mathbf{S}$ whereas Meuwissen et al. (2001) and our presentation consider the parameters to be fixed.

The generalization to the multivariate case of Bayesian method B (and its variants) or Bayesian method Cπ is similarly; see Jia and Jannink (2012) for a short presentation.

5.2 Accuracy compared to single-trait methods

Jia and Jannink (2012) compared the accuracy of single-trait models (STMs) to the accuracy of multi-trait models (MTMs) using simulated data. The genome had seven chromosomes (150cM each) with 5000 SNPs. Out of those SNPs, 20 or 200 QTLs were selected at random. All QTLs affected two traits and were assumed to be pleiotropic. The effects of the QTLs were drawn from a bivariate normal distribution with correlation 0.5. Trait I had low heritability, whereas trait II had high heritability.

A univariate and multivariate version of the pedigree-BLUP model (31), the genomic-BLUP model (20)(36), Bayesian method A, and Bayesian method Cπ (see Section 3.4) was analyzed: The MTMs performe much better than the STMs at trait I (low heritability) if the trait is determined by 20 QTLs (instead of 200). Then, the accuracy at trait I increases from the STMs to the MTMs by 5%, 4%, 22%, and 36%, in the pedigree-BLUP model, the genomic BLUP model, Bayesian method A, and Bayesian method Cπ, respectively (Figure 3a). However, the accuracies of the STMs and the MTMs at trait II (high heritability) are rather similar (Figure 3b,c). Furthermore, if 200 QTLs determine the traits, the prediction

accuracy of the STMs and the MTMs are almost the same (Figure 3c,d). Finally, we note that the accuracy increases from the pedigree-BLUP model to the genomic BLUP model to Bayesian method A and to Bayesian method Cπ if there is a low number of QTLs (Figure 3a,b), which is in accordance with our findings from Section 4.3(iv). If a large number of QTLs determines the trait, the genomic BLUP model performs best (Figure 3c,d).

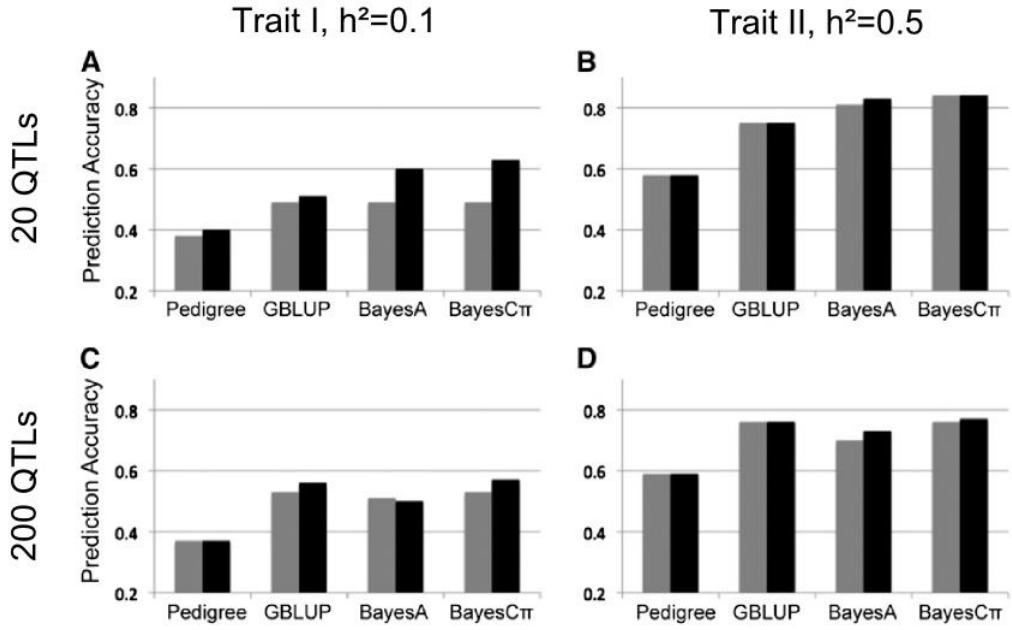

Figure 3: The STMs (gray) and the MTMs (black) are compared for two correlated traits with correlation factor 0.5 and different numbers of QTLs, i.e., different genetic architectures. (The Figure is taken from Jia and Jannink (2012).)

