Scale effects on genomic modelling and prediction

Dissertation

for the Doctoral Degree

at the Faculty of Agricultural Sciences,

Department of Animal Sciences,

Georg-August-University Göttingen

presented by

Swetlana Berger

née Miller

born in Temirtau

Göttingen, February 2015

1th Referee: Prof. Dr. Henner Simianer
Animal Breeding and Genetics Group
Department of Animal Sciences
Georg-August-University Göttingen

2nd Referee: Prof. Dr. Heike Bickeböller
Department of Genetic Epidemiology
University Medical Centre Göttingen
Georg-August-University Göttingen

Date of Disputation: 3rd February 2015

TABLE OF CONTENS

Summary

In this thesis, a novel method for scale corrected comparisons of LD structure in different genomic regions is suggested. Several aspects of scale-caused problems – from precision of marker effect estimates to accuracy of predictions for new individuals - are investigated. Furthermore, based on a comparison of the performance of different approaches, recommendations on the application of examined methods are given.

In the **first chapter** a general introduction to fundamental genetics and quantitative genetics methods is given. In the **second chapter** the impact of different marker density in terms of resulting higher LD between the markers on errors in estimates of marker effects is investigated. In order to quantify this impact, genotypes with a pre-defined LD structure are needed. For this purpose, four different simulation techniques were compared and the most reliable method - in terms of reproduction of sought LD structure in marker data - was used to generate a pool of genotype records with a pre-defined LD structure. The effect of the magnitude of LD between the markers on marker effect estimates was investigated in three linear models - *Single Marker Regression* (SMR), *Multiple Marker Regression* (MMR) and *Linear Mixed Model* (LMM) using different simulation scenarios that reflect differences in MAF (varying from 0.05 to 0.5 in steps of 0.025) and heritability fixed at 0.3, 0.5 or 0.7. A clear dependence was observed between the increase of LD in the data and the increase of errors in the effect estimates. A high amount of LD, above a threshold of harmful multicollinearity, had a large impact on the estimates of marker effects, whilst LD below this threshold had no influence on precision of estimates. The threshold of harmful multicollinearity was observed to depend on the model: in MMR a negative impact on the precision of estimates was observed when the amount of LD (measured in squared correlation, r^2) exceeded a value of 0.7, while in LMM, an even higher negative impact was detected for values of $r^2 \geq$ 0.6. Observed impact was more pronounced for SNPs with lower MAF and phenotypes with lower heritability. All in all, high LD level in marker data led to a bias in estimates from all the considered models that are routinely used when genomic data comprises thousands of markers.

A further scale-caused problem lies in the varying degrees of relatedness in different species and populations. The accuracy of genomic prediction in three whole genome regression (WGR) methods, performing variable selection or penalized estimation of marker effects, is the subject of the **third chapter**. The *Genomic Best Linear Unbiased Prediction* (GBLUP*)* represents a classical infinitesimal model, where the trait is described as the weighted sum of SNP effects and where all marker effect estimates are penalized equally. We performed two GBLUP methods, which differ in the calculation of genomic relationship matrix (Van Raden and LD-corrected matrices). The next evaluated model was the Bayesian hierarchical model *Bayes A*, where the prior distribution of marker effects (*scaled-t* distribution) induces differential shrinkage of marker effect estimates. Finally, in *Bayesian Sparse Linear Mixed Models* (BSLMM) the total effect at each SNP is the sum of a small and a potential sparse large effect. The BSLMM is a general model: if the variance of sparse effects is close to zero and variance of small effects is large, an infinitesimal model is applied, however, if the variance of small effects is close to zero and only a few SNPs with sparse effects are present, the Bayesian Sparse Variable Selection model is applied.

In order to investigate the accuracy of genomic predictions, extensive simulation studies that reflect different genetic architectures as well as the analysis of a real phenotype – human height – were performed. Data sets for both analyses were obtained from the GENEVA study, containing nominally unrelated individuals. After quality control the remaining 673.197 SNPs were divided into two subsets: randomly sampled 350.000 SNPs were assigned as markers and from the remaining 323.197 SNPs, a quantity of 5.000 SNPs was sampled in each repetition as *Quantitative Trait Loci* (QTL). Five different scenarios were considered to reflect different genetic architectures. In further simulation scenarios, the distribution of MAF in QTL and in markers was either identical or not. In all introduced scenarios, the genomic models were applied using different subsets of SNPs: 1) only markers or 2) markers and QTL or 3) only QTL. For the real data analysis as well as for the analysis of simulated data, 500 individuals were assigned randomly to the validation data set and the rest to the training, thus 30 replications were performed for each scenario. The correlation between true and predicted phenotype $cor(\mathbf{y},\hat{\mathbf{y}})$ was used to quantify the predictive ability (PA).

In each of the considered scenarios, the BSLMM outperformed both Bayes A and GBLUP methods and showed higher prediction accuracy. The averaged predictive ability of BSLMM ranged between 0.08 and 0.58 across the simulation scenarios and was in average 16% higher than in Bayes A and 123% higher than in GBLUP. In contrast to GBLUP, the prediction accuracy in BLSMM and Bayes A was improved by 10% by including QTL in addition to markers in the data set. When only a few genes were involved in the manifestation of a trait, the BSLMM provided very good results (PA of 0.55 ± 0.04) even when the degree of relatedness in the data set was low. The prediction accuracy corresponding to an infinitesimal trait was low for all considered methods (averaged PA ranged from 0.07 to 0.18), however BSLMM delivered good results and did not perform worse than GBLUP. For the analysis of genomic data from less related individuals and pertaining to traits with unknown genetic architecture, BSLMM proved to be a more robust and effective approach.

In the **fourth chapter** the causes of the phenomena observed in studies described in the second and third chapters are investigated: the LD structures in different genomic regions are explored. A method was introduced to enable a scale independent comparison of LD structure in different genomic regions. This method accounted not only for the MAF distribution in the regions under comparison, but also for the distribution of pair-wise physical distances and the pair-wise differences in MAFs. In the present work, a comparison of LD structure between a genic region (G) and a non-genic region (IG) was performed as well as a control comparison between two similar non-genic regions IG and IG'. To quantify the observed differences in all G/IG and IG/IG' pairs, the medians of squared correlations r^2 and standardized squared correlation r^2/r_{max}^2 were compared a) genome-wide as well as b) chromosome-wise by using Wilcoxon signed rank tests. Comparative studies were performed in three different species: an Arabidopsis data set (*A. thaliana*, genotyped using Affymetrix 250K SNP-tiling array), a human data set (*H .sapiens,* genotyped using 780 K Affymetrix Genome-Wide Human SNP Array 6.0) from GENEVA study and a white layer data set (*G. g. domesticus,* genotyped using 600 K Affymetrix Axiom® Genome-Wide Chicken Genotyping Array) from the Synbreed project. After the quality control procedure, 199 double haploid Arabidopsis inbred lines with 216 K SNPs, 5.827 human individuals with 685 K SNPs and 673 chickens with 278 K SNPs were available. Gene annotations were based on „Ensemble genes 74“ for human and chicken data sets and on „Ensemble plant genes 21“ for the Arabidopsis data set. In total 3.721 genic regions in *A. thaliana,* 7.180 in *H. sapiens* and 3.033 in *G. g. domesticus* were tested. Genome-wide comparison detected 31,2% more LD in genic compared to non-genic regions in *A. thaliana*, followed by 13,6% in *H. sapiens* and 6,0% in *G. g. domesticus*. Chromosome-wide comparison discovered significant differences on all 5 chromosomes in *Arabidopsis thaliana,* on one quarter of the human and one third of the chicken chromosomes. The control comparisons of LD structure in similar non-genic regions showed almost no significant differences in any species.

Chapter five presents a discussion on the influence of LD on the performance of the considered models and possibilities for mitigating the severity of consequences. An additional real data analysis of predictive ability of BSLMM is introduced, using British Cohort 1958 data set, which consists of records of unrelated individuals born in one week in March 1958. Furthermore, the sensitivity of Bayesian methods to the choice of hyper parameters and number of iterations is discussed and results of sensitivity analysis are presented.

Zusammenfassung

In dieser Arbeit wird eine neue Methode für den skalenunabhängigen Vergleich von LD-Strukturen in unterschiedlichen genomischen Regionen vorgeschlagen. Verschiedene Aspekte durch Skalen verursachter Probleme – von der Präzision der Schätzung der Markereffekte bis zur Genauigkeit der Vorhersage für neue Individuen - wurden untersucht. Darüber hinaus, basierend auf den Leistungsvergleichen von unterschiedlichen statistischen Methoden, wurden Empfehlungen für die Verwendungen der untersuchten Methoden gegeben.

Im **ersten Kapitel** wurde eine allgemeine Einführung in genetische Grundlagen und in die Methoden der quantitativen Genetik gegeben. Im **zweiten Kapitel** wurden die Auswirkungen der unterschiedlichen Markerdichten, in Form von daraus resultierenden höheren LD zwischen den Markern, auf Fehler bei der Schätzung der vorliegenden Markereffekte untersucht. Um diese Auswirkungen zu quantifizieren, wurden Genotypen mit einer vorgegebenen LD-Struktur benötigt. Zu diesen Zweck wurden vier mögliche Simulationsmethoden verglichen und die zuverlässigste Methode – im Sinne der Wiedergabe der gewünschten LD-Struktur in Markerdatensatz - wurde genutzt, um einen Datenpool mit Genotypen in einem vordefinierten LD zu erstellen. Die Auswirkung des unterschiedlichen Ausmaßes von LD zwischen den Markern auf die Schätzung der Markereffekte wurde in drei verschiedenen linearen Modellen - der *Single Marker Regression* (SMR), der *Multiple Marker Regression* (MMR) und der *Linear Mixed Model* (LMM) – untersucht. Dafür wurden Simulationsstudien mit Szenarien, die unterschiedliche MAF (zwischen 0.05 und 0.5 in 0.025 Schritten variierend) und die Heritabilitätswerte von 0.3, 0.5 oder 0.7 wiederspiegeln, verwendet. Eine deutliche Abhängigkeit der Korrelation zwischen den größeren Schätzfehlern und einem höheren Ausmaß von LD (oder Multikolliniarität) in den Daten konnte festgestellt werden. Ein höheres LD über einen Schwellenwert für unbedenklichen Multikollinearität im Datensatz hatte einen gravierenden Einfluss auf die Schätzungen von Markereffekten, wärend ein LD unterhalb dieses Schwellenwertes keine Auswirkung auf die Genauigkeit der Schätzung hatte. Eine Abhängigkeit dieses Schwellenwertes von dem Modell wurde beobachtet: für MMR wurde eine Verringerung der Schätzgenauigkeit für LD-Werte (gemessen als quadrierte Korrelation r^2) über 0.7 beobachtet, während für LMM größere Genauigkeitsverluste für LD-Werte $r^2 \geq 0.6$ festgestellt wurden. Die beobachtete Auswirkung war stärker ausgeprägt für SNPs mit niedrigerem MAF und für Merkmale mit niedrigerer Heritabilität. .Zusammenfassend lässt sich sagen, dass ein höheres LD-Niveau in den Markerdaten zu einer Verzerrung der Schätzung der Markereffekte bei allen untersuchten Modellen, die üblicherweise bei den Analyse von genomischen Daten angewandt werden, führte.

Ein weiteres Skalenproblem liegt im unterschiedlichen Ausmaß von Verwandtschaft in unterschiedlichen Populationen und Spezies. Die Genauigkeit der genomischen Vorhersage in drei genomweiten Regressionsmodellen (WGR), die sowohl Modellselektion als auch unterschiedliche Penalisierung (Bestrafung) der Markereffekte durchführen, war der Gegenstand des **dritten Kapitels**. Durch *Genomic Best Linear Unbiased Prediction* (GBLUP*)* wird ein klassisches, infinitesimales Modell repräsentiert: Hier wird das Merkmal als gewichtete Summe der SNP-Effekte dargestellt und die Bestrafung der Effektgröße ist für alle Marker gleich. Zwei verschieden GBLUP Methoden wurden betrachtet, die sich in der Berechnung der genomischen Verwandschaftsmatrix **G** unterscheiden (Van Raden **G** und LD-korrigierte Matrix **G-ldak**). Bei dem zweiten Modell handelt es sich um *Bayes A*, welches eine a-priori Annahme an die Verteilung von Markereffekte stellt (*scaled-t* Verteilung) und diese entsprechend ihrer Effektgröße bestraft. Im *Bayesian Sparse Linear Mixed Models* (BSLMM) wird der gesamte Effekt von jedem SNP durch die Summe von einem kleinen und - bei einem bestimmten Anteil der SNPs - einem zusätzlichen großen Effekt dargestellt, folglich ist BSLMM eine neue Implementierung von einem Spike-Slab Modell (SS). Bei dem SS handelt es sich um ein verallgemeinertes Modell: Ist der Anteil an SNPs mit zusätzlichem Effekt gleich Null, so liegt ein infinitesimales Modell vor, wenn die Varianz der kleinen Effekte gegen Null geht und nur wenige SNPs mit großen Effekten vorhanden sind, so liegt ein Bayesian Sparse Variable Selection Modell vor.

Um die Genauigkeit der genomischen Vorhersage zu untersuchen, wurden sowohl die Simulationsstudien, die unterschieche genetische Architekturen wiederspiegeln, als auch Analysen der realen Phänotypen (menschliche Körpergröße) durchgeführt. Für die Analysen standen die Humandaten aus der GENEVA Studie zur Verfügung, welche 5.758 nominal unverwandte Individuen umfassen. Nach der Qualitätskontrolle, wurden die verbliebenen 673.197 SNPs in zwei Teildatensätze aufgeteilt: 350.000 SNPs wurden zufällig als Marker ausgewählt und aus den restlichen 323.197 SNPs wurden 5.000 SNPs bei jeder Wiederholung als Quantitative Trait Loci (QTL) zufällig ausgewählt. Fünf unterschiedliche Szenarien spiegelten unterschiedliche genetische Architektur von Merkmal wieder. In einem weiteren Simulationsszenario waren die Verteilungen von Frequenzen der seltenen Allele (MAF) in QTL und Marker gleich oder unterschiedlich. Alle Szenarien wurden mit unterschiedlich zusammengesetzten genomischen Datensätzen analysiert: 1) nur Marker, 2) nur QTLs und 3) Marker und QTLs. Sowohl für die Analyse von simulierten als auch für die Analyse von den realen Daten wurden 500 Individuen zufällig in die Validierungsgruppe eingeteilt und der Rest in die Trainigsgruppe; insgesamt wurden 30 Wiederholungen durchgeführt. Die Korrelation zwischen den wahren und vorhergesagten Phänotypen $cor(\mathbf{y},\hat{\mathbf{y}})$ wurde benutzt um die Vorhersagegenauigkeit (PA) zu quantifizieren.

In jedem der untersuchten Szenarien zeigte SS eine höhere Vorhersagegenauigkeit als Bayes A und GBLUP. Die mittlere PA von SS lag zwischen 0.08 und 0.58 über alle Simulationsszenarien hinweg und war im Schnitt 16% höher als von Bayes A und 123% höher als PA von GBLUP. Im Gegensatz zu GBLUP war die Genauigkeit der Vorhersage in SS und Bayes A 10% höher, wenn zusätzlich zu den Markern die QTL im Datensatz enthalten waren. Im Falle, dass nur wenige Gene an der Ausbildung des Merkmals beteiligt waren, lieferte SS sehr gute Ergebnisse (PA von0.55 ± 0.04) auch für wenig verwandte Individuen. Unter einem infinitesimalen Modell, war die Vorhersagegenauigkeit war niedrig bei allen betrachteten Methoden (mittlere PA von 0.07 bis 0.18), aber SS lieferte gute Ergebnisse und war nicht schlechter als GBLUP. Für die Analyse von genomischen Daten von wenig verwandten Individuen oder von Merkmalen mit unbekannter genetischer Architektur, erwies sich SS als eine besser geeignete und robustere Methode

Im **vierten Kapitel** wurden die Ursachen der in Kapitel zwei und drei beschriebenen Phänomene detailliert untersucht: Vergleiche der LD-Strukturen in unterschiedlichen genomischen Regionen wurden durchgeführt. Eine Methode wurde vorgestellt, die einen skalenunabhängigen Vergleich von LD-Strukturen in unterschiedlichen genomischen Regionen ermöglicht. Diese Methode berücksichtigt nicht nur die Verteilung von MAF in den zu vergleichenden genomischen Regionen, sondern auch die Verteilung der paarweisen physikalischen Distanz und Differenzen in den MAFs. Vergleiche der LD-Struktur wurden zwischen ähnlichen Gen- und Nicht-Genregionen (G und IG), sowie Kontrollvergleiche zwischen zwei ähnlichen Nicht-Genregionen (IG und IG‘) durchgeführt. Um die beobachteten Unterschiede zu quantifizieren, wurden für die Mediane der quadrierten Korrelationen (r^2) und den Ausschöpfungskoeffizienten ($r_s^2 = r^2/r_{max}^2$) aller G/IG und IG/IG‘ Paare a) chromosomenweise sowie b) genomweite Vorzeichenrangtests von Wilcoxon durchgeführt. Vergleichsstudien wurden in drei verschiedene Spezies durchgeführt: Arabidopsisdaten (*A. thaliana,* typisiert mit Affymetrix 250K SNP-tiling array), Humandaten (*H. sapiens,* typisiert mit 780K Affymetrix Genome-Wide Human SNP Array 6.0) aus der GENEVA-Studie und Weißlegerdaten (*G. g. domesticus,* typisiert mit 600K Affymetrix Axiom® Genome-Wide Chicken Genotyping Array) aus dem Projekt „Synbreed“ wurden benutzt. Nach der Qualitätskontrolle standen für die folgenden Analysen 199 homozygote Arabidopsis-Inzuchtlinien mit 216 K SNPs, 5,827 Menschen mit 685 K SNPs und 673 Hühner mit 278 K SNPs zur Verfügung. Genannotationen basierten auf der Version „Ensemble genes 74“ für die Human- und Hühnerdaten bzw. auf „Ensemble plant genes 21“ für die Arabidopsisdaten. Insgesamt wurden 3,721 Genregionen in *A .thaliana,* 7.180 in *H. sapiens* und 3,033 in *G. g. domesticus* getestet. In einem genomweiten Vergleich wurde in *A. thaliana* ca. 31,2% mehr LD in Genregionen als in Nicht-Genregionen entdeckt, in *H. sapiens* ca. 13,6% und in *G. g. domesticus* ca. 6,0%. In den chromosomweisen Vergleichen wurden signifikante Differenzen an allen 5 Chromosomen in *Arabidopsis thaliana* entdeckt, an einem Viertel von den Chromosomen in *H. sapiens* und an einem Drittel der Chromosomen in G. g. domesticus. Die Vergleiche von IG mit IG‘ zeigten so gut wie keine signifikanten Unterschiede.

Das fünfte Kapitel beinhaltet eine Diskussion über die Auswirkung von LD auf die Leistungsfähigkeit der betrachteten Modelle und Möglichkeiten zur Begrenzung der negativen Konsequenzen. Eine zusätzliche SS Analyse von neuen realen Merkmalen von British Cohort 1958 Datensatz, welcher Daten von unverwandten Individuen beinhaltet, die in einer einzigen Woche in März 1958 geboren sind. Darüber hinaus wurde eine Sensitivitätsanalyse bezüglich der Wahl der Hyperparameter in Bayesianischen Methoden und die Zahl der benötigten Iterationen präsentiert.

1ST CHAPTER

General Introduction

Uniqueness of each individual, either human or animal, is created by small deviations in genetic materials inherited. The stature and performance as well as the susceptibility to particular diseases depend on a specific base pair manifestation in the deoxyribonucleic acid (DNA) chain. The ultimate goals of quantitative genetics are firstly, to identify regions that play an important role in the inheritance of particular traits and secondly, to predict those traits for new individuals using the available genomic information. Since the rapid development of genome sequencing and genotyping techniques in the last decades, a variety of informative markers covering the whole genome are now available. These markers, which are specific variations in the sequence of the bases in the DNA, as well as the phenotypic records are the input used for statistical analysis. Many parametric and non-parametric statistical models and approaches have been proposed for assignment of genomic data to the phenotypes.

Until a few years ago, only a small number of genetic variants were available for modeling but in the last few years, genotypes from thousands of individuals with hundreds of thousands of markers each have become available. However, computational and methodological problems arise and approaches functioning well with a small number of variants need to be verified and if necessary adapted to high-density data.

Genomic data

Molecular genetics background

Firstly, a short introduction to some fundamental genetics is presented, based on genetics book by Henning (2001).

DNA contains genetic information, stored as a sequence of four nucleotides (**A**denine, **C**ytosine, **G**uanine and **T**hymine), which build base pairs A with T and G with C. These base pairs are arranged in two strands that form a kind of spiral, called *double helix*. Due to pairing of complementary bases, the replication of DNA during the division of a cell is enabled. In higher organisms, the genome is organized in sets of *chromosomes* that represent DNA sections of different length, and the number of chromosomes varies across species. In general, in a *diploid* organism like humans or most animals, the genome consists of pairs of chromosomes that comprise two identical copies (*autosomes*) and two copies of non-identical sex chromosomes (*allosomes*) that determine the sex of the individual. For instance, humans are diploid and possess 46 chromosomes: a double set of 22 autosomes and one set of allosomes XX (for female) and XY (male), while wheat is hexaploid and possesses 42 chromosomes in total with six copies each of 7 chromosomes. Hereafter only diploid organisms will be considered and the two copies of a chromosome will be referred to as the inherited maternal or paternal chromosome. Since humans are diploid, there are 2^{23} possibilities of combining the maternal and paternal haploid chromosome sets.

A *gene* is a unit of heredity which carries the information for construction functional molecules, called proteins. The position of a specific location of a gene or a single base pair on the genome, called *locus*, is the analogue to a physical address. For instance, in sugar beet the base pair manifestations at about 98.7% of $5.5 \cdot 10^{8}$ loci are identical in humans and only 1.3% of loci have different variants, called *alleles*. Variation in the genome occurs spontaneously during cell division or as an error in genetic recombination. Errors in duplication of a DNA strand might result in changing a single nucleotide, which is called *point mutation*. In case a point mutation increases the fitness of the organism, it has a chance to remain in the population. If the new allele appears in up to 1% of individuals, it is called a *rare variant*. One or more extra nucleotides added during the replication process are called *insertions*, and extra nucleotides that are removed are called *deletions.* Structural variants that occur repeatedly, for instance insertion or deletion will occur one, two or three times in a population, the different numbers of structural variation are called *copy number variations* (CNV). The last structural rearrangement of DNA that we will mention here is *crossing over,* which refers to the exchange of genetic material between the paternal and maternal copies of a chromosome when the two sister chromatids overlap. This exchange alters the constellation of parental origin upstream and downstream of the site where the crossing over has taken place and thus is referred to as *recombination*. For instance, in human an average probability of occurrence of recombination is 10^{-6} (Malats and Calafell, 2003), although the recombination rate varies greatly across the genome.

A locus with occurrence of different nucleotides among individuals is called *single nucleotide polymorphism* (SNP). Most commonly, SNPs have only two alleles, the less frequent allele is called the *minor allele*. Accordingly, the frequency of the minor allele is referred to as the *minor allele frequency* (MAF). A set of SNPs at a single chromosome copy is referred to as a haplotype. The summaries of observed alleles at both copies, which are, e.g., AA, AG or GG, are called *genotypes.* At any given locus, genotypes with the same set of alleles (e.g. A/A or G/G) are referred to as *homozygous* and genotypes with different set of alleles (e.g. A/G or G/A) are referred to as *heterozygous*. Note that most modern genotyping methods cannot assign the realization of alleles to the original haplotype strand; however, plenty of approaches exist that can reconstruct haplotypes from the observed genotypes (e.g. Scheet and Stephens, 2006; Browning and Browning, 2009; Roach et al., 2011; Delaneau et al., 2012).

Without recombination, loci situated on one chromosome would be inherited together from generation to generation. Other evolutionary forces like random mating, selection or genetic drift also influence the linkage between two or more loci. The non-random association between alleles at different loci is referred to as *linkage disequilibrium* (LD) (this association can be interpreted as a measure of correlation between pairs of loci), while two alleles occurring absolutely independently are in *linkage equilibrium*.

Marker genotype data

In our studies we restrict ourselves to the most common type of genomic polymorphism, the SNP, which is for our purpose the most informative of all markers (Middleton et al., 2004). The scientific importance of SNPs arises because of their high frequency, e.g. in human $3.8 \cdot 10^7$ SNPs exist, which corresponds to 1.3% of the total of $3.3 \cdot 10^9$ base pairs (Kersey, 2014), as well as their availability in a wide range of species at relatively low genotyping costs.

In the present study, SNP chip arrays from Affymetrix Inc. were used. The information from the SNP chip, denoted for instance as A/B or as A/T/G/C, was re-coded numerically for the statistical analysis of a quantitative trait as 0, 1 or 2, according to the number of minor alleles. Affymetrix and Illumina are two largest commercial producer of the SNP arrays, whereby Affymetrix produced the first commercial SNP array containing 1494 SNPs (Wang et al., 1998). Albeit the differences in how both genotyping platforms are designed, both SNP arrays share the same basic principle of complementary binding of nucleotides, namely A to T and C to G. Both genotyping method utilize hybridization of single-strand DNA sequences to prepared arrays, containing plenty nucleotide probe sequences. The intensity of signal can be measured and, assuming that signal intensity depends on the amount of target DNA, translated to genotypes AA, AB or BB. Both manufactures report genotyping accuracy about 99.5 % (LaFramboise, 2009). A comparative study involving 12 different SNP arrays (Ha et al., 2014) have shown that performance in terms of coverage and cost-efficiency of different population-optimized SNP arrays varies across populations and the choice of a SNP array should be done depending on genetic background of the sample.

In recent years a new sequencing technique called *next-generation sequencing* (Mardis, 2008) has rapidly developed. The key aspect of the next-generation sequencing is the ability to simultaneously sequence millions of DNA fragments.

Genomic predictions

Prediction of phenotypes for new individuals proceeds in two steps: 1) a genomic model is fitted to the training data set and 2) the phenotype or the breeding value, often used in animal breeding, for a new individual is predicted based on the genotype readings of this individual and the estimated marker effects from the fitted model. The evaluation of prediction accuracy can be performed using *training-testing validation design* (Hastie et al., 2005). For this purpose the data set is split many times into training and testing data sets; the assignment of individuals to either one of the subsets occurs randomly. In each repetition of the design, the correlation between the predicted and true phenotype for individuals in the training subset is calculated. This allows us to obtain the distribution of correlation coefficients with corresponding confidence bounds (Fisher, 1915; Hawkins, 1989).

Genomic models and approaches

Genomic models are needed to create a link between the phenotype or trait of interest and the genomic marker data, in order to estimate the marker effects or to predict an unobserved phenotype for a new individual. Challenges in the study of association between genomic markers and traits of interest typically include computational problems associated with large datasets and the over parameterization of models due to the large number of genomic variants. The causal loci for a trait are referred to as quantitative trait loci (QTL); in the simplest case each causal locus affects the trait (positively or negatively) and the sum over effects of all QTL results in the observed manifestation of the trait. The relationship between the QTL may deviate from pure additive nature and the underlying genetic architecture of a complex trait may consist of an additive component as well as the interaction between different genomic regions. Although classical regression models like multiple regression are simple to perform, they can only assume additive effects and will fail in case the number of predictors is larger than the number of individuals in the sample, which is the so called *small-n-large-p* problem. Many regression models, based on different penalization procedures of marker effect estimates, like *ridge regression* (Hoerl and Kennard, 1976) or *LASSO* (Tibshirani, 1996) cope with the *small-n-large-p* problem but still ignore the potential interaction between genes or between genomic and environment data. To capture these potentially non-linear components arising from interactions within the genome, non-parametric methods like *reproducing kernel Hilbert spaces* regression (RKHS) (de los Campos et al., 2010; Ober et al., 2011), the *radial basis functions* model (Long et al., 2010; González-Camacho et al., 2012) or artificial *neural networks* (Ehret et al., 2014) are often used. The diversity of available approaches is considerable, most of these methods are parametric. A short outline of the genomic models often used in quantitative genetics is presented below.

Linear Regression models

The *Single Marker Regression* is a standard approach used in *genome wide association studies* (GWAS), where the observed phenotype is modeled against each individual locus separately. Consequently, the problem of multiple testing of marker effects arises and the significance level needs to be corrected. For instance, one can apply the *Bonferroni correction* (Dunn, 1961), which is based on penalization of the global significance level by the number of comparisons. The Bonferroni correction is the simplest but most conservative approach to control the family-wise error rate. An alternative method to control the Type I error, the *false discovery rate* (FDR) (Benjamini and Hochberg, 1995), is characterized by less conservative behavior and consequently by higher statistical power. This method is based on considering the proportion of expected false discoveries, thus a posteriori adjusting of the significance level as performed by Bonferroni correction is not needed.

In *multiple marker regression,* marker effects can be assumed to be fixed and the phenotype is modeled as the weighted sum of genotypes, where the weights correspond to the marker effects (Meuwissen et al., 2001). This approach has no unique solution in situations where the number of predictors exceeds the sample size, which is a common situation in genomic analysis. To overcome this limitation, the *Least-Square Regression* proposed by Meuwissen et al. (2001) or the *Least Angle Regression* proposed by Efron et al. (2004) perform a stepwise forward selection procedure for inclusion of most informative SNPs. A similar approach, the *Partial Least Square Regression* (Helland, 1990), constructs orthogonal predictors by transforming the original genotype matrix. Another possibility to cope with this over-parameterization problem is to penalize the effect estimates. Plenty of penalized estimation methods exist, and the main difference between these methods lies in the choice of penalty. Most of methods make predictions with the sum of estimated effects weighted by the new individual observed genotypes. The so called shrinkage methods, for instance *ridge regression* proposed by Hoerl and Kennard (1976) or *LASSO* proposed by Tibshirani (1996), tend to have less prediction error in comparison to model selection approaches. An approach proposed by Zou and Hastie (2005), called *Elastic Net*, suggests a compromise between model selection and shrinkage. Penalized estimation is a rapidly developing research field with many approaches being proposed (Shen et al., 2013; Burnaev and Vovk, 2014; Fan et al., 2014; Beran, 2014)

The *linear mixed model* (Henderson, 1950; Henderson, 1963; Goldberger, 1962) simultaneously models fixed covariates as well as the random SNP effects. A widely used approach in animal breeding, the *genomic best linear unbiased predictor (GBLUP)* (Henderson, 1984; Meuwissen et al., 2001), is as special form of linear mixed model in which the covariance structure is modeled from the relatedness within the sample. This model can be viewed as a ridge regression model when performing uniform shrinkage of estimates, with a shrinkage parameter equal to the ratio of residual and genetic variance components.

