Genomic Prediction for Quantitative Traits: Using Kernel Methods and Whole Genome Sequence Based Approaches

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen

vorgelegt von
Ulrike Ober
aus Wiesbaden

Göttingen 2012

Referent: Prof. Dr. Martin Schlather
Korreferent: Prof. Dr. Henner Simianer
Tag der mündlichen Prüfung: 28. September 2012

Abstract

Predicting genetic values is important in animal and plant breeding, personalized medicine and evolutionary biology. Traditionally, prediction is based on a best linear unbiased prediction (BLUP) approach within a linear mixed model framework, with covariance structures obtained from relationship measures between individuals. Nowadays, single nucleotide polymorphism (SNP) data allow to incorporate genomic information into the model (genomic BLUP (GBLUP)).

Prediction is also the principal topic in geostatistics in the framework of correlated data. Here, the so-called “kriging” approach performs BLUP using parameterized covariance functions. In this thesis, the kriging concept to perform genomic prediction using the family of Matérn covariance functions is adopted and kriging is compared to GBLUP in a whole-genome simulation study. The results of the simulation study suggest that kriging is superior over GBLUP in non-additive gene-action scenarios.

The methodological development of genome-based prediction methods has become even more important with the increasing availability of whole genome *sequence* data. This thesis provides the world-wide first application of phenotype prediction based on sequence data in a higher eukaryote using the “*Drosophila melanogaster* Genetic Reference Panel”, which comprises sequences and phenotypic data of 157 inbred lines of the model organism *Drosophila melanogaster*. For the traits “starvation resistance” and “startle response” moderate predictive abilities are obtained performing GBLUP, utilizing 2.5 million SNPs to infer genomic relationships between individuals. The predictive ability of a Bayesian method with internal SNP selection is not higher than the one obtained with GBLUP, and predictive ability of GBLUP decreases only when fewer than 150,000 SNPs are used.

For a third trait (“chill coma recovery”) the GBLUP approach fails completely. Based on differentiated analyses and a corresponding two-marker genome-wide association study, two possible reasons for this failure are identified: the bimodal phenotypic distribution and an extensive network of epistatic interactions between SNPs.

The accuracy of genomic prediction is also affected by the underlying structure of linkage disequilibrium (LD) between SNPs. Several formulae for the expected levels of LD in finite populations have been proposed in the literature, most of them being approximate. In this thesis, an alternative recursion formula for the development of LD over time is proposed. A simulation study illustrates that for all parameter constellations under consideration the proposed formula performs better than the widely used formula of Sved. The theory of discrete-time Markov chains further allows the derivation of the expected amount of LD at equilibrium, leading to a formula for the effective population size N_e. By analyzing the effect of non-exactness of the recursion formula on the steady-state, it is demonstrated that the resulting error in expected LD can be substantial. Using the human HapMap data, it is further illustrated that the N_e-estimate strongly depends on the distribution of minor allele frequencies taken as a basis to select SNPs for the analyses.

Comprising a wide spectrum of investigations at the interface between statistics, animal breeding and genetics, the findings of this thesis are of interest from a practical as well as from a methodical statistical point of view.

Zusammenfassung

Die Vorhersage genetischer Werte ist von großer Bedeutung in der Tier- und Pflanzenzucht, der personalisierten Medizin und der Evolutionsbiologie. Traditionell werden genetische Werte durch eine beste lineare unverzerrte Vorhersage (BLUP) im Rahmen eines linearen gemischten Modells ermittelt, dessen Kovarianzstrukturen aus Verwandtschaftsmaßen zwischen Individuen berechnet werden können. Heutzutage ermöglichen Single Nucleotide Polymorphism (SNP) Marker die Einbeziehung genomischer Informationen in das Model (genomisches BLUP (GBLUP)).

Die Vorhersage von Zufallsvariablen auf Basis korrelierter Daten ist auch eines der wichtigsten Gebiete in der Geostatistik. Dabei wird der sogenannte „Kriging"-Ansatz verwendet, bestehend aus einem BLUP-Ansatz mit parametrisierten Kovarianzfunktionen. In der vorliegenden Arbeit wird das Kriging Konzept auf die genomische Vorhersage übertragen. Unter Verwendung der Familie der Matérn Kovarianzfunktionen wird Kriging mit dem GBLUP-Ansatz in einer genomweiten Simulationsstudie verglichen. Die Ergebnisse der Simulationsstudie lassen darauf schließen, dass Kriging dem GBLUP-Ansatz in nicht-additiven Genwirkungs-Szenarien überlegen ist.

Mit der zunehmenden Verfügbarkeit genomweiter *Sequenzdaten* hat die methodologische Entwicklung genom-basierter Vorhersagemethoden erneut an Bedeutung gewonnen. Diese Arbeit enthält die weltweit erste Studie zur phänotypischen Vorhersage unter Verwendung von Sequenzdaten in einem höheren eukaryotischen Organismus. Der „*Drosophila melanogaster* Genetic Reference Panel" dient dabei als Datengrundlage und umfasst Sequenzen sowie phänotypische Daten von 157 Inzuchtlinien des Modellorganismus *Drosophila melanogaster*. Für die beiden Merkmale „starvation resistance" und „startle response" können unter Verwendung von 2.5 Millionen SNPs moderate Vorhersagegenauigkeiten mit GBLUP beobachtet werden. Die Vorhersagegenauigkeit einer Bayesschen Methode mit interner SNP-Selektion ist nicht größer als die durch GBLUP erzielte Genauigkeit, und die Vorhersagegenauigkeit des GBLUP-Ansatzes nimmt erst ab, wenn weniger als 150.000 SNPs verwendet werden.

Für ein drittes Merkmal („chill coma recovery") erzielt der GBLUP-Ansatz nur sehr geringe Genauigkeiten. Mit Hilfe differenzierter Analysen und einer genomweiten Assoziationsstudie, welche paarweise Interaktionen zwischen Markern miteinbezieht, werden zwei mögliche Ursachen für das Scheitern des GBLUP-Ansatzes identifiziert: die bimodale phänotypische Verteilung sowie ein extensives Netzwerk epistatischer Interaktionen zwischen SNPs.

Es ist bekannt, dass die Genauigkeit der genomischen Vorhersage auch durch die zugrunde liegende Struktur des Kopplungsungleichgewichtes (linkage disequilibrium (LD)) zwischen SNPs beeinflusst wird. Mehrere, meist approximative Formeln für die erwartete Höhe an LD in Populationen endlicher Größe existieren bereits in der Literatur. In dieser Arbeit wird eine alternative Rekursionsformel vorgeschlagen, welche die zeitliche Entwicklung des LDs beschreibt, und in einer Simulationsstudie wird gezeigt, dass die vorgeschlagene Formel der vielfach verwendeten Formel von Sved in allen betrachteten Parameterkonstellationen überlegen ist. Die Theorie zu zeit-diskreten Markovketten erlaubt weiterhin die Herleitung des erwarteten LDs im Gleichgewichtszustand, was wiederum zu einer Formel für die effektive Populationsgröße N_e führt. Durch die Analyse des Effektes der Nicht-Exaktheit der

Rekursionsformel auf den Gleichgewichtszustand kann gezeigt werden, dass der resultierende Fehler an erwartetem LD beachtlich sein kann. Unter Verwendung des humanen HapMap Datensatzes wird außerdem deutlich gemacht, dass der N_e-Schätzer stark von der Verteilung der Allelhäufigkeit des selteneren Allels abhängt, die den zur Analyse ausgewählten SNPs zugrunde liegt.

Die vorliegende Arbeit umfasst ein weites Spektrum an Untersuchungen an Schnittstellen der Statistik, Tierzucht und Genetik. Die vorgestellten Ergebnisse sind sowohl aus praktischer als auch aus methodisch-statistischer Sicht von Interesse.

Preface

First of all, I would like to thank my supervisors Prof. Martin Schlather and Prof. Henner Simianer for their constant support, their never-ending encouragement, their belief in me and their endless list of inspiring ideas, full of intuition for statistics and biology. Thank you for always being available for my questions, even in busy times, and for giving me advice whenever I requested it.

I warmly thank Prof. Daniel Gianola for drawing my attention to kernel-based methods in animal breeding and genetics and for giving me the opportunity to spend four months at the University of Wisconsin-Madison to work on my first paper (Ober *et al.*, 2011). Thank you for your steady support, for sharing your scientific creativity, for being my roommate in the summer of 2011, for your continuous interest in my manuscript drafts and for regularly asking how I am doing. Thanks to Hayrettin Okut for making my stay in Madison as comfortable as possible.

I am grateful to Prof. Trudy Mackay for sharing the DGRP data (Mackay *et al.*, 2012) with me before publication, allowing to perform the first study on the use of full genome sequence data in prediction (Ober *et al.*, 2012*a*), and for giving me the opportunity to spend three weeks at her impressive lab to work on the chill coma recovery data. During this stay at the North Carolina State University in Raleigh, I was lucky to collaborate with Wen Huang and Michael Magwire, whom I want to thank for their time and their help with SAS and all kinds of chill coma recovery related problems. Thanks to Laura Duncan for her warm welcome and for letting me look over her shoulder when working with real *Drosophila* flies.

I am also indebted to my coauthors for joint work on my papers. Above all, I would like to mention Malena Erbe from my working group, being my number one contact person for all questions related to animal breeding and genetics; Nanye Long who shared her R-script for simulating non-additive and epistatic effects with me; Julien Ayroles, who promptly answered a lot of DGRP data related questions despite moving to Boston; and Christian Stricker, who provided the “GenSel” program and lots of Swiss server capacity. Thank you all for contributing in various ways to my research and for accompanying me on my journey.

My research was made possible through the financial support I received from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” (FKZ 0315528C) in association with the Deutsche Forschungsgemeinschaft (DFG) research training group “Scaling problems in statistics” (GRK 1644) and the “Centre for Statistics” in Göttingen.

I also want to thank my fellow PhD students for regular lunch times, for quality times on courses and conferences, for frequent cheering up, for their patient companionship and for their friendship. My special thanks go to Heidi Signer-Hasler, Johannes Martini and Marco Oesting for always being available for lively discussions on my research projects.

Finally, I would like to thank my family who always encouraged me to pursuit my aims and supported me wherever they could. Ultimately, this PhD would not have been possible without Alexander Malinowski: Thank you for your love and your never-ending patience.

Contents

1 Introduction: Challenges in Animal Breeding and Genetics in the Face of Genomic Revolution 1

2 Kriging, Best Linear Unbiased Prediction (BLUP) and the Mixed Model Equations (MME) 9
- 2.1 Basic concepts of kriging 9
- 2.2 BLUP and the MME 11

3 Predicting Genetic Values: a Kernel-Based Best Linear Unbiased Prediction with Genomic Data 17
- 3.1 Background 17
- 3.2 Prediction methods 19
- 3.3 Simulation study 24
- 3.4 Results of the simulation study 27
- 3.5 Discussion 30
- 3.6 Relation between the Matérn covariance function and the covariance matrix of VanRaden (2008) 35
- 3.7 Using linear transformations of covariance matrices leads to linearly transformed predicted GVs 37

4 Using Whole Genome Sequence Data to Predict Quantitative Trait Phenotypes in *Drosophila melanogaster* 39
- 4.1 Introduction 39
- 4.2 Results 41
- 4.3 Discussion 52
- 4.4 Materials and methods 56
- 4.5 More details on the expected LD and M_e 64
- 4.6 The expected value of the genomic relationship matrix of VanRaden (2008) . 66

5 Analyses of Chill Coma Recovery Data: Evidence of Epistatic Interactions 69
- 5.1 Introduction 69
- 5.2 Investigations in analogy to the analyses of starvation resistance and startle response 70
- 5.3 Observation I: two clusters of lines in relation to a large region on chromosome $2L$ 74
- 5.4 Observation II: bimodal phenotypic distribution 81
- 5.5 Epistatic interactions 83
- 5.6 Clustering and epistatic interactions as a basis for further investigations . . . 85
- 5.7 Discussion 87

6 The Expected Linkage Disequilibrium in Finite Populations Revisited 91
6.1 Introduction . . . 91
6.2 A new recursion formula . . . 94
6.3 Simulation study to analyze the performance of the new recursion formula . . 95
6.4 The expected LD at equilibrium based on the theory of discrete-time Markov chains . . . 103
6.5 More details on the expected LD at equilibrium based on Markov chain theory 107
6.6 Application based on the HapMap data . . . 112
6.7 Discussion . . . 118

Bibliography 123

A Supplementary Tables 137

B Supplementary Figures 143

C Curriculum Vitae 151

1 Introduction: Challenges in Animal Breeding and Genetics in the Face of Genomic Revolution

The prediction of phenotypic or genetic values (GVs) is one of the most important subjects in animal and plant breeding and has recently gained relevance in other areas of research like personalized medicine and evolutionary biology. The most prevalent methodological approach in this field is best linear unbiased prediction (BLUP) applied in a linear mixed model framework, dating back to the works of Henderson in the 1950s (Henderson *et al.*, 1959). In this approach, random components of the linear mixed model used for prediction are usually assumed to be multivariate normally distributed, with predetermined covariance structures based on relationship measures between individuals.

Over the last decades, a "genomic revolution" has found its way into both research and practical applications, since the available amount of genomic data has risen exponentially. Single nucleotide polymorphism (SNP) data provide a valuable new source of information, which can be used for prediction purposes, as well as for genome-wide association studies (GWAS), whose aim is the identification of genomic regions with potential influence on the considered trait. Starting with only a few available SNPs at the outset of this advent in the late 1990s, SNP arrays comprising several tens or hundreds of thousands of markers have been developed commercially in the meantime and their use has become standard practice.

Different BLUP methods have been proposed to take genomic marker information into account (Meuwissen *et al.*, 2001), and constructing "genomic relationship" matrices (VanRaden, 2008) has led to so-called genomic best linear unbiased prediction (GBLUP) approaches. In this context, Gianola *et al.* (2006) were the first to propose a non-parametric treatment of genomic information, using reproducing kernel Hilbert space regression methods. However, only the case of Gaussian covariance functions was considered in their analyses.

Prediction is also the principal topic in geostatistics in the framework of correlated data. Here, the so-called "kriging" concept has been developed by Matheron in the 1960s (Matheron, 1962, 1963), which is also based on a BLUP approach for the prediction of regionalized random variables in a low-dimensional space, and the corresponding covariance structure is typically determined by a parameterized covariance function. Based on a given (limited) set of measurements, the prediction of the variable realization in any point of the considered space is of interest. In principle, kriging consists of two steps: (i) estimation of the unknown parameters and hidden variables (in particular by (restricted) maximum likelihood (ML) methods) and (ii) prediction of the values of the regionalized variables by performing a BLUP, under the auxiliary assumption that the parameter values and hidden variables estimated in the first step are the true ones. While in geostatistics the application of kriging is naturally limited to few dimensions, the basic approach is rather universal (Schölkopf *et al.*, 2004).

The first goal of this PhD project was the aggregation of both areas of research – animal breeding on the one hand and geostatistics on the other hand – in the course of the genomic revolution. It is well-known that solving the kriging system (which is in fact a BLUP system based on a specific linear mixed model) is equivalent to solving the so-called Mixed Model Equations (MME) established by Henderson (1963). Piepho (2009) was the first performing genomic prediction with stationary covariance functions in a genomic context by performing genome-wide selection in maize, but it has not been investigated so far, whether classical covariance functions from the geostatistical framework like the widely used family of Matérn functions can also be employed for genomic modeling in animal breeding.

Having both concepts at hand – the classical BLUP approach from animal breeding and the geostatistical kriging method – the following questions arise, which are of interest both from a statistical as well as from a biological point of view:

- How does the kriging approach perform in comparison to the classical GBLUP approach, especially when a very flexible class of covariance functions like the family of Matérn covariance functions is used or when different gene-action scenarios underly the considered trait?
- Is there a mathematical relationship between covariance structures based on the family of Matérn functions and the widely used genomic relationship matrix according to VanRaden (2008)?

One way to investigate these questions is by means of a whole-genome simulation study considering additive, additive-dominance and epistatic gene-action models. Simulation studies are commonly used in the animal breeding community, especially in the framework of model and method comparison (see *e.g.* Meuwissen *et al.* (2001); Long *et al.* (2010)). Along these lines, it is sensible to investigate different kriging approaches, for which parameters and hidden variables are estimated via ML, with the aim to compare the predictive performance of the kriging methods to the standard GBLUP method on the basis of simulated genomic and phenotypic data.

Piepho (2009) used the Gaussian and the exponential covariance function in his analyses, and the Gaussian kernel was also applied by Gianola *et al.* (2006) and de los Campos *et al.* (2009, 2010*a*,*b*). Both kernels are special cases of the family of Matérn covariance functions. It is therefore natural to consider the whole class of covariance functions to reflect the functional dependency of the observed covariances from the distance of genotypes expressed as Euclidean norm, after giving a detailed description of the underlying statistical theory. Furthermore, it can be shown that in a limiting case the genomic covariance structure proposed by VanRaden (2008) may be considered as a covariance function with corresponding quadratic variogram and it can be proven that predicted GVs are only scaled by a factor if the covariance structures are linearly transformed.

The results of this simulation study and the related findings have been published in Ober *et al.* (2011) and form chapter 3.

Beginning in the year 2000, first full genome sequences have been released, *e.g.* for the model species *Drosophila melanogaster* (Adams *et al.*, 2000) and *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) as well as for humans (International Human Genome Sequencing Consortium, 2001). Since then, the incremental decoding of full genome sequences for more and more species has been revealed to be an additionally challenging aspect of the genomic revolution, and full genome sequencing projects like the HapMap project for humans (The International HapMap Consortium, 2003) have induced a supplemental increase in genome-wide data available for research purposes. The methodological development of prediction methods taking these data into account has therefore become even more important.

While it has become standard to apply genomic-based prediction using SNPs from genotyping platforms, the application has been limited with respect to the number of SNP markers (usually fewer than 100,000) entering the model, and GBLUP methods have not been applied using complete genome sequences. However, it has been expected that the availability of sequence data through next generation sequencing technologies would also revolutionize the field of genomic prediction in terms of predictive ability (Meuwissen & Goddard, 2010).

Besides the GBLUP method, a second class of prediction methods has been developed by Meuwissen *et al.* (2001) within a Bayesian framework: The widely used “BayesB” method is also based on a linear mixed model for the phenotypic values, but the random component is a vector of SNP effects whose genetic variances are modeled via a prior distribution.

GBLUP approaches are based on a linear mixed model for the phenotypic values, which includes a vector of random GVs of individuals whose covariance structure is inferred from genomic data, and the GBLUP model is equivalent to the so-called “Random Regression BLUP” (RRBLUP) model under certain conditions. This model includes a vector of random marker effects (instead of a vector of GVs) which are assumed to be drawn from the same normal distribution and uncorrelated. GBLUP and RRBLUP both rely on the assumption that all SNPs are contributing equally to the GVs, which is obviously not true if there are only a few quantitative trait loci (QTL) underlying the trait. The BayesB method therefore includes only a predefined fraction of the available markers into the model to conform with the alternative assumption that most loci are expected to have zero effect on the phenotype, and the remaining non-zero marker effects are drawn from normal distributions with random variances.

It has been suggested that differences between the two prediction methods will become more pronounced in terms of predictive ability with the availability of full genome sequence data, and simulation studies (Meuwissen & Goddard, 2010) have shown that RRBLUP or equivalent GBLUP procedures do not take full advantage of high-density marker data if the number of causal SNPs is small, while approaches with an implicit feature selection such as BayesB might be more accurate. If, in contrast, the number of causal loci is large, RRBLUP or GBLUP methods may yield accurate predictions because the assumption that every SNP has an effect is more realistic. With the availability of whole genome sequence data for prediction, this issue can finally be investigated using real instead of simulated data.

In this regard, it is important to note that genome-based prediction follows a different paradigm than genome-wide association studies (GWAS). While prediction methods are based on linear mixed models with random components describing SNP effects or GVs, GWAS usually identify single molecular variants associated with phenotypic variability using statistical tests, typically based on standard ANOVA models including only one SNP effect at a time as fixed factor variable which is individually tested for significance. The target of this approach is not to predict phenotypes or GVs but to gain knowledge about the biological mechanisms underlying the trait.

So far, sequencing techniques have only been applied to individuals or cohorts of limited size, but initiatives to sequence larger panels are under way for humans (The 1000 Genomes Project Consortium, 2010; Elshire *et al.*, 2011) and cattle (The 1000 Bull Genomes Project, `www.1000bullgenomes.com`), and genotyping by whole genome resequencing will become a standard technology in the foreseeable future. In this context, the first study enabling sequence based prediction has been awaited curiously by the scientific community, and the following problems are of importance from a statistical as well as from a practical point of view:

- Can genomic prediction be efficiently implemented for the GBLUP method using whole genome sequence data?
- Is there a gain in using sequence data as opposed to moderate or high density SNP array data?
- How does the classical GBLUP method perform compared to the BayesB method, when whole genome sequence data are included?
- Is it possible to increase predictive ability by a pre-selection of SNPs or models with an internal feature selection?
- How comparable are the results of genomic prediction and GWAS? Do areas encompassing SNPs with large estimated effects based on the RRBLUP approach coincide with areas including significant SNP positions identified in a corresponding GWAS?

A suitable data set to answer these as well as other closely related questions is the recently published "*Drosophila melanogaster* Genetic Reference Panel" (DGRP, Mackay *et al.* (2012)), which comprises whole genome sequence data for 192 inbred lines of *Drosophila melanogaster*. The *Drosophila* "fruit fly" is the number one model organism in genetics research (Brookes, 2001) and the DGRP data set comprises the first substantial sample of sequences of a higher eukaryotic organism.

The release of the DGRP data has been considered to be a landmark publication, providing an enormously valuable community resource for genetics research and the publication Ober *et al.* (2012*a*), included in chapter 4, is in fact one of the first studies working with this data set. At the same time, Ober *et al.* (2012*a*) provide the first application of genomic prediction using whole genome *sequence* data. In the corresponding investigations, approximately 2.5 million SNPs derived from the sequences of 157 DGRP lines are used to predict GVs for two traits (starvation resistance and startle-induced locomotor behavior) based on different cross-validation procedures.

One important feature of the genetic architecture of quantitative traits is epistasis (Falconer & Mackay, 1996), which occurs when the effect at one locus is modified by the genotype at another locus. The dynamics of epistatic interactions in natural populations though are poorly understood (Swarup *et al.*, 2012).

Since the GBLUP method is predominantly designed for traits with a purely additive genomic background, it is canonical to suspect underlying epistatic gene-action scenarios when observing a poor predictive ability. While predictive abilities for starvation resistance and startle response turned out to be moderately high (Ober *et al.*, 2012*a*), almost no predictive ability was observed for a third trait (chill coma recovery). As the study in Ober *et al.* (2012*a*) does not comprise the chill coma data, it is interesting to investigate possible special characteristics of the chill coma trait in search for potential (biological) explanations for the failure of the GBLUP approach in this case. In this thesis (chapter 5), we specifically focus on the influence of

- a region on chromosome *2L* dividing the DGRP lines into two clusters, which can be revealed using the basic idea of the geostatistical variogram,
- and the bimodality of the phenotypic data.

The latter turns out to be the critical point in identifying extensive epistatic interactions between SNPs affecting the chill coma phenotype on the basis of extended ANOVA-models which account for this bimodality.

While it is standard to carry out association studies including *single* SNP effects (as done in Mackay *et al.* (2012) for starvation resistance, startle response and chill coma recovery), epistatic interactions between SNPs are not included into association studies by default. However, hints of epistasis are manifold for diverse species (see *e.g.* Flint & Mackay (2009); Steinmetz *et al.* (2002)). In a recent study, Swarup *et al.* (2012) found extensive epistasis for olfactory behavior, sleep and waking activity in *D. melanogaster*, and also gave an overview of previously reported studies investigating epistatic interactions for other traits and species. In light of this, it is interesting to pursue hints of epistatic interactions, whenever they occur, since they might allow new insights into the complex biology underlying the considered trait.

In this thesis we present the basic procedures used in the corresponding statistical analyses to detect these interactions for chill coma recovery. The proposed procedures rely on a two-step GWAS, whose ANOVA-models account for both the phenotypic bimodality (in the first step) and for possible interaction terms (in a second step). We also sketch the first findings of this research, which comprise meaningful hints of epistasis underlying the chill coma recovery trait. Since these results were promising with respect to the epistatic findings, they gave rise to many possible subsequent investigations.

To fully understand the complexity of the chill coma recovery trait from a biological point of view, further analyses have been carried out in collaboration with the working group of Prof. Mackay, including Gene Ontology enrichment analyses as well as genetic network investigation. Based on the results presented in chapter 5, a complex genetic architecture of the chill coma fitness trait could be revealed by confirming extensive epistasis and identifying alleles with large effects (Ober *et al.*, 2012*b*). This enabled novel insights into the genetic architecture of chill coma recovery time.

It is well-known that the accuracy of prediction methods based on marker data depends on many factors: the heritability of the trait, its genetic architecture (number of loci affecting trait variation, mode of inheritance, and distribution of allelic effects (Hayes *et al.*, 2010)), the size of the genome, the marker density and the sample size used in the statistical analysis (Daetwyler *et al.*, 2010). Another important factor determining the accuracy is the underlying pattern of stochastic correlation between SNP markers (usually termed “linkage disequilibrium” (LD), Hill & Robertson (1968)).

The concept of LD forms the basis of the promise of genomic prediction methods, since the inclusion of marker information into genomic prediction methods is based upon the assumption that genotyped genetic markers entering the statistical model are in LD with (possibly not genotyped) QTL affecting the trait. The description of patterns of LD in the considered population is therefore obligated when performing a study using real or simulated data. LD is also one of the most important measures considered in population genetics (*cf.* Hedrick (2011) and the references therein), since it is related to the so-called "effective population size" N_e, which can be estimated based on the LD structure in the population of interest.

Substantial efforts have been made to describe the evolution of LD over time mathematically, and computing its expectation in the population exactly has remained an intriguing open problem (Song & Song, 2007). One standard approach to estimate N_e from LD is based on a formula proposed by Sved (1971) for the expected LD "at equilibrium", which relies on a recursion formula for the development of expected LD from one generation to the next. The estimation of N_e based on LD has become common practice in animal breeding (de Roos *et al.*, 2008; Flury *et al.*, 2010; Qanbari *et al.*, 2010), plant breeding (Remington *et al.*, 2001) and human genetics (Tenesa *et al.*, 2007; McEvoy *et al.*, 2011), since it also allows to describe the evolution of N_e over time (Hayes *et al.*, 2003).

Several formulae for the expected levels of LD in populations of finite size have been proposed during the last decades and their plausibility has been shown empirically. However, most of their derivations contain heuristic parts, so that they remain questionable from a mathematical point of view. Therefore, a sound mathematical approach to describe the development of LD in a finite population is urgently needed. On that account, we propose a clearer approach in chapter 6 which is based on an alternative linear recursion formula for the expected LD. In fact, the *exact* formula for the expected LD, which depends on the distribution of allele frequencies, can be calculated only theoretically. We give an approximate solution and analyze its validity extensively in a simulation study. Compared to the widely used formula of Sved (1971), the proposed formula turns out to perform better for all parameter constellations under consideration.

The mathematical theory underlying these formulae assumes that the underlying population is "ideal", *i.e.* there is random mating in non-overlapping generations, a constant population size, no selection, no migration and no mutation. Populations considered in practice usually do not fulfill these conditions. It is therefore of interest to calculate the effective population size N_e based on the average LD-value observed from the population. By definition, N_e is the size of an ideal population "at equilibrium" with the same structure of LD as the population under consideration. In previous studies, "equilibrium" was defined as the point in time at which the expected LD of the next generation equals the LD of the previous one (see *e.g.* Sved (1971); Tenesa *et al.* (2007)). Using this definition and assuming a linear recursion formula for the development of LD from one generation to the next, the expected LD at equilibrium can easily be calculated. However, two major statistical problems arise from this definition:

- It is not clear whether this equilibrium will ever be achieved.
- One cannot infer from this definition how the formula for the expected LD at equilibrium

is affected if the recursion formula is not exact but only approximate.

We address these problems by a novel approach in chapter 6 and analyze the expected LD at equilibrium using the theory of discrete-time Markov chains for the development of the vector of gamete frequencies in the population over time, with equilibrium being defined as the steady-state of the chain. This allows the mathematical derivation of the expected amount of LD at equilibrium based on a linear recursion formula under the assumption that the recursion is exact. An additional analysis considers the effect of non-exactness of a recursion formula on the steady-state, demonstrating that the resulting error in expected LD can be substantial.

Another issue in this context is related to the distribution of minor allele frequencies (MAFs) of the SNPs used for N_e-estimation: While the natural MAF distribution is usually skewed with a substantial excess of small MAF values, commercial SNP arrays are often constructed such that the MAF distribution is uniform (*cf.* Matukumalli *et al.* (2009)), leading to the question how this affects N_e-estimation and how reliable previously reported estimates are. In an application to the HapMap data of two human populations (The International HapMap Consortium, 2003) we therefore illustrate the dependency of the N_e-estimate on the MAF distribution, showing that estimates can vary by up to 30% when a uniform instead of a skewed MAF distribution is taken as a basis to select SNPs for the analyses.

Based on a rigorous statistical approach, our analyses enable new insights into the mathematical complexity of LD-evolution.

This PhD thesis comprises a wide spectrum of investigations at the interface between statistics, animal breeding and genetics. Its findings are of interest from a practical as well as from a methodological and statistical point of view. Parts of this thesis have been published in international peer-reviewed journals, making valuable contributions to ongoing research questions.

Chapter 2 provides a brief description of the basic principles of the kriging concept and the BLUP approach in a linear mixed model framework with a focus on the equivalence of the kriging system and the MME. The investigation of the relationship between both methods taking genomic data into account is included in chapter 3 and published in Ober *et al.* (2011). The first application of the GBLUP approach using full genome sequence data of the DGRP lines is described in chapter 4 and published in Ober *et al.* (2012*a*). Chapter 5 analyzes the special characteristics of the chill coma recovery trait. A manuscript with continuative analyses, which is not contained in this thesis, is currently in revision for *PLoS Genetics* (Ober *et al.*, 2012*b*). Finally, chapter 6 investigates the evolution of LD in a finite population. A corresponding publication is in preparation (Ober *et al.*, 2012*c*).

In order to allow a selective reading of the single chapters, they are coherent but not constitutive (except for chapter 5 which relies on chapter 4).

2 Kriging, Best Linear Unbiased Prediction (BLUP) and the Mixed Model Equations (MME)

In chapters 3 and 4, the basic principle of best linear unbiased prediction (BLUP) is repeatedly used in the context of the geostatistical kriging approach and within the framework of linear mixed model theory. The BLUP approach in linear mixed model theory is closely related to the so-called “Mixed Model Equations” (MME). In this chapter, we will therefore briefly discuss the basic concepts of kriging and the equivalence of the BLUP approach and the MME.

2.1 Basic concepts of kriging

In geostatistics, kriging is nowadays the standard approach whenever spatial prediction of a so-called regionalized variable (Matheron, 1989) has to be performed based on a few isolated measurements of the quantity. To this end, it is assumed that the regionalized variable is a realization of a random function with a certain covariance structure. Mostly, the latter is given by a parameterized covariance function (Cressie, 1993). The kriging approach usually consists of two steps: (i) estimation of the unknown parameters and hidden variables and (ii) prediction of the values of the regionalized variables by performing a BLUP approach, under the auxiliary assumption that the parameters and hidden variables estimated in the first step are the true parameters.

Depending on the specific side conditions of the kriging procedure, many variants of the unique kriging principle have been published (Chilès & Delfiner, 1999; Wackernagel, 2003; Cressie, 1993). The type of kriging is implied by the unbiasedness condition: In “simple kriging” it is assumed that the underlying regionalized variable (which is used for prediction) has known mean, whereas in “universal kriging”, a linear model for the unknown mean of the underlying regionalized variable is assumed.

Let us consider the different kriging concepts more detailed:

Let $Z(\cdot)$ be a regionalized random variable with covariance function C, *i.e.*

$$C(x_i, x_j) = \mathrm{Cov}(Z(x_i), Z(x_j)),$$

with locations $x_i, x_j \in \mathbb{R}^d$. Our aim is to predict the value of $Z(x_0)$ conditional on the observed values of $Z(x_1), \dots, Z(x_n)$. We perform a BLUP by predicting $Z(x_0)$ as

$$\hat{Z}(x_0) = \sum_{i=1}^{n} a_i(x_0) Z(x_i) = (Z(x_1), \dots, Z(x_n))\mathbf{a}$$

with $\mathbf{a} := (a_1(x_0), \dots, a_n(x_0))^T$ and minimizing

$$\begin{aligned}\operatorname{Var}(\hat{Z}(x_0) - Z(x_0)) &= \sum_{i,j=1}^{n} a_i(x_0) a_j(x_0) C(x_i, x_j) + \operatorname{Var}(Z(x_0)) - 2\sum_{i=1}^{n} a_i(x_0) C(x_i, x_0)\\ &= \mathbf{a}^T \mathbf{C} \mathbf{a} + C(x_0, x_0) - 2\mathbf{a}^T \mathbf{C}_0\end{aligned}$$

with $\mathbf{C} := \{C(x_i, x_j)\}_{i,j=1}^{n}$ and $\mathbf{C}_0 := (C(x_1, x_0), \dots, C(x_n, x_0))^T$ subject to the unbiasedness condition

$$0 \overset{!}{=} \mathbb{E}(\hat{Z}(x_0) - Z(x_0)) = (\mathbb{E}(Z(x_1)), \dots, \mathbb{E}(Z(x_n)))\mathbf{a} - \mathbb{E}(Z(x_0)).$$

2.1.1 Simple kriging

In simple kriging, we may assume $\mathbb{E}(Z(x)) = 0$. Then, the unbiasedness condition is automatically fulfilled and the function to be minimized becomes

$$\Phi := \mathbf{a}^T \mathbf{C} \mathbf{a} + C(x_0, x_0) - 2\mathbf{a}^T \mathbf{C}_0.$$

Taking the derivative with respect to $\mathbf{a}$ leads to

$$\frac{\partial \Phi}{\partial \mathbf{a}} = 2\mathbf{C}\mathbf{a} - 2\mathbf{C}_0 \overset{!}{=} \mathbf{0} \quad \Leftrightarrow \quad \mathbf{a} = \mathbf{C}^{-1}\mathbf{C}_0,$$

provided that $C(\cdot, \cdot)$ is a strictly positive definite function.

2.1.2 Universal kriging

In universal kriging, we assume $\mathbb{E}(Z(x)) = \sum_{i=1}^{m} \beta_i f_i(x) = (f_1(x), \dots, f_m(x))\boldsymbol{\beta}$ with a vector $\boldsymbol{\beta} \in \mathbb{R}^m$ and known functions $f_1, \dots, f_m$. Then, the unbiasedness condition amounts to

$$\mathbf{a}^T \mathbf{F} \boldsymbol{\beta} - \mathbf{F}_0^T \boldsymbol{\beta} = 0 \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^m$$

with $\mathbf{F} := \begin{pmatrix} f_1(x_1) & \cdots & f_m(x_1) \\ \vdots & & \vdots \\ f_1(x_n) & \cdots & f_m(x_n) \end{pmatrix}$ and $\mathbf{F}_0^T := (f_1(x_0), \dots, f_m(x_0))$, which is fulfilled if and only if $\mathbf{a}^T \mathbf{F} = \mathbf{F}_0^T$. The function to be minimized becomes

$$\Phi := \mathbf{a}^T \mathbf{C} \mathbf{a} + C(x_0, x_0) - 2\mathbf{a}^T \mathbf{C}_0 + 2(\mathbf{a}^T \mathbf{F} - \mathbf{F}_0^T)\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the corresponding Lagrange multiplier. Taking the derivatives with respect to $\mathbf{a}$ and $\boldsymbol{\lambda}$ leads to

$$\frac{\partial \Phi}{\partial \mathbf{a}} = 2\mathbf{C}\mathbf{a} - 2\mathbf{C}_0 + 2\mathbf{F}\boldsymbol{\lambda} \overset{!}{=} \mathbf{0} \quad \text{and} \quad \frac{\partial \Phi}{\partial \boldsymbol{\lambda}} = \mathbf{F}^T \mathbf{a} - \mathbf{F}_0 \overset{!}{=} \mathbf{0},$$

which finally yields the universal kriging system

$$\begin{pmatrix} \mathbf{F} & \mathbf{C} \\ \mathbf{0} & \mathbf{F}^T \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_0 \\ \mathbf{F}_0 \end{pmatrix}. \tag{2.1}$$

2.2 BLUP and the MME

While geostatisticians usually solve the kriging system to obtain a BLUP of their random variable of interest, animal breeders mostly solve the MME introduced by Henderson *et al.* (1959) to obtain a BLUP of total genetic values of individuals. Although both approaches can obviously be embedded into a mixed model framework, it is not quite evident at first sight that the two different systems of equations are in fact closely related. More precisely, the BLUP approach applied to certain components in a mixed model context (which naturally leads to a linear system of the same form as the kriging system) can be shown to be equivalent to solving the MME. Since both systems are often used without reference to each other in the literature, a derivation of the equivalence is given in the following. Note that the basic idea of this derivation has been established by Henderson (1963) and independently by Goldberger (1962).

2.2.1 The BLUP approach in the mixed model framework

Consider the following linear mixed model:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

with $\boldsymbol{\beta}$ being an m-vector of fixed effects, $\mathbf{u}$ being a p-vector of random effects with $\mathbb{E}(\mathbf{u}) = \mathbf{0}$ and variance-covariance matrix $\mathrm{Cov}(\mathbf{u}) = \mathbf{A}$, and $\mathbf{e}$ being an n-vector of residual effects with $\mathbb{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{e}) = \mathbf{R}$. Further assume that $\mathbf{u}$ and $\mathbf{e}$ are uncorrelated. $\mathbf{W}$ and $\mathbf{Z}$ are supposed to be known incidence matrices of dimensions $n \times m$ and $n \times p$. Let

$$\mathbf{V} := \mathbf{Z}\mathbf{A}\mathbf{Z}^T + \mathbf{R}$$

be the variance-covariance matrix of $\mathbf{y}$. In the following, we will assume that $\mathbf{A}$ and $\mathbf{R}$ are positive definite, which implies that $\mathbf{A}^{-1}, \mathbf{R}^{-1}$ and $\mathbf{V}^{-1}$ exist. (Note that strictly positive definite functions are defined in analogy to positive definite matrices and that positive definite functions are defined in analogy to positive semi-definite matrices.) We will further assume that the rank of $\mathbf{W}$ equals m (which implies that $\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W}$ is invertible). Our aim is to predict

$$\mathbf{K}^T\boldsymbol{\beta} + \mathbf{M}^T\mathbf{u}$$

via a BLUP approach, with $\mathbf{K}$ and $\mathbf{M}$ being coefficient matrices (both having q rows) and with "best prediction" characterized by simultaneously minimized variances of all q components of $\mathbf{K}^T\boldsymbol{\beta} + \mathbf{M}^T\mathbf{u} - \mathbf{K}^T\widehat{\boldsymbol{\beta} + \mathbf{M}}^T\mathbf{u}$. That is we want to find an estimator

$$\mathbf{K}^T\widehat{\boldsymbol{\beta} + \mathbf{M}}^T\mathbf{u} = \mathbf{L}^T\mathbf{y} \tag{2.2}$$

for some coefficient matrix $\mathbf{L}$, provided that $\mathbf{K}^T\boldsymbol{\beta}$ is *estimable* for all $\boldsymbol{\beta} \in \mathbb{R}^m$.
By definition, $\mathbf{K}^T\boldsymbol{\beta}$ is *estimable* if it has a linear unbiased estimate, *i.e.* if there exists a matrix $\tilde{\mathbf{K}}^T$ with

$$\begin{aligned} & \mathbb{E}(\tilde{\mathbf{K}}^T\mathbf{y}) = \mathbf{K}^T\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^m \\ \Leftrightarrow \quad & \tilde{\mathbf{K}}^T\mathbf{W}\boldsymbol{\beta} = \mathbf{K}^T\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^m \\ \Leftrightarrow \quad & \tilde{\mathbf{K}}^T\mathbf{W} = \mathbf{K}^T. \end{aligned}$$

If $\mathbf{K}^T\boldsymbol{\beta}$ is estimable, the Aitken Theorem (Aitken, 1934), which is a generalization of the Gauss-Markov Theorem, ensures that a best linear unbiased estimate (BLUE) of $\mathbf{K}^T\boldsymbol{\beta}$ exists and is unique. Since $\mathbb{E}(\mathbf{L}^T\mathbf{y}) = \mathbf{L}^T\mathbf{W}\boldsymbol{\beta}$ and $\mathbb{E}(\mathbf{K}^T\boldsymbol{\beta} + \mathbf{M}^T\mathbf{u}) = \mathbf{K}^T\boldsymbol{\beta}$, the prediction (2.2) is unbiased if and only if

$$\mathbf{L}^T\mathbf{W}\boldsymbol{\beta} = \mathbf{K}^T\boldsymbol{\beta} \text{ for all } \boldsymbol{\beta} \in \mathbb{R}^m \quad \Leftrightarrow \quad \mathbf{L}^T\mathbf{W} = \mathbf{K}^T.$$

Subject to this condition, we have to minimize

$$\mathrm{Var}((\mathbf{K}^T\boldsymbol{\beta} + \mathbf{M}^T\mathbf{u} - \mathbf{L}^T\mathbf{y})_i)$$

for $i = 1, \ldots, q$, where the subscript i indicates the ith row of a matrix. In the following, the ith columns of $\mathbf{K}, \mathbf{L}$ and $\mathbf{M}$ are denoted by k_i, l_i and m_i.
We first note that

$$\begin{aligned} & \mathrm{Var}((\mathbf{K}^T\boldsymbol{\beta} + \mathbf{M}^T\mathbf{u} - \mathbf{L}^T\mathbf{y})_i) \\ & = (\mathrm{Cov}(\mathbf{M}^T\mathbf{u} - \mathbf{L}^T\mathbf{y}))_{ii} \\ & = (\mathrm{Cov}(\mathbf{M}^T\mathbf{u} - \mathbf{L}^T(\mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e})))_{ii} \\ & = (\mathrm{Cov}((\mathbf{M}^T - \mathbf{L}^T\mathbf{Z})\mathbf{u} + \mathbf{L}^T\mathbf{e}))_{ii} \\ & = ((\mathbf{M}^T - \mathbf{L}^T\mathbf{Z})\mathbf{A}(\mathbf{M} - \mathbf{Z}^T\mathbf{L}) + \mathbf{L}^T\mathbf{R}\mathbf{L})_{ii} \\ & = m_i^T\mathbf{A}m_i - m_i^T\mathbf{A}\mathbf{Z}^T l_i - l_i^T\mathbf{Z}\mathbf{A}m_i + l_i^T\mathbf{Z}\mathbf{A}\mathbf{Z}^T l_i + l_i^T\mathbf{R}l_i. \end{aligned}$$

Minimizing this expression with respect to l_i subject to the side condition $l_i^T\mathbf{W} = k_i^T$ from $\mathbf{L}^T\mathbf{W} = \mathbf{K}^T$ finally leads to the system of equations

$$\begin{pmatrix} \mathbf{W} & \mathbf{Z}\mathbf{A}\mathbf{Z}^T + \mathbf{R} \\ \mathbf{0} & \mathbf{W}^T \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\lambda}_i \\ l_i \end{pmatrix} = \begin{pmatrix} \mathbf{Z}\mathbf{A}m_i \\ k_i \end{pmatrix},$$

for $i = 1, \ldots, q$, where $\boldsymbol{\lambda}_i$ is the corresponding Lagrange multiplier. Summarizing these systems for $i = 1, \ldots, q$ yields

$$\begin{pmatrix} \mathbf{W} & \mathbf{V} \\ \mathbf{0} & \mathbf{W}^T \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{L} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}\mathbf{A}\mathbf{M} \\ \mathbf{K} \end{pmatrix} \tag{2.3}$$

with $\boldsymbol{\lambda} := (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_q)$. Note that (2.3) is of the same form as q (combined) universal kriging systems specified in (2.1). Hence we have to solve

$$\mathbf{VL} - \mathbf{ZAM} + \mathbf{W}\boldsymbol{\lambda} \overset{!}{=} \mathbf{0} \tag{2.4}$$

$$\text{and} \quad \mathbf{W}^T\mathbf{L} - \mathbf{K} \overset{!}{=} \mathbf{0}. \tag{2.5}$$

From (2.4) we get

$$\mathbf{L} = \mathbf{V}^{-1}(\mathbf{ZAM} - \mathbf{W}\boldsymbol{\lambda}). \tag{2.6}$$

Plugging this into (2.5), we obtain

$$\begin{aligned} & \mathbf{W}^T(\mathbf{V}^{-1}\mathbf{ZAM} - \mathbf{V}^{-1}\mathbf{W}\boldsymbol{\lambda}) - \mathbf{K} = \mathbf{0} \\ \Leftrightarrow \quad & \mathbf{W}^T\mathbf{V}^{-1}\mathbf{W}\boldsymbol{\lambda} = \mathbf{W}^T\mathbf{V}^{-1}\mathbf{ZAM} - \mathbf{K} \\ \Leftrightarrow \quad & \boldsymbol{\lambda} = (\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{V}^{-1}\mathbf{ZAM} - \mathbf{K}). \qquad (2.7) \end{aligned}$$

Plugging this formula into (2.6), we arrive at

$$\begin{aligned} \mathbf{L}^T &= (\mathbf{M}^T\mathbf{AZ}^T - \boldsymbol{\lambda}^T\mathbf{W}^T)\mathbf{V}^{-1} \\ &= \mathbf{M}^T\mathbf{AZ}^T\mathbf{V}^{-1} - (\mathbf{W}^T\mathbf{V}^{-1}\mathbf{ZAM} - \mathbf{K})^T(\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{V}^{-1} \\ &= \mathbf{M}^T\mathbf{AZ}^T\mathbf{V}^{-1} + \mathbf{K}^T(\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{V}^{-1} \qquad (2.8) \\ &\quad - \underbrace{(\mathbf{W}^T\mathbf{V}^{-1}\mathbf{ZAM})^T(\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{V}^{-1}}_{=\mathbf{M}^T\mathbf{AZ}^T\mathbf{V}^{-1}\mathbf{W}(\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{V}^{-1}}. \qquad (2.9) \end{aligned}$$

Indeed, $\mathbf{L}$ and $\boldsymbol{\lambda}$ from (2.7) and (2.8) solve the system (2.3).
Now, let

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{V}^{-1}\mathbf{y},$$

which is the generalized least square (GLS) solution for $\boldsymbol{\beta}$. Then, the BLUP of $\mathbf{K}^T\boldsymbol{\beta} + \mathbf{M}^T\mathbf{u}$ is given by

$$\mathbf{L}^T\mathbf{y} = \mathbf{K}^T\hat{\boldsymbol{\beta}} + \mathbf{M}^T\mathbf{AZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}). \tag{2.10}$$

Particularly, the BLUP of $\mathbf{u}$ is given by

$$\hat{\mathbf{u}} = \mathbf{AZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}})$$

and the BLUE of $\boldsymbol{\beta}$ equals $\hat{\boldsymbol{\beta}}$, which can easily be seen by choosing $\mathbf{K} = \mathbf{0}$ and $\mathbf{M} = \mathbf{I}$ (and $\mathbf{K} = \mathbf{I}$ and $\mathbf{M} = \mathbf{0}$, respectively) in equation (2.10).

2.2.2 Equivalence to the MME

Consider now the following linear system of equations, also known as the *Mixed Model Equations* (MME):

$$\begin{pmatrix} \mathbf{W}^T\mathbf{R}^{-1}\mathbf{W} & \mathbf{W}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{W} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{A}^{-1} \end{pmatrix} \cdot \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{W}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \tag{2.11}$$

In the following, we will show the equivalence of the BLUP for $\mathbf{u}$ and the BLUE for $\boldsymbol{\beta}$ to the solution of the MME. First of all, the Sherman-Morrison-Woodbury formula (Henderson & Searle, 1981) states that

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^T\mathbf{R}^{-1} \quad \text{with} \quad \mathbf{T} = (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{A}^{-1})^{-1}. \tag{2.12}$$

From the MME we get

$$\begin{aligned} & (\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{A}^{-1})\hat{\mathbf{u}} = \mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\beta}}) \\ \Leftrightarrow \quad & \hat{\mathbf{u}} = \mathbf{T}\mathbf{Z}^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\beta}}). \end{aligned}$$

Using (2.12), we also have

$$\begin{aligned} \mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1} &= \mathbf{A}\mathbf{Z}^T\mathbf{R}^{-1} - \mathbf{A}\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^T\mathbf{R}^{-1} \\ &= \mathbf{A}(\mathbf{T}^{-1} - \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z})\mathbf{T}\mathbf{Z}^T\mathbf{R}^{-1} \\ &= \mathbf{T}\mathbf{Z}^T\mathbf{R}^{-1}. \end{aligned} \tag{2.13}$$

Hence, we get

$$\hat{\mathbf{u}} = \mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\beta}}), \tag{2.14}$$

which is the BLUP of $\mathbf{u}$. From the MME we also have

$$\mathbf{W}^T\mathbf{R}^{-1}\mathbf{W}\hat{\boldsymbol{\beta}} + \mathbf{W}^T\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{u}} = \mathbf{W}^T\mathbf{R}^{-1}\mathbf{y}.$$

By plugging in $\hat{\mathbf{u}}$ from eq. (2.14), we finally get

$$\begin{aligned} & \mathbf{W}^T\mathbf{R}^{-1}\mathbf{W}\hat{\boldsymbol{\beta}} + \mathbf{W}^T\mathbf{R}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\beta}}) = \mathbf{W}^T\mathbf{R}^{-1}\mathbf{y} \\ \Leftrightarrow \quad & \mathbf{W}^T \underbrace{(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\underbrace{\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}}_{\overset{(2.13)}{=}\mathbf{T}\mathbf{Z}^T\mathbf{R}^{-1}})}_{=\mathbf{V}^{-1}} \mathbf{W}\hat{\boldsymbol{\beta}} = \mathbf{W}^T \underbrace{(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1})}_{=\mathbf{V}^{-1}} \mathbf{y} \\ \Leftrightarrow \quad & \hat{\boldsymbol{\beta}} = (\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{V}^{-1}\mathbf{y}, \end{aligned}$$

which is the GLS solution for $\boldsymbol{\beta}$. Indeed, $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\beta}}$ solve the MME (2.11).
Thus, we have shown that solving the linear system (2.3) for the two cases (i) $\mathbf{W} = \mathbf{0}$ and $\mathbf{Z} = \mathbf{I}$ and (ii) $\mathbf{W} = \mathbf{I}$ and $\mathbf{Z} = \mathbf{0}$ to obtain a BLUP of $\mathbf{u}$ and a BLUE of $\boldsymbol{\beta}$, respectively, is

equivalent to solving the MME (2.11).

Note that the solution of the MME can also be obtained by maximizing the likelihood function of $(\mathbf{y},\mathbf{u})$ with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ under the assumption that $\mathbf{u}$ and $\mathbf{e}$ are multivariate normally distributed.

3 Predicting Genetic Values: a Kernel-Based Best Linear Unbiased Prediction with Genomic Data

This chapter is based on the article Ober *et al.* (2011).