Jia and Jannink (2012) mention that the performance of the MTMs was slightly inferior to the performance of the STMs if the two traits were uncorrelated. The authors think that this arises from the sampling process in the training population, which may suggest correlations between traits where there are none. Similar issues can also occur if the traits are correlated; see Figure 3c where the MTM performed slightly worse than the STM in Bayesian method A.

We note that the results in Figure 3 for Bayesian method A were generated with a version of the method which is slightly different to the one presented in Section 3.3 and Section 5.1. For the results in Figure 3 the shape and scale parameter ν and $\mathbf{S}$ of the prior distribution (38) were assumed to be unknown. Therefore, they were sampled in the Metropolis-Hasting algorithm which slightly improved Bayesian method A (see Appendix of Jia and Jannink 2012).

The influence of heritability

If the heritability at trait I is low ($h^2 = 0.1$ in Table 7), higher heritabilities at trait II increase the accuracy of GEBVs at trait I. However, if the heritability of trait II is high ($h^2 = 0.8$ in Table 7), an increase of heritability at trait I does barely increase the accuracy at trait II. The results in Table 7 were obtained for the MTMs with Bayesian method Cπ. However, qualitatively, the conclusions also hold for the other models.

Heritability		r_{BV}	
Trait I	Trait II	Trait I	Trait II
0.1	0.5	0.63 ± 0.10	0.86 ± 0.05
0.1	0.8	0.70 ± 0.08	0.94 ± 0.02
0.5	0.8	0.89 ± 0.04	0.93 ± 0.03
0.8	0.8	0.93 ± 0.03	0.94 ± 0.03

Table 7: The accuracy of the multivariate Bayesian method Cπ for different heritabilities at trait I and trait II. The accuracy of the STM for trait I with $h^2 = 0.1$ in this simulation was 0.49. The genetic correlation between the traits was 0.5. (The Table is obtained from Table 2 in Jia and Jannink (2012).)

The influence of correlations between traits

As the genetic correlation between the traits increases, the accuracy increases for trait I (low heritability), whereas the accuracy at trait II (high heritability) is barely affected (Figure 4).

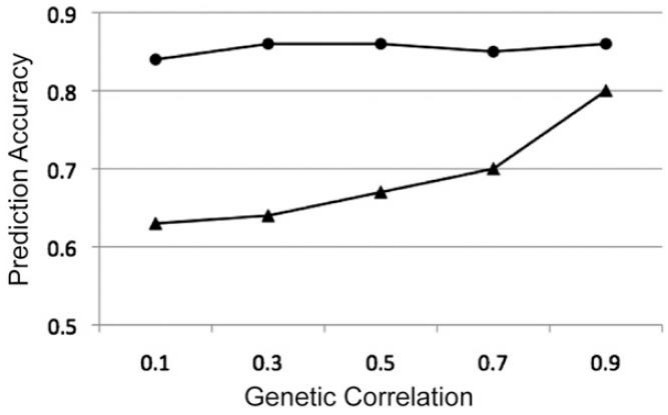

Figure 4: The accuracy of trait I (▲) and trait II (•) as a function of the genetic correlation for the multivariate Bayesian method Cπ. The heritabilities at trait I and trait II were set to $h^2 = 0.1$ and $h^2 = 0.5$, respectively. (The Figure is taken from Jia and Jannink (2012).)

The consequences of missing phenotypic records

Guo et al. (2014) compared the accuracy of the single-trait BLUP model (20) with the accuracy of the multi-trait BLUP model (36) with two traits. To reflect the common practical problem of missing phenotypic records, the authors considered three scenarios with and without missing phenotypic data: (i) no missing data, (ii) 90% missing data for trait I, (iii) 90% missing data for trait II.