Bayesian linear regressions

A large number of Bayesian methods have arisen in the last decade; here, only a short outline is given that is not claimed to be complete. Bayesian variable selection and shrinkage estimation approaches require a priori assumptions on the distribution of marker effects. Different Bayesian approaches vary in their a-priori assumptions and in handling the hyperparameters of the prior distribution, which are a further hierarchical level in the model and can be modeled as either fixed or random. The prior beliefs specify whether variable selection, shrinkage or both – variable selection and shrinkage - will be performed. For instance, Bayes A and B proposed by Meuwissen et al. (2001) perform different regularization of estimates: Bayes A performs a marker specific shrinkage of estimates, whilst Bayes B performs differential shrinkage and does variable selection in addition to the regularization procedure. New implementations of the spike-slab model (Mitchell and Beauchamp, 1988), which is equivalent to a wide class of Bayesian methods called the Bayes C, have been proposed recently (Zhou et al., 2013; Goodfellow et al., 2013; Hernández-Lobato et al., 2013). In Bayes C, a two-point mixture distribution made up of a flat distribution and a distribution concentrated around zero, is assigned as a prior distribution of marker effects. Using this type of prior induces variable selection. Bayesian Lasso or Bayes L, proposed by Park and Casella (2008) presents an analogue to LASSO regression mentioned above. In contrast to the non-Bayesian version, it does not remove markers from the model; rather markers with small effects are regularized even stronger. In Bayes R, proposed by Erbe et al. (2012), a four component mixture distribution is assigned as a prior distribution of marker effects. In addition to the prior beliefs about the distribution of marker effects, an a priori assumption on genetic variance is made that leads to an improvement in predictive ability. The key aspect here is the usage of prior knowledge, gained from prior cross-validation study, for setting the prior genetic variance parameter.

In all Bayesian settings, the impact of prior distribution decreases with the growing sample size (Gianola, 2013) but for small samples the choice of prior is crucial for the performance of the model (Lehermeier et al., 2013). The estimates of unknown hyperparameters as well as the estimates of marker effects in all Bayesian approaches are sampled from a posteriori distribution, achieved in a sampling procedure. Some of the widely used Markov chain Monte Carlo (MCMC) methods are the Gibbs sampler (George and McCulloch, 1993) and Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) sampling algorithms.

Non-parametric methods

Predictive functions are used in machine learning techniques to obtain genomic predictions. Such predictive functions result from a training process that is based on a generalization algorithm. The training set consists of phenotype and genotype records and is used to predict the phenotype of a new individual not included to the training set. In contrast to additive models described above, non-parametric methods provide an opportunity to capture both, additive as well as non-additive effects.

For instance, in RKHS regression the effects are described by a real valued function of genotypes and a kernel defines an a priori correlation structure of outputs of this function. The choice of kernel is crucial for the performance of the model. In case a Gaussian kernel is chosen, RKHS regression is equivalent to the ridge regression and consequently equivalent to GBLUP method (de los Campos et al., 2010). An advantage of using RKHS method is the reduction of dimensionality from number of SNPs to the sample size, this method then models genetic values assigned to the individuals in the sample.

A *neural network* (NN) (Hastie et al., 2005; Ehret et al., 2014) is made up of components that are called *layers* in the context of NN: the input layer containing the genotype records, the output layer containing the phenotypes and hidden layers in-between them both. NN is as a system of interconnected neurons or nodes, where in the hidden layers at each node the inputs, weighted by connection specific constants are summed up. Thus hidden layers can be understood as a system of weighted paths between the inputs and outputs. Predictions performed using NN are based on predictive functions, which might be expressed analytically or result from approximation processes. NN can be viewed as a non-linear regression model that is trained using Markov Chain Monte Carlo methods.

The *support vector machine* (SVM) (Cortes and Vapnik, 1995; Long et al., 2011) is an algorithm developed from statistical learning theory that can be used for estimating unknown regression coefficients or unknown maker effects in context of quantitative genetics. Applying SVM regression, the relationship between the observed phenotypes and genotypes can be mapped using linear as well as the non-linear mapping functions. The regularization parameter, which penalizes the complexity of the model, and the choice of loss function as a measure of quality of estimates defines the SVM model.

A guide over this thesis

Scale problems are omnipresent in quantitative genetic analysis; different scales in relatedness among individuals in the data set, different marker densities or different numbers of markers – from the single marker to the whole genome data - used as input in a genomic model can have an impact on the performance of genomic models. In particular, the rapid development of molecular genetics, especially of high throughput sequencing and genotyping techniques, gives us a large amount of genotypes. Scale related problems arise with growing data sizes and the computational ability of classical approaches reaches its limits. A crucial point is whether the methods, which perform well in low-density data sets, will maintain the quality of estimation and prediction when applied to a high-density data set.

This study aims at investigating the impact of different scales in genomic data as well as different scales in the input data of widely used methods on the precision of estimates of genomic effects and on the accuracy of genomic predictions.

Chapter 2 reports the impact of multicollinearity on the performance of three different models: single marker regression, multiple marker regression and linear mixed model. A detailed insight into the nature of the problem is provided, and the consequences of variation in the amount of LD on effect estimates at each single SNP are investigated. For this reason, a technique to simulate genotype data with a pre-defined LD structure is developed and compared with other approaches so as to assess the reliability of generated LD structure.

Chapter 3 deals with comparison of the accuracy of predictions in unrelated individuals, obtained from different statistical methods: GBLUP, Bayes A and a new implementation of the spike-slab model. Extensive simulations are designed to assess the effects of important factors such as the extent of LD between markers and QTL and trait complexity on prediction accuracy. Additionally, a real data analysis comparing the predictive performance of different methods on human height is performed.

Chapter 4 introduces a new method for comparison of LD in different genomic regions. This method enables us to control the differences in minor allele frequencies as well as the differences in spatial structures of genomic regions under comparison, thus a scale corrected comparison is performed. Further, an upper limit for squared correlation is achieved using known allele frequencies and boundaries for gametic frequencies, derived using the Fréchet-Hoeffding bounds. This upper limit is needed for construction of a MAF independent measure of LD. This method is used for the investigation of differences in magnitude of the LD between genic and non-genic regions. A significantly higher LD level is detected in genic regions compared to non-genic regions in all considered data sets: in human, animals (chicken) and plants (*Arabidopsis thaliana*).

In **Chapter 5** comprises a general discussion on the impact of different marker densities and methods chosen on scales.

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 289–300.

Beran, R. (2014). Hypercube estimators: Penalized least squares, submodel selection, and numerical stability. Comput. Stat. Data Anal. *71*, 654–666.

Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.

Burnaev, E., and Vovk, V. (2014). Efficiency of conformalized ridge regression. ArXiv Prepr. ArXiv14042083.

De los Campos, G., Gianola, D., Rosa, G.J., Weigel, K.A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. *92*, 295–308.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Dunn, O.J. (1961). Multiple comparisons among means. J. Am. Stat. Assoc. *56*, 52–64.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., and others (2004). Least angle regression. Ann. Stat. *32*, 407–499.

Ehret, A., Tusell, L., Gianola, D., and Thaller, G. (2014). Artificial neural networks for genome-enabled prediction in animal and plant breeding: A review.

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

Fan, J., Xue, L., Zou, H., and others (2014). Strong oracle optimality of folded concave penalized estimation. Ann. Stat. *42*, 819–849.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 507–521.

George, E.I., and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. J. Am. Stat. Assoc. *88*, 881–889.

Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics.

Goldberger, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. J. Am. Stat. Assoc. *57*, 369–375.

González-Camacho, J.M., De Los Campos, G., Pérez, P., Gianola, D., Cairns, J.E., Mahuku, G., Babu, R., and Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. *125*, 759–771.

Goodfellow, I.J., Courville, A., and Bengio, Y. (2013). Scaling up spike-and-slab models for unsupervised feature learning. Pattern Anal. Mach. Intell. IEEE Trans. On *35*, 1902–1914.

Ha, N.-T., Freytag, S., and Bickeboeller, H. (2014). Coverage and efficiency in current SNP chips. Eur. J. Hum. Genet.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. Math. Intell. *27*, 83–85.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika *57*, 97–109.

Hawkins, D.L. (1989). Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. Am. Stat. *43*, 235–237.

Helland, I.S. (1990). Partial least squares regression and statistical models. Scand. J. Stat. 97–114.

Henderson, C.R. (1950). Estimation of genetic parameters. In Biometrics, , pp. 186–187.

Henderson, C.R. (1963). Selection index and expected genetic advance. Stat. Genet. Plant Breed. *982*, 141–163.

Henderson, C.R. (1984). Applications of linear models in animal breeding (University of Guelph, Guelph, ON, Canada).

Henning, W. (2001). Genetik (Springer).

Hernández-Lobato, D., Hernández-Lobato, J.M., and Dupont, P. (2013). Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. J. Mach. Learn. Res. *14*, 1891–1945.

Hoerl, A.E., and Kennard, R.W. (1976). Ridge regression iterative estimation of the biasing parameter. Commun. Stat.-Theory Methods *5*, 77–88.

Kersey, P.J. (2014). Ensembl Plants-an Integrative Resource for Plant Genome Data. In Plant and Animal Genome XXII Conference, (Plant and Animal Genome),.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. gkp552.

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

Long, N., Gianola, D., Rosa, G.J., Weigel, K.A., Kranis, A., and Gonzalez-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. Genet. Res. *92*, 209–225.

Long, N., Gianola, D., Rosa, G.J., and Weigel, K.A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. Theor. Appl. Genet. *123*, 1065–1074.

Malats, N., and Calafell, F. (2003). Basic glossary on genetic epidemiology. J. Epidemiol. Community Health *57*, 480–482.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet *9*, 387–402.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys. *21*, 1087–1092.

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

Middleton, F.A., Pato, M.T., Gentile, K.L., Morley, C.P., Zhao, X., Eisener, A.F., Brown, A., Petryshen, T.L., Kirby, A.N., Medeiros, H., et al. (2004). Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide–polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. Am. J. Hum. Genet. *74*, 886–897.

Mitchell, T.J., and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. J. Am. Stat. Assoc. *83*, 1023–1032.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., and Simianer, H. (2011). Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics *188*, 695–708.

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

Roach, J.C., Glusman, G., Hubley, R., Montsaroff, S.Z., Holloway, A.K., Mauldin, D.E., Srivastava, D., Garg, V., Pollard, K.S., Galas, D.J., et al. (2011). Chromosomal haplotypes by genetic phasing of human families. Am. J. Hum. Genet. *89*, 382–397.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.

Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science *280*, 1077–1082.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.

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

2ND CHAPTER

Scale Dependency in the Estimation of Quantitative Trait Loci Effects

SWETLANA BERGER, HENNER SIMIANER

Animal Breeding and Genetics Group, Department of Animal Sciences,

Georg-August-University Goettingen,

Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany

Publication in preparation

Introduction

Due to rapid development of gene sequencing methods, a huge amount of genomic data is now available, accompanied by lower genotyping costs: for example, the Next-Generation Sequencing technology allows the production of millions of DNA sequence reads in a single run. In animal breeding, until a few years ago, genomic data containing a few hundred microsatellites or low-density SNP-chips with about 3.000 markers were used (Beuzen et al., 2000) and were subsequently replaced by SNP-chips with about 54.000 markers. Currently, high density SNP-chips comprising approximately between 600.000 and 2 million SNPs, respectively, are used in in animal breeding and in human genetics, not to mention the growing number of sequence data sets across these and other species. This explosion of information leads to the question whether the performance of genomic models will change given the increase in marker density. High-density data provided by modern methods of high throughput sequencing or genotyping are characterized by a high degree of non-random association between the markers (de los Campos at al., 2009). This association is known as linkage disequilibrium (LD) and can be interpreted as a measure of correlation between pairs of loci.

Modeling the relationship between the available genomic information and phenotypes of interest is one of the most important aspects of quantitative genetics. In animal breeding, a response or target variable, such as milk yield, fat percentage or the widely used breeding value, is described using a set of predictors. In genomics, these predictors are represented using molecular markers, usually SNPs. Multiple regression methods are powerful tools used for gaining quantitative insights into genetic research as long as the assumptions and limitations of those methods are understood and recognized. One of the main assumptions is the independence of predictors, which is very hard to hold in practice given the redundancy of information from correlated predictors. This problem, called multicollinearity, is well-known in many scientific fields (Gunst and Webster, 1975; Kockläuner, 1984; Graham, 2003; Tu et al., 2005; Wheeler and Tiefelsdorf, 2005). Lack of awareness of this fact can lead to wrong results; for instance, the estimated parameters are often of incorrect magnitude or sign. Most of the methods that deal with this multicollinearity problem are two-step procedures that include a diagnostic step and various ad hoc procedures. For instance, Slinker and Glantz (1985) discussed experimental designs that would minimize the extent of multicollinearity in the analysis of physiological data, Mason and Brown (1975) investigated the bias caused by multicollinearity upon performing ridge regression (RR) on sociological data, and Ofir and Khuri (1986) addressed the subject of handling multicollinearity in marketing data. However, all of these approaches used small data sets with few predictors and cannot be directly applied to the problems in quantitative genetics where the number of predictors is in several hundreds of thousands.

To develop approaches that resolve the problem of multicollinearity in quantitative genetics, the initial step is to understand whether methods that work reliably with low density SNP data give trustworthy results with high-density SNP data. Hence, this study investigates the impact of multicollinearity on the performance of linear models used in quantitative genetics. One of the major aims is to provide sufficiently detailed insight into the pattern and severity of consequences on the marker effect estimates caused by multicollinearity in genomic data. Impact of different levels of LD on each SNP effect estimate was investigated using three different models: Single Marker Regression (SMR), Multiple Marker Regression (MMR) and Linear Mixed Model (LMM).

Material and Methods

Linear Models

How the genomic information (in our study, SNP data) is used in the estimation of marker effects and prediction depends on the choice of a model. For example, candidate gene approaches, which utilize only a pre-specified part of the genome, are based on knowledge from previous studies about the particular trait and are widely used in human genetics. For Mendelian traits with a simple genetic architecture (where genetic variance is explained by a small number of variants), such approaches are the method of choice. However, most productive traits (e.g. meat and milk yield) are not influenced by a small subset of variants, rather a large number of genomic variants with moderate and small effects (Robertson, 1967). In practice, lack of knowledge about the genetic architecture of the majority of traits coerces us to use an infinitesimal model, which is based on the assumption that an infinitesimal number of small effects are widespread across the genome. The SNPs are coded as 0, 1, or 2, according to the number of minor alleles at each locus, which corresponds to the additive modelling of marker effects.

In our studies three common linear statistical models are compared: Single Marker Regression (SMR), Multiple Marker Regression (MMR) and Linear Mixed Model (LMM).

Single Marker Regression

Generally, in a linear model a response Y is explained as a linear combination of predictors (or functions of them) and an error term containing unused or unknown information that is not included in the model as well as the remaining random effects on Y. In an SMR model (Grapes et al., 2004), the response (in genetic context often a phenotype or trait) is individually fit against each SNP while the unknown marker effects are assumed to be fixed. For a specific SNP data set consisting of p SNPs, p different linear equations for the same n-dimensional vector of phenotypes $\mathbf{Y}$ can be formed:

$$\begin{array}{l} \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \\ \vdots \\ \mathbf{Y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}_p \end{array} \text{, for } j = 1,..., p\text{,} \qquad (1)$$

where a $n \times 2$ design matrix $\mathbf{X}_j$ contains, for all n individuals, a vector of ones and genotype readings $\mathbf{Z}_j$ for the j^{th} SNP, $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \mathbf{V}_j)$ is a vector of errors in each model and $\boldsymbol{\beta}_j$ contains the population mean and effect of j^{th} SNP. A $n \times n$ matrix $\mathbf{V}_j$ is the residual variance-covariance matrix $\mathbf{V}_j = \mathrm{Var}(\mathbf{Y})$ in the model for j^{th} SNP, which is also the phenotypic variance-covariance, since the effects are assumed to be fixed in SMR. The marker effect at the current SNP is estimated for each equation, independent of the results for the rest of SNPs. The information contained at other markers is aggregated into the error term; thus predictions from an SMR-model are not usually exact and just give a basic idea about the genetic effects.

The impact of association between the markers on the precision of estimates can be comprehended using a simple example for $p = 2$. In this case we would have two linear equations to describe the relationship between the vector of phenotypes and markers 1 and 2 separately:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1\,Z_{11} \\ 1\,Z_{21} \\ \vdots \\ 1\,Z_{n1} \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} \\ \beta_{11} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n1} \end{pmatrix} \text{ and } \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1\,Z_{12} \\ 1\,Z_{22} \\ \vdots \\ 1\,Z_{n2} \end{pmatrix} \cdot \begin{pmatrix} \beta_{02} \\ \beta_{12} \end{pmatrix} + \begin{pmatrix} \varepsilon_{12} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n2} \end{pmatrix}$$

or in matrix notation

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \text{ and } \mathbf{Y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2\,.$$

The effect of the 1st SNP is β_{11} and effect of the 2nd SNP is β_{12}; the population mean, estimated in each model, will have different estimates: $\hat{\beta}_{01}$ and $\hat{\beta}_{02}$ from models at SNPs 1 and 2, respectively.

The fixed SNP-effects $\beta_{11},\ldots,\beta_{1p}$ are estimated by using the unbiased *Generalized Least Squares* (GLS) estimator, under the model assumption that it is the Best Linear Unbiased Estimator (BLUE) (Henderson, 1984):

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\hat{\beta}_{01},\hat{\beta}_{11})' = \left(\mathbf{X}_1'\mathbf{V}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{V}_1^{-1}\mathbf{Y}\\ &\vdots \\ \hat{\boldsymbol{\beta}}_p &= (\hat{\beta}_{0p},\hat{\beta}_{1p})' = \left(\mathbf{X}_p'\mathbf{V}_p^{-1}\mathbf{X}_p\right)^{-1}\mathbf{X}_p'\mathbf{V}_p^{-1}\mathbf{Y}.\end{aligned} \tag{2}$$

This estimates are unbiased, $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$.

For evaluation of the performance of SMR, the correlation between the estimates of marker effects from different equations $\mathrm{Cor}(\hat{\beta}_{1j},\hat{\beta}_{1k})$ for $j,k = 1,\ldots,p$, can be calculated, using the covariance matrix of both estimates $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j,\hat{\boldsymbol{\beta}}_k)$. The variance-covariance matrices $\mathrm{Var}(\hat{\boldsymbol{\beta}}_j)$ estimates can be derived analytically, using the assumptions of the SMR-model. For detailed derivation see Appendix A1.1.

In ***Multiple Marker Regression*** (Cohen, 1968; Kearsey and Farquhar, 1998; Meuwissen et al., 2001), similar to SMR-Model, the unknown marker effects are assumed to be fixed, but in contrast to the SMR-Model, all SNPs are included into one linear equation: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the design matrix $\mathbf{X}$ contains a vector of ones and genotype readings of all SNPs, $\boldsymbol{\beta}$ is the vector of SNP effects:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_{11} & \cdots & Z_{1p} \\ 1 & Z_{21} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1} & \cdots & Z_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3}$$

The estimation of all SNP effects is done simultaneously and prediction makes use of the complete genomic information, thus errors in estimates and prediction in the MMR-model are expected to be lower than in the SMR-model.

Both models SMR and MMR assume genomic effects to be fixed and both have similar model assumptions: residuals $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\mathbf{V})$ and $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0},\mathbf{V}_j)$ are normally distributed. The residual variance-covariance matrices $\mathbf{V} = \sigma^2\mathbf{I}_n$ and $\mathbf{V}_j = \sigma_j^2\mathbf{I}_n$ are assumed to be diagonal matrices with identical σ^2 and σ_j^2 on the diagonals, respectively. In the analysis of real data sets the unknown variance components σ^2 and σ_j^2 should be estimated from the data (mostly using maximum-likelihood procedures), while in the simulation studies we choose the magnitude of variance components. A further assumption is that design matrices $\mathbf{X}_j$ and $\mathbf{X}$ are non-stochastic and non-singular, meaning the determinants $\left|\mathbf{X}_j'\mathbf{X}_j\right| \neq 0$ and $\left|\mathbf{X}'\mathbf{X}\right| \neq 0$. Note, if some of the predictors are in perfect LD (or in mathematical terms in perfect collinearity), the rank of design matrix $\mathbf{X}$ will be smaller than *p* and the determinant of $\mathbf{X}'\mathbf{X}$ will be equal to zero.

Furthermore, a strong limitation of the MMR model is the restriction of the number of explanatory variables – in our case number of genomic markers *p* – which must not exceed the number of individuals *n*. Nowadays, the genomic data sets are often very large, thus *large-p-small-n* problem ($p >> n$) is omnipresent in genomic analysis. In case the number of predictors *p* exceeds the number of observations *n,* this assumption is violated, a unique solution could not be obtained in this situation.

Under the MMR model assumptions, the marker effects $\boldsymbol{\beta}$ can be estimated by using BLUE

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)' = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{4}$$

The expectation of these estimates is the vector of true effects $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and the variance of estimates can be computed analytically, as long as the phenotypic variance-covariance matrix $\mathbf{V}$ is known: $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$. For fixed effects, the variance-covariance matrix of the error in estimates $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is equal to the variance-covariance of estimates itself, i.e. $\mathrm{Var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathrm{Var}(\hat{\boldsymbol{\beta}})$. For comparisons with other linear models the correlation matrix $\mathrm{Cor}(\hat{\boldsymbol{\beta}})$ was also calculated. For detailed derivation see Appendix A1.2.

A ***Linear Mixed Model*** (Henderson, 1984) provides possibilities to model fixed effects as well as random genomic effects simultaneously:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \beta + \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1p} \\ Z_{21} & Z_{22} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{21} & \cdots & Z_{np} \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{5}$$

or in matrix notation:

$$\mathbf{Y} = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{\text{fixed effects}} + \underbrace{\mathbf{Z}\mathbf{u}}_{\text{randomeffects}} + \boldsymbol{\varepsilon}\,,$$

where $\boldsymbol{\beta}$ contains fixed effects and $\mathbf{X}$ is the corresponding design matrix of fixed effects and random marker effects are contained in vector $\mathbf{u}$ and corresponding genotypes are contained in a $n \times p$ matrix $\mathbf{Z}$.

Application of LMM to genomic data opens up the opportunity to account for various confounding factors, such as genetic relatedness, population structure or familial relatedness. For simplicity reasons just the population mean is modeled as fixed effects. Thus, in our studies, vector of fixed effects in LMM $\boldsymbol{\beta} \equiv \mu$ is one-dimensional. However it is possible to include more fixed covariates like age, gender, herd or time into the analysis.

The assumptions of the LMM are following:

- Variance matrices of random effects $\mathrm{Var}(\mathbf{u}) = \mathbf{G} = \sigma_u^2\mathbf{I}$ and for the error term $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{R} = \sigma_\varepsilon^2\mathbf{I}$ are known.
- Residuals $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\mathbf{R})$ and marker effects $\mathbf{u} \sim N(\mathbf{0},\mathbf{G})$ follow normal distributions and are stochastically independent.

Using these assumptions, the phenotypic variance matrix $\mathrm{Var}(\mathbf{Y}) =: \mathbf{V} \in R^{n \times n}$ can be derived analytically from the model: $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$.

While the fixed effects $\boldsymbol{\beta}$ can be estimated by using BLUE: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'V}^{-1}\mathbf{Y}$ with expectation $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and variance $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1}$ (e.g. Henderson, 1984), random effects in the LMM can be predicted by using the Best Linear Unbiased Predictor (BLUP) (Henderson, 1953):

$$\hat{\mathbf{u}} = \mathbf{GZ'}\,\mathbf{V}^{-1}\mathbf{QY}, \text{ with } \mathbf{Q} := \mathbf{I} - \mathbf{X}\left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'V}^{-1} \qquad (6)$$

Expectation of random marker effects $\mathbf{u}$ and of its prediction $\hat{\mathbf{u}}$ is equal to zero and the variance-covariance matrix of predictions is of the form $\mathrm{Var}(\hat{\mathbf{u}}) == \mathbf{GZ'}\,\mathbf{V}^{-1}\mathbf{QZG}$ and is equal to the covariance between the true random marker effects and their predictions $\mathrm{Cov}(\hat{\mathbf{u}},\mathbf{u})$. In case number of parameters is large, BLUP can still be used instead of BLUE if there are indications for fixed SNP effects. Furthermore, BLUP is able to capture the relatedness in sample and improve in that way the accuracy of prediction (Piepho et al., 2008).

Applying these results, the variance-covariance of the difference $\hat{\mathbf{u}}-\mathbf{u}$ between the true and predicted random effects $\mathrm{Var}(\hat{\mathbf{u}}-\mathbf{u}) = \mathbf{G}-\mathbf{GZ'V}^{-1}\mathbf{QZG}$, the covariance $\mathrm{Cov}(\hat{\mathbf{u}}-\mathbf{u},\hat{\mathbf{u}}) = 0$ between the random effects, prediction $\hat{\mathbf{u}}$, predictive error $\hat{\mathbf{u}}-\mathbf{u}$, and the corresponding correlation matrices were derived. For detailed derivation see Appendix A1.3.

Note, that the design matrices $\mathbf{X}$ in SMR, MMR and LMM are different.

Evaluation of performance of SMR, MMR and LMM in estimations and predictions

To quantify the impact of LD on effect estimates at each individual SNP, correlations $\mathrm{Cor}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_j)$ in SMR and $\mathrm{Cor}(\hat{\boldsymbol{\beta}})$ in MMR, and $\mathrm{Cor}(\hat{\mathbf{u}}-\mathbf{u})$ and $\mathrm{Cor}(\hat{\mathbf{u}})$ in LMM were applied. Correlation matrixes corresponding to the variance-covariance matrices in all models were obtained by standardizing the covariance by square root of product of the appropriate variances.

The correlation between predicted and true phenotype $\mathrm{Cor}(\mathbf{Y},\hat{\mathbf{Y}})$ and the mean squared error $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ was used to evaluate the goodness of fit of considered models.

Simulation of Genomic Data with a predefined LD structure

To compare all three models introduced above we used simulations. A SNP data set with a predefined LD structure was required to investigate the impact of association between the SNPs on the estimates and prediction in different statistical models. The SNPs were generated for different values of minor allele frequency (MAF): MAFs were varied in steps of 0.05 in the range from $p = 0.05$ to $p = 0.5$. For each combination of parameters we generated a data set Z of 100.000 independent individuals with a 15-SNP sequence per individual.

The genotypes were generated so that LD estimates (measured in r^2) between the first SNP and SNPs 2 to 15 were fixed; so that the highest LD was between the first and second SNP whereas the lowest LD was between the first and last (15th) SNP.

The simulation of genomic data in our study was performed by using a method, based on interpretation of random uniformly distributed variable as a gamete. For a given squared correlation between two loci (r^2) and known minor allele frequencies (p_1 and p_2) the resulting disequilibrium coefficient becomes $D(p_1, p_2, r^2) = \sqrt{r^2 p_1(1-p_1)p_2(1-p_2)}$, which was used to generate genotypes in pre-defined LD. Further, a representation of gametic frequencies using a uniformly distributed random variable on a unit interval leads to the needed genotypes with a fixed degree of association $D(p_1, p_2, r^2)$. In this way we generate two loci that are in pre-defined LD by using independent uniformly distributed random variables. We extend this method for more than 2 SNPs by shifting the limits on the unit interval. This method has been demonstrated as most reliable of four considered methods.

To be sure that the desired LD structure was imparted to the simulated data, four different methods for generating SNP data were tested. Detailed description of all four simulation methods as well as the performance (in terms of correlation structure of generated SNPs) of simulation methods mentioned above is given in Appendix A2.1-A2.4. The methods for generating correlated genotypes were compared for their precision in reproducing the given correlation structure in simulated marker data sets.

Simulation of Phenotypes

The next step was to construct the phenotypes for comparisons of linear regression models. Two different true effect models were considered for the construction of phenotypes: a random homoscedastic (the variances σ_j^2 at different SNPs are equal) true model (*RAND)* and a fixed true model (*FIX)*. A heteroscedastic (variance components σ_j^2 may vary across different loci) random model was also applied for the purpose of sensitivity analysis. Results of comparisons using this true model do not differ very much from *RAND*-scenarios.

Random true model: Assuming that the SNP effects were random, we chose LMM as the true model. Using the R-package *mvtnorm* (Genz et al., 2014), a normally distributed vector of effects term $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ and an independent vector of random errors $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$ were generated, where $\mathbf{G} = \sigma_u^2 \mathbf{I}_{15}$ and $\mathbf{R} = \sigma_\varepsilon^2 \mathbf{I}_{500}$ are the variance-covariance matrices of SNP effects and error term, respectively. We added to the random effect at SNP 5 a value of $\alpha = 1$. Finally we set fixed effect to $\boldsymbol{\beta} = \mu = 1$, so that vector of phenotypes $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ and its variance-covariance matrix $\mathrm{Var}(\mathbf{Y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ could be derived from the LMM according to equation (5).

Fixed true model: In *FIX*-scenario the SNP-effects were assumed to be fixed, therefore MMR was stated as the true model. All marker effects were set to zero, except the effect at the SNP 5, which was set to $\alpha = 1$. Assuming a population mean $\mu = 1$, vector of true marker effects becomes $\boldsymbol{\beta} = (1,0,0,0,0,1,0,\ldots,0)' \in R^{1+15}$. According to equation (3), vector of phenotypes $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ was constructed as a sum of the product of design matrix $\mathbf{X}$ and the vector of true effects $\boldsymbol{\beta}$, and the normally distributed vector of errors $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\mathbf{V})$.

For all scenarios, variance components were calculated based on the heritability: we chose σ_u^2 and σ_ε^2 so that heritability $h^2 = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$ took different values of 0.3, 0.5 and 0.7 . In each simulation loop a sample of genotypes $\mathbf{Z}$ of size $n = 500$ was taken from the generated data set Z and phenotypes were calculated according to the true models. Then we estimated $\boldsymbol{\beta}$ and $\mathbf{u}$, the variance-covariance matrix of predictor $\mathrm{Var}(\hat{\mathbf{u}})$ and that of errors in prediction $\mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u})$ in LMM, variance-covariance matrix of estimates $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ in MMR, the covariance between the estimates of marker effects $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_j)$ for $j = 1,\ldots,p$ in SMR, as well as the corresponding correlation matrices. Empirical sampling variance-covariance and correlation matrices for estimates $\hat{\boldsymbol{\beta}}$, $\hat{\beta}_{1j}$ and predictions $\hat{\mathbf{u}}$ and 95% confidence intervals are obtained from $n_{sim} = 2500$ repetitions (see in appendix A3) and compared with variance-covariance and correlation matrices expected in each model.

Statistical analysis as well as generation of genotype and phenotype data were performed using R (R Core Team, 2014). For generating multivariate normal distributed vectors in normal-truncated method the R-package *mvtnorm* (Genz et al., 2014) was used and for creating genotypes in copula-based method the R-package *copula* (Hofert et al., 2014) was used.

Results and Discussion

Impact of LD on estimates and predictions of marker effects in different models

In all considered models and across all scenarios, a clear impact of the amount of LD between the loci on precision of estimates of marker effects at each single locus was observed. The results achieved in a RAND scenario with heritability of $h^2 = 0.5$ and $\mathrm{MAF} = 0.05$ at all loci are represented in Figure 2.1.