3.1 Background

Predicting genotypes and phenotypes plays an important role in many areas of life sciences. Both in animal and plant breeding, it is essential to predict the genetic quality (the so-called total genetic value (GV)) of individuals or lines, based on different sources of knowledge. Often, phenotypic measures for various traits are available and the additive genetic relationship between individuals (Wright, 1922) can be derived, based on the known pedigree. Best linear unbiased prediction (BLUP, Henderson (1973)) of breeding values is a well established methodology in animal breeding (Mrode, 2005) and has recently gained relevance in plant breeding (Piepho *et al.*, 2008). In both areas, the main interest is in complex traits with a quantitative genetic background.

In human medicine, the interest is in predicting phenotypes, rather than genotypes, for simple or complex traits (*e.g.* the probability/risk to encounter a certain disease). Genetic prediction is mainly applied in the context of genetic counseling by predicting the risk of genetic disorders with known mono- or oligogenetic modes of inheritance and a certain history of cases in a known family structure, but accurate predictions of genetic predispositions to human diseases should also be useful for preventive and personalized medicine (de los Campos *et al.*, 2010*a*). Wray *et al.* (2007) discuss the potential use of prediction of the genetic liability for traits with a complex quantitative genetic background in a human genetics context, and the variety of possible methods, including linear models, penalized estimation methods and Bayesian approaches was reviewed by de los Campos *et al.* (2010*a*).

With the availability of high-throughput genotyping facilities (Ranade *et al.*, 2001), genotypes for massive numbers of single nucleotide polymorphisms (SNPs) are available and can be used as an additional source of information for predicting GVs. Meuwissen *et al.* (2001) have suggested to include SNP information in a statistical model of prediction. They used three statistical models: a model assigning random effects to all available SNPs (later termed “genomic BLUP”), assuming all SNP effects to be drawn from the same normal distribution, and two Bayesian models, where all (“BayesA”) or a subset (“BayesB”) of the random SNP effects are drawn from distributions with different variances. Various modifications of these methods and additional models have been subsequently suggested (Gianola *et al.*, 2009).

Gianola *et al.* (2006) and Gianola & van Kaam (2008) have suggested a non-parametric treatment of genomic information by using Reproducing Kernel Hilbert Spaces (RKHS) regression, which has already been demonstrated with real data (González-Recio *et al.*, 2008, 2009). As was argued by de los Campos *et al.* (2009), the RKHS regression approach to genomic modeling represents a generalized class of estimators and provides a framework for genetic evaluation of quantitative traits that can be used to incorporate information on pedigrees, markers, or any other ways of characterizing the genetic background of individuals.

Opportunities to enhance genetic analyses by using non-parametric kernel-based statistical methods are enormous and these methods have been considered in different areas of genetic research. Schaid (2010*a*,*b*) provides an overview of measures of genomic similarity based on kernel-methods and describes how kernel functions can be incorporated into different statistical methods like *e.g.* non-parametric regression, support vector machines or regularization in a mixed model context. Only recently, kernel-based methods have also been used in association studies (Yang *et al.*, 2008; Kwee *et al.*, 2008) and QTL mapping for complex traits (Zou *et al.*, 2010), which demonstrates their great potential and flexibility.

Prediction is also relevant in other areas of research: In large parts of geostatistics, the spatial distribution of variables (like temperature, humidity, ore concentration, etc.) is considered. Based on a given (limited) set of measurements, the prediction of the variable realization in any point of the considered space is of interest. A standard approach for prediction in this case is the so-called "kriging" (Chilès & Delfiner, 1999) which makes use of a parameterized covariance function of the regionalized variables.

While in geostatistics the application of kriging is naturally limited to few dimensions, the basic approach is rather universal (Schölkopf *et al.*, 2004). In this chapter we apply kriging to the genomic prediction problem. Here, one dimension reflects genotype realizations at one SNP. In the genomic context, with p SNPs, realizations are in an p-dimensional orthogonal hypercube. Due to the biallelic nature of SNPs, only three genotype realizations (coded *e.g.* as 0, 1 and 2) are possible in each dimension, so that the number of possible genotype constellations over p SNPs is 3^p.

The concept of kriging is closely related to the concept of best linear unbiased prediction (BLUP). Cressie (1990) provides a "historical map of kriging" up to 1963 in which he also refers to Henderson (1963) who introduced BLUP in animal breeding. The steps of kriging are equivalent to "empirical BLUP"-procedures known in other frameworks, and kriging can be viewed as a "spatial BLUP". The conceptual equivalence of geostatistical kriging and BLUP has already been discussed by Harville (1984). Robinson (1991) provides a detailed review on the history of estimation of random effects via BLUP and its various derivations. He also points out the similarities between BLUP and kriging.

The equivalence of kriging with BLUP in a space spanned by *genomic* data was first noted by Piepho (2009), who also discusses relationships with other estimation principles, like ridge regression (Whittaker *et al.*, 2000) and least squares support vector machines (Suykens *et al.*, 2002). Comparing the performance of spatial mixed models to ridge regression with maize data, he found that spatial models provide an attractive alternative for prediction. He also points out that the BLUP model used in Meuwissen *et al.* (2001) has an interpretation as a spatial model with quadratic covariance function. Spatial models for genomic prediction were also used by Schulz-Streeck & Piepho (2010).

Moreover, kriging is known to be closely related to radial basis function (RBF) regression methods (Myers, 1992). Long *et al.* (2010) showed with real and simulated data that non-parametric RBF regression methods can outperform BayesA when predicting total GVs in the presence of non-additive effects using SNP markers.

In this chapter we will demonstrate the potential of the kriging approaches applied to genomic data: As a novelty, we will suggest the family of Matérn covariance functions to reflect the functional dependency of the observed covariances from the distance of genotypes expressed as Euclidean norm. Based on this model and the assumed covariance function, we will suggest two kriging approaches. Under both models, parameters and hidden variables are estimated via maximum likelihood (ML) and BLUP of the unknowns is established by solving the corresponding linear kriging systems. All predictions can also be implemented in the form of the so-called mixed model equations (Henderson, 1973). The predictive performance of the two models will be compared to a common genomic BLUP as a reference method in a whole-genome simulation study considering various gene-action models.

Furthermore, we will show that in a limiting case the genomic covariance structure proposed by VanRaden (2008) can be considered as a covariance function with corresponding quadratic variogram. Besides we will prove theoretically that predicted GVs are only scaled by a factor if the covariance structures are linearly transformed. Finally, we will discuss further options for a more differentiated modeling using the suggested methodological approach.

3.2 Prediction methods

3.2.1 Kriging

The term kriging stems from the prediction of ore concentrations in deposits and was mainly developed by Matheron (1962, 1963) based on the master's thesis of Krige (1951). In geostatistics, kriging is nowadays the standard approach whenever spatial prediction of a so-called regionalized variable (Matheron, 1989), *e.g.* temperature, ozone concentration or soil moisture, has to be performed based on a few isolated measurements of the quantity. It is assumed that the regionalized variable is a realization of a random function with a certain covariance structure. Mostly, the latter is given by a parameterized covariance function (Cressie, 1993), and the random function is assumed to be Gaussian.

The kriging approach consists of two steps: (i) estimation of the unknown parameters and hidden variables (in particular by ML or REML) and (ii) prediction of the values of the regionalized variables by performing a BLUP, under the auxiliary assumption that the parameter values and hidden variables estimated in the first step are the true ones.

Many variants of the general kriging principle have been discussed (Cressie, 1993). The type of kriging is implied by the unbiasedness condition: In "simple kriging" it is assumed that the underlying regionalized variable has zero-mean, whereas in "universal kriging" a linear model for the mean of the underlying regionalized variable is assumed.

3.2.2 The model for polygenic and genomic data

In our further studies, we assume to have q individuals with pedigree information, n of them being genotyped and having phenotype measurements of a certain quantitative trait.

Typically, GVs have to be predicted for individuals that are genotyped, but have no phenotype data.

We use the following model for the given data:

$$y_i = \mathbf{w}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u} + g(\mathbf{x}_i) + e_i, \quad i = 1, \ldots n,$$

where y_i is a measurement of the phenotype for individual i, $\boldsymbol{\beta}$ is an f-vector of nuisance location parameters, $\mathbf{x}_i$ is a p-vector of dummy SNP instance variates (genotype) observed on individual i, and g is an unknown, random function as described below. Let $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathbf{A})$ be a q-vector of additive genetic effects of q individuals where σ_u^2 is the additive genetic variance due to unmarked polygenes, and $\mathbf{A}$ is the numerator relationship matrix. The entries of the numerator relationship matrix are twice the coefficients of coancestry between individuals. The vectors $\mathbf{w}_i^T$ and $\mathbf{z}_i^T$ are known incidence vectors; $\mathbf{z}_i$ is a unit vector with one component being 1 and all the others zero, indicating the respective position in the pedigree. Let $\mathbf{e} = (e_1, \ldots, e_n)^T$ be the vector of environmental residual effects with $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I})$, where σ_e^2 is the environmental variance.

We assume that $\{g(\mathbf{x}_i), \mathbf{x}_i \in \mathbb{R}^p\}$ is a Gaussian random field (Lifshits, 1995) with $\mathbb{E}(g(\mathbf{x}_i)) = 0$ and covariance structure given by $\mathrm{Cov}(g(\mathbf{x}_i), g(\mathbf{x}_j)) = \mathbb{E}(g(\mathbf{x}_i) g(\mathbf{x}_j)) = K_{\nu,h,\sigma_K}(\mathbf{x}_i, \mathbf{x}_j)$, where $K_{\nu,h,\sigma_K}(\cdot\,,\,\cdot)$ is a covariance function depending on parameters ν, h, and σ_K. Let $\mathbf{K}_{\nu,h,\sigma_K} = (K_{\nu,h,\sigma_K}(\mathbf{x}_i, \mathbf{x}_j))_{1 \le i,j \le n}$ be the corresponding covariance matrix.

The family of Matérn covariance functions: For the covariance structure we suggest to use the so-called family of Matérn covariance functions, which was introduced by Matérn (1960) and Handcock & Wallis (1994), and which is defined by

$$\mathrm{Cov}(g(\mathbf{x}_i), g(\mathbf{x}_j)) = K_{\nu,h,\sigma_K}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_K^2 \cdot \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\|\mathbf{x}_i - \mathbf{x}_j\|/h\right)^{\nu} \mathfrak{K}_\nu \left(\sqrt{2\nu}\|\mathbf{x}_i - \mathbf{x}_j\|/h\right).$$

Here, $\|\cdot\|$ is the Euclidean norm, $\nu > 0$ is a smoothness parameter, h is a scale parameter, σ_K^2 is the variance parameter and $\mathfrak{K}_\nu(\cdot)$ is a modified Bessel function of the second kind of order ν (Abramowitz & Stegun, 1984). The Matérn function is isotropic, in that $\mathrm{Cov}(g(\mathbf{x}_i), g(\mathbf{x}_j))$ only depends on the Euclidean norm of the separation vector $\mathbf{x}_i - \mathbf{x}_j$.

Matérn covariance functions build a very general class of covariance functions including special cases like the exponential ($\nu = 1/2$) and the Gaussian ($\nu = \infty$) covariance function, the ones that have also been used by Piepho (2009). If the smoothness parameter ν is of the form $m + 1/2$, where m is an integer, the Matérn function factorizes into the product of an exponential function and a polynomial of degree m, *cf.* Table 3.1 and Figure 3.1. The best fitting parameter value ν is determined through the model-fitting approaches described below.

In matrix notation, the statistical model is

$$\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{g}(\mathbf{X}) + \mathbf{e}, \tag{3.1}$$

where $\mathbf{W} = (\mathbf{w}_1^T, \ldots, \mathbf{w}_n^T)^T$ is an $(n \times f)$- and $\mathbf{Z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$ is an $(n \times q)$-incidence matrix and $\mathbf{g}(\mathbf{X}) = (g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n))^T$. Finally, we assume that the random vectors $\mathbf{u}$, $\mathbf{e}$

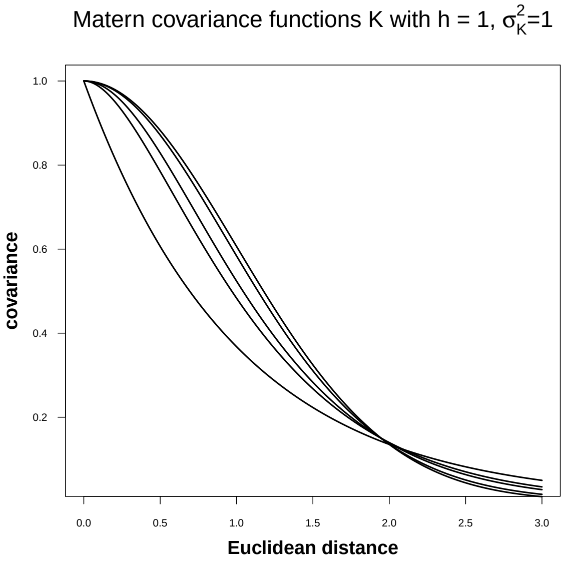

Figure 3.1: Matérn covariance functions for $h = 1, \sigma_K^2 = 1$ and different values of ν. From top to bottom $\nu = \infty, 10, 2.5, 1.5, 0.5$.

Table 3.1: Special cases of Matérn covariance functions

	ν	h	$K_{\nu,h,\sigma_K}(\mathbf{x}_i, \mathbf{x}_j)$
Exponential	0.5	1	$\sigma_K^2 \cdot \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|)$
	1.5	1	$\sigma_K^2 \cdot \exp(-\sqrt{3}\|\mathbf{x}_i - \mathbf{x}_j\|) \cdot \left(1 + \sqrt{3}\|\mathbf{x}_i - \mathbf{x}_j\|\right)$
	2.5	1	$\sigma_K^2 \cdot \exp(-\sqrt{5}\|\mathbf{x}_i - \mathbf{x}_j\|) \cdot \left(1 + \sqrt{5}\|\mathbf{x}_i - \mathbf{x}_j\| + \frac{5}{3}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$
Gaussian	∞	1	$\exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$

and $\mathbf{g}(\mathbf{X})$ are independent.

3.2.3 Two kriging approaches and a reference model

We consider two models to predict the total genetic value $\mathbf{z}_0^T\mathbf{u} + g(\mathbf{x}_0)$ of a certain genotyped individual indexed by 0. This individual belongs to the set of q individuals, but it does not have to be phenotyped. The models differ in the size of the sets of quantities that are estimated in the first kriging step and subsequently used for predictions.

Universal Kriging: Modeling of $\mathbf{y}$: We exploit the fact that $\mathbf{y}$ has a multivariate normal distribution,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\beta}, \sigma_u^2 \mathbf{Z}\mathbf{A}\mathbf{Z}^T + \mathbf{K}_{\nu,h,\sigma_K} + \sigma_e^2 \mathbf{I}),$$

and estimate the parameters $\beta, \sigma_u, \sigma_e, \nu, h$ and σ_K by maximizing the loglikelihood of the corresponding density function.

Then, we perform a best linear unbiased prediction of $g(\mathbf{x}_0)$ and $\mathbf{z}_0^T\mathbf{u}$, *i.e.* we apply the BLUP principle: To obtain $\hat{g}(\mathbf{x}_0)$ we minimize

$$\mathbb{E}(\hat{g}(\mathbf{x}_0) - g(\mathbf{x}_0))^2 \longrightarrow \text{ min!}$$

with the linear predictor $\hat{g}(\mathbf{x}_0) = \mathbf{a}_\mathbf{g}^T\mathbf{y}$ under the condition $\mathbf{a}_\mathbf{g}^T\mathbf{W} = \mathbf{0}$. This approach is called "universal kriging" in other areas of research (Cressie, 1993). In fact, the condition assures $\mathbf{a}_\mathbf{g}^T\mathbf{W}\boldsymbol{\beta} = 0$ and therefore $\mathbb{E}g(\mathbf{x}_0) = 0 = \mathbf{a}_\mathbf{g}^T\mathbf{W}\boldsymbol{\beta} = \mathbb{E}\hat{g}(\mathbf{x}_0)$, *i.e.* $\hat{g}(\mathbf{x}_0)$ is unbiased. Let $\mathbf{K}_0 = (K_{\nu,h,\sigma_K}(\mathbf{x}_1,\mathbf{x}_0), \ldots, K_{\nu,h,\sigma_K}(\mathbf{x}_n,\mathbf{x}_0))^T$. The approach results in the following kriging system of equations:

$$\begin{bmatrix} \mathbf{W} & \sigma_u^2\mathbf{Z}\mathbf{A}\mathbf{Z}^T + \mathbf{K}_{\nu,h,\sigma_K} + \sigma_e^2\mathbf{I} \\ \mathbf{0} & \mathbf{W}^T \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{a}_\mathbf{g} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_0 \\ \mathbf{0} \end{bmatrix}.$$

Note that this linear system does not depend on $\boldsymbol{\beta}$. Analogously, $\mathbf{z}_0^T\mathbf{u}$ can be predicted by the universal kriging estimator $\widehat{\mathbf{z}_0^T\mathbf{u}} = \mathbf{a}_\mathbf{u}^T\mathbf{y}$, where $\mathbf{a}_\mathbf{u}$ satisfies

$$\begin{bmatrix} \mathbf{W} & \sigma_u^2\mathbf{Z}\mathbf{A}\mathbf{Z}^T + \mathbf{K}_{\nu,h,\sigma_K} + \sigma_e^2\mathbf{I} \\ \mathbf{0} & \mathbf{W}^T \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{a}_\mathbf{u} \end{bmatrix} = \begin{bmatrix} \sigma_u^2\mathbf{Z}\mathbf{A}\mathbf{z}_0 \\ \mathbf{0} \end{bmatrix},$$

and one gets $\widehat{\mathbf{z}_0^T\mathbf{u}} + \hat{g}(\mathbf{x}_0)$ as BLUP of $\mathbf{z}_0^T\mathbf{u} + g(\mathbf{x}_0)$.

Mixed Model Equations (MME). In the animal breeding context it is well-known that a BLUP-approach for the model $\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{g}(\mathbf{X}) + \mathbf{e}$ is equivalent to solving the MME

$$\begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} & \mathbf{W}^T \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{A}^{-1} & \mathbf{Z}^T \\ \mathbf{W} & \mathbf{Z} & \mathbf{I} + \sigma_e^2\mathbf{K}_{\nu,h,\sigma_K}^{-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \widehat{\mathbf{g}(\mathbf{X})} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{y} \end{bmatrix} \tag{3.2}$$

for given variance components estimated *e.g.* by ML. For a derivation of the MME from the kriging system compare section 2.2 or Dempfle (1982).

Simple Kriging: Joint modeling of $\mathbf{y}$, $\mathbf{u}$ and $\mathbf{g}(\mathbf{X})$: In the second approach we model the hidden variables $\mathbf{u}$ and $\mathbf{g}(\mathbf{X})$ explicitly and consider the joint density function $f_{\mathbf{y},\mathbf{u},\mathbf{g}}$ of $\mathbf{y}, \mathbf{u}$ and $\mathbf{g}(\mathbf{X})$ which equals

$$\begin{aligned} f_{\mathbf{y},\mathbf{u},\mathbf{g}(\mathbf{X})}(\mathbf{y},\mathbf{u},\mathbf{g}(\mathbf{X})) &= f_{\mathbf{y}|\mathbf{u},\mathbf{g}(\mathbf{X})}(\mathbf{y}) \cdot f_\mathbf{u}(\mathbf{u}) \cdot f_\mathbf{g}(\mathbf{g}(\mathbf{X})) \\ &= f_\mathbf{e}(\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \mathbf{g}(\mathbf{X})) \cdot f_\mathbf{u}(\mathbf{u}) \cdot f_\mathbf{g}(\mathbf{g}(\mathbf{X})) \end{aligned}$$

$$= c \cdot \exp\left(-\frac{1}{2} \cdot \left[\frac{1}{\sigma_e^2}\|\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Zu} - \mathbf{g}(\mathbf{X})\|^2\right]\right)$$
$$\cdot \exp\left(-\frac{1}{2} \cdot \left[\frac{1}{\sigma_u^2}\mathbf{u}^T\mathbf{A}^{-1}\mathbf{u}\right]\right) \cdot \exp\left(-\frac{1}{2} \cdot \left[\mathbf{g}(\mathbf{X})^T\mathbf{K}_{\nu,h,\sigma_K}^{-1}\mathbf{g}(\mathbf{X})\right]\right)$$

with

$$c^{-1} = (2\pi)^{n+q/2}\sigma_e^n \cdot \sigma_u^q(\det \mathbf{A})^{1/2} \cdot (\det \mathbf{K}_{\nu,h,\sigma_K})^{1/2}.$$

Here, we have to estimate the parameters $\boldsymbol{\beta}, \sigma_u, \sigma_e, \nu, h, \sigma_K$ and the hidden variables $\mathbf{u}$ and $\mathbf{g}(\mathbf{X})$. Note that in this approach we consider $\mathbf{u}$ and $\mathbf{g}(\mathbf{X})$ to be parameters that have to be estimated via ML in the first kriging step. Therefore, we maximize the loglikelihood J of the density function $f_{\mathbf{y},\mathbf{u},\mathbf{g}}$, *i.e.* we maximize

$$J = \log(c) - \frac{1}{2} \cdot \left[\frac{1}{\sigma_e^2}\|\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Zu} - \mathbf{g}(\mathbf{X})\|^2 + \frac{1}{\sigma_u^2}\mathbf{u}^T\mathbf{A}^{-1}\mathbf{u} + \mathbf{g}(\mathbf{X})^T\mathbf{K}_{\nu,h,\sigma_K}^{-1}\mathbf{g}(\mathbf{X})\right] \tag{3.3}$$

with respect to $\boldsymbol{\beta}, \mathbf{u}$ and $\mathbf{g}(\mathbf{X})$. Taking the derivatives with respect to $\beta, \mathbf{u}$ and $\mathbf{g}(\mathbf{X})$ leads to the linear system given in eq. (3.2) which yields estimators for $\boldsymbol{\beta}, \mathbf{u}$ and $\mathbf{g}(\mathbf{X})$. When using these estimates in eq. (3.3), the value of J depends only on $\sigma_u, \sigma_e, \nu, h$ and σ_K. Thus, J can be maximized numerically with respect to these parameters, leading to estimates for $\boldsymbol{\beta}, \sigma_u, \sigma_e, \nu, h, \sigma_K, \mathbf{u}$ and $\mathbf{g}(\mathbf{X})$. According to the kriging philosophy, we now assume the values of the estimators (especially the value of the estimator for $\mathbf{g}(\mathbf{X})$) to be the true ones, and $g(\mathbf{x}_0)$ is predicted via $\hat{g}(\mathbf{x}_0) = \mathbf{a}_\mathbf{g}^T\mathbf{g}(\mathbf{X})$ by the BLUP principle. That is, we minimize

$$\mathbb{E}(\hat{g}(\mathbf{x}_0) - g(\mathbf{x}_0))^2 \longrightarrow \text{ min!}$$

with the linear estimator

$$\hat{g}(\mathbf{x}_0) = \mathbf{a}_\mathbf{g}^T\mathbf{g}(\mathbf{X}).$$

This approach is called “simple kriging” (Cressie, 1990, 1993; Chilès & Delfiner, 1999). Note that $\hat{g}(\mathbf{x}_0)$ is always unbiased. The solution is

$$\hat{g}(\mathbf{x}_0) = \mathbf{K}_0^T\mathbf{K}_{\nu,h,\sigma_K}^{-1}\mathbf{g}(\mathbf{X}). \tag{3.4}$$

Finally, the predicted GV is given by $\widehat{g(\mathbf{x}_0) + \mathbf{z}_0^T\mathbf{u}} = \hat{g}(\mathbf{x}_0) + \mathbf{z}_0^T\hat{\mathbf{u}}$, where $\hat{\mathbf{u}}$ is the estimator obtained in the iterative procedure described above.

Reference model (genomic BLUP): This approach performs a genomic BLUP based on the model

$$\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Zu} + \tilde{\mathbf{X}}\mathbf{g} + \mathbf{e},$$

which leads to the kriging system

$$\begin{bmatrix} \mathbf{W} & \sigma_u^2\mathbf{ZAZ}^T + \sigma_g^2\tilde{\mathbf{X}}\mathbf{G}\tilde{\mathbf{X}}^T + \sigma_e^2\mathbf{I} \\ \mathbf{0} & \mathbf{W}^T \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \sigma_u^2\mathbf{ZAz}_0 + \sigma_g^2\tilde{\mathbf{X}}\mathbf{G}\tilde{\mathbf{x}}_0 \\ \mathbf{0} \end{bmatrix}$$

and predicting $\widehat{\mathbf{z}_0^T\mathbf{u} + \tilde{\mathbf{x}}_0^T\mathbf{g}} = \mathbf{a}^T\mathbf{y}$.

Here, $\boldsymbol{\beta}, \mathbf{e} \sim \mathcal{N}(0,\sigma_e^2\mathbf{I}), \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2\mathbf{A}), \mathbf{W}$ and $\mathbf{Z}$ are defined as in the previous approaches. The vector $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{G})$ is multivariate normal with $\mathbf{G}$ being a genomic relationship matrix calculated by using the approach of VanRaden (2008). (For the definition of the genomic relationship matrix see the formulas in section 3.6.) The matrix $\tilde{\mathbf{X}}$ is a known incidence matrix whose rows consist of unit vectors with one component being 1 and all the others zero, indicating the respective position in the $\mathbf{g}$-vector. Variance components for this model are estimated via ML.

3.3 Simulation study

In a first step, four types of simulations were performed differing in the hypothetical gene-action scenario: “additive”, “additive-dominance” with two different ratios of dominance variance to additive variance and “epistasis”. For each scenario 50 independent simulations were run, resulting in 50 data sets per scenario.

The simulation process basically followed that of Meuwissen *et al.* (2001), Solberg *et al.* (2008) and Long *et al.* (2010).

3.3.1 Population and genome

In each scenario, the population evolved during 1,000 generations of random mating and random selection with a population size of 100 (50 males and 50 females) in each generation to reach a mutation-drift balance. After 1,000 generations, the population size was increased to 500 at generation $t = 1{,}001$ by mating each male with 10 females, with one offspring per mating pair. In generations $t = 1{,}002; \ldots ; 1{,}011$ offspring were born from random mating of individuals of the previous generation. The 1,500 individuals of generations 1,008, 1,009 and 1,010 were used as estimation set, the 500 individuals of generation 1,011 formed the validation set for which total GVs were predicted. Pedigree data were recorded for individuals of the last 10 generations. SNP data of individuals were recorded both for the estimation- and the validation set. Phenotypes were only stored for individuals of the estimation set.

The simulated genome consisted of 1 chromosome of length 1 Morgan, containing 100 equally spaced putative QTL. Each QTL was flanked by 30 equally spaced SNP markers resulting in 3,030 markers (M) in total. The layout of the chromosome was therefore given by

$$M_1 - M_2 - \ldots - M_{30} - QTL_1 - M_{31} - \ldots - M_{60} - QTL_2 - \ldots - QTL_{100} - M_{3{,}001} - \ldots - M_{3{,}030}.$$

Starting with monomorphic loci in the base generation, mutation rates at QTL and SNP markers were 2.5×10^{-3} per locus per generation ($t = 1; \ldots ; t = 1{,}000$), to obtain an adequate number of segregating (biallelic) loci. On average, simulation resulted in 2,745 segregating markers and 98 segregating QTL in generation $t = 1{,}001$. Only segregating markers and QTL were considered in the following generations. True total GVs were obtained by summing up the QTL effects resulting from the following three gene-action models.

3.3.2 Three different gene-action models

Additive scenario A: Each QTL locus had an additive effect only, without dominance or epistasis. The additive effect (a) was equal to the allele substitution effect, such that for genotypes QQ, Qq and qq their GVs were $2a, a$ and 0, respectively. The value of a at each QTL locus was sampled from a normal distribution $\mathcal{N}(0{,}0.1)$.

Additive-dominance scenarios AD1 and AD2: Each QTL locus had both an additive and a dominance effect. Two different scenarios were considered, setting the ratio of dominance variance to additive variance at each QTL to $\delta = 1$ or $\delta = 2$. The additive effects (a) were obtained as in the additive scenario. Given the additive effect a_i and allele frequency p_i at the ith locus, its dominance effect (d_i) was determined by solving the equation

$$\delta = \frac{\sigma^2_{D,i}}{\sigma^2_{A,i}} = \frac{(2p_i(1-p_i)d_i)^2}{2p_i(1-p_i)[a_i + ((1-p_i)-p_i)d_i]^2},$$

see Falconer & Mackay (1996). Genetic values at that locus were then given by $2a, a+d$ and 0 for genotypes QQ, Qq and qq respectively.

For simplicity, independence between QTL was assumed and, as a result, the total additive (dominance) variance was summed over all loci.

Epistasis scenario E: In this model there was no additive or dominance effect at any of the individual QTL. Epistasis existed only between pairs of QTL. The forms of epistasis included additive × dominance ($A \times D$), dominance × additive ($D \times A$), and dominance × dominance ($D \times D$). Additive and ($A \times A$) epistatic effects were excluded, to prevent the additive variance from dominating the total genetic variance.

All segregating QTL were involved in epistatic interactions. QTL were randomly chosen to form pairs and each pair was assigned an ($A \times D$) interaction effect ℓ_{AD}, a ($D \times A$) interaction effect ℓ_{DA} and a ($D \times D$) interaction effect ℓ_{DD}, which were all equal and sampled from a normal distribution $\mathcal{N}(0{,}4)$. Given a pair of QTL ($i = 1{,}2$), its epistatic value was given by

$$\ell_{AD} x_1 z_2 + \ell_{DA} z_1 x_2 + \ell_{DD} z_1 z_2,$$

where x_i and z_i were additive and dominance codes at locus i, respectively. For genotype QQ at locus i, $x_i = 1, z_i = -0.5$; for $Qq, x_i = 0, z_i = 0.5$; and for $qq, x_i = -1, z_i = -0.5$, compare Cordell (2002). The total GV was the sum of the epistatic values produced by the QTL pairs.

Note that although no additive, dominance and ($A \times A$) epistatic effects were explicitly simulated, the model still generated additive (σ^2_A), dominance (σ^2_D) and epistatic ($\sigma^2_{A\times A}, \sigma^2_{A\times D}, \sigma^2_{D\times A}, \sigma^2_{D\times D}$) variances. The procedure of estimating these variance components followed Cockerham (1954), assuming independence between two loci of each QTL pair and between QTL pairs.

On average, simulation in the epistatic scenario resulted in a broad-sense heritability of 0.84. Furthermore, 30% of the total genetic variance was attributed to additive effects, 27%

was due to dominance effects, 14% was due attributed to $(A \times A)$-effects, 25% was due to $(D \times A)$- and $(A \times D)$-effects and 4% was due to $(D \times D)$-effects.

In all scenarios phenotypic records were obtained by adding a normally distributed $\mathcal{N}(0, \sigma_e^2)$ residual term to the total GVs of the individuals. The environmental variance σ_e^2 was obtained such that the narrow sense heritability was 0.25 in all scenarios.

3.3.3 Additional scenarios

Four additional scenarios based on scenario AD1 were simulated, to analyze the influence of the number of chromosomes, the QTL architecture, the SNP density and a polygenic effect on the prediction accuracy:

- *Scenario AD1.2*: Three chromosomes of length 1/3 Morgan were simulated, each containing 33 equally spaced QTL and 1,000 SNPs.
- *Scenario AD1.3*: Three chromosomes of length 1/3 Morgan were simulated, each of them containing 1,000 SNPs and the first two of them containing 50 equally spaced QTL. The third chromosome contained no QTL.
- *Scenario AD1.4*: The same as scenario AD1.2 but with each chromosome containing 33 equally spaced QTL and 3,000 SNPs.
- *Scenario AD1.5*: The same as scenario AD1, but additionally a polygenic effect u was simulated, starting from generation 1,006. Here, the ratio of additive QTL variance to polygenic variance was set to 3. The polygenic effect u of an offspring was calculated as $0.5 \cdot (u_{mother} + u_{father}) + m$, where m is its Mendelian sampling term drawn from a normal distribution

  $$\mathcal{N}(0, 0.25 \cdot (2 - (F_{mother} + F_{father})) \cdot \sigma_{poly}^2),$$

  with F_{mother} and F_{father} being the inbreeding coefficients of the corresponding mother and father. Here, the true total GV was obtained by summing up the QTL effects and the polygenic effect.

3.3.4 Statistical analyses

The three methods were compared for their accuracy of predicting the true GVs of the individuals in generation $t = 1{,}011$. For this we applied the three approaches described in section 3.2.3 to the 50 simulated data sets consisting of 5,500 individuals, the last 5,000 of them having pedigree information and the last 2,000 of them being fully genotyped, as described in the previous section. Total GVs of the non-phenotyped individuals in generation $t = 1{,}011$ (validation set) were predicted. Thereby, parameters and hidden variables were estimated with the help of 1,500 individuals (generations 1,008–1,010, estimation set).

All approaches were implemented using R software (R Development Core Team, 2012; Ihaka & Gentleman, 1996). The ML estimation of the parameters and hidden variables was done using the R-package “RandomFields”, Version 2.0.23 (Schlather, 2001–2009), and its function “fitvario”. The function “fitvario” determines the ML by the function “optim” of R with automatically created starting values.

All models were run on a 1.9 GHz PC running Linux. On average, computing times per data set ranged from approximately 20 minutes (genomic BLUP) over 77 minutes (universal kriging) to 227 minutes (simple kriging), but no special efforts were made to achieve computational efficiency at this stage.

For each method and each gene-action scenario, we computed the correlation between the predicted and the true GVs. This was done both for the estimation set of 1,500 individuals and for the validation set of 500 individuals. In addition, we calculated the average true GV of the 50 individuals with the highest predicted GVs in the validation set. Finally, results were summarized by averaging over the 50 data sets and a paired t-test was applied to test for significant differences between each pair of characteristics at the 1% significance level.

One data set and the corresponding R-code for the prediction of GVs are available on `http://www.stochastik.math.uni-goettingen.de/~schlather/genoKriging/`.

3.4 Results of the simulation study

The results of 50 replicates for the different gene-action models and scenarios are shown in Tables 3.2–3.3.

Table 3.2: Average correlations between predicted and true GVs

scenario	set	universal kriging	simple kriging	genomic BLUP
A	estimation set	$0.801_{\alpha}{}^{1,2}$ (0.005)	0.772_{β} (0.009)	0.815_{γ} (0.004)
	validation set	0.773_{α} (0.005)	0.731_{β} (0.008)	0.776_{γ} (0.005)
AD1	estimation set	0.754_{α} (0.004)	0.652_{β} (0.009)	0.670_{β} (0.004)
	validation set	0.571_{α} (0.006)	0.530_{β} (0.010)	0.558_{γ} (0.007)
AD2	estimation set	0.854_{α} (0.004)	0.624_{β} (0.013)	0.621_{β} (0.005)
	validation set	0.490_{α} (0.007)	0.447_{β} (0.009)	0.457_{β} (0.007)
E	estimation set	0.910_{α} (0.009)	0.631_{β} (0.015)	0.681_{γ} (0.006)
	validation set	0.468_{α} (0.006)	0.411_{β} (0.008)	0.437_{γ} (0.007)

[1] Results were averages of 50 replicates. Standard errors of the means in parentheses.
[2] Different small Greek letters in the rows indicate significant differences (1 % level of significance).

In the additive scenario, universal kriging yields a correlation between predicted and true simulated GVs which is almost as high as the correlation obtained by the reference method genomic BLUP, both in the estimation and in the validation set (*cf.* Table 3.2), while simple kriging yields the lowest correlations both in the estimation and in the validation set.
These results are similar to the findings of Piepho (2009) and Schulz-Streeck & Piepho (2010) who also report that for an additive true genetic model the prediction accuracies for ridge regression (with covariance-structures based on relationship matrices) and spatial models (with covariance-structures based on covariance functions) are similar.

In the AD and E scenarios, universal kriging outperforms genomic BLUP in both estimation and validation set by showing the highest average correlations. The difference in correlations of universal kriging and genomic BLUP is highest in the E scenario and the scenario with the higher ratio of dominance to additive variance (≈ 0.03 for the results of the validation set, which is an increase of accuracy by approximately 7%).

Scatterplots of the correlations of the 50 replicates for the different methods and scenarios are shown in Figure 3.2, which also demonstrate the better performance of universal kriging in the presence of dominance and epistasis. With the degree of non-additivity ((E, AD2) > AD1 > A) the accuracy of prediction in the validation set compared to the estimation set deteriorates.

Comparing the average true GV of the 50 individuals (10%) ranked best by prediction in the validation set (*cf.* Table 3.3), universal kriging and genomic BLUP yield results which are not significantly different from each other both in the A and AD scenarios, while universal kriging outperforms genomic BLUP in the E scenario. Again, simple kriging performs worst in all scenarios apart from AD1.

Table 3.3: Average true GVs of the 50 highest ranked individuals (validation set)

scenario	universal kriging	simple kriging	genomic BLUP
A	2.420_α[1,2] (0.259)	2.291_β (0.261)	2.432_α (0.258)
AD1	1.754_α (0.182)	1.648_α (0.186)	1.728_α (0.177)
AD2	1.720_α (0.172)	1.563_β (0.178)	1.612_α (0.171)
E	6.410_α (0.502)	5.847_β (0.476)	5.893_β (0.485)

[1] Results were averages of 50 replicates. Standard errors of the means in parentheses.
[2] Different small Greek letters in the rows indicate significant differences (1 % level of significance).

All three methods, being unbiased by definition, show almost no empirical bias of total GVs (results not shown).

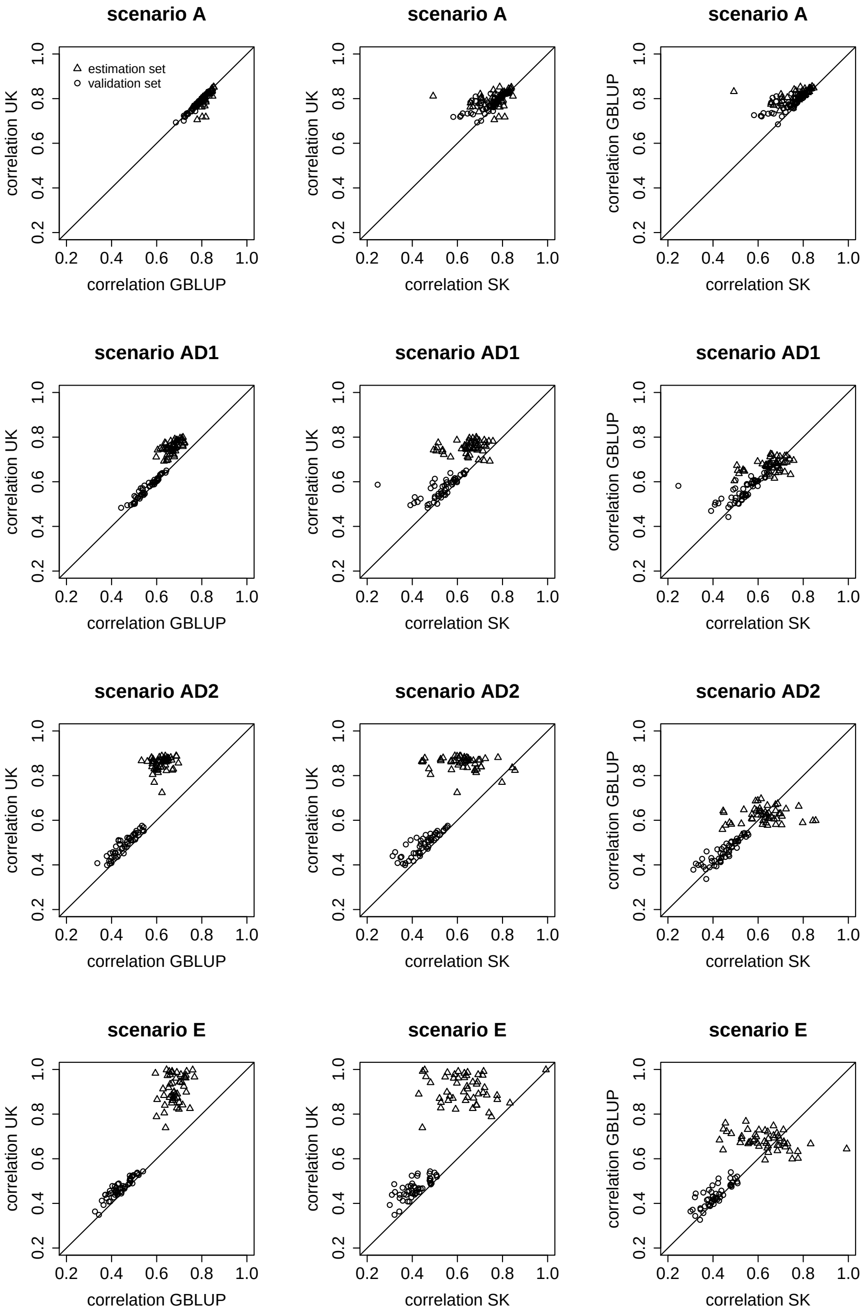

Figure 3.2: Scatterplot of the correlations between true and predicted GVs both for the estimation and the validation set and for the different scenarios (additive (A), additive-dominance with ratio of dominance to additive variance of 1 or 2 (AD1 and AD2) and epistasis (E)) to compare. Scatterplots are produced to compare universal kriging (UK) with genomic BLUP (GBLUP), UK with simple kriging (SK) and UK with GBLUP.

The results of the additional scenarios AD1.2 to AD1.5 indicate that the predictive ability of the universal kriging approach is robust with respect to the number of chromosomes, the QTL distribution, the SNP density and the inclusion of a polygenic effect (*cf.* Table 3.4). In scenario AD1.4 with higher SNP density the absolute values of correlations between true and predicted GVs are slightly higher compared to scenario AD1.2 with lower SNP density. In scenario AD1.5 the absolute values of correlations between true and predicted GVs are lower for all three methods.

Table 3.4: Additional scenarios: average correlations between predicted and true GVs

	universal kriging		simple kriging		genomic BLUP	
scenario	est. set[1]	val. set[2]	est. set	val. set	est. set	val. set
AD1	0.754_{α}[3,4] (.004)	0.571_{α} (.006)	0.652_{α} (.009)	0.530_{α} (.010)	0.670_{α} (.004)	0.558_{α} (.007)
AD1.2	0.751_{α} (.004)	0.550_{α} (.006)	0.627_{α} (.007)	0.511_{α} (.008)	0.666_{α} (.005)	0.541_{α} (.007)
AD1.3	0.753_{α} (.005)	0.554_{α} (.010)	0.630_{α} (.009)	0.518_{α} (.011)	0.670_{α} (.006)	0.543_{α} (.010)
AD1.4	0.758_{α} (.004)	0.567_{α} (.007)	0.642_{α} (.007)	0.531_{α} (.008)	0.677_{α} (.005)	0.558_{α} (.007)
AD1.5	0.718_{β} (.004)	0.528_{β} (.006)	0.623_{α} (.009)	0.496_{α} (.008)	0.666_{α} (.005)	0.518_{β} (.007)

[1] Estimation set
[2] Validation set
[3] Results were averages of 50 replicates. Standard errors of the means in parentheses.
[4] Different small Greek letters in the columns indicate significant differences (1 % level of significance).

3.5 Discussion

Overall, results indicate the superiority of universal kriging over genomic BLUP in the presence of non-additive effects. Simple kriging was shown to have a poorer predictive ability compared to universal kriging and genomic BLUP in all considered gene-action models and scenarios.

The poorer predictive ability of simple kriging is most likely due to the high number of parameters estimated in the first kriging step and the resulting numerical difficulties in optimization. In simple kriging 3,505 parameters $(\mathbf{u}, \mathbf{g}(\mathbf{X}), \sigma_e^2, \sigma_u^2, \sigma_K^2, \nu, h)$ are estimated compared to only 5 parameters in universal kriging and 3 parameters in genomic BLUP. The poor performance of simple kriging and the influence of the high-dimensional parameter space need further investigations, especially, as simple kriging is known to work well in low-dimensional geostatistical frameworks.

The simulation study is primarily meant as a “proof of concept”. Results demonstrate that the suggested kriging procedures based on the Matérn function are able to yield competitive results, despite the fact that the modeling of the genomic part of the data by use of the

Matérn function follows a completely different reasoning than in the usual methods. This also demonstrates the flexibility of the basic kriging principle.

The importance of the Matérn family is highlighted by Stein (1999), who recommends the use of the Matérn model in the context of prediction of spatial data. The Matérn model has been widely used in other areas of research, see Guttorp & Gneiting (2006) for a historical excursion. One of the most important reasons for adopting the Matérn model is the inclusion of the parameter ν in the model which controls the smoothness of the underlying random field. Whereas Stein (1999) advocates the simultaneous estimation of all relevant parameters via (restricted) maximum likelihood, Ruppert *et al.* (2003) and Nychka (2000) remark that the likelihood-based estimation of h and ν may lead to problems as both parameters enter in a nonlinear fashion which may cause the ML fitting to be computationally intensive. Our experience so far indicates that the simultaneous estimation of all relevant parameters is feasible.

As an alternative to the ML estimation of parameters, one could also use REML (Patterson & Thompson, 1971) to adjust for the loss of degrees of freedom caused by the fixed effects and to produce less biased estimates. In our simulation study there is only one fixed effect (*i.e.* $\boldsymbol{\beta}$ is a scalar and $\mathbf{W} = (1, \ldots, 1)^T$), such that there will be little difference between REML and ML estimates for variance components in the reference method GBLUP (Abney *et al.*, 2000; Webster *et al.*, 2006; Bonate, 2006; Ruppert *et al.*, 2003). This is also mostly the case in practical applications, where highly accurately predicted GVs are used as phenotypes and only an overall mean is included in the model. With respect to the parameter estimates in the kriging approaches using the Matérn function, it is not clear whether REML is preferable to ML, as the parameters h and ν enter in a nonlinear fashion.

3.5.1 Relation between the Matérn covariance function and the covariance matrix of VanRaden (2008)

To investigate the general relationship between covariance matrices based on the Matérn function and the genomic relationship matrix of VanRaden (2008), we consider the so-called variograms which are often used in spatial statistics (*cf.* Wackernagel (2003); Chilès & Delfiner (1999) for instance).

For a random field $\{g(\mathbf{x}), \mathbf{x} \in \mathbb{R}^s\}$, the theoretical variogram is defined by $\gamma(\mathbf{x}_i, \mathbf{x}_j) = 0.5\mathbb{E}((g(\mathbf{x}_i) - g(\mathbf{x}_j))^2)$ for $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^s$. If $\mathrm{Var}(g(\mathbf{x}_i)) = \sigma_g^2$ and $\mathbb{E}(g(\mathbf{x}_i)) = 0$ for all $\mathbf{x}_i \in \mathbb{R}^s$, the variogram is given by

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = \sigma_g^2 - \mathrm{Cov}(g(\mathbf{x}_i), g(\mathbf{x}_j))$$

for $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^s$. If further $\mathrm{Cov}(g(\mathbf{x}_i), g(\mathbf{x}_j))$ only depends on the Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|$, the variogram γ can be considered as a function on $[0, \infty)$.

In section 3.6 we show that in a limiting case (in which the number of SNPs tends to infinity) the covariance structure of VanRaden (2008) only depends on the Euclidean distance between the SNP vectors and that the corresponding variogram is a quadratic function on $[0, \infty)$.

In all kriging procedures, ν was estimated to be larger than 5, indicating an approximately

Gaussian form of the covariance function. In fact $K_{\nu,h,\sigma_K}(x_i,x_j) = \sigma_K^2 \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2h^2}\right)$ for $\nu \to \infty$. The corresponding variogram for the Matérn function is then given by

$$\gamma_{\text{Matern}}(x) = \sigma_K^2\left(1 - \exp\left(-\frac{x^2}{2h^2}\right)\right) \approx \frac{\sigma_K^2}{2h^2}x^2$$

for $x \in [0,\infty)$ by Taylor expansion up to the second derivative around zero. This means the corresponding variogram of the Matérn function is approximately quadratic for $\frac{x}{h}$ small and for $\nu \to \infty$. If both variograms, the one induced by VanRaden's covariance structure and γ_{Matern}, were exactly quadratic, the corresponding covariance matrices would be linear transformations of each other. The equivalence of a quadratic covariance function and the second order Taylor expansion of the Gaussian model has also been noted by Piepho (2009).

Note that the Matérn covariance function is at least three times differentiable for $\nu > 1.5$ (Guttorp & Gneiting, 2006), such that it is still possible to derive a second order Taylor expansion for $1.5 < \nu < \infty$, leading to a quadratic variogram for small distances x as well.

3.5.2 Using linear transformations of covariance matrices leads to linearly transformed predicted genetic values

In this context another interesting relation can be shown: There is a linear relation between the predicted GVs, if there is a linear relation between the phenotypic covariance matrices $\mathbf{B}$ and $\tilde{\mathbf{B}}$ and a linear relation between the covariance vectors $\mathbf{B}_0$ and $\tilde{\mathbf{B}}_0$ on the right hand sides of the kriging systems under the assumption that $\mathbf{W} = (1,\ldots,1)^T = \mathbf{j}$ and that $\mathbf{V} := \begin{bmatrix}\mathbf{W} & \mathbf{B}\\ 0 & \mathbf{W}^T\end{bmatrix}$ is invertible: In detail, it can be shown

$$\tilde{\mathbf{a}} = \frac{\tilde{d}}{d}\cdot\mathbf{a}$$

for the linear (kriging) systems

$$\begin{bmatrix}\mathbf{j} & \mathbf{B}\\ 0 & \mathbf{j}^T\end{bmatrix}\cdot\begin{bmatrix}\lambda\\ \mathbf{a}\end{bmatrix} = \begin{bmatrix}\mathbf{B}_0\\ 0\end{bmatrix} \tag{3.5}$$

and

$$\begin{bmatrix}\mathbf{j} & d\mathbf{B}+c\mathbf{J}\\ 0 & \mathbf{j}^T\end{bmatrix}\cdot\begin{bmatrix}\tilde{\lambda}\\ \tilde{\mathbf{a}}\end{bmatrix} = \begin{bmatrix}\tilde{d}\mathbf{B}_0+\tilde{c}\mathbf{j}\\ 0\end{bmatrix} \tag{3.6}$$

with $d \neq 0$ and $\mathbf{J} = (\mathbf{j},\ldots,\mathbf{j})$, from which we get $\tilde{\text{GV}} = \frac{\tilde{d}}{d}\cdot\text{GV}$. The proof of this result is given in section 3.7.

This general result has important practical implications: It is shown that predictions resulting from the two systems (3.5) and (3.6) are identical although a constant (c and $\tilde{c}$) is added to the phenotypic covariance matrix or the covariance vector on the right hand side of the kriging system, or to both. In the genetic context, such a modification changes relevant population parameters, like heritabilities as well as genetic and phenotypic correlations. Despite this, predicted GVs remain completely unaffected.

Scaling the phenotypic covariance matrix and the covariance vectors by a factor (d and $\tilde{d}$) also changes the heritability, but is shown to lead to a mere linear transformation of the GVs, thus providing an identical ranking of individuals according to their predicted GVs. However, results obtained from such a scaled system might lead to a higher or lower level of mean squared errors.