Trait I had high heritability ($h^2 = 0.3$), whereas trait II had low heritability ($h^2 = 0.05$). A genome consisting of five chromosomes (100 cM each) with 5000 biallelic SNPs at distances of 0.1cM was simulated. Three hundred randomly located QTLs were assumed and divided randomly into three groups: A hundred QTLs affected only trait I, 100 affected only trait II, and 100 affected both traits. The latter pleiotropic QTLs were assumed to have the same effect on both traits, wherefore the traits were correlated with factor 0.5. As in Meuwissen et al. (2001) the QTL effects were drawn from a Gamma distribution but slightly different shape parameters were assumed.

The main finding of Guo et al. (2014) was that the MTMs gave more accurate GEBVs than the STMs for the trait with missing phenotypic records (Table 8). When no data was missing, Table 8 shows barely an increase in accuracy when switching from the STM to the MTM, even for the low heritability trait. This is in contrast to the results from Jia and Jannink (2012) displayed in Figure 3a. One reason may be that the two authors assumed different genetic architectures. In contrast to Guo et al. (2014), Jia and Jannink (2012) assumed that all QTLs affected all traits.

Trait	Data	r_{BV} of the STM	r_{BV} of the MTM
I	No missing data	0.873 ± 0.002	0.874 ± 0.001
I	Missing data for trait I	0.473 ± 0.008	0.616 ± 0.004
II	No missing data	0.723 ± 0.006	0.725 ± 0.001
II	Missing data for trait II	0.338 ± 0.013	0.554 ± 0.009

Table 8: The accuracy of the univariate and multivariate BLUP model for two traits in the study of Guo et al. (2014). Heritability for trait I and trait II were $h^2 = 0.3$ and $h^2 = 0.05$, respectively, and the correlation between the traits was 0.5. (The Table is obtained from Table 2 in Guo et al. (2014).)

Real data

A few authors have applied MTMs to real data. Jia and Jannink (2012) considered two disease-resistance traits in pines. They found that STMs and MTMs, as well as BLUP and

the Bayesian methods, performed similarly. This would imply that both traits have high heritability and a moderate number of underlying QTLs.

Aguilar et al. (2011) used the multivariate BLUP model to predict GEBVs for fertility traits in Holstein cattle. Low heritability in the selected traits led to moderate improvements of accuracy when using a MTM compared to a STM (see Table 4 in Aguilar et al. 2011). Similarly, Tsuruta et al. (2011) selected for 18 traits in Holsteins and observed that the increase of accuracy depends strongly on the trait when switching from STMs to MTMs.

Conclusion

Multi-trait models (MTMs) have the potential to increase the accuracy of genomic selection, since they use the available data more efficiently than single-trait models (STMs) by incorporating correlations between traits into the model. It was shown that MTMs outperform the STMs for low-heritability traits and traits with missing phenotypic records.

6 Discussion

In genomic selection one determines the breeding values of selection candidates solely from genetic data. Beforehand, genetic effects have to be assigned to markers by phenotyping a reference population and associating genetic with phenotypic variants. To this aim a variety of methods has been developed which estimate genomic breeding values. In Section 3 we presented four of these methods which were developed by Meuwissen et al. (2001) and assume that fitness is determined by a single trait. Mainly they differ in their assumptions on the underlying genetic architecture of the trait. On the one hand, the BLUP model assumes that the effects of the markers are independently normally distributed with the same variance, whereas on the other hand, the Bayesian methods allow for different variances of effect sizes at each SNP. Bayesian method B also incorporates the possibility that SNPs have zero effects, which is the reason that Bayesian method B usually performs slightly better than Bayesian method A.

In Section 4 we compared the accuracy of genomic estimated breeding values obtained from the different methods. The overall accuracy can be split into two components. First, the accuracy of a method depends on its capability to reflect genetic relationships between individuals (Section 3.5, Section 4.3(ii)). Second, the accuracy depends on the amount of LD between markers and QTLs, which is captured by the models with variable precision.