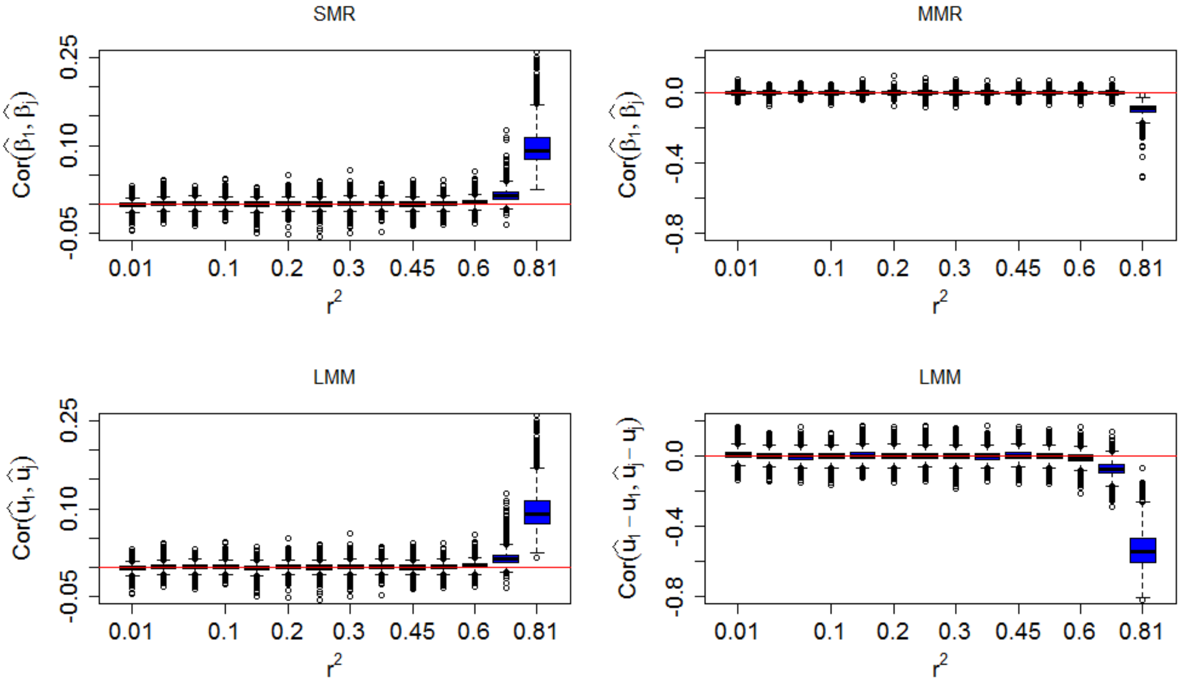


Figure 2.1. Boxplots of correlation between estimates and between predictions of marker effects, achieved from SMR, MMR and LMM in the RAND scenario with $\mathrm{MAF}=0.05$ **and heritability** $h^2=0.5$**.** The correlation coefficients between the estimates $\mathrm{Cor}(\hat{\beta}_1,\hat{\beta}_j)$ in SMR and MMR, and correlation coefficients between the predictors $\mathrm{Cor}(\hat{u}_1,\hat{u}_j)$ and errors in predictions $\mathrm{Cor}(\hat{u}_1-u_1,\hat{u}_j-u_j)$ in LMM at *1st* locus and at *jth* locus, $j=2,\ldots,15$ are plotted against the corresponding amount of LD denoted by r^2.

In all models, no impact of LD was detected on the estimates and predictions of marker effects, as long as amount of LD did not exceed the level of $r^2=0.7$. Depending on the model, LD higher than a model specific limit value had a noticeable effect on estimates and predictions and led to a decrease in their precision. The correlation between the estimates in SMR $\mathrm{Cor}(\hat{\beta}_1,\hat{\beta}_j)$ and between the predictions in LMM $\mathrm{Cor}(\hat{u}_1,\hat{u}_j)$ on average took values of about 0.1 and seemed to capture LD structure in the data when the LD level exceeded $r^2\approx 0.6$. The correlation in MMR $\mathrm{Cor}(\hat{\beta}_1,\hat{\beta}_j)$, which reflect errors in estimates, as well as the correlation of predictive errors $\mathrm{Cor}(\hat{u}_1-u_1,\hat{u}_j-u_j)$ in LMM turned negative as soon as the threshold of harmful LD level was exceeded. The negative correlation in errors of estimation and prediction indicate that the overestimation at one locus will be followed by underestimation at the second locus and *vice versa*. The thresholds for harmful LD levels were different in both multi-locus methods: in LMM the influence of collinearity between the loci was noted for $r^2\geq 0.6$, while in the MMR model this influence was observed when the value of $r^2 \approx 0.8$ was reached. While in the MMR model 95% of the correlation coefficients were situated between -0.03 and -0.18, in the LMM about 95% of the correlation coefficients were observed between -0.25 and -0.8.

The correlations between estimates or predictions of marker effects, visualized in Figure 2.1, were derived based on assumptions in each model, discussed in Material and Methods section, and on utilization of the known phenotypic variances and its components (residual and random effects variances). Figure 2.2 displays the same boxplots as in Figure 2.1 with the sample correlation coefficient and its confidence intervals drawn in addition.

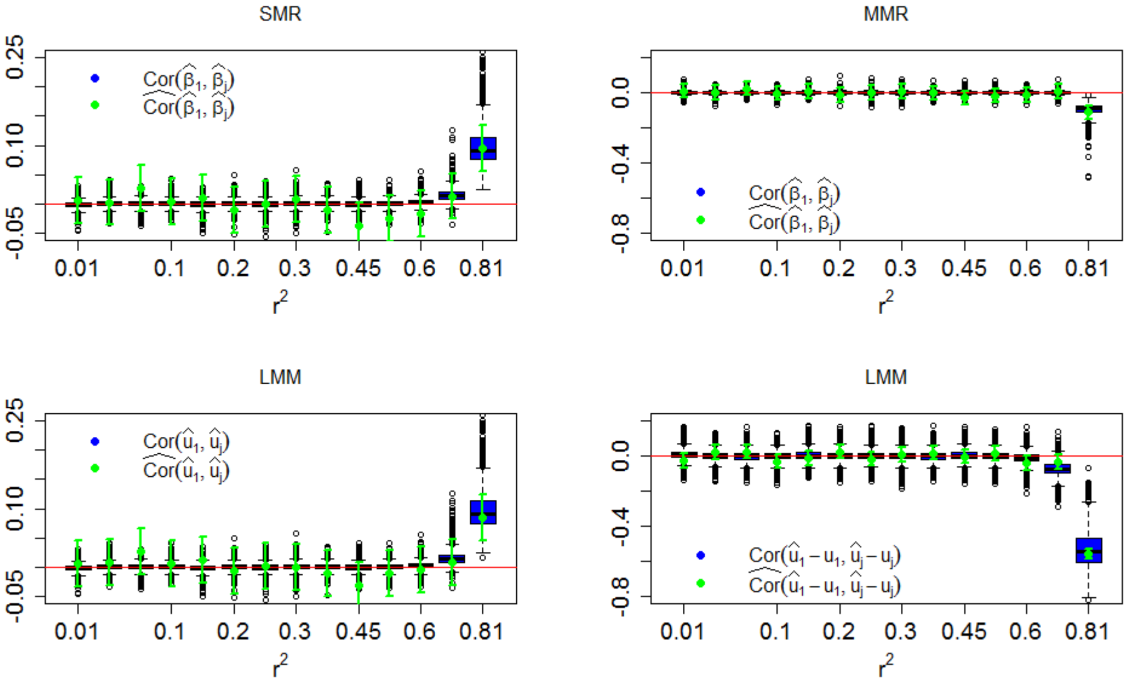


Figure 2.2. Boxplots of correlation between estimates and between predictions of marker effects, achieved from SMR, MMR and LMM and the sample correlation coefficients with corresponding 95 % confidence intervals. The correlation coefficients between the estimates $\mathrm{Cor}(\hat\beta_1, \hat\beta_j)$ in SMR and MMR, and correlation coefficients between the predictors $\mathrm{Cor}(\hat u_1, \hat u_j)$ and errors in predictions $\mathrm{Cor}(\hat u_1 - u_1, \hat u_j - u_j)$ in LMM at *1st* locus and at *jth* locus, $j = 2,\ldots,15$ are plotted against the corresponding amount of LD denoted by r^2. Results are achieved in the RAND scenario with $\mathrm{MAF} = 0.05$ and heritability $h^2 = 0.5$. The sample correlation coefficients and corresponding 95% confidence intervals are drawn in green.

The sample correlation coefficient and corresponding 95% confidence intervals are calculates using samples from $n_{sim} = 2{,}500$ repetitions and known true marker effects (calculation procedure and more details in Appendix A2). In all regression models, the expected correlation coefficients were confirmed by the empirical ones. For all models, the sample correlation coefficient was clearly scattered around zero and without exception, zero was included into the confidence intervals for all pairs of SNPs with values of $r^2 < 0.8$.

One of the parameters varied across the scenarios was the minor allele frequency, because MAF was expected to affect the severity of consequences of LD. Figures 2.1 and 2.2 pertain to the simulation scenarios with MAF fixed at 0.05, whilst in our studies different scenarios with MAF increasing in steps of 0.05 from 0.05 to 0.5 were performed. In Figure 2.3, results for MMR and LMM for scenarios with two extreme MAF values and heritability $h^2 = 0.5$ are shown, which are representative for the trends observed across all models and scenarios.

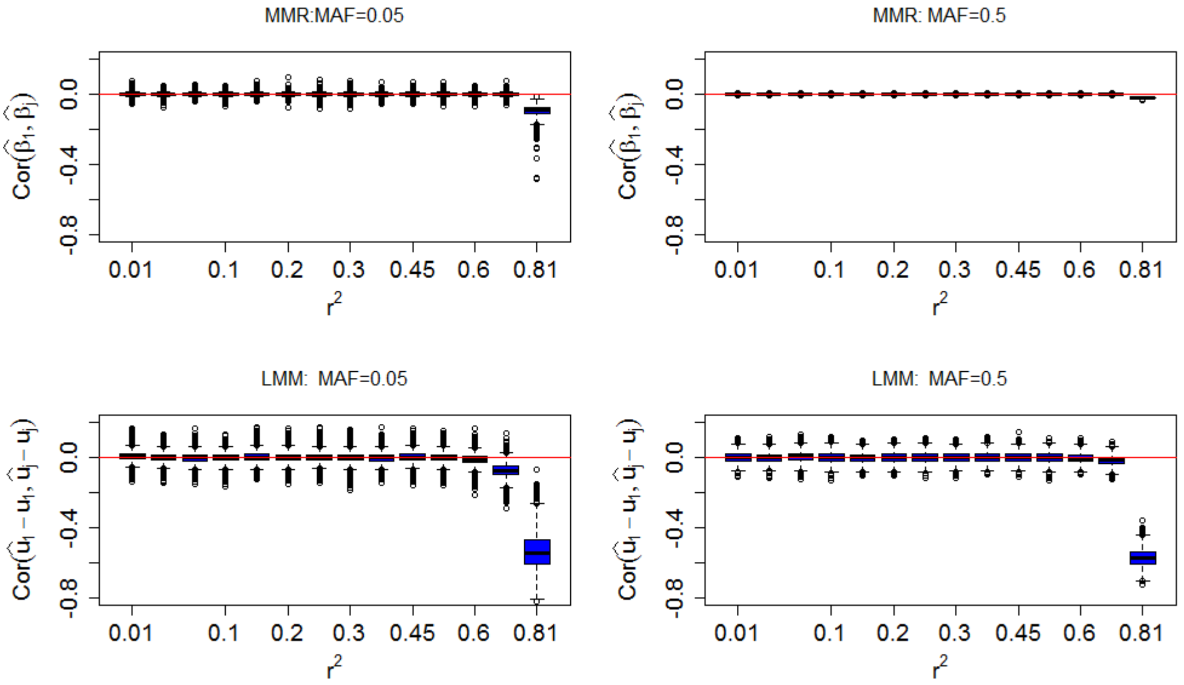


Figure 2.3. Boxplots of correlation between estimates and between predictions of marker effects, achieved from MMR and LMM in the RAND scenario with heritability $h^2 = 0.5$ for $\mathrm{MAF} = 0.05$ and $\mathrm{MAF} = 0.5$. The correlation coefficients between the estimates $\mathrm{Cor}(\hat\beta_1, \hat\beta_j)$ in MMR, and correlation coefficients between the errors in predictions $\mathrm{Cor}(\hat u_1 - u_1, \hat u_j - u_j)$ in LMM at *1st* locus and at *jth* locus, $j = 2, \ldots, 15$ are plotted against the corresponding amount of LD denoted by r^2.

The extent of LD influenced the precision of estimates much more strongly in the lower MAF scenarios in all three models; also the threshold for the extent of harmful LD increased with increasing MAF. The 95% of correlation coefficients between the estimates from MMR were observed between -0.03 and -0.18 when MAFs were fixed at 0.05, while this interval shrunk to $[-0.01,-0.035]$ when MAFs were fixed at 0.5. Same trends were observed also in the SMR model. While the impact of allele frequencies was less pronounced in the LMM, the influence of LD on estimates was still high for common variants when MAFs were equal to 0.5, and 95% of correlations of predictive errors at two loci $\mathrm{Cor}(\hat{u}_1 - u_1, \hat{u}_j - u_j)$ took values from -0.42 to -0.67. However, for common variants, the threshold for harmful LD shifted to $r^2 \approx 0.8$ in LMM and the intensity of dispersion was clearly lower than that when MAF=0.05.

Another factor which may influence extent of losses in precision of effect estimates caused by multicollinearity in the data, is the heritability of the trait. We considered three different scenarios for heritability $h^2 \in \{0.3,\ 0.5,\ 0.7\}$. In Figure 2.4, comparison of results for all values of heritability and MAF=0.05 is shown. In both regression models that assume the marker effects to be fixed - the SMR and MMR models - traits with higher heritability were less affected by the multicollinearity between the regressors. In MMR, the correlation between the estimators decreased with increasing heritability: for a trait with heritability of $h^2 = 0.3$, 95% of correlations between errors of estimates are located between -0.05 and -0.35 with a mean at -0.18 (central panel of Figure 2.4, left), whereas for a trait with much higher heritability of $h^2 = 0.7$, the correlations were observed between -0.005 and -0.09 with a mean at -0.04 (central panel of Figure 2.4, right). Analogous results were observed in the SMR. In contrast to the MMR model, the correlation between the errors in prediction from the LMM model were not affected by the different heritabilities of the traits and remained at a high level: about 50% of correlation coefficients were situated between -0.4 and -0.6.

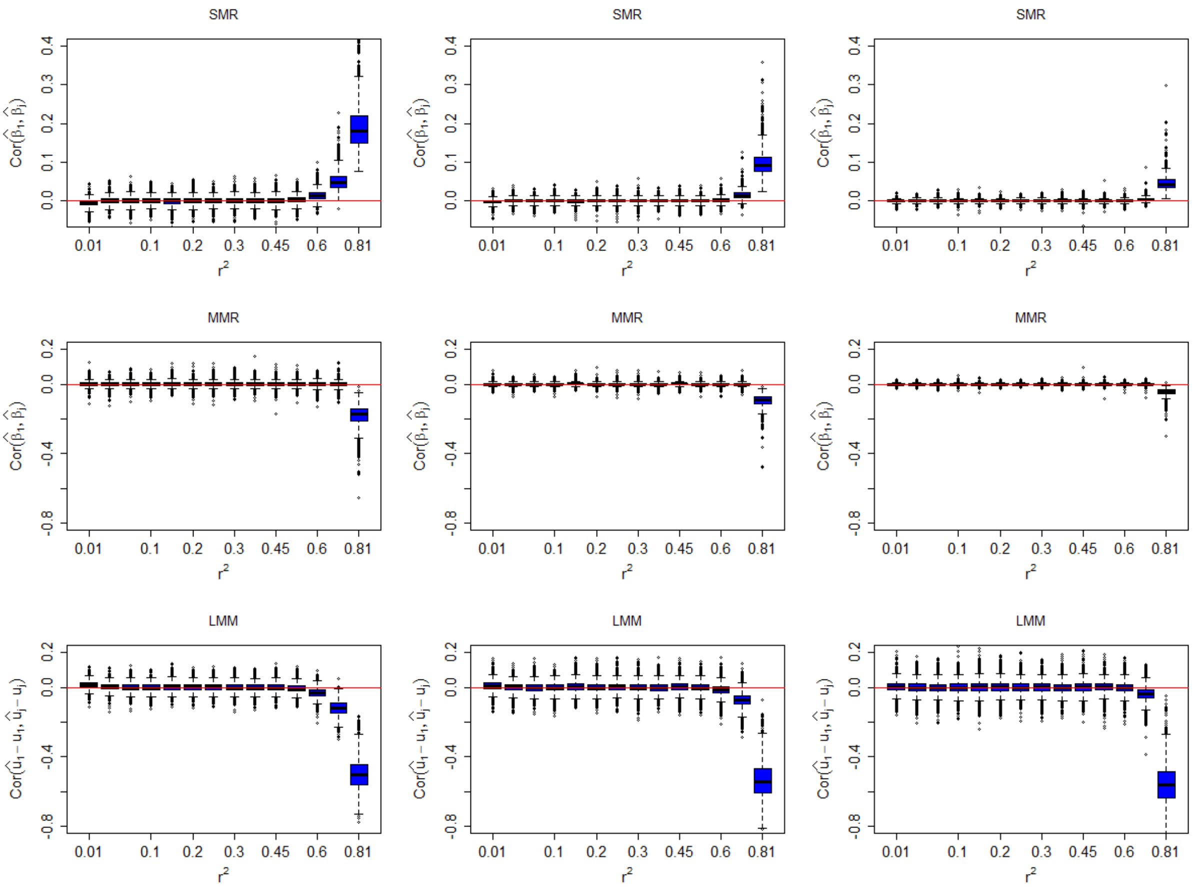


Figure 2.4: Boxplots of correlation between errors in estimates and between predictions of marker effects for different values of heritability. The correlation between errors in estimates $\mathrm{Cor}(\hat{\beta}_1, \hat{\beta}_j)$ from SMR are shown in the upper panel and $\mathrm{Cor}(\hat{\beta}_1, \hat{\beta}_j)$ from MMR in the central panel, in lower panel the correlations between predictive errors $\mathrm{Cor}(\hat{u}_1 - u_1, \hat{u}_j - u_j)$ in LMM are presented. All results are achieved in a RAND scenario with $\mathrm{MAF} = 0.05$ and values of heritability $h^2 = 0.3$ (left), $h^2 = 0.5$ (center) and $h^2 = 0.7$ (right). $\mathrm{Cor}(\hat{\beta}_1, \hat{\beta}_j)$ and $\mathrm{Cor}(\hat{u}_1 - u_1, \hat{u}_j - u_j)$ at *1st* locus and at *jth* locus, $j = 2, \ldots, 15$ are plotted against the corresponding amount of LD denoted by r^2.

Until now only results from simulation studies based on RAND scenario were reported. In Figure 2.5 results based on RAND or on FIX scenarios are introduced.

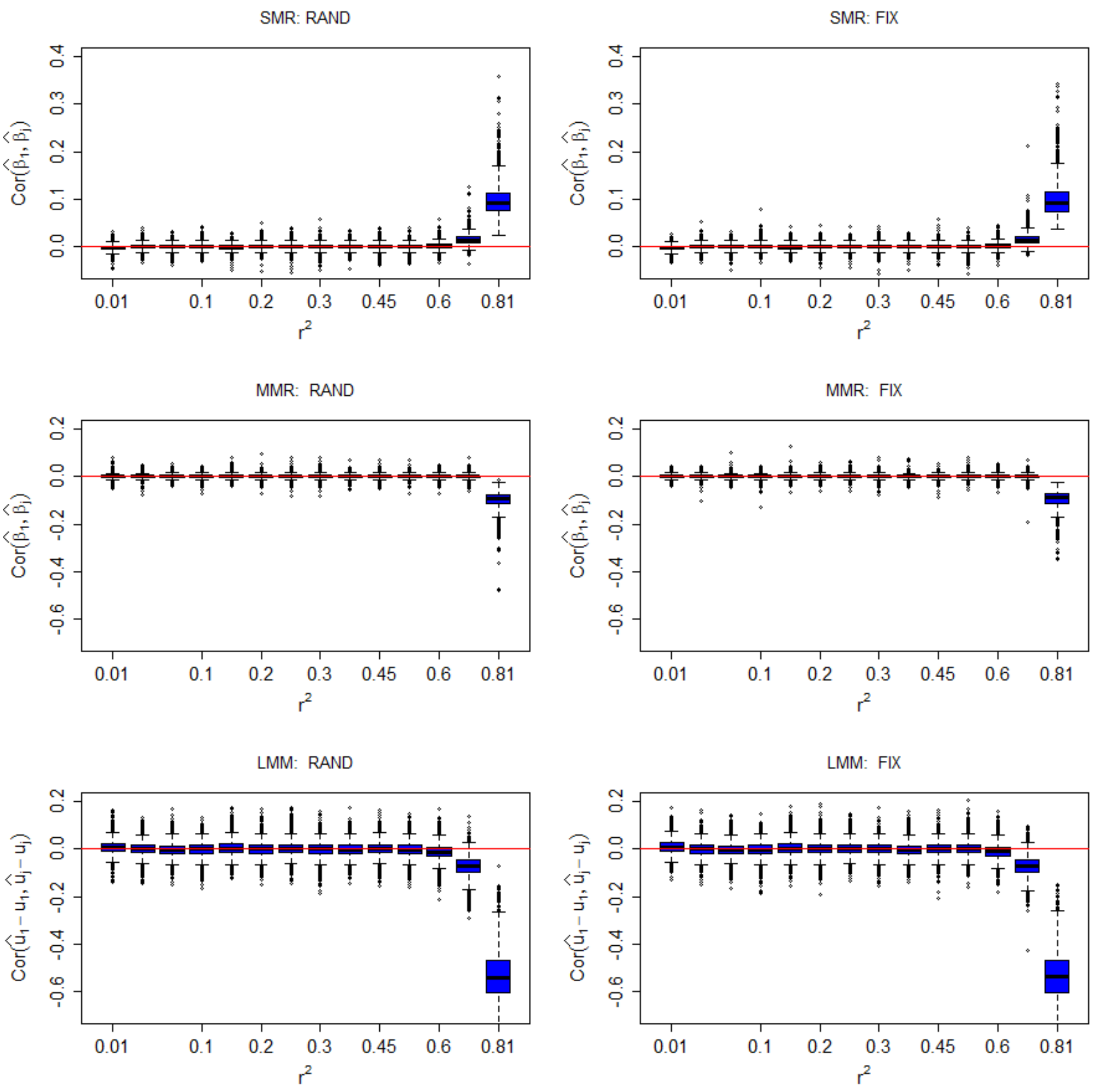


Figure 2.5. RAND versus FIX scenarios: boxplots of correlation between estimates of marker effects and error in predictions of marker effects from LMM, with heritability $h^2 = 0.5$ **for** $\mathrm{MAF} = 0.05$. The correlation between errors in estimates $\mathrm{Cor}(\hat{\beta}_1, \hat{\beta}_j)$ from SMR are shown in the upper panel and $\mathrm{Cor}(\hat{\beta}_1, \hat{\beta}_j)$ from MMR in the central panel, in lower panel the correlation between predictive errors $\mathrm{Cor}(\hat{u}_1 - u_1, \hat{u}_j - u_j)$ in LMM are presented. $\mathrm{Cor}(\hat{\beta}_1, \hat{\beta}_j)$ and $\mathrm{Cor}(\hat{u}_1 - u_1, \hat{u}_j - u_j)$ at *1st* locus and at *jth* locus, $j = 2, \ldots, 15$ are plotted against the corresponding amount of LD denoted by r^2.

The scenario with MAF=0.05 and $h^2 = 0.5$ was chosen as representative given that in other scenarios with different values of heritability or MAF the same trends were observed: no perceptible effect of a chosen true model on the performance of considered models was detected either in model derived correlations of estimates and predictions or in sample correlation coefficients

Impact of LD amount in data on goodness of fit in different models

In the interest of completeness, the potential impact of LD between the loci on goodness of fit of all three models under different simulation scenarios was investigated. In Figure 2.6 the MSE of predictions under a heritability $h^2 = 0.5$ are plotted against MAF, the MSE in RAND scenario is illustrated in the upper panel, whilst MSE in FIX scenario is shown in the lower panel.

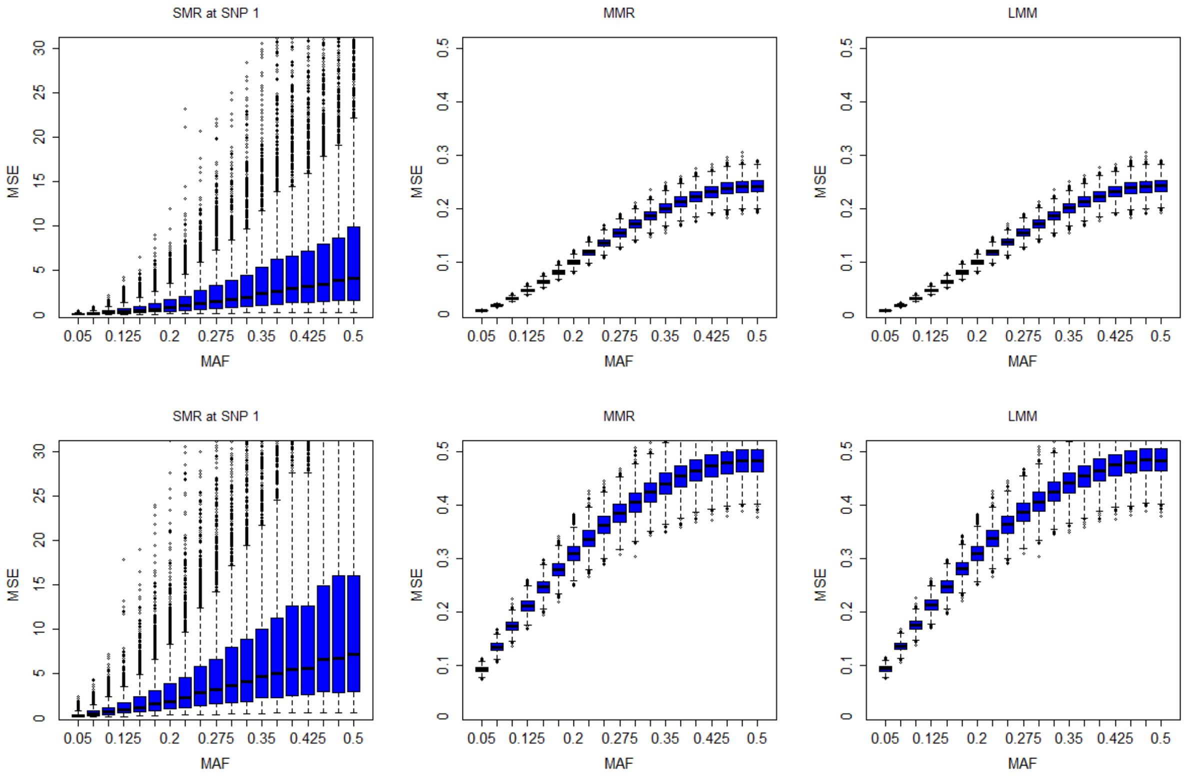


Figure 2.6. Boxplots of MSE in RAND (upper panel) versus FIX (lower panel) true models. MSE was plotted against the MAF for SMR (left diagrams), for MMR (central diagrams) and for LMM (right diagrams). Scenarios with heritability $h^2 = 0.5$ were considered.

Obviously, allele frequency of markers had a strong impact on goodness of fit of all considered models: the MSE is smaller for infrequent variants compared to the MSE for common variants. While the magnitude of MSE in LMM and MMR models is comparable. The choice of the true model had an impact on goodness of fit of all regression models;

with fixed true effects up to two times higher MSE was measured across compared models and MAFs, in comparison to random true effects.

Also, the dependence of MSE on heritability of a trait was investigated, which is illustrated in Figure 2.7 on behalf of an example of MMR as representative for all three models.

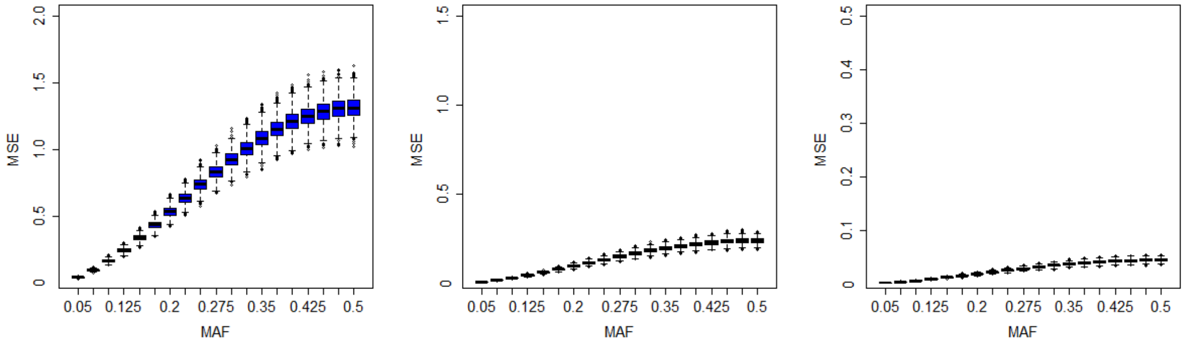


Figure 2.7. Boxplots of MSE obtained from in RAND scenario in a MMR model, plotted against the MAF for heritability $h^2 = 0.3$ (left), $h^2 = 0.5$ (center) and $h^2 = 0.7$ (right).

All three models showed similar trends for MSE in dependence on different MAFs, accompanied by different absolute values of MSE across the range of MAFs. Obviously, the goodness of fit of all models is strongly influenced by the heritability of the trait: the higher the heritability of the trait, the smaller the MSE of predictions $\hat{Y}$. The goodness of fit improved in all compared models if the heritability of the trait was greater, however this effect was less pronounced in the SMR model compared to LMM and MMR models.

Finally, the correlation between the true and predicted phenotype was investigated. The $Cor(Y,\hat{Y})$, plotted against the MAF, for scenarios with heritability of the trait fixed to 0.5 across models is represented in Figure 2.8 for RAND scenarios (upper panel) and FIX scenarios (lower panel). No differences between the RAND and FIX scenarios were observed in the SMR model: the SMR performed poorly, in contrast to comparable goodness of fit in LMM and MMR models. The whole genome models MMR and LMM showed small differences for MAFs up to a value of 0.2, for more frequent variants with MAF greater than 0.2 no differences in goodness of fit between RAND and FIX scenarios were detected.

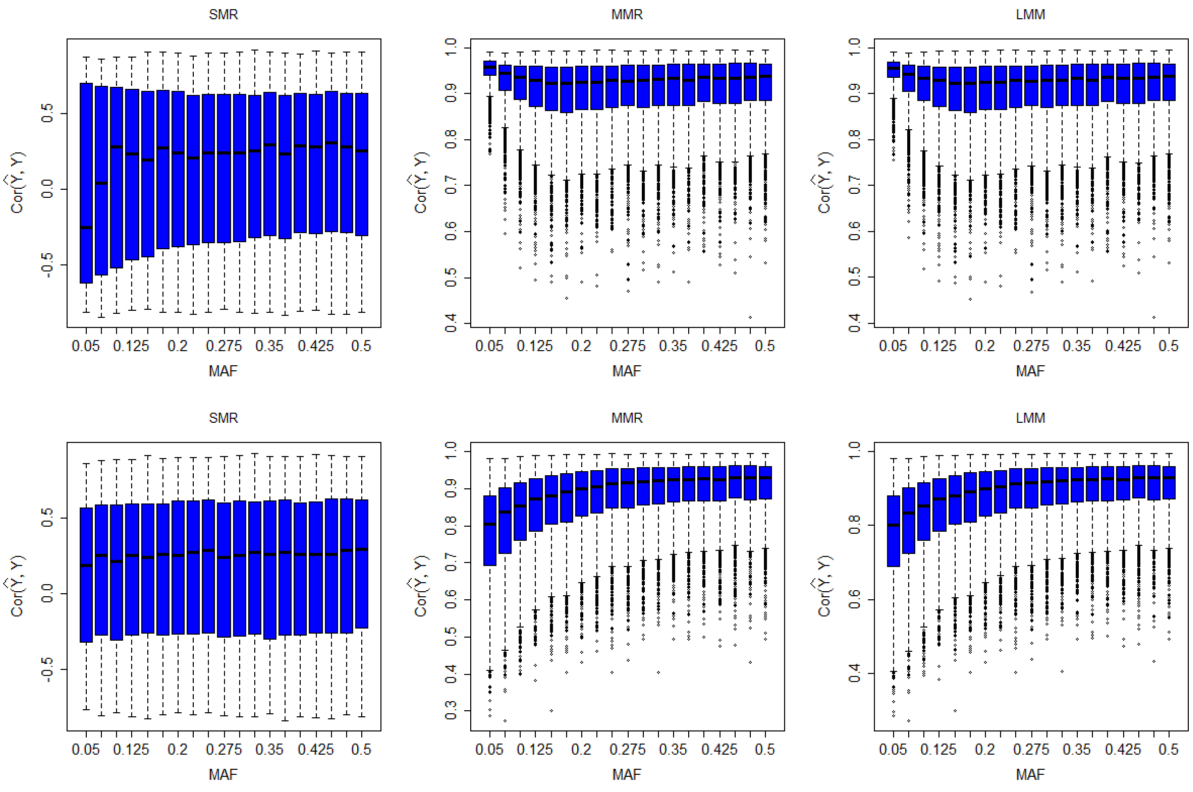


Figure 2.8. Boxplots of correlation between true and predicted phenotypes in RAND (upper panel) versus FIX (lower panel) scenarios. MSE was plotted against the MAF for SMR (left diagrams), for MMR (central diagrams) and for LMM (right diagrams). Scenarios with heritability $h^2 = 0.5$ were considered.