As stated before, solving the kriging systems is equivalent to solving the corresponding MME. Hence, we have also proved that the solutions $\hat{\mathbf{u}}$ and $\widehat{\mathbf{g}(\mathbf{X})}$ of the MME are scaled by the factor $\frac{\tilde{d}}{d}$, if the phenotypic covariance matrix and the covariance matrix of $\mathbf{Zu} + \mathbf{g}(\mathbf{X})$ are linearly transformed.

To our knowledge, the above theoretical result (including the scaling factors d and $\tilde{d}$) has not been proved elsewhere in this explicit form, but some authors refer to the invariance of the predictions to the addition of a multiple of the matrix $\mathbf{J}$: It is well-known that in ordinary kriging with constant mean one only needs to know the covariance function up to a constant (Matheron, 1971; Christensen, 1990). Kitanidis (1993) discusses in the context of so-called “generalized covariance functions” the variability among the covariance functions that behave identically in terms of prediction. The invariance to the addition of a multiple of $\mathbf{J}$ in a mixed model context is also mentioned in Piepho (2009).

3.5.3 Reproducing Kernel Hilbert Space approach

In this subsection we contrast our approach to the Reproducing Kernel Hilbert Spaces approach of Gianola & van Kaam (2008). Stein (1999) strongly advocates use of the Matérn family because of the wide range of smoothness controlled by the smoothness parameter $0 < \nu < \infty$. In our study ν was estimated to be larger than 5 in all kriging procedures, indicating an approximately Gaussian form of the covariance function, the one which has been used by Gianola & van Kaam (2008). Gianola & van Kaam (2008) use the same model as in (3.1) except for the assumption that g is a Gaussian random function. They consider the functional

$$J(\mathbf{g}|s) = \frac{1}{\tilde{\sigma}_e^2} \sum_{i=1}^{n} (y_i - \mathbf{w}_i^T \boldsymbol{\beta} - \mathbf{z}_i^T \mathbf{u} - g(\mathbf{x}_i))^2 + \frac{s}{2} \|\mathbf{g}(\cdot)\|_{\mathfrak{H}}$$

where g and $y_i - \mathbf{w}_i^T \boldsymbol{\beta} - \mathbf{z}_i^T \mathbf{u}$ are implicitly assumed to be elements of a Reproducing Kernel Hilbert Space $\mathfrak{H}$ for fixed $\boldsymbol{\beta}$ and $\mathbf{u}$. Then, the representer theorem (Schölkopf *et al.*, 2001) states that the minimizer of $J(\mathbf{g}|s)$ has the form

$$\hat{g}(\mathbf{x}_0) = \sum_{j=1}^{n} \alpha_j K(\mathbf{x}_0, \mathbf{x}_j) = \boldsymbol{\alpha}^T \mathbf{K}_0, \tag{3.7}$$

where the α_i's are unknown coefficients. The function to be minimized becomes

$$J(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\alpha}|s) = \frac{1}{2\sigma_e^2} \|\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Zu} - \mathbf{K}\boldsymbol{\alpha}\|^2 + \frac{s}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}.$$

Gianola & van Kaam (2008) state further that a random-effects treatment of $\mathbf{u}$ leads to the functional

$$J(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\alpha}|s) = \frac{1}{2\tilde{\sigma}_e^2}\|\mathbf{y} - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \mathbf{K}\boldsymbol{\alpha}\|^2 + \frac{1}{2\tilde{\sigma}_u^2}\mathbf{u}^T\mathbf{A}^{-1}\mathbf{u} + \frac{s}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha},$$

which then has to be minimized. Taking the gradients of $J(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\alpha}|s)$ with respect to $\boldsymbol{\beta}, \mathbf{u}$ and $\boldsymbol{\alpha}$ and setting them to zero leads to the following linear system of equations:

$$\begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} & \mathbf{W}^T\mathbf{K} \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z} + \frac{\tilde{\sigma}_e^2}{\tilde{\sigma}_u^2}\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{K} \\ \mathbf{K}^T\mathbf{W} & \mathbf{K}^T\mathbf{Z} & \mathbf{K}^T\mathbf{K} + s\tilde{\sigma}_e^2\mathbf{K} \end{bmatrix} \cdot \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{K}^T\mathbf{y} \end{bmatrix} \tag{3.8}$$

By equating $\widehat{\mathbf{g}(\mathbf{X})} = \mathbf{K}\hat{\boldsymbol{\alpha}}$, $\sigma_e^2 = s\tilde{\sigma}_e^2$ and $\sigma_u^2 = s\tilde{\sigma}_u^2$, eqs. (3.2) and (3.8) are obviously identical, as well as eqs. (3.4) and (3.7). Finally, Gianola & van Kaam (2008) proceed with embedding the above approach into a Bayesian framework.

The approach of Gianola & van Kaam (2008) and our approach are different in that we maximize the full likelihood whereas they drop the summand $\log(c)$ in eq. (3.3). Note that c depends on the unknown parameters, *i.e.* the variance components and the parameters of the Matérn covariance function. Dropping the summand $\log(c)$ therefore leads to different estimates of the parameters. Scheuerer (2011) argues that the factor c might be included even in the framework of Reproducing Kernel Hilbert Spaces. Hence, maximizing J in eq. (3.3), is partially justified even if the normal assumption for the e_i does not hold.

3.5.4 Further options

The general non-parametric approach of basing the prediction on a covariance function offers a number of possibilities for more differentiated modeling. While in spatial statistics using the Euclidean distance is a natural choice, other distance metrics (Reif *et al.*, 2005) may be more adequate in the genomic context. With dense marker maps it is found that the genome is structured in haplotype blocks of varying length (The International HapMap Consortium, 2005; Qanbari *et al.*, 2010) within which the loci are in high linkage disequilibrium, *i.e.* genotypes are highly correlated. Here, it might be adequate to account for this non-independence in the definition of the scale, since otherwise highly correlated loci will lead to a massive double counting. A further option is to implement a feature selection which could *e.g.* give a higher weight to SNPs that are positioned in genomic regions which are found to be relevant for the physiological pathways (Wang *et al.*, 2007) underlying the studied trait complex.

3.5.5 Total GVs

Prediction of the total GV of an individual, including non-additive components, is of different relevance in different fields. In animal breeding, the value of a breeding animal is mostly determined by its so-called breeding value which is purely additive. While it is possible to predict non-additive genetic components even in pedigree-based estimation procedures (see *e.g.* Hoeschele (1991); de Boer & Hoeschele (1993)), these components are in general not transmitted to the offspring and therefore are mostly considered as nuisance parameters in animal breeding.

In plant breeding, prediction of the total GV as part of the phenotype is more relevant, especially since the biological nature of some crop species and/or reproductive biotechnologies allow an identical reproduction (cloning) of given genotypes. Complex gene models including dominance and epistasis might be especially useful in predicting crossbred performance, but the relevance is rather diverse across the agriculturally used plants (Holland, 2001).

It was recently suggested that under polygenic inheritance the additive part is the dominating genetic component (Hill *et al.*, 2008) and that under directional selection the rate of change is largely determined by the additive genetic variance, so that attempts to include non-additive terms in prediction might be, at best, useless or even harmful (Crow, 2010). These arguments pertain both to animal and plant breeding and need careful consideration based on empirical evidence.

Predicting the genetic disposition in humans in the context of preventive and personalized medicine using whole genome markers is a relatively new and controversial topic (see de los Campos *et al.* (2010*a*) for a review). The main motivation to consider such approaches comes from the phenomenon that even in extremely large scale studies the genetic background of complex diseases cannot be sufficiently determined with classical mapping approaches (the so-called "case of the missing heritability"; Maher (2008)). Disposition for complex diseases is assumed to be affected to a considerable extent by non-additive allelic interactions, and hence models allowing for such interactions are expected to yield improved predictions compared to purely additive models.

3.6 Relation between the Matérn covariance function and the covariance matrix of VanRaden (2008)

We show that the covariance structure of VanRaden (2008) leads to a quadratic variogram γ in a limiting case. The covariance matrix of VanRaden (2008) is defined as

$$\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})^T}{2\sum_{j=1}^{s} p_j(1-p_j)},$$

where $\mathbf{M}$ is the $(n \times s)$-matrix of SNP vectors for the n animals with s SNPs coded by -1,0,1 and the jth column of $\mathbf{P}$ is $(2(p_j-0.5),\ldots,2(p_j-0.5))^T$, where p_j is the frequency of the second allele at locus j.

Let $\tilde{\mathbf{P}} = (2(p_1-0.5),\ldots,2(p_s-0.5))$ and let $D = 2\sum_{j=1}^{s} p_j(1-p_j)$. In the genomic BLUP model we assumed $\mathbf{g} \sim \mathcal{N}(0, \sigma_g^2 \mathbf{G})$. It follows easily that

$$\begin{aligned}\mathrm{Cov}(g_i, g_j) &= \frac{\sigma_g^2}{D}(\mathbf{m}_{i\bullet} - \tilde{\mathbf{P}})(\mathbf{m}_{j\bullet}^T - \tilde{\mathbf{P}}^T) \\ &= \frac{\sigma_g^2}{D}\left(-\frac{1}{2}\|\mathbf{m}_{i\bullet} - \mathbf{m}_{j\bullet}\|^2 + \frac{1}{2}\|\mathbf{m}_{i\bullet} - \tilde{\mathbf{P}}\|^2 + \frac{1}{2}\|\mathbf{m}_{j\bullet} - \tilde{\mathbf{P}}\|^2\right),\end{aligned} \tag{3.9}$$

where $\mathbf{m}_{i\bullet}$ denotes the ith row of $\mathbf{M}$ and $\|\cdot\|$ is the Euclidean norm. Consider $\mathbf{M}_{ij}$ as a random variable with values -1,0,1 and corresponding probabilities $(1-p_j)^2, 2p_j(1-p_j), p_j^2$. Then $\mathbb{E}(\mathbf{M}_{ij}) = 2(p_j - 0.5)$ and $\mathrm{Var}(\mathbf{M}_{ij}) = 2p_j(1-p_j)$ for all $i = 1, \ldots, n$. With $Y_j = (\mathbf{M}_{ij} - 2(p_j - 0.5))^2$ we have $\mathbb{E}(Y_j) = \mathrm{Var}(\mathbf{M}_{ij}) = 2p_j(1-p_j)$ and

$$\frac{1}{D}\|\mathbf{m}_{i\bullet} - \tilde{\mathbf{P}}\|^2 = \left(\sum_{j=1}^{s} Y_j\right)\left(\sum_{j=1}^{s} \mathbb{E}(Y_j)\right)^{-1}. \tag{3.10}$$

Now consider the limiting case $s \to \infty$ and assume the series $p_1, p_2, \ldots$ and $(1-p_1), (1-p_2), \ldots$ to be uniformly bounded away from zero, which implies

$$c \leq \frac{\sum_{j=1}^{s} \mathbb{E}(Y_j)}{s} \leq 0.5 \tag{3.11}$$

for some $c > 0$ and for all s. Assume further that $Y_1, Y_2, \ldots$ are uncorrelated. Because of $\mathrm{Var}(Y_i) < \infty$ we can apply Rajchman's version of the strong law of large numbers (Rajchman (1932), cited by Krengel (2005), p. 154) which yields

$$\frac{\sum_{j=1}^{s}(Y_j - \mathbb{E}(Y_j))}{s} \quad \longrightarrow \quad 0$$

with probability 1 for $s \to \infty$. Because of eq. (3.11) we also have

$$\frac{\sum_{j=1}^{s} Y_j}{\sum_{j=1}^{s} \mathbb{E}(Y_j)} - 1 = \left(\frac{\sum_{j=1}^{s}(Y_j - \mathbb{E}(Y_j))}{s}\right)\left(\frac{\sum_{j=1}^{s} \mathbb{E}(Y_j)}{s}\right)^{-1} \quad \longrightarrow \quad 0$$

with probability 1 for $s \to \infty$, from which we get that the left hand side of eq. (3.10) converges to 1 with probability 1 for $s \to \infty$. Together with eq. (3.9) it follows

$$\mathrm{Cov}(g_i, g_j) + \frac{\sigma_g^2}{2D}\|\mathbf{m}_{i\bullet} - \mathbf{m}_{j\bullet}\|^2 \quad \longrightarrow \quad \sigma_g^2(0.5 + 0.5) = \sigma_g^2$$

with probability 1 for $s \to \infty$, *i.e.* $\mathrm{Cov}(g_i, g_j) \sim \sigma_g^2\left(1 - \frac{\|\mathbf{m}_{i\bullet} - \mathbf{m}_{j\bullet}\|^2}{2D}\right)$ and $\mathrm{Var}(g_i) \sim \sigma_g^2$ for s large. Hence, $\mathrm{Cov}(g_i, g_j)$ only depends on the Euclidean distance $\|\mathbf{m}_{i\bullet} - \mathbf{m}_{j\bullet}\|$ of the SNP vectors for s large. If we consider g_i as the value of a random field on $\mathbb{R}^s$ at position $m_{i\bullet}$, then the corresponding variogram is

$$\gamma_g(\mathbf{m}_{i\bullet}, \mathbf{m}_{j\bullet}) = \sigma_g^2 - \mathrm{Cov}(g_i, g_j) = \frac{\sigma_g^2}{2D}\|\mathbf{m}_{i\bullet} - \mathbf{m}_{j\bullet}\|^2$$

for s large, *i.e.*

$$\gamma_g(x) = \frac{\sigma_g^2}{2D}x^2$$

for $x \in [0, \infty)$.

3.7 Using linear transformations of covariance matrices leads to linearly transformed predicted GVs

In this section we show that using linear transformations of covariance matrices in the universal kriging system leads to linearly transformed predicted genetic values, as discussed in section 3.5.2. The proof starts with calculating

$$
\begin{aligned}
(3.6) \quad &\Leftrightarrow \quad \begin{bmatrix} \mathbf{j} & d\mathbf{B} \\ 0 & \mathbf{j}^T \end{bmatrix} \cdot \begin{bmatrix} \tilde{\lambda} \\ \tilde{\mathbf{a}} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & c\mathbf{J} \\ 0 & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \tilde{\lambda} \\ \tilde{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \tilde{d}\mathbf{B}_0 + \tilde{c}\mathbf{j} \\ 0 \end{bmatrix} \\
&\Leftrightarrow \quad \begin{bmatrix} \mathbf{j} & d\mathbf{B} \\ 0 & \mathbf{j}^T \end{bmatrix} \cdot \begin{bmatrix} \tilde{\lambda} \\ \tilde{\mathbf{a}} \end{bmatrix} + c \cdot \underbrace{\sum_i \tilde{a}_i}_{=0} \begin{bmatrix} \mathbf{j} \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{d}\mathbf{B}_0 + \tilde{c}\mathbf{j} \\ 0 \end{bmatrix} \\
&\Leftrightarrow \quad \underbrace{\begin{bmatrix} \mathbf{j} & \mathbf{B} \\ 0 & \mathbf{j}^T \end{bmatrix}}_{=\mathbf{V}} \cdot \begin{bmatrix} \frac{\tilde{\lambda}}{d} \\ \tilde{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \frac{\tilde{d}}{d}\mathbf{B}_0 + \frac{\tilde{c}}{d}\mathbf{j} \\ 0 \end{bmatrix}.
\end{aligned}
$$

Here we used the unbiasedness condition $\mathbf{j}^T\tilde{\mathbf{a}} = \sum_i \tilde{a}_i = 0$. Hence we get

$$
\begin{bmatrix} \frac{\tilde{\lambda}}{d} \\ \tilde{\mathbf{a}} \end{bmatrix} = \mathbf{V}^{-1} \cdot \begin{bmatrix} \frac{\tilde{d}}{d}\mathbf{B}_0 + \frac{\tilde{c}}{d}\mathbf{j} \\ 0 \end{bmatrix} \overset{(3.5)}{=} \frac{\tilde{d}}{d} \cdot \begin{bmatrix} \lambda \\ \mathbf{a} \end{bmatrix} + \mathbf{V}^{-1} \cdot \frac{\tilde{c}}{d} \begin{bmatrix} \mathbf{j} \\ 0 \end{bmatrix}.
$$

Furthermore, we have

$$
\begin{bmatrix} \mathbf{j} \\ 0 \end{bmatrix} = \mathbf{V} \cdot \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \quad \Leftrightarrow \quad \frac{\tilde{c}}{d} \cdot \begin{bmatrix} \mathbf{j} \\ 0 \end{bmatrix} = \mathbf{V} \cdot \begin{bmatrix} \frac{\tilde{c}}{d} \\ \mathbf{0} \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{V}^{-1} \cdot \frac{\tilde{c}}{d} \begin{bmatrix} \mathbf{j} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{\tilde{c}}{d} \\ \mathbf{0} \end{bmatrix}.
$$

Thus, we get

$$
\begin{bmatrix} \frac{\tilde{\lambda}}{d} \\ \tilde{\mathbf{a}} \end{bmatrix} = \frac{\tilde{d}}{d} \cdot \begin{bmatrix} \lambda \\ \mathbf{a} \end{bmatrix} + \begin{bmatrix} \frac{\tilde{c}}{d} \\ \mathbf{0} \end{bmatrix} \quad \text{and therefore} \quad \tilde{\mathbf{a}} = \frac{\tilde{d}}{d} \cdot \mathbf{a}
$$

which finishes the proof.

4 Using Whole Genome Sequence Data to Predict Quantitative Trait Phenotypes in *Drosophila melanogaster*

This chapter is based on the article Ober *et al.* (2012*a*).

4.1 Introduction

Most efforts to understand the genetic architecture of quantitative traits have focused on mapping the variants causing phenotypic variation in quantitative trait locus (QTL) mapping populations derived from crosses between lines genetically divergent for the trait, or in association mapping populations, with the goal of understanding the biological underpinnings of trait variation (Mackay *et al.*, 2009). However, the ability to accurately predict quantitative trait phenotypes from information on genotypic variation in the absence of knowledge of causal variants will revolutionize evolutionary biology, medicine and human biology, and breeding of agriculturally important plant and animal species. The premise of personalized medicine is based on prediction of individual genetic risk to disease from genome-wide association studies (Wray *et al.*, 2007; de los Campos *et al.*, 2010*a*), and the ability to select individuals or lines in animal and plant breeding programs based on genotypic information circumvents the costly process of progeny testing and reduces the generation interval in applied breeding programs, leading to greater efficiency (Hayes *et al.*, 2009; Lorenz *et al.*, 2011).

In classical animal and plant breeding, the genetic quality of individuals or lines is predicted from phenotypic values of selection candidates and their relatives. The widely used Best Linear Unbiased Prediction (BLUP, Henderson (1973)) method models the covariance structures between individuals via the numerator relationship matrix, which is constructed from known pedigree information and thus reflects expected relationships between individuals (*i.e.* the proportion of shared alleles of identical ancestral origin) given the pedigree. The advent of high-throughput genotyping platforms for many agronomic species (Ranade *et al.*, 2001) enabled genotyping large numbers of individuals for dense panels of single nucleotide polymorphisms (SNPs) spanning the genome. The expected, pedigree-based numerator relationship matrix can then be replaced by a realized, genome-based relationship matrix (often called the "genomic" relationship matrix, VanRaden (2008)). This approach is equivalent to a random regression approach in which all SNP genotypes are simultaneously accounted for as explanatory variables in a multiple regression model (Goddard, 2009). In animal and plant breeding, selection based on genome-based predictions of genetic values is expected to massively increase genetic progress (Meuwissen *et al.*, 2001; Hayes *et al.*, 2009) and has quickly found its way into widespread practical application (see Hayes *et al.* (2009); Lorenz *et al.* (2011) for reviews).

Genome-based prediction follows a different paradigm than genome-wide association studies (GWAS). GWAS identify single molecular variants associated with phenotypic variability using individual statistical tests for significance of each variant. Genome-based prediction uses the entire genomic variability captured by the available marker set to explain the observed phenotypic variation, and does not rely on selection of single loci based on significance tests. Standard prediction methods are thought to work for traits with a highly polygenic or even infinitesimal (Fisher, 1918) genetic architecture, where the effect of a single variant is too small to be captured by a statistical test in a GWAS. There is strong empirical evidence that many quantitative traits have such a highly polygenic genetic architecture in farm animals (Pimentel *et al.*, 2011), agriculturally used plants (Schön *et al.*, 2004), model organisms and humans (Mackay, 2004; Flint & Mackay, 2009).

With the advent of next generation sequencing technologies, it is now feasible to implement genomic prediction based on complete genome sequences of higher organisms. While these techniques have only been applied to individuals or cohorts of limited size (Eck *et al.*, 2009) to date, initiatives to sequence larger panels are under way (The 1000 Genomes Project Consortium, 2010; Elshire *et al.*, 2011), and genotyping by whole genome resequencing will become a standard technology in the foreseeable future.

The accuracy of prediction methods based on marker data depends on the heritability of the trait, its genetic architecture (number of loci affecting trait variation, mode of inheritance, and distribution of allelic effects, Hayes *et al.* (2010)), the LD reflecting effective population size, the size of the genome, the marker density and the sample size used in the statistical analysis (Daetwyler *et al.*, 2010). Various methods of prediction incorporating genomic information have been studied on real and simulated data, including Genomic Best Linear Unbiased Prediction (GBLUP) approaches with genomic relationship matrices (VanRaden, 2008), Random Regression BLUP (RRBLUP), Bayesian linear regression methods (Meuwissen *et al.*, 2001; Gianola *et al.*, 2009) or fully non-parametric approaches (Gianola & van Kaam, 2008; de los Campos *et al.*, 2009; Long *et al.*, 2010; Ober *et al.*, 2011).

As elucidated in chapter 1 and already applied in chapter 3, GBLUP approaches are based on a linear model for the phenotypic values, which encompasses a vector of random genetic values of individuals whose covariance structure is inferred from genomic data. The linear model underlying the RRBLUP approach includes a vector of random marker effects (instead of a vector of genetic values) which are assumed to be drawn from the same normal distribution and uncorrelated. This model primarily provides estimates of SNP effects, but estimated genetic values of individuals can be derived as linear combinations of the estimated SNP effects, yielding the same predictions of individual genotypic or phenotypic values as GBLUP. The BayesB method (Meuwissen *et al.*, 2001), on the other hand, fits only a small fraction of the available markers to conform with the assumption that most loci are expected to have zero effect on the phenotype, and the remaining non-zero marker effects are drawn from normal distributions with random variances.

It has been proposed (Meuwissen & Goddard, 2010) that differences between prediction methods will become more pronounced with the availability of full genome sequence data. According to a study with simulated data (Meuwissen & Goddard, 2010), RRBLUP and equivalent GBLUP procedures do not take full advantage of high-density marker data if the number of causal SNPs is small, while approaches with an implicit feature selection such as BayesB might be more accurate. If, on the other hand, the number of causal loci is large, RRBLUP or GBLUP methods may yield accurate predictions because the assumption that every SNP has an effect is closer to reality.

Implementing genomic prediction with full genome sequence data raises a number of questions. What is the most efficient way to incorporate the complete genomic information in prediction? How much predictive ability is gained by using whole genome sequence data compared to high density SNP panels? Is it possible to increase predictive ability by a pre-selection of SNPs or models with an internal feature selection? How comparable are the results of genomic prediction and genome-wide association? In this chapter we address these questions empirically based on full genomic sequences of a population of *Drosophila melanogaster* inbred lines. The inbred lines have been sequenced, and constitute the "*Drosophila melanogaster* Genetic Reference Panel" (DGRP, Mackay *et al.* (2012)), a new community resource for genetic studies of complex traits.

We report the results of a full sequence based genomic prediction for two quantitative traits, starvation stress resistance and locomotor startle response, both of which display considerable genetic variation in natural populations and respond rapidly to artificial selection (Ayroles *et al.*, 2009; Harbison *et al.*, 2004; Jordan *et al.*, 2007). We used whole-genome sequences determined on the Illumina platform for 157 (155) DGRP lines for starvation resistance (startle response) (Mackay *et al.*, 2012). Our reference method is a GBLUP approach in which $\approx$ 2.5 million polymorphic SNPs are used to derive a genomic relationship matrix (VanRaden, 2008). We evaluated predictive ability via cross-validation (CV), and compared prediction within *vs.* across sexes, various SNP densities, and training set sizes. We assessed whether BayesB is superior over GBLUP given full genome sequence data (Meuwissen & Goddard, 2010), and compared our genomic prediction results with those of GWAS conducted on the same DGRP lines (Mackay *et al.*, 2012).

To our knowledge, this is the first application of genomic prediction on empirical whole genome sequence in a substantial sample of a higher organism. However, this study, as well as all previous association studies, only assesses the effects of common SNPs, since the effects of rare alleles cannot be estimated due to the small sample of sequenced lines. The results illustrate both the potential of the approach and challenges to be addressed in the future.

4.2 Results

4.2.1 Genomic Best Linear Unbiased Prediction (GBLUP)

We constructed a genomic relationship matrix (VanRaden, 2008) from $\approx$ 2.5 million SNPs for which the minor allele was present in at least four of the DGRP lines (Mackay *et al.*, 2012). A histogram of the off-diagonal elements of this matrix for 157 DGRP lines used in the GBLUP analyses (Figure 4.1) and a corresponding heatmap (Figure 4.2) show that there were no large blocks of high genomic relationship among the lines.

The average genomic relationship is close to zero, as expected, but there is considerable variance around this average (Figure 4.1), as indicated by two block of lines with average

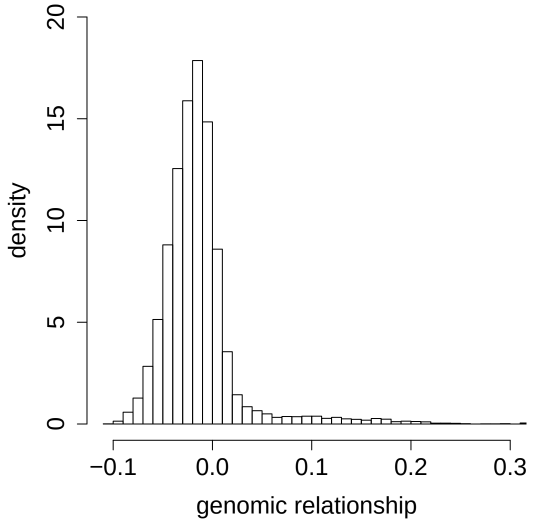

Figure 4.1: Histogram of the offdiagonal elements of the genomic relationship matrix G. The genomic relationship matrix **G** was calculated according to VanRaden (2008) using 157 lines and 2.5 million SNPs.

genomic relationships within each block of 0.25 and 0.34 (Figure 4.2). We performed genomic prediction for starvation stress resistance and locomotor startle response. The phenotypes used were the medians of many (40–52) individually tested males and females for each line, or the average of the male and female medians (Table 4.1).

Table 4.1: Mean and standard deviation of phenotypic values and of the number of individual records per line. Phenotypic values were calculated as the averages of the medians of male and female records ("all") or as the medians of female or male records separately.

	starvation resistance		startle response	
	phen. value[1]	# rec. per line[2]	phen. value	# rec. per line
all	52.5 ± 10.7	104.1 ± 21.5	29.4 ± 6.6	80.1 ± 7.4
female	44.9 ± 10.0	52.2 ± 11.2	29.2 ± 6.7	40.2 ± 3.9
male	60.2 ± 13.4	51.8 ± 10.8	29.5 ± 6.7	39.8 ± 4.3

[1] Phenotypic values.
[2] Number of records per line.

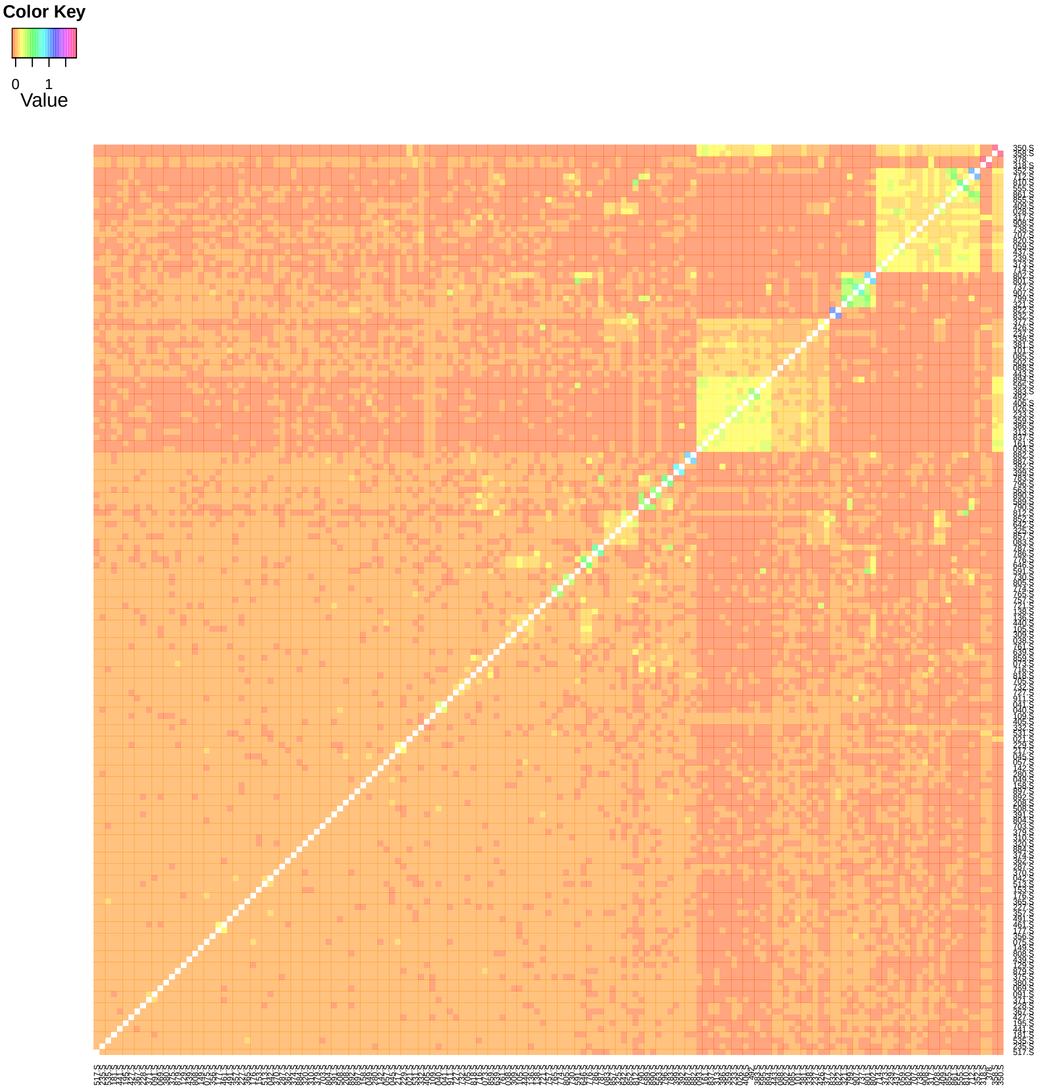

Figure 4.2: Heatmap of the genomic relationship matrix G. The genomic relationship matrix **G** was calculated according to VanRaden (2008) using 157 lines and 2.5 million SNPs. The "S" after the line-ID indicates that the line belongs to the set of lines for which phenotypic records for startle response were also available (in addition to the phenotypic records of starvation resistance).

We used several cross-validation (CV) procedures for each trait (Table 4.2). In the 5-fold CV, predictive ability was 0.239 ± 0.008 for starvation resistance and 0.230 ± 0.012 for startle response. In human studies the efficiency of a predictor is reported as the squared correlation r^2 rather than r (Makowsky *et al.*, 2011), so that in terms of variance explained the estimates

Table 4.2: Average correlations between predicted genetic values and observed phenotypes for different CV procedures with GBLUP and different traits.

type of CV	starvation resistance	startle response
(4:1)-CV[1] all[2]	0.239[3] (0.008)	0.230 (0.012)
(3:2)-CV all	0.213 (0.006)	0.216 (0.011)
(2:3)-CV all	0.176 (0.006)	0.181 (0.010)
(1:4)-CV all	0.124 (0.006)	0.128 (0.006)
(4:1)-CV male – female[4]	0.164 (0.007)	0.217 (0.011)
(4:1)-CV female – male	0.182 (0.007)	0.235 (0.012)
(4:1)-CV male – male	0.203 (0.008)	0.230 (0.012)
(4:1)-CV female – female	0.254 (0.009)	0.216 (0.011)

[1] “$(t : v)$-CV” means: t parts are used as training set and v parts are used as validation set.

[2] The average of the medians of male and female measurements was used to predict line phenotypes. Predicted phenotypes were then correlated with the averages of the medians of male and female measurements.

[3] Average correlation between predicted genetic values and observed phenotypes. Results are averages over 20 replicates. Standard errors of the means in parentheses.

[4] “CV sex_1 – sex_2” means: Medians of measurements of sex_1 were used in the training set, medians of sex_2 were used in the validation set.

were 0.074 ± 0.005 for starvation resistance and 0.080 ± 0.005 for startle response. The observed accuracy depends on the size of the training set (Figure 4.3), with decreasing accuracies obtained with smaller training sets. Predictive abilities are roughly halved for both traits when using only 20% instead of 80% of the data to train the model. Maximum likelihood estimates of narrow-sense heritabilities based on the GBLUP model using the genomic relationship matrix were 1.0 in all analyses (Table 4.3), reflecting the fact that phenotypes are averages over many replicates and thus residual variance is minimal. Hence, the phenotypes used represent the line genotypes with maximum accuracy, which is the ideal case for training the genomic model.

Using male performance data to train the model and using the results to predict the female performance (or vice versa) does not affect the predictive ability for startle response, but substantially reduces the predictive ability for starvation resistance, reflecting a higher degree of genotype by sex interaction in this trait (Mackay *et al.* (2012), and see below). Prediction is more accurate in females than in males (0.254 *vs.* 0.203) for starvation resistance, while there is little difference for startle response.

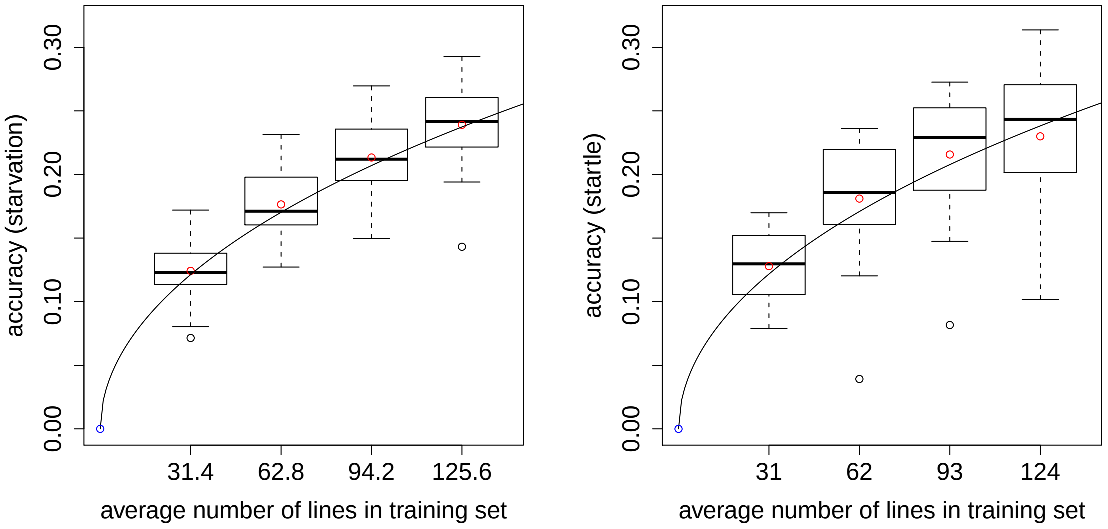

Figure 4.3: Accuracy of prediction of GBLUP for CVs with different numbers of lines in the training set. Each boxplot illustrates the average accuracies for 20 replicates of the CV procedure using GBLUP. The left (right) plot shows accuracies for starvation resistance (startle response). The solid line is the curve of Daetwyler *et al.* (2010) fitted to the empirical data, which results in estimates of $N_e = 8{,}747$ and $N_e = 8{,}676$ for starvation resistance and startle response. All 2.5 million SNPs were used to construct the genomic relationship matrix in the GBLUP model.

Table 4.3: Variance components and heritabilities estimated from GBLUP using all 157 (155) lines. Variance components were estimated by maximum likelihood using the R-package "RandomFields" and its function "fitvario" and the averages of the medians of male and female records ("all") or the medians of female or male records separately as phenotypic data.

	starvation resistance			startle response		
	$\hat{\sigma}_g^2$	$\hat{\sigma}_e^2$	$\hat{h}^2_{\text{GBLUP}}$	$\hat{\sigma}_g^2$	$\hat{\sigma}_e^2$	$\hat{h}^2_{\text{GBLUP}}$
all	62.6	0	1	21.7	0	1
female	91.2	0	1	22.4	0	1
male	57.9	0	1	22.5	0	1

A series of 5-fold CVs for starvation resistance using different SNP densities showed that predictive ability remained almost constant if every 16th SNP (≈ 150,000 SNPs) was used to construct the genomic relationship matrix (Figure 4.4). The predictive ability began to deteriorate when fewer than 150,000 SNPs were used, but only vanished completely when as few as ≈ 2,500 SNPs (every 1,024th SNP) were used.

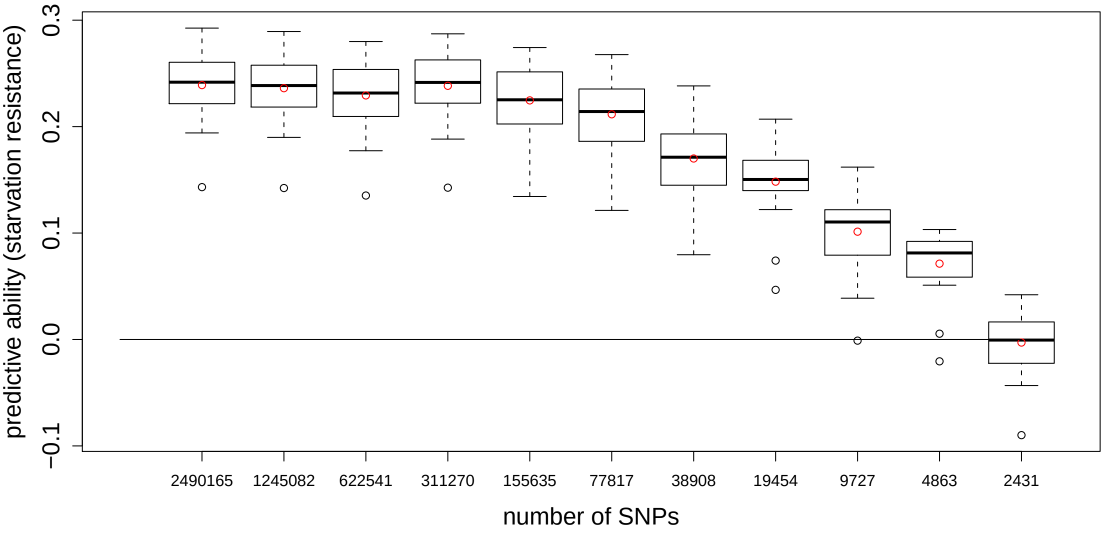

Figure 4.4: Predictive ability of 5-fold CV with GBLUP for starvation resistance using different numbers of SNPs. Each boxplot shows the average predictive abilities for 20 replicates of 5-fold CV using GBLUP. For the CVs leading to the $(k+1)$-th boxplot, every 2^k-th SNP was used to build the genomic relationship matrix **G** according to VanRaden (2008). This was done for the thinning factors $k = 0, \ldots, 10$. The red dots indicate the average predictive abilities.

The corresponding LD distribution for SNP neighbors for different SNP densities is shown in Figure 4.5, illustrating the extreme short-range extent of LD in the *D. melanogaster* genome. The average LD between SNPs (after imputation) whose distance lay in the interval [10,50] ([100,200] , [900,1000]) bp was $r^2 = 0.24\,(0.14, 0.07)$ for the autosomes and $r^2 = 0.38\,(0.23, 0.10)$ for the X-chromosome. Long-range LD between pairs of loci at the opposite ends of chromosome arms or across different chromosome arms was on average 0.007 both for the autosomes and the X-chromosome.

For starvation resistance, the influence of the minor allele frequency of the SNPs used on the predictive ability was assessed with a series of 5-fold CVs using SNP sets with different average minor allele frequency. We find that the variability of the predictive ability increases when the average minor allele frequency of the SNPs used to construct the genomic relationship matrix is decreased (Figure 4.6). In 20 replicates of an additional 5-fold CV, in which we *randomly* chose 77,817 SNPs to build the genomic relationship matrix, an average predictive ability of 0.221 ± 0.009 was obtained, which is in the range obtained when every 32^{nd} SNP (≈ 77,817 SNPs) was used (0.211 ± 0.008, Figure 4.4).

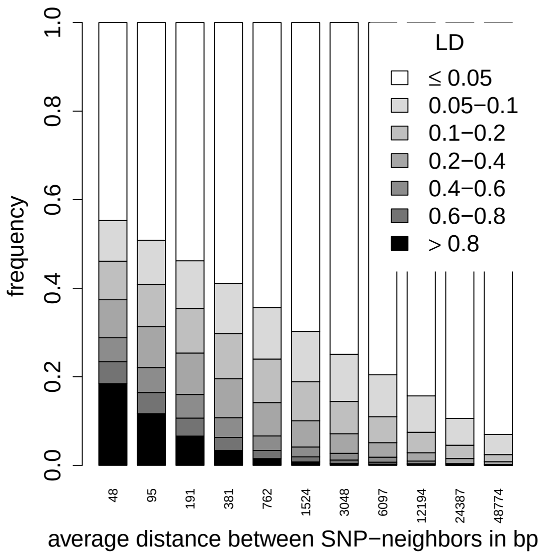

Figure 4.5: The distribution of r^2 between SNP neighbors for different SNP densities. For the $(k+1)$-th stacked bar, every 2^k-th SNP was used, $k = 0, \ldots, 10$. Then, the distribution of r^2 for the resulting SNP neighbors was calculated.

Running 20 replicates of a 5-fold CV using 10 randomly chosen blocks of adjacent SNPs (each block consisting of 7,781 SNPs) led to an average predictive ability of 0.210 ± 0.011.

To analyze whether the predictive ability is due to lines which are more highly related, we ran an additional 5-fold CV with 20 replicates in which the two groups of higher overall relatedness (Figure 4.2) were excluded. Here we found an average predictive ability of 0.290 ± 0.008 for starvation resistance, which is larger than the average predictive ability we obtained using all lines (0.239 ± 0.008). For startle response, excluding the two groups led to a decrease in predictive ability (0.168 ± 0.017 in comparison to 0.230 ± 0.012).

4.2.2 Effective population size derived from empirical accuracies of genomic prediction

The accuracy of genomic prediction is a function of a number of quantities, including the size of the training set and the effective population size N_e (Daetwyler *et al.*, 2010). N_e has an effect on the number of independently segregating chromosome segments, M_e, in a population (the larger N_e, the larger M_e); and the predictive ability of GBLUP is higher when the number of segments is small. By varying the size of the training set in a series of CVs, we can estimate N_e by fitting a curve through the empirical accuracies obtained (Figure 4.3).

We estimated $\hat{N}_e = 8{,}748$ for starvation resistance and $\hat{N}_e = 8{,}676$ for startle response. The coefficient of determination of the fitted curve was $R^2 = 0.70\,(0.44)$ for starvation resistance (startle response). The bias corrected empirical 95% confidence intervals for the N_e estimates obtained with bootstrapping (Efron & Tibshirani, 1986) were [8,173; 9,474] for

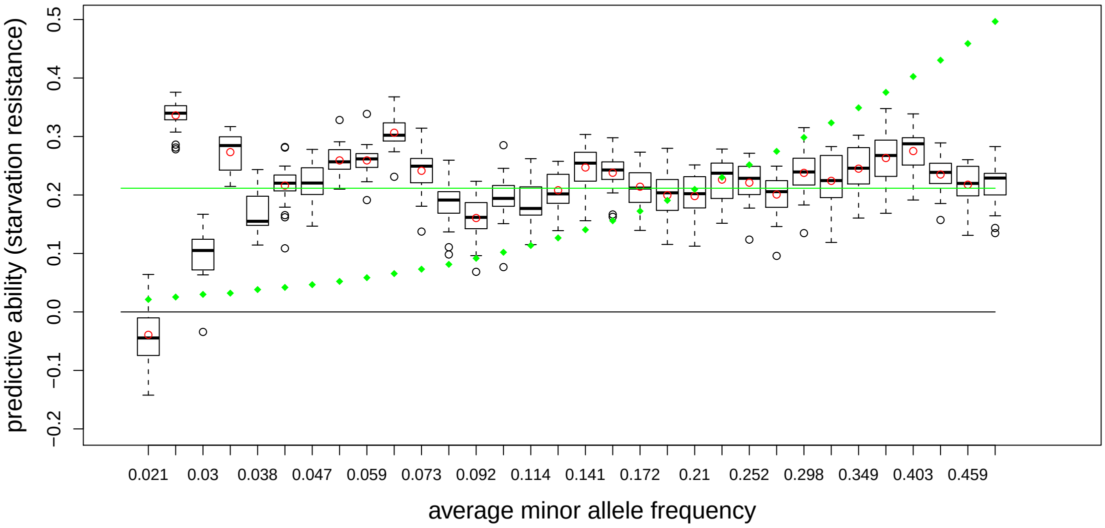

Figure 4.6: Predictive ability of 5-fold CV with GBLUP for starvation resistance using different set of SNPs with different average minor allele frequencies. Each boxplot shows the average predictive abilities for 20 replicates of 5-fold CV using GBLUP and SNPs with different average minor allele frequencies. The different average minor allele frequencies are plotted as green dots. To choose the SNPs for each bin of minor allele frequency the SNPs were sorted by minor allele frequency and then divided into 32 blocks, *i.e.* each bin contained ≈ 77,817 SNPs. The horizontal green line indicates the average predictive ability obtained using every 32^{nd} SNP (resulting in 77,817 SNPs as well), which was 0.212 ± 0.008.

starvation resistance and [7,716; 9,925] for startle response.

The effective population size in the Raleigh population (from which the DGRP lines were drawn) was estimated to be ≈ 19,000 in 1984, with a massive fluctuation between years (Kusakabe *et al.*, 2000). Our estimates of $N_e \approx 8{,}700$ correspond to $M_e = \frac{N_e L_f}{\ln(2N_e L_f)} \approx 2{,}000$ independently segregating chromosome segments. In this formula L_f is the length of the female genome in Morgans (there is no recombination in male Drosophila). Since the sequenced animals resulted from 20 generations of full-sib mating following the original sampling from the Raleigh population, the DGRP lines are not expected to have the same M_e as the original population and are consequently expected to have a different N_e.

We can use the curves fitted through the empirical accuracies (Figure 4.3), to predict the expected accuracy of prediction for an arbitrarily large size of the training set: If 1,000 lines were available in the training set, the curve would predict accuracies of ≈ 0.58 for starvation resistance and startle response. This value was obtained by using $\hat{N}_e$ and $\hat{h}^2_{\text{GBLUP}} = 1$ as well as $N_p = 1{,}000$ and $L_f = 2.451$ in the modified formula of Daetwyler *et al.* (2010).

4.2.3 Effective population size derived directly from linkage disequilibrium

We also estimated the effective population size based on LD directly. For a distance bin of 0.02 Morgan we obtained average LD-values of $0.010\,(0.009, 0.008, 0.011, 0.008)$ for chromosome *2L* (*2R*, *3L*, *3R*, *X*). These values correspond to an estimated effective population size of $\hat{N}_e = 3{,}415\,(5{,}541,\ 10{,}663,\ 2{,}811,\ 9{,}710)$, approximately 25 generations ago. The average estimated effective population size is $\hat{N}_e = 6{,}428$, which is in the range of the estimates based on the observed accuracies.

4.2.4 Genomic prediction with SNP selection

Genomic prediction might be improved if we only fit SNPs which are associated with variance in a trait, because we then concentrate on the biologically relevant genomic regions, and excluding SNPs which are not associated with the trait reduces statistical noise. We tested this hypothesis using the starvation resistance data. We identified the 5% SNPs with the highest absolute estimated effect or the highest estimated genetic variance, respectively, in the training set of the respective 80% of the folds in a 5-fold CV. We then used these subsets of selected SNPs to predict the phenotype in the remaining 20% of the fold. Predictive ability was improved by 3.3% over the reference scenario when using the 5% SNPs with largest effects (average predictive ability of 0.247 ± 0.008 in comparison to 0.239 ± 0.008). Using the 5% SNPs with greatest variance explained, predictive ability was improved by 2.1% (average predictive ability of 0.244 ± 0.008). In both cases, the improvement is marginal and provides little support for the idea of SNP pre-selection.

We also compared our GBLUP results to those from a method which does not assume that all SNP effects are drawn from the same normal distribution and carries out an internal feature selection. We ran 20 replicates of a 5-fold CV for starvation resistance using BayesB (Meuwissen *et al.*, 2001). In each round of the Markov Chain Monte Carlo based procedure (see section 4.4.11), 99.5% of the SNPs were assumed to have no effect and the effects of the remaining 0.5% of the SNPs were drawn from normal distribution with random variances. In most folds of each single CV and for all replicates of CV, the observed predictive abilities differed only marginally between BayesB and GBLUP (Figure 4.7). The average predictive ability obtained with BayesB was 0.238 ± 0.008 which is not appreciably different from the result obtained with GBLUP (0.239 ± 0.008).

4.2.5 Genomic prediction *vs.* GWAS

Although genomic prediction follows a different paradigm than genome-wide association studies, it is informative to compare significant SNP positions from the GWAS to areas of large estimated SNP effects resulting from the GBLUP model. Previously (Mackay *et al.*, 2012), a GWAS of 168 DGRP lines (of which the material used here is a subset) identified 115 SNPs associated with starvation resistance and 75 SNPs associated with startle response at a nominal p-value $\leq 10^{-5}$ in the analyses of sex-averaged data. We estimated SNP effects using RRBLUP and compared them to the significant SNPs from the GWAS study (Suppl. Figures S1 and S2). There is excellent concordance of signals from both approaches in some regions (*e.g.* the genome-wide largest SNP effects on chromosome *3L* for starvation resistance

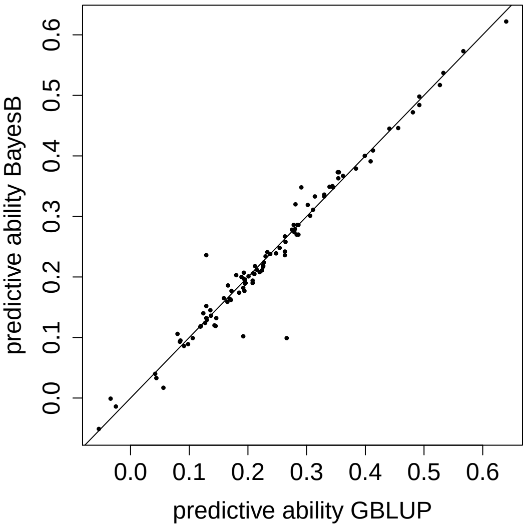

Figure 4.7: Predictive ability for GBLUP *vs.* BayesB using phenotypic values of starvation resistance. Predictive abilities are plotted for 20 replicates of a 5-fold CV, each replicate consisting of 5 corresponding folds of CV.

and *2L* for startle response), while concordance is poor in other regions, especially on the *X* chromosome.