We found that the least-squares method performs worst in almost all aspects and thus focused on the comparison of BLUP with the Bayesian methods. Most factors, such as the number of marker loci (Table 3) or the number of phenotypic records (Table 6) have a similar influence on the accuracy of all methods. The genetic architecture of the trait (in particular the number of QTLs), however, affects BLUP and the Bayesian methods differently, which is in accordance with the assumptions posed in the models. Following Daetwyler et al. (2010) we documented that the accuracy of Bayesian method B clearly exceeds the accuracy of BLUP if only a few QTLs determine the trait (see Figure 2 and equations (32), (33)). If the trait is determined by many loci with small effects, BLUP has the potential to outperform the Bayesian methods.

This behavior reflects the fact that BLUP assigns positive effects to all chromosome segments, irrespective of whether they contribute to the trait, whereas Bayesian method B tries to determine the set of chromosome segments with an effect on the trait via the parameter π. When the number of QTLs is lower than the number of independent chromosome segments

($Q < M_e$), the advantage of choosing a subset is clear, whereas it vanishes otherwise.

In Meuwissen and Goddard (2010) it was shown that the difference in accuracy between BLUP and Bayesian method B increases with an increasing number of markers (or when full sequence data is used instead of SNP arrays). In contrast, the difference in accuracy between the two methods decreases with the number of QTLs (Goddard 2009). It would be interesting to combine the two studies and to investigate the performance of BLUP and Bayesian method B, both, as a function of the number of markers and the number of QTLs.

When examining the accuracy of the methods on real data, the differences between the methods are often even smaller than in simulation studies (Section 4.2, Hayes et al. 2009a). This suggests that the number of QTLs assumed in simulations is usually too small (most studies assume 50 or fewer QTLs; see Habier et al. 2007), which makes it likely that the Bayesian methods perform better than BLUP (e.g., see Zhong et al. (2009) for a study with barely). An exception is the relative fat content in cattle (Table 2) for which we know that it is determined by a few QTLs with large effects.

In Section 5 we generalized BLUP and Bayesian method A to a multivariate version, estimating genetic effects of markers for multiple traits. The MTMs can deal with the phenotypic data in a more efficient way than the STMs by incorporating genetic correlations between traits. Especially the estimation of breeding values at low heritability traits which are correlated to high heritability traits profits from the MTMs. Second, Guo et al. (2014) showed that MTMs increase the accuracy of GEBVs at traits with few phenotypic records which are correlated to traits with many phenotypic records.

In Figure 3a it was shown that the relative benefit of MTMs to STMs is highest for the Bayesian methods. This suggests that the MTMs can capture the correlations between traits best if a few major QTLs are underlying the traits. However, in contrast to the STMs there is not much literature on MTMs and the joint dependence of accuracy on correlations between traits, heritability, and the genetic architecture is not well understood. Therefore, it would be desirable to generalize the approaches of Daetwyler et al. (2010) and Hayes et al. (2009b) (see Section 4.3 (iii, iv)) to multiple traits.

All methods presented in this thesis are concerned with estimating additive gene effects. However, the proportion of the additive genetic variance to the total genetic variance is subject to ongoing discussion and investigation. Multiple studies suggested that epistasis plays a crucial role in maintaining quantitative genetic variation (e.g., Carlborg and Haley 2004),

wherefore it may be necessary to estimate haplotype effects instead of additive gene effects. Gianola et al. (2006) pursued this direction by including nonlinear terms in the genotype-phenotype map of the models presented in Section 3.

In this thesis we focused on theoretical aspects of genomic selection by comparing the accuracy of the genomic estimated breeding value for different methods. In the following we note a few practical issues which should be kept in mind when implementing genomic selection into breeding programs.

(i) In applications it is often not the absolute accuracy that matters, but rather the accuracy per year or accuracy per Euro which is of most interest. Therefore, even if the selective gain of genomic selection is lower than the selective gain of phenotypic selection, it may be economically advantageous.