Finally, the impact of different levels of heritability of the trait on correlation between true and predicted phenotype was considered. In Figure 2.9, the correlations $Cor(Y,\hat{Y})$ for LMM at different values of heritability are plotted against the MAFs. The higher heritability had a positive effect on the goodness of fit and also minimized the dispersion of correlation coefficients: for heritability of $h^2 = 0.3$ the 95% of correlations $Cor(Y,\hat{Y})$ are observed from 0.3 to 0.98, while for $h^2 = 0.5$ this interval shrunk to 0.75 - 0.99.

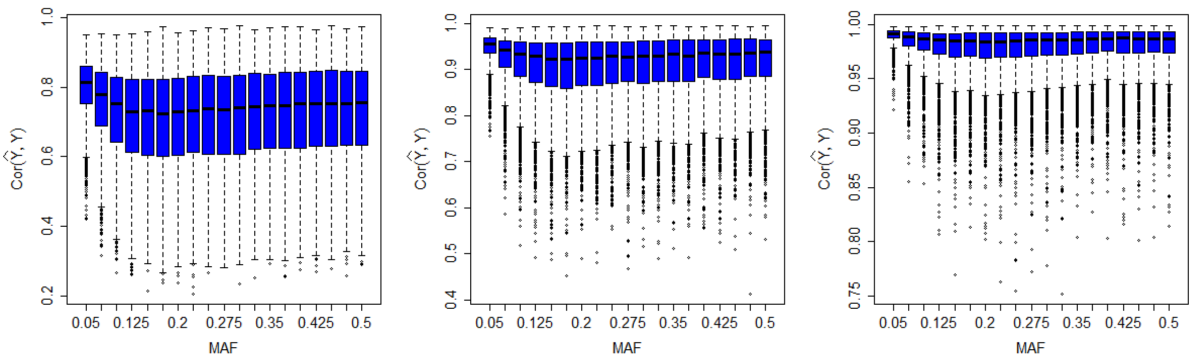


Figure 2.9. Boxplots of correlation between the predicted and true phenotype obtained from LMM in RAND scenario, plotted against the MAF for heritability $h^2 = 0.3$ (left), $h^2 = 0.5$ (center) and $h^2 = 0.7$ (right).

The instability of estimations due to the degree of multicollinearity detected in the present study and consequently the integrity of estimated genomic models is a serious issue. The results of this simulation study suggest that the multiple marker regression model was more robust against the multicollinearity in the data, and marker effect estimates from MMR were less affected by increased LD than those from the LMM. Also in comparison to SMR, MMR provided more reliable estimates and the threshold of harmful LD level between the loci was much lower.

This led to the conclusion that the MMR is a better approach to estimate the marker effects and consequently to map the quantitative trait loci (QTL). The main limitation of MMR that inhibited its application as a QTL mapping tool, is the restriction that the number of explanatory variables must be smaller than the sample size.

Limitations of simulated genotype data

The simulation method of our choice does have some minor limitations. In reality, minor allele frequencies aren’t the same at all loci. This assumption was made since a large impact of differences in MAFs on measures of LD will complicate the assignment of observed effects on estimates only to the association between the loci. Furthermore, it is well-known that MAF, especially the difference in MAFs, strongly influences the range of achievable LD. In our preliminary studies, a two-locus model was considered and also a scenario with different MAF at both loci. No general difference was observed in comparison to scenarios with the same MAF at both loci, until the whole spectrum of r^2 was not available. Another disadvantage of chosen simulation method is the unrealistic structure of the data: the wanted correlation structure between the markers is obtained by shifting the δ parameter, so that genotypes at each individual are increasing (e.g. 0 0 0 0 0 1 1 1 1 1 2) or decreasing (e.g. 2 2 2 1 1 0 0 0 0 0 0). This prompted us initially to look for a method for creating SNP-data that captures the pre-defined correlation structure and has a more realistic appearance of the genotypes. However, all other considered simulation methods showed less reliable results and did not capture LD structure as well as the method based on definition of gametic disequilibrium. Thus, we decided to use a method with less realistic appearance of genotypes, but with exactly reproduced LD structure.

Implications

While the rapid development of molecular genetics has resulted in high density genomic data, this is accompanied by methodological and computational difficulties associated with handling this amount of information. The other issue with high dimensionality of genomic data is multicollinearity, which plays a significant role in the performance of estimators of marker effects. The eigenvalues of the genotype matrix provide the possibility of not only detecting but also addressing the magnitude of multicollinearity in the real data sets. For instance, the influence of multicollinearity in MMR can be examined by using eigenvalues or a ratio of eigenvalues, so-called condition numbers, of $\mathbf{X'X}$ or $\mathbf{X'V}^{-1}\mathbf{X}$ (Wang et al., 1990). Several historical approaches, such as variable selection or principal components regression have been proposed to minimize and overcome the multicollinearity in the data. Methods aimed at reducing the model complexity could be summarized so as to help make a decision about which markers should be kept in the model. Therefore, there is a need to have a statistical method which guarantees reliable effect estimates and predictions independent of the amount of multicollinearity present without ad-hoc adjusting.

MMR has been shown to be a better approach than the SMR, which is a classical method for genome wide association studies (GWAS), as well as the LMM, which is often used for predictions for new individuals but not for QTL mapping.

The main problem with applying MMR as a QTL mapping tool is the assumption $p \leq n$. In most cases, this assumption cannot be fulfilled in a quantitative genetic context, where the data extends to several hundred thousands of markers and sample sizes of no more than a few thousand individuals. This so-called *large-p-small-n* problem and proposals for solutions are discussed by Ishwaran and Rao (2014). However, methods like ridge regression suggested by Hoerl and Kennard (1976), LASSO proposed by Tibshirani (1996) and hybrids of both like elastic net (Zou and Hastie, 2005) are able to cope with the multicollinearity problem and can be the method of choice for QTL mapping using the whole genome approach. However, further studies are needed to establish which of these methods is the most reliable.

It should be noted that the performance of estimators and predictors in linear regression models was examined only by using simulated data. The results of our studies indicate a strong impact of LD between the markers on predictions of random marker effects in linear mixed model. For instance, in a data set consisting on about 6,000 unrelated individuals of Caucasian origin the LD level at 95% of SNP pairs $r^2 \leq 0.47$, while in a data set consisting on 673 individuals of a highly selected White Leghorn chicken line 30% of SNP pairs $r^2 \geq 0.60$ and about 10% of SNP pairs $r^2 \geq 0.80$. Additional research using real genomic data can help us establish this hypothesis.

Appendix

A1: Variance-covariance matrices and corresponding correlation matrices in linear models

1.1. Variance-covariance matrices and corresponding correlation matrices derived from the SMR model

In a simple case of $p = 2$ two models for the same vector of phenotypes are described by:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1Z_{11} \\ 1Z_{21} \\ \vdots \\ 1Z_{n1} \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} \\ \beta_{11} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n1} \end{pmatrix} \text{ and } \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1Z_{12} \\ 1Z_{22} \\ \vdots \\ 1Z_{n2} \end{pmatrix} \cdot \begin{pmatrix} \beta_{02} \\ \beta_{12} \end{pmatrix} + \begin{pmatrix} \varepsilon_{12} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n2} \end{pmatrix}.$$

The fixed SNP-effects β_{11} and β_{12} are estimated by

$$\hat{\boldsymbol{\beta}}_1 = (\hat{\beta}_{01}, \hat{\beta}_{11})' = \left(\mathbf{X}_1^{'}\mathbf{V}^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{'}\mathbf{V}^{-1}\mathbf{Y} \text{ and } \hat{\boldsymbol{\beta}}_2 = (\hat{\beta}_{02}, \hat{\beta}_{12})' = \left(\mathbf{X}_2^{'}\mathbf{V}^{-1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{'}\mathbf{V}^{-1}\mathbf{Y}.$$

The variance-covariance matrices for each estimate as well as covariance matrix of both estimates can be derived analytically, by using the assumptions of the SMR-model:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_1) = \left(\mathbf{X}_1^{'}\mathbf{V}^{-1}\mathbf{X}_1\right)^{-1} = \begin{pmatrix} \mathrm{Var}(\hat{\beta}_{01}) & \mathrm{Cov}(\hat{\beta}_{01}, \hat{\beta}_{11}) \\ \mathrm{Cov}(\hat{\beta}_{11}, \hat{\beta}_{01}) & \mathrm{Var}(\hat{\beta}_{11}) \end{pmatrix}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_2) = \left(\mathbf{X}_2^{'}\mathbf{V}^{-1}\mathbf{X}_2\right)^{-1} = \begin{pmatrix} \mathrm{Var}(\hat{\beta}_{02}) & \mathrm{Cov}(\hat{\beta}_{02}, \hat{\beta}_{12}) \\ \mathrm{Cov}(\hat{\beta}_{12}, \hat{\beta}_{02}) & \mathrm{Var}(\hat{\beta}_{12}) \end{pmatrix}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2)=\left(\mathbf{X}_1^{'}\mathbf{V}^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{'}\mathbf{V}^{-1}\mathbf{X}_2\left(\mathbf{X}_2^{'}\mathbf{V}^{-1}\mathbf{X}_2\right)^{-1}=\begin{pmatrix}\mathrm{Cov}(\hat{\beta}_{01},\hat{\beta}_{02}) & \mathrm{Cov}(\hat{\beta}_{01},\hat{\beta}_{12})\\ \mathrm{Cov}(\hat{\beta}_{11},\hat{\beta}_{02}) & \mathrm{Cov}(\hat{\beta}_{11},\hat{\beta}_{12})\end{pmatrix}$$

To get a measure which is standardized for variance in estimates, the correlation between the estimates was calculated as $\mathrm{Cor}(\hat{\beta}_{11},\hat{\beta}_{12})=\dfrac{\mathrm{Cov}(\hat{\beta}_{11},\hat{\beta}_{12})}{\sqrt{\mathrm{Var}(\hat{\beta}_{11})\mathrm{Var}(\hat{\beta}_{12})}}$.

The calculations for 15 SNPs are done analogously for $\mathrm{Cor}(\hat{\beta}_{1,1},\hat{\beta}_{1,j})$, where $\hat{\beta}_{1,j}$ correspond to the estimate of marker effect at jth SNP for $j=1,\ldots,15$:

$$\mathrm{Cor}(\hat{\beta}_1,\hat{\beta}_j)=\frac{\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_j)}{\sqrt{\mathrm{Var}(\hat{\beta}_1)\mathrm{Var}(\hat{\beta}_j)}}$$

1.2. Variance-covariance matrices and corresponding correlation matrices derived from the MMR model

In a simple case of $p=2$, the marker effects $\boldsymbol{\beta}$ could be estimated from linear equation $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ by using BLUE: $\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2)'=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$. The expectation of these estimates is the vector of true effects $E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$ and variance of estimates is available analytically, as long as the phenotypic variance-covariance matrix $\mathbf{V}$ is known: $\mathrm{Var}(\hat{\boldsymbol{\beta}})=\mathrm{Var}\left(\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\right)=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathrm{Var}\mathbf{Y}\left(\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\right)^{'}=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$. For fixed effects, the variance-covariance matrix of the error in estimates $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ is equal to the variance-covariance of estimates itself:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\mathrm{Var}(\hat{\boldsymbol{\beta}})+\underbrace{\mathrm{Var}(\boldsymbol{\beta})-2\cdot\mathrm{Cov}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})}_{=0}=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}.$$

The correlation between the estimates was calculated similar to that in SMR:

$$\mathrm{Cor}(\hat{\beta}_1,\hat{\beta}_2)=\frac{\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)}{\sqrt{\mathrm{Var}(\hat{\beta}_1)\mathrm{Var}(\hat{\beta}_2)}}\ \text{or}\ \mathrm{Cor}(\hat{\beta}_1,\hat{\beta}_j)=\frac{\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_j)}{\sqrt{\mathrm{Var}(\hat{\beta}_1)\mathrm{Var}(\hat{\beta}_j)}},\ \text{for}\ j=2,\ldots,15,$$

where $\hat{\beta}_j$ correspond to the estimate of marker effect at jth SNP.

1.3. Variance-covariance matrices and corresponding correlation matrices derived from the LMM

In LMM fixed effects and random genomic effects are modeled simultaneously: $\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ contains fixed effects and $\mathbf{X}$ is the corresponding matrix of fixed effects and random marker effects are contained in vector $\mathbf{u}$ and corresponding genotypes are contained in a $n \times p$ matrix $\mathbf{Z}$. Using known $\mathrm{Var}(\mathbf{u}) =: \mathbf{G} = \sigma_u^2\mathbf{I}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) =: \mathbf{R} = \sigma_\varepsilon^2\mathbf{I}$, the phenotypic variance-covariance matrix could be derived analytically from the model: $\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}) = \mathbf{ZGZ'} + \mathbf{R} =: \mathbf{V} \in R^{n \times n}$

The fixed effects $\boldsymbol{\beta}$ could be estimated by using BLUE: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'V}^{-1}\mathbf{Y}$ with expectation $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and variance-covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1}$, similar to MMR and SMR models. The random effects in the LMM could be predicted by using the Best Linear Unbiased Predictor (BLUP):

$$\hat{\mathbf{u}} = \mathbf{GZ'\,V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{GZ'\,V}^{-1}\underbrace{\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'V}^{-1}\right)}_{=:\mathbf{Q}}\mathbf{Y} = \mathbf{GZ'\,V}^{-1}\mathbf{QY}$$

Expectation of random effect $\mathbf{u}$ and consequently of its prediction $\hat{\mathbf{u}}$ is equal to zero and the variance-covariance matrix is of the form:

$$\mathrm{Var}(\hat{\mathbf{u}}) = \mathrm{Var}(\mathbf{GZ'\,V}^{-1}\mathbf{QY}) = \mathbf{GZ'\,V}^{-1}\mathbf{Q}\mathrm{Var}(\mathbf{Y})[\mathbf{GZ'\,V}^{-1}\mathbf{Q}]' = \cdots = \mathbf{GZ'\,V}^{-1}\mathbf{QZG}.$$

The covariance between the true random effects $\mathbf{u}$ and its predictor $\hat{\mathbf{u}}$ is equal to the variance of predictor:

$$\mathrm{Cov}(\hat{\mathbf{u}},\mathbf{u}) = \mathrm{E}(\hat{\mathbf{u}}\cdot\mathbf{u}') - \mathrm{E}(\hat{\mathbf{u}})\cdot\underbrace{\mathrm{E}(\mathbf{u}')}_{=0} = \mathbf{GZ'\,V}^{-1}\mathbf{Q}\;\underbrace{\mathrm{E}(\mathbf{Y}\cdot\mathbf{u}')}_{=\mathbf{ZG}} = \mathbf{GZ'\,V}^{-1}\mathbf{QZG} = \mathrm{Var}(\hat{\mathbf{u}}),$$

here we used $\mathrm{E}(\mathbf{Y}\cdot\mathbf{u}') = \mathbf{X\beta}\;\underbrace{\mathrm{E}(\mathbf{u}')}_{=0} + \mathbf{Z}\;\underbrace{\mathrm{E}(\mathbf{uu}')}_{=\mathbf{G}} + \underbrace{\mathrm{E}(\boldsymbol{\varepsilon})\mathrm{E}(\mathbf{u}')}_{=0} = \mathbf{ZG} = \mathrm{Cov}(\mathbf{Y},\mathbf{u}')$.

Applying these results, the variance-covariance of the difference $\hat{\mathbf{u}} - \mathbf{u}$ between the true and predicted random effects $\mathrm{Var}(\hat{\mathbf{u}} - \mathbf{u}) = \mathrm{Var}(\hat{\mathbf{u}}) + \cdot\mathrm{Var}(\mathbf{u}) - 2\mathrm{Cov}(\hat{\mathbf{u}},\cdot\mathbf{u}') = \mathbf{G} - \mathbf{GZ'\,V}^{-1}\mathbf{QZG}$ as well as the covariance between the random effects prediction $\hat{\mathbf{u}}$ and predictive error $\hat{\mathbf{u}} - \mathbf{u}$ $\mathrm{Cov}(\hat{\mathbf{u}} - \mathbf{u},\hat{\mathbf{u}}) = \underbrace{\mathrm{Cov}(\hat{\mathbf{u}},\hat{\mathbf{u}})}_{=\mathrm{Var}(\hat{\mathbf{u}})} - \cdot\underbrace{\mathrm{Cov}(\hat{\mathbf{u}},\cdot\mathbf{u})}_{=\mathrm{Var}(\hat{\mathbf{u}})} = 0$ were derived.

The correlation matrixes $\mathrm{Cor}(\hat{\mathbf{u}}-\mathbf{u})$ and $\mathrm{Cor}(\hat{\mathbf{u}})$ are obtained by standardizing with appropriate variances. In our studies, the fixed effects $\boldsymbol{\beta}$ in LMM are represented only by population mean, correlation between the estimates of fixed effects $\mathrm{Cor}(\hat{\boldsymbol{\beta}})$ is not considered.

A2: Simulation methods for generating SNP-data with pre-defined LD structure

1.1. Simulation of SNP-data: definition of gametic disequilibrium-based method

Two biallelic loci with minor allele frequencies p_1 und p_2 are considered, which are in linkage disequilibrium with disequilibrium coefficient D. The gametic probabilities for all possible combinations of alleles at both loci are presented in Table A2.1:

Locus 2

		1	0	
Locus 1	1	$\underbrace{p_1 \cdot p_2 + D}_{=:a}$	$\underbrace{p_1 \cdot (1-p_2) - D}_{=:b}$	p_1
	0	$\underbrace{(1-p_1) \cdot p_2 - D}_{=:c}$	$\underbrace{(1-p_1) \cdot (1-p_2) + D}_{=:d}$	$(1-p_1)$
		p_2	$1-p_2$	1

Table A2.1. Gametic frequencies expressed by minor allele frequencies p_1 and p_2 and the disequilibrium coefficient D, the appearance of minor allele is coded as 1.

The relationship $a = p_1 - b$, $d = 1 - p_2 - b$ and $c = 1 - p_1 - d = p_2 - p_1 + b$ between the gametic frequencies a, b, c, d and allele frequencies p_1 and p_2 represented in Table A2.1, can be used for rewriting the expression $a \cdot d - b \cdot c$ as

$$a \cdot d - b \cdot c = (p_1 - b)(1 - p_2 - b) - b(p_2 - p_1 + b) = \cdots = D$$

Thus, the squared correlation between both loci is expressed by

$$r^2 = \frac{(a \cdot d - b \cdot c)^2}{\sqrt{p_1 \cdot (1-p_1) \cdot p_2 \cdot (1-p_2)}} = \frac{D^2}{\sqrt{p_1 \cdot (1-p_1) \cdot p_2 \cdot (1-p_2)}}.$$

For a desired squared correlation between two loci r^2 and known minor allele frequencies p_1 and p_2 the resulting disequilibrium coefficient can be expressed as $D(p_1, p_2, r^2) = \sqrt{r^2 p_1(1-p_1)p_2(1-p_2)}$.

For the purpose of simulation of genotypes, the gametic frequencies can be expressed in terms of uniformly distributed random numbers $U_j \sim Unif[0,1]$. To this end, the unit interval $(0,1)$ was divided by thresholds for gametes in four disjunctive segments:

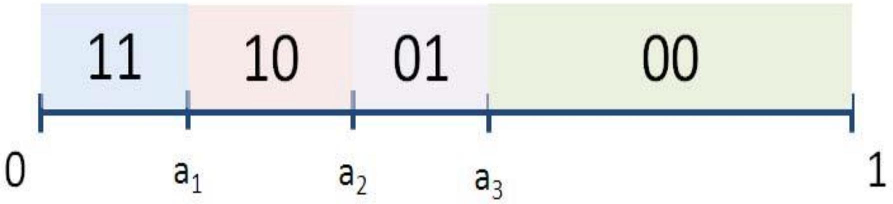


Figure A2.1. Unit interval, divided in four segments according to gametic frequencies.

The probability for a random variable U_j to take values between 0 and a_1 corresponds to the gametic frequency of the gamete**11**: $P(11) = a_1 - 0 = p_1 p_2 + D$. Thus, the threshold a_1 can be expressed by using disequilibrium coefficient $D(p_1, p_2, r^2)$ and minor allele frequencies p_1 and p_2 as $a_1 = p_1 p_2 + D$.

In analogy the thresholds a_2 and a_3 can be expressed by using D and p_1 and p_2 as

$$P(10) = a_2 - a_1 = p_1(1-p_2) - D \quad \Rightarrow \quad a_2 = p_1$$

$$P(01) = a_3 - a_2 = (1-p_1)p_2 - D \quad \Rightarrow \quad a_3 = p_1 + (1-p_1)p_2 - D$$

$$P(00) = 1 - a_3 = (1-p_1)(1-p_2) + D \quad \Rightarrow \quad a_3 = p_1 + (1-p_1)p_2 - D$$

Depending on the value of a randomly sampled uniform variable, it is located in one of the segments of unit interval and in this way we specified the gamete as **11**, **01**, **10** or **00**: two correlated haplotypes are obtained, viewed in genetic context as alleles at two different loci on one copy of the chromosome. Correlated haplotypes from the second chromosome copy could be obtained in the same way and the sum of minor allele counts at the two positions separately yields the desired genotypes with pre-defined correlation.

This procedure can be explained using a small example, where we assume that two loci are in LD so that $r^2 = 0.70$ and MAFs at both loci are set to $p_1 = p_2 = 0.4$.

The first step is to calculate the disequilibrium coefficient $D(p_1, p_2, r^2) = 0.202$ and the corresponding thresholds: $a_1 = 0.362$, $a_2 = 0.4$ and $a_3 = 0.439$.

In the second step two random uniform variable $U_1 = 0.21$ and $U_2 = 0.47$ are generated, the first one is smaller than a_1 which leads to the gamete **11** and the second random variable is larger than a_3, consequently the gamete **00** is obtained for the "second copy".

The genotype at locus 1 results in $1+0=1$ and genotype at locus 2 results in $1+0=1$, both are in LD so that $r^2 = 0.70$.

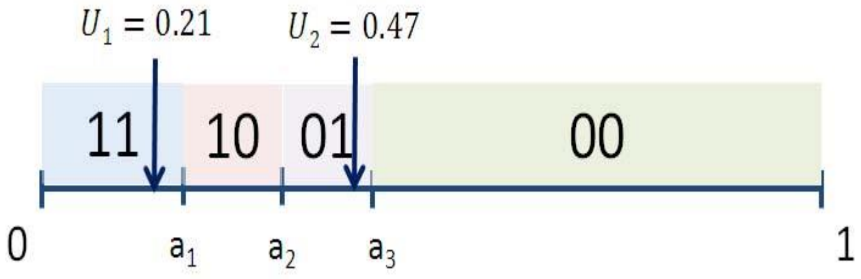


Figure A2.2. Example for generation of more than two SNPs with the predefined LD by using two uniform distributed random variables U_1 and U_2.

To extend this method to more than two SNPs, different thresholds a_1, a_2 and a_3 should be applied to two fixed uniformly distributed random variables U_1 and U_2. In the example above, two genotypes: **1** at locus 1 and **1** at locus 2 are created using realizations of random variable $U_1 = 0.21$ and $U_2 = 0.47$, so that the squared correlation between the genotypes at both loci is equal to $r^2 = 0.36$. We calculate new thresholds $a_1 = 0.304$, $a_2 = 0.4$ and $a_3 = 0.496$, corresponding to $r^2 = 0.36$. Now $U_2 = 0.47$ is located in the segment belonging to the gamete **01,** thus the genotype at locus 3 is$1+1=2$, while the genotype at locus 1 remains $1+0=1$, both genotypes are in LD so the squared correlation between the genotypes at both loci is equal to $r^2 = 0.36$. Obviously the genotype at the 1st SNP never changes; it is possible to generate any number of SNPs with a predefined correlation with the 1st SNP.

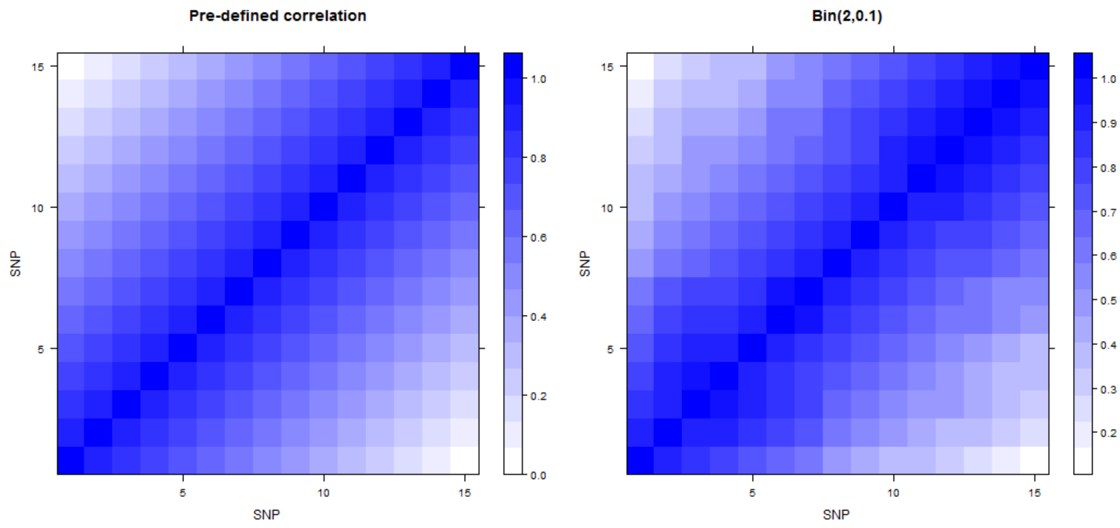


Figure A2.3: Heatmap of predefined correlation matrix G (left), empirical correlation between generated genotypes (right) over 1000 independent samples, MAF=0.1 for all loci.

In Figure A2.3 a comparison between the wanted correlation structure of data (left panel) and the realized correlation in simulated data. Simulation approach seems to be reliable and creates a data set that is congruent to the pre-defined correlation structure.

2.2. Simulation of SNP-data: truncated normal distribution method (TN)

The main idea of this approach is to generate a vector of correlated random variables that follow a multivariate normal distribution in the first step and to transform those continuous variables to discrete Bernoulli distributed variables by using quantiles of the normal distribution in the second step.

Independent normal vectors $\mathbf{X}_i \sim N_p(\mathbf{0},\mathbf{G})$ for $i=1,\ldots,n$ were generated by using *mvtnorm*-R-package (Genz et al., 2014).

The correlation structure between the entries in each vector is predefined by a matrix $\mathbf{G}$. In Figure A2.4 the wanted correlation structure is presented in left panel and the realized correlation structure of a sample of $n=1000$ independent normal vectors $\mathrm{Cor}(\mathbf{X})$ in the right panel. Obviously, data created using the mvtnorm package follows predefined correlation structure.

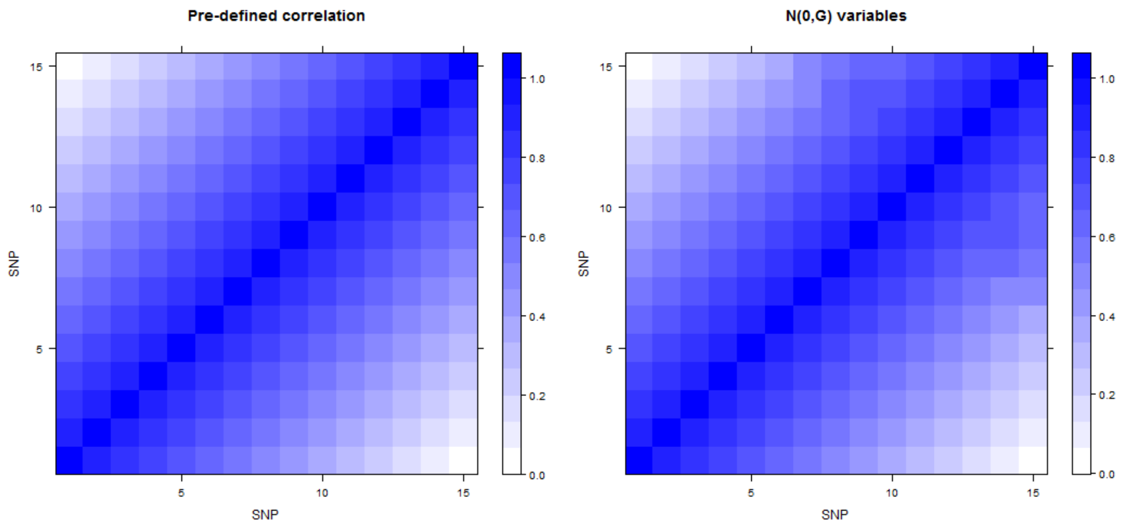


Figure A2.4: **Heatmap of predefined correlation matrix G (left), empirical correlation between the SNPs in a normal distributed random vector (right) over 1000 independent samples.**

Desired haplotypes (0/1 variables) are obtained from the normally distributed random vectors by applying a threshold, which corresponds to MAF p_j at each locus: $\mathbf{z} = (z_1, \ldots, z_p)'$ is the vector of quantiles of normal distribution $N_p(\mathbf{0}, \mathbf{G})$, so that $P(X_{ij} \le z_j) = p_j$. We used the same MAF at each locus $(p_j = p_{j'} \;\; \forall j, j' = 1, \ldots, p)$, but it is possible to generate loci with different MAFs. A haplotype could be viewed as a Bernoulli distributed variable $Y_{ij}^{TN} \sim Ber(p_j)$ with success probability equal to MAF (observation of a minor allele is defined as a success). Haplotypes variables are defined as $Y_{ij}^{TN} = 1$ if $X_{ij} \le z_j$, otherwise $Y_{ij}^{TN} = 0$. The genotypes are obtained as a sum of two independent samples of $\mathbf{Y}_i^{TN}$ – corresponding to two copies of a chromosome.

In the Figure A2.5 the empirical correlation matrix of a sample of generated genotypes $\mathbf{Y}^{TN} \in R^{n \times p}$ (right panel) is compared with the desired correlation matrix **G**, which is represented in the left panel. It can be seen, that predefined correlation structure is not fully captured by random variables $\mathbf{Y}_i^{TN}$. The reason for this is the loss of information due to transforming a continuous variable X_{ij} to a discrete variable Y_{ij}^{TN}.