We further investigated whether the most significant SNPs detected in the GWAS are reflected by large SNP effects in the GBLUP study using a different approach. For each significant SNP position from the GWAS we took the 100 neighboring SNPs (50 on each side) and calculated the sum of the absolute values of their estimated effects using the GBLUP model. To avoid an effect of different sample size, we used the 75 most significant loci from the GWAS for both traits. We compared these sums to the sums of the absolute values of estimated SNP effects in $\approx$ 250,000 sliding windows spanning the whole genome (with each window containing 100 neighboring SNPs). We observed a clear separation of the density functions of these sums for both startle response and starvation resistance (Figure 4.8).

The density resulting from the sliding window approach reflects the overall distribution of the suggested statistic in the sample. For starvation resistance (startle response) a threshold value of 0.0076 (0.0046), *cf.* Figure 4.8, cuts off the upper 10% of the respective distribution. Applying the same threshold with the density function reflecting the statistic for the *significant* GWAS positions, 66.7% (74.7%) of the distribution exceeds the threshold, indicating that signals found in the GWAS are also associated with large estimates of the SNP effects in the genomic model.

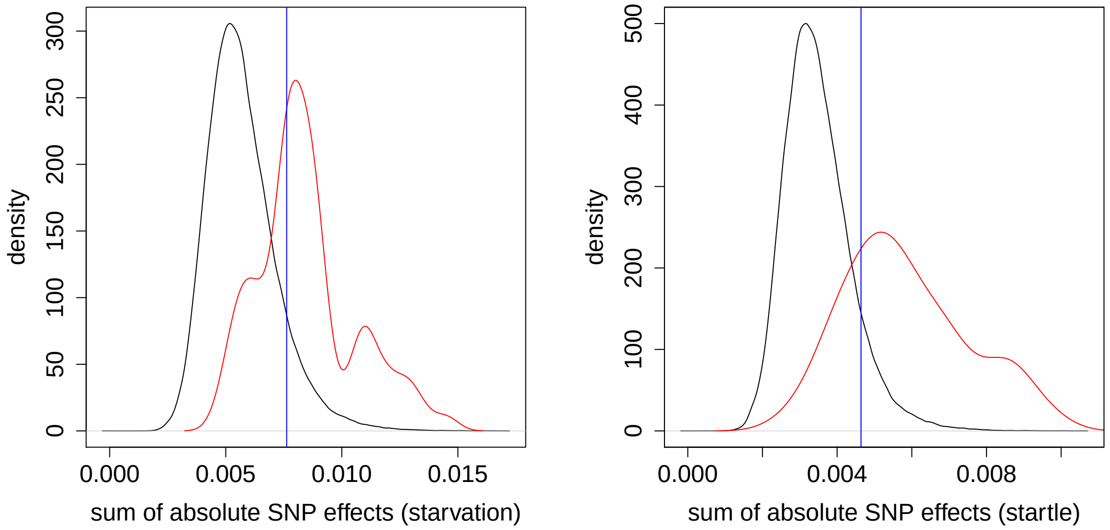

Figure 4.8: Distribution of absolute SNP effects. The density of the sum of the absolute SNP effects from GBLUP is plotted for sliding windows of 100 adjacent SNPs covering the whole genome (black) and for windows around the 75 most significant SNPs (red) according to the GWAS of Mackay *et al.* (2012). The left (right) plot shows the densities for starvation resistance (startle response). The blue vertical line indicates the 90% quantile of the black density function.

4.2.6 Analyses of individual trait data

In addition to the line medians we also analyzed individual records (104 ± 21 individual flies per line tested for starvation resistance and 80 ± 7 for startle response) to assess whether the variance between lines can be fully explained by additive gene effects or if non-additive mechanisms have an impact. This was done by modeling the covariance structure between lines based on the additive and additive × additive genomic relationship matrix and testing the goodness of fit of the respective models. Most applications of genomic prediction are for outbred populations, for which the additive genetic variance and corresponding narrow-sense heritability determine the extent to which phenotypes in the next generation can be predicted from information obtained on the current generation. However, the variance *among* DGRP lines is the total genetic variance, and is possibly inflated by additive by additive epistatic variance (Falconer & Mackay, 1996). Therefore, we performed several analyses on measurements of *individual* flies to determine the nature of the total genetic variance, especially to what extent the presence of non-additive genetic variance might have affected predictive abilities. We fitted three different models to the individual phenotype data: Model 1 contained a random line effect, and lines were assumed to be unrelated. In Model 2, a random additive line effect g was added, whose covariance structure was modeled via the genomic relationship matrix $\mathbf{G}$. In Model 3, an additional random additive × additive epistatic effect $g \times g$ was included, whose covariance structure was modeled via the Hadamard product $\mathbf{G} \circ \mathbf{G}$. Since the between line variance relates to inbred lines, while the additive and additive × additive variance component pertain to the non-inbred base population (or a hypothetical random mating $F2$ produced from the inbred lines), the variance between inbred lines in Model 1 is expected to be twice the additive genetic variance in Model 2 or 3 under a fully additive model.

We estimated variance components for all three models pooled across sexes and separately for males and females (Suppl. Tables S1 and S2). We find little evidence of non-additive genetic variance for these traits. The estimate of σ_g^2 from Model 2 is $\approx \frac{1}{2}\sigma_{\text{line}}^2$ from Model 1, and Model 2 gave a significantly better fit than Model 1 when applying the likelihood ratio test, again indicating that the observed between line variance is due to additive gene action. Inclusion of the $g \times g$ component was not significant for either of the traits. We found significant sex by line interaction variance for starvation resistance, but not for startle response (Suppl. Tables S1 and S2), which is in accordance with the findings of the genomic prediction across sexes (Table 4.2) and previous analyses of these data (Mackay *et al.*, 2012).

4.3 Discussion

We report the first (to our knowledge) application of genomic prediction to a real set of full genomic sequencing data in a eukaryotic organism. Although predictive abilities obtained with starvation resistance and startle behavior are only moderate to low, and although we limited our analysis to SNPs that are common due to the small sample size of lines, this study can be seen as a proof of concept for this approach. There are several reasons for the limited predictive ability obtained in this study. First, the training set is small, with a maximum of ≈ 120 observations in the 5-fold CV, and the accuracy of genomic prediction is a function of the size of the training set (Daetwyler *et al.*, 2010). Using the curves fitted through the empirical accuracies (Figure 4.3), we predict accuracies of ≈ 0.58 for starvation resistance and startle response, if 1,000 sequenced lines were available for the training set.

The second important factor affecting accuracy of prediction is the number of independently segregating chromosome segments, M_e (Daetwyler *et al.*, 2010). In our study we obtained $M_e \approx 2{,}000$. This is larger than usually observed for Holstein cattle ($M_e \approx 640$ with $N_e \approx 100$ and genome length $L \approx 30$ Morgans (Qanbari *et al.*, 2010)), but is smaller than the corresponding value in the human genome ($M_e \approx 14{,}000$ with $N_e \approx 3{,}000$; $L \approx 30$ Morgans, (Tenesa *et al.*, 2007)). (Note that in mammalian species, there is recombination in both sexes and $M_e = \frac{2N_eL}{\ln(4N_eL)}$ (Goddard, 2009).)

Accuracy of genomic prediction is thought to come from two sources: (i) SNPs in useful LD with causal loci; and (ii) SNPs reflecting the relationship structure between the training set and the set to be predicted (Habier *et al.*, 2007). Due to the very fast decay of LD in the *D. melanogaster* genome, few SNPs are in useful LD with any causal polymorphism. Even if we define “useful LD” very conservatively as $r^2 > 0.2$, then on average only a region of 120 bp around a causal polymorphism was in useful LD on an autosome (400 bp on the X chromosome). This means that on average 3 (6) SNPs were in useful LD with a causal autosomal (X-linked) polymorphism, as the average distance between neighboring SNPs was 45 bp (66 bp) on an autosome (X chromosome). If predictive ability was mainly driven by SNPs in LD with causal polymorphisms, reducing the SNP density should lead to a massive decay of predictive ability of the models, which was not observed. Little decrease in predictive ability was seen, even if every 32^{nd} SNP was used in the model, in which case hardly any SNP would be in useful LD with causal polymorphisms. The underlying mechanism therefore seems to depend on a sufficient number of SNPs being in low LD with causal polymorphisms, rather than few SNPs in close physical association and high LD. In the DGRP population, LD approaches a small but positive baseline level with increasing physical distance (Mackay *et al.*, 2012), so that even with large physical distances a minimum level of LD is maintained, which was on average $0.007 \approx \frac{1}{n}$ with $n = 157$ being the sample size.

The number of SNPs for maximal accuracy of genomic prediction with unrelated individuals has been estimated as $10N_eL$ (Meuwissen, 2009), corresponding to ≈ 110,000 SNPs in the present study.

For starvation resistance, we find that the empirical accuracy levels off when approximately every 16th SNP is used, which is equivalent to ≈ 155,000 or $7.3N_eL_f = 14.6N_eL$ SNPs. Adding more SNPs beyond this value does not lead to any improvement in the genomic prediction of starvation resistance, but also does not reduce accuracy, which one might expect when using more SNPs than actually needed. While fitting large numbers of "superfluous" SNPs may be considered as noise in the RRBLUP model, these SNPs can also be seen to provide a better basis to estimate the realized relationship matrix in the GBLUP model, which leads to a higher accuracy of the estimated realized relationships. Since both models are fully equivalent (Goddard, 2009) no penalty is expected in the prediction of genomic values.

Since pedigree information for the founders of the inbred lines was not available, our estimates of heritability and genomic prediction are based on the actual degree of identity-by-descent sharing between relatives (Visscher *et al.*, 2006). There is little pedigree structure in the DGRP lines, with the exception of two distinct blocks of higher relatedness, comprising 18 and 13 lines, respectively, with a genomic relationship within blocks of ≈ 0.25 and 0.34. When these blocks were excluded from the data, predictive ability in a 5-fold CV increased (decreased) for starvation resistance (startle response), suggesting that prediction in the DGRP population does not rely on distinct family structures. Given this together with the short-range extent of LD in the *D. melanogaster* genome and the robustness of the accuracy of genomic prediction with reduced marker density, we conclude that the observed accuracy of prediction for starvation resistance and startle response is primarily due to the long-range LD in the population, or equivalently, the subtle relationship structure as reflected by the genomic relationship matrix.

We restricted our analyses to SNPs for which the minor allele was present in at least four DGRP lines (a minor allele frequency of 0.025). We applied this threshold to avoid computational limitations, especially when applying the BayesB method; and for consistency with the GWAS in the DGRP (Mackay *et al.*, 2012), which used the same filtering criterion. Thus, we did not utilize the ≈ 2 million SNPs with minor allele frequencies less than this, nor did we take other forms of molecular variation into account.

Structural variations such as transposable elements have been repeatedly reported to be associated with phenotypic variation (González & Petrov, 2009), therefore we must consider to what extent not including these variants in the models affected prediction accuracy. Given that we do not observe an increase in predictive ability when increasing the number of SNPs from ≈ 150,000 to 2.5 million, we do not expect that increasing the marker density by adding more SNPs and other variants will have a significant effect on predictive ability. Additionally, SNPs with low minor allele frequencies were shown to be highly variable in predictive ability, so that the potential amount of information possibly added by the 2 million low frequency SNPs is limited. However, accounting for all polymorphisms in the model means that some fraction of the genetic variants must causally affect the trait. Simulations (Meuwissen & Goddard, 2010) including the causal polymorphism in the model improves the predictive ability over models based only on neutral SNPs in LD with the causal variants. Further research is needed to understand these mechanisms in the context of genomic prediction based on empirical data.

The accuracy of BayesB has outperformed that of GBLUP in several simulation studies (Meuwissen *et al.*, 2001; Habier *et al.*, 2007). Simulation results have suggested that GBLUP did not take full advantage of genome sequence data, suggesting that Bayesian methods are needed to obtain maximum accuracy (Meuwissen & Goddard, 2010). The superiority of BayesB over GBLUP is expected to increase with marker density, and decrease when the size of the training data set is increased (Meuwissen, 2009). However, we did not find that BayesB yielded a significantly higher predictive ability than GBLUP in the 20 replicates of 5-fold CV with starvation resistance implemented in the present study. We used a very high marker density and a small training set, and yet GBLUP performed as well as BayesB. These conclusions should be taken with caution, since the available size of the training set was extremely small in our study due to the limited availability of fully sequenced lines. In Daetwyler *et al.* (2010), BayesB yielded a higher accuracy than GBLUP, when the number of simulated QTL was low; but GBLUP slightly outperformed BayesB, when the number of QTL became large, since the GBLUP model is equivalent to RRBLUP, in which all SNPs are assumed to have an effect drawn from the same normal distribution. Although this model may not seem biologically plausible, it performed as well as BayesB in the present study, consistent with several studies on real data from dairy cattle for different traits (Hayes *et al.*, 2009; VanRaden *et al.*, 2009).

The finding that BayesB did not outperform GBLUP in the present study is consistent with a quasi-infinitesimal genetic architecture; and results indicate that starvation resistance and startle response are complex traits with a highly polygenic genetic architecture rather than being driven by a few major causal genes. This is in agreement with previous studies stating that starvation resistance and startle response can be considered to be model traits with a complex (*i.e.* quasi-infinitesimal) genetic background (Ayroles *et al.*, 2009; Harbison *et al.*, 2004; Jordan *et al.*, 2007); and it is also in line with the results from the GWAS (Mackay *et al.*, 2012). One reasonable conclusion might be that there are so many causal polymorphisms, each with a small effect, that the ≈ 2,000 effective chromosome segments are saturated with causal variants and the effects of segments follow a normal distribution. Under this circumstance, GBLUP is expected to perform as well as BayesB. However, these hypotheses clearly need further investigation. More systematic model comparisons based on the available data were not considered here due to the prohibitive computing time required for BayesB.

Previously, gene centered multiple regression and partial least square (PLS) regression models were used to predict starvation resistance and startle response phenotypes from genotypic data (Mackay *et al.*, 2012). In both cases only SNPs that had nominal significance levels of $P < 10^{-5}$ from the GWAS were used. The gene centered prediction models found that a few SNPs explained a large fraction of the genetic and phenotypic variance of the traits, while the PLS models found that the significant SNPs explained a high fraction of the phenotypic variance. The purpose of these studies was a comparison with human association studies, in which the faction of the variance explained by significant variants in the entire sample is commonly quoted. These approaches are fundamentally different from the BLUP approach used in this study. The BLUP approach includes random components and their covariance structure in the model, whereas regression models do not incorporate random terms except from the residuals; and the BLUP approach does not rely on a pre-selection of SNPs based on a GWAS. Most critically, we evaluated the robustness of the BLUP predictions using 5-fold cross-validation; whereas the previous analyses only tested the explanatory power of the most significant associated SNPs using the entire sample. Had we done the same analysis using GBLUP, we would be able to predict 100% of the variance.

The imperfect concordance of the positions of the most significant SNPs from the GWAS and the largest estimates of SNP effects from RRBLUP is a consequence of the different objectives of the two approaches. A sequence-based GWAS is conducted to identify causal polymorphisms and provide estimates of allelic effects and frequencies. Also, the GWAS suffers from estimating one effect at a time and so does not necessarily position the QTL accurately. The goal of RRBLUP is to predict the phenotype using all available SNP information simultaneously. Here, estimated SNP effects are a by-product and mapping causal variants is not the primary objective. Given that the number of SNP effects to estimate is much larger than the number of observations, effects are estimated using penalized multiple regression approaches, shrinking estimated effect sizes towards zero. In addition, the magnitude of estimated SNP effects from RRBLUP is a function of the marker density. The higher the marker density, the more SNPs will be in LD with a causal mutation; therefore, the true allele substitution effect of a causal polymorphism will be split up and assigned in parts to a series of SNPs in the respective haplotype block. This can mask both the effect size, because one large effect may come in many small pieces; and the mapping position, because any SNP in LD with the causal polymorphism may have a substantial estimated effect. Nevertheless, some of the largest SNP effects from RRBLUP are in the proximity of prominent SNPs identified in the GWAS, so that to some extent positional information can still be retrieved from the RRBLUP results.

A methodology combining the strengths of both approaches – unbiased effect estimates and high positional resolution of GWAS with the simultaneous analysis of all SNPs, high predictive power and quality control via CV of genomic approaches – still needs to be developed. Results obtained in our study cannot be directly compared to predictive abilities in human studies due to the extremely small training set size (120 in CV), and Drosophila has much larger N_e and rapid decline of LD compared to humans. When genomic prediction in human studies was based on large training sets (thousands), substantial SNP panels (400k) and a highly heritable trait ($h^2 = 0.80$), predictive ability of genomic models was found to exceed what has been previously reported using a reduced number of markers pre-selected based on GWAS (Makowsky *et al.*, 2011) and genomic prediction based on pre-selected SNPs was found to be of limited use in human studies of height (Aulchenko *et al.*, 2009).

In the near future individual whole genome sequences will become increasingly available for large numbers of individuals in many species (The 1000 Genomes Project Consortium, 2010; Elshire *et al.*, 2011). Sequence-based predictions will therefore be relevant for prediction of risk disease and individualized medicine in humans, and for genome-based selection in farm animals and crops. The main findings of our study are: (i) genomic prediction can be efficiently implemented via GBLUP with full genome sequence data; (ii) there is little, if any, gain in predictive ability if the number of SNPs is increased above $14.6N_eL$ (equivalent to $\approx$ 43,000 in Holstein cattle and 1,300,000 in humans); and (iii) approaches based on external or internal (BayesB) selection of subsets of SNPs were not found to provide a substantial gain in predictive ability compared to GBLUP. All findings must be seen against the background of the small sample size and the specific genetic constellation, with almost unrelated inbred lines and highly accurate phenotypes. Nevertheless, these results provide a realistic assessment of the potential benefits of sequenced-based prediction applied to non-model organisms and indicate avenues for future research.

4.4 Materials and methods

4.4.1 The "*Drosophila melanogaster* Genetic Reference Panel" (DGRP)

The full "*Drosophila melanogaster* Genetic Reference Panel" (DGRP) (Mackay *et al.*, 2012), a recently developed new community resource for genetic studies of complex traits, consists of 192 *D. melanogaster* lines derived by 20 generations of full-sib mating from wild-caught females from the Raleigh, North Carolina population. Whole genome sequence data of 168 DGRP lines (Freeze 1.0) have been obtained using a combination of Illumina and 454 next generation sequencing technology, which are available from the Baylor College of Medicine, `http://www.hgsc.bcm.tmc.edu/project-species-i-DGRP_lines.hgsc`. We used the Illumina sequences for 157 DGRP lines in this study.

4.4.2 Data preprocessing

SNPs were called from the raw sequence data as described previously (Mackay *et al.*, 2012). We used SNPs with a coverage greater than 2X but less than 30X, for which the minor allele frequency was present in at least four lines, and for which SNPs were called in at least 60 lines. This series of filters gave a total of 2,490,165 SNPs for this analysis; 582,024 on *2L*, 478,218 on *2R*, 563,094 on *3L*, 534,979 on *3R* and 331,850 on the *X* chromosome. We did not consider the few SNPs on the very short chromosome *4*. In total there were 18,077,784 missing SNP genotypes (4.6%), which we imputed using Beagle Version 3.3.1 software (Browning & Browning, 2009).

4.4.3 Phenotypic values

Phenotypic measurements for starvation resistance were available for all 157 DGRP lines, and for startle response on 155 lines (Mackay *et al.*, 2012). We used the average of the medians of measurements for each trait in males and females as the phenotypic value y_i of the ith line, *i.e.* $y_i = 0.5((z_f)_i + (z_m)_i)$, where $(z_f)_i$ and $(z_m)_i$ are the medians of the measurements for female and male individuals of the ith line. We used medians because of the skewed distribution of traits; however, medians are highly correlated with line means. For starvation resistance (startle response) there were on average $52 \pm 11\,(40 \pm 4)$ measurements for females, and $52 \pm 11(40 \pm 4)$ measurements for males (Table 4.1). Measurements were taken in several replicates for each trait (Mackay *et al.*, 2012).

4.4.4 Cross-validation

We used different cross-validation (CV) procedures (Stone, 1974, 1977; Allen, 1974) to assess the predictive ability of different methods. In one replicate of a CV, the lines are randomly divided into a training set, which is used for parameter estimation; and a validation set, for which genetic values are predicted. The CV procedures differ in the ratios of the numbers of lines belonging to the training and validation sets: In a $(t : v)$-CV (with integers t and v), the lines are randomly divided into $(t + v)$ groups. The t groups build the training set, and the remaining v groups build the validation set. For this classification, there are $\binom{t+v}{t}$ possibilities. For each of these possibilities ("folds"), total genetic values for the lines of the validation set are predicted and the corresponding predictive ability is calculated. The $\binom{t+v}{t}$ predictive abilities are then averaged to obtain one average correlation per CV replicate. For example, one (3:2)-CV, consists of $\binom{3+2}{3} = 10$ CV folds, over which predictive abilities are averaged. A $(t : 1)$-CV is also called $(t + 1)$-fold CV.

We used (4:1)-, (3:2)-, (2:3)- and (1:4)-CVs to analyze the effect of decreasing training set size. The CVs also differed in the constellations of phenotypic records used for the training and validation set. For example, the notation "(4:1) male – female" indicates that only the medians of male records were used in the training set, and that the predicted genetic values were correlated with the medians of female records of the validation set to obtain the predictive ability in a (4:1)-CV. CVs were also run for different marker densities, using every 2^k-th SNP ($k = 0,1,\dots,10$). Additionally, 5-fold CVs using only the 5% SNPs with the largest absolute values of estimated effects (obtained in the training set), or using only the 5% SNPs with the largest SNP variances (obtained in the training set) were performed. The additive genetic variance marked by the ith SNP was calculated as $4p_i(1 - p_i)\hat{s}_i^2$ with allele frequency p_i and estimated SNP effect $\hat{s}_i$. In another series of 5-fold CVs we randomly chose 77,817 SNPs to build the genomic relationship matrix or we randomly chose 10 blocks of adjacent SNPs (each block consisting of 7,781 SNPs). In an additional 5-fold CV we excluded the lines in the two blocks of higher relatedness (Figure 4.2) from the data. Each type of CV was replicated 20 times, resulting in 20 average predictive abilities.

We also analyzed the influence of minor allele frequency on the predictive ability by another series of 5-fold CV. For this, we sorted all SNPs by their minor allele frequency and divided the sorted vector into 32 blocks. For each block we ran 20 replicates of a 5-fold CV using

GBLUP and the corresponding ≈ 78,000 SNPs.

4.4.5 Predictive ability and accuracy

Predictive ability was measured in terms of correlation between predicted genetic values and observed phenotypic values. The corresponding accuracy ρ, defined as the correlation between true and predicted genetic value, was obtained by dividing the observed predictive ability by the square root of the observed heritability h^2 (Legarra *et al.*, 2008). The heritability was based on the GBLUP model (see below).

4.4.6 Genomic prediction with GBLUP

The underlying statistical model is

$$\mathbf{y} = \mathbf{W}\mu + \mathbf{Z}\mathbf{g} + \mathbf{e}. \tag{4.1}$$

In this model, the ith component of the q-vector $\mathbf{y}$ is the phenotypic value of the ith line that is used for prediction, *i.e.* the average of the medians of the phenotypic measurements for males and females for this line. Moreover, $\mathbf{W} = (1, \ldots, 1)^T, \mu$ is the overall mean; $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{G})$ is assumed to be multivariate normal, with $\mathbf{G}$ the genomic relationship matrix of all n lines (VanRaden, 2008) and σ_g^2 the additive genetic variance among lines. The matrix $\mathbf{Z}$ is an $(q \times n)$-incidence matrix, whose rows consist of unit vectors with one component being 1 and all the others zero, indicating the respective positions of lines used for prediction in the $\mathbf{g}$-vector of genetic values of all lines. The term $\mathbf{e} \sim \mathcal{N}(0,\sigma_e^2\mathbf{I})$ is the residual, where σ_e^2 is the residual variance. Following the approach of VanRaden (2008), $\mathbf{G}$ was defined as

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})^T}{2\sum_{j=1}^{s} p_j(1 - p_j)},$$

where $\mathbf{M}$ is the $(n \times s)$-matrix of SNP genotype vectors for the n lines with the s SNPs coded as -1,1 and the jth column of $\mathbf{P}$ is $(2(p_j - 0.5), \ldots, 2(p_j - 0.5))^T$, where p_j is the frequency of the second allele at locus j.

Note that the GBLUP approach is the same as the reference approach considered in section 3.2.3, but without including a polygenic component $\mathbf{u}$.

Variance components were estimated via maximum likelihood (ML) using the R-package “RandomFields”, Version 2.0.46, and its function “fitvario”. The BLUP approach to obtain the vector of genetic values is equivalent to solving the following *Mixed Model Equations* (MME), *cf.* section 2.2:

$$\begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W} & \mathbf{Z}^T\mathbf{Z} + \frac{\sigma_e^2}{\sigma_g^2}\mathbf{G}^{-1} \end{bmatrix} \cdot \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}$$

A narrow-sense heritability based on the GBLUP model (4.1) was calculated as

$$\hat{h}^2_{\text{GBLUP}} = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \hat{\sigma}^2_e}.$$

4.4.7 Estimation of SNP effects

The GBLUP model (4.1) is equivalent to the following linear model (also termed *random regression* BLUP model) in which all SNPs are assumed to have an effect drawn from the same normal distribution (Goddard, 2009):

$$\mathbf{y} = \mathbf{W}\mu + \mathbf{Z}(\mathbf{M} - \mathbf{P})\mathbf{s} + \mathbf{e},$$

where $\mathbf{Z}, \mathbf{M}$ and $\mathbf{P}$ are as described above and $\mathbf{s} \sim \mathcal{N}(0, \sigma^2_s \mathbf{I})$ is the vector of SNP effects with $\sigma^2_s = \frac{\sigma^2_g}{2\sum_{j=1}^{s} p_j(1-p_j)}$. Using this equivalence, the SNP effects can be predicted as

$$\begin{aligned}\hat{\mathbf{s}} &= \hat{\sigma}^2_s \mathbf{I}(\mathbf{M} - \mathbf{P})^T \mathbf{Z}^T (\hat{\sigma}^2_s \mathbf{Z}(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})^T \mathbf{Z}^T + \hat{\sigma}^2_e \mathbf{I})^{-1}(\mathbf{y} - \mathbf{W}\hat{\mu}) \\ &= \frac{\hat{\sigma}^2_g}{2\sum_{j=1}^{s} p_j(1-p_j)} (\mathbf{M} - \mathbf{P})^T \mathbf{Z}^T (\hat{\sigma}^2_g \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \hat{\sigma}^2_e \mathbf{I})^{-1}(\mathbf{y} - \mathbf{W}\hat{\mu}).\end{aligned}$$

To estimate the SNP effects resulting from GBLUP for a single trait, we used all of the available lines, *i.e.* $\mathbf{y}$ in model (4.1) contained the phenotypic values of all lines so that $\mathbf{Z} = \mathbf{I}$ in the corresponding formulas. Note that only the inversion of a matrix of size equal to the number of sequenced lines is required.

4.4.8 Distribution of linkage disequilibrium

We used r^2 (Hill & Weir, 1994) as a measure of LD between a pair of loci. With two biallelic loci A and B with alleles A_1, A_2, B_1, and B_2 and frequencies p_{A_1},p_{A_2}, p_{B_1}, and p_{B_2}, we denote the frequencies of the genotypes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 as x_{11}, x_{12}, x_{21}, and x_{22} respectively. Then,

$$r^2 = \frac{(x_{11}x_{22} - x_{12}x_{21})^2}{p_{A_1}p_{A_2}p_{B_1}p_{B_2}}.$$

We performed the LD analyses using the imputed SNP matrix of ≈ 2.5 million SNPs for the 157 lines. We calculated the distribution of LD between all pairs of neighboring SNPs for different marker densities, using every 2^k-th SNP ($k = 0, 1, \ldots, 10$). The extent of long-range LD was calculated for 20,000 pairs of SNPs randomly sampled from the first and the last 50,000 SNPs per chromosome arm. Moreover, the average LD was calculated between SNPs on different chromosome arms, by sampling 10,000 pairs of SNPs for each combination of chromosome arms.

4.4.9 Effective population size derived from empirical accuracies of genomic prediction:

We modified the formula of Daetwyler *et al.* (2010) for the expected accuracy, $\mathbb{E}(\rho)$, of GBLUP given different population parameters (see section 4.5 for more details on the derivation in the case of *D. melanogaster*):

$$\mathbb{E}(\rho) = \sqrt{\frac{N_p h^2}{N_p h^2 + \frac{N_e L_f}{\ln(2N_e L_f)}}} \tag{4.2}$$

N_e is the effective population size, N_p is the size of the training set, L_f is the length of the female genome in Morgans and h^2 is the narrow-sense heritability of the trait estimated from model (4.1). The term $M_e = \frac{N_e L_f}{\ln(2N_e L_f)}$ describes the number of independently segregating genome segments (Goddard, 2009).

We ran CVs with different numbers of lines ($N_{p,1} = 31.4, N_{p,2} = 62.8, N_{p,3} = 94.2, N_{p,4} = 125.6$ for starvation resistance and $N_{p,1} = 31, N_{p,2} = 62, N_{p,3} = 93, N_{p,4} = 124$ for startle response) in the training set (20 replicates each). Average numbers of lines in the training set are reported, which are non-integer values for starvation resistance because in a $(t + v)$-CV, division of 157 lines into $t + v$ groups may give unequal numbers of lines in the different partitions. Given the corresponding average accuracies $\rho_{ij}, i = 1, \dots, 4, j = 1, \dots, 20$ for the CV replicates, we estimated N_e by fitting a curve to the points $(N_{p,i}, \rho_{ij})$. To fit the curve, we chose N_e such that the sum of the squared differences of the observed accuracies and the accuracies obtained by (4.2) was minimized:

$$\hat{N}_e = \underset{N_e}{\operatorname{argmin}} \left[\sum_{i,j} \left(\rho_{ij} - \sqrt{\frac{N_{p,i} h^2}{N_{p,i} h^2 + \frac{N_e L_f}{\ln(2N_e L_f)}}} \right)^2 \right],$$

using $\hat{h}^2 = \hat{h}^2_{\text{GBLUP}} = 1$ and $L_f = 2.451$ Morgan. We calculated the length of the female genome in Morgans by summing the lengths of the chromosomes in base-pairs (23.0 (21.4, 24.4, 28.0, 21.8) Mbp for chromosome *2L* (*2R*, *3L*, *3R*, *X*), Adams *et al.* (2000)) and multiplying by the average recombination rates of females for the different chromosomes in Morgans per base-pair (Fiston-Lavier *et al.*, 2010).

After performing bootstrapping (1,000 replicates), the bias corrected empirical 95% confidence intervals (2.5% error in each tail) for the N_e estimates (Efron & Tibshirani, 1986; Efron, 1987) were calculated as

$$\left[\hat{G}^{-1}(\Phi(2z_0 + z^{(\alpha)}), \hat{G}^{-1}(\Phi(2z_0 + z^{(1-\alpha)})\right],$$

where $\hat{G}^{-1}(\alpha)$ is the 100α-percentile of the bootstrap cumulative distribution function, $z^{(\alpha)}$ is the 100α-percentile of the standard normal distribution function Φ, $\alpha = 0.025$ and $z_0 = \Phi^{-1}(\hat{G}(\hat{N}_e))$.

4.4.10 Effective population size derived directly from linkage disequilibrium

To estimate the effective population size based on LD, the following formula was used (Sved, 1971):

$$\mathbb{E}(r^2) = \frac{1}{1+2N_e c_f} + \frac{1}{n} \quad \Leftrightarrow \quad N_e = \frac{\frac{1}{\mathbb{E}(r^2)-\frac{1}{n}} - 1}{2c_f},$$

where n is the number of lines and c_f is the recombination rate in female individuals, *cf.* section 4.5 for more details on this formula.

4.4.11 Genomic prediction with BayesB (Meuwissen *et al.*, 2001)

The underlying model for the Markov Chain Monte Carlo based BayesB method is

$$\mathbf{y} = \mathbf{W}\mu + \mathbf{M}\mathbf{s} + \mathbf{e},$$

where $\mathbf{y}, \mathbf{W}, \mu, \mathbf{M}$ and $\mathbf{e}$ are as defined previously and $\mathbf{s}$ is the vector of normally distributed and independent SNP effects. The variance of the ith SNP effect, $\sigma^2_{s_i}$, is assigned an informative prior. The prior distribution of the genetic variances aims to resemble a situation where there are many loci with zero variance and only some loci with variance not equal to zero. Therefore, the prior distribution of the variance of a marker effect is a mixture of distributions which is given by

$$\sigma^2_{s_i} \begin{cases} = & 0 \text{ with probability } \pi \\ \sim & \chi^{-2}(\nu, S) \text{ with probability } (1-\pi). \end{cases}$$

Note that this implies that the unconditional distribution of each single marker effect is a mixture of a point mass at 0 (with probability π) and of a t-distribution with zero mean, ν degrees of freedom and scale parameter S (Gianola *et al.*, 2009), *i.e.* BayesB assigns the same unconditional prior distribution to each marker effect.

In our studies, we used $\nu = 4$ and the scale parameter S was calibrated as

$$S = \frac{(\nu-2)\sigma^2_{\text{genetic}}}{(1-\pi)\nu\sum_{j=1}^{s} 2p_j(1-p_j)}.$$

We chose $\pi = 0.995$, such that approximately 125,000 markers were contributing to the additive genetic variance. For the residual variance, σ^2_e, the prior distribution was $\chi^{-2}(\nu_{\text{res}}, S_{\text{res}})$, with $\nu_{\text{res}} = 10$ and

$$S_{\text{res}} = \frac{(\nu_{\text{res}}-2)\sigma^2_{\text{res}}}{\nu_{\text{res}}}.$$

Values for $\sigma^2_{\text{genetic}}$ and σ^2_{res} were chosen in the order of magnitude of the variance components of the GBLUP model (4.1), which were estimated using all lines and “fitvario”. The BayesB procedure is described in detail in Meuwissen *et al.* (2001). It consists of running a Gibbs chain, where additionally a Metropolis-Hastings algorithm (10 iterations) is used to sample from $p(\sigma^2_{s_i}|\mathbf{y}^*)$, where $\mathbf{y}^*$ denotes the data $\mathbf{y}$ corrected for the mean μ and all genetic effects other than the marker effect s_i. Following graphical inspection, we ran BayesB with a chain length of 40,000 iterations including a burn in of 5,000 iterations that were discarded. To perform the BayesB approach, we used GenSel (Fernando & Garrick, 2009), which is implemented in C++. BayesB is computationally very intensive. The analyses were run on a Mac Pro 2x 2.93 GHz 6-Core Intel Xeon with 64 GB RAM running Mac OS X Server 10.6.7. One fold of a 5-fold CV for starvation resistance took approximately 70 hours.

4.4.12 Comparing areas with large SNP effects with significant SNP positions

A genome-wide association study (GWAS) revealed 203 (90) significant SNP positions for starvation resistance (startle response) (Mackay *et al.*, 2012), where a SNP position was considered significant if at least one of the three p-values, obtained using only male, only female or sex-pooled phenotypic records, was $\leq 10^{-5}$. We considered the subset of SNPs for which p-values of SNP effects of pooled data were $\leq 10^{-5}$, to be more conservative and to be consistent with the previous analyses, leading to 115 (75) significant SNPs for starvation resistance (startle response).

We compared genomic regions for which GBLUP estimated large SNP effects to these significant SNP positions of the GWAS. To avoid an effect of different sample sizes, we chose the 75 most significant SNPs from the GWAS analysis for each trait. For each of these SNPs, we chose the 100 closest (neighboring) SNPs (50 on each side) and calculated the sums of absolute values of the corresponding 100 SNP effects (resulting from the GBLUP model). We compared the distribution of these sums to the distribution of the sums of the absolute values of estimated SNP effects in $\approx$ 250,000 windows of 100 neighboring SNPs covering the whole genome by plotting the corresponding density functions. To obtain the sums of the absolute values of estimated SNP effects covering the whole genome, the windows were overlapping, displaced by 10 SNP positions. If the genomic regions for which GBLUP estimated large SNP effects coincide with the significant SNP positions of the GWAS, we expect the density functions to be separated.

4.4.13 Variance component estimation using ASReml (Gilmour *et al.*, 2006) and individual trait records

For each trait, we fitted three different models using *individual* trait records. The first model included a fixed sex effect, a random line effect, a random line-sex-interaction term and a random term accounting for the different replicates in which measurements of the traits were taken:

$$\text{phenotype} = \mu + \text{sex} + \text{line} + \text{sex} * \text{line} + \text{replicate}(\text{sex} * \text{line}) + \text{residual} \qquad \text{(Model 1)}$$

In the second model, an additional random genetic effect g was added for each line. The variance-covariance matrix of the vector of these genetic effects was assumed to be given by the genomic relationship matrix **G** of VanRaden (2008):

$$\text{phenotype} = \mu + \text{sex} + \text{line} + \text{sex} * \text{line} + \text{replicate}(\text{sex} * \text{line}) + g + \text{residual} \qquad \text{(Model 2)}$$

In the third model, an additional random additive × additive epistatic effect $g \times g$ was included for each line. The variance-covariance matrix of the vector of these genetic effects was given by the Hadamard product $\mathbf{G} \circ \mathbf{G}$ (Henderson, 1984) of the genomic relationship matrix **G** of VanRaden (2008):

$$\begin{aligned}\text{phenotype} = \mu + \text{sex} + \text{line} + \text{sex} * \text{line} + \text{replicate}(\text{sex} * \text{line})\\ + g + (g \times g) + \text{residual} \qquad \text{(Model 3)}\end{aligned}$$

Other two-way epistatic interactions, like additive × dominance or dominance × dominance, should not exist in inbred lines, provided inbreeding is complete. Variance components and their standard errors were estimated using ASReml 2.0 (Gilmour *et al.*, 2006). The analyses were done pooled across sexes as well as separately for males and females. The analyses of separate sexes did not include the sex term, and the replicate(sex∗line) term was reduced to replicate(line).

4.4.14 Heritabilities

The broad-sense heritability for Model 1 was calculated as

$$\hat{H}^2_{\text{Model 1}} = \frac{\hat{\sigma}^2_{\text{line}} + \hat{\sigma}^2_{\text{sex}*\text{line}}}{\hat{\sigma}^2_{\text{line}} + \hat{\sigma}^2_{\text{sex}*\text{line}} + \hat{\sigma}^2_{\text{residual}}},$$

cf. Ayroles *et al.* (2009). Narrow sense heritabilities for Models 2 and 3 were calculated as

$$\hat{h}^2_{\text{Model 2}} = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_{\text{line}} + \hat{\sigma}^2_{\text{sex}*\text{line}} + \hat{\sigma}^2_g + \hat{\sigma}^2_{\text{residual}}}$$

and

$$\hat{h}^2_{\text{Model 3}} = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_{\text{line}} + \hat{\sigma}^2_{\text{sex}*\text{line}} + \hat{\sigma}^2_g + \hat{\sigma}^2_{g\times g} + \hat{\sigma}^2_{\text{residual}}}.$$

These heritabilities are based on individual trait records.

Unless stated otherwise, all statistical analyses were performed using R software (R Development Core Team, 2012; Ihaka & Gentleman, 1996). The R-package “ff” (Adler *et al.*, 2012) was used to handle the large amount of SNP data efficiently in terms of memory capacity.

4.5 More details on the expected LD and M_e

When working with *D. melanogaster*, we have to pay attention to a specific characteristic: Male individuals do not recombine, *i.e.* the overall recombination rate c equals $\frac{1}{2}c_f$, where c_f is the recombination rate in female individuals. Moreover, the genome length in Morgans is $L = 0.5L_f$, where L_f is the length of the female genome in Morgans.

4.5.1 The formula of Sved (1971) for the expected linkage disequilibrium

The following formula for the expected LD at equilibrium in a population based on the effective population size N_e was proposed by Sved (1971):

$$\mathbb{E}(r^2) = \frac{1}{1 + 4N_e c} \quad \Leftrightarrow \quad N_e = \frac{\frac{1}{\mathbb{E}(r^2)} - 1}{4c} \tag{4.3}$$

Here, N_e corresponds to an effective population size $t = \frac{1}{2c}$ generations ago (Hayes *et al.*, 2003). Using $c = \frac{1}{2}c_f$ we obtain

$$\mathbb{E}(r^2) = \frac{1}{1 + 2N_e c_f} \quad \Leftrightarrow \quad N_e = \frac{\frac{1}{\mathbb{E}(r^2)} - 1}{2c_f},$$

$t = \frac{1}{c_f}$ generations ago.

If this formula is used to estimate N_e based on a finite sample of individuals, one should adjust for the chromosome sample size (Weir & Hill, 1980), which equals the number of individuals n in the case of inbred lines. Then,

$$\mathbb{E}(r^2) = \frac{1}{1 + 2N_e c_f} + \frac{1}{n} \quad \Leftrightarrow \quad N_e = \frac{\frac{1}{\mathbb{E}(r^2) - \frac{1}{n}} - 1}{2c_f}.$$

Note that when applying this formula to the DGRP population, the estimated N_e is not the effective population size of the local wild population the actual lines were sampled from, but the effective population size of an idealized population having the same structure of LD as the DGRP inbred lines. This means that we consider the 157 independent gametes of the DGRP inbred lines as a random sample of this idealized population.

Several derivations of the above formula have been suggested in the last forty years (Sved, 1971; Sved & Feldmann, 1973; Tenesa *et al.*, 2007; Sved, 2008, 2009) and simulation studies have shown that the simulated values of r^2 agree reasonably well with the expectations based on this formula. However, we found that all derivations mentioned above have serious shortcomings from a mathematical point of view. Similar concerns over the exact validity of the formula and their derivations were recently raised by Sved (2008), p. 185, *cf.* also the manuscript published on John Sved's personal homepage (`http://www.handsongenetics.com/PIFFLE/LinkageDisequilibrium.pdf`). We clearly think that further research is needed to find a substantiated derivation and that results based on this formula should therefore be taken with caution. We will deal with this issue in detail in chapter 6, also proposing an alternative formula for the expected LD in a finite population.

4.5.2 More details on the derivation of the number of independently segregating chromosome segments M_e and the expected accuracy of prediction $\mathbb{E}(\rho)$:

The formula of Daetwyler *et al.* (2010) for the expected accuracy of genomic prediction $\mathbb{E}(\rho)$ with GBLUP depends on the number of independently segregating genome segments M_e (Goddard, 2009):

$$\mathbb{E}(\rho) = \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}}$$

We will give more details on how M_e can be calculated in the case of *D. melanogaster*. The derivation of M_e for a diploid population is given in Goddard (2009) and based on the formula of Sved for the expected LD $\mathbb{E}(r^2)$ at equilibrium (Sved, 1971). Central in Goddard's derivation is the calculation of the double integral over the formula for $\mathbb{E}(r^2)$. In general, one can verify that

$$\begin{aligned}&\frac{1}{a_1^2}\int_0^{a_1}\int_0^{a_1}\frac{1}{a_3+a_2|x_1-x_2|}dx_1dx_2\\ =&\frac{2(a_3+a_1a_2)\ln(a_3+a_1a_2)}{a_1^2a_2^2}-\frac{2a_3\ln(a_3)}{a_1^2a_2^2}-\frac{2\ln(a_3)}{a_1a_2}-\frac{2}{a_1a_2},\end{aligned}$$

for arbitrary constants a_1, a_2, a_3 with $a_1, a_2 > 0$. If $a_3 \in \{1,2\}$ and if a_2 is large enough, the double integral is approximately

$$\frac{1}{a_1^2}\int_0^{a_1}\int_0^{a_1}\frac{1}{a_3+a_2|x_1-x_2|}dx_1dx_2 \approx \frac{2(a_1a_2)\ln(a_1a_2)}{a_1^2a_2^2} = \frac{2\ln(a_1a_2)}{a_1a_2}.$$

Following the derivation of Goddard (2009), we need to calculate the double integral over eq. (4.3) and displace c by the distance $|x_1 - x_2|$ which leads to

$$\begin{aligned}\frac{1}{L^2}\int_0^{L}\int_0^{L}\frac{1}{1+4N_e|x_1-x_2|}dx_1dx_2 &= \frac{1}{L_f^2}\int_0^{L_f}\int_0^{L_f}\frac{1}{1+2N_e|x_1-x_2|}dx_1dx_2\\ &\approx \frac{2\ln(L_f2N_e)}{L_f2N_e} = \frac{\ln(L_f2N_e)}{L_fN_e}.\end{aligned}$$

Here, the first equality holds because of the transformation formula and the identity $L = \frac{1}{2}L_f$ in the case of *D. melanogaster*. Using this result, M_e can be derived as in Goddard (2009), leading to

$$M_e = \frac{N_eL_f}{\ln(2N_eL_f)}.$$

Hence, the formula of Daetwyler *et al.* (2010) for the expected accuracy of prediction in the case of *D. melanogaster* equals

$$\mathbb{E}(\rho) = \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}} = \sqrt{\frac{N_p h^2}{N_p h^2 + \frac{N_e L_f}{\ln(2N_e L_f)}}},$$

where N_p is the size of the training set and h^2 is the narrow-sense heritability of the trait estimated from the GBLUP model.

Note that we do not claim at this point that Goddard's approach to derive the formula for M_e is correct.

4.6 The expected value of the genomic relationship matrix of VanRaden (2008)

In this section we will show that the expected value of the genomic relationship matrix $\mathbf{G}$ of VanRaden (2008) is given by the additive relationship matrix $\mathbf{A}$, *i.e.*

$$\mathbb{E}(\mathbf{G}) = \mathbf{A}.$$

Following VanRaden (2008), $\mathbf{G}$ is defined as

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})^T}{2\sum_{k=1}^{s} p_k(1 - p_k)},$$

where $\mathbf{M}$ is the $(n \times s)$-matrix of SNP genotype vectors for the n lines with the s SNPs coded as -1,1 and the kth column of $\mathbf{P}$ is $(2(p_k - 0.5), \ldots, 2(p_k - 0.5))^T$, where p_k is the frequency of the second allele at locus k for $k = 1, \ldots, s$.

Let $\mathbf{m}_i$ be the vector of SNP genotypes of individual i, *i.e.* $\mathbf{m}_i = (m_{i1}, \ldots, m_{is})$. Then, $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_n)^T$. We consider the case of fully homozygous individuals due to full-sib mating. Then, the genotype m_{ik} of individual $i = 1, \ldots, n$ at locus $k = 1, \ldots, s$ can be considered as a discrete random variable with values -1,1 and probabilities $(1 - p_k), p_k$, and it is

$$\mathbb{E}(m_{ik}) = -(1 - p_k) + p_k = 2p_k - 1$$

for all $i = 1, \ldots, n$. Moreover, we have

$$\sum_{k=1}^{s} \mathrm{Cov}(m_{ik}, m_{jk}) = a_{ij} \sum_{k=1}^{s} \sigma^2_{\mathbf{m}_{\bullet k}},$$

where a_{ij} is the coefficient of relationship between individuals i and j, and $\sigma^2_{\mathbf{m}_{\bullet k}}$ is the variance of the genotype variable $m_{\bullet k}$ at locus k of the original base-population, see Cornelius & Dudley (1975) for a derivation of the covariance between relatives under full-sib mating. The variance of $m_{\bullet k}$ in the base population is equal to the variance of a random variable with values -1,0,1 and probabilities $(1-p_k)^2, 2p_k(1-p_k), p_k^2$, which equals

$$\begin{aligned}\sigma^2_{\mathbf{m}_{\bullet k}} &= \mathbb{E}(\mathbf{m}^2_{\bullet k}) - \mathbb{E}(\mathbf{m}_{\bullet k})^2 \\ &= (-1)^2 \cdot (1-p_k)^2 + 0^2 \cdot 2p_k(1-p_k) + 1^2 \cdot p_k^2 \\ &\qquad - \Big(-1 \cdot (1-p_k)^2 + 0 \cdot 2p_k(1-p_k) + 1 \cdot p_k^2\Big)^2 \\ &= 2p_k(1-p_k).\end{aligned}$$

This leads to

$$\sum_{k=1}^{s} \mathrm{Cov}(m_{ik}, m_{jk}) = a_{ij} \sum_{k=1}^{s} 2p_k(1-p_k). \tag{4.4}$$

Define $D := 2\sum_{k=1}^{s} p_k(1-p_k)$. The expected value of $\mathbf{G}$ can now be calculated as

$$\begin{aligned}[\mathbb{E}(\mathbf{G})]_{ij} &= \left[\mathbb{E}\left(\frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})^T}{D}\right)\right]_{ij} \\ &= \frac{1}{D}\mathbb{E}\left[(\mathbf{m}_i - (2(p_1-0.5),\ldots,2(p_s-0.5))) \cdot (\mathbf{m}_j - (2(p_1-0.5),\ldots,2(p_s-0.5)))^T\right] \\ &= \frac{1}{D}\mathbb{E}\left[((m_{i1},\ldots,m_{is}) - \mathbb{E}(m_{i1},\ldots,m_{is})) \cdot ((m_{j1},\ldots,m_{js}) - \mathbb{E}(m_{j1},\ldots,m_{js}))^T\right] \\ &= \frac{1}{D}\sum_{k=1}^{s} \mathbb{E}\left[(m_{ik} - \mathbb{E}(m_{ik})) \cdot (m_{jk} - \mathbb{E}(m_{jk}))\right] \\ &= \frac{1}{D}\sum_{k=1}^{s} \mathrm{Cov}(m_{ik}, m_{jk}) \\ &= \frac{1}{2\sum_{k=1}^{s} p_k(1-p_k)} \left(a_{ij}\sum_{k=1}^{s} 2p_k(1-p_k)\right), \text{using eq. (4.4)} \\ &= a_{ij}\end{aligned}$$

for $i,j = 1,\ldots,n$, *i.e.* $\mathbb{E}(\mathbf{G}) = \mathbf{A}$.

The derivation presented above was for the case of fully homozygous individuals due to full-sib mating. The identity $\mathbb{E}(\mathbf{G}) = \mathbf{A}$ can analogously be derived for a non-homozygous population. Then, the genotype m_{ik} of individual i at locus k can be considered as a discrete random variable with values -1,0,1 and probabilities $(1-p)^2, 2p_k(1-p_k), p_k^2$.

5 Analyses of Chill Coma Recovery Data: Evidence of Epistatic Interactions

5.1 Introduction

In addition to phenotypes of starvation resistance and startle response, the DGRP data also comprised phenotypic records of the trait "chill coma recovery" (Jordan *et al.*, 2007; Ayroles *et al.*, 2009; Mackay *et al.*, 2012) for 147 lines. This trait describes the time to recover from a chill-induced coma and builds a component of fitness in *Drosophila* and other insects.

We analyzed this trait using the same procedure as described in the previous chapter. In contrast to the results reported for starvation resistance and startle response, we found that for chill coma recovery genomic-based prediction had essentially no predictive ability in the analyses of sex-averaged and male data, and that it had very low predictive ability for the female data.