(ii) Until now, GEBVs are rarely the only source for the choice of selection candidates but are combined with pedigree or phenotypic information. One method to combine traditional selection methods with genomic selection is to calculate breeding values bases on phenotypic and pedigree information, determine GEBVs, and combine both breeding values in a selection index.

(iii) When implementing methods for genomic selection, one should keep in mind that the models come with different computational costs. As reported by Verbyla et al. (2009), the BLUP model has usually the lowest computational costs and is also easiest implemented. Bayesian method A is slightly slower but Bayesian method B is much slower. Increasing the speed of Bayesian method B was also the motivation for combining the Bayesian method with variable selection algorithms (e.g., Bayesian method SSVS; see Verbyla et al. 2009).

(v) Although the costs for genotyping and full genome sequences have rapidly declined in the last years and are supposed to decline further, it is unlikely that the whole reference population can be sequenced. Therefore, it is necessary to take a sample (of size l) from the reference population which should maximize the accuracy of genomic predictions. To this aim, algorithms have been developed which choose individuals based on the relationship matrix $\mathbf{A}$ and intend to capture the largest proportion of the genetic variance which can be explained by l individuals.

(iv) Finally, we note that genomic selection brings along bioinformatical problems such as the the determination of haplotypes from a pool of sequences or the imputation of missing genotypic records (arising from sequencing errors). Statistical models for the imputation of data are reviewed in Marchini and Howie 2010.

Glossary

We provide a glossary of symbols and abbreviations that occur in multiple sections. Roman and Greek letters are listed separately. Uppercase letters precede lower case ones and listing is in alphabetical order. The plus + refers to the text below the equation.

Symbol	Reference	Definition
$\mathbf{1}_l$	(15)	A column vector with l 1s
b_{BV}	Sec. 4.1	The regression of the TBV (or the PEBV) on the GEBV
c	(2)	Recombination rate between two loci
$\mathbf{e}$	(7)	Vector of environmental effects
$\mathbf{G}$	(10)	Covariance matrix of random (genetic) effects
$\mathbf{g}_i$	(16)	Vector of genotypic effects at marker i
h^2	(6)	Heritability of the trait
$\mathbf{I}_l$	Sec. 2.1	$l \times l$ identity matrix
k_i	(15)+	Number of effects at a marker i
$\mathbb{M}^{l,p}$	Sec. 2.1	Set of $l \times p$ matrices
M_e	(32)+	Number of independent chromosome segments
m	(15)	Number of markers
N	(3)	Effective population size
n	(7)	Number of (phenotypic) observations
L	Sec. 4.3(i)	Length of the genome in Morgans
o	(5.1)	Number of traits
Q	(32)	Number of QTLs
$\mathbf{R}$	(7)	Covariance matrix of residual effects
r^2	(1)	Measure of LD
r_{BV}	(4.1)	Correlation of the TBV (or the PEBV) and the GEBV
S	(24)	Scale parameter for the inverted chi-square distribution
$\mathbf{X}$, $\mathbf{X}_i$	(7), (16)	Incidence matrices
$\mathbf{y}$	(7)	Vector of (phenotypic) observations
$\mathbf{Z}$, $\mathbf{Z}_i$	(10), (20)	Incidence matrices
μ	(15)	Mean value of the phenotypic records
ν	(24)	Degrees of freedom in the inverted chi-square distribution
π	(30)	Proportion of markers with vanishing variance
$\mathbf{\Sigma}_g$, $\mathbf{\Sigma}_e$	(6)+	The $o \times o$ covariance matrices between traits
σ_g^2	(4)	Genetic variance
σ_e^2	(4)	Environmental variance
σ_a^2	(5)	Additive genetic variance
$\hat{}$	(8)	Indicates an estimator
^{T}	(9)	Indicates a transposed matrix

Abbreviation	Full term
cM	Centi-Morgan
BLUP	Best linear unbiased prediction
GEBV	Genomic estimated breeding value
GWAS	Genome wide association studies
LE	Linkage equilibrium
LD	Linkage disequilibrium
LS	Least-squares
QTL	Quantitative trait locus
MTM	Multi-trait model
PEBV	Phenotypically evaluated breeding value
SSVS	Stochastic search variable selection
SNP	Single-nucleotide polymorphism
STM	Single-trait model
TBV	True breeding value

References

Aguilar, I., Misztal, I., Tsuruta, S., Wiggans, G., and Lawlor, T. (2011). Multiple trait genomic evaluation of conception rate in Holsteins. *Journal of Dairy Science*, 94:2621–2624.