In this approach the discrete variable is created by considering a threshold, which indicate the values of the 0/1 variable. A further possibility to truncate the normally distributed variables is to define the top $p_j/2$ and the lower $p_j/2$ as success and the rest in-between these two thresholds as 0. This two-sided version of the truncated normal approach has the same loss of correlation in the generated data set.

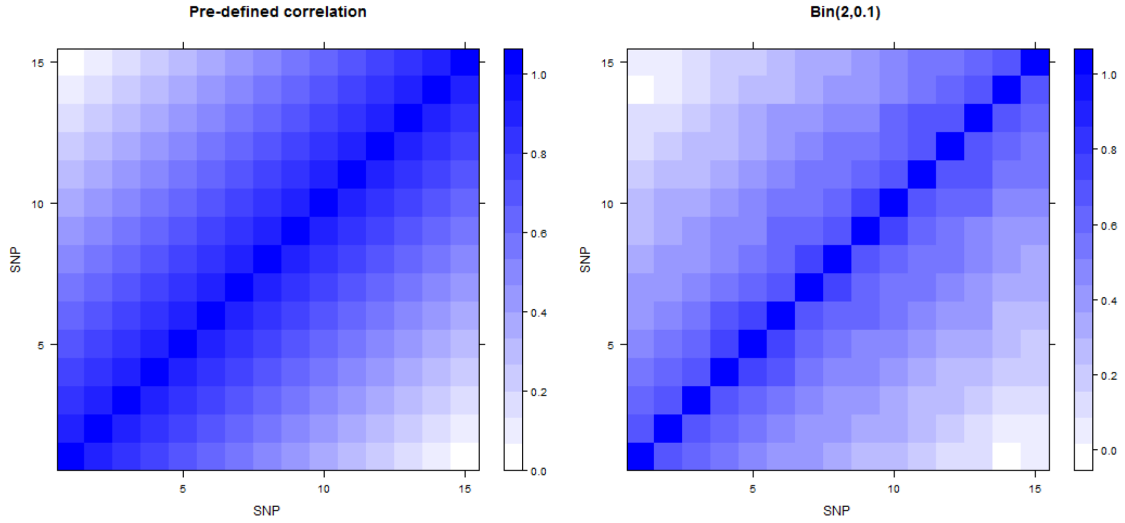


Figure A2.5. **Heatmap of predefined correlation matrix G (left), empirical correlation matrix of generated genotypes $\mathbf{Y}_i^{TN}$ (right) over 1000 independent samples.** Minor allele frequencies of all SNP are equal to 0.1.

2.3 Simulation of SNP-data: Cholesky decomposition based method (Chol)

To create binomial variables with a predefined correlation **G**, a vector $\mathbf{X}_i$ of independent identically distributed (iid) binomial variables $X_{ij} \sim Bin(p_j, 2),\ \ j = 1, \dots, p$ was created in the first step. In Figure A2.6 the empirical correlation matrix of these iid binomial variables $\mathrm{Cor}(\mathbf{X})$ is shown (upper panel, right). As expected, the correlations between the variables are very close to zero. In the second step $\mathbf{X}_i$ were transformed by using the Cholesky decomposition of correlation matrix $\mathbf{G} = \mathbf{QQ'}$ to $\mathbf{Y}_i^{Chol} = \mathbf{Q'}{\cdot}\mathbf{X}_i$. The empirical correlation of transformed vectors $\mathbf{Y}^{Chol} \in R^{n \times p}$ is represented in Figure A2.6 (lower panel, right). Transformed variables seem to capture the desired correlation structure; through the transformation process, the initially natural number variables (or integers) X_{ij} changed to floating point (or real) numbers. For our purpose, the simulated data should contain numbers of observed minor alleles at each locus, thus if the variables turned to be continuous, they should be rounded to 0, 1 and 2 in the last step. After the discretization process the empirical correlation of $\mathbf{Y}^{Chol}$ shows losses in the amount of captured predefined correlation and is presented in Figure A2.6 (lower panel, right).

For the same reason as in the truncated normal approach, this simulation method cannot capture the predefined correlation structure of genotypes: the predefined association between the variables is stronger than measured empirical correlation. Losses in association actually incurred are caused by the loss of information due to the transformation of a continuous variable to a discrete variable.

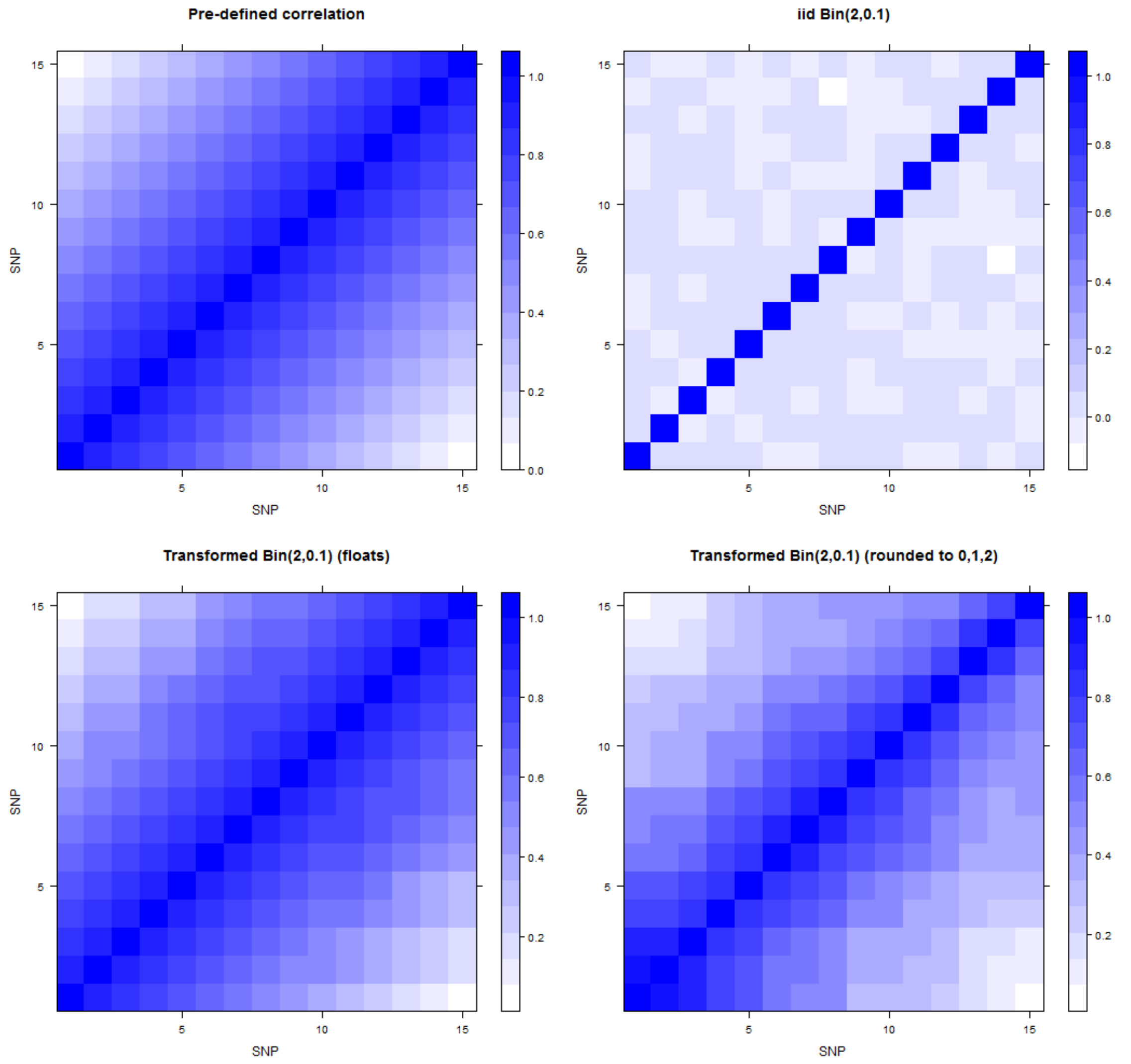


Figure A2.6. Heatmap of predefined correlation matrix G (upper panel, left), empirical correlation matrix (over 1000 independent samples) of iid binomially distributed variables $\mathbf{X}$ (upper panel, right), transformed continuous variables $\mathbf{Y}^{Chol}$ (lower panel, left) and those rounded to integers (lower panel, right). Minor allele frequencies of all SNPs are equal to 0.1.

2.4 Simulation of SNP-data: Normal-Copula based method (NC)

Another possible method to construct correlated genotypes along a given correlation structure is the Gaussian Copula, which creates the joint distribution of the correlation structure if the marginal distributions are known. A copula C could be thought as a function that joins multivariate distribution $C(F_1,\ldots,F_p)=F(Y_1,\ldots,Y_p)$ to their marginal distributions $F_j=F(Y_j)$, $j=1,\ldots,p$. In case the random variables describe genotypes, independent binomial distributions $F_1=\cdots=F_p=Bin(p_1,2)$ with equal success probabilities $p_1=p_2=\cdots=p_p$ are considered. However, it is also possible to choose different marginal distributions $F_j=Bin(p_j,2)$ if required. For the first step, an R-package *copula* (Hofert et al., 2014) was used to obtain the margins with desired correlation structure. For the second step, the genotypes $\mathbf{Y}^{NC}$ are sampled from the joint distribution. In Figure A2.7, the desired correlation structure is shown on the left panel, while the realized amount of correlation in the generated data is shown on the right. Obviously there are very large losses in the correlation. This method performed the least well in capturing the pre-defined correlation structure compared to the other methods considered.

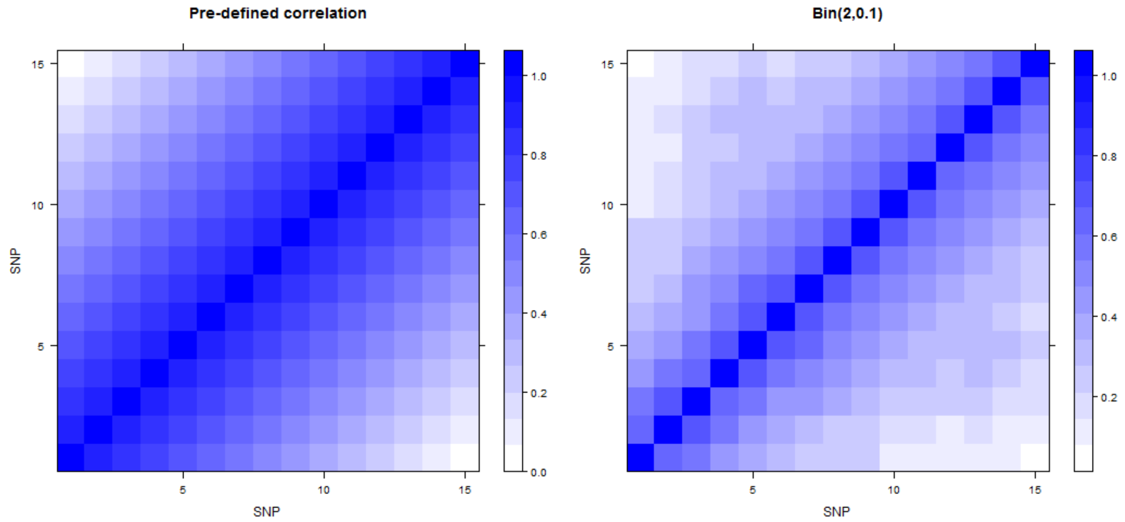


Figure A2.7: **Heatmap of predefined correlation matrix G (left) and correlation in sampled variables $\mathbf{Y}^{NC}$ (right)**

A3: Calculation of sample correlation coefficients and corresponding confidence intervals

Marker effects $\boldsymbol{\beta}$, β_{1j} and $\mathbf{u}$ were estimated in SMR and MMR and predicted in LMM repeatedly for $n_{sim} = 2500$ random sampled genotype data sets with sample size $n = 500$. For estimates $\hat{\boldsymbol{\beta}}$, $\hat{\beta}_{1j}$ and for predictive error $\hat{\mathbf{u}} - \mathbf{u}$ empirical correlation coefficients as well as the corresponding 95% confidence intervals were calculated.

For the k^{th} repetition, the estimates from SMR and MMR $\hat{\boldsymbol{\beta}}_k = \left(\hat{\beta}_{0k}, \hat{\beta}_{1k}, \ldots, \hat{\beta}_{pk}\right)'$. The empirical coefficient between the estimates at loci *j* and *j'* was calculated according to following formula:

$$\mathrm{Cor}(\hat{\beta}_{\mathrm{j}}, \hat{\beta}_{\mathrm{j'}}) = \frac{\sum_{k=1}^{2500}\left(\hat{\beta}_{\mathrm{jk}} - \frac{1}{2500}\sum_k \hat{\beta}_{\mathrm{jk}}\right)\left(\hat{\beta}_{\mathrm{j'k}} - \frac{1}{2500}\sum_k \hat{\beta}_{\mathrm{j'k}}\right)}{\sqrt{\sum_{k=1}^{2500}\left(\hat{\beta}_{\mathrm{jk}} - \frac{1}{2500}\sum_k \hat{\beta}_{\mathrm{jk}}\right)^2 \sum_{k=1}^{2500}\left(\hat{\beta}_{\mathrm{j'k}} - \frac{1}{2500}\sum_k \hat{\beta}_{\mathrm{j'k}}\right)^2}}$$

The empirical correlation coefficients for the predictive error $\hat{\mathbf{u}} - \mathbf{u}$ were obtained analogously. We define $\hat{\mathbf{u}}_k - \mathbf{u}_k := \mathbf{d}_k = (d_{1k}, \ldots, d_{pk})'$ as the deviation of predictions from LMM from true marker effects in the k^{th} repetition and the empirical correlation between the predictive errors at loci *j* and *j'* is obtained thusly:

$$\mathrm{Cor}(\hat{\beta}_{\mathrm{j}}, \hat{\beta}_{\mathrm{j'}}) = \frac{\sum_{k=1}^{2500}\left(\hat{d}_{\mathrm{jk}} - \bar{d}_{\mathrm{j}\bullet}\right)\left(\hat{d}_{\mathrm{j'k}} - \bar{d}_{\mathrm{j'}\bullet}\right)}{\sqrt{\sum_{k=1}^{2500}\left(\hat{d}_{\mathrm{jk}} - \bar{d}_{\mathrm{j}\bullet}\right)^2 \sum_{k=1}^{2500}\left(\hat{d}_{\mathrm{j'k}} - \bar{d}_{\mathrm{j'}\bullet}\right)^2}},$$

where $\bar{d}_{\mathrm{j}\bullet} = \frac{1}{2500}\sum_{k=1}^{2500} d_{\mathrm{j}k}$ stay for the average over the deviations at *j*th locus.

A confidence interval for sample correlation coefficient $\hat{r}$ (e.g., $\hat{r} = \hat{r}_{jj'} = \mathrm{Cor}(\hat{\beta}_{\mathrm{j}}, \hat{\beta}_{\mathrm{j'}})$) was obtained by using the Fisher transformation $\hat{\xi} := \Phi(\hat{r}) = \frac{1}{2}\ln\left(\frac{1+\hat{r}}{1-\hat{r}}\right)$ (Fisher, 1915; Hawkins, 1989). For increasing sample size $n \to \infty$, $\hat{\xi}$ tends to very quickly converge to a normal distribution $N\left(0, \frac{1}{n-3}\right)$. A two-sided confidence interval $\left[\hat{\xi}_{low}\ ,\ \hat{\xi}_{up}\right]$ for $\hat{\xi}$ is obtained by applying the upper 2.5% quantile $z_{0.975}$ of standard normal distribution to calculate a lower limit $\hat{\xi}_{low} = \hat{\xi} - z_{0.975} \cdot {}^{1}\!/\!{\sqrt{n-3}}$ and an upper limit $\hat{\xi}_{up} = \hat{\xi} + z_{0.975} \cdot {}^{1}\!/\!{\sqrt{n-3}}$.

Finally, the calculated upper and lower limits are transformed back to derive the confidence limits for sample correlation coefficient $\hat{r}$:

$$\hat{r}_{low} = \Phi^{-1}\left(\hat{\xi}_{low}\right) = \frac{e^{2\hat{\xi}_{low}} - 1}{e^{2\hat{\xi}_{low}} + 1} \text{ and } \hat{r}_{up} = \Phi^{-1}\left(\hat{\xi}_{up}\right) = \frac{e^{2\hat{\xi}_{up}} - 1}{e^{2\hat{\xi}_{up}} + 1}.$$

References

Beuzen, N.D., Stear, M.J., and Chang, K.C. (2000). Molecular markers and their use in animal breeding. Vet. J. *160*, 42–52.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychol. Bull. *70*, 426.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 507–521.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Hothorn, T., and Hothorn, M.T. (2014). mvtnorm: Multivariate Normal and t Distributions.

Graham, M.H. (2003). Confronting multicollinearity in ecological multiple regression. Ecology *84*, 2809–2815.

Grapes, L., Dekkers, J.C.M., Rothschild, M.F., and Fernando, R.L. (2004). Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. Genetics *166*, 1561–1570.

Gunst, R.F., and Webster, J.T. (1975). Regression analysis and problems of multicollinearity. Commun. Stat.-Theory Methods *4*, 277–292.

Hawkins, D.L. (1989). Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. Am. Stat. *43*, 235–237.

Henderson, C.R. (1953). Estimation of variance and covariance components. Biometrics *9*, 226–252.

Henderson, C.R. (1984). Applications of linear models in animal breeding (University of Guelph, Guelph, ON, Canada).

Hoerl, A.E., and Kennard, R.W. (1976). Ridge regression iterative estimation of the biasing parameter. Commun. Stat.-Theory Methods *5*, 77–88.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2014). copula: Multivariate Dependence with Copulas. *R package version 0.999-10*

Ishwaran, H., and Rao, J.S. (2014). Geometry and properties of generalized ridge regression in high dimensions. Perspect. Big Data Anal. Methodol. Appl. *622*, 81.

Kearsey, M.J., and Farquhar, A.G.L. (1998). QTL analysis in plants; where are we now? Heredity *80*, 137–142.

Kockläuner, G. (1984). Multicollinearity and Biased Estimation: Proceedings of a Conference at the University of Hagen, September 8-10, 1980 (Vandenhoeck & Ruprecht).

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

Ofir, C., and Khuri, A. (1986). Multicollinearity in marketing models: diagnostics and remedial measures. Int. J. Res. Mark. *3*, 181–205.

Piepho, H.P., Möhring, J., Melchinger, A.E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. Euphytica *161*, 209–228.

R Core Team (2014). R: a language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing.

Robertson, A. (1967). The nature of quantitative genetic variation. Herit. Mendel 265–280.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.

Tu, Y.K., Kellett, M., Clerehugh, V., and Gilthorpe, M.S. (2005). Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. Br. Dent. J. *199*, 457–461.

Wang, S.-G., Tse, S.-K., and Chow, S.-C. (1990). On the measures of multicollinearity in least squares regression. Stat. Probab. Lett. *9*, 347–355.

Wheeler, D., and Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J. Geogr. Syst. *7*, 161–187.

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

3RD CHAPTER

Effectiveness of Shrinkage and Variable Selection Methods for the Prediction of Complex Human Traits Using Data from Distantly Related Individuals

SWETLANA BERGER[1*], PAULINO PÉREZ-RODRÍGUEZ[2], YOGASUDHA VETURI[3], HENNER SIMIANER[1], GUSTAVO DE LOS CAMPOS[3]

1. Animal Breeding and Genetics Group, Department of Animal Sciences,
   Georg-August-University Goettingen,
   Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany
2. Colegio de Postgraduados,
   Carretera México-Texcoco Km. 36.5,
   Montecillo , Texcoco 56230, Estado de México, México
3. Department of Biostatistics, University of Alabama at Birmingham,
   RPHB 317C, Ryals School of Public Health,
   1665 University Boulevard, Birmingham, AL 35205, US

Published in Annals of Human Genetics, 79.2 (2015): 122-135: doi: 10.1111/ahg.12099

SUMMARY. Genome-Wide Association Studies have detected large numbers of variants associated with complex human traits and diseases. However, the proportion of variance explained by GWAS-significant SNPs has been usually small. This brought interest in the use of Whole-Genome Regression (WGR) methods. However, there has been limited research on the factors that affect prediction accuracy (PA) of WGRs when applied to human data of distantly related individuals. Here, we examine, using real human genotypes and simulated phenotypes, how trait complexity, marker-QTL LD and the model used affect the performance of WGRs. Our results indicated that the estimated rate of missing heritability is dependent on the extent of marker-QTL LD. However, this parameter was not greatly affected by trait complexity. Regarding PA our results indicated that: (a) under perfect marker-QTL LD WGR can achieve moderately high prediction accuracy, and with simple genetic architectures variable selection methods outperform shrinkage procedures. (b) Under imperfect marker-QTL LD, variable selection methods can achieved reasonably good PA with simple or moderately complex genetic architectures; however the PA of these methods deteriorated as trait complexity increases and with highly complex traits variable selection and shrinkage methods both performed poorly. This was confirmed with an analysis of human height.

Introduction

The availability of genomic data has revolutionized the statistical analysis of human diseases and traits. The development of methods that can accurately predict the genetic risk associated with these diseases and complex human traits can have a great impact on public health (e.g. Guttmacher et al., 2002; Simon-Sanchez et al., 2009). Modern genotyping and sequencing technologies can deliver massive amounts of information about the human genome, which are necessary for the prediction of genetic risk. However, the incorporation of genomic data into prediction remains challenging.

In recent years, a large number of genome-wide association studies (GWAS) have been conducted (e.g. http://www.genome.gov/gwastudies/). These studies have identified unprecedented numbers of variants associated with important complex traits and diseases. In some cases the variants identified so far explain a sizable proportion of the variance of the trait or disease. Examples of these include Crohn's disease, age-related macular degeneration and Type I diabetes (Manolio et al., 2008). However, for the great majority of traits and diseases, the variance accounted for by GWAS hits is small, regardless of whether they are moderately or highly heritable (Allen et al., 2010). Consequently, the use of genomic information for prediction of risk for diseases with complex genetic architectures remains limited. This problem, the so-called "missing heritability" of complex traits, has been discussed extensively by multiple authors (e.g. Maher, 2008; Manolio et al., 2009; Eichler et al., 2010).

Although several factors contribute to the "missing heritability" problem, a major explanation resides in the lack of power of standard GWAS to detect small-effect variants. Recent studies have shown that prediction accuracy can be improved by including in risk scores information of allele content at variants that show suggestive, albeit not statistically significant, association with the trait or disease being studied (Allen et al., 2010). However, most risk score methods are still based on a limited number of loci and alleles at different loci that are either equally weighted or weighted using statistics derived from single-marker-based association tests. Several authors (Yang et al., 2010) have suggested that a potentially better approach may consist of regressing phenotypes on whole-genome markers simultaneously using a Whole-Genome Regression (WGR) approach like the one originally proposed by Meuwissen et al. (2001).

Whole-Genome Regression has been used with human data for estimation of the proportion of variance that can be explained by regression of phenotype on markers (Yang et al., 2010; Speed et al., 2012) and for the assessment of prediction accuracy (Makowsky et al., 2011; de los Campos et al., 2013a). Using a GBLUP (Genomic Best Linear Unbiased Predictor) model and data from distantly related individuals, Yang et al. (2010) showed that simultaneous regression on a large set of ~300,000 common Single Nucleotide

Polymorphisms (SNPs) could explain roughly 50% of the heritability of human height. This encouraging result suggested that a large fraction of the missing heritability could be recovered by using regression methods based on large panels of whole-genome markers.

Accuracy of prediction of yet-to-be observed phenotypic or disease outcomes is arguably one of the most important features of a model when it comes to potential use of the method for precision medicine. It is well established that prediction accuracy of WGR methods is highly affected by genetic relationships (e.g. Makowsky et al., 2011) and it is not clear whether WGR methods that have been proved accurate for prediction of complex traits with family data (VanRaden et al., 2009; Crossa et al., 2010; Makowsky et al., 2011) will also be effective when applied to distantly related individuals, which are often of interest in human genetic applications.

According to Goddard (Goddard and Hayes, 2009), when WGR is applied to distantly related individuals, the prediction accuracy depends on two main factors: 1) the proportion of variance that can be explained by regression on the marker set (this depends largely on the extent of linkage disequilibrium (LD) between alleles at the markers and those at causal loci and, according to Yang et al. (2010) could be estimated using variance components), and 2) the accuracy of estimates of marker effects. These are two opposing forces: as we add more markers in the prediction equation the proportion of variance explained by markers potentially increases; however, more marker effects need to be estimated and the individual accuracy of estimates of effects will typically decrease. Therefore, in finite samples is not exactly clear that methods that have a higher proportion of variance explained in the training data will also be best for prediction of yet-to-be-observed outcomes. For example, in a recent study on prediction of human height using GBLUP, de los Campos et al. (2013a) showed that, with distantly related individuals, prediction accuracy increased as markers were added to the model up to a saturation point beyond which it decreased. This result suggests that the analysis and prediction of complex traits may benefit from the use of models that combine variable selection and shrinkage within a single framework.

In the last two decades, important developments in the area of penalized and Bayesian estimation procedures have led to a number of methods for implementing *large-p-small-n* regressions, including various methods that combine shrinkage estimation and variable selection. An overview of different penalized methods can be found in Hastie et al. (2005) and an overview of Bayesian methods for variable selection and shrinkage estimation (with a focus on genetic applications) is given by Gianola (2013) and de los Campos et al. (2013b). In animal and plant breeding, use of these methods has led to a substantial improvement in prediction accuracy (Habier et al., 2011; Heslot et al., 2012). Several studies have compared shrinkage and variable selection methods from a predictive perspective in animal and plant breeding applications (e.g. Habier et al., 2007; Calus et al., 2008;Verbyla et al., 2009; Daetwyler et al., 2010; Gao et al., 2013; Wimmer et al., 2013). Simulation studies have suggested superiority of variable selection methods over shrinkage estimation procedures. However, real data have not always confirmed that (de los Campos et al., 2013b) and in empirical analyses the predictive performance of different regression methods has been very similar, perhaps reflecting the fact that the architecture of most traits is more complex than often assumed in simulation studies. Most of the studies in plant and animal breeding are based on family data. The few studies (e.g. Habier et al. (2007), Gao et al. (2013) in breeding populations and Makowsky et al. (2011) or de los Campos et al. (2013a) with human data) that have assessed prediction accuracy with distant relatives have found that the prediction accuracy of WGRs models deteriorates quickly as the genetic distance between training and testing populations increases. In principle, variable selection methods are better suited to detect variants that are in strong LD with QTL, and this should make these methods more robust with respect to the effects of genetic distance on prediction accuracy (e.g. Habier et al., 2007).

However, the performance of these methods for prediction with human data so far has not been studied in detail. Indeed, in applications involving human data, most of the studies (Yang et al., 2010; Makowsky et al., 2011; de los Campos et al., 2013a) have used ridge-regression type estimators that do not involve variable selection or differential shrinkage of estimated effects. Zhou et al. (2013) used WGR models that combine variable selection and shrinkage using data from distantly related individuals; unfortunately the study did not evaluate the prediction accuracy. Importantly, the factors that affect prediction accuracy in the analysis of family data can be different than those that affect prediction accuracy when training and validation samples are distantly related. Indeed, with family data, co-segregation of alleles at markers and at quantitative trait loci (QTL) plays a major role, and can induce linkage between markers and QTL at distant positions. Under these conditions, variable selection is difficult to perform and may not be needed because signals generated by QTL can be tracked by markers that are far apart from a QTL. This type of linkage is not present when training and validation samples are distantly related, and we lack research about the relative effectiveness of shrinkage and variable selection methods with data from distantly related individuals.

Therefore, the main goal of this study was to assess the predictive performance of different types of WGR methods, including both shrinkage estimation procedures and methods that perform variable selection, when used for prediction of complex traits and with distantly related individuals. We considered three statistical methods that differ in the prior distribution of marker effects and consequently yield different types of estimates. Firstly, a model with

Gaussian distribution of marker effects (the GBLUP) was used; this ridge-regression-type method induces homogeneous shrinkage of marker effects. Secondly, a *scaled-t* prior for marker effects (labeled as Bayes A by Meuwissen et al. (2001)) was used; a method that induces an effect-size dependent shrinkage of estimates (Gianola, 2013). Finally, a Spike-Slab model (e.g. George and McCulloch, 1993; Ishwaran and Rao, 2005) was used, which combines variable selection and shrinkage. Recent methodological developments introduced by Zhou et al. (2013) allow implementation of a Spike-Slab model even with a very large numbers of markers.

The performance of these methods was assessed with simulated and real data. Our simulation comprised different scenarios pertaining to the complexity of the trait (in terms of number of large-effect loci) and the pattern of linkage disequilibrium between markers and causal or quantitative trait loci. The results obtained from simulation studies were validated by analysis of human height measured on distantly related individuals.

Materials and Methods

In the classical quantitative genetic model, a continuous trait y_i is described as a sum of three components: the population mean (μ), a random component reflecting the genetic factors, the so-called genetic value u_i, and a random model residual (ε_i) usually assumed to be identically and independently normal distributed with zero mean and variance σ_ε^2.

In genomic models, the genomic values u_i are approximated using regressions on marker genotypes. For instance, in an additive model one can set $u_i = \sum_{j=1}^{p} X_{ij}\beta_j$, where $X_{ij} \in \{0,1,2\}$ represents the allele dosage at the *j*th locus of the *i*th individual and β_j represents the corresponding marker effect. Thus, the model for p markers can be expressed as:

$$y_i = \mu + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i,\ \ i = 1,...,n \qquad (1)$$

In WGR methods the number of effects to be estimated can vastly exceed the number of data points (i.e., *p*>>*n*). Thus, the estimation of effects in the model described above requires the use of some type of regularized regression procedure such as penalized or Bayesian regression. In Bayesian regressions, the type and extent of shrinkage of estimates of effects is controlled by the choice of prior for marker effects.

To cover a wide range of methods, in this study we considered two extreme approaches (GBLUP a shrinkage estimation procedure and the Spike-Slab, a method that combines variable selection and shrinkage) and an intermediate one (Bayes A) that induce differential shrinkage of estimates of effects.

The **GBLUP** model is obtained by assigning independent identically distributed (IID) normal priors to the marker effects, that is: $\beta_j \sim N(0,\ \sigma_\beta^2),\quad j=1,...p$. This approach yields estimates equivalent to those from ridge regression, where all effects are shrunk towards zero to a similar extent. Using the expectation of *i*th phenotype y_i (given the genotypes and marker effects), and the genomic value $u_i = \sum_{j=1}^{p} X_{ij}\beta_j$, we rewrite equation (1) as $y_i = u_i + \varepsilon_i,\ i=1,...,n$. Thus the genomic value is also normal: $\mathbf{u} \sim N(\mathbf{0},\ \sigma_u^2\mathbf{G})$ with a genomic relationship matrix, which is obtained as a cross product of genotype readings $\mathbf{G} = \{G_{ik}\} = \frac{1}{\sum_j 2p_j(1-p_j)}\mathbf{XX}'$ (p_j is the minor allele frequency (MAF) at the *j*th locus) and a genomic variance component $\sigma_u^2 = \sum_{j=1}^{p} 2p_j(1-p_j)\sigma_\beta^2$. Therefore, the GBLUP could be implemented in Bayesian settings as a random effect model with a variance-covariance structure represented by $\sigma_u^2\mathbf{G} + \sigma_\varepsilon^2\mathbf{I}$, assuming for example a scaled inverse χ^2-density as a prior distribution for variance components σ_u^2 and σ_ε^2.