In search for possible explanations for this behavior, we encountered several characteristics of the DGRP data:

- a region on chromosome *2L* dividing the lines into two clusters
- the bimodality of the phenotypic chill coma recovery data

While further analyzing both characteristics, we found strong hints of

- numerous pairwise epistatic interactions of SNPs underlying the chill coma recovery trait.

In the following, we will describe the approaches taken to get to these findings. The results presented in this chapter form a basis for numerous possible further investigations. To fully understand the complexity of the chill coma recovery trait from a biological point of view, further analyses have been carried out in collaboration with the working group of Prof. Mackay, including Gene Ontology enrichment analyses as well as genetic network investigations, revealing a complex genetic architecture of the chill coma fitness trait by confirming extensive epistasis and identifying alleles with large effects. This finally led to novel insights into the underlying biology of chill coma recovery time. A joint manuscript including these continuative analyses is currently in revision for *PLoS Genetics* (Ober *et al.*, 2012*b*); its content is briefly summarized in section 5.6.

5.2 Investigations in analogy to the analyses of starvation resistance and startle response

In line with the study using whole genome sequence data for prediction described in chapter 4, the same analyses as applied to the starvation resistance and startle response data were carried out for the chill coma recovery data, using the same set of ≈ 2.5 million SNPs as before. The corresponding results are presented in the following subsections.

5.2.1 The chill coma recovery data

Phenotypic records of coma chill recovery were available for 148 out of the 157 DGRP lines. For details on the sampling procedure we refer to Mackay *et al.* (2012). There were on average 101 ± 15 measurements of female individuals, and 100 ± 16 measurements of male individuals per line. One extreme outlier-line ("RAL-879") was excluded from further analyses for this trait, in line with Mackay *et al.* (2012). The mean and standard deviation of the phenotypic values for the three traits are shown in Table 5.1.

Table 5.1: Mean and standard deviation of phenotypic values and of the number of records per line for chill coma recovery. Phenotypic values were calculated as the averages of the medians of male and female records ("all") or as the medians of female or male records separately.

	chill coma	
	phen. value[1]	# rec. per line[2]
all	16.3 ± 4.8	200.7 ± 30.7
female	16.1 ± 5.2	100.8 ± 15.4
male	16.5 ± 4.7	99.9 ± 16.0

[1] Phenotypic values.
[2] Number of records per line.

Lines for which phenotypic records of chill coma recovery were available are also marked by a "C" in the heatmap of the genomic relationship matrix according to VanRaden (2008) (Suppl. Figure S3).

5.2.2 Results of the GBLUP approach

The results in terms of predictive ability obtained with various CV procedures using the GBLUP approach (with covariance structure given by the genomic relationship matrix according to VanRaden (2008)) are shown in Table 5.2. We found that genomic-based prediction for chill coma recovery had essentially no predictive ability when using a 5-fold

CV and the sex-averaged records or the median of male records only, but that it worked with low predictive ability, if only the medians of female records were used.

Table 5.2: Average correlations between predicted genetic values and observed phenotypes of chill coma recovery for different CV procedures using GBLUP.

type of CV	correlation
(4:1)-CV all[1]	−0.038[2] (0.010)
(4:1)-CV male – female[3]	−0.053 (0.011)
(4:1)-CV female – male	−0.041 (0.008)
(4:1)-CV male – male	−0.148 (0.011)
(4:1)-CV female – female	0.051 (0.008)
(3:2)-CV female – female	0.041 (0.009)
(2:3)-CV female – female	0.023 (0.008)
(1:4)-CV female – female	0.016 (0.006)

[1] The average of the medians of male and female measurements was used to predict line phenotypes. Predicted phenotypes were then correlated with the averages of the medians of male and female measurements.

[2] Average correlation between predicted genetic values and observed phenotypes. Results are averages over 20 replicates. Standard errors of the means in parentheses.

[3] "CV sex_1 – sex_2" means: Medians of measurements of sex_1 were used in the training set, medians of sex_2 were used in the validation set.

This low predictive ability for chill coma recovery was not an artifact but was systematic, as illustrated by a series of CVs with reduced size of the training set (*cf.* Table 5.2 and Figure 5.1), where a decline of accuracy could be observed, when the size of the training set decreased. This series of CVs was performed using female measurements only, as no predictive ability could be observed for chill coma recovery with sex-averaged and male measurements even with the largest training set used in the 5-fold CV (Table 5.2).

The low predictive ability for chill coma recovery was also consistent with the fact that the narrow sense heritability estimated from the GBLUP model was 0 using sex-averaged records or only the medians of male records, while heritability was 0.09 when using the medians of female records only (Table 5.3).

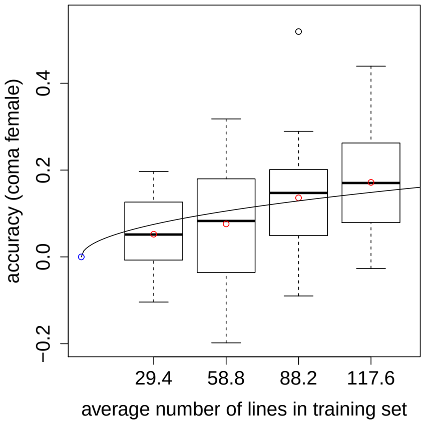

Figure 5.1: Accuracy of prediction of GBLUP for CVs with different numbers of lines in the training set using female measurements of chill coma recovery. Each boxplot illustrates the average accuracies for 20 replicates of the CV procedure using GBLUP. The solid line is the curve of Daetwyler *et al.* (2010) fitted to the empirical data. All 2.5 million SNPs were used to construct the genomic relationship matrix in the GBLUP model.

Table 5.3: Variance components and heritabilities for chill coma recovery estimated from GBLUP using all 147 lines. Variance components were estimated by maximum likelihood using the R-package "RandomFields" and its function "fitvario" and the averages of the medians of male and female records ("all") or the medians of female or male records separately as phenotypic data.

	chill coma		
	$\hat{\sigma}_g^2$	$\hat{\sigma}_e^2$	$\hat{h}^2_{\text{GBLUP}}$
all	0	22.6	0
female	2.2	22.8	0.09
male	0	21.5	0

5.2.3 Analyses of individual trait data

As for the other two traits, we also analyzed *individual* trait data to assess whether the variance between lines can be fully explained by additive genetic effects or if non-additive mechanisms have an impact. We considered the same three linear models already used in section 4.4.13. Results of these analyses are summarized in Suppl. Table S3: When including the additive × additive epistatic component $g \times g$ into the model, the estimate of the between line variance σ^2_{line} was zero, while σ^2_{line} was 25.6 when the $g \times g$ component was not included. For a trait with largely additive × additive epistatic variation, we expect $\sigma^2_{\text{line}} = 4\sigma^2_{g\times g}$, where $\sigma^2_{g\times g}$ is the additive × additive epistatic variance in the non-inbred base population. Since this is indeed what we observed ($\sigma^2_{\text{line}} = 28.6$, $\sigma^2_{g\times g} = 7.21 \approx 28.6/4$ using all records), it might be that additive × additive epistasis is an important feature of the genetic architecture of chill coma resistance. It has to be noted that all three models had basically the same likelihood (*cf.* Table S3), also stressing that the line differences cannot be assigned to additive genetic effects. In consequence, genomic prediction based on an additive model is bound to fail, which is consistent with what we observed using the GBLUP approach.

5.2.4 Comparing areas with large SNP effects with significant SNP positions

In Mackay *et al.* (2012), a GWAS revealed 235 significant SNP positions for chill coma recovery, where a SNP position was considered as “significant”, if at least one of the three p-values, obtained using only male, only female or pooled phenotypic records, was $\leq 10^{-5}$. Here, we only considered SNP positions showing a p-value $\leq 10^{-5}$ with female phenotypic records to be more conservative and to be consistent with the previous analyses of starvation resistance and startle response (*cf.* section 4.4.12), leading to 145 significant SNPs. For the 75 most significant putative QTLs from the GWAS of Mackay *et al.* (2012), we considered the 100 neighboring SNP positions and calculated the sum of the absolute values of their estimated SNP effects (using the GBLUP model), along the lines of the analyses of the other two traits starvation resistance and startle response in section 4.4.12. These sums were compared to the sums of the absolute values of estimated SNP effects in ≈ 250,000 windows of 100 neighboring SNPs covering the whole genome, *cf.* Figure 5.2 for the density functions of these sums. For chill coma recovery, the separation of the densities is small (using female records only), as opposed to what we observed for the other two traits (*cf.* Figure 4.8).

A Manhattan plot of the estimated SNP effects obtained with GBLUP is shown in Suppl. Figure S4, also indicating the positions of significant SNPs according to the GWAS.

Overall, results indicate that the proportion of causative genetic factors captured by the GWAS is only poorly corresponding to the estimated SNP effects from the genomic model, and the accordance of large estimated SNP effects with significant markers is less pronounced for chill coma recovery in comparison to starvation resistance and startle response.

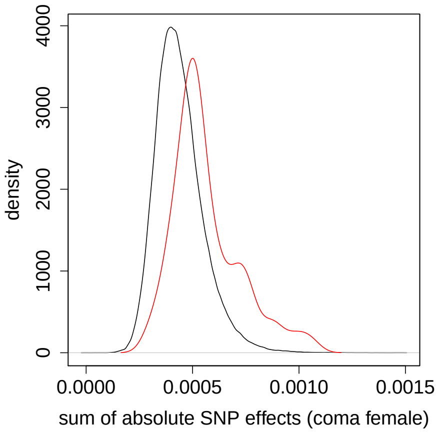

Figure 5.2: Distribution of estimated SNP effects for chill coma recovery. The density of the sum of the absolute values of the SNP effects (estimated from the GBLUP model) for chill coma recovery is plotted for sliding windows of 100 adjacent SNPs covering the whole genome (black) and for windows around the 75 most significant SNPs (red) according to the GWAS of Mackay *et al.* (2012). Only female measurements were used.

5.3 Observation I: two clusters of lines in relation to a large region on chromosome *2L*

5.3.1 Phenotypic differences

To further assess the potential for genomic prediction, we applied a diagnostic tool based on the expectation, that if a trait is inherited under an additive model, pairs of individuals with higher additive genomic relationship coefficient should be more similar in phenotype. This hypothesis was tested by fitting a linear regression of the squared differences of the standardized phenotypic values on the additive genomic relationship coefficient for all pairs of lines. This was done for sex-averaged records as well as for male and female measurements separately and for all three traits. For the pair of lines i and j, for example, $(y_i - y_j)^2$ was plotted against the entry g_{ij} of the genomic relationship matrix $\mathbf{G}$, where y_i and y_j denote the phenotypic records of lines i and j after standardization. This was repeated for all possible combinations of i and j. A linear regression was fitted and the one-sided null hypothesis $b > 0$ (positive slope) was tested using a corresponding t-test. Phenotypic dissimilarity would decrease with genomic relationship under the alternative hypothesis.

Results are illustrated in Figure 5.3: While the hypothesis could be rejected for starvation resistance and startle response, this was not the case for chill coma recovery. Here, the estimated slope was positive for sex-averaged or male traits, and slightly negative when using female records only.

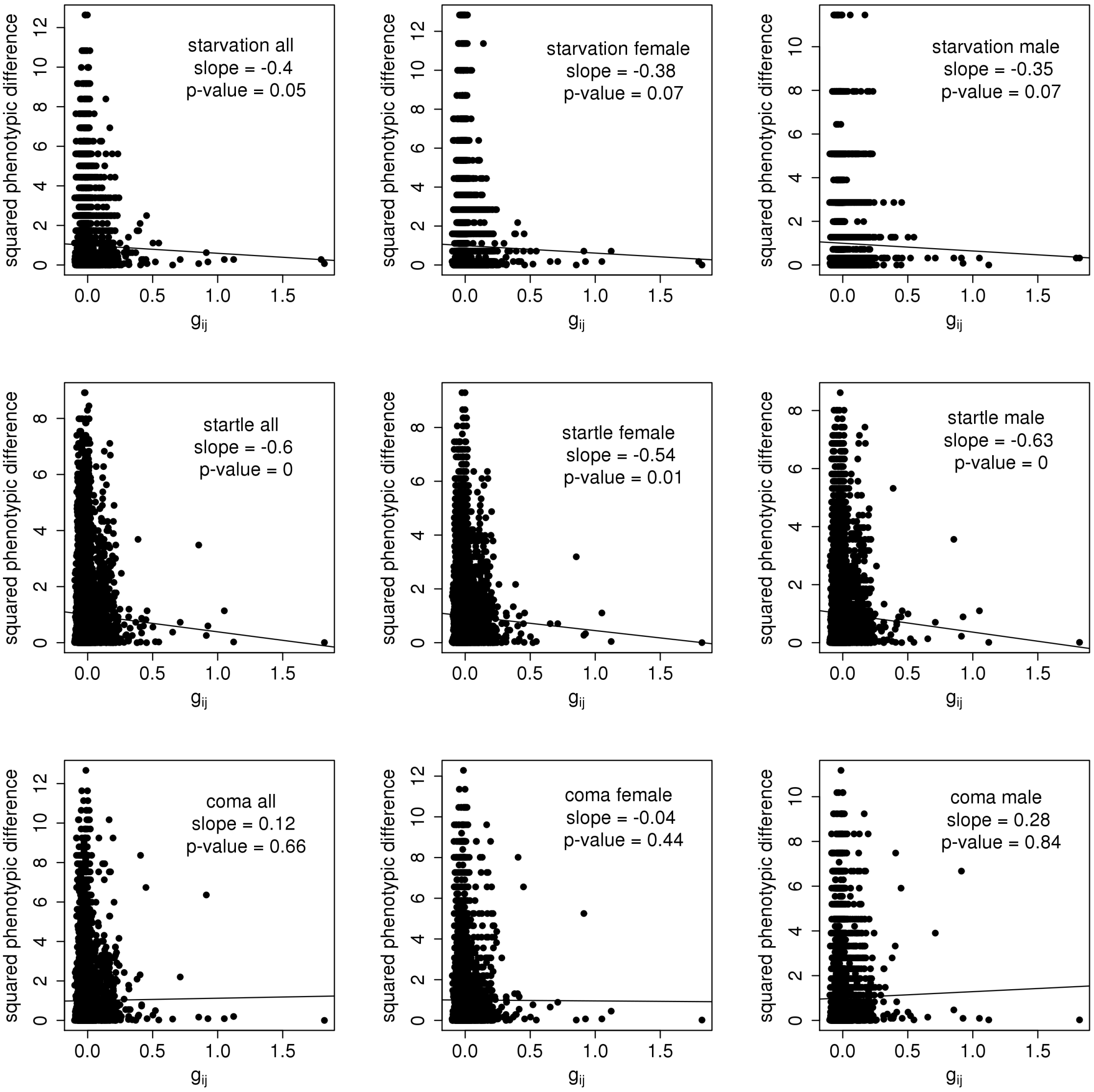

Figure 5.3: Plot of the standardized squared phenotypic differences for different traits against the genomic relationship coefficients g_{ij}. The genomic relationship coefficients were calculated according to VanRaden (2008). A regression line was fitted and the one-sided null hypothesis $b > 0$ (positive slope) was tested. From left to right: average medians of male and female records, medians of female records, medians of male records were used as phenotypic values. From top to bottom: starvation resistance, startle response, chill coma recovery.

5.3.2 Empirical variogram – evidence of two clusters of lines

Another way to reflect this dependency is the so-called "empirical variogram", a geostatistical tool usually applied in the context of the analysis of stochastic random fields (*cf.* Wackernagel (2003) and chapter 3, in which we considered the variogram corresponding to the genomic relationship matrix of VanRaden (2008) in a limiting case). To obtain the empirical variogram, the average squared differences between the standardized phenotypic values were plotted against the average Euclidean difference of the SNP vectors for pairs of lines falling into different bins of (Euclidean) distances. We chose 20 distance bins such that each bin contained the same number of pairs of lines.

The empirical variograms for all three traits are displayed in Figure 5.4. While the empirical variograms for starvation resistance and startle response show a monotone increasing trend, this trend cannot be observed for chill coma recovery, and especially the points belonging to the last 6 bins of distances (marked in red in Figure 5.4) are conspicuous.

Schlather & Tawn (2003) used a similar tool (the extremal coefficient cloud) in an exploratory analysis of daily rainfall data, and by identifying outliers they were able to reveal inconsistencies in their underlying data set.

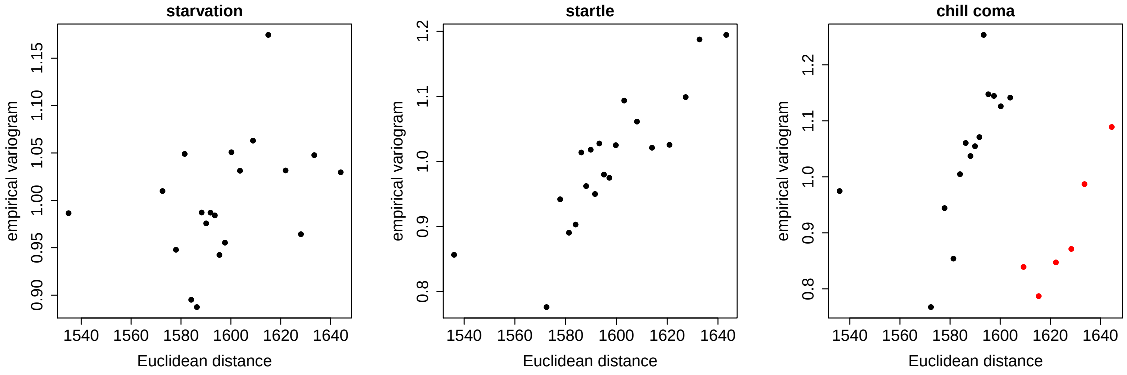

Figure 5.4: Empirical variogram of the standardized phenotypic values of different traits. From left to right: starvation resistance, startle response, chill coma recovery. On the x-axis: Euclidean distance between SNP vectors. On the y-axis: average squared difference of the standardized phenotypic values (using average medians of male and female records) for pairs of lines lying in the corresponding bin of Euclidean distance between SNP vectors. The distance bins were chosen such that 20 bins containing the same number of points were created, for which the average squared difference of the standardized phenotypic values was calculated.

To investigate the untypical form of the empirical variogram for chill coma recovery, we plotted a histogram for each distance bin showing how often the different lines were contributing to its variogram-point (results not shown). We identified 25 lines that were extraordinarily frequently contributing to the variogram-points of the last 6 distance bins (compared to the other lines). Based on this, the 147 lines having phenotypic records for chill coma recovery could be divided into two clusters consisting of 122 and 25 lines. In the following, we will denote these two clusters by "C_1" and "C_2". The IDs of lines belonging to cluster C_2 are listed in Suppl. Table S4. Most lines of cluster C_2 also belong to one of the two blocks of higher relationship displayed in the heatmap (Suppl. Figure S3). Excluding cluster C_2 from the variogram calculations finally led to a more typical (monotonically increasing) empirical variogram (Figure 5.5).

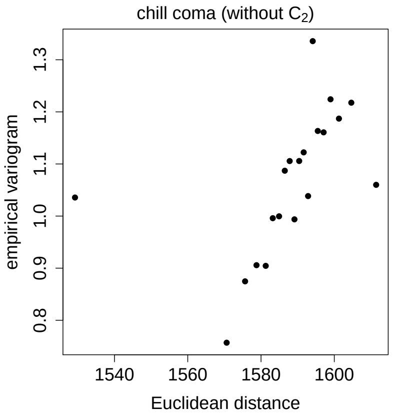

Figure 5.5: Variogram of the standardized phenotypic values of chill coma recovery after excluding lines of cluster C_2.

5.3.3 Indices of fixation (F_{ST}-values)

To further analyze the clusters C_1 and C_2 obtained in the previous section, we considered a measure of population differentiation, the so-called "fixation index", which was developed in the 1920s by Sewall Wright (*cf.* Wright (1984)) and which measures the diversity of randomly chosen alleles within the same subpopulation relative to that found in the entire population.

Following Gianola *et al.* (2010), the fixation index (also called F_{ST}-value) of the ℓ-th locus can be obtained as

$$F_{\mathrm{ST},\ell} = \frac{p^2_{C_1,\ell} + p^2_{C_2,\ell} - 0.5 \cdot (p_{C_1,\ell} + p_{C_2,\ell})^2}{p_{C_1,\ell} + p_{C_2,\ell} - 0.5 \cdot (p_{C_1,\ell} + p_{C_2,\ell})^2},$$

where $p_{C_1,\ell}$ and $p_{C_2,\ell}$ are the allele frequencies of the second allele at locus ℓ for the clusters C_1 and C_2, respectively. Its values range from 0 to 1, the latter implying that the two considered populations are completely separate.

Using C_1 and C_2 obtained from the variogram analyses, we calculated the average fixation indices for sliding windows of 100 adjacent loci, displaced by 10 SNP positions, across the whole genome. The results for the five chromosomes are displayed in Figure 5.6.

Interestingly, there is a region on chromosome *2L* of approximate length of 15,000,000bp showing extraordinarily high F_{ST}-values in the range of [0.1,0.2]. In Figure 5.6, we also indicated the significant SNP positions revealed in the GWAS by Mackay *et al.* (2012) for starvation resistance, startle response and chill coma recovery. For chill coma recovery, we checked whether the region with high F_{ST}-values included extraordinarily many significant SNP positions revealed in the GWAS, but this was not the case. However, it cannot be ruled out at this stage of research that this region might have a special effect on the trait.

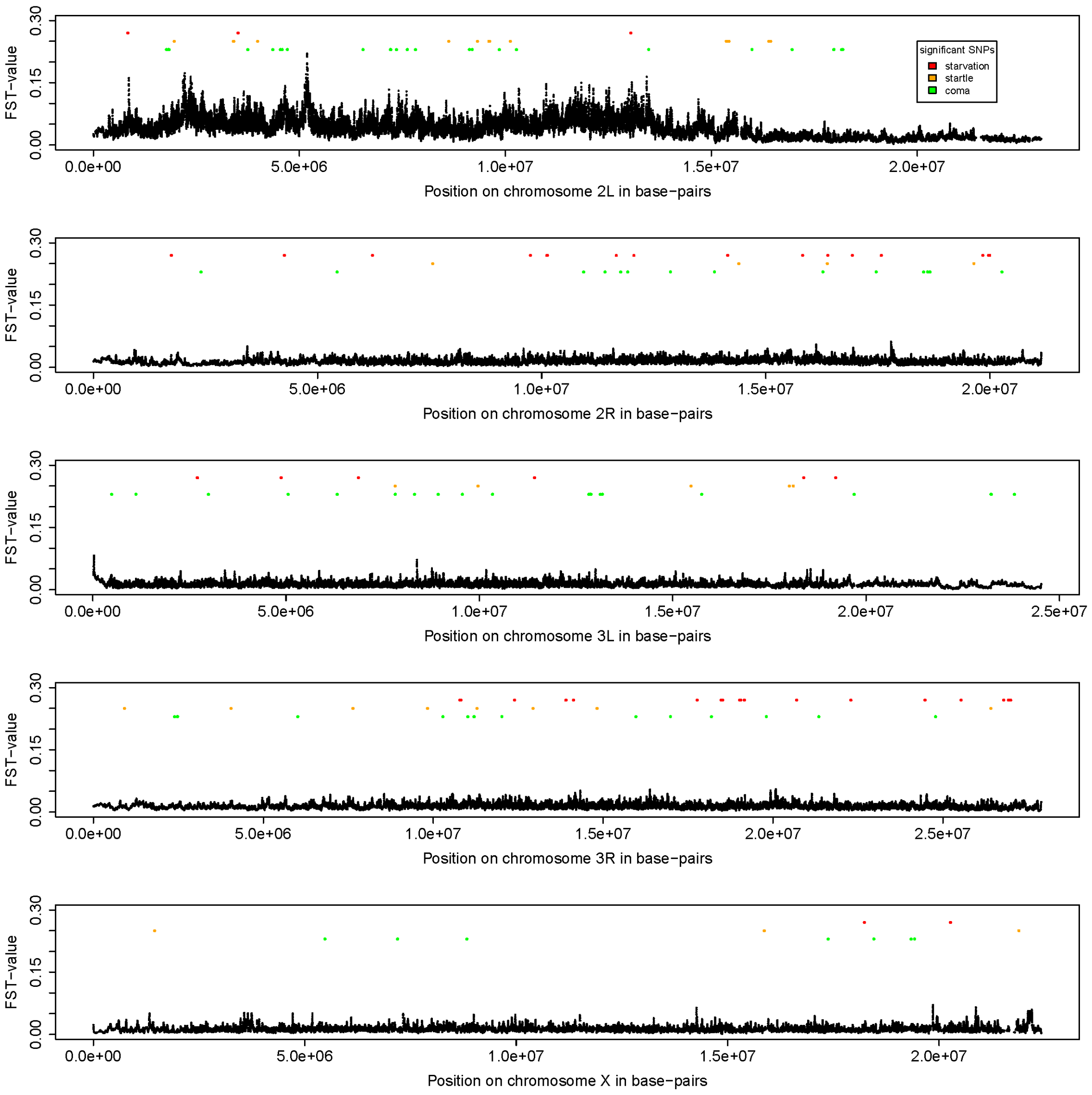

Figure 5.6: Average F_{ST}-values for sliding windows of 100 adjacent loci between the two clusters C_1 and C_2 of lines. Significant SNP positions (p-value $< 10^{-5}$) according to the GWAS of Mackay *et al.* (2012) are indicated for starvation resistance (red), startle response (orange) and chill coma recovery (green).

5.3.4 Conjecture: presence of the “Nova Scotia” inversion

Sturtevant (1931) was the first reporting the so-called “Nova Scotia”-inversion in *D. melanogaster*, denoted by “In(2L)NS”, which is a rare cosmopolitan inversion lying on chromosome *2L*, whose computed breakpoints include $23E2 - 23E3$ and $35F1 - 35F2$ (see `www.flybase.org`). It might be that the 25 lines of C_2 (or the 122 lines of C_1, respectively) are carriers of this inversion, which could be an explanation for the extraordinarily high F_{ST}-values of the two clusters in this region.

5.3.5 Cross-validations for different clusters

As a first analysis of whether the large region on chromosome *2L* has an effect on the trait, we ran several types of 5-fold CVs (20 replicates each) using GBLUP and different sets of SNPs and lines. The set of SNPs used to construct the genomic relationship matrix **G** according to VanRaden (2008) could be:

- all 2.5 million SNPs (“all SNPs”)
- the 2.5 million SNPs without the SNPs in the region with high F_{ST}-values on chromosome *2L* (“SNPs without *2L*-region”)
- or the SNPs in the region with high F_{ST}-values on chromosome *2L* (“SNPs of *2L*-region only”).

The set of lines used in the CV could be:

- all 147 lines (“all lines”)
- the 122 lines belonging to C_1
- or the 22 lines belonging to C_2.

Hence, there were in total 9 possible different scenarios, each of which was investigated using the average median of male and female records (“all”), the median of male records only (“male”) or the median of female records only (“female”) for each line as phenotypic value.

The predictive abilities, the estimated variance components and the corresponding heritabilities for the different CV procedures are shown in Table 5.4. Predictive abilities increased, when only the 25 lines of C_2 were considered, with another increase in predictive ability when SNPs from the *2L*-region were excluded. It is also noticeable, that predictive abilities are especially poor when all lines but no SNPs from the *2L*-region are used.

The relatively high predictive abilities using the 25 lines of C_2 could be an artifact due to the small sample size. Consider *e.g.* two standard normally distributed uncorrelated random variables 0. Then, the probability to obtain an empirical correlation greater or equal to 0.25 when drawing 25 realizations of this bivariate distribution is approximately 0.11, illustrating that the high predictive abilities achieved in the study should not be overvalued.

In summary, these analyses did not uncover the causations of the low accuracies obtained with the GBLUP approach applied to the chill coma recovery data entirely, and it remains in large parts unclear, which effect the clustering based on the *2L*-region has on this trait.

Table 5.4: Results of (4:1)-CVs of GBLUP for chill coma recovery, using different sets of lines and SNPs: average correlations, estimated variance components and corresponding heritabilities. Each scenario was investigated using the average median of male and female records (“all”), the median of male records only (“male”) or the median of female records only (“female”) for each line as phenotypic value. Variance components were estimated using “fitvario” from the R-package “RandomFields”, version 2.0.46.

				set of lines	
phenotypes	SNP set		all lines	C_1 (122 lines)	C_2 (25 lines)
all	all SNPs	cor[1]	−0.036 (0.009)	−0.140 (0.011)	0.248 (0.041)
		$\hat{\sigma}_g^2$	0	0	7.32
		$\hat{\sigma}_e^2$	22.57	23.98	0
		$\hat{h}^2$	0	0	1
	SNPs without $2L$-region	cor	−0.092 (0.010)	−0.169 (0.012)	0.331 (0.027)
		$\hat{\sigma}_g^2$	0	0	7.07
		$\hat{\sigma}_e^2$	22.58	23.98	0
		$\hat{h}^2$	0	0	1
	SNPs of $2L$-region only	cor	0.064 (0.007)	0.025 (0.017)	0.087 (0.046)
		$\hat{\sigma}_g^2$	0.42	0.55	0
		$\hat{\sigma}_e^2$	21.72	22.93	15.30
		$\hat{h}^2$	0.02	0.03	0
female	all SNPs	cor	0.048 (0.008)	−0.008 (0.013)	0.311 (0.037)
		$\hat{\sigma}_g^2$	2.24	0	9.06
		$\hat{\sigma}_e^2$	22.84	28.95	0
		$\hat{h}^2$	0.09	0	1
	SNPs without $2L$-region	cor	0.020 (0.010)	−0.019 (0.014)	0.368 (0.028)
		$\hat{\sigma}_g^2$	1.13	0	8.74
		$\hat{\sigma}_e^2$	24.72	28.95	0
		$\hat{h}^2$	0.05	0	1
	SNPs of $2L$-region only	cor	0.094 (0.005)	0.041 (0.013)	0.178 (0.042)
		$\hat{\sigma}_g^2$	0.91	1.78	0
		$\hat{\sigma}_e^2$	25.50	25.54	18.99
		$\hat{h}^2$	0.03	0.07	0
male	all SNPs	cor	−0.147 (0.011)	−0.226 (0.012)	0.164 (0.042)
		$\hat{\sigma}_g^2$	0	0	6.91
		$\hat{\sigma}_e^2$	21.50	22.94	0
		$\hat{h}^2$	0	0	1
	SNPs without $2L$-region	cor	−0.208 (0.012)	−0.268 (0.011)	0.272 (0.027)
		$\hat{\sigma}_g^2$	0	0	6.70
		$\hat{\sigma}_e^2$	21.50	22.94	0
		$\hat{h}^2$	0	0	1
	SNPs of $2L$-region only	cor	0.014 (0.009)	0.018 (0.020)	−0.051 (0.050)
		$\hat{\sigma}_g^2$	0	0	0
		$\hat{\sigma}_e^2$	21.50	22.47	14.08
		$\hat{h}^2$	0	0.01	0

[1] Average correlation between predicted genetic values and observed phenotypes. Results are averages over 20 replicates. Standard errors of the means in parentheses.

5.4 Observation II: bimodal phenotypic distribution

In the last sections it is described how we identified two clusters of lines in relation to a large region on chromosome *2L*. Apart from that, we discovered that the phenotypic distribution of chill coma recovery is in fact a mixture of two normal distributions.

5.4.1 Bimodal distribution of phenotypes

Given the phenotypic values for the chill coma recovery trait, we applied the R-package "mixtools" (Young *et al.*, 2010) to analyze whether the corresponding distribution was a mixture of two or more normal distributions. We used the function "boot.comp" to chose the number of components of the mixture distribution. This function is based on a parametric bootstrap approach to test the number of components in the mixture model sequentially. We found that two components were the optimal choice for the chill coma recovery data. In a second step, the function "normalmixEM" was used to determine the parameters of the two distributions, which has implemented an Expectation Maximization (EM) algorithm which maximizes the conditional expected complete-data loglikelihood at each step.

Using the average of medians of male and female records for each line as phenotypic data, we found that the distribution is a mixture of the two normal distributions $\mathcal{N}(13.53, 1.87^2)$ and $\mathcal{N}(20.39, 4.77^2)$ with weights given by 0.597 and 0.403. Hence, the phenotypic distribution is indeed bimodal. The corresponding density functions are shown in Figure 5.7 (left plot).

The same scenario can be observed when only the male (female) medians of records are used as phenotypic values (middle (right) plot of Figure 5.7).

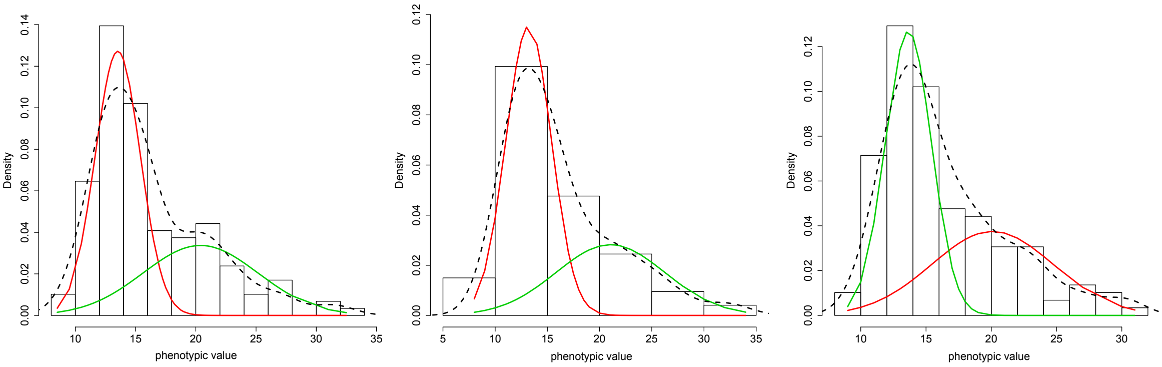

Figure 5.7: Bimodality of the chill coma recovery phenotypes. Histograms of the phenotypic distribution for chill coma recovery are displayed using the averages of the medians of male and female records (left plot), the medians of female records (middle plot) or the medians of male records (right plot) as phenotypic values. The red and green lines are the density functions of the two components of the mixture distribution. The dashed black line is the density function of the mixture distribution.

Given the two components of the mixture distributions, the lines could be allocated to one of these two components based on their posterior probabilities to belong to the distributions, which were given as part of the output of "normalmixEM". The posterior probabilities for each line are listed in Suppl. Table S5. Clustering the lines based on their posterior probabilities finally led to two different populations denoted by "Pop1" and "Pop2" (consisting of 99 and 48 lines). All lines with a posterior probability greater than 0.5 to belong to the second distribution were assigned to Pop2.

5.4.2 First analyses using Pop1 and Pop2

In a first step, we redid the 5-fold CVs using GBLUP separately for Pop1 and Pop2. The results are summarized in Table 5.5. Performing a 5-fold CV using GBLUP and only female records of Pop1 led to a moderate predictive ability of 0.288 (0.014), indicating that the lack of accuracy of GBLUP (using all lines and sex pooled records) might stem from a complex structure of population-sex-interactions underlying the trait, possibly in combination with epistasis as indicated in previous sections. This suggests the hypothesis that the genomic relationship matrix **G** is not able to model this complexity adequately.

Table 5.5: Average correlations between predicted genetic values and observed phenotypes for different CV procedures with GBLUP using chill coma recovery data.

type of CV	correlation
(4:1)-CV all[1] Pop1[2]	0.127[3] (0.014)
(4:1)-CV all Pop2	−0.375 (0.024)
(4:1)-CV female – female[4] Pop1	0.228 (0.014)
(4:1)-CV female – female Pop2	−0.338 (0.025)
(4:1)-CV male – male Pop1	−0.047 (0.017)
(4:1)-CV male – male Pop2	−0.181 (0.030)

[1] The average of the medians of male and female measurements was used to predict line phenotypes. Predicted phenotypes were then correlated with the averages of the medians of male and female measurements.

[2] "Pop1" means: Only lines of Pop1 were used in the estimation and in the validation set.

[3] Average correlation between predicted genetic values and observed phenotypes. Results are averages over 20 replicates. Standard errors of the means in parentheses.

[4] "CV sex_1 – sex_2" means: Medians of measurements of sex_1 were used in the training set, medians of sex_2 were used in the validation set.

We also calculated the F_{ST}-values for Pop1 and Pop2, but there were no regions with higher F_{ST}-values like in the case of C_1 and C_2 (results not shown). There is also no extraordinary overlap between the clusters C_1 and C_2 and the clusters Pop1 and Pop2, as one might have suspected: In total, $\frac{48}{147} \approx 33\%$ of the lines belong to Pop2. Assuming that there is no connection between C_1/C_2 and Pop1/Pop2, we would expect that $33\% \cdot 25 = 8$ lines out of the 25 lines from C_2 belong to Pop2. Indeed, there are 8 lines from C_2 belonging to Pop2 (*cf.* Suppl. Tables S4 and S5). Thus, there is no hint for a connection between C_1/C_2 and Pop1/Pop2.

5.5 Epistatic interactions

Due to the results of section 5.2.3, we suspected that the chill coma recovery trait is driven by many epistatic interactions between pairs of SNPs. To test for possible epistatic interactions, we developed the following approach, which also accounts for the Pop1/Pop2-clustering.

5.5.1 Candidate list of SNP positions

Starting point was the set of *individual* trait records, which we averaged over the replicates (for each line/sex combination). To get a candidate list of SNP positions that could possibly be involved in epistatic interactions, we ran a linear mixed model using SAS software (SAS Institute, 2002-2008) separately for each SNP position, applying the following `proc GLM`-procedure and using the averaged individual trait data as response variable `coma`:

```
proc GLM data=dat;
by position;
class line sex pop SNP;
model coma = sex sex*pop sex*SNP sex*pop*SNP
  line(pop*SNP) pop SNP SNP*pop;
random line(pop*SNP) / test;
ods output RandomModelANOVA = result_GWAS;
run;
```

In this model, `pop` was a dummy variable with values 1 and 2 indicating whether a line belonged to Pop1 or Pop2. The terms `SNP`, `sex` and `line` were categorical variables describing the SNP value (coded with 0 and 2), the sex and the line-ID, all treated as fixed effects. The nested interaction term `line(pop*SNP)` was treated as a random effect. The several components of this linear mixed model were tested for significance (type III sum of squares). Based on the results of these genome-wide ANOVA-tests, we looked for SNP positions

- with significant `pop`- and `pop*SNP`-interactions:
  There were 1,508 SNP positions with a p-value $< 1 \cdot 10^{-5}$ for both terms.

- with significant `pop`- and `pop*SNP*sex`-interactions:
  There were 6,555 SNP positions with a p-value $< 1 \cdot 10^{-5}$ for both terms.

- with significant differences in allele frequencies between two populations Pop1 and Pop2, based on Fisher's Exact Test (Fisher, 1922):
  There were 521 SNP positions for which the difference was significant (p-value $< 5 \cdot 10^{-6}$).

These SNP positions built a candidate list which was used in further analyses.

5.5.2 The "Flyland" data set

The third category of candidate SNPs was amended by several SNP positions identified using the so-called "Flyland" data set, which was produced in the laboratory of Prof. Mackay.

The Flyland data set consists of a synthetic outbred, intercross population based on a subset of 40 DGRP lines, which were crossed in a "round robin" manner, followed by 70 generations of random mating. After doing different assays to assess phenotypes of 2,000 flies for various traits, a pooled DNA sequencing of flies belonging to the top and bottom 15% of the phenotypic distribution curve for a given trait was performed in a second step.

From the 40 DGRP lines forming the basis of the Flyland data, 23 (6) lines belonged to Pop1 (Pop2). Using the corresponding allele frequencies from the pooled sequencing in the two subsets forming the top and bottom 15% of the phenotypic distribution curve for chill coma recovery, we looked for SNP positions with significant differences between these two pools (using again Fisher's Exact Test, p-value $< 5 \cdot 10^{-4}$). This finally resulted in 170 additional SNP positions, which were added to the above candidate list of SNP positions.

5.5.3 Tests for significant pairwise interactions

The candidate list finally consisted of 8,750 SNP positions in total. In a next step, all possible pairs of SNPs from the candidate list were tested for significant pairwise interactions, using SAS software (SAS Institute, 2002-2008) and applying the following `proc GLM`-procedure, using again the average of the medians of male and female measurements as response variable in a linear model (as already done in the GBLUP approach). The linear model included the fixed categorical variables `SNP_A` and `SNP_B` for the genotypes of the two considered SNP positions as well as the corresponding fixed interaction term `SNP_A*SNP_B`:

```
proc GLM data=dat;
by position;
class line SNP_A SNP_B;
model av_median = SNP_A SNP_B SNP_A*SNP_B ;
ods output ModelANOVA = result_Epistasis;
run;
```

The interaction terms `SNP_A*SNP_B` in the ANOVA-models were tested for significance (type III sum of squares). As expected, we found many significant epistatic interactions, confirming our hypothesis that there are complex structures underlying the trait, which involve both the population structure and epistasis. In total, we found

15 (15; 46; 184; 897; 4,508; 18,856; 52,353)

significant pairwise interactions using a significance threshold for the p-value of

$$10^{-12}(10^{-11};\ 10^{-10};\ 10^{-9};\ 10^{-8};\ 10^{-7};\ 10^{-6};\ 10^{-5}).$$

The 897 significant interactions corresponding to a threshold of 10^{-8} are exemplarily displayed in Figure 5.8. To obtain this network representation, all positions showing a significant interaction with at least one other position were placed in a circle as equally spaced nodes. Significant interactions were then drawn as edges between the nodes, using the R-package "diagram" (Soetaert, 2011) and its function "plotweb". Note that a lot of LD between neighboring SNP positions can be observed from this network plot.

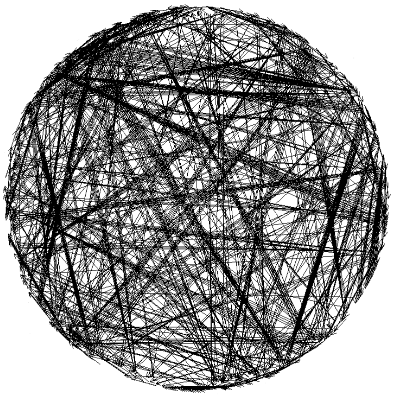

Figure 5.8: Network plot of the 897 significant epistatic interactions between SNPs (p-value 10^{-8}), which were found based on the candidate list of SNP positions.

5.6 Clustering and epistatic interactions as a basis for further investigations

The results on the chill coma recovery trait presented in the previous sections allow a first insight into the complex biological background of the trait and give rise to many possible research directions. Possible approaches include the further analysis of epistasis, *e.g.*, by assigning the epistatic SNP positions to biological relevant pathways. This seems to be the most promising course of action. As noted before, another potential route is the analysis of whether the lines of cluster C_1 (or C_2) carry the Nova Scotia inversion. The procedural method and the results of this chapter can therefore serve as a stepping stone for further studies, which may lead to a better understanding of the complex trait. Some of these analyses have already been carried out in collaboration with the working group of Prof. Mackay and the results are described in a joint manuscript which is currently in revision for *PLoS Genetics* (Ober *et al.*, 2012*b*). We will briefly report the results of Ober *et al.* (2012*b*) in the following section. Note that parts of the analyses presented in Ober *et al.* (2012*b*) to infer epistatic interactions for the chill coma trait differ only slightly from the analyses presented in the previous sections. Ober *et al.* (2012*b*) additionally perform gene ontology and network analyses based on the inferred interactions.

5.6.1 Results of Ober *et al.* (2012*b*)

In summary, Ober *et al.* (2012*b*) found an unexpectedly complex genetic architecture underlying the chill coma recovery trait, comprising a few alleles with large additive effects as well as complex networks of epistatic interactions, leading to failure of genomic prediction and misleading results from single marker genome-wide association studies. The methodological strategy and the corresponding results of Ober *et al.* (2012*b*) are displayed schematically in Figure 5.9.

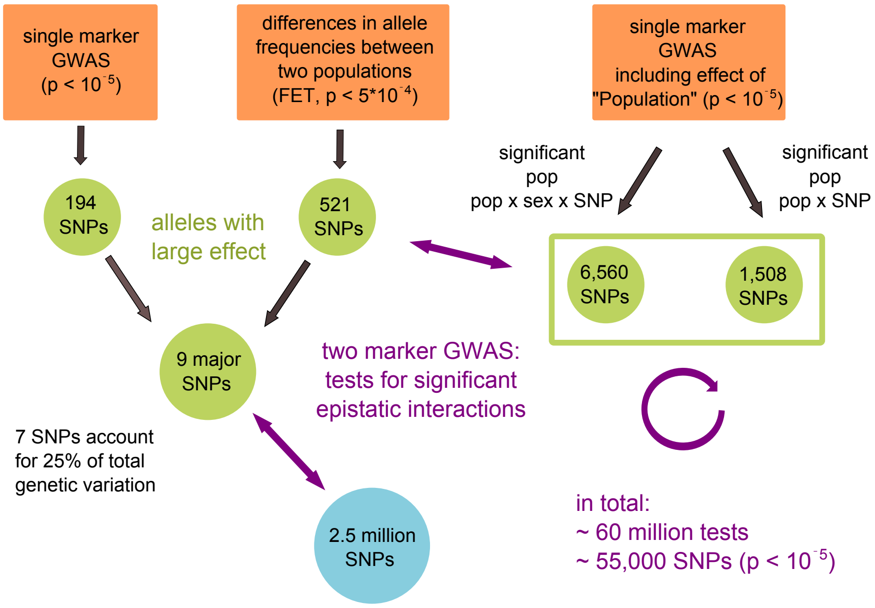
single marker GWAS ($p < 10^{-5}$)
differences in allele frequencies between two populations (FET, $p < 5*10^{-4}$)
single marker GWAS including effect of "Population" ($p < 10^{-5}$)
significant pop pop x sex x SNP
significant pop pop x SNP
194 SNPs
alleles with large effect
521 SNPs
6,560 SNPs
1,508 SNPs
9 major SNPs
two marker GWAS: tests for significant epistatic interactions
7 SNPs account for 25% of total genetic variation
2.5 million SNPs
in total:
~ 60 million tests
~ 55,000 SNPs ($p < 10^{-5}$)

Figure 5.9: Schematic illustration of the GWAS approaches used in Ober *et al.* (2012*b*).

As shown in section 5.4.1, the DGRP lines partition into two populations with different means. Using this result, Ober *et al.* (2012*b*) further illustrate that these two populations exhibit patterns of sexual dimorphism as well as different magnitudes of genetic and environmental variance. Based on the hypothesis that the difference in mean between the two populations is in part due to variants with large additive effects on the trait, Ober *et al.* (2012*b*) looked for significant SNPs from a single-marker GWAS and for SNPs with significant differences in allele frequencies between the two populations, revealing 6–7 SNPs which contributed to 22–26% of the total genetic variance, depending on the analysis, and in contrast to the GBLUP predictions in which the additive genetic variance was estimated to be zero. It was further hypothesized that the remaining genetic variance was due to epistasis and different genetic architectures in the two populations. Indeed, when performing a GWAS for chill coma recovery time in line with the analysis in section 5.5 and including the main effect of `population` and its interactions with the fixed factor variables `SNP` and `sex`, Ober *et al.* (2012*b*) identified 8,086 (2,453) SNPs for which the effect of `population` was highly significant and for which one of the interaction terms `population*SNP` or `population*sex*SNP` was significant at a nominal p-value less than 10^{-5} (10^{-6}). Based on these findings, Ober *et al.* (2012*b*) sought to identify

- pairs of significantly interacting SNPs among these variants
- significant interactions between the 8,086 SNPs and the SNPs showing significant differences in allele frequencies between the two populations
- and significant interactions between the ≈ 2.5 million SNPs and the SNPs with large effects.

The corresponding two-marker GWAS (including almost 60 million linear two-marker models) finally revealed $\approx 55{,}000$ significant interactions (p-value less than 10^{-5}).

After removal of pairs of SNPs in perfect LD and concentrating only on significant interactions with p-values less than 10^{-7}, 2,515 interactions involving 961 SNPs within 483 annotated genes remained. These genes were enriched for Gene Ontology terms linked to signaling and metabolic pathways, and comprised a tightly woven genetic interaction network. Ober *et al.* (2012*b*) further found that the discovered genes affecting the time to recover from chill coma were involved in nervous system development and signaling, which is biologically plausible. Besides, it is intriguing that many intergenic SNPs far away from annotated genes participate in numerous interactions, potentially defining novel regulatory regions.

5.7 Discussion

5.7.1 Evidence of epistasis in the literature

Epistasis is known to be an important component of the genetic architecture of quantitative traits, as reviewed *e.g.* by Flint & Mackay (2009), Mackay *et al.* (2009) and Anholt (2010). According to Mackay *et al.* (2009), epistasis “refers to the masking of genotypic effects at one locus by genotypes of another locus (Phillips, 2008) and also to any statistical interaction between genotypes at two or more loci (Falconer & Mackay, 1996; Lynch & Walsh, 1998)”.

Various studies have already reported substantial evidence of epistatic interactions among QTLs affecting quantitative traits in different species like *Drosophila* and mice (Flint & Mackay, 2009), chickens (Carlborg *et al.*, 2006), *Arabidopsis* (Kroymann & Mitchell-Olds, 2005) and yeast (Sinha *et al.*, 2008; Steinmetz *et al.*, 2002). As outlined by Mackay *et al.* (2009) and Swarup *et al.* (2012), epistatic interactions in *Drosophila* have been documented to affect metabolic activity as well as olfactory and locomotor behaviors, and epistatic interactions have been found for traits measuring bristle number, wing shape and longevity. This diversity of results stresses the importance of epistasis with respect to the genetic architecture of quantitative traits, and it also highlights the need for prediction models allowing to account for epistasis.

One challenging aspect in this regard is the fact that the existing forms of epistasis are manifold. Mackay *et al.* (2009) explain: "Epistatic effects can be as large as main QTL effects, and can occur in opposite directions between different pairs of interacting loci and between loci without significant main effects on the trait. Epistasis can also occur between closely linked QTLs (Kroymann & Mitchell-Olds, 2005; Steinmetz *et al.*, 2002; Sinha *et al.*, 2008) and even between polymorphisms at a single locus (Stam & Laurie, 1996)."

Ober *et al.* (2012*b*) further state that "if epistasis is present but not accounted for, estimates of allelic effects from association studies will be biased (Carlborg *et al.*, 2006; Phillips, 2008), potentially accounting for missing heritability (Zuk *et al.*, 2012) and leading to a failure of genomic prediction. On the other hand, knowledge of epistatic interactions can be used to infer genetic networks affecting complex traits (Phillips, 2008)."

5.7.2 The animal breeder's point of view

The lack of accuracy of prediction for chill coma recovery using GBLUP is bewildering from an animal breeder's point of view: While 38% to 39% of the phenotypic variation based on individual measurements are due to line differences (Suppl. Table S3), only a small proportion of the variance (0% to 3%) can be assigned to additive genetic causes, and adding an additive × additive component does not improve the model at all based on the loglikelihood.