Bürger, R. (2000). *The mathematical theory of selection, recombination and mutation.* Wiley Series in Mathematical and Computational Biology. Chichester: Wiley.

Calus, M., De Roos, A., Veerkamp, R., et al. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, 178:553–561.

Calus, M. P. and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selelection Evolution*, 43:1–14.

Carlborg, Ö. and Haley, C. S. (2004). Epistasis: Too often neglected in complex trait studies? *Nature Reviews Genetics*, 5:618–625.

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 Roos, A., Hayes, B. J., Spelman, R., and Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. *Genetics*, 179:1503–1512.

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

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

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, 13:135–145.

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

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics*, 15:30.

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

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

Hayashi, T. and Iwata, H. (2013). A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics*, 14:34.

Hayes, B., Bowman, P. J., Chamberlain, A., and Goddard, B. J. (2009a). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92:433–443.

Hayes, B. and Daetwyler, H. (2013). Genomic selection in the era of genome sequencing. *Course Notes from Piacenza, Italy*.

Hayes, B. and Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution*, 33:209–230.

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

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

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

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44:1066–1071.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., and Jannink, J.-L. (2011). Genomic selection in plant breeding: Knowledge and prospects. *Advances in Agronomy*, 110:77.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Incorporated.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11:499–511.

Meuwissen, T. (2009). Accuracy of breeding values of unrelated individuals predicted by dense SNP genotyping. *Genetics Selection Evolution*, 41:35.

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

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

Mrode, R. A. (2014). *Linear models for the prediction of animal breeding values*. Cabi.

Muir, W. (2007). Comparison of genomic and traditional BLUP–estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, 124:342–355.

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6 (Suppl 2):S10.

Pryce, J., Gredler, B., Bolormaa, S., Bowman, P., Egger-Danner, C., Fuerst, C., Emmerling, R., Sölkner, J., Goddard, M., and Hayes, B. (2011). Short communication: Genomic selection using a multi-breed, across-country reference population. *Journal of Dairy Science*, 94:2625–2630.

Schaeffer, L. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123:218–223.

Schulz-Streeck, T. and Piepho, H.-P. (2010). Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. *BMC Proceedings*, 4 (Suppl 1):S8.

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44:825–830.

Slatkin, M. (2008). Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9:477–485.

Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, 2:125–141.

Tsuruta, S., Misztal, I., Aguilar, I., and Lawlor, T. (2011). Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *Journal of Dairy Science*, 94:4198–4204.

Verbyla, K. L., Hayes, B. J., Bowman, P. J., and Goddard, M. E. (2009). Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetics Research*, 91:307–311.

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

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569.

Zhong, S., Dekkers, J. C., 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.

Curriculum Vitae

Personal Data

Name:	Ludwig Geroldinger
Date of birth:	27.9.1987
e-mail:	ludwig_geroldinger@gmx.at

Working Experience

1-2015 – 3-2015:	Researcher (Post-Doc position) at the University of Vienna
9-2010 – 12-2014:	Researcher (PhD-position) at the University of Vienna

Education

12-2014:	Doctoral degree in mathematics with distinction
9-2010 – 12-2014:	PhD Student at the University of Vienna in Biomathematics, employed by the FWF and member of the Vienna Graduate School of Population Genetics
1-2010:	Master degree in mathematics with distinction
2-2007:	First section in physics (1. Studienabschnitt) with distinction
2005–2010:	Study of Mathematics and Physics (Diplom) at the University of Vienna
1997–2005:	BRG Carnerigasse, Graz