Above we described the GBLUP model that one obtains by regressing phenotypes on markers using IID normal priors for marker effects. This model can be fitted by either regressing phenotypes on markers explicitly, or using an equivalent model based on a genomic relationship matrix $\mathbf{G} \propto \mathbf{XX}'$. Some authors (Speed et al., 2012) have proposed alternative ways of computing genomic relationships that account for LD; therefore, we also fitted the GBLUP model applying the method proposed by Speed et al. (2012) to compute **G** using the LDAK software (available at www.dougspeed.com); we refer to this method as to GBLUP-ldak.

In **Bayes A** markers are assumed to follow IID scaled-t densities (an example for t-scaled prior with 5 degrees of freedom is given in Figure S1). In practice it is convenient to represent this density as an infinite mixture of scaled-normal densities: $t(\beta_j|df,S) = \int N(\beta_j|0,\sigma_{\beta_j}^2)\chi^{-2}(\sigma_{\beta_j}^2|df,S)\partial\sigma_{\beta_j}^2$, where $N(\beta_j|0,\sigma_{\beta_j}^2)$ is a normal density with null mean and variance $\sigma_{\beta_j}^2$ and $\chi^{-2}(\sigma_{\beta_j}^2|df,S)$ is a scaled-inverse χ^2-density with degree of freedom *df* and scale parameter *S* (e.g. Gianola, 2013; Gianola et al., 2009).

In the **Spike-Slab** model, the prior assigned to marker effects is a mixture of two distributions: one (the spike) with small variance concentrated around zero that corresponds to small or no effects and the other (the slab) is a flat distribution with large variance that is linked to large marker effects. The spike can be represented by a continuous distribution centered at zero and with very small variance or by a point mass at zero. We concentrate on the prior introduced by George and McCulloch (1993), a mixture of two normal distributions. Conditional on the proportion of large effects, π, and on variance parameters, the distribution of marker effects is given by $p\left(\beta_j \middle| \pi, \sigma^2_{\beta_1}, \sigma^2_{\beta_2}\right) = \pi\ N(\beta_j|0,\ \sigma^2_{\beta_1}) + (1-\pi)\ N(\beta_j|0,\ \sigma^2_{\beta_2})$, where $\sigma^2_{\beta_1}$ reflects the variability in large effects and $\sigma^2_{\beta_2}$ is the variance component of small effects. An example for $\pi = 0.15$ is represented in Figure S3.1.

Recently, Zhou et al. (2013) proposed an efficient method to implement the Spike-Slab model. In their approach, called Bayesian Sparse Linear Mixed Model (BSLMM), they represent marker effects as the sum of two components: small effects $\alpha_j \sim N(\alpha_j|0,\sigma^2_\alpha)$, assigned to all markers and sparse effects $\gamma_j \sim \pi\ N(\gamma_j|0,\sigma^2_\gamma) + (1-\pi)\delta_0$ (a mixture of a normal and a point-mass-at-zero distribution), which are assigned to a proportion of markers π, so that the total effect of the j^{th} SNP $\beta_j = \alpha_j + \gamma_j$ is a mixture of normal distributions $\pi N(\beta_j|0,\sigma^2_\alpha + \sigma^2_\gamma) + (1-\pi)N(\beta_j|0,\sigma^2_\alpha)$. Zhou et al. (2013) specified this model using a reparameterization which greatly facilitates computations.

All simulations as well as subsequent statistical analyses of simulated and real data were implemented in R (R Core Team, 2014). In this study, the GBLUP and Bayes A methods were fitted using the Gibbs Sampler algorithm implemented in the R package, BGLR (Pérez and de los Campos, 2014). The Spike-Slab model was fitted using the BSLMM method, which is included in the GEMMA software package (http://stephenslab.uchicago.edu/software.html).

Simulation and Real Data Analysis

Data

The genotypes used for simulation and in the real data analysis came from by NIH-funded Gene-Environment Association Studies (GENEVA, http://www.genevastudy.org), which is a consortium of sixteen genome wide association studies. We used a subset of GENEVA consisting of data from the Nurses' Health Study and the Health Professionals' Follow-up Study studies. Samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 with about 780 K SNPs. The GENEVA data set contains phenotypic and genotypic records of n=5,961 individuals (3,391 women and 2,570 men) with average age of 57.2 years (SD=7.7 years) and average height 170.2 cm (SD=9.6 cm). For the real data analysis we used adult height (adjusted for age, sex and affiliation to case or control group) as the phenotype.

Quality control procedures

We removed all markers with proportion of missing genotypes per SNP ≥ 0.01 and all individuals with a proportion of missing genotypes per individual ≥ 0.05. Further, on the basis of the available pedigree information, we also removed all nominally related individuals and individuals with a Hispanic genomic background such that only individuals of Caucasian origin remained in the data set. We also set a lower threshold of 0.01 for MAF, so that after quality control of the genomic data sample size was 5,758 individuals and 673,197 SNPs loci remained.

Simulation

We aimed at investigating the performance of three models, which apply different types of shrinkage of effect estimates, under different genetic architectures and varying levels of LD between markers and QTL. The simulation was conducted using true genotypes (see details above) and simulated phenotypes.

Markers and QTL. SNPs were randomly divided into two subsets: 350K SNPs were designated as markers and the rest (~323K) were used as a pool for sampling subsets of QTL (5K, in each replicate). The 5K QTL were sampled from the pool of 323K loci either completely at random (RAND) or by oversampling among the loci with low minor allele frequency (LOW-MAF). In this case sampling probabilities were set to target 75% of the QTL with MAF < 0.05, 25% of the QTL with MAF between 0.05 and 0.15, no QTL had a MAF > 0.15. In the LOW-MAF scenario the distributions of allele frequencies at markers and at QTL were expected to be different, and this was expected to influence the extent of LD between markers and QTL. Therefore, for each replicate, we used PLINK (Purcell et al., 2007) to compute the pairwise squared correlation r^2 between genotypes at the QTL and those at the two flanking markers.

Genetic architecture. We assumed that only a subset of QTL had large effects, while the rest of them had small effects. We considered three different scenarios: in the first one all QTL effects were sampled from IID normal densities $N(\beta_j | 0, \sigma_\beta^2)$. In the second and third scenarios we randomly chose *p*=50 or *p*=250 SNPs, respectively, and sampled their effects from a normal density with a large (see next) variance, the rest of the QTL effects were sampled from a normal density with a smaller variance. We set the variance parameters of the two normal densities used to sample effects in scenarios 2 and 3 to target a heritability (h^2) of 0.5 and a partition of the genetic variance (hereinafter called *pve*) where large effect QTL explain either 25% or 75% of genetic variance in scenarios 2 and 3.

Simulation of phenotypes. The phenotypes were constructed according to an additive model $y_i = \sum_{j=1}^{5000} Z_{ij}\beta_j + \varepsilon_i$ for $i = 1, \ldots, n$, where model error ε_i and marker effects β_j follow normal distributions with zero mean and Z_{ij} are the genotype readings at causal loci. The variance of the residual term $V(\varepsilon_i) = 0.5$ was kept fixed across all scenarios, while the variance of marker effects $V(\beta_j)$ varied from scenario to scenario, depending on the number of large effect QTL, amount of genetic variance explained by these large effects QTL, and the distribution of MAFs in QTL.

Data Analyses.

We analyzed the simulated data using markers, QTL or markers and QTL. The first scenario involved imperfect LD between markers and QTL, the last two contained the causal variants in the panel and therefore were perfect LD scenarios.

Genomic Heritability. For the GBLUP, the estimated genomic heritability $h_G^2 = \dfrac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$ was defined as the ratio between the variance explained by genomic factors, σ_g^2, and the phenotypic variance, $\sigma_p^2 = \sigma_g^2 + \sigma_\varepsilon^2$; in the G-BLUP h_G^2 was estimated based on posterior samples collected using the BGLR-package.

For Bayes A the BGLR-package did not provide the estimates of genomic heritability directly. In this model, a scaled-inverse □□ distribution is assigned to the variance of the effects β_j. Therefore, we have $E(\sigma_\beta^2) = \dfrac{S_0}{df - 2}$; using this we can define the genomic variance as follows: $\sigma_g^2 = \sum_{j=1}^{p} 2p_j(1-p_j)\dfrac{S_0}{df-2}$,where p_j stands for allele frequency at locus *j*. With this, the genomic heritability can be defined as $h_G^2 = \dfrac{\sum_{j=1}^{p} 2p_j(1-p_j)\frac{S_0}{df-2}}{\sum_{j=1}^{p} 2p_j(1-p_j)\frac{S_0}{df-2} + \sigma_\varepsilon^2}$.

We also estimated this parameter using posterior samples collected using the BGLR-package.

GEMMA provided posterior samples of $PVE(\boldsymbol{\beta},\mathbf{u},\tau^{-1})=\dfrac{V(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})}{V(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})+\tau^{-1}}$ (Zhou et al., 2013) which describes total proportion of variance in phenotype explained by the sum of the 'sparse' ($\mathbf{X}\boldsymbol{\beta}$) and random effect ($\mathbf{u}$). Essentially this quantity meets definition of genomic heritability, we used posterior mean of PVE to obtain the estimate of genomic heritability. In addition to estimates of genomic heritability we report the R^2 between phenotypes and predictions in the training data set as a measure of goodness of fit. This was only done for the GBLUP and Bayes A because GEMMA does not provide predictions for the training data set.

Assessment of Prediction Accuracy.

To assess prediction accuracy, in both the simulated and real data, we replicated 30 times a **training-testing** (TRN-TST) **validation design** (Hastie et al., 2005). In each TRN-TST experiment, data were randomly split into two disjoint sets, 5,258 data points in the TRN and from the remaining 500 individuals, we retained for validation only the ones whose genomic pairwise relationships with individuals in the TRN group did not exceed $1/8$; these were typically ~400 individuals. In the analysis of real phenotype (adjusted human height) we used the same subset of SNPs that were used in the 'only marker' scenario in simulation studies and the same mapping of individuals to TRN/TST groups. We assessed prediction accuracy using the Pearson's product-moment correlation between the true and predicted phenotypes $cor(\mathbf{y},\hat{\mathbf{y}})$ in the validation set.

Results

Results from simulation studies

The empirical quantiles of the distribution of MAF at different sets of loci are given in Table 3.1. In the RAND scenario, the empirical distribution of the MAF at QTL and markers were very similar; this was expected because both sets of loci were sampled at random. However, as intended, the empirical distribution of MAF at QTL in the LOW-MAF scenario had, relative to the same distribution at the marker loci, an over representation of loci in the low MAF spectra.

Table 3.1. *Empirical percentiles of the distribution of minor allele frequency for markers and for QTL in simulated data in both sampling scenarios.*

Set (Scenario)	**Quantiles of the distribution of minor allele frequency**				
	5%	10%	25%	50%	95%
Markers	0.0298	0.0498	0.1115	0.2268	0.4713
QTL (RAND)	0.0302	0.0501	0.1117	0.2273	0.4713
QTL (LOW-MAF)	0.0133	0.0169	0.0279	0.0461	0.1383

The 5%, 10%, 25%, 50%, and 95% percentiles for marker data set and for QTL in both sampling scenarios, averaged over 30 replicates.

Linkage disequilibrium is allele-frequency dependent; therefore, based on results of Table 3.1 one would expect that the extent of Marker-QTL LD will vary between scenarios. Table 3.2 provides a summary of estimates of LD between QTL and the two flanking markers by scenario.

Table 3.2. *Summary statistics of pairwise LD measure in both sampling scenarios.*

Scenario	**Average r^2**	**Quantiles**				
		5%	25%	50%	75%	95%
RAND	0.624 (0.286)	0.223	0.344	0.609	0.941	0.996
LOW-MAF	0.206 (0.333)	0.001	0.007	0.029	0.203	0.982

Summary statistics of pairwise LD, measured as squared correlation r^2 between the QTL and markers, flanking markers on either side in the RAND- and LOW-MAF- scenarios; r^2 is averaged over 30 Monte-Carlo replicates, with standard deviation given in parentheses and 5%, 25%, 50%, 75% and 95% quantiles.

The average of r^2 over 30 Monte-Carlo (MC) replicates in the RAND-scenario was 0.624 with a standard deviation (SD) of 0.286. On the other hand, the average of pairwise r^2 in the LOW-MAF-scenario was three times smaller.

Estimated Genomic Heritability and Goodness of Fit

The average (over MC replicates) estimated genomic heritabilities obtained by simulation scenario (RAND in the upper panel, LOW-MAF in the lower panel), statistical method (Bayes A, Spike-Slab, GBLUP and GBLUP-ldak), information used (markers, markers+QTL and QTL) and genetic architecture are shown in Figure 3.1.

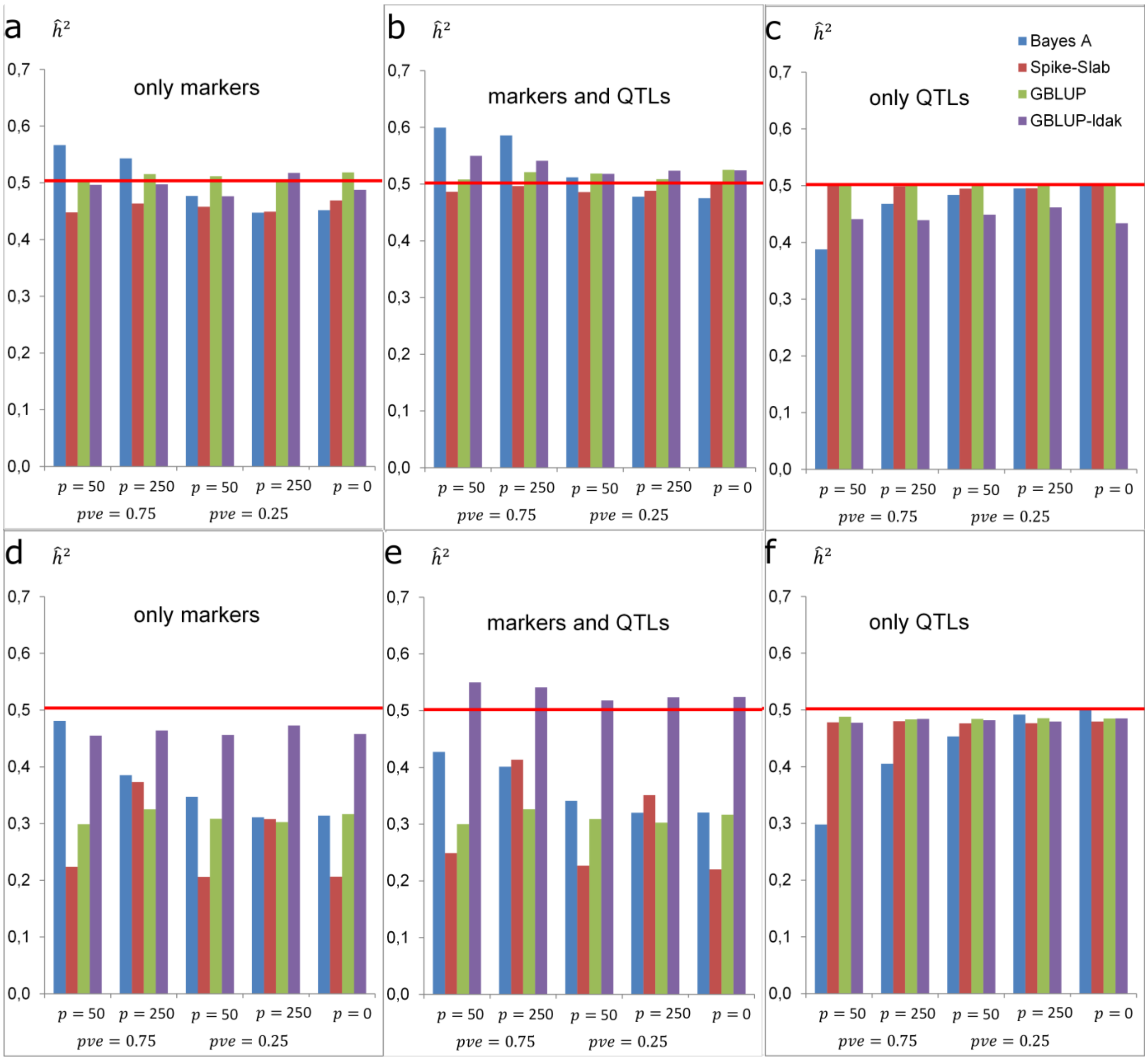


Figure 3.1. Estimates of Genomic Heritability. Averages of estimates of genomic heritability over Monte-Carlo (MC) replicates obtained by simulation scenario (RAND upper panel: **a**, **b**, **c**; LOW-MAF in lower panel: **d**, **e**, **f**), genetic architecture (*p*=number of large effect QTL, *pve*=proportion of genetic variance explained by large effect QTL), model (GBLUP, GBLUP-ldak, Bayes A, and Spike-Slab) and data used (only markers, markers and QTL or only QTL).

QTL-based analysis. When only QTL genotypes were used to fit models to data simulated with the RAND scenario (Figure 3.1, panel **c**) the GBLUP and Spike-Slab models gave an average estimate of genomic heritability that was very close to the simulated heritability, suggesting that these two methods have almost no bias with the sample size used in this study. GBLUP-ldak generally under-estimated heritability and Bayes A yielded downwardly biased estimates when the genetic architecture had a few markers explaining a sizeable proportion of genetic variance (e.g., pve=0.75 p=50 in Figure 3.1 panel **c**). In the LOW-MAF scenario (Figure 3.1, panel **f**), GBLUP, Spike-Slab and GBLUP-ldak showed almost un-biased estimates, but Bayes A continued to deliver downwardly biased estimates in scenarios where large-effect QTL explained a sizable fraction of genetic variance (e.g. pve=0.75 p=50 in Figure 3.1 panel **f**).

Marker-based analysis. It is important to note that, due to imperfect marker-QTL LD when only markers are used in the analysis, the true proportion of variance that can be explained by regression on markers, the so-called genomic heritability (de los Campos et al., 2014), can be lower than the trait heritability. Therefore, even in simulations, the population value of the genomic heritability is unknown and therefore we can compare results across models but we cannot assess bias. In the RAND scenario the estimates derived with the GBLUP models (see Figure 3.1 **a**) were very close to the simulated trait heritability. However, the estimates obtained with the Spike-Slab model suggested some extent (of the order of 10%) of missing heritability. Bayes A yielded estimates similar to those of the Spike-Slab with complex genetic architectures but tended to over-estimate the genomic heritability with simpler genetic architectures.

In the LOW-MAF scenario (See Figure 3.1 **d**) estimates of genomic heritability varied substantially between methods and genetic architectures: the GBLUP and Bayes A yielded a great extent of missing heritability. In comparison GBLUP-ldak yielded a much smaller extent of missing heritability and Spike-Slab estimated an extent of missing heritability that was small in scenarios in which large effect QTL contributed a sizeable proportion of variance and increased - to the point of getting very close to GBLUP- as trait complexity increased.

Finally, as one could expect, the analysis based on markers and QTL (panels **b** and **e** in Figure 3.1) yielded estimates that were intermediate between the QTL only and marker only cases in the RAND scenario and were very close to the analysis based on markers in the LOW-MAF scenario.

The R^2 between true and the predicted phenotypes in the training data sets, averaged over 30 MC replicates, is represented in Figure S3.2. We do not present results for GEMMA because this software does not provide predictions for the training data set. In the perfect LD scenario (only QTL genotypes used, Figure S3.2, panels **c** and **f**) the R^2 was between 60-70%, suggesting some over-fitting (the simulated heritability was 0.5). The evidence of over-fitting increased slightly when markers were used. The clearest sign of over-fitting was observed with Bayes A in the LOW-MAF scenario. In the analysis based on markers only (Figure S3.2, panels **a** and **d**) the three models behaved very differently: GBLUP showed the lowest R^2, and this statistic did not vary much between scenarios. On the other hand, GBLUP-ldak showed much higher R^2 than GBLUP and the value of this goodness of fit statistics for this model was also very stable across simulation scenarios. Finally, Bayes A showed a pattern with higher R^2 than GBLUP in scenarios involving large-effect QTL with sizeable contribution to additive variance. However, the R^2 in the training data set of Bayes A decreased as the genetic architecture of the simulated trait became more complex, to a point that the R^2 of Bayes A approached GBLUP when there were no large effect QTL.

Prediction accuracy

Figure 3.2 displays the correlation (average over 30 MC replicates) between phenotypes and predictions in testing data sets.

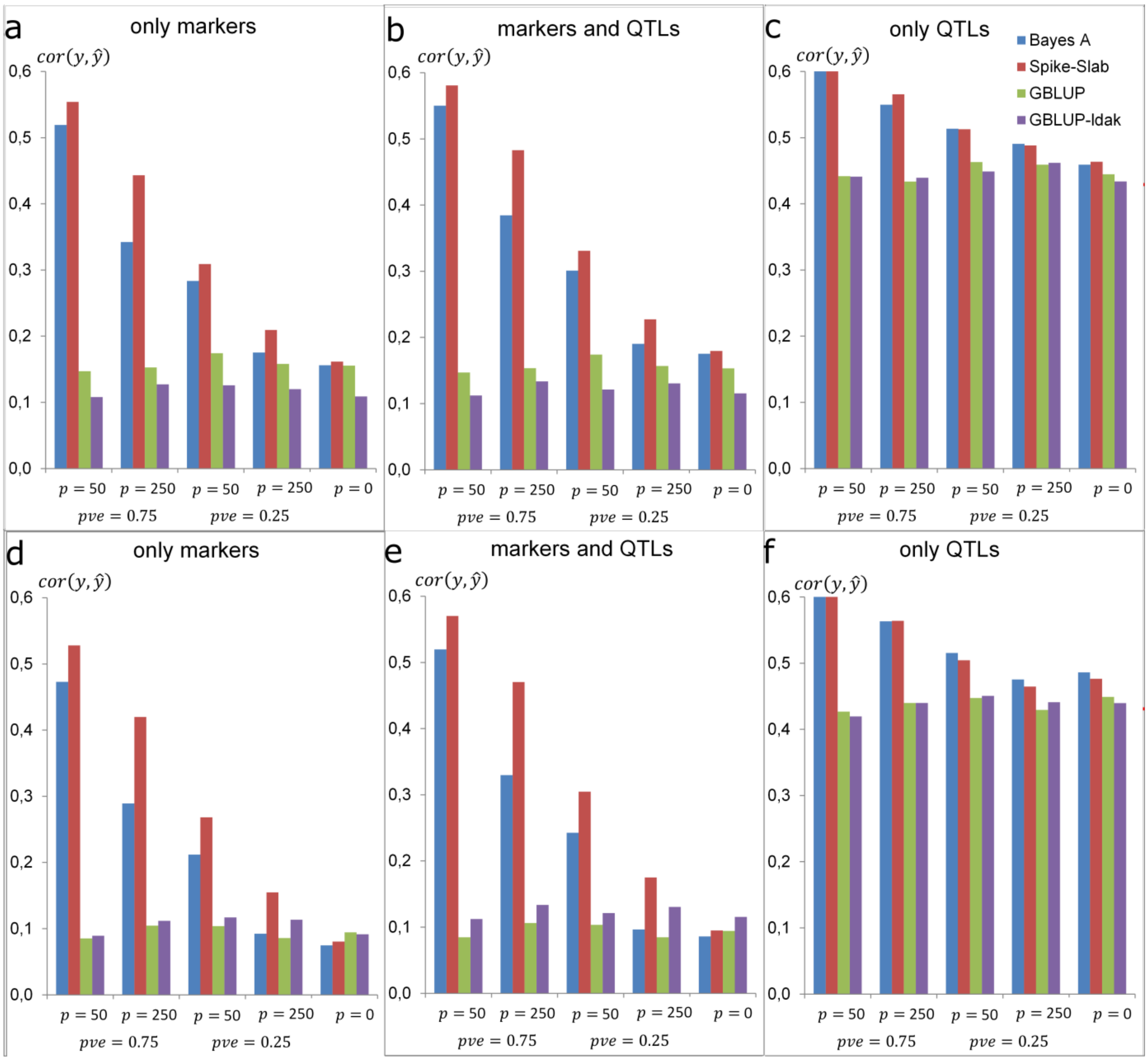


Figure 3.2. Correlation between phenotypes and genomic predictions in training data sets. Correlation (average over MC replicates) between phenotypes and genomic predictions in training data sets, by simulation scenario (RAND upper panel: **a**, **b**, **c**; LOW-MAF in lower panel: **d**, **e**, **f**), genetic architecture (*p*=number of large effect QTL, *pve*=proportion of genetic variance explained by large effect QTL) data used (only markers, markers and QTL or only QTL) and analysis method (GBLUP, GBLUP-Idak, Bayes A, and Spike-Slab).

Plots were sorted, by simulation scenario (RAND or LOW-MAF), genetic architecture (number of large effect-QTL and proportion of genetic variance explained by large effect QTL), data used (QTL, markers or markers+QTL) and analysis methods (Bayes A, Spike-Slab, GBLUP and GBLUP-ldak).

Impacts of LD. The comparison of the prediction accuracy achieved using only QTL (Figure 3.2, panels **c** and **f**) and those obtained using only markers (Figure 3.2, panels **a** and **d**) sheds light on the impacts of LD on prediction accuracy. As expected, the maximum prediction accuracy across methods and simulation scenarios was achieved when only QTL genotypes were used for model fitting and prediction (perfect LD scenario). When markers in imperfect LD with QTL were introduced, prediction accuracy was reduced markedly. The adverse effects of imperfect LD between markers and QTL were more marked in the **GBLUP** and **GBLUP-ldak** and less adverse for model Spike-Slab and Bayes A and in scenarios with simpler genetic architectures; however as the genetic architecture of the trait become more complex, the superiority of these two methods, relative to GBLUP diminished.

Statistical Method. Overall, GBLUP and GBLUP-ldak had the worst predictive performance; this was particularly clear when only markers or markers and QTL were used. Bayes A performed considerably better than the GBLUP and the Spike-Slab performed even better than Bayes A indicating clear benefits of methods inducing differential shrinkage of estimates relative to methods like the GBLUP that induce homogeneous shrinkage of estimates.

Genetic Architecture. The highest prediction accuracy was obtained in scenarios where a small number of QTL with large effects (*p*=50) explained a large proportion of the genetic variance (*pve*=75%). The superiority of the Spike-Slab or Bayes A over the GBLUP was maximum when the genetic architecture was simple; however the differences between the prediction accuracy of Bayes A and Spike-Slab, relative to GBLUP methods diminished as the trait architecture became more complex. Although, the prediction accuracy of the GBLUPs was not greatly affected by the genetic architecture of the trait, in analyses based on markers or markers and QTL, there was a small but systematic trend suggesting that GBLUP outperformed GBLUP-ldak in the RAND scenario and the opposite was true in the LOW-MAF scenarios.

For each MC replicate we computed differences in prediction accuracy, measured by differences in correlations $cor(\mathbf{y},\hat{\mathbf{y}})$, between different simulations or data analysis scenarios and studied the distribution of these differences (boxplots with pairwise differences in prediction accuracy (by method) are provided in Figure S3.3). In analyses including markers, (either markers only or markers+QTL), adding QTL to the set of loci used to compute the **G** matrix increased prediction accuracy when Bayes A or Spike-Slab were used, while the GBLUP methods did not benefit from having the QTL loci within the set of markers used to compute the **G** matrix. As expected, the prediction accuracy obtained in the RAND scenario was higher than the one obtained in the LOW-MAF scenario; this pattern was observed across statistical methods.

Figure 3.3 gives boxplots of the differences in prediction accuracy by pair of models, across simulation scenarios. The Spike-Slab models and Bayes A were significantly better than the GBLUP; the superiority of the Spike-Slab over Bayes A was also systematic, but very small in magnitude.

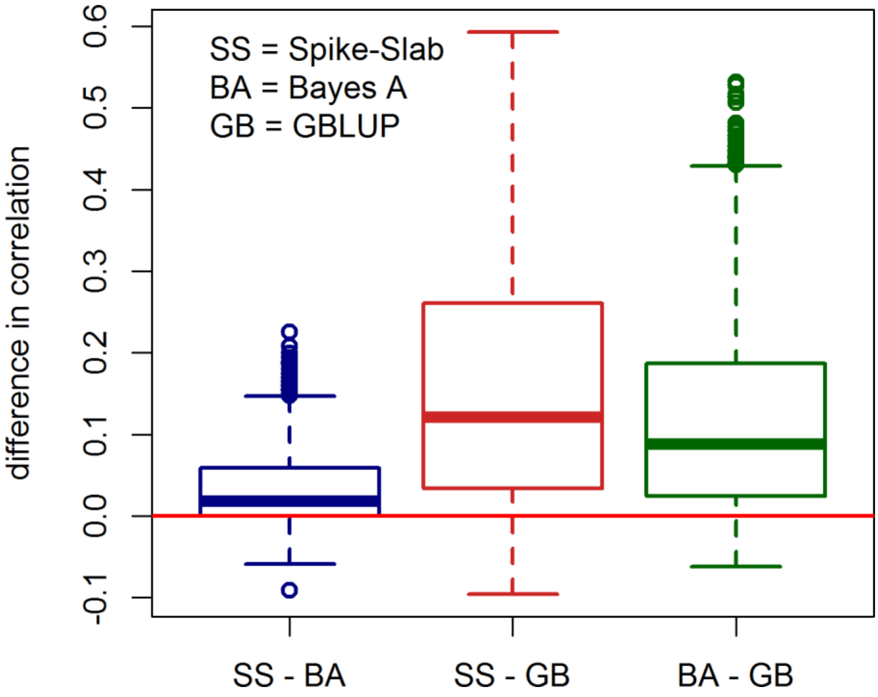


Figure 3.3. Pairwise difference in prediction accuracy across methods. Boxplots of the pairwise differences (across MC replicates and simulation scenarios) in prediction accuracy by pair of models.

Results from Real Data Analysis

The estimates of genomic heritability and of prediction accuracy in testing data sets, averaged over 30 training-testing partitions, are displayed in Table 3.3. The estimated genomic heritability ranged from 0.367 (Spike-Slab) to 0.561 (GBLUP-ldak). The GBLUP had an intermediate estimate of genomic heritability (0.435). Our estimates are in line with previous reports for human height using common SNPs (e.g. Yang et al., 2010; de los Campos et al., 2013a). These results are also in agreement with what we observed in the LOW-MAF setting, in scenarios for traits without major QTL and using only marker genotypes for computing **G** (see Figure 3.1 **d** for *pve*=0). The correlations between phenotypes and predictions were low (0.16-0.17) for all methods, and only slightly higher for the GBLUP methods. These correlations are in agreement with what we obtained in the simulation study in the LOW-MAF scenario when QTL were not used in the model (see Figure 3.2 **d**).

Figure 3.4 provides box-plots of the difference in prediction accuracy obtained, within each TRN-TST partition, between methods. Although the average difference in prediction accuracy between methods was small, the analysis of pair-wise differences in prediction accuracy (by using the Wilcoxon signed rank test) suggested a statistically significant, albeit small, superiority of the GBLUP methods over Bayes A; the differences between the Spike-Slab and GBLUP are non-significant.