Different predictive abilities for different traits have also been obtained in practical applications of genomic prediction to agriculturally relevant species. In an application to genomically predict testcross performance across families in maize (Albrecht *et al.*, 2011), predictive ability for grain dry matter yield was 0.48 while it was 0.64 for grain dry matter content, although both traits have very similar heritabilities. In Hayes *et al.* (2010), such differences are explained with differences in the architecture (mainly the number of QTL with very large effects) of the quantitative traits studied.

Based on the bimodality of the distribution of phenotypic values, which we could demonstrate for the chill coma trait, we retrospectively know that the underlying assumption of the GBLUP model – the normal distribution of the phenotypic values, all having approximately the same phenotypic mean and variance – is violated in chill coma recovery records of the DGRP data set. Hence, it is not astonishing that the GBLUP approach did not yield high predictive abilities. More importantly, the fact that we found many epistatic interactions between SNPs in this study and in Ober *et al.* (2012*b*) provides a reasonable explanation for the failure of the GBLUP approach, since the covariance matrix $\mathbf{G}$ used in the GBLUP approach is based on the additive relationship between lines and does not account for possible epistatic interactions between SNPs.

5.7.3 Conclusion

In summary, our analyses and the study in Ober *et al.* (2012*b*) demonstrate that epistasis can be of great importance for the specification of a quantitative trait and that it can even lead to a failure of prediction approaches. Additionally, "epistasis poses considerable statistical challenges like estimating the contribution of entire networks of interacting loci to genetic variance, predicting network responses to natural or artificial selection or incorporating such networks into genomic prediction models”, as stated in Ober *et al.* (2012*b*).

6 The Expected Linkage Disequilibrium in Finite Populations Revisited

In chapter 4, we applied a formula proposed by Sved (1971) for the expected linkage disequilibrium (LD) at equilibrium to estimate the effective population size for the DGRP data. In section 4.5, we further indicated that the exact validity of this formula is questionable from a mathematical point of view and that further research is necessary to investigate the formula and its derivations more detailed. We will deal with this subject comprehensively in the subsequent sections.

6.1 Introduction

In genetics research, the decay of LD as a function of the distance of the considered loci is an important characteristic of a population. One measure of LD between two loci which has widely been used in the literature is r^2 (*cf.* Hill & Weir (1994)), which depends on the frequencies of gametes in the considered population.

Moreover, it is commonly assumed that a finite population of size N with constant recombination rate c achieves a state of "equilibrium" after a certain time. Usually, this state of equilibrium is said to be reached when the expected amount of LD does not change from one generation to the next.

The effective population size N_e, which is defined as the size of an ideal population at equilibrium with the same structure of LD as the population under consideration (*cf.* Hedrick (2011)), is an important population parameter when considering how real populations evolved over time. In practice, N_e cannot be measured but LD can. Hence, efforts have been made to link the two quantities by formulae of the form $\mathbb{E}(r^2) \approx f(c, N_e)$, with a function f depending on c and N_e.

6.1.1 Sved's formula for the expected linkage disequilibrium (Sved, 1971)

The following formula for the expected LD at equilibrium in a population was proposed by Sved (1971) and has been used extensively to estimate N_e:

$$\mathbb{E}(r^2) = \frac{1}{1 + 4N_e c} \qquad \text{(Sved's formula)} \tag{6.1}$$

The equality can be written as

$$N_e = \frac{\frac{1}{\mathbb{E}(r^2)} - 1}{4c} = \frac{1 - \mathbb{E}(r^2)}{4c\mathbb{E}(r^2)},$$

and by using an empirically estimated $\mathbb{E}(r^2)$, the effective population size N_e can be calculated, as done in chapter 4 for the DGRP lines. Then, the estimated N_e corresponds to an effective population size $\frac{1}{2c}$ generations ago (Hayes *et al.*, 2003). In the following, we will refer to formula (6.1) as "Sved's formula". Sved (1971) derived this formula based on the following recursion formula for the conditional probability Q_T of identity by descent (IBD) at the second locus, given that two sampled gametes from the population are IBD at the first locus in generation T:

$$Q_T = \left(1 - \frac{1}{2N}\right)(1-c)^2 Q_{T-1} + \frac{1}{2N}(1-c)^2 \qquad \text{(Sved's recursion formula)} \qquad (6.2)$$

Note that this recursion formula is of linear form $Q_T = aQ_{T-1} + b$ with constants a and b. Sved claims that $Q_T = \mathbb{E}(r_T^2)$, where r_T^2 is the LD after T generations. Additionally, equilibrium is considered to be the point in time for which $Q_{T+1} = Q_T$. Based on this definition, the equation $Q_T = \mathbb{E}(r_T^2)$ combined with eq. (6.2) yields approximately eq. (6.1) for small values of c and after replacing N with N_e.

Sved's formula has been used in different areas of research and applications, ranging from animal breeding (Meuwissen *et al.*, 2001; de Roos *et al.*, 2008; Flury *et al.*, 2010; Qanbari *et al.*, 2010) and plant breeding (Remington *et al.*, 2001) to human genetics (Tenesa *et al.*, 2007; McEvoy *et al.*, 2011), and it has become one of the standard approaches for N_e-estimation.

6.1.2 Mathematical shortcomings of previous derivations

Several other derivations of the formula have been suggested in the last forty years (Sved & Feldmann, 1973; Tenesa *et al.*, 2007; Sved, 2008, 2009). We found that all derivations are in some parts of heuristic nature, including mathematical gaps or unsound conclusions, as already indicated in section 4.5. Indeed, concerns over the validity of the formula and their derivations have already been raised by Sved (*cf.* Sved (2008), p. 185, and a manuscript published on Sved's personal homepage `http://www.handsongenetics.com/PIFFLE/LinkageDisequilibrium.pdf`). In the following, we will sketch some of the mathematical concerns unfolding in these derivations.

Derivations of Sved (Sved, 1971; Sved & Feldmann, 1973; Sved, 2008, 2009)

In the manuscript mentioned above Sved reports a misunderstanding in the original derivation (Sved, 1971), in which eq. (6.2) is derived, stating that the recursion formula (6.2) should have been $Q_T = \left(1 - \frac{1}{N}\right)(1-c)^2 Q_{T-1} + \frac{1}{N}(1-c)^2$. But this would not lead to Sved's formula at equilibrium. It is further argued that a second misunderstanding seems to cancel out the first one leading to eq. (6.2) again, but some uncertainty about the correctness of the equations remains, as stated by Sved in the manuscript mentioned above.

A second key step in this derivation is the equation $Q_T = \mathbb{E}(r_T^2)$ which finally leads to eq. (6.1) at equilibrium. To justify this equation, the following argumentation is used: Imagine, a gamete is sampled at random from the population. A second gamete with the same genotype at the first locus is sampled afterwards. The genes at the first locus are said to be identical by descent (IBD) per definition. Then, Sved uses the formula $p_{B_1}^2 + p_{B_2}^2$ for the probability of homozygosity at the second locus, where p_{B_1} and p_{B_2} are the corresponding allele frequencies. The expression $p_{B_1}^2 + p_{B_2}^2$ is the *unconditional* probability of homozygosity, not taking into account the homozygosity at the first locus in LD, while the *conditional* probability is expected to be greater than $p_{B_1}^2 + p_{B_2}^2$.

Sved & Feldmann (1973) rediscuss this approach and propose a modified recursion formula which is $Q_T = \left(1 - \frac{1}{2N}\right)(1-c)^2 Q_{T-1} + \frac{1}{2N}$, but the proof of $\mathbb{E}(r_T^2) = Q_T$ is still lacking.

Finally, another approach is presented in Sved (2008, 2009) by combining the concepts of correlation of two loci and probability of IBD. The critical point in these derivations is that correlations are assumed to be additive. However, this assumption is only verified for the one-locus case, and a proof for the required two-locus case is missing.

Derivation of Tenesa *et al.* (2007)

Tenesa *et al.* (2007) provide a shorter derivation of Sved's formula using the equation $\mathbb{E}(r_{t+1}) = (1-c)r_t$. Here, the left-hand side is a constant, whereas the right-hand side is a random variable. Additionally, $\mathrm{Var}(r) \approx \frac{(1-\mathbb{E}(r)^2)}{n}$ is used as a general expression for the sampling variance of an estimate of a correlation coefficient r with sample size n. In this context, it is not distinguished between the true underlying correlation ρ and the empirical correlation coefficient r. It is not stated either for which underlying distribution this formula can be applied. According to Hotelling (1953), $\mathrm{Var}(r) \approx \frac{(1-\rho^2)^2}{n}$ holds for a bivariate normal distribution. Note that the numerator is squared, whereas this is not the case in the formula used by Tenesa *et al.* (2007). It is unclear, whether and how the formula used by Tenesa *et al.* (2007) is related to the result of Hotelling (1953), since in the case of LD the underlying distribution is bivariate Bernoulli, and approximation by a bivariate normal distribution is questionable in this case.

All points of critique mentioned so far stress the need for a clearer approach and an extensive empirical analysis of the existing formulae.

6.1.3 Organization of the chapter

The rest of this chapter is organized as follows: In section 6.2, we propose an alternative linear recursion formula for the expected LD in a finite population and analyze its validity in an extensive simulation study (section 6.3). The new formula is also compared to Sved's recursion formula, and the dependency of the precision of both formulae on the constellation of allele frequencies is analyzed.

In section 6.4, we consider the expected LD at equilibrium in the mathematical framework of the theory of discrete-time Markov chains. On the basis of a (linear) recursion formula, we derive a formula for the expected amount of LD at equilibrium, leading to a formula for the effective population size N_e. First, the derivation is given under the assumption that the recursion formula is exact. We then analyze how the non-exactness of a linear recursion formula affects the result for the expected LD at equilibrium.

In section 6.6, we estimate effective population sizes for the human HapMap data (The International HapMap Consortium, 2003) using records of two populations. To illustrate the impact of the allele frequency spectrum used, this is done for different sampling schemes based on minor allele frequencies.

We finally discuss the practical implications of our findings in section 6.7.

6.2 A new recursion formula

6.2.1 Basic principles and assumptions

Hill & Robertson (1968) proposed r^2 as a measure of LD between a pair of loci. With two biallelic loci A and B with alleles A_1, A_2, B_1, B_2 and frequencies $p_{A_1}, p_{A_2}, p_{B_1}, p_{B_2}$, we denote the frequencies of the genotypes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 by x_{11}, x_{12}, x_{21}, and x_{22}, respectively. Then,

$$r^2 = \frac{(x_{11}x_{22} - x_{12}x_{21})^2}{p_{A_1}p_{A_2}p_{B_1}p_{B_2}}. \tag{6.3}$$

Note that if we consider the allelic states at the two loci as Bernoulli variables with parameters p_{A_1} and p_{B_1}, then r^2 is the square of the correlation coefficient of these two random variables.

In the following, we consider a diploid population of finite size N at some arbitrary point $T = t_0$ in time and two biallelic loci A and B as described above, with gamete frequencies $\mathbf{x}_{t_0} := (x_{t_0,11}, x_{t_0,12}, x_{t_0,21}, x_{t_0,22})$. Assuming random mating and a constant recombination rate c, we can calculate the probabilities $\mathbf{x}'_{t_0} := (x'_{t_0,11}, x'_{t_0,12}, x'_{t_0,21}, x'_{t_0,22})$ for receiving the four different genotypes when producing an offspring gamete as

$$\begin{aligned} x'_{t_0,11} = x_{t_0,11} - cD_0, \quad x'_{t_0,12} = x_{t_0,12} + cD_0, \quad x'_{t_0,21} = x_{t_0,21} + cD_0 \\ \text{and} \quad x'_{t_0,22} = x_{t_0,22} - cD_0, \end{aligned} \tag{6.4}$$

with $D_0 := x_{t_0,11}x_{t_0,22} - x_{t_0,12}x_{t_0,21}$. For a detailed derivation we refer to Hedrick (2011), p. 528ff and the references therein. We are now interested in the expected squared correlation coefficient $\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})$ of the two random variables (the allelic states at the two loci) in $T = t_0 + 1$, given $\mathbf{x}_{t_0}$ (and hence $r^2_{t_0}$) from $T = t_0$. Since a constant population size is assumed, the population in $T = t_0 + 1$ is formed by $2N$ gametes, and the absolute frequencies of the four types of gametes $(n'_{11}, n'_{12}, n'_{21}, n'_{22}) := 2N\mathbf{x}_{t_0+1}$ follow a multinomial distribution with parameters $2N$ and $p = (x'_{t_0,11}, x'_{t_0,12}, x'_{t_0,21}, x'_{t_0,22})$.

6.2.2 Analytic expression for the expected LD in the next generation

Based on the above assumptions, the exact expected LD in $T = t_0 + 1$ conditional on $\mathbf{x}_{t_0}$ is given by:

$$\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1}) = \mathbb{E}_{\mathbf{x}_{t_0}}\left(\frac{(n'_{11}n'_{22} - n'_{12}n'_{21})^2}{(n'_{11}+n'_{12})(n'_{21}+n'_{22})(n'_{11}+n'_{21})(n'_{12}+n'_{22})}\right), \tag{6.5}$$

where $\mathbb{E}_{\mathbf{x}_{t_0}}$ denotes the expectation with respect to the multinomial distribution with parameters $2N$ and $p = \mathbf{x}'_{t_0}$ as described above. Analytical treatment of this expectation (*i.e.*, expressing it in terms of the probabilities $x'_{t_0,ij}$) does not seem to be feasible for general N. The open question is now how to deal with the complex formula. Even if one tried to approximate the expectation of the ratio by the ratio of expectations (*cf. e.g.* Ohta & Kimura (1971); Hill (1977)), the result would still depend on $\mathbf{x}_{t_0}$ in a very complex manner. Therefore, it is reasonable to work with an approximation of this expression, involving only $r^2_{t_0}$ on the right-hand side of eq. (6.5).

6.2.3 The alternative recursion formula for the LD

According to Sved's approach and based on the assumptions of the previous sections, we propose the following form of an approximate recursion formula for the expected LD in the population, given the gamete frequencies $\mathbf{x}_{t_0}$ in $T = t_0$:

$$\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1}) = ar^2_{t_0} + b = ar^2(\mathbf{x}_{t_0}) + b, \tag{6.6}$$

where a and b are functions of c and N. Note that $r^2_{t_0}$ is in fact a function of $\mathbf{x}_{t_0}$, which we indicate sometimes by writing $r^2(\mathbf{x}_{t_0})$. We further choose

$$a = (1-c)^2\left(1 - \frac{1}{2N}\right) \quad \text{and} \quad b = \frac{1}{2N-1-c}. \tag{6.7}$$

Note that this choice differs from Sved's recursion formula only in the value of b (*cf.* eq. (6.2)), what we will justify in the subsequent sections. The coefficients a and b were determined heuristically followed by a systematic validation.

6.3 Simulation study to analyze the performance of the new recursion formula

6.3.1 Simulation set-up

The general idea of the simulation study is the following: For a given combination $(N, c, \mathbf{x}_{t_0})$ in $T = t_0$, we randomly draw N_{sample} samples of $2N$ gametes according to the above multinomial distribution with parameters $2N$ and $p = \mathbf{x}'_{t_0}$. For each of these samples, $\mathbf{x}_{t_0+1}$ and the allele frequencies are obtained as empirical gamete and allele frequencies in $T = t_0 + 1$, and $r^2_{t_0+1}$ is calculated according to eq. (6.3). Then, $\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})$ is approximated by averaging over the N_{sample} values of $r^2_{t_0+1}$. Given all tuples $(N, c, r^2_{t_0}, \widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})})$, we can systematically analyze the fit of eq. (6.6) in combination with eq. (6.7), as described below.

The simulation was done for all combinations of N, c and $\mathbf{x}_{t_0}$, where

$$\begin{aligned} N &\in \left\{2^2, 2^3, \ldots, 2^{14}\right\} \\ c &\in \{0, 0.001, 0.002, \ldots, 0.01, 0.02, \ldots, 0.5\} \\ x_{t_0,11} &\in \{0, 0.05, 0.1, \ldots, 1\} \end{aligned}$$

$$x_{t_0,12} \in \{0, 0.05, 0.1, \ldots, (1 - x_{t_0,11})\}\,, \text{ for given } x_{t_0,11}$$
$$x_{t_0,21} \in \{0, 0.05, 0.1, \ldots, (1 - x_{t_0,11} - x_{t_0,12})\}\,, \text{ for given } x_{t_0,11}, \text{ and } x_{t_0,12}.$$

Note that $x_{t_0,22}$ is determined by $x_{t_0,22} = 1 - x_{t_0,11} - x_{t_0,12} - x_{t_0,21}$. For each parameter constellation, the number of realizations N_{sample} was chosen dynamically, as described below. Further note that 0.05 was chosen as grid-length for $\mathbf{x}_{t_0}$, although each component of $\mathbf{x}_{t_0}$ can theoretically take only values in $\left\{0, \frac{1}{2N}, \frac{2}{2N}, \ldots, 1\right\}$. For $N < 32$, we simulated according to this real grid, but got almost identical results. For large N, the computational costs are too high to simulate from the true grid $\left\{0, \frac{1}{2N}, \frac{2}{2N}, \ldots, 1\right\}$.

Parameter constellations causing at least one allele frequency to be zero were excluded from the analyses, because in this case $r^2_{t_0}$ cannot be calculated.

6.3.2 Measures of the goodness of fit

We propose the following characteristic as a measure of goodness of fit of eq. (6.6):

$$F := \frac{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1}) - a r^2_{t_0}}{b} - 1,$$

for given a and b. If equalities (6.6) and (6.7) were exact, we would observe $F = 0$ for all possible parameter constellations. In section 6.5.5 we show that F is closely related to the relative error of the expected LD at equilibrium due to the non-exactness of the recursion formula. Note that F is especially sensitive for a misspecification of b in eq. (6.7). Since $\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})$ is unknown, we use

$$\hat{F} = \frac{\widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})} - a r^2_{t_0}}{b} - 1,$$

where $\widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})}$ is obtained from the simulation study.

Dynamic sampling: In the simulation process described above, N_{sample} was chosen dynamically such that the standard deviation of $\hat{F}$ was approximately constant over all combinations $(N,c,\mathbf{x}_{t_0})$: It is

$$\begin{aligned} \text{s.d.}(\hat{F}) &= \text{s.d.}\left(\frac{\widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})} - a r^2_{t_0}}{b} - 1\right) \\ &= \frac{\text{s.d.}\left(\widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})}\right)}{b} \\ &= \frac{\text{s.d.}(r^2_{t_0+1})}{b \cdot \sqrt{N_{\text{sample}}}}. \end{aligned}$$

The right-hand side is constant if

$$N_{\text{sample}} = \left(\frac{\text{s.d.}(r^2_{t_0+1})}{b} \cdot d \right)^2$$

with any constant $d > 0$. To obtain $\text{s.d.}(r^2_{t_0+1})$, we performed a preliminary simulation study according to the same simulation set-up, but with a constant sample size of 10,000, and the empirical standard deviation of $r^2_{t_0+1}$ was calculated for all $(N,c,\mathbf{x}_{t_0})$-constellations. The value of d was chosen such that the maximal value of N_{sample} equaled $5 \cdot 10^6$. This led to an average sample size of 462,800 with a median of 15,000.

All statistical analyses were performed using R software (R Development Core Team, 2012). The R-package "multicore" (Urbanek, 2011) was used to parallelize the simulation.

6.3.3 Results of the simulation study

We found that $\hat{F}$ was centered around 0 (Supp. Figure S5). When all values of $\hat{F}$ below the 2.5% and above the 97.5% quantiles were excluded, $\hat{F}$ ranged between -1 and 1 indicating that the recursion formula fits the simulated data reasonably well. Values of $\hat{F}$ below the 2.5% quantiles and above the 97.5% quantiles were found to be generated by parameter constellations for which

$$P := x_{t_0,11} x_{t_0,12} x_{t_0,21} x_{t_0,22}$$

was close to zero, *i.e.* for constellations in which at least one gamete frequency in $T = t_0$ was close to zero (results not shown).

We used boxplots to display $\hat{F}$ for different parameter constellations. Boxplots were created separately for different values of N, c, P and $S := x_{t_0,11} + x_{t_0,12}$. Note that S equals the allele frequency p_{A_1} at the first locus and for symmetry arguments is representative for all other allele frequencies. Values of P (and S) were subdivided into 20 (and 15) equidistant bins, respectively. Outliers (*i.e.* values which lie beyond the extremes of the whiskers) are not displayed in any of the plots.

From Figure 6.1 we can see that the proposed recursion formula fits the data reasonably well, both for varying N and c. The bias as a function of N is almost constant, and it decays with increasing c. The goodness of fit depends heavily on P and S: $\hat{F}$ is larger and more variable for small P and for extreme S, but $\hat{F}$ still ranges between -1.5 and 1.5 in all considered boxplots.

Figure 6.2 shows that Sved's recursion formula does not fit the simulated data as well as the new formula, especially for $c > 0.01$ and for $N < 30$. This insufficient fit for $c > 0.01$ also pertains to the boxplots for varying P and S (as $\hat{F}$ is averaged over all constellations of $(N,c,\mathbf{x}_{t_0})$ for a fixed bin of P and S, respectively). For $c < 0.01$, there are only marginal differences between $\hat{F}$ based on Sved's recursion formula and the new recursion formula.

Contourplots were drawn for the empirical mean of $\hat{F}^2$, with the mean calculated using all values of $\hat{F}$ obtained for a given combination of values on the vertical and horizontal axis of the contourplot. For example, in the contourplot with axes (N,c) (*cf.* Figure 6.3) $\hat{F}^2$ values were averaged over all possible combinations of $(x_{t_0,11},x_{t_0,12},x_{t_0,21},x_{t_0,22})$ for a fixed

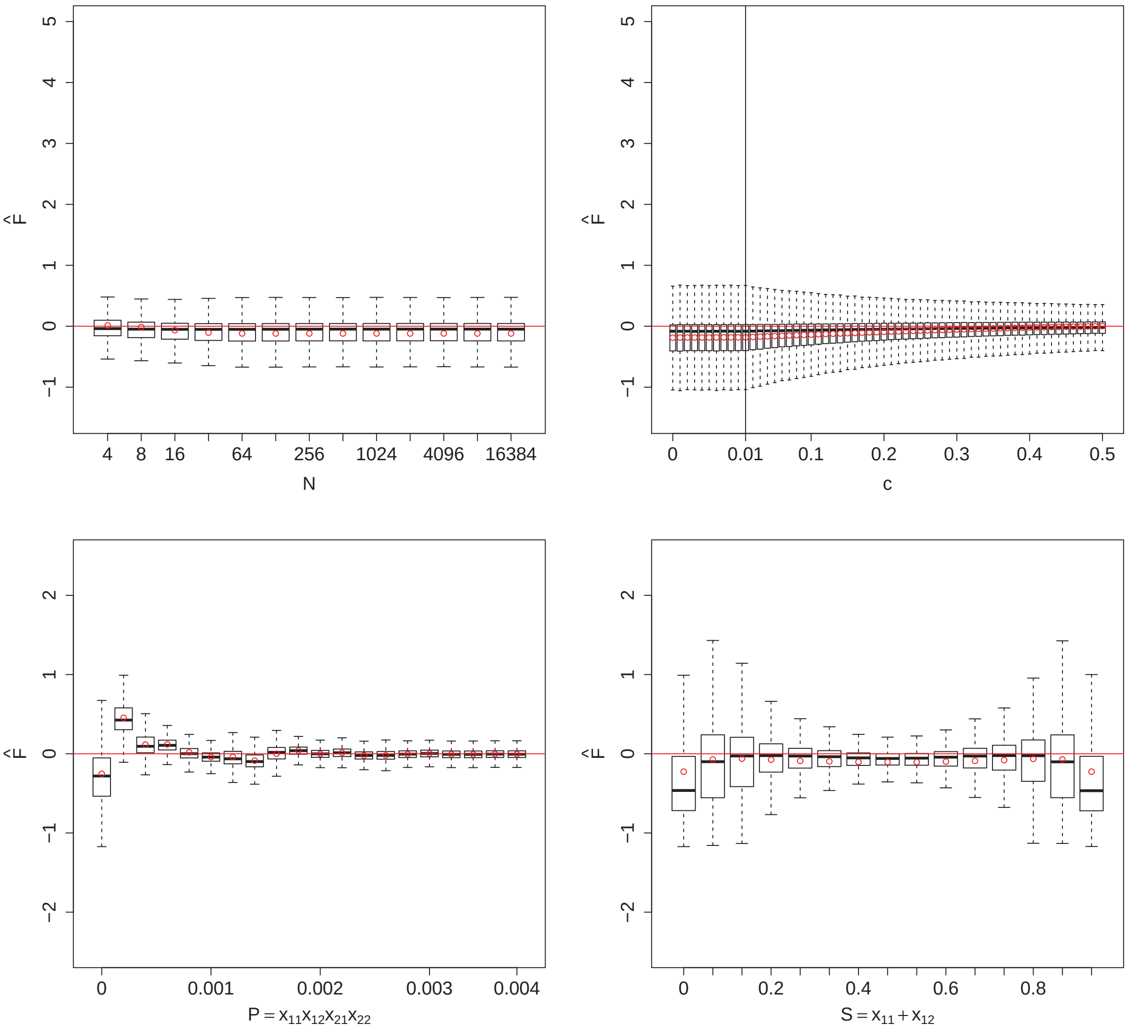

Figure 6.1: Boxplots of $\hat{F}$, separately for different bins of $N, c, P := x_{11}x_{12}x_{21}x_{22}$ and $S := x_{11} + x_{12}$, based on the new recursion formula. Here, P is the product of gamete frequencies and S is the allele frequency of the first allele at the first locus. $\hat{F}$ was calculated according to the new recursion formula. Outliers (*i.e.* values which lie beyond the extremes of the whiskers) are not shown.

combination (N,c). For a clearer representation of the contourplots, we excluded all values of $\hat{F}$ below the 2.5% and above the 97.5% quantiles beforehand.

The contourplot of Figure 6.3 shows that $\hat{F}^2$ depends only slightly on N and c, emphasizing the adequate fit of the new recursion formula. Figure 6.4 and Suppl. Figures S6 and S7 approve the previous results on the dependency of the goodness of fit on gamete and allele frequencies in $T = t_0$: The quality of the fit is reduced for $S < 0.2, S > 0.8, P < 0.0004$ as well as for

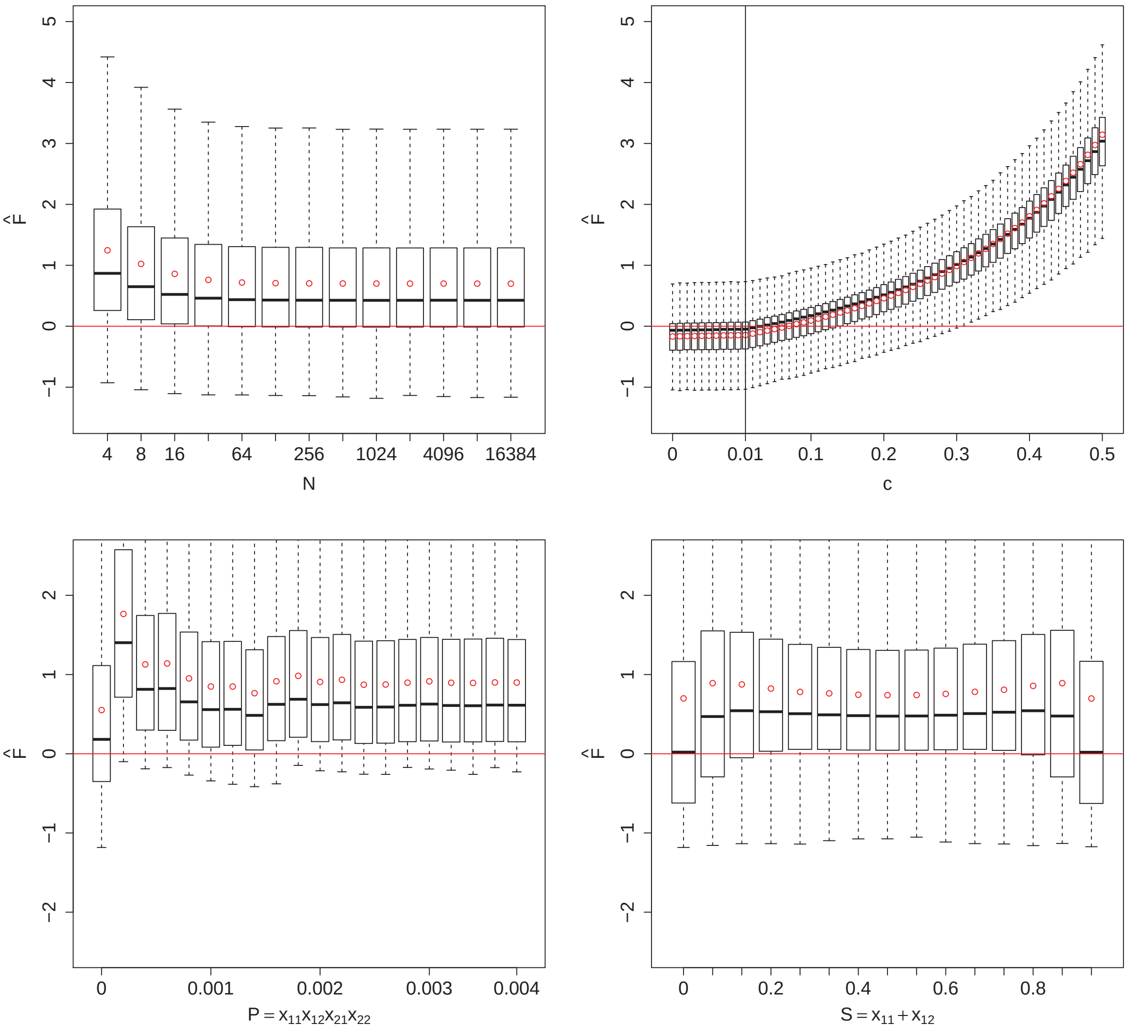

Figure 6.2: Boxplots of $\hat{F}$, separately for different bins of $N, c, P := x_{11}x_{12}x_{21}x_{22}$ and $S := x_{11} + x_{12}$, based on Sved's recursion formula. Here, P is the product of gamete frequencies and S is the allele frequency of the first allele at the first locus. $\hat{F}$ was calculated according to Sved's recursion formula. Outliers (*i.e.* values which lie beyond the extremes of the whiskers) are not shown.

$\varDelta\mathrm{MAF} := |\min(p_{A_1}, p_{A_2}) - \min(p_{B_1}, p_{B_2})| < 0.2$. The term $\varDelta$MAF describes the absolute difference in minor allele frequencies (MAFs) of both loci, with $p_{A_2} = 1 - p_{A_1} = x_{t_0,21} + x_{t_0,22}$ and $p_{B_2} = 1 - p_{B_1} = x_{t_0,12} + x_{t_0,22}$. To obtain the contourplots in this case, the values of $\varDelta$MAF were subdivided into 10 equally spaced bins.

Note that similar structures in the contourplots with respect to the dependencies on the allele frequencies can be observed for the goodness of fit of Sved's recursion formula (results

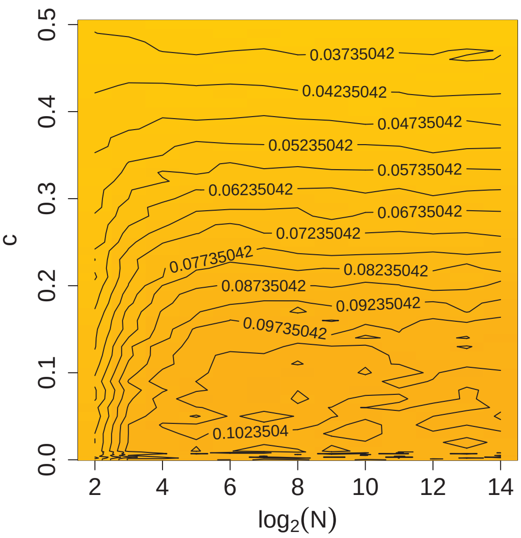

Figure 6.3: Contourplot of the average values of $\hat{F}^2$. For a given combination of $(\log_2(N), c)$, the values of $\hat{F}^2$ were averaged over all possible combinations of $\mathbf{x}_{t_0}$. Contourplots were created after excluding the extreme 2.5% quantiles of $\hat{F}$.

not shown).

Overall, the results confirm that an exact recursion formula must depend on the gamete frequencies. However, this would lead to complex formulae, especially for the state of equilibrium, which we will discuss in section 6.5.

6.3.4 Comparison of slope and intercept of the recursion formulae to empirical values

We compared the new recursion formula (eq. (6.6) in combination with eq. (6.7)) to the recursion formula of Sved (eq. (6.2)) by plotting the slope a against c for a given N. The same was done for the intercept b. Given N and c, we also fitted a linear regression model to the tuples $(r^2_{t_0}, \widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})})$ from the simulation study and added the points $(c, \hat{a})$ (and $(c, \hat{b})$, respectively) to the plots, where $\hat{a}$ and $\hat{b}$ were the estimated slope and intercept from the regression model. The results are shown in Figure 6.5.

The slopes a for the two recursion formulae are identical and coincide well with the empirical ones, with a better agreement for larger N (Figure 6.5). The intercepts, though, differ greatly remarkable between the two approaches, especially for $c > 0.1$ and for small N. The intercepts according to Sved's formula are not in agreement with the empirical ones for small N and $c > 0.1$. This is also reflected in the large values of $\hat{F}$ for increasing c (*cf.* Figure 6.2). Differences between the intercepts according to Sved's recursion formula and the new recursion formula become less pronounced for increasing N and for decreasing c.

We tried to improve the fit of a and b based on Figure 6.5, especially for small N, by

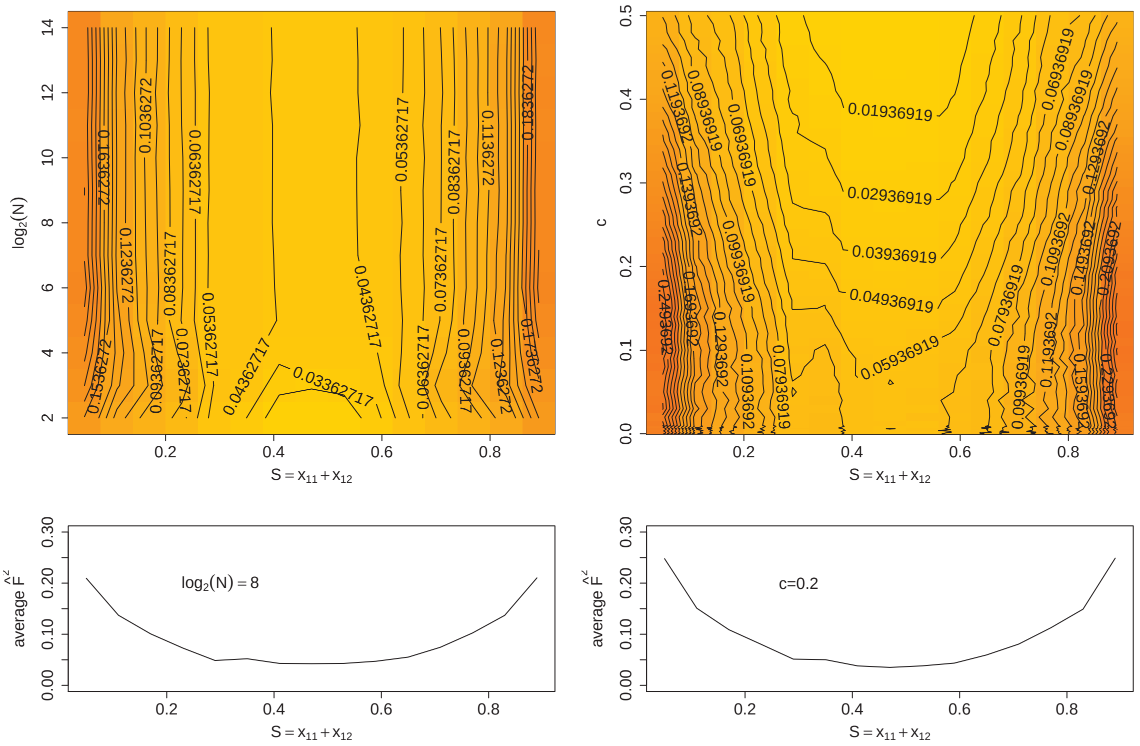

Figure 6.4: Contourplot of the average values of $\hat{F}^2$. Left plots: For a given combination of $(S, \log_2(N))$, the values of $\hat{F}^2$ were averaged over all possible values of c and $\mathbf{x}_{t_0}$ (upper plot). The lower plot illustrates the average value of $\hat{F}^2$ as a function of P for $\log_2(N) = 8$. Right plots: For a given combination of (S, c), the values of $\hat{F}^2$ were averaged over all possible values of $\log_2(N)$ and $\mathbf{x}_{t_0}$ (upper plot). The lower plot illustrates the average value of $\hat{F}$ as a function of P for $c = 0.2$. Contourplots were created after excluding the extreme 2.5% quantiles of $\hat{F}$.

using different formulae for a and b in eq. (6.7), *e.g.* $a = \left(1 - \frac{1}{3N}\right)\left(1 - c\left(1 - \frac{1}{3N}\right)\right)^2$ and $b = \frac{1}{2N+1-c}$. However this did not lead to a significant improvement in terms of $\hat{F}$ (results not shown). Since the true relation between $\mathbb{E}(r^2_{t_0+1})$ and $r^2_{t_0+1}$ is not linear, optimizing a (weighted) average of $\hat{F}$ is relevant. Thus, we would recommend to use Figures 6.1 and 6.2 as a basis for the assessment of adequacy.

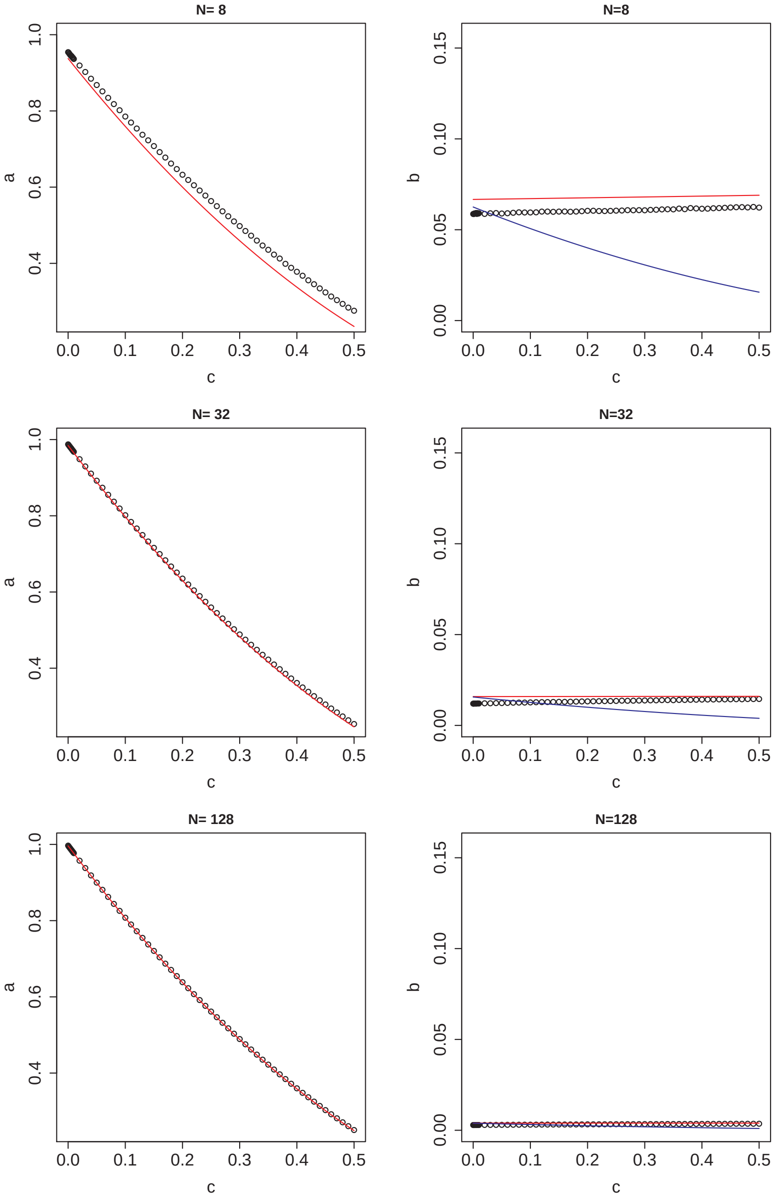

Figure 6.5: Comparison of slope and intercepts of Sved's and the new recursion formula to simulated values. Left plots: The slope a (identical for Sved's and the new recursion formula) is plotted against c for different values of N. The black dots are the slopes empirically obtained via linear regression. Hereby, the average LD values obtained in $T = t_0 + 1$ were regressed against $r^2_{t_0}$. Right plots: The intercepts b of the recursion formulae are plotted against c for different values of N. Blue (red) lines indicate Sved's (the new) recursion formula. The black dots are the slopes empirically obtained via linear regression.

6.4 The expected LD at equilibrium based on the theory of discrete-time Markov chains

6.4.1 Assuming that the recursion formula is exact

In previous studies, "equilibrium" was defined as the point in time at which the expected LD of the next generation equals the LD of the previous one (see *e.g.* Sved (1971); Tenesa *et al.* (2007)). Using this definition and assuming a linear recursion formula with coefficients a and b (eq. (6.6)), the expected LD at equilibrium equals $\frac{b}{1-a}$.

Two major problems arise from this definition: Firstly, it is not clear whether this equilibrium will ever be achieved. Secondly, one cannot infer from this definition how the formula for the expected LD at equilibrium is affected if the recursion formula is not exact but only approximate.

To overcome these problems, a mathematically deeper definition of equilibrium can be given based on the theory of Markov chains, since the sequence of gamete frequencies $\mathbf{x}_T, T = t_0, t_0 + 1, \ldots$ forms a homogeneous Markov chain with transition probabilities given by the multinomial distribution of the number of gametes of the four types in each generation. In this framework equilibrium is defined as the steady-state of the considered Markov chain, and the expected LD at equilibrium can be calculated as expectation of r^2 under the steady-state distribution.

Under the assumption that the underlying recursion formula is exact, the expected LD at equilibrium based on this approach turns out to be

$$R := \mathbb{E}(r_\infty^2) = \frac{b}{1-a} \tag{6.8}$$

for $|a| < 1$, in concordance with the above formula. A detailed derivation based on the Markov chain theory is given in section 6.5.2. Note that despite the apparent coincidence with formulae currently used in practice, usually no reference to the Markov chain theory is made. The same Markov chain model for the evolution of gamete frequencies has been used by Karlin & McGregor (1968) (in a study on the ascertainment of fixation probabilities) and by Littler (1973) (in a study on the LD measure $D = x_{11}x_{22} - x_{12}x_{21}$).

Furthermore, the Markov chain theory has the advantage that it also allows the calculation of the expected LD at equilibrium in the case of a non-exact recursion formula. We will come back to this issue in section 6.4.2, in which we will analyze how this non-exactness affects the formula for the expected LD at equilibrium.

Using a and b of eq. (6.7) in eq. (6.8) yields the following formula for the expected LD at equilibrium:

$$\begin{aligned} R &= \frac{\frac{1}{2N-1-c}}{1-(1-c)^2(1-\frac{1}{2N})} \\ &= \frac{1}{(2N-1-c)-(2N-1-c)(1-c)^2(1-\frac{1}{2N})} \end{aligned} \tag{6.9}$$

This formula differs from Sved's formula (eq. (6.1)). Solving eq. (6.9) for N yields

$$N = \frac{1}{2(8cR - 4c^2R)} \left(Y + \sqrt{-4(8cR - 4c^2R)(-R + cR + c^2R - c^3R) + Y^2} \right), \qquad (6.10)$$

with $Y := 2 - 2R + 8cR - 2c^3R$, taking into account that N cannot be negative.

To compare the formulae for the expected LD at equilibrium according to eqs. (6.1) and (6.9), we plotted $\mathbb{E}(r_\infty^2)$ based on both recursion formulae against c for $N \in \{4,16,64,256\}$. Results are shown in Figure 6.6.

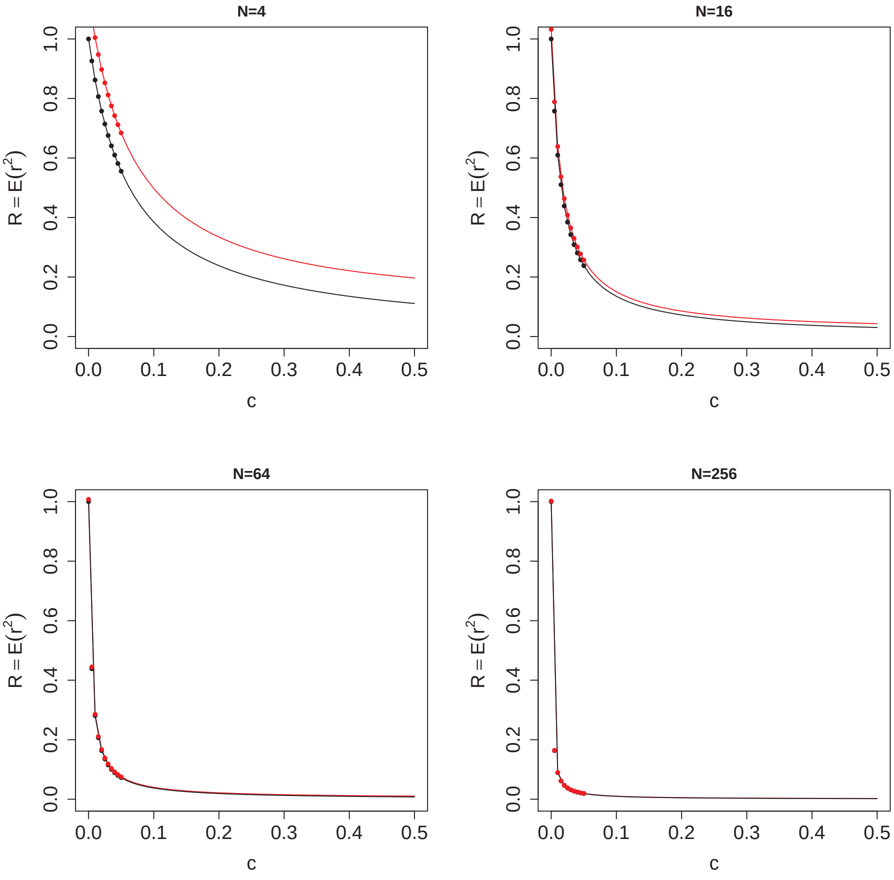

Figure 6.6: Comparison of Sved's formula and the new formula for the expected LD R at equilibrium: R is plotted against c for different values of N. Red (black) lines show R according to the new (Sved's) formula. The dots indicate the values of R for $c = 0, 0.005, \ldots, 0.05$.

Only small differences between both recursion formulae are observed for large N in Figure 6.6, whereas the difference gets more pronounced, when N is small (*cf.* Figure 6.5 where similar effects can be realized). The new recursion formula predicts higher values for the expected LD at equilibrium for small values of N.

Real populations usually do not fulfill the implicit assumptions of ideal populations. It is therefore of interest to calculate the effective population size N_e based on the average LD-value observed from the population for a given value of c. By definition, N_e is the size of an ideal population at equilibrium with the same structure of LD as the population under consideration. In practice, N_e is obtained from the right-hand side of eq. (6.10), using the average LD-value observed from the data as value of R.

6.4.2 Non-exactness of the recursion formulae

As indicated above, another problem which has not been discussed in the literature so far arises from the non-exactness of the recursion formulae. This problem can also be overcome by the Markov chain theory.

In the case of a non-exact recursion formula, the Markov chain theory allows to transfer an error term in the recursion formula to the state of equilibrium. In section 6.5.3, we provide the corresponding calculations and show that in case of a non-exact recursion formula

$$R^{\varepsilon} := \mathbb{E}_{\boldsymbol{\mu}}(r_{\infty}^2) = \frac{b + \boldsymbol{\pi}^* \boldsymbol{\varepsilon}}{1-a}$$

for $|a| < 1$, where $\boldsymbol{\pi}^*$ is the stationary distribution of the considered Markov chain and

$$\begin{aligned} \boldsymbol{\varepsilon} &= (\varepsilon(s_1), \ldots, \varepsilon(s_z))^T \\ \text{with} \quad \varepsilon(s_i) &:= \mathbb{E}_{s_i}(r_{t_0+1}^2) - ar^2(s_i) - b \end{aligned}$$

is the residual term of the recursion formula depending on the different possible values s_i of $\mathbf{x}_{t_0}$. Then, the term $\frac{R^{\varepsilon}}{R} - 1 = \frac{\boldsymbol{\pi}^* \boldsymbol{\varepsilon}}{b}$ measures the relative influence of $\boldsymbol{\pi}^* \boldsymbol{\varepsilon}$ on the expected LD.

To analyze the effect of the non-exactness, we calculated $\frac{\boldsymbol{\pi}^* \boldsymbol{\varepsilon}}{b}$ for different (N,c)-combinations. The results are listed in Table 6.1 for $N = 4,8,16$ and $c = 0.001,0.01,0.05,0.1,0.2,0.3$. More details on this are given in sections 6.5.4, 6.5.5 and 6.5.6.

Table 6.1 illustrates that the error-term can lead to a deviance of expected LD up to 25% suggesting that the effect of non-exactness may be non-negligible. These analyses were restricted to small values of N, since the calculation times increase rapidly with N.

To get a first impression on the development of $\frac{\boldsymbol{\pi}^* \boldsymbol{\varepsilon}}{b}$ for $N > 16$, we also plotted the negative mean and the maximum value of $\boldsymbol{\varepsilon}$ divided by b, given by the terms $S_1 = \frac{-\frac{1}{z}\sum_i \varepsilon_i}{b}$ and $S_2 = \frac{\max_i |\varepsilon_i|}{b}$, for more values of N and c (Figure 6.7). Note that this does not incorporate the stationary distribution $\boldsymbol{\pi}^*$ and that $\boldsymbol{\varepsilon}$-values in these plots were based on the simulation study using the grid of $(x_{11}, x_{12}, x_{21}, x_{22})$-values with fixed grid-distance of 0.05, which is not as fine as the "true" grid if $N > 20$ so that results at this point have to be taken with caution. More detailed analyses and a comprehensive simulation study are needed to underpin the quantitative results on the influence of $\boldsymbol{\varepsilon}$ on the expected LD at equilibrium.