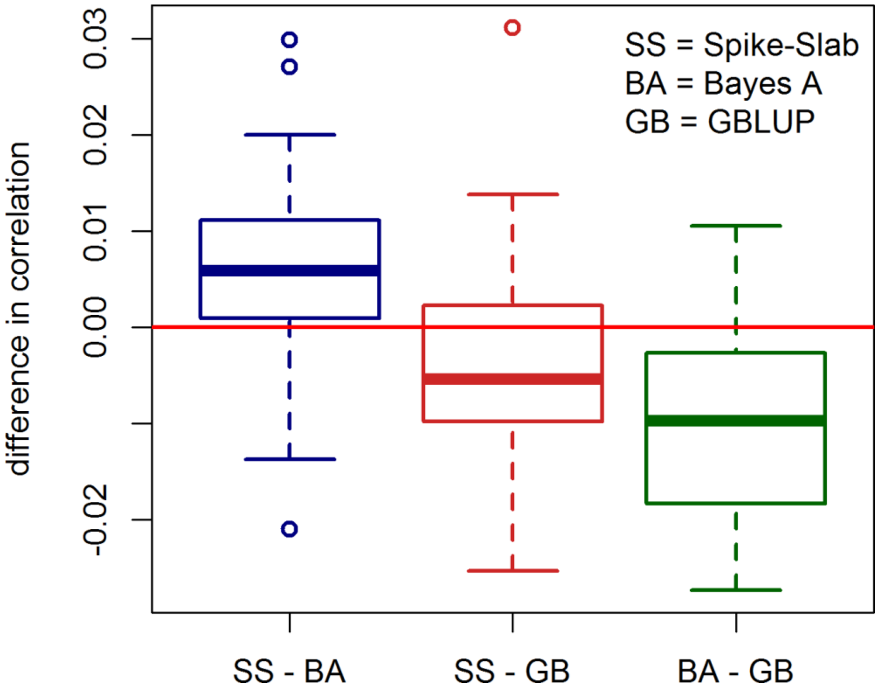


Figure 3.4. Pairwise difference in prediction accuracy across methods. Boxplots of the difference in prediction accuracy, within TRN-TST partition, between methods.

Discussion

In recent years, Genome Wide Association Studies have found an unprecedented number of variants associated with important human traits and diseases (http://gds.nih.gov/). However, for complex traits and diseases, the variants identified so far usually explain a small fraction of inter-individual differences in a trait or in disease risk, a problem referred to as the missing heritability of complex traits (Maher, 2008; Manolio et al., 2009; Eichler et al., 2010; Gibson, 2010; Makowsky et al., 2011). This problem has been partially attributed to the lack of power of GWAS to detect small-effect variants, and some studies (e.g. Allen et al., 2010; Ober et al., 2012) have shown that the proportion of marker-driven variance and prediction accuracy could be improved when prediction models include variants that show strong, but not GWAS-significant association.

Several authors (e.g. de los Campos et al., 2010; Yang et al., 2010) have suggested the use of Whole-Genome Regression methods (Meuwissen et al., 2001), where phenotypes are regressed on potentially hundreds of thousands of variants concurrently, for analysis and prediction of complex human traits and diseases. In human genetic applications, the most commonly used WGR method has been the GBLUP (Gondro et al., 2013). This method has been used primarily for the estimation of missing heritability (e.g. Eichler et al., 2010; Yang et al., 2010; Speed et al., 2012). Only a few studies have assessed these methods from a prediction perspective. These studies have reported poor prediction performance of GBLUP when training and validation samples were distantly related (e.g. de los Campos et al., 2013a). This leaves open the question of what avenues should be pursued to improve the prediction performance of WGR methods when used for the prediction of phenotypes for distantly related individuals.

The prediction accuracy of WGR is known to be affected by many important factors, including genetic relationship (e.g. VanRaden et al., 2009; Crossa et al., 2010), trait heritability (e.g. Hayes et al., 2009; Daetwyler et al., 2010), marker density (e.g. Vazquez et al., 2010; Makowsky et al., 2011; Ober et al., 2012; Erbe et al., 2013; Vazquez et al., 2010), the genetic architecture of the model (e.g. the number of QTL, the distribution of effects (VanRaden et al., 2009; Wimmer et al., 2013), the extent of LD between markers and QTL (Habier et al., 2007; Calus et al., 2008), the sample size (Hayes et al., 2009; Makowsky et al., 2011) and the method used (e.g. Habier et al., 2007; Hayes et al., 2009; VanRaden et al., 2009; Verbyla et al., 2009; Gao et al., 2013; Wimmer et al., 2013; Zhang et al., 2014). The vast majority of studies that have compared the predictive performance of shrinkage and variable selection methods have used family data from populations with intensive history of recent selection. Indeed, there has been little, if any, assessment of the factors that affect the prediction accuracy of WGRs using human data from distantly related individuals. In this article we contributed towards filling this gap by conducting an extensive simulation study where we assessed the impact on estimated missing heritability and on prediction accuracy of: (a) the extent of LD between markers and QTL, (b) the complexity of the trait architecture, and (c) the statistical model used.

Missing heritability can be attributed to imperfect LD between marker and QTL genotypes (e.g. Goddard and Hayes, 2009; Yang et al., 2010; de los Campos et al., 2013a).

Therefore, in scenarios where QTL genotypes were used for analysis (either when QTL only or when both markers and QTL were used) there is no missing heritability because the causal loci were included in the set of genotypes used for data analysis. In these analysis scenarios (only QTL or markers and QTL), estimates of genomic heritability above or below the simulated heritability (0.5) reflect bias of the estimation method.

When the analysis was carried out using QTL genotypes only, the Spike-Slab and GBLUP methods yielded estimates very close to the simulated heritability, while Bayes A and GBLUP-ldak yielded substantial biases. In the case of Bayes A the estimate was downwardly biased in scenarios where a few QTL made a substantial contribution to genetic variance (e.g., p=50, pve=0.75) and GBLUP-ldak showed a clearly downwardly biased estimate in the RAND scenario.

When markers and QTL were used for analysis the results differed between the RAND and LOW-MAF scenarios: in the RAND scenario GBLUP and Spike-Slab yielded almost unbiased estimates, while Bayes A and GBLUP-ldak yielded upwardly biased estimates under simple genetic architectures. In the LOW-MAF scenario, GBLUP, Spike-Slab and Bayes A yielded downwardly biased estimates while estimates from GBLUP-ldak were slightly biased upwards.

Finally, in scenarios using only markers the estimated genomic heritability was very close to the trait heritability in the RAND scenario, while in the LOW-MAF scenario estimates revealed a substantial extent of missing heritability.

The observation that having a different distribution of allele frequencies at markers and at QTL can induce a large extent of missing heritability is in line with the reasoning and results presented in some studies (Goldstein, 2009; Yang et al., 2010; Lee et al., 2012; de los Campos et al., 2013a). This result is also in agreement with the fact that the extent of LD between markers and QTL in the LOW-MAF scenarios was much weaker than in the RAND scenarios (see Table 3.2). It should be noted that in all simulation scenarios considered in our study, including the LOW-MAF scenario, the frequency of rare variants among the QTL was limited relative to what one could have with sequence data, because the genotypes used in our study were all obtained from a panel of common SNPs. Therefore, one could speculate that the extent of differences in distribution of allele frequency between markers and causal loci and the corresponding extent of missing heritability may be even more extreme with real phenotypes than the one observed in our LOW-MAF scenario.

Importantly, within any scenario we found remarkable differences in estimates of genomic heritability across models, and there was no single method with smallest bias across all genetic architectures and analysis scenarios (QTL, markers+QTL or only markers).

The GBLUP and Spike-Slab methods performed well in the RAND scenario, but had clear problems in the LOW-MAF scenarios (both had seriously downwardly biased estimates in the analysis based on markers and QTL). On the other hand, GBLUP-ldak exhibited some clear problems in the RAND scenarios (downwardly biased estimates when analysis was based on QTL only) or upwardly biased estimates in the LOW-MAF analysis based on markers and QTL). Finally, Bayes A showed somewhat erratic behavior, especially with simple genetic architectures (e.g., p=50, pve=0.75); we believe that this is not a limitation of the model per-se but a consequence of the degree-of-freedom parameter being fixed. Estimating this parameter from the data, as done, for instance in (Yi and Xu, 2008), is likely to confer more flexibility to Bayes A to cope with different genetic architectures.

Prediction Accuracy. When the **analysis** was carried out **using only QTL genotypes** ('perfect LD', panels **c** and **f** of Figure 3.2) all methods achieved relatively high prediction accuracy (correlation of about 0.5 or greater, that is an R^2 50% or more of the trait heritability); this indicates that if one is able to narrow down the influential genetic regions of a trait to a limited number (5,000 loci in our simulation) regularized regressions like the one used here can yield relatively high prediction accuracy. In these scenarios, the prediction accuracy of the GBLUP and GBLUP-ldak methods was not affected by the genetic architecture and tended to be poorer than that of Bayes A and the Spike-Slab methods. Bayes A and Spike-Slab performed similarly and clearly better than any of the GBLUP methods in scenarios where a limited number of QTL (e.g., 50 or 250) explained a sizeable proportion of the genetic variance. However, with increase in trait complexity there was a decrease in prediction performance of these two methods, to the point that the three methods performed very similarly when the most complex genetic architecture was considered (5,000 QTL without any 'major effect' one). Overall, our results are in agreement with previous studies in animal and plant breeding (Daetwyler et al., 2010 and Wimmer et al., 2013) that have reported that: (a) the prediction accuracy of GBLUP is largely independent of the genetic architecture of the trait, and (b) with simple genetic architectures there are benefits of using methods such as Bayes B, Spike-Slab, Bayes C or Bayes A, relative to ridge-regression type-methods. However as the trait architecture became more complex, these differences disappeared.

When markers and QTL were jointly used (panels **b** and **e** of Figure 3.2) or when only markers were used (panels **a** and **d** in Figure 2), important changes in prediction accuracy were observed. The prediction accuracy of any of the GBLUP methods was reduced from correlation levels of the order of 0.45 (QTL-only analysis) to 0.15 when both markers and QTL were used, and to levels below 0.1 when only markers were used. This reflects the limitations of using methods such as GBLUP or GBLUP-ldak where the effects of all predictors are homogeneously shrunk, especially in situations where a large number of markers do not have effects.

In scenarios where 50 or 250 QTL explained a sizeable proportion (e.g., 0.75) of the genetic variance, the benefits of using methods that perform variable selection (Spike-Slab) or differential shrinkage of estimated effects (Bayes A) relative to the GBLUP methods were pronounced. In the scenario with the simplest genetic architecture (50 QTL explaining 75% of the genetic variance) these methods, especially the Spike-Slab were able to achieve levels of prediction accuracy comparable to those obtained when only QTL genotypes were used, illustrating the 'oracle' property (e.g. Ishwaran and Rao, 2005; Scheipl et al., 2013) that these methods have. However, as the complexity of the trait increased, the predictive performance of these methods decreased and in the most complex scenario (5,000 small QTL) all methods performed similarly.

Real data analysis. Human height is believed to be a trait affected by a very large number of small-effect QTL (e.g. Allen et al., 2010; Yang et al., 2010). The analysis conducted with human height data from the GENEVA data set very closely matched the results from the simulation for scenarios with large numbers of small effect QTL, where the distributions of allele frequency at markers and at QTL were different. We estimated a sizeable proportion of missing heritability, given a trait heritability of 0.8, the estimates of missing heritability ranged from 0.24 with GBLUP-ldak to 0.54 with Spike-Slab and very poor prediction accuracy (correlation of about 0.16-0.17, and very similar across methods).

Implications

The results presented in this study have several implications. Firstly, estimates of missing heritability derived from distantly related individuals using WGR methods need to be treated with caution; although they are indicative of how imperfect LD between markers and QTL can limit the ability of a model to capture the genetic signal, some of the results presented here indicate that under some circumstances estimates can have a sizeable bias. Additionally, we observed that in some scenarios these estimates of heritability can vary significantly between methods. This is not surprising because the proportion of variance explained by a model depends both on the input information (markers/QTL, etc.) and on the statistical model used. We believe that this model-genetic architecture dependency has been overlooked so far. Importantly, the model that yields the highest estimated genomic heritability is not necessarily the one that yields the best prediction accuracy.

Secondly, the assessment of prediction accuracy suggests that for traits in which a limited number of regions explain a sizeable proportion of genetic variance, the use of WGR methods that perform variable selection or differential shrinkage of estimates of effects is strongly recommended over ridge-regression type methods such as the GBLUP. On the other hand, for very complex traits such as human height all the methods evaluated yield low prediction accuracy. It remains to be determined whether significant increases in sample size (which likely should be by orders of magnitude) will also yield substantial gains in prediction accuracy.

Acknowledgments

The authors wish to thank the participants of the GENEVA study and Ana I. Vazquez and Malena Erbe for useful comments provided. During the review process we benefited from valuable contributions made by three anonymous reviewers.

This study was financially supported by Short-term International Mentorship Program and RTG 1644 ‘Scaling problem in statistics’, both financed by German Research Foundation (DFG). Gustavo de los Campos received financial support from NIH grants: R01GM099992 and R01GM101219.

The authors have declared that no competing interests exist.

References

Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., and Raychaudhuri, S. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

Calus, M.P.L., Meuwissen, T.H.E., Roos, A.P.W. de, and Veerkamp, R.F. (2008). Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics *178*, 553–561.

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

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

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

De los Campos, G., Sorensen, D., and Gianola, D. (2014). Genomic Heritability: What Is It? (Vancouver, BC, Canada),.

Crossa, J., de los Campos, G., Pérez-Rodrigues, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., and Yan, J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics *186*, 713–724.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., and Woolliams, J.A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics *185*, 1021–1031.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.

Gao, H., Su, G., Janss, L., Zhang, Y., and Lund, M.S. (2013). Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. J. Dairy Sci. *96*, 4678–4687.

George, E.I., and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. J. Am. Stat. Assoc. *88*, 881–889.

Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics.

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

Gibson, G. (2010). Hints of hidden heritability in GWAS. Nat. Genet. *42*, 558–560.

Goddard, M.E., and Hayes, B.J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. *10*, 381–391.

Goldstein, D.B. (2009). Common genetic variation and human traits. N. Engl. J. Med. *360*, 1696.

Gondro, C., Van der Werf, J., and Hayes, B. (2013). Genome-wide Association Studies and Genomic Prediction (Springer).

Guttmacher, A.E., Collins, F.S., Guttmacher, A.E., and Collins, F.S. (2002). Genomic Medicine — A Primer. N. Engl. J. Med. *347*, 1512–1520.

Habier, D., Fernando, R.L., and Dekkers, J.C.M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics *177*, 2389–2397.

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

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. Math. Intell. *27*, 83–85.

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

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

Ishwaran, H., and Rao, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. 730–773.

Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., Wray, N.R., and Consortium, S.P.G.-W.A.S. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.

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., and de los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. PLoS Genet. *7*, e1002051.

Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. *118*, 1590–1605.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., and Chakravarti, A. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

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

Ober, U., Ayroles, J.F., Stone, E.A., Richards, S., Zhu, D., Gibbs, R.A., Stricker, C., Gianola, D., Schlather, M., and Mackay, T.F. (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. *8*, e1002685.

Pérez, P., and de los Campos, G. (2014). Genome-wide regression & prediction with the BGLR statistical package. Genetics genetics – 114.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

R Core Team (2014). R: a language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing.

Scheipl, F., Kneib, Thomas, and Fahrmeir, L. (2013). Penalized likelihood and Bayesian function selection in regression models - Springer. Advances in Statistical Analysis *97*, 349–385.

Simon-Sanchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., and Hernandez, D.G. (2009). Genome-wide association study reveals genetic risk underlying Parkinson’s disease. Nat. Genet. *41*, 1308–1312.

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., and Schenkel, F.S. (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. *92*, 16–24.

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

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

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H.-J., Wang, Y., and Schön, C.-C. (2013). Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. Genetics *195*, 573–587.

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

Yi, N., and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. Genetics *179*, 1045–1055.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., Li, J., and Simianer, H. (2014). Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. PLoS ONE *9*, e93017.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.

Supporting Information

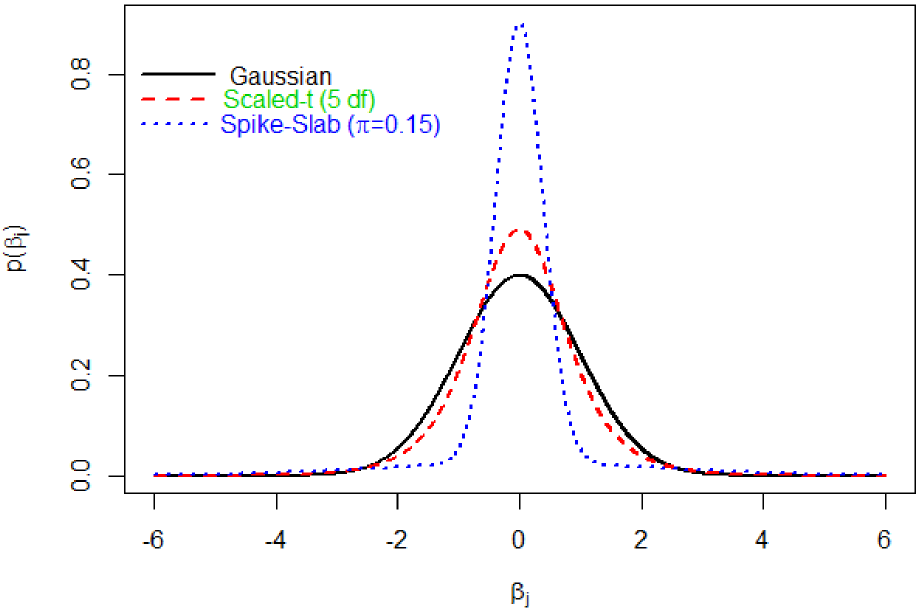


Figure S3.1. Prior distributions in Bayesian settings. Commonly used prior distributions for regression coefficients in Bayesian models (all with null mean and unit variance): Gaussian, Bayes A (scaled-t) and Spike-Slab (mixture of two normal distributions) models.

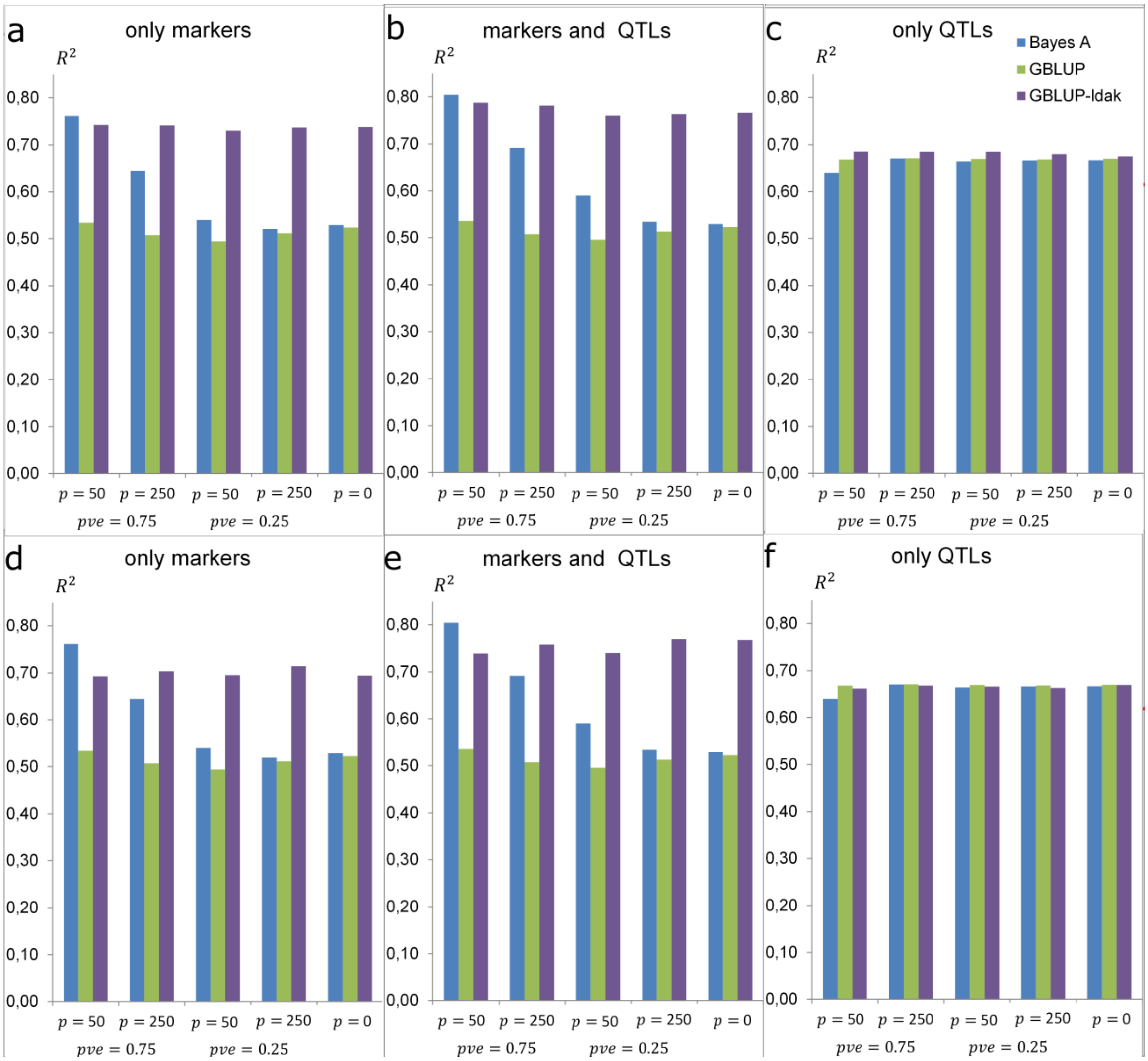


Figure S3.2. R-squared statistic in training data sets. R^2 (averaged over 30 MC replicates) between phenotype and genomic predictions in training data sets, by simulation scenario (RAND upper panel; LOW-MAF in lower panel), genetic architecture (*p*=number of large effect QTL, *pve*=proportion of genetic variance explained by large effect QTL) data used (only markers, markers and QTL or only QTL) and analysis method (GBLUP, GBLUP-ldak or Bayes A).

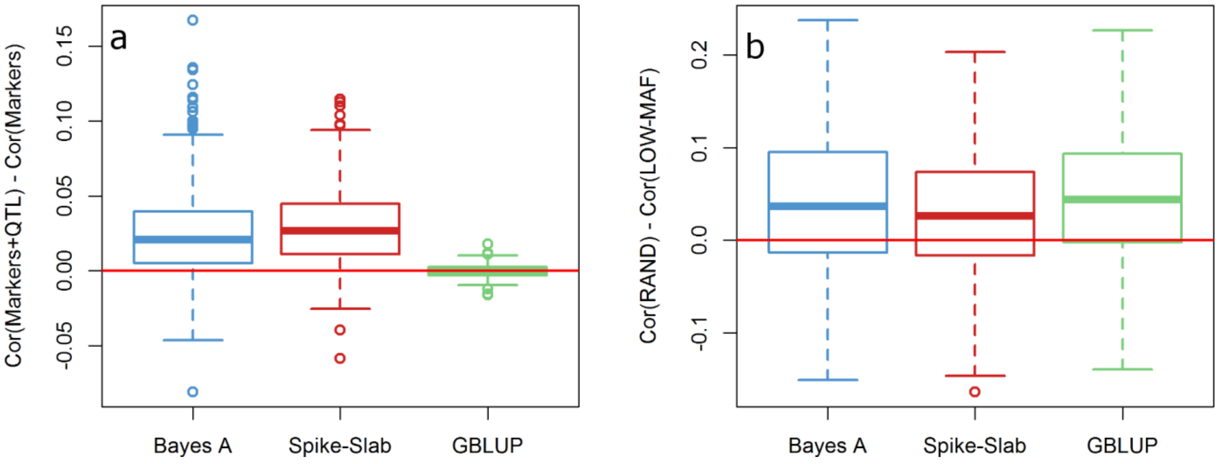


Figure S3.3. Difference in prediction accuracy in sampling scenarios and different types of data used. Difference in prediction accuracy obtained using markers and QTL minus that obtained using markers only (panel **a**) and the prediction accuracy obtained in the RAND scenario minus that obtained in the LOW-MAF scenario (panel **b**), by model.

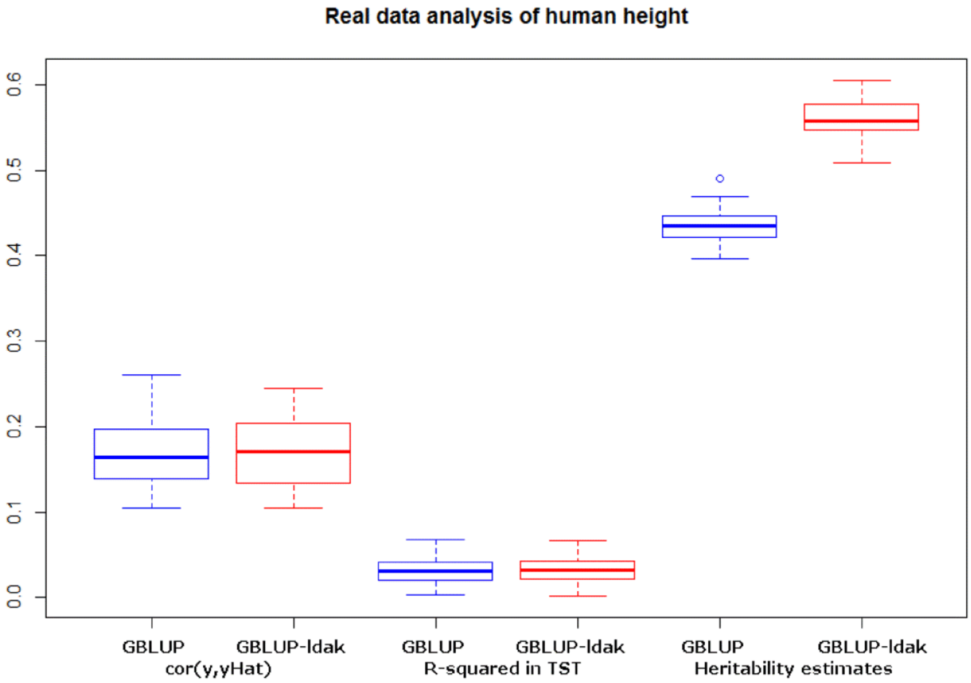


Figure S3.4. Differences between both GBLUP methods in the real data analysis of human height. Prediction accuracy, measured as the correlation between the true and predicted phenotype, Proportion of genetic variance explained as R-squared in TST and heritability estimates, obtained in GBLUP or in GBLUP-ldak.

Table S3.1. Heritability estimates $\hat{h}_g^2$ in GBLUP from 30 Monte Carlo replicates, across all configurations of effects in RAND scenario and genetic information used.

Data used	only markers					markers and QTLs					only QTLs				
LEQTLs*	50		250		none	50		250		none	50		250		none
pve**	25%	75%	25%	75%	-	25%	75%	25%	75%	-	25%	75%	25%	75%	-
run 1	0.58	0.43	0.61	0.49	0.53	0.59	0.44	0.62	0.49	0.54	0.49	0.50	0.49	0.53	0.49
run 2	0.53	0.53	0.55	0.58	0.58	0.53	0.54	0.55	0.58	0.59	0.50	0.47	0.47	0.51	0.48
run 3	0.42	0.64	0.58	0.41	0.37	0.43	0.65	0.59	0.41	0.38	0.52	0.50	0.51	0.51	0.46
run 4	0.55	0.42	0.50	0.54	0.62	0.54	0.42	0.51	0.55	0.62	0.50	0.50	0.49	0.48	0.52
run 5	0.53	0.50	0.47	0.47	0.52	0.54	0.51	0.49	0.48	0.52	0.47	0.50	0.47	0.49	0.48
run 6	0.51	0.64	0.53	0.49	0.51	0.51	0.64	0.53	0.49	0.51	0.50	0.49	0.51	0.49	0.49
run 7	0.50	0.52	0.52	0.51	0.48	0.50	0.52	0.54	0.51	0.49	0.51	0.50	0.50	0.49	0.50
run 8	0.55	0.39	0.56	0.52	0.50	0.55	0.40	0.57	0.54	0.51	0.49	0.47	0.49	0.50	0.51
run 9	0.50	0.44	0.44	0.49	0.54	0.52	0.45	0.45	0.49	0.55	0.53	0.49	0.50	0.49	0.49
run 10	0.47	0.64	0.43	0.63	0.56	0.49	0.65	0.43	0.64	0.57	0.52	0.52	0.46	0.51	0.52
run 11	0.51	0.53	0.43	0.47	0.49	0.52	0.53	0.44	0.48	0.50	0.49	0.49	0.45	0.49	0.45
run 12	0.49	0.53	0.37	0.60	0.55	0.50	0.54	0.38	0.61	0.56	0.47	0.50	0.51	0.50	0.52
run 13	0.41	0.65	0.63	0.54	0.51	0.41	0.66	0.64	0.55	0.51	0.51	0.49	0.51	0.49	0.50
run 14	0.50	0.44	0.63	0.41	0.51	0.50	0.44	0.64	0.41	0.52	0.50	0.53	0.50	0.50	0.49
run 15	0.49	0.46	0.48	0.53	0.61	0.49	0.48	0.49	0.53	0.62	0.49	0.52	0.52	0.50	0.50
run 16	0.46	0.51	0.53	0.48	0.51	0.46	0.52	0.54	0.49	0.52	0.50	0.51	0.49	0.49	0.49
run 17	0.52	0.42	0.49	0.39	0.51	0.53	0.43	0.50	0.40	0.51	0.50	0.48	0.47	0.50	0.48
run 18	0.51	0.60	0.44	0.51	0.68	0.51	0.61	0.44	0.52	0.69	0.48	0.50	0.50	0.49	0.52
run 19	0.51	0.56	0.46	0.53	0.57	0.52	0.59	0.47	0.53	0.58	0.51	0.49	0.49	0.49	0.50
run 20	0.61	0.50	0.50	0.49	0.54	0.63	0.49	0.51	0.48	0.54	0.47	0.51	0.51	0.52	0.48
run 21	0.42	0.57	0.61	0.48	0.54	0.43	0.57	0.62	0.48	0.55	0.47	0.49	0.48	0.48	0.51
run 22	0.51	0.50	0.45	0.45	0.62	0.52	0.51	0.45	0.46	0.63	0.49	0.50	0.50	0.48	0.51
run 23	0.45	0.39	0.47	0.54	0.53	0.45	0.38	0.48	0.55	0.53	0.49	0.48	0.51	0.50	0.53
run 24	0.37	0.62	0.49	0.57	0.42	0.37	0.63	0.49	0.57	0.42	0.49	0.47	0.51	0.49	0.51
run 25	0.50	0.54	0.45	0.57	0.48	0.50	0.54	0.46	0.57	0.50	0.47	0.52	0.48	0.48	0.51
run 26	0.49	0.43	0.67	0.44	0.46	0.49	0.43	0.68	0.43	0.46	0.50	0.50	0.50	0.51	0.49
run 27	0.53	0.50	0.49	0.62	0.40	0.55	0.50	0.49	0.62	0.41	0.49	0.50	0.52	0.50	0.51
run 28	0.53	0.47	0.62	0.42	0.43	0.54	0.48	0.63	0.44	0.43	0.50	0.51	0.51	0.49	0.48
run 29	0.52	0.47	0.44	0.39	0.42	0.53	0.48	0.44	0.39	0.42	0.53	0.48	0.51	0.50	0.51
run 30	0.53	0.60	0.46	0.48	0.48	0.53	0.61	0.47	0.50	0.49	0.50	0.51	0.50	0.48	0.51
aver-	0.50	0.52	0.51	0.50	0.52	0.51	0.52	0.52	0.51	0.52	0.50	0.50	0.50	0.50	0.50
sd	0.05	0.07	0.07	0.06	0.06	0.05	0.08	0.07	0.06	0.06	0.01	0.01	0.02	0.01	0.01

*: Number of Large Effect QTL **: % of Genetic Variance Explained by Large Effect QTL

Table S3.2. Heritability estimates $\hat{h}_g^2$ in GBLUP from 30 Monte Carlo replicates, across all configurations of effects in LOW-MAF scenario and genetic information used.