Table 6.1: Values of $\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}$ for different (N,c)-combinations. Absorbing states were excluded beforehand, and $\boldsymbol{\pi}^*$ was rescaled so that its entries summed up to 1 afterwards.

c	$N = 4$	$N = 8$	$N = 16$
0.001	−0.264	−0.199	−0.067
0.01	−0.255	−0.165	−0.023
0.05	−0.176	−0.088	0.096
0.1	−0.116	−0.018	0.195
0.2	−0.074	0.053	0.297
0.3	−0.025	0.097	0.324

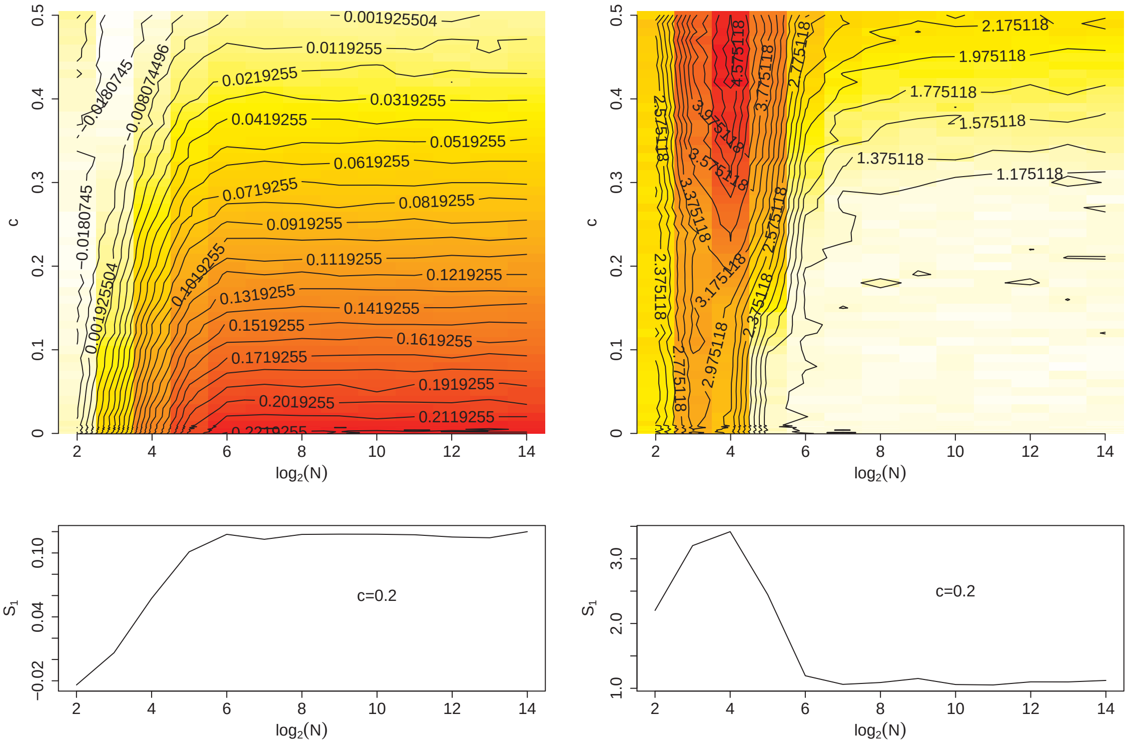

Figure 6.7: Values of $S_1 = \frac{-\frac{1}{z}\sum\varepsilon_i}{b}$ (upper left plot) and $S_2 = \frac{\max_i|\varepsilon_i|}{b}$ (upper right plot) for different values of N and c, calculated on the basis of the simulation study. Note that the simulation study was based on a grid of $(x_{t_0,11}, x_{t_0,12}, x_{t_0,21}, x_{t_0,22})$-values with fixed grid-length 0.05. Absorbing states were excluded beforehand. The left (right) lower plot illustrates the values of S_1 (S_2) as a function of $\log_2(N)$ for $c = 0.2$.

6.5 More details on the expected LD at equilibrium based on Markov chain theory

6.5.1 The expected LD at equilibrium based on a recursion formula

Given a recursion formula like eq. (6.6), we will derive a formula for the expected LD at equilibrium which is based on the theory of Markov chains (for introductory books on Markov chains we refer to Grimmett & Stirzaker (2001) or Norris (1997), for instance). Note that the derivation pertains to all recursion formulae with arbitrary coefficients a and b with $|a| < 1$ and that we will provide a general mathematical description of the term "equilibrium" which will be defined as the steady-state or "equilibrium state" of the considered Markov chain.

According to the multinomial model for the development of the population of gametes (see section 6.2), the sequence of gamete frequencies $\mathbf{x}_T, T = t_0, t_0 + 1, \ldots$ forms a homogeneous Markov chain with transition matrix $\mathbf{P}$ which is given by the multinomial distribution of the number of gametes of the four types in each generation. The parameters of the multinomial distribution are $2N$ and $p = (x'_{T,11}, x'_{T,12}, x'_{T,21}, x'_{T,22})$. Since the population size is finite, the Markov chain has a finite set S of states $s_1, \ldots, s_z$. Here, the s_i are quadruples of frequencies $(x_{11}, x_{12}, x_{21}, x_{22})$, each of which describes a possible partition of the $2N$ gametes into the four types of gametes. In total, there are $\binom{2N+3}{2N}$ possible states (Karlin & McGregor, 1968). Let $\boldsymbol{\pi}_T, T \geq t_0$, denote the probability vector of $\mathbf{x}_T$. Then

$$\boldsymbol{\pi}_{t_0+n} = \boldsymbol{\pi}_{t_0}\mathbf{P}^n$$

for $n = 1, 2, \ldots$. We write $r_T^2 := r^2(\mathbf{x}_T)$, $\mathbb{E}_{s_j}(r_T^2) := \mathbb{E}(r_T^2|\mathbf{x}_{t_0} = s_j)$ for all $T \geq t_0$, and $\mathbf{e}_j$ for the j-th unit vector ($\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ where the 1 is at the j-th position).

6.5.2 Step I: Assuming that the recursion formula is exact

Let us first assume that the recursion formula (6.6) with coefficients a and b holds for some statistic r^2 depending on the time T and on the state $\mathbf{x}_{t_0}$ in $T = t_0$. Note that the following derivation is valid for arbitrary values of a and b with $|a| < 1$. From the recursion formula we get

$$\sum_j p_j \mathbb{E}_{s_j}(r_{t_0+1}^2) = a \sum_j p_j r^2(s_j) + b$$

for all probability vectors $\boldsymbol{\mu} = (p_1, \ldots, p_z)$ with $p_j \geq 0$ and $\sum_j p_j = 1$. With a slight abuse of notation, we also write $\mathbb{E}_{\boldsymbol{\mu}}(r_T^2)$ for the expectation of r_T^2, given that the initial probability vector $\boldsymbol{\pi}_{t_0}$ equals $\boldsymbol{\mu}$. Then, the last equation is equivalent to

$$\sum_j p_j \left(\sum_i (\mathbf{e}_j \mathbf{P})_i r^2(s_i) \right) = a\mathbb{E}_{\boldsymbol{\mu}}(r_{t_0}^2) + b.$$

The left-hand side equals

$$\sum_i \sum_j p_j (\mathbf{e}_j \mathbf{P})_i r^2(s_i) = \sum_i (\boldsymbol{\mu}\mathbf{P})_i r^2(s_i) = \mathbb{E}_{\boldsymbol{\mu}}(r^2_{t_0+1}).$$

Hence, we have

$$\mathbb{E}_{\boldsymbol{\mu}}(r^2_{t_0+1}) = a\mathbb{E}_{\boldsymbol{\mu}}(r^2_{t_0}) + b$$

for an arbitrary initial probability vector $\boldsymbol{\mu}$, and the weak Markov property yields

$$\mathbb{E}_{\boldsymbol{\mu}}(r^2_{T+1}) = a\mathbb{E}_{\boldsymbol{\mu}}(r^2_{T}) + b$$

for all $T \geq t_0$. If $\boldsymbol{\pi}_{t_0} = \boldsymbol{\mu}$, then $\boldsymbol{\pi}_T = \boldsymbol{\mu}\mathbf{P}^{T-t_0}$, and the last equation is equivalent to

$$\sum_j (\boldsymbol{\pi}_{T+1})_j r^2(s_j) = a \sum_j (\boldsymbol{\pi}_T)_j r^2(s_j) + b.$$

If the Markov chain is regular, the convergence theorem for regular discrete Markov chains yields $\boldsymbol{\pi}_T \to \boldsymbol{\pi}^*$ for $T \to \infty$ with $\boldsymbol{\pi}^*$ being the unique stationary distribution, *i.e.* both sides converge and we get

$$\mathbb{E}_{\boldsymbol{\mu}}(r^2_{\infty}) = a\mathbb{E}_{\boldsymbol{\mu}}(r^2_{\infty}) + b \quad \Leftrightarrow \quad R := \mathbb{E}_{\boldsymbol{\mu}}(r^2_{\infty}) = \frac{b}{1-a} \tag{6.11}$$

for $|a| < 1$. Note that we need regularity of the Markov chain when applying the convergence theorem and that a chain is called "regular" if some power of $\mathbf{P}$ contains only positive elements. In our setting with $\mathbf{P}$ based on the multinomial distribution, this is a priori not true since "absorbing states" in the Markov chain exist. These absorbing states reflect situations in which one or more alleles are fixated. We will deal with this problem in section 6.5.4.

6.5.3 Step II: Dealing with non-exactness of the recursion formula

Since we know that eq. (6.6) with a and b only depending on N and c is not correct, we will now analyze how the non-exactness of eq. (6.6) affects the formula $R = \frac{b}{1-a}$ (*cf.* eq. (6.11)). For each state $\mathbf{x}_{t_0}$, let

$$\varepsilon(\mathbf{x}_{t_0}) := \mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1}) - ar^2(\mathbf{x}_{t_0}) - b \tag{6.12}$$

be the residual term of the proposed recursion formula, and let further $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_z)^T = (\varepsilon(s_1), \ldots, \varepsilon(s_z))^T$ be the vector of residual terms for the different states s_i of the Markov chain. As before, we can calculate

$$\sum_j p_j \left(\sum_i (\mathbf{e}_j \mathbf{P})_i r^2(s_i) \right) = a\mathbb{E}_{\boldsymbol{\mu}}(r^2_{t_0}) + b + \sum_j p_j \varepsilon(s_j),$$

leading to

$$\mathbb{E}_{\boldsymbol{\mu}}(r^2_{t_0+1}) = a\mathbb{E}_{\boldsymbol{\mu}}(r^2_{t_0}) + b + \boldsymbol{\mu\varepsilon}$$

for an arbitrary initial probability vector $\boldsymbol{\mu}$. The weak Markov property yields

$$\mathbb{E}_{\boldsymbol{\mu}}(r^2_{T+1}) = a\mathbb{E}_{\boldsymbol{\mu}}(r^2_{T}) + b + \boldsymbol{\mu}\mathbf{P}^{T-t_0}\boldsymbol{\varepsilon}$$

for all $T \geq t_0$. If $\boldsymbol{\pi}_{t_0} = \boldsymbol{\mu}$, then $\boldsymbol{\pi}_T = \boldsymbol{\mu}\mathbf{P}^{T-t_0}$, and the last equation is equivalent to

$$\sum_j (\boldsymbol{\pi}_{T+1})_j r^2(s_j) = a \sum_j (\boldsymbol{\pi}_T)_j r^2(s_j) + b + \boldsymbol{\pi}_T\boldsymbol{\varepsilon}.$$

Using $\boldsymbol{\pi}_T \to \boldsymbol{\pi}^*$ for $T \to \infty$ as before, this finally leads to

$$\mathbb{E}_{\boldsymbol{\mu}}(r^2_{\infty}) = a\mathbb{E}_{\boldsymbol{\mu}}(r^2_{\infty}) + b + \boldsymbol{\pi}^*\boldsymbol{\varepsilon} \quad \Leftrightarrow \quad R^{\varepsilon} := \mathbb{E}_{\boldsymbol{\mu}}(r^2_{\infty}) = \frac{b + \boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{1-a} \tag{6.13}$$

for $|a| < 1$. Hence the formula for the expected LD at equilibrium differs from eq. (6.11) by the summand $\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{1-a}$.

6.5.4 Dealing with absorbing states

In the setting so far, the Markov chain contains several “absorbing” states (states which force the chain to move in a certain subset of the set of states). These absorbing states correspond to situations in which one or two alleles at the considered two loci are already fixed. Hence, the Markov chain is not regular and the convergence theorem for Markov chains cannot be applied. Furthermore, r^2 is not defined in case one or more allele frequencies are equal to zero. In practice, pairs of SNPs with fixed alleles are not considered when estimating the expected LD in the population. Therefore, we propose to modify the transition matrix $\mathbf{P}$ of the chain by enforcing at least one immediate mutation of an allele in case this allele has become fixed. The corresponding rows of $\mathbf{P}$ are modified for the absorbing states by choosing the transition probabilities in these rows as indicated in Table 6.2 mimicking the enforced mutations to leave the absorbing state.

Note that, since $r^2_{t_0}$ could also not be calculated in the simulation study when one or more allele frequencies were equal to zero, this modification of the Markov chain does not influence the recursion formula and the results of the simulation study with respect to the goodness of fit.

Further note that this modification is biologically inspired by the event of mutations and that this modification is only one possibility among others to deal with the problem of absorbing states. One could *e.g.* also discard columns and rows of absorbing states in the $\mathbf{P}$-matrix and rescale the rows so that their sums are equal to 1. Yet, it is a priori not clear which effect different procedures have on the resulting stationary distribution $\boldsymbol{\pi}^*$ and what they mean in terms of a stochastic model underlying the chain. Further research is needed in this area.

In the following, we will concentrate on the first possibility described above mimicking

Table 6.2: Absorbing states and their modified transition probabilities.

"Absorbing state"				Transition	State after enforced mutation			
x_{11}	x_{12}	x_{21}	x_{22}	probability	x_{11}	x_{12}	x_{21}	x_{22}
1	0	0	0	1	$1-\frac{2}{2N}$	$\frac{1}{2N}$	$\frac{1}{2N}$	0
e^*	$1-e$	0	0	0.5	$e-\frac{1}{2N}$	$1-e$	$\frac{1}{2N}$	0
				0.5	e	$1-e-\frac{1}{2N}$	0	$\frac{1}{2N}$
0	1	0	0	1	$\frac{1}{2N}$	$1-\frac{2}{2N}$	0	$\frac{1}{2N}$
0	e	0	$1-e$	0.5	$\frac{1}{2N}$	$e-\frac{1}{2N}$	0	$1-e$
				0.5	0	e	$\frac{1}{2N}$	$1-e-\frac{1}{2N}$
0	0	1	0	1	$\frac{1}{2N}$	0	$1-\frac{2}{2N}$	$\frac{1}{2N}$
e	0	$1-e$	0	0.5	$e-\frac{1}{2N}$	$\frac{1}{2N}$	$1-e$	0
				0.5	e	0	$1-e-\frac{1}{2N}$	$\frac{1}{2N}$
0	0	0	1	1	0	$\frac{1}{2N}$	$\frac{1}{2N}$	$1-\frac{2}{2N}$
0	0	$1-e$	e	0.5	$\frac{1}{2N}$	0	$e-\frac{1}{2N}$	$1-e$
				0.5	0	$\frac{1}{2N}$	e	$1-e-\frac{1}{2N}$

[*] with arbitrary constant $e \in (0, 1)$

biological mutations. If $c > 0$, all transition probabilities are strictly larger than zero for some power of the modified transition matrix, and the Markov chain is regular allowing for the calculation of expected LD at equilibrium as described in the previous sections. From now on, we will restrict to the modified transition matrix.

6.5.5 The term $\frac{\pi^*\varepsilon}{b}$ as parameter of interest

Let R be as before and let R^ε denote the expected LD at equilibrium taking into account the error term $\boldsymbol{\varepsilon}$. Then, the relative difference between these two values is given by

$$\frac{R^\varepsilon}{R} - 1 = \frac{\frac{b+\pi^*\varepsilon}{1-a}}{\frac{b}{1-a}} - 1 = \frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}.$$

Hence, $\left|\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}\right|$ measures the relative influence of $\boldsymbol{\pi}^*\boldsymbol{\varepsilon}$ on the expected LD. Note that $\frac{\boldsymbol{\varepsilon}}{b} = (F(s_1), \ldots, F(s_z))$ and that $\boldsymbol{\pi}^*$ depends on N and c. If we were able to obtain the stationary distribution $\boldsymbol{\pi}^*$ for a fixed combination of N and c, we could quantify $\left|\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}\right|$. The identity $\frac{\boldsymbol{\varepsilon}}{b} = (F(s_1), \ldots, F(s_z))$ motivates the choice of F as a measure of goodness of fit of the recursion formula since we are especially interested in the expected LD at equilibrium.

The following statistics give a first glance at the behavior of $\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}$:

$$S_1 := \frac{-\frac{1}{z}\sum_i \varepsilon_i}{b} \quad \text{and} \quad S_2 := \frac{\max_i |\varepsilon_i|}{b}$$

S_1 is closely related to Figure 6.1 and corresponds to the negative mean of values illustrated in each boxplot. S_2 gives an upper bound for $\left|\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}\right|$.

6.5.6 Empirical analysis based on the new recursion formula

To analyze the term $\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}$ for the new recursion formula, we repeated the simulation study described in section 6.3 for $N = 4{,}8{,}16$ and $c = 0.001{,}0.01{,}0.05{,}0.1{,}0.2{,}0.3$ using the following (true) grid for $\mathbf{x}_{t_0} = (x_{t_0,11}, x_{t_0,12}, x_{t_0,21}, x_{t_0,22})$:

$$\begin{aligned}
x_{t_0,11} &\in \left\{0, \frac{1}{2N}, \frac{2}{2N}, \ldots, 1\right\} \\
x_{t_0,12} &\in \left\{0, \frac{1}{2N}, \frac{2}{2N}, \ldots, (1 - x_{t_0,11})\right\}, \text{ for given } x_{t_0,11} \\
x_{t_0,21} &\in \left\{0, \frac{1}{2N}, \frac{2}{2N}, \ldots, (1 - x_{t_0,11} - x_{t_0,12})\right\}, \text{ for given } x_{t_0,11}, \text{ and } x_{t_0,12}.
\end{aligned}$$

As mentioned before, this grid comprises the exact and full set of states of the Markov chain. We chose N_{sample} so that it had approximately the same magnitude as in the previous simulations.

To obtain the stationary distribution $\boldsymbol{\pi}^*$, we built the transition matrix $\mathbf{P}$ of the Markov chain according to the multinomial distribution. The absorbing states listed in Table 6.2 were modified as described above. Then, we calculated $\mathbf{P}^n$ for $n = 2^1, \ldots, 2^{15}$. At equilibrium, each column of $\mathbf{P}^n$ is constant. By graphical inspection, it could be observed that this situation was always reached within $n = 2^{15}$ generations so that all rows of the power $\mathbf{P}^n$ were equal to the stationary distribution $\boldsymbol{\pi}^*$. In practice, $\widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})}$ is estimated using SNP pairs with non-fixed alleles in the population. Hence, we are interested in $\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}$ where $\boldsymbol{\pi}^*$ and $\boldsymbol{\varepsilon}$ only contain non-absorbing states. Therefore, we calculated $\varepsilon_i(\mathbf{x}_{t_0})$ for all non-absorbing states $\mathbf{x}_{t_0}$ based on $\widehat{\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})}$ obtained from the simulation and rescaled $\boldsymbol{\pi}^*$ so that its sum equaled 1 after excluding all absorbing states. Then, $\frac{\boldsymbol{\pi}^*\boldsymbol{\varepsilon}}{b}$ could be calculated for different (N,c)-combinations, to analyze the influence of the non-exactness of the recursion formula.

6.6 Application based on the HapMap data

As an application of the equilibrium-formula based on the proposed recursion formula, we estimated N_e from LD using human data from the HapMap project (The International HapMap Consortium, 2003, 2007) and applying eq. (6.10) as described in section 6.4.1. We also investigated, how the MAF distribution of SNP pairs, used to estimate the expected LD in the population, influences the average LD values and hence also the estimates of N_e.

6.6.1 The HapMap data set

The HapMap data set comprises 270 samples from four populations. In this study, we consider two different populations, the Yoruba in Ibadan, Nigeria (YRI) and Utah residents with Northern and Western European ancestry from the CEPH collection (CEU). For each population, the data comprises 30 trios of individuals. The data are available from `http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en`. For both populations, we used allele frequencies from phases II and III (release #27) as well as LD data from phases I, II and III (release #27) for single nucleotide polymorphisms (SNPs) lying on the 22 autosomes, and a corresponding genetic map (from phase II, estimated from phased haplotypes in release #22 (NCBI 36)) (The International HapMap Consortium, 2007). LD values were available for markers up to 200kb apart. For autosome 22, *e.g.*, there were $\approx$ 38,000 (34,000) SNPs occurring in 10,133,060 (8,130,042) LD values for the YRI (CEU) population, for which the genetic distance between the corresponding SNP pairs was available. Summing over the 22 autosomes, there were in total $\approx$ 2,868,000 (2,560,000) SNPs occurring in 701,820,000 (563,239,000) LD values for the YRI (CEU) population.

6.6.2 Estimation of N_e for the YRI and CEU population

We estimated N_e separately for the YRI and the CEU population for each of the 22 autosomes, using eq. (6.10) for the expected LD at equilibrium, with $R = \mathbb{E}(r^2_\infty)$, estimated as average LD value obtained from the data for given c, and replacing N with N_e.

Following Weir & Hill (1980) we adjusted for the chromosome sample size n by subtracting $\frac{1}{n}$ from the sample-based LD values. This is necessary, since even in the case of independent loci $\mathbb{E}(r^2) = \frac{1}{n}$. It has been shown by Bishop *et al.* (1975), p. 382, that nr^2 has an approximate χ^2_1 distribution for a bivariate Bernoulli distribution with independent components, and hence $\mathbb{E}(r^2) = \frac{1}{n}$ in this case. With this adjustment,

$$\hat{N}_e = \frac{1}{2(8c\tilde{R} - 4c^2\tilde{R})}\left(\tilde{Y} + \sqrt{-4(8c\tilde{R} - 4c^2\tilde{R})(-\tilde{R} + c\tilde{R} + c^2\tilde{R} - c^3\tilde{R}) + \tilde{Y}^2}\right), \qquad (6.14)$$

with $\tilde{Y} := 2 - 2\tilde{R} + 8c\tilde{R} - 2c^3\tilde{R}$ and $\tilde{R} = \widehat{\mathbb{E}(r^2)} - \frac{1}{n}$.

For the HapMap data of the YRI and CEU population, $n = 120$, since sequences of 30 trios for each population were available, comprising 4 independent parental gametes for each trio. Estimates of N_e and average LD values were obtained for different bins of the recombination rate c. To classify the pairs of SNPs to the bins, c was approximated by the genetic distance in Morgan. Note that this approach is admissible for small distances. For each autosome, 100 equidistant bins of c ranging from 0 to the maximal genetic distance occurring in the data were chosen. For each bin of c, the average r^2 value minus $\frac{1}{120}$ was calculated and used in eq. (6.14) to obtain an estimate of N_e. The estimated N_e was plotted against $\log_{10}(\frac{1}{2c})$ (Figure 6.8). Additionally, a plot of the adjusted average LD value against c was generated (Figure 6.8). Results were plotted for all bins of c containing at least 1,000 LD values.

The decay of LD with genetic distance for the YRI and CEU population can be seen in the upper plots of Figure 6.8, estimates of N_e are displayed in the lower plots. Note that, since N_e of human populations is large ($N_e > 1{,}000$), eqs. (6.9) and (6.1) basically lead to the same estimates (results not shown). N_e is smaller for the CEU population (lower right plot of Figure 6.8) and increasing from $\approx 5{,}000$ to $\approx 10{,}000$ for $\frac{1}{2c}$ ranging from 1,500 to 200, whereas N_e for the YRI population is decreasing for these values of $\frac{1}{2c}$. Hayes *et al.* (2003) argue that the N_e estimate for a fixed c corresponds to an estimated N_e approximately $\frac{1}{2c}$ generations ago. Applying this concept and assuming a generation interval of 25 years, the above time frame encompasses 37,500 to 5,000 years ago, and we find $\hat{N}_e \approx 15{,}000\,(5{,}800)$ for the YRI (CEU) population $\approx 1{,}000$ generations ($= 25{,}000$ years) ago, as well as $\hat{N}_e \approx 20{,}500\,(10{,}000)$ for YRI (CEU) $\approx 8{,}000$ generations ($= 200{,}000$ years) ago (*cf.* Figure 6.8).

For values of c with $\log_{10}(\frac{1}{2c}) < 1.75$, a high variability of $\hat{N}_e$ values can be observed (Figure 6.8). We hypothesize that the corresponding values of LD observed from the data are in the order of magnitude one would expect if loci were independent, in which case it would not make sense to estimate N_e. This hypothesis can also be warranted by the approximate χ^2 distribution of nr^2 for independent loci.

6.6.3 Influence of MAF distribution on the N_e-estimates

As the detailed analysis of $\hat{F}$ indicated that the expected LD also depends on the distribution of allele frequencies, it is important to investigate how the underlying MAF distribution affects the estimation of N_e.

In commercial SNP array construction for animal breeding purposes, the use of SNPs with uniform MAF distribution is common practice. A uniform MAF distribution is in general not pursued in human genetics, but may still occur, *e.g.* in studies using phase I data of the HapMap project (The International HapMap Consortium, 2003), where an ascertainment bias can be observed (Nielsen, 2000; Nielsen *et al.*, 2004; Clark *et al.*, 2005; Pe'er *et al.*, 2006).

An enforced uniform MAF distribution may introduce a systematic and substantial downward bias in N_e estimates, especially for historical effective population sizes, which we demonstrated with the YRI and the CEU population using data of autosome 22:

Histograms for the MAF values for all SNPs occurring in SNP pairs for which LD values and the genetic distance were available (Figure 6.9) show that in both populations low MAFs are overrepresented.

For each population we sampled 10,000 SNP positions out of the $\approx 38{,}000\,(36{,}000)$ SNPs available for YRI (CEU) according to two different scenarios:

(1) The 10,000 positions were sampled randomly, *i.e.* from the true skewed MAF distribution (*cf.* Figure 6.9).

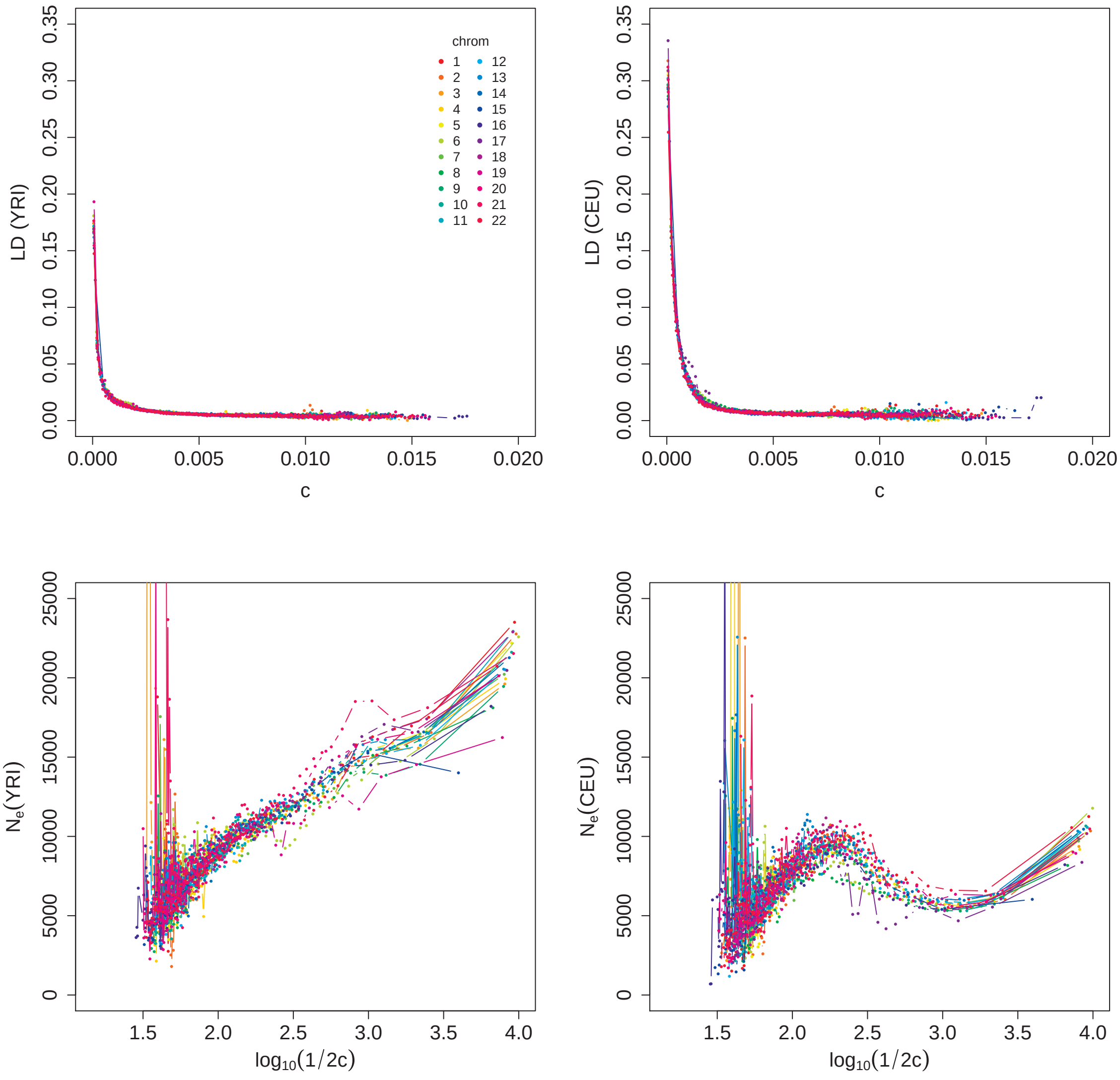

Figure 6.8: LD and estimates of N_e for the YRI and CEU population based on the new recursion formula displayed for the 22 autosomes. The upper plots show the decay of LD for varying c, estimated from SNPs on different autosomes. In the lower plots, the corresponding estimates of N_e are plotted against $\log_{10}(\frac{1}{2c})$. The left (right) plots are for the YRI (CEU) population. N_e estimates based on a given value of c correspond to the point in time "$\frac{1}{2c}$ generations ago" (Hayes *et al.*, 2003).

(2) MAFs were divided into 10 equidistant bins between 0 and 0.5. Then, 1,000 SNPs from each bin were sampled to mimic a uniform MAF distribution, and all those LD values of pairs of SNPs were kept for which the positions of both SNPs were among

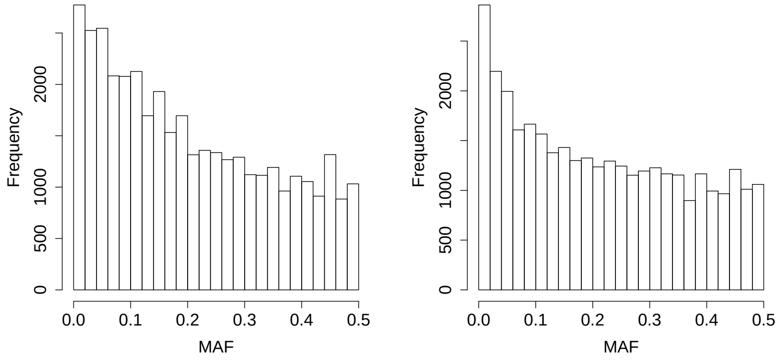

Figure 6.9: Histograms of the MAF distribution for all SNPs occurring in the LD data of autosome 22. The left (right) plot shows the histogram for the YRI (CEU) population.

the 10,000 sampled positions.

For each scenario, N_e was estimated for different bins of c. We chose 24 equidistant bins of c ranging from 0 to ≈ 0.002 and 25 equidistant bins of c ranging from ≈ 0.002 to ≈ 0.02. The whole sampling process was repeated 100 times. This resulted in 100 estimated N_e values per scenario and per bin of c, for which boxplots were created to display the distributions of N_e for both scenarios. Boxplots were only created for bins in which on average (averaged with respect to the 100 replicates) at least 1,000 LD-values were available.

Figure 6.10 illustrates the influence of the MAF distribution of SNP pairs used for LD estimation on the N_e estimates in the two populations, respectively. N_e was estimated for different time points "$\frac{1}{2c}$ generations ago" (Hayes *et al.*, 2003), as described above. The plots demonstrate that the N_e estimates using a skewed MAF distribution are up to 30% larger than the ones using a uniform MAF distribution for large values of $\frac{1}{2c}$. For example, for $\frac{1}{2c} = 500$, the estimated N_e ranges from $\approx 9{,}000\,(5{,}400)$ using a skewed MAF distribution to $\approx 12{,}100\,(6{,}900)$ using a uniform MAF distribution and the YRI (CEU) population.

6.6.4 Comparison with recent results of other studies

Tenesa *et al.* (2007) used the phase I HapMap data to estimate N_e in the YRI and CEU population based on $\approx 1{,}000{,}000$ SNPs from 23 chromosomes. The intermarker distance was in the range of 5kb to 100kb for all SNP pairs. Using only SNPs on autosome 22 and estimating recombination rates from a nonlinear model, Tenesa *et al.* (2007) estimated $\hat{N}_e = 3{,}246$ for the YRI and $\hat{N}_e = 1{,}459$ for the CEU population. Overall, their estimates appeared to be much lower than the usually quoted value of 10,000 (Takahata, 1993; Harding *et al.*, 1997). Using a model-free method to estimate recombination rates however changed the estimate of N_e between $+33\%$ and -45%. Results for the YRI population indicated an ancestral population size of $\approx 7{,}000$ followed by expansion in the last 20,000 years ($\approx 1{,}000$

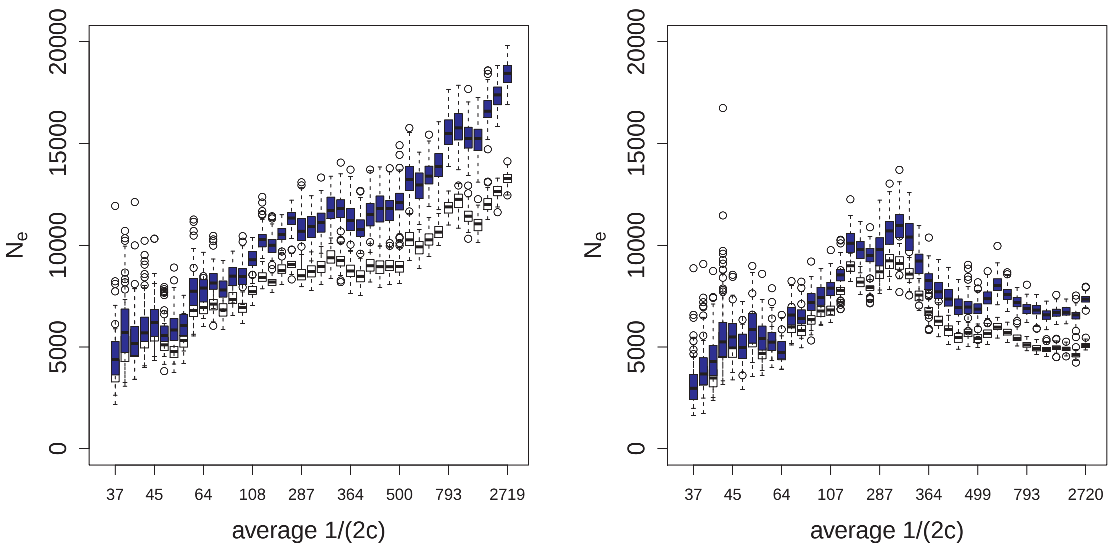

Figure 6.10: Estimates of N_e for the YRI and CEU population for different distributions of MAFs. The left (right) plot is for the YRI (CEU) population. Each boxplot represents the distribution of N_e estimates for a given bin of distance c between SNPs in Morgan. N_e was estimated based on the new recursion formula. Estimates of N_e were obtained for two different scenarios: (1) SNP positions were randomly sampled, *i.e.* from a skewed MAF distribution. (2) SNP positions were randomly sampled, so that the distribution of corresponding MAFs was uniform. The sampling process was replicated 100 times and N_e was estimated for each replicate, resulting in 100 N_e-estimates per scenario and per bin of c. Blue boxplots represent the distribution of N_e-estimates for scenario (1), black boxplots represent the distribution of N_e-estimates for scenario (2). Only SNPs on chromosome 22 were considered. Note that the scale of the x-axis is not linear. N_e estimates based on a given value of c correspond to the point in time "$\frac{1}{2c}$ generations ago" (Hayes *et al.*, 2003).

generations), whereas results for the CEU data supported recent dramatic population growth from an ancestral population size of $\approx$ 2,500.

Our study, based on all 22 autosomes, also indicates a population growth for the CEU population (*cf.* Figure 6.8), but no growth can be observed for the YRI population over the period 37,500 to 5,000 years ago. One possible reason for the discrepancy of the results may be that Tenesa *et al.* (2007) used a smaller SNP set and different recombination rates, due to the different methods of obtaining these rates. More importantly, the results were based on a release of phase I, whereas we used phase II data of the HapMap project. Additionally, Tenesa *et al.* (2007) excluded all SNPs with MAF < 0.05 for the LD estimation, whereas no filtering was performed in the present study.

Tenesa *et al.* (2007) also analyzed the effect of a possible ascertainment bias in the HapMap phase I data (Nielsen, 2000; Nielsen *et al.*, 2004; Clark *et al.*, 2005; Pe'er *et al.*, 2006) by simulating SNPs with complete ascertainment and simulating SNPs according to a uniform MAF distribution (still excluding SNPs with MAF < 0.05). They found that N_e estimates were biased downwards by 18% in the second scenario (which is supposed to mimick the MAF distribution of the HapMap phase I data) and concluded that this was also true for their estimated N_e of the HapMap population. As opposed to this, in the present study a skewed MAF distribution for SNPs occurring in the LD data was observed, as illustrated in Figure 6.9, which is due to the $\approx$ 2.1 million additional SNPs of phase II data compared to the phase I data, which comprised only 1.3 million SNPs (The International HapMap Consortium, 2007). However, the simulation results of Tenesa *et al.* (2007) qualitatively confirm our results of the previous section with respect to the influence of the MAF distribution on the N_e-estimates.

Park (2011) used HapMap phase III data to estimate N_e of the current human population based on two different methods, one using the deviation from linkage equilibrium (LE), the other based on the deviation from Hardy-Weinberg Equilibrium (HWE). For the YRI population, estimates fluctuated between 1,275 and 7,729, depending on the method, whereas estimates for the CEU population ranged between 1,331 and 10,437, illustrating again the great variability of results. Park (2011) argued that the HWE-based method presented the N_e of the current generation, whereas the LE-based method reflected values of "current and recent" generations. By considering the ratio of HWE- and LE-based estimates it was found that both populations experienced a recent population growth, which was more distinct for the CEU population.

McEvoy *et al.* (2011) also estimated N_e based on the HapMap phase III data set. It was found that the CEU population experienced a population growth from $N_e \approx 5{,}000$ to $N_e \approx 11{,}000$ between 800 and 240 generations ago, whereas N_e of the YRI population stayed fairly constant during this time. To decrease a possible ascertainment bias, McEvoy *et al.* (2011) only used SNPs that were segregating in all populations.

Our estimates of N_e are highly variable in size for recent points in time (< 50 generations ago, corresponding to $< 1{,}250$ years ago). A similar variability is reported in Tenesa *et al.* (2007), whereas no results are presented in McEvoy *et al.* (2011) for these points in time.

In summary, our results agree reasonably well with previously reported findings. Similar to the findings of McEvoy *et al.* (2011), we found the YRI population to be $\approx$ 2.5 times as large as the CEU population 1000 generations (25,000 years) ago, while the effective sizes of the two populations converge when considering more recent points in time. The observed increase of the European population between 15,000 to 10,000 years before present in the so-called neolithic expansion is in agreement with archaeological findings and coincides well with findings from independent sources, such as the estimations of Fu *et al.* (2012) based on mitochondrial genomes. However, it has been shown that the margin of fluctuation is large and that results should always be seen in relation to other existing studies.

6.6.5 Limitations

As indicated in section 6.4.2, results of the HapMap application have to be taken with caution, since the underlying recursion formula is not exact, contrary to what is assumed in the derivation of formula (6.8), which underlies formula (6.10). Complications arising from the non-exactness have neither been considered in previous studies so that our results are comparable to the results of other studies from this perspective. The non-exactness also pertains to the apparent dependency of the development of LD on allele frequencies, for which current recursion formulae do not account. All findings therefore have to be considered against the background of the implicit assumptions underlying eq. (6.10).

6.7 Discussion

6.7.1 The influence of SNP array designs on N_e-estimates

We showed in the simulation study as well as in the application to the HapMap data, that allele frequencies have a strong influence on the performance of the recursion formula and on the estimation of N_e. While the true MAF distribution in practical applications (*e.g.*, in sequencing studies) is usually skewed with a substantial excess of small MAF values, commercial SNP arrays are often constructed such that the MAF distribution is uniform, *i.e.*, alleles with extreme MAFs are systematically underrepresented (see *e.g.* (Matukumalli *et al.*, 2009) for the construction of a density SNP genotyping array for cattle). Hence, using LD values based on such a SNP array can have a major impact on estimates of N_e and may result in biased estimates of N_e compared to a situation in which the MAF distribution is not uniform. A similar bias may appear if a SNP array is constructed to reflect the allele frequency spectrum in one population but then is used to estimate N_e in other populations.

6.7.2 Analytic expression vs. approximate recursion formula

The proposed recursion formula for $\mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1})$ is still not completely unbiased, which can be seen in the boxplots of Figure 6.1. One possibility to reduce the bias is to use $\tilde{b} = (1+m)b$ instead of b where m is chosen such that

$$F - m = 0 \quad \Leftrightarrow \quad \mathbb{E}_{\mathbf{x}_{t_0}}(r^2_{t_0+1}) = ar^2_{t_0} + \tilde{b}.$$

The bias is in fact a function of the gamete frequencies, which can be seen when the upper plots of Figure 6.1 are created separately for different bins of P and S (results not shown). This leads back to the problem that an *exact* recursion formula will depend on the frequencies as well.

Even if it was possible to derive an exact recursion for a specific pair of loci with given allele frequencies, many pairs of loci are used to estimate the expected LD, and one would have to account not only for the allele frequencies of a single pair of loci but for the whole distribution of underlying frequencies, which is simply not feasible.

6.7.3 Obtaining the expected LD at equilibrium directly

One general way of obtaining the expected LD at equilibrium, without using any recursion formula, is to consider the matrix $\mathbf{P}$ of transition probabilities of the Markov chain and to calculate the limit of $\mathbf{P}^n$ for $n \to \infty$ to obtain the stationary distribution of gamete frequencies. From this, the expected LD at equilibrium can be calculated directly. However, a problem with this approach is that the size of $\mathbf{P}$ is $\binom{2N+3}{2N} \times \binom{2N+3}{2N}$ (there are $\binom{2N+3}{2N}$ possible states of the Markov chain (Karlin & McGregor, 1968)), which makes numerical calculation impossible even for moderately high N. We have already encountered this problem when analyzing the term $\frac{\pi^*\varepsilon}{b}$ relating to the non-exactness of the recursion formula.

As an alternative, one could also simulate the Markov chain of gamete frequencies directly (instead of calculating $\mathbf{P}^n$ for $n \to \infty$) and determine the stationary distribution based on the realization of the Markov chain. This could be done for different values of N and c, and the expected LD could be calculated based on the empirically obtained stationary distribution. Afterwards, the expected LD could be expressed as a function of N and c which then could be used for the estimation of N_e. This approach is left for future work.

6.7.4 Consequences of non-exact recursion formula

Previous studies are based on the implicit assumption that the underlying recursion formula is exact and the formula for the expected LD at equilibrium does not incorporate an error-term of the recursion formula. For $N < 16$, we showed that the error-term in the recursion formula can lead to a non-negligible deviance of expected LD at equilibrium. These analyses were restricted to small values of N due to the limited calculating capacity and only illustrate the effect qualitatively. It might be that the error-term becomes negligible for increasing N (and small values of c) so that results from previous studies remain reliable. The critical question remains, how reliable estimates of N_e are if they are based on a non-exact recursion formula, and further research is needed in this field.

6.7.5 Alternative approaches in the literature

In the literature, there are several other references with alternative approaches to derive formulae for N_e based on LD. Hayes *et al.* (2003), *e.g.*, state that N_e can be estimated based on the chromosome segment homozygosity (CSH) by using the relation $\text{CSH} = \frac{1}{4N_ec+1}$, which is the same formula one obtains based on Sved's recursion from eq. (6.2). However, in the course of their derivation it is assumed that the two considered loci behave independently, which is equivalent to Sved's questionable calculation of homozygosity at the second locus, given the alleles on the first locus are IBD.

Ohta & Kimura (1971) derived an approximate formula for the expected LD at equilibrium using the theory of diffusion process approximation. Here, the ratio of expectations instead of the expectation of the ratio is used to calculate the expected LD, resulting in

$$\mathbb{E}(r^2) \approx \frac{5 + 2N_ec}{11 + 26N_ec + 8(N_ec)^2}. \tag{6.15}$$

McVean (2007) demonstrated that the main difference between Sved's formula for the expected LD at equilibrium and eq. (6.15) is for small values of N_ec: While the expected LD based on Sved's formula approaches 1 for c tending to zero, eq. (6.15) tends to a value considerably less than 1. Comparing both estimates from Monte Carlo coalescent simulation, McVean (2007) found that neither of the formulae provides a particularly accurate prediction for the expected value of LD at equilibrium, unless rare variants (MAF < 0.1) are excluded. But eq. (6.15) still predicts the general shape of the decrease in LD with increasing N_ec, and it fits the simulated data better than Sved's formula when compared to a sliding average of simulated LD values.

Song & Song (2007) also used diffusion process approximation to derive a formula for the expected LD at equilibrium for a model with recurrent mutation, genetic drift and recombination. Note that the considered process in diffusion approximation is continuous in both time and space. Diffusion processes possess many nice properties which allow the calculation of certain expectations at stationarity with little effort (Song & Song, 2007). Song & Song (2007) were able to express the LD at equilibrium as infinite sum over certain terms, which in turn can be evaluated using the diffusion approximation, finally enabling a numerical calculation of the expected LD. One drawback of this approach is that the diffusion approximation is only valid for sufficiently large populations. For $Nc \to \infty$ Song & Song (2007) derived a closed-form expression for the expected LD at equilibrium which is the same as obtained by Ohta & Kimura (1971) for the expectation of the ratio.

Song & Song (2007) provide an approximate formula for the expected LD at equilibrium *directly*, without making a detour via a recursion formula, and despite the fact that derivations based on diffusion approximations are a priori valid for sufficiently large populations only, it might be that their approach constitutes a reasonable approximation even for small values of N. So far, we have not compared the validity of the proposed formula for the expected LD at equilibrium in this study with the results obtained by Song & Song (2007), nor have we compared our N_e-estimates with estimates based on coalescent approaches, as *e.g.* proposed by Li & Durbin (2011), who use both local homozygosity and LD information to estimate N_e for all past times via a "pairwise sequentially Markovian coalescent model". These comparisons are left for future research. "Direct" approaches as applied by Song & Song (2007) or Li & Durbin (2011) can easily incorporate mutation and recombination rates and do not rely on the formula of Hayes *et al.* (2003) for the determination of the corresponding time in point an N_e-estimates refers to, whose derivation was in fact based on the concept of "chromosome segment homozygosity" instead of LD.

6.7.6 Conclusions

In this study, we provide a theoretical basis for modeling the evolution of LD in a finite population using the framework of Markov chain theory with underlying multinomial distribution. On the basis of simulation studies, the HapMap application and the analyses of the state of equilibrium, we can summarize the following points:

The proposed recursion formula seems to provide a better overall fit than Sved's recursion formula. If N is large or if c is small, differences become marginal.

The performance of such recursion formulae heavily depends on allele frequencies, and LD is in general a function of the allele frequencies and the gamete frequencies. Hence, estimates of average LD in the population considerably depend on the MAF distribution of the SNP pairs used for estimation. Therefore, if the formula for the expected LD at equilibrium is used to estimate N_e, this estimate will also depend on the MAF distribution of the SNPs used to calculate the average value of LD for a given genetic distance c. This effect was illustrated in the HapMap application. It is important to keep in mind that SNP arrays used in certain populations not necessarily will reflect the allele frequency spectrum of this population, which can bias resulting estimates of N_e.

The currently used formulae for the expected LD at equilibrium are based on the assumption that the recursion formulae are correct. As shown in this study, one can theoretically account for the non-exactness of the recursion formula when deriving a formula for the expected LD at equilibrium, but exact solutions can only be obtained for small values of N due to computational limitations. For small values of N, the expected bias at equilibrium is non-negligible, and we have indicated how the effect can be approximated for larger values of N. Since the effect of the non-exactness might have a substantial influence on the resulting formula, as we have demonstrated in our empirical analyses, this might also be relevant for practical applications. In any case, the mathematical complexity of the problem studied warrants some caution when using the results. Estimates of N_e based on this method should always be confirmed by some other, independent method (*cf.* section 6.7.5) and possible sources of bias should critically be monitored.

Bibliography

Abney, M., McPeek, M. S. & Ober, C. (2000). Estimation of variance components of quantitative traits in inbred populations. *American Journal of Human Genetics* **66**(2):629–650.

Abramowitz, M. & Stegun, I. A. (1984). *Pocketbook of mathematical functions.* Verlag Harri Deutsch, Frankfurt/Main.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A. *et al.* (2000). The genome sequence of *Drosophila melanogaster. Science* **287**(5461):2185–2195.

Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J. & Zucchini, W. (2012). *R-package ff 2.2-5: memory-efficient storage of large data on disk and fast access functions.* URL `http://CRAN.R-project.org/package=ff`.

Aitken, A. C. (1934). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh, Section: A Mathematics* **55**:42–47.

Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M., Simianer, H. & Schön, C.-C. (2011). Genome-based prediction of testcross values in maize. *TAG Theoretical and Applied Genetics* **123**(2):339–350.

Allen, D. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics* **16**(1):125–127.

Anholt, R. (2010). Making scents of behavioural genetics: lessons from *Drosophila. Genetics Research* **92**(5–6):349–359.

Aulchenko, Y. S., Struchalin, M. V., Belonogova, N. M., Axenovich, T. I., Weedon, M. N., Hofman, A., Uitterlinden, A. G., Kayser, M., Oostra, B. A., van Duijn, C. M., Janssens, A. C. J. W. *et al.* (2009). Predicting human height by Victorian and genomic methods. *European Journal of Human Genetics* **17**:1070–1075.

Ayroles, J. F., Carbone, M. A., Stone, E. A., Jordan, K. W., Lyman, R. F., Magwire, M. M., Rollmann, S. M., Duncan, L. H., Lawrence, F., Anholt, R. R. H. & Mackay, T. F. C. (2009). Systems genetics of complex traits in *Drosophila melanogaster. Nature Genetics* **41**(3):299–307.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete multivariate analysis.* MIT Press, Cambridge, Massachusetts.

Bonate, P. L. (2006). *Pharmacokinetic-pharmocodynamic modeling and simulation.* Springer, New York.

Brookes, M. (2001). *Fly: The unsung hero in the history of genetics.* Harper-Collins, New York.

Browning, B. L. & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**(2):210–223.

Carlborg, O., Jacobsson, L., Åhgren, P., Siegel, P. & Andersson, L. (2006). Epistasis and the release of genetic variation during long-term selectiong. *Nature Genetics* **38**:418–420.

Chilès, J. P. & Delfiner, P. (1999). *Geostatistics. Modeling spatial uncertainty.* John Wiley & Sons, New York, Chichester.

Christensen, R. (1990). The equivalence of predictions from universal kriging and intrinsic random-function kriging. *Mathematical Geology* **22**(5):655–664.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* **15**:1496–1502.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**(6):859–882.

Cordell, J. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11**(20):2463–2468.

Cornelius, P. L. & Dudley, J. W. (1975). Theory of inbreeding and covariances between relatives under fullsib-mating in diploids. *Biometrics* **31**(1):169–187.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology* **22**(3):239–252.

Cressie, N. A. C. (1993). *Statistics for spatial data.* John Wiley & Sons, New York, Chichester.

Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**(1544):1241–1244.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B. & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* **185**(3):1021–1031.

de Boer, I. J. M. & Hoeschele, I. (1993). Genetic evaluation methods for populations with dominance and inbreeding. *TAG Theoretical and Applied Genetics* **86**(2–3):245–258.

de los Campos, G., Gianola, D. & Rosa, G. J. M. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science* **87**(6):1883–1887.

de los Campos, G., Gianola, D. & Allison, D. B. (2010*a*). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics* **11**(12):880–886.

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A. & Crossa, J. (2010*b*). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research* **92**(4):295–308.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J. & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* **179**(3):1503–1512.

Dempfle, L. (1982). *Zuchtwertschätzung beim Rind mit einer ausführlichen Darstellung der BLUP-Methode*. Fortschritte der Tierzüchtung und Züchtungskunde. Paul Parey Verlag, Hamburg, Berlin.

Eck, S. H., Benet-Pagès, A., Flisikowski, K., Meitinger, T., Fries, R. & Strom, T. M. (2009). Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. *Genome Biology* **10**(R82). doi:10.1186/gb-2009-10-8-r82.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**(397):171–185.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**(1):54–75.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**(5):e19379. doi:10.1371/journal.pone.0019379.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Pearson, Harlow, England.

Fernando, R. & Garrick, D. (2009). *GenSel: user manual of genomic selection related analyses*. Iowa State University, Ames, Iowa.

Fisher, R. A. (1918). The correlation between relatives under the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**:399–433.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85**(1):87–94.

Fiston-Lavier, A.-S., Singh, N. D., Lipatov, M. & Petrov, D. A. (2010). *Drosophila melanogaster* recombination rate calculator. *Gene* **463**(1–2):18–20.

Flint, J. & Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* **19**:723–733.

Flury, C., Tapio, M., Sonstegard, T., Drögemüller, C., Leeb, T., Simianer, H., Hanotte, O. & Rieder, S. (2010). Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *Journal of Animal Breeding and Genetics* **127**(5):339–347.

Fu, Q., Rudan, P., Pääbo, S. & Krause, J. (2012). Complete mitochondrial genomes reveal neolithic expansion into Europe. *PLoS ONE* **7**(3):e32473. doi:10.1371/journal.pone.0032473.

Gianola, D. & van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* **178**(4):2289–2303.

Gianola, D., Fernando, R. L. & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* **173**(3):1761–1776.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* **183**(1):347–363.

Gianola, D., Simianer, H. & Qanbari, S. (2010). A two-step method for detecting selection signatures using genetic markers. *Genetics Research* **92**(2):141–155.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. & Thompson, R. (2006). *ASReml user guide release 2.0*. VSN International Ltd., Hemel Hempstead, UK.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long-term response. *Genetica* **136**(2):245–257.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* **57**(298):369–375.

González, J. & Petrov, D. A. (2009). The adaptive role of transposable elements in the *Drosophila* genome. *Gene* **448**(2):124–133.

González-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M. & Avendaño, S. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* **178**(4):2305–2313.

González-Recio, O., Gianola, D., Rosa, G. J. M., Weigel, K. A. & Kranis, A. (2009). Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genetics Selection Evolution* **41**(3). doi:10.1186/1297-9686-41-3.

Grimmett, G. & Stirzaker, D. (2001). *Probability and Random Processes*. Oxford University Press, Oxford, 3rd edn.

Guttorp, P. & Gneiting, T. (2006). Studies in the history of probability and statistics XLIX: on the Matérn correlation family. *Biometrika* **93**(4):989–995.

Habier, D., Fernando, R. L. & Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**(4):2389–2397.

Handcock, M. S. & Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meterological fields. *Journal of the American Statistical Association* **89**(426):368–378.

Harbison, S. T., Yamamoto, A. H., Fanara, J. J., Norga, K. K. & Mackay, T. F. C. (2004). Quantitative trait loci affecting starvation resistance in *Drosophila melanogaster*. *Genetics* **166**(4):1807–1823.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S. & Clegg, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics* **60**(4):772–789.

Harville, D. A. (1984). Interpolation and estimation: discussion. In *Statistics: an appraisal* (eds. H. D. David & H. T. David), pp. 281–286. The Iowa State University Press, Ames, Iowa.

Hayes, B. J., Visscher, P. M., McPartland, H. C. & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective size. *Genome Research* **13**:635–643.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science* **92**(2):433–443.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J. & Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics* **6**(9):e1001139. doi:10.1371/journal.pgen.1001139.

Hedrick, P. W. (2011). *Genetics of populations*. Jones and Bartlett Publishers, Sudbury, Massachusetts, 4th edn.

Henderson, C. R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding* (eds. W. D. Hanson & H. F. Robinson), vol. 141-163. Publication 982, National Academy of Sciences, National Research Council, Washington, D.C.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science* **1973**:10–41.

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. University of Guelph, Guelph, Canada.

Henderson, C. R., Kempthorne, O., Searle, S. R. & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to cullings. *Biometrics* **15**(2):192–218.

Henderson, H. V. & Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review* **23**(1):53–60.

Hill, W. G. (1977). Correlation of gene frequencies between neutral linked genes in finite populations. *Theoretical Population Biology* **11**(2):239–248.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics* **38**(6):226–231.

Hill, W. G. & Weir, B. S. (1994). Maximum likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics* **54**(4):704–714.

Hill, W. G., Goddard, M. E. & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* **4**(2):e1000008. doi:10.1371/journal.pgen.1000008.

Hoeschele, I. (1991). Additive and nonadditive genetic variance in female fertility of Holsteins. *Journal of Dairy Science* **74**(5):1743–1752.

Holland, J. B. (2001). Epistasis and plant breeding. *Plant Breeding Reviews* **21**:27–92.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **15**(2):193–232.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**(3):299–314.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

Jordan, K. W., Carbone, M. A., Yamamoto, A., Morgan, T. J. & Mackay, T. F. C. (2007). Quantitative genomics of locomotor behavior in *Drosophila melanogaster*. *Genome Biology* **8**(R172). doi:10.1186/gb-2007-8-8-r172.

Karlin, S. & McGregor, J. (1968). Rates and probabilities of fixation for two locus random mating finite populations without selection. *Genetics* **58**(1):141–159.

Kitanidis, P. K. (1993). Generalized covariance functions in estimation. *Mathematical Geology* **25**(5):525–540.

Krengel, U. (2005). *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Vieweg, Wiesbaden, 8th edn.

Krige, D. G. (1951). A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master's thesis, University of Witwatersrand, Johannesburg, South Africa.

Kroymann, J. & Mitchell-Olds, T. (2005). Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435**:95–98.

Kusakabe, S., Yamaguchi, Y., Baba, H. & Mukai, T. (2000). The genetic structure of the Raleigh natural population of *Drosophila melanogaster* revisited. *Genetics* **154**(2):679–685.

Kwee, L. C., Liu, D., Lin, X., Gosh, D. & Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics* **82**(2):386–397.

Legarra, A., Robert-Granié, C., Manfredi, E. & Elsen, J.-M. (2008). Performance of genomic selection in mice. *Genetics* **180**(1):611–618.

Li, H. & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**:493–496.

Lifshits, M. A. (1995). *Gaussian random functions.* Kluwer, Dordrecht.

Littler, R. A. (1973). Linkage disequilibrium in two-locus, finite, random mating models without selection or mutation. *Theoretical Population Biology* **4**(3):259–275.

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. A., Kranis, A. & González-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genetics Research* **92**(3):209–225.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E. & Jannink, J. L. (2011). Genomic selection in plant breeding: knowledge and prospects. *Advances in Agronomy* **110**:77–123.

Lynch, M. & Walsh, B. (1998). *Genetics and analysis of quantitative traits.* Sinauer Associates, Inc., Sunderland, Massachusetts.

Mackay, T. F. C. (2004). The genetic architecture of quantitative traits: lessons from *Drosophila. Current Opinion in Genetics & Development* **14**(3):253–257.

Mackay, T. F. C., Stone, E. A. & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**:565–577.

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F. *et al.* (2012). The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**:173–178.

Maher, B. (2008). Personal genomes: the case of the missing heritability. *Nature* **456**:18–21.

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B. & de los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. *PLoS Genetics* **7**(4):e1002051. doi:10.1371/journal.pgen.1002051.

Matérn, B. (1960). Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden fran Statens Skogsforskningsinstitut* **49**(5):1–144.

Matheron, G. (1962). *Traité de geostatisque appliquée, Vol. I. Memoires du Bureau de Recherches Géologiques et Miniéres, no. 14.* Editions Technip, Paris.

Matheron, G. (1963). *Traité de geostatistique appliquée, Vol. II: Le krigeage. Memoires du Bureau de Recherches Géologiques et Miniéres, no. 24.* Editions Bureau de Recherche Géologiques et Miniéres, Paris.

Matheron, G. (1971). *The theory of regionalized random variables and its applications.* Ećole des Mines, Fountainbleau.

Matheron, G. (1989). *Estimating and choosing – an essay on probability in practice.* Springer, Berlin, Heidelberg.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J., Allan, M. F. M., Heaton, M. P., O'Connel, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. & Van Tassell, C. P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* **4**(4):e5350. doi:10.1371/journal.pone.0005350.

McEvoy, B. P., Powell, J. E., Goddard, M. E. & Visscher, P. M. (2011). Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Research* **21**:821–829.

McVean, G. (2007). Linkage disequilibrium, recombination and selection. In *Handbook of statistical genetics*, vol. 2, chap. 27, pp. 909–944. Wiley & Sons, Ltd, 3rd edn.

Meuwissen, T. & Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* **185**(2):623–631.

Meuwissen, T. H. E. (2009). Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genetics Selection Evolution* **41**(35). doi:10.1186/1297-9686-41-35.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4):1819–1829.

Mrode, R. A. (2005). *Linear models for the prediction of animal breeding values.* CABI Publishing, Oxfordshire, UK, 2nd edn.

Myers, D. E. (1992). Kriging, cokriging, radial basis functions and the role of positive definiteness. *Computers and Mathematics with Applications* **24**(12):139–148.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**(2):931–942.

Nielsen, R., Hubisz, M. J. & Clark, A. G. (2004). Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* **168**(4):2373–2382.

Norris, J. R. (1997). *Markov chains.* Cambridge University Press, Cambridge, United Kingdom.

Nychka, D. W. (2000). In *Smoothing and regression: approaches, computation, and application* (ed. M. G. Schimek), chap. Spatial process estimates as smoothers, pp. 393–424. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Ober, U., Erbe, E., Long, N., Porcu, E., Schlather, M. & Simianer, H. (2011). Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics* **188**(3):695–708.

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., Stricker, C., Gianola, D., Schlather, M., Mackay, T. F. C. & Simianer, H. (2012*a*). Using whole genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genetics* **8**(5):e1002685. doi:10.1371/journal.pgen.1002685.

Ober, U., Magwire, M., Huang, W., Schlather, M., Simianer, H. & Mackay, T. F. C. (2012*b*). Complex genetic architecture of a *Drosophila* fitness trait. Unpublished manuscript. In revision for *PLoS Genetics*.

Ober, U., Malinowski, A., Schlather, M. & Simianer, H. (2012*c*). The expected linkage disequilibrium in finite populations revisited. Unpublished manuscript.

Ohta, T. & Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* **68**(4):571–580.

Park, L. (2011). Effective population size of current human population. *Genetics Research* **93**(2):105–114.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**(3):545–554.

Pe'er, I., Chretien, Y. R., de Bakker, P. I. W., Barrett, J. C., Daly, M. J. & Altshuler, D. M. (2006). Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *American Journal of Human Genetics* **78**(4):588–603.

Phillips, C. P. (2008). Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9**:855–867.

Piepho, H. P. (2009). Ridge regression and extensions for genome-wide selection in maize. *Crop Science* **49**(4):1165–1176.

Piepho, H. P., Möhring, J., Melchinger, A. E. & Buchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* **161**:209–228.

Pimentel, E., Erbe, M., König, S. & Simianer, H. (2011). Genome partitioning of genetic variation for milk production and composition traits in Holstein cattle. *Frontiers in Genetics* **2**(19). doi:10.3389/fgene.2011.00019.

Qanbari, S., Pimentel, E., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R. & Simianer, H. (2010). The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics* **41**(4):346–356.

R Development Core Team (2012). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Rajchman, A. (1932). Zaostrzone prawo wielkich liczb. *Mathesis Polska* **6**:145–161.

Ranade, K., Chang, M. S., Ting, C. T., Pei, D., Hsiao, C. F., Olivier, M., Pesich, R., Hebert, J., Chen, Y. I. & Dzau, V. J. (2001). High-throughput genotyping with single nucleotide polymorphisms. *Genome Research* **11**:1262–1268.

Reif, J. C., Melchinger, A. E. & Frisch, M. (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science* **45**(1):1–7.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M. & Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *PNAS* **98**(20):11479–11484.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**(1):15–51.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric regression.* Cambridge University Press, New York.

SAS Institute (2002-2008). *SAS Software, version 9.2.* SAS Institute Inc., Cary, NC, USA.

Schaid, D. J. (2010*a*). Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Human Heredity* **70**(2):109–131.

Schaid, D. J. (2010*b*). Similarity and kernel methods II: methods for genomic information. *Human Heredity* **70**(2):132–140.

Scheuerer, M. (2011). An alternative procedure for selecting a good value for the parameter *c* in RBF-interpolation. *Advances in Computational Mathematics* **34**(1):105–126.

Schlather, M. (2001–2009). RandomFields: contributed extension package to R for the simulation of Gaussian and max-stable random fields. `http://cran.r-project.org`; Version 2.0.23 available at `http://www.stochastik.math.uni-goettingen.de/~schlather/genoKriging`.

Schlather, M. & Tawn, J. A. (2003). A dependence measure for multivariate and spatial extreme values: properties and inference. *Biometrika* **90**(1):139–156.

Schölkopf, B., Herbrich, R., Smola, A. J. & Williamson, R. C. (2001). A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory*, vol. 2111 of *Lecture Notes in Computer Science*, pp. 416–426. Springer.

Schölkopf, B., Tsuda, K. & Vert, J. P. (eds.) (2004). *Kernel methods in computational biology.* MIT Press, Massachusetts.

Schön, C.-C., Utz, H. F., Groh, S., Truberg, B., Openshaw, S. & Melchinger, A. E. (2004). Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics* **167**(1):485–498.

Schulz-Streeck, T. & Piepho, H. P. (2010). Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. In *BMC Proceedings 2010*, vol. 4(Suppl 1):S8. 13th European workshop on QTL mapping and marker assisted selection, Wageningen, The Netherlands.

Sinha, H., David, L., Pascon, R., Clauder-Münster, S., Krishnakumar, S., Nguyen, M., Shi, G., Dean, J., W, D. R., Oefner, P. J., McCusker, J. H. *et al.* (2008). Sequential elimination of major-effect contributors identifies additional quantitative trait loci conditioning high-temperature growth in yeast. *Genetics* **180**(3):1661–1670.

Soetaert, K. (2011). *R-package diagram 1.6: functions for visualising simple graphs (networks), plotting flow diagrams.* URL `http://cran.r-project.org/package=diagram`.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A. & Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *Journal of Animal Science* **86**(10):2447–2454.

Song, Y. S. & Song, J. S. (2007). Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. *Theoretical Population Biology* **71**(1):49–60.

Stam, L. F. & Laurie, C. C. (1996). Molecular dissection of a major gene effect on a quantitative trait: the level of alcohol dehydrogenase expression in *Drosophila melanogaster*. *Genetics* **144**(4):1559–1564.

Stein, M. L. (1999). *Interpolation of spatial data.* Springer, Heidelberg, New York.

Steinmetz, L. M., Sinha, J., Richards, D. R., Spiegelman, J. I., Oefner, P. J., McCusker, J. H. & Davis, R. W. (2002). Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**:326–330.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **36**(2):111–147.

Stone, M. (1977). An aymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **39**(1):44–47.

Sturtevant, A. H. (1931). Known and probable inverted sections of the autosomes of *Drosophila melanogaster. Carnegie Inst Washington Pub* **421**:1–27.

Suykens, J. A. K., Gestel, T. V., de Brabanter, J., de Moor, B. & Vandewalle, J. (2002). *Least squares support vector machines.* World Scientific, Singapore.

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

Sved, J. A. (2008). Linkage disequilibrium and its expectation in human populations. *Twin Research and Human Genetics* **12**(1):35–43.

Sved, J. A. (2009). Correlation measures for linkage disequilibrium within and between populations. *Genetics Research* **91**(3):183–192.

Sved, J. A. & Feldmann, M. W. (1973). Correlation and probability methods for one and two loci. *Theoretical Population Biology* **4**(1):129–132.

Swarup, S., Harbison, S. T., Hahn, L. E., Morozova, T. V., Yamamoto, A., Mackay, T. F. C. & Anholt, R. R. H. (2012). Extensive epistasis for olfactory behavior, sleep and waking activity in *D. melanogaster*. *Genetics Research* **94**(1):9–20.

Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution* **10**(1):2–22.

Tenesa, A., Navarro, P. & Hayes, B. J. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**:520–526.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**:1061–1073.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**(6814):796–815.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**:789–796.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**:1299–1320.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**:851–861.

Urbanek, S. (2011). *R-package multicore 0.1-7: parallel processing of R code on machines with multiple cores or CPUs*. URL `http://CRAN.R-project.org/package=multicore`.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**(11):4414–4423.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. & Schenkel, F. S. (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**(1):16–24.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* **2**(3):e41. doi:10.1371/journal.pgen.0020041.

Wackernagel, H. (2003). *Multivariate geostatistics*. Springer, Berlin, Heidelberg, 3rd edn.

Wang, K., Li, M. & Bucan, M. (2007). Pathway-based approach for analysis of genomewide association studies. *American Journal of Human Genetics* **81**(6):1278–1283.

Webster, J., Welham, S. J., Potts, J. M. & Oliver, M. A. (2006). Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. *Computer & Geosciences* **32**(9):1320–1333.

Weir, B. S. & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**(2):477–488.

Whittaker, J. C., Thompson, R. & Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetical Research* **75**(2):249–252.

Wray, N. R., Goddard, M. E. & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17**:1520–1528.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist* **56**(645):330–338.

Wright, S. (1984). *Evolution and the genetics of populations.* University of Chicago Press, Chicago.

Yang, H. C., Hsieh, H. Y. & Fann, C. S. J. (2008). Kernel-based association test. *Genetics* **179**(2):1057–1068.

Young, D., Benaglia, T., Chauveau, D., Elmore, R., Hettmansperger, T., Hunter, D., Thomas, H. & Xuan, F. (2010). *R-package mixtools 0.4.4: tools for analyzing finite mixture models.* URL `http://cran.r-project.org/package=mixtools`.

Zou, F., Huang, H., Lee, S. & Hoeschele, I. (2010). Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene-environment interaction. *Genetics* **186**(1):385–394.

Zuk, O., Hechtera, E., Sunyaeva, S. R. & Lander, E. S. (2012). The mystery of missing heritability: genetic interactions create phantom heritability. *PNAS* **109**(4):1193–1198.

A Supplementary Tables

Table S1: Results of variance component estimation for starvation resistance using ASReml. Different linear models for individual trait records were investigated.

Starvation	$\hat{\sigma}^2_{\text{line}}$	$\hat{\sigma}^2_{\text{sex}*\text{line}}$	$\hat{\sigma}^{2,1}_{\text{rep(sex}*\text{line)}}$	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_{g \times g}$	$\hat{\sigma}^2_{\text{residual}}$	$\ln(L)^2$	$\hat{H}^{2,3}_{\text{Model 1}}$	$\hat{h}^{2,4}_{\text{Model 2/3}}$
Model 1	87.97[5] (12.74)	39.65 (5.06)	17.00 (1.01)			88.00 (1.03)	-46097.51	0.59 (0.02)	
Model 2	0.00	39.35 (4.99)	17.00 (1.01)	43.13 (6.32)		88.00 (1.03)	-46093.63**,[6]		0.25 (0.03)
Model 3	0.00	39.35 (4.98)	17.00 (1.01)	43.13 (6.32)	0.00	88.00 (1.03)	-46093.63		0.25 (0.03)
Model 1 f[7]	167.62 (19.63)		18.22 (1.64)			113.92 (1.88)	-24174.10	0.60 (0.03)	
Model 2 f	9.07 (28.35)		18.22 (1.64)	78.24 (18.02)		113.92 (1.88)	-24170.31**		0.39 (0.11)
Model 3 f	9.08 (28.34)		18.22 (1.64)	78.24 (18.02)	0.00	113.92 (1.88)	-24170.31		0.39 (0.11)
Model 1 m[8]	87.57 (10.40)		15.73 (1.21)			61.89 (1.02)	-21617.19	0.59 (0.03)	
Model 2 m	5.23 (15.49)		15.73 (1.21)	40.51 (9.68)		61.89 (1.02)	-21613.34**		0.38 (0.11)
Model 3 m	5.23 (15.49)		15.73 (1.21)	40.51 (9.68)	0.00	61.89 (1.02)	-21613.34		0.38 (0.11)

[1] or “rep(line)” if factor sex is not included

[2] loglikelihood

[3] broad-sense heritability, standard errors in parentheses

[4] narrow-sense heritability, standard errors in parentheses

[5] estimated variance components, standard errors in parentheses

[6] The superscript ** indicates the 1%-significance of Model 2 compared to the Model 1 without g-component based on a likelihood ratio test.

[7] Only measurements of female *Drosophila* were used.

[8] Only measurements of male *Drosophila* were used.

Table S2: Results of variance component estimation for startle response using ASReml. Different linear models for individual trait records were investigated.

Startle	$\hat{\sigma}^2_{\text{line}}$	$\hat{\sigma}^2_{\text{sex}*\text{line}}$	$\hat{\sigma}^{2,1}_{\text{rep(sex}*\text{line)}}$	$\hat{\sigma}^2_g$	$\hat{\sigma}^2_{g\times g}$	$\hat{\sigma}^2_{\text{residual}}$	$\ln(L)^2$	$\hat{H}^{2,3}_{\text{Model 1}}$	$\hat{h}^{2,4}_{\text{Model 2/3}}$
Model 1	33.49[5] (4.38)	0.00	17.81 (1.26)			25.70 (0.33)	-27320.93	0.57 (0.03)	
Model 2	0.00	0.00	17.79 (1.26)	16.45 (2.17)		25.70 (0.33)	-27317.45**,[6]		0.39 (0.03)
Model 3	0.00	0.00	17.78 (1.26)	12.98 (11.47)	1.68 (5.52)	25.70 (0.33)	-27317.41		0.32 (0.24)
Model 1 f[7]	24.07 (4.72)		27.96 (3.33)			25.26 (0.46)	-13746.53	0.49 (0.05)	
Model 2 f	0.00		27.61 (3.25)	11.89 (2.34)		25.26 (0.46)	-13743.71*		0.32 (0.04)
Model 3 f	0.00		27.61 (3.25)	11.89 (2.34)	0.00	25.26 (0.46)	-13743.71		0.32 (0.04)
Model 1 m[8]	27.26 (4.74)		23.23 (2.81)			26.13 (0.48)	-13682.01	0.51 (0.04)	
Model 2 m	0.00		23.03 (2.77)	13.34 (2.35)		26.13 (0.48)	-13678.87*		0.34 (0.04)
Model 3 m	0.00		23.03 (2.77)	13.34 (2.35)	0.00	26.13 (0.48)	-13678.87		0.34 (0.04)

[1] or "rep(line)" if factor sex is not included

[2] loglikelihood

[3] broad-sense heritability, standard errors in parentheses

[4] narrow-sense heritability, standard errors in parentheses

[5] estimated variance components, standard errors in parentheses

[6] The superscripts $*$ and $**$ indicate the 5%- and 1%-significance of Model 2 compared to the Model 1 without g-component based on a likelihood ratio test.

[7] Only measurements of female *Drosophila* were used.

[8] Only measurements of male *Drosophila* were used.

Table S3: Results of variance component estimation for chill coma recovery using ASReml. Different linear models for individual trait records were investigated.

Coma[1]	$\hat{\sigma}^2_{\text{line}}$	$\hat{\sigma}^2_{\text{line*sex}}$	$\hat{\sigma}^{2,}_{\text{rep(line*sex)}}$[2]	$\hat{\sigma}^2_{g}$	$\hat{\sigma}^2_{g\times g}$	$\hat{\sigma}^2_{\text{residual}}$	$\log(L)$[3]	$\hat{H}^{2,}_{\text{Model 1}}$[4]	$\hat{h}^{2,}_{\text{Model 2/3}}$[5]
Model 1	28.60[6] (3.60)	1.92 (0.52)	3.56 (0.37)			48.55 (0.40)	-72763.03	0.39 (0.03)	
Model 2	28.60 (3.60)	1.92 (0.52)	3.56 (0.37)	0.00		48.55 (0.40)	-72763.03		0.00
Model 3	0.00	1.92 (0.52)	3.56 (0.37)	0.00	7.21 (0.91)	48.55 (0.40)	-72762.97		0.00
Model 1 f[7]	34.07 (4.17)		2.01 (0.36)			52.64 (0.62)	-37175.16	0.39 (0.03)	
Model 2 f	29.07 (14.20)		2.01 (0.36)	2.48 (6.92)		52.64 (0.62)	-37175.07		0.03 (0.08)
Model 3 f	0.00		2.01 (0.36)	0.99 (8.54)	8.10 (4.36)	52.64 (0.62)	-37175.09		0.02 (0.14)
Model 1 m[8]	26.94 (3.53)		5.14 (0.70)			44.43 (0.52)	-35631.82	0.38 (0.03)	
Model 2 m	26.94 (3.53)		5.14 (0.70)	0.00		44.43 (0.52)	-35631.82		0.00
Model 3 m	0.00		5.14 (0.70)	0.00	6.79 (0.89)	44.43 (0.52)	-35631.80		0.00

[1] without outlier
[2] or “rep(line)” if factor sex is not included
[3] loglikelihood
[4] broad-sense heritability, standard errors in parentheses
[5] narrow-sense heritability, standard errors in parentheses
[6] estimated variance components, standard errors in parentheses
[7] Only measurements of female *Drosophila* were used.
[8] Only measurements of male *Drosophila* were used.

Table S4: Clustering based on the variogram analyses using the chill coma recovery data. Listed are the IDs of the 25 lines belonging to C_2.

Line ID				
26	28	93	105	138
161	233	313	350	358
359	377	383	386	426
492	595	642	646	721
776	786	837	852	894

Table S5: Clustering based on the bimodality of the phenotypic distribution of chill coma recovery. The table contains the line IDs and the corresponding posterior probabilities to belong to the two distributions of Pop1 and Pop2.

line ID	prob1*	prob2**	line ID	prob1	prob2
101	0.01	0.99	502	0.92	0.08
105	0.84	0.16	508	0	1
109	0.79	0.21	517	0.79	0.21
129	0.05	0.95	531	0.6	0.4
153	0.79	0.21	535	0.92	0.08
136	0.47	0.53	555	0.84	0.16
138	0.2	0.8	563	0.92	0.08
142	0.93	0.07	57	0	1
149	0.79	0.21	589	0	1
158	0.92	0.08	59	0.88	0.12
161	0	1	591	0.88	0.12
176	0	1	595	0.93	0.07
177	0.66	0.34	639	0	1
181	0	1	642	0.84	0.16
195	0.92	0.08	646	0	1
208	0.88	0.12	69	0.91	0.09
21	0.05	0.95	703	0	1
217	0.93	0.07	705	0	1
227	0.71	0.29	707	0.02	0.98
228	0.93	0.07	712	0.91	0.09
229	0.9	0.1	714	0.93	0.07
233	0.92	0.08	716	0.84	0.16
235	0.47	0.53	721	0.71	0.29
237	0.66	0.34	727	0	1
239	0	1	73	0	1
26	0.79	0.21	730	0.82	0.18
28	0.79	0.21	732	0.88	0.12
280	0.9	0.1	737	0.93	0.07
287	0.9	0.1	738	0.71	0.29
309	0.9	0.1	75	0	1
313	0	1	757	0.75	0.25
310	0.88	0.12	761	0.71	0.29
318	0.79	0.21	765	0.71	0.29
325	0.9	0.1	774	0.92	0.08
332	0.92	0.08	776	0.05	0.95
338	0.93	0.07	783	0.82	0.18
350	0.93	0.07	786	0.6	0.4
352	0.92	0.08	787	0.15	0.85
356	0.7	0.3	790	0	1
357	0.91	0.09	796	0.01	0.99
358	0.93	0.07	799	0.32	0.68
359	0.92	0.08	801	0.9	0.1
362	0.79	0.21	802	0.93	0.07
365	0.91	0.09	804	0	1
367	0.92	0.08	805	0.9	0.1
370	0.71	0.29	808	0.71	0.29
371	0.88	0.12	810	0.9	0.1
373	0.93	0.07	812	0	1
374	0.84	0.16	818	0.05	0.95
375	0.93	0.07	820	0.92	0.08
377	0	1	822	0.02	0.98
379	0.9	0.1	83	0.08	0.92
38	0.6	0.4	832	0.6	0.4
380	0.93	0.07	837	0.02	0.98
381	0.93	0.07	85	0.92	0.08
383	0.92	0.08	852	0.93	0.07
386	0.91	0.09	855	0	1
391	0.92	0.08	857	0.93	0.07
392	0.39	0.61	859	0.9	0.1
399	0.88	0.12	861	0	1
409	0.93	0.07	88	0	1
41	0	1	882	0	1
42	0	1	884	0	1
426	0.92	0.08	887	0.9	0.1
427	0.9	0.1	890	0.32	0.68
437	0	1	892	0	1
440	0.91	0.09	94	0.88	0.12
441	0.9	0.1	897	0	1
443	0.93	0.07	907	0	1
45	0.9	0.1	908	0	1
461	0.9	0.1	91	0.92	0.08
49	0.92	0.08	911	0.92	0.08
491	0.92	0.08	93	0.91	0.09
492	0.84	0.16			

[*] Posterior probability to belong to Pop1 based on the EM-algorithm of the R-package “mixtools”.

[**] Posterior probability to belong to Pop2.

B Supplementary Figures

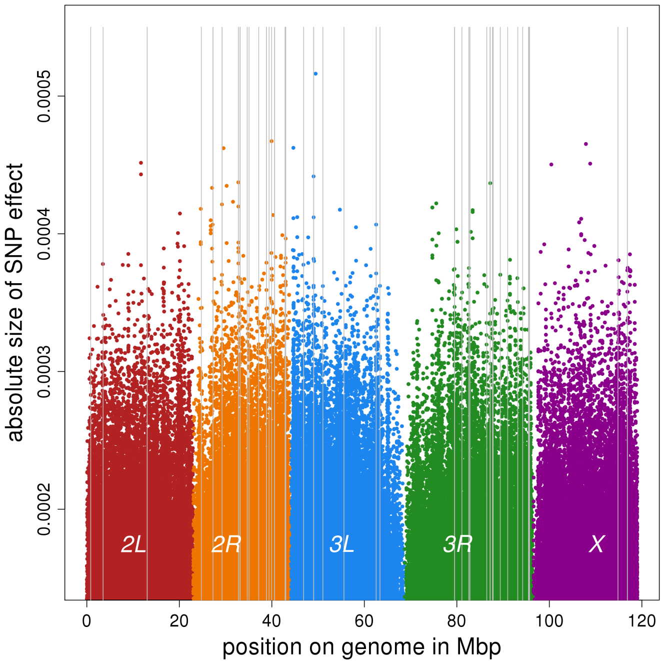

Figure S1: Manhattan plot of the estimated SNP effects for starvation resistance for different chromosomes. The SNP effects were estimated using the GBLUP approach and sex-averaged phenotypic values of 157 lines. Vertical lines indicate the 115 significant SNP positions according to the GWAS of Mackay *et al.* (2012) using sex-pooled records.

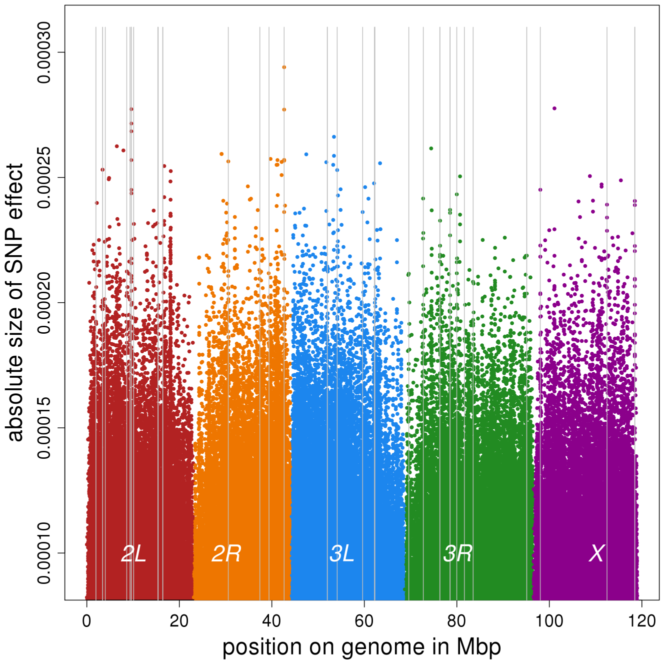

Figure S2: Manhattan plot of the estimated SNP effects for startle response for different chromosomes. The SNP effects were estimated using the GBLUP approach and sex-averaged phenotypic values of 155 lines. Vertical lines indicate the 75 significant SNP positions according to the GWAS of Mackay *et al.* (2012) using sex-pooled records.

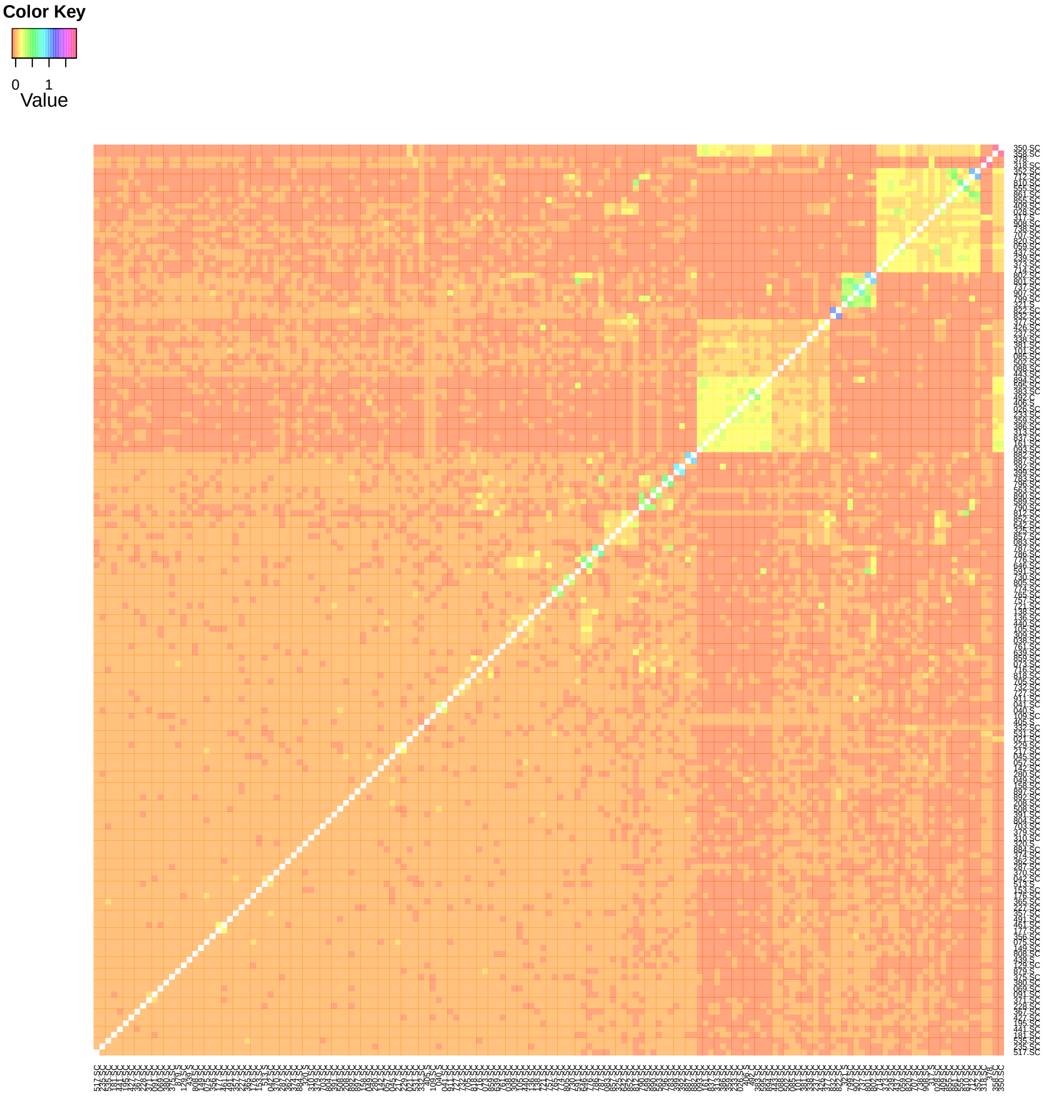

Figure S3: Heatmap of the genomic relationship matrix G. The genomic relationship matrix **G** was calculated according to VanRaden (2008) using 157 lines and 2.5 million SNPs. The “S” (“C”) after the line-ID indicates that the line belongs to the set of lines for which phenotypic records for startle response (chill coma recovery) were also available.

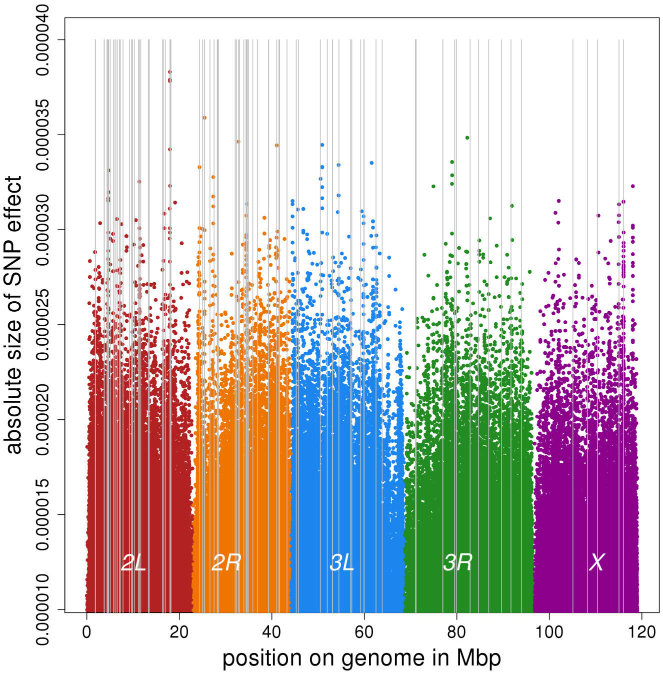

Figure S4: Manhattan plot of the estimated SNP effects for chill coma recovery for different chromosomes. The SNP effects were estimated with the GBLUP approach. As phenotypic values of the 147 lines, only female records were used. Vertical lines indicate the 145 most significant SNP positions according to the GWAS of Mackay *et al.* (2012) using female records only.

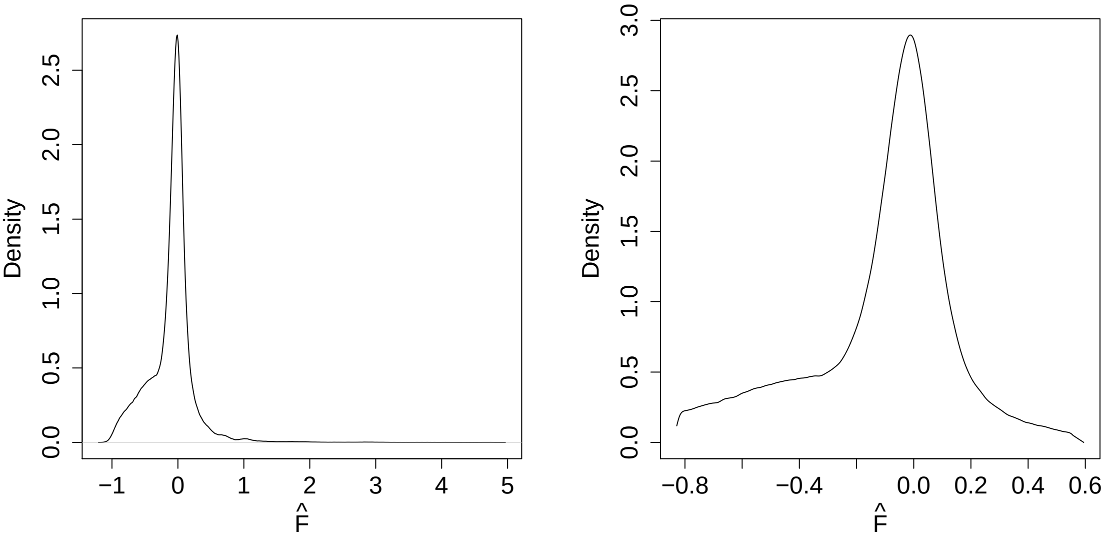

Figure S5: Density function of $\hat{F}$. The left plot shows the density for all obtained values of $\hat{F}$. To obtain the right density plot, values of $\hat{F}$ below the 2.5% quantile and above the 97.5% quantile of the distribution of $\hat{F}$ were excluded.

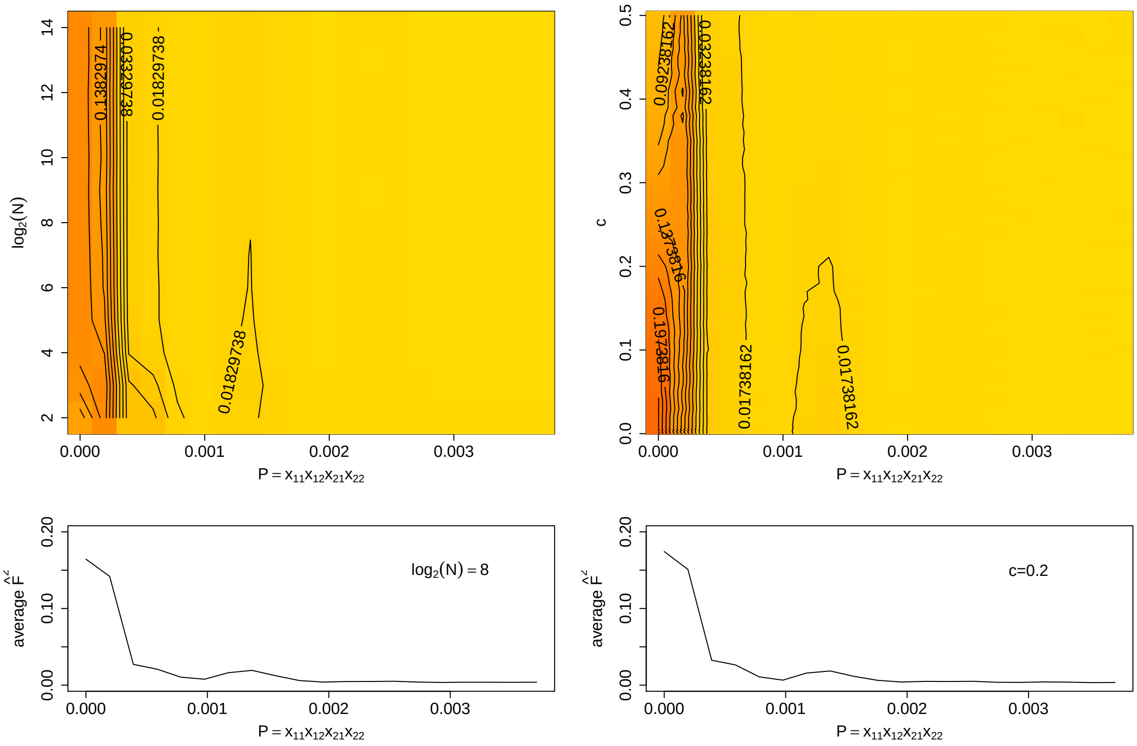

Figure S6: Contourplot of the average values of $\hat{F}^2$. Left plots: For a given combination of $(P, \log_2(N))$ the values of $\hat{F}^2$ were averaged over all possible values of c and $\mathbf{x}_{t_0}$ (upper plot). The lower plot illustrates the average value of $\hat{F}^2$ as a function of P for $\log_2(N) = 8$. Right plots: For a given combination of (P,c) the values of $\hat{F}^2$ were averaged over all possible values of $\log_2(N)$ and $\mathbf{x}_{t_0}$ (upper plot). The lower plot illustrates the average value of $\hat{F}^2$ as a function of P for $c = 0.2$. Contourplots were created after excluding the extreme 2.5% quantiles of $\hat{F}$.

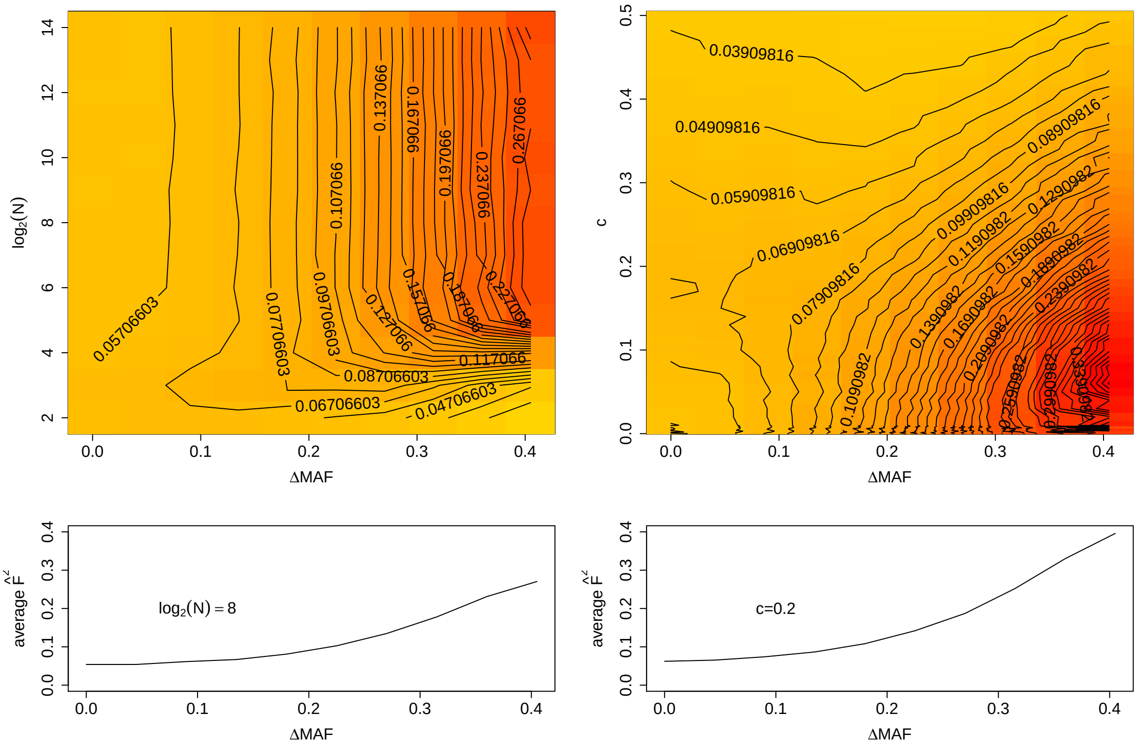

Figure S7: Contourplot of the average values of $\hat{F}^2$. Left plots: For a given combination of $(\Delta\text{MAF}, \log_2(N))$ the values of $\hat{F}^2$ were averaged over all possible values of c and $\mathbf{x}_{t_0}$ (upper plot). The lower plot illustrates the average value of $\hat{F}^2$ as a function of P for $\log_2(N) = 8$. Right plots: For a given combination of (ΔMAF,c) the values of $\hat{F}^2$ were averaged over all possible values of $\log_2(N)$ and $\mathbf{x}_{t_0}$ (upper plot). The lower plot illustrates the average value of $\hat{F}^2$ as a function of ΔMAF for $c = 0.2$. Contourplots were created after excluding the extreme 2.5% quantiles of $\hat{F}$.

C Curriculum Vitae

Ulrike Ober born July 27, 1985, in Wiesbaden, Germany

Education

10/2011–11/2011	Research stay at the North Carolina State University, Raleigh, USA (with Prof. T. F. C. Mackay)
08/2010–12/2010	Research stay at the University of Wisconsin-Madison, Madison, USA (with Prof. D. Gianola)
since 10/2010	Associate member of the Deutsche Forschungsgemeinschaft (DFG) Research Training Group “Scaling Problems in Statistics” (GRK 1644)
since 10/2009	Member of the PhD program “Applied Statistics and Empirical Methods” at the Centre for Statistics in Göttingen
since 10/2009	**PhD position** at the Department for Animal Sciences, Animal Breeding and Genetics Group, Georg-August-Universität Göttingen; supervised by Prof. M. Schlather and Prof. H. Simianer
10/2009	**Diploma in Mathematics**
04/2007	Intermediate Diploma in Mathematics
10/2005–10/2009	Studies of Mathematics; minor: Business Studies; Georg-August-Universität Göttingen
06/2005	**Academic High School Diploma** (German “Abitur”)
07/1992–06/2005	School education

Scholarship

10/2005–10/2009	Scholarship holder of the “Studienstiftung des deutschen Volkes”

Publications

in preparation	**U. Ober, A. Malinowski, M. Schlather, H. Simianer** *The Expected Linkage Disequilibrium in Finite Populations Revisited.* Unpublished manuscript.
	U. Ober*, M. Magwire*, W. Huang, M. Schlather, H. Simianer, T. F. C. Mackay *Complex Genetic Architecture of a* Drosophila *Fitness Trait.* Unpublished manuscript. In revision for PLoS Genetics. *equal contribution
2012	**U. Ober, J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, H. Simianer** *Using Whole Genome Sequence Data to Predict Quantitative Trait Phenotypes in* Drosophila melanogaster*.* PLoS Genetics 8(5):e1002685, doi:10.1371/journal.pgen.1002685
2011	**U. Ober, M. Erbe, N. Long, E. Porcu, M. Schlather, H. Simianer** *Predicting Genetic Values: a Kernel-Based Best Linear Unbiased Prediction with Genomic Data.* Genetics 188(3), 695–708
2010	**U. Ober, M. Erbe, M. Schlather, H. Simianer** *Kernel-based BLUP with Genomic Data.* Proceedings: 9th World Congress on Genetics Applied to Livestock Production (WCGALP)
2009	**U. Ober** *Zum Problem, ob zentrale Divisionsalgebren von Primzahlgrad stets zyklisch sind.* Unpublished Diploma thesis, Mathematical Institute, Georg-August-Universität Göttingen, supervised by Prof. I. Kersten

Göttingen, October 2, 2012