Data used	only markers					markers and QTLs					only QTLs				
LEQTLs .	50		250		none	50		250		none	50		250		none
pve**	25%	75%	25%	75%	-	25%	75%	25%	75%	-	25%	75%	25%	75%	-
run 1	0.34	0.19	0.32	0.32	0.281	0.34	0.19	0.32	0.33	0.287	0.51	0.48	0.45	0.46	0.470
run 2	0.35	0.28	0.25	0.25	0.366	0.35	0.27	0.25	0.25	0.372	0.45	0.47	0.47	0.49	0.469
run 3	0.20	0.38	0.32	0.20	0.335	0.19	0.38	0.33	0.20	0.333	0.49	0.47	0.50	0.48	0.480
run 4	0.32	0.32	0.29	0.30	0.416	0.32	0.33	0.29	0.30	0.413	0.50	0.47	0.49	0.49	0.484
run 5	0.27	0.29	0.27	0.39	0.293	0.27	0.30	0.27	0.39	0.291	0.47	0.50	0.46	0.49	0.498
run 6	0.36	0.40	0.34	0.32	0.236	0.36	0.40	0.34	0.32	0.240	0.51	0.47	0.48	0.48	0.520
run 7	0.27	0.32	0.33	0.37	0.375	0.27	0.33	0.33	0.37	0.373	0.51	0.48	0.52	0.50	0.505
run 8	0.36	0.23	0.26	0.35	0.462	0.36	0.23	0.26	0.34	0.463	0.49	0.45	0.48	0.49	0.495
run 9	0.25	0.30	0.38	0.27	0.274	0.25	0.31	0.39	0.27	0.269	0.52	0.51	0.48	0.47	0.495
run 10	0.31	0.44	0.29	0.26	0.325	0.31	0.44	0.29	0.26	0.320	0.49	0.49	0.47	0.48	0.470
run 11	0.27	0.27	0.27	0.21	0.248	0.27	0.27	0.28	0.20	0.247	0.48	0.48	0.47	0.49	0.482
run 12	0.26	0.33	0.27	0.28	0.435	0.26	0.33	0.27	0.29	0.436	0.46	0.49	0.47	0.48	0.492
run 13	0.28	0.26	0.45	0.32	0.255	0.29	0.27	0.45	0.31	0.250	0.49	0.48	0.46	0.43	0.491
run 14	0.28	0.34	0.24	0.29	0.428	0.28	0.35	0.25	0.29	0.427	0.47	0.47	0.48	0.46	0.511
run 15	0.26	0.34	0.33	0.34	0.424	0.25	0.34	0.33	0.34	0.416	0.47	0.47	0.49	0.50	0.502
run 16	0.22	0.29	0.37	0.26	0.332	0.22	0.30	0.37	0.26	0.343	0.51	0.51	0.48	0.46	0.475
run 17	0.29	0.35	0.24	0.41	0.230	0.29	0.35	0.24	0.40	0.229	0.47	0.48	0.48	0.48	0.454
run 18	0.25	0.27	0.22	0.32	0.288	0.26	0.27	0.23	0.33	0.284	0.48	0.50	0.50	0.48	0.443
run 19	0.28	0.34	0.23	0.27	0.277	0.28	0.34	0.24	0.27	0.279	0.49	0.46	0.46	0.46	0.490
run 20	0.38	0.52	0.26	0.22	0.292	0.38	0.51	0.27	0.23	0.291	0.48	0.50	0.47	0.49	0.472
run 21	0.34	0.36	0.41	0.35	0.282	0.35	0.36	0.41	0.35	0.289	0.51	0.48	0.50	0.45	0.471
run 22	0.22	0.32	0.23	0.37	0.406	0.22	0.32	0.24	0.37	0.408	0.46	0.48	0.48	0.48	0.505
run 23	0.22	0.25	0.36	0.33	0.424	0.23	0.26	0.36	0.32	0.422	0.47	0.45	0.45	0.49	0.493
run 24	0.23	0.31	0.27	0.30	0.196	0.24	0.31	0.27	0.30	0.196	0.47	0.47	0.50	0.47	0.463
run 25	0.28	0.30	0.27	0.29	0.342	0.28	0.30	0.27	0.29	0.339	0.49	0.47	0.48	0.47	0.489
run 26	0.36	0.40	0.40	0.27	0.337	0.36	0.39	0.40	0.27	0.335	0.47	0.47	0.46	0.50	0.484
run 27	0.38	0.26	0.36	0.34	0.249	0.39	0.26	0.35	0.34	0.252	0.48	0.48	0.48	0.48	0.484
run 28	0.33	0.34	0.34	0.31	0.260	0.34	0.33	0.35	0.31	0.260	0.48	0.47	0.49	0.51	0.472
run 29	0.27	0.34	0.35	0.31	0.304	0.27	0.34	0.34	0.31	0.300	0.47	0.46	0.47	0.50	0.449
run 30	0.31	0.35	0.22	0.24	0.231	0.31	0.35	0.23	0.24	0.233	0.47	0.50	0.49	0.47	0.517
aver-	0.29	0.32	0.31	0.30	0.320	0.29	0.32	0.31	0.30	0.320	0.48	0.48	0.48	0.48	0.484
sd	0.05	0.06	0.06	0.05	0.072	0.05	0.06	0.05	0.05	0.072	0.01	0.01	0.01	0.01	0.019

*: Number of Large Effect QTL **: % of Genetic Variance Explained by Large Effect QTL

Table S3.3. Correlation between true and predicted phenotype $cor(\mathbf{y}, \hat{\mathbf{y}})$: Average (SD) over all 30 replications.

Simulation Scenarios			Data Analysis Method & Information Used								
			GBLUP			BayesA			Spike-Slab		
Number of Large Effect QTL	% of Genetic Variance Explained by Large Effect QTL	Sampl-ing of QTL	Markers	Markers+QTL	QTL	Markers	Markers+QTL	QTL	Markers	Markers+QTL	QTL
50	25	UNIF	0.174(0.04)	0.174 (0.04)	0.463 (0.04)	0.283 (0.06)	0.303 (0.05)	0.513 (0.05)	0.309 (0.05)	0.331 (0.05)	0.513 (0.04)
		LOW-MAF	0.104 (0.05)	0.104 (0.05)	0.447 (0.04)	0.200 (0.08)	0.236 (0.05)	0.516 (0.04)	0.268 (0.05)	0.305 (0.05)	0.504 (0.04)
	75	UNIF	0.147 (0.05)	0.147 (0.05)	0.442 (0.05)	0.512 (0.04)	0.547 (0.04)	0.604 (0.04)	0.554 (0.04)	0.581 (0.03)	0.601 (0.03)
		LOW-MAF	0.085 (0.04)	0.085 (0.04)	0.427 (0.04)	0.473 (0.05)	0.524 (0.04)	0.607 (0.03)	0.528 (0.04)	0.570 (0.03)	0.601 (0.03)
250	25	UNIF	0.158 (0.05)	0.157 (0.05)	0.459 (0.04)	0.178 (0.06)	0.193 (0.05)	0.492 (0.04)	0.209 (0.06)	0.227 (0.05)	0.488 (0.04)
		LOW-MAF	0.086 (0.05)	0.085 (0.05)	0.429 (0.04)	0.110 (0.05)	0.111 (0.06)	0.485 (0.05)	0.155 (0.06)	0.175 (0.06)	0.465 (0.04)
	75	UNIF	0.153 (0.04)	0.154 (0.04)	0.434 (0.04)	0.330 (0.06)	0.376 (0.05)	0.550 (0.04)	0.443 (0.05)	0.483 (0.05)	0.565 (0.04)
		LOW-MAF	0.105 (0.04)	0.106 (0.04)	0.440 (0.04)	0.286 (0.05)	0.325 (0.03)	0.564 (0.04)	0.420 (0.05)	0.470 (0.04)	0.564 (0.04)
None	---	UNIF	0.155 (0.06)	0.153 (0.06)	0.445 (0.05)	0.143 (0.07)	0.162 (0.06)	0.447 (0.04)	0.161 (0.06)	0.179 (0.06)	0.464 (0.05)
		LOW-MAF	0.094 (0.05)	0.094 (0.05)	0.449 (0.04)	0.069 (0.05)	0.077 (0.05)	0.475 (0.04)	0.080 (0.05)	0.095 (0.05)	0.476 (0.04)

Table S3.4. R-squared in validation group: Average (SD) over 30 replications.

Simulation Scenarios			Data Analysis Method & Information Used								
			GBLUP			BayesA			Spike-Slab		
Number of Large Effect QTL	% of Genetic Variance Explained by Large Effect QTL	Sampl-ing of QTL	Markers	Markers+QTL	QTL	Markers	Markers+QTL	QTL	Markers	Markers+QTL	QTL
50	25	UNIF	0.031 (0.01)	0.031 (0.01)	0.216 (0.03)	0.082 (0.03)	0.093 (0.03)	0.264 (0.05)	0.097 (0.03)	0.111 (0.03)	0.264 (0.04)
		LOW-MAF	0.012 (0.01)	0.012 (0.01)	0.201 (0.03)	0.037 (0.05)	0.058 (0.03)	0.267 (0.04)	0.073 (0.03)	0.094 (0.03)	0.255 (0.03)
	75	UNIF	0.025 (0.02)	0.025 (0.02)	0.197 (0.04)	0.264 (0.04)	0.300 (0.04)	0.366 (0.05)	0.309 (0.04)	0.339 (0.04)	0.373 (0.04)
		LOW-MAF	0.011 (0.01)	0.010 (0.01)	0.185 (0.04)	0.227 (0.04)	0.277 (0.04)	0.370 (0.03)	0.281 (0.04)	0.327 (0.04)	0.362 (0.03)
250	25	UNIF	0.027 (0.01)	0.026 (0.01)	0.212 (0.04)	0.304 (0.02)	0.039 (0.02)	0.244 (0.04)	0.046 (0.02)	0.054 (0.02)	0.240 (0.04)
		LOW-MAF	0.009 (0.01)	0.009 (0.01)	0.185 (0.04)	0.014 (0.01)	0.014 (0.02)	0.235 (0.05)	0.027 (0.02)	0.033 (0.02)	0.217 (0.04)
	75	UNIF	0.025 (0.01)	0.026 (0.01)	0.189 (0.04)	0.111 (0.04)	0.144 (0.04)	0.304 (0.04)	0.198 (0.04)	0.235 (0.05)	0.321 (0.05)
		LOW-MAF	0.016 (0.01)	0.016 (0.01)	0.197 (0.03)	0.086 (0.03)	0.108 (0.02)	0.321 (0.04)	0.180 (0.05)	0.224 (0.04)	0.321 (0.04)
None	---	UNIF	0.026 (0.02)	0.025 (0.02)	0.199 (0.04)	0.013 (0.04)	0.028 (0.02)	0.200 (0.04)	0.028 (0.02)	0.034 (0.02)	0.216 (0.05)
		LOW-MAF	0.011 (0.01)	0.011 (0.01)	0.203 (0.03)	0.005 (0.01)	0.007 (0.01)	0.227 (0.03)	0.008 (0.01)	0.010 (0.01)	0.229 (0.03)

TableS3.5. Correlation and R-squared between human height and genomic predictions in testing data sets by method and testing set.

	Correlation			R-squared		
Method	Bayes A	Spike-Slab	GBLUP	Bayes A	Spike-Slab	GBLUP
run 1	0.238	0.244	0.247	0.068	0.065	0.067
run 2	0.107	0.106	0.109	0.001	0.003	0.007
run 3	0.122	0.133	0.130	0.003	0.014	0.012
run 4	0.153	0.155	0.180	0.021	0.025	0.034
run 5	0.138	0.146	0.148	0.016	0.022	0.023
run 6	0.254	0.269	0.261	0.061	0.058	0.057
run 7	0.231	0.228	0.233	0.053	0.050	0.052
run 8	0.131	0.137	0.146	0.011	0.019	0.021
run 9	0.142	0.152	0.166	0.012	0.021	0.027
run 10	0.205	0.232	0.219	0.045	0.053	0.049
run 11	0.170	0.176	0.194	0.031	0.035	0.041
run 12	0.157	0.158	0.160	0.029	0.033	0.034
run 13	0.117	0.146	0.115	0.004	0.020	0.007
run 14	0.128	0.126	0.133	0.010	0.014	0.016
run 15	0.174	0.153	0.178	0.035	0.029	0.038
run 16	0.143	0.155	0.164	0.024	0.033	0.036
run 17	0.210	0.221	0.227	0.044	0.048	0.052
run 18	0.199	0.214	0.217	0.040	0.045	0.047
run 19	0.176	0.178	0.200	0.034	0.035	0.043
run 20	0.083	0.103	0.109	-0.004	0.010	0.011
run 21	0.089	0.096	0.105	-0.011	0.000	0.002
run 22	0.126	0.128	0.141	0.005	0.012	0.016
run 23	0.171	0.175	0.185	0.030	0.034	0.037
run 24	0.209	0.195	0.204	0.043	0.037	0.041
run 25	0.124	0.129	0.122	0.026	0.033	0.030
run 26	0.120	0.136	0.145	0.010	0.020	0.023
run 27	0.134	0.139	0.137	0.014	0.021	0.019
run 28	0.160	0.166	0.160	0.021	0.027	0.025
run 29	0.187	0.181	0.179	0.034	0.033	0.032
run 30	0.174	0.172	0.163	0.033	0.034	0.031
average	0.159	0.165	0.169	0.025	0.029	0.031
sd	0.044	0.043	0.043	0.019	0.016	0.016

4TH CHAPTER

A scale-corrected comparison of linkage disequilibrium level between genic and non-genic regions

SWETLANA BERGER[1], MARTIN SCHLATHER[2], GUSTAVO DE LOS CAMPOS[3],
STEFFEN WEIGEND[4], RUDOLF PREISINGER[5], MALENA ERBE[1],
HENNER SIMIANER[1]

1. Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University, Goettingen, Germany
2. School of Business Informatics and Mathematics, University of Mannheim, Mannheim, Germany
3. Biostatistics Department, University of Alabama at Birmingham, Birmingham, Alabama, US
4. Institut of Farm Animal Genetics, Friedrich Loeffler Institut, Neustadt-Mariensee, Germany
5. Lohmann Tierzucht GmbH, Cuxhaven, Germany

In revision by PLoS One

Abstract

The understanding of non-random association between loci, termed linkage disequilibrium (LD), plays a central role in genomic research. Since causal mutations are generally not included in genomic marker data, LD between those and available markers is essential for capturing the effects of causal loci on localizing genes responsible for traits. Thus, the interpretation of association studies requires a detailed knowledge of LD patterns. It is well-known that most LD measures depend on minor allele frequencies (MAF) of the considered loci and the magnitude of LD is influenced by the physical distances between loci.

In the present study, a procedure to compare the LD structure between genomic regions comprising several markers each is suggested. The approach accounts for different scaling factors, namely the distribution of MAF, the distribution of pair-wise differences in MAF, and the physical extent of compared regions, reflected by the distribution of pair-wise physical distances. In the first step, genomic regions are matched based on similarity in these scaling factors. In the second step, chromosome- and genome-wide significance tests for differences in medians of LD measures in each pair are performed.

The proposed framework was applied to test the hypothesis that the average LD is different in genic and non-genic regions. This was tested with a genome-wide approach with data sets for humans (*Homo sapiens*), a highly selected chicken line (*Gallus gallus domesticus*) and the model plant *Arabidopsis thaliana*. In all three data sets we found a significantly higher level of LD in genic regions compared to non-genic regions. About 31% more LD was detected genome-wide in genic compared to non-genic regions in *Arabidopsis thaliana*, followed by 13.6% in human and 6% chicken. Chromosome-wide comparison discovered significant differences on all 5 chromosomes in *Arabidopsis thaliana* and on one third of the human and of the chicken chromosomes.

Author Summary

Non-random association between loci, termed linkage disequilibrium (LD), is a central parameter in genetic studies. Most LD measures are highly affected by the constellation of minor allele frequencies (MAF) and physical distances of the considered loci. In this study, we suggest a novel procedure to compare the LD structure between genomic regions comprising several markers each, which accounts for different scaling factors. To avoid a scale-caused bias, the distribution of MAF, the distribution of pair-wise differences in MAF, and the distribution of pair-wise physical distances were considered. In the first step we matched genomic regions based on similarity in these scaling factors and in the next step we applied significance tests for differences in LD measures in each matched pair. We hypothesized a difference in LD average in genic compared to non-genic regions and tested this hypothesis with real data sets for humans, a highly selected chicken line and the model plant *Arabidopsis thaliana*. In genome-wide comparisons we detected 31% more genic LD in *Arabidopsis thaliana*, followed by 13.6% in human and 6% in chicken. In chromosome-wide comparisons we discovered significant differences on all chromosomes in *Arabidopsis thaliana* and on one third of the human and of the chicken chromosomes.

Introduction

In genomic studies, associations between traits of interest and genomic polymorphisms are sought. In most whole genome marker data sets, the causal variants are generally not included but the effects of quantitative loci are reflected by markers that are in linkage disequilibrium (LD) with the causal loci (e.g. Jorde, 1995). For this reason, LD has become particularly instrumental in mapping genes that cause diseases (McVean et al., 2004; Meyer-Lindenberg et al., 2006; Lin et al., 2004). LD patterns also reflect the demographic development and demographic processes like migration and admixture and can be used to infer respective parameters (e.g. McVean et al., 2004; Ardlie et al., 2002; Smith et al., 2005). Awareness of LD patterns in the genome is thereby essential for correctly interpreting results from Genome-Wide Association Studies (GWAS). Rare variants will only be captured if they are in high LD with observable markers, which is only possible if the MAF of the causal variant and the marker are of similar magnitude (Meuwissen et al., 2002; Zondervan and Cardon, 2004). In populations with a limited effective population size, such as breeding populations, high LD extends over long physical distances. In such cases, methods utilizing LD mapping allow for more efficient usage of low density single nucleotide polymorphism (SNP) chips already available for genomic selection (Meuwissen et al., 2002; Zhao et al., 2007; Xu et al., 2013).

Large-scale data from high density SNP chips provide fine scale resolution LD maps for many species (Kruglyak, 1999; La Chapelle and Wright, 1998; Kim et al., 2007) and can be used to analyze the genome-wide LD structure. A wide range of scientific insights or groundbreaking findings based on LD patterns has been gained in human genetics (Huttley et al., 1999; Conrad et al., 2006; Smith et al., 2006) and in population genetics (La Chapelle and Wright, 1998; Hill, 1981; Mueller et al., 2005).

Factors like mutation, recombination, selection, or genetic drift have a strong impact on the development and dynamics of the non-random association between loci. Influence of MAF on LD is disturbing the genetic analysis. Both, the decay of the non-random association between the SNPs with growing physical distance (e.g. La Chapelle and Wright, 1998) and the dependency of most measures of LD on minor allele frequency (MAF) are well known (Mueller, 2004). Hence, different remedies have been suggested. For instance, Garner and Slatkin (2003) used a subset of markers selected on the basis of allele frequencies for association studies, other methods (e.g. Lewontin, 1988; Morton et al., 2001) are based on various kinds of standardization to minimize the influence of MAF on LD measures. For example, the dependency of the disequilibrium coefficient D on MAF is reduced by standardizing with its maximum, but the resulting measure reaches its maximum value only if less than four gametes are observed. Other less MAF dependent methods need haplotype data (e.g. index of association, homozygosity of haplotypes (Agapow and Burt, 2001), normalized entropy difference (Zhao et al., 2005) or are of parametric nature (e.g. Kullback-Leibler distance (Gianola et al., 2012))).

Deeper insight into the LD structure of the genome, especially in genic regions, will also help to identify relationships between traits of interest and genetic variants, to improve the understanding of biological processes and also may increase the accuracy of estimating genomic effects. Many studies investigating the association between the loci compare the LD level in different populations (e.g. Conrad et al., 2006; Reich et al., 2001), but only a few studies compared the magnitude of the LD in genic versus non-genic regions. McVean et al. (2004) indicated higher recombination rates outside of genic regions in the human genome, suggesting a higher rate of LD within genes. Smith et al. (2005) reported the proportion of genes in different quartiles of LD, while Kim et al. (2007) presented the proportion of genic markers in LD hotspots. Eberle et al. (2006) evaluated the decay of LD in genic and inter-genic regions by assessing the number of perfectly correlated SNPs. To avoid the bias due to differences in MAF, the authors used only a small subset of available SNPs for the analysis that had identical MAF. Eberle et al. (2006) observed a higher fraction of perfectly correlated SNPs in genic regions compared to intergenic regions, however these observations are valid only for the specific subset of SNPs and cannot be automatically generalized to other not pre-selected sets of SNPs. So far, a general procedure for comparing LD levels between different genomic regions that uses the comprehensive information and accounts for various potential sources of bias is missing. A key challenge when comparing LD patterns between different regions in the genome is to eliminate the impact of MAF on LD. An additional difficulty is that the density of markers varies across chromosomes and different SNP chips (Simianer and Erbe, 2014) and is different for genic and non-genic regions, which may lead to a structural bias on LD measures.

To overcome the MAF driven limitations of LD measures and the bias caused by genome topology variations we propose a general framework for comparison of LD magnitude in different genomic regions by applying the following methodology, which is structurally similar to matched pairs design used in clinical studies (e.g. Laska et al., 1975): (a) identification of pairs of regions with most similar characteristics (MAFs, pairwise MAF differences, pairwise physical distances), (b) determination of the LD levels for each matched pair of regions, and (c) application of the Wilcoxon signed rank test to the paired LD measures at chromosome-wide or genome-wide level. Best matching regions are identified by comparing the empirical cumulative distribution functions (ECDF) of the considered variables in both regions. To assess the extent of linkage disequilibrium we used the squared correlation (r^2) derived from phased haplotypes, a widely used statistic describing the association between two loci

(Mueller, 2004). We rescaled r^2 using the bounds given by VanLiere and Rosenberg (2008) to achieve a less MAF dependent measure of LD. The suggested approach was applied to test the hypothesis that the level of LD is higher in genic than in non-genic regions. We applied our approach to three real data sets: for humans (*Homo sapiens*), a highly selected chicken line (*Gallus gallus domesticus*) and the model plant *Arabidopsis thaliana*.

Materials and Methods

Statistical methods

In a diploid organism, there are four possible combinations of alleles at two bi-allelic loci (locus 1 with major allele *A* or minor allele *a* and locus 2 with major allele *B* or minor allele *b*) called gametes *AB*, *Ab*, *aB* or *ab*. For ease of notation, only the frequencies of minor alleles p_1 at locus 1 and p_2 at locus 2 were used, since the major allele frequencies can be expressed as 1-p_1 and 1-p_2, respectively. The coefficient of gametic (phase) disequilibrium D, also called disequilibrium coefficient, measures the differences between the observed frequency p_{12} of gamete *ab* and its expectation under independence, yielding $D = p_{12} - p_1 p_2$.

The disequilibrium coefficient D builds a basis for several measures of allelic association. Pearson's correlation coefficient r for a 2x2 contingency table representing gametic frequencies can be rewritten as $r = \frac{D}{\sqrt{p_1(1-p_1)p_2(1-p_2)}}$. Note that the absolute value, but not the sign of r is insensitive to an arbitrary labeling of alleles, and thus the Pearson's squared correlation coefficient r^2 is an appropriate measure of LD which was first used by Hill and Robertson (1968) to describe the extent of LD in finite populations. The authors also recognized that the range (and other characteristics) of this statistic depend on the allele frequencies, which was intensively considered in later studies (e.g. Devlin and Risch, 1995; Hedrick, 2005; Wray, 2005). VanLiere and Rosenberg (2008) suggested $r_S^2 = r^2 / r_{\max}^2$, where $r_{\max}^2$ is the maximum possible value of r^2 given the respective MAFs at the two loci considered. For our studies, squared correlations r^2 as well as r_S^2 were used to determine the amount of LD in compared genomic regions.

For the calculation of the upper limit $r_{\max}^2$ we extended the results presented by VanLiere and Rosenberg (2008) and provided a formal derivation of limiting bounds for gametic frequency p_{12}. For this reason the manifestation of different alleles at one locus was treated as a realization of a Bernoulli random variable, where the appearance of the minor allele was defined as a success. Thus, the bounds for p_{12} are obtained by applying Fréchet-Hoeffding bounds

(Fréchet, 1960; Rüschendorf, 1981) on Bernoulli distributed random variables $X_1 \sim B(p_1)$ and $X_2 \sim B(p_2)$ with success probabilities $p_1 = P(X_1 = 1)$ and $p_2 = P(X_2 = 1)$, for details see Appendix 1.

For known minor allele frequencies p_1 and p_2 with $p_2 \geq p_1$ and the difference $\delta = p_2 - p_1$, the upper limit for r^2 was given by

$$r_{max}^2(\delta, p_2) = 1 - \frac{\delta}{p_2\,(\delta + 1 - p_2)}$$

which equals to the upper limit suggested by VanLiere and Rosenberg (2008). Note that this upper limit equals the odds ratio, which is commonly used in the survey research or in case-control studies in the human medicine.

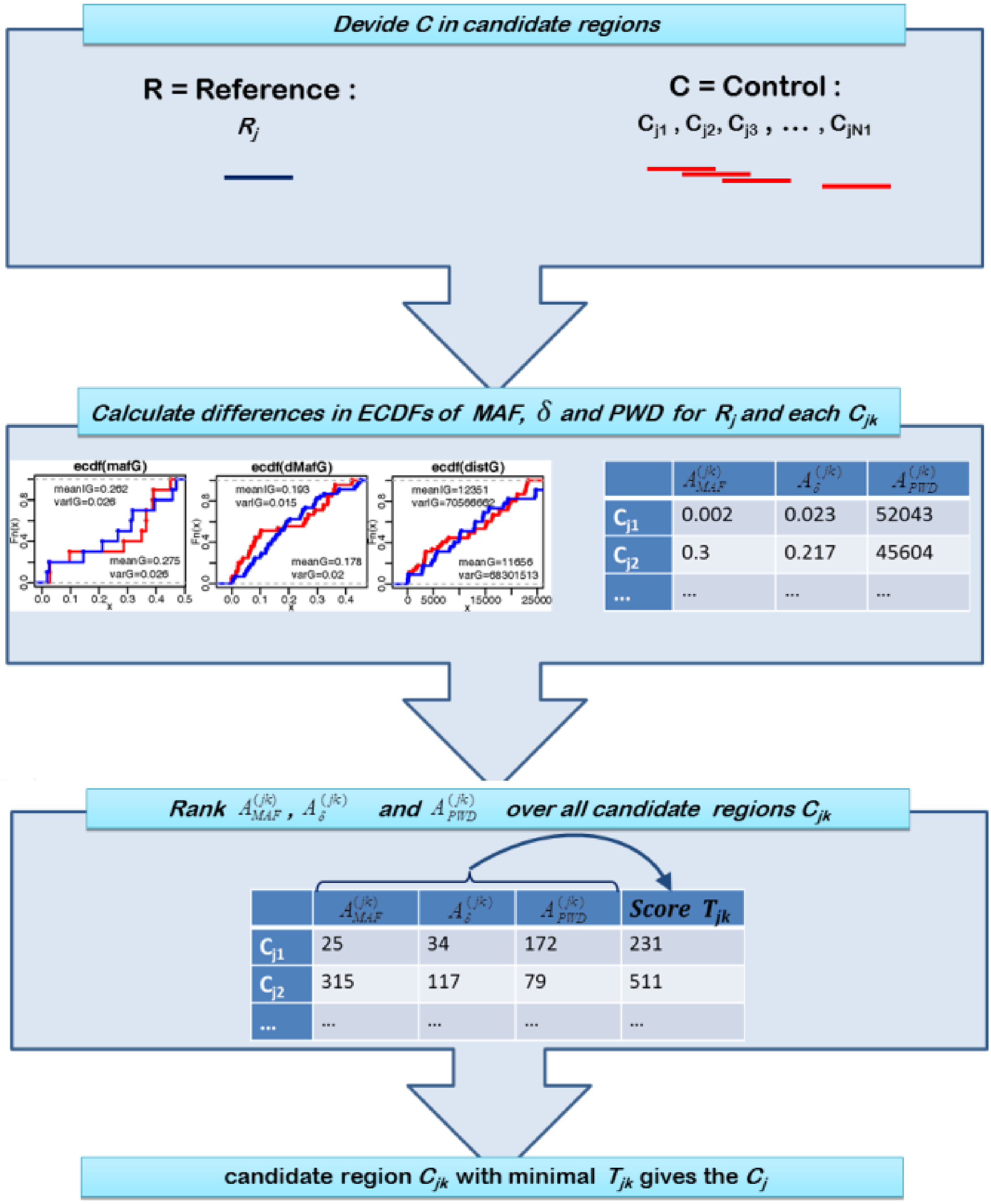
A more general upper limit, based only on the differences in MAFs δ (for details see Appendix 1), is given by

$$r_{max}^2(\delta) = 1 - \frac{4\delta}{2\delta+1}.$$

Accounting for scale effects

We consider the general problem of testing whether the LD structure differs between certain genomic regions, such as genic vs. non-genic regions, each region being represented by a number of sets of SNPs (a set may e.g. represent all SNPs in a gene). The basic idea of our approach is, similar to the matched pairs design (Laska et al., 1975), for a given reference set of SNPs to find a best matching control set (a set may e.g. represent SNPs in a non-genic chromosomal region) with the same number of SNPs that is most similar in all characteristics known to affect the LD measures. For each pair of matching sets, LD measures were calculated and averaged. Finally statistical tests were performed across all pairs of sets to verify whether the median differences are significantly different.

Identifying best matching sets. We denoted a reference set (for example a gene) consisting of m_j SNPs as R_j, and the best matching set of markers with the most similar characteristics on the chosen scales as the control set C_j (for example subset of markers from a non-genic region). We used MAFs, pairwise differences between the MAFs (δ), and pairwise physical distances (PWD) as most relevant characteristics to identify similarity between genomic regions. To identify this best matching control set C_j, the control region was divided into N_j candidate subsets $C_{j1},\ldots,C_{jk},\ldots C_{jN_j}$ by sliding windows of size m_j SNPs (see Fig. 4.1).


Devide C in candidate regions
R = Reference :
R_j
C = Control :
C_{j1} , C_{j2}, C_{j3} , … , C_{jN1}
Calculate differences in ECDFs of MAF, δ and PWD for R_j and each C_{jk}
ecdf(mafG)
meanlG=0.262
varlG=0.026
meanG=0.275
varG=0.026
ecdf(dMafG)
meanlG=0.193
varlG=0.015
meanG=0.178
varG=0.02
ecdf(distG)
meanlG=12351
varlG=70566662
meanG=11656
varG=68301513
Fn(x)
x
Rank $A_{MAF}^{(jk)}$, $A_{\delta}^{(jk)}$ and $A_{PWD}^{(jk)}$ over all candidate regions C_{jk}
candidate region C_{jk} with minimal T_{jk} gives the C_j

	$A_{MAF}^{(jk)}$	$A_{\delta}^{(jk)}$	$A_{PWD}^{(jk)}$
C_{j1}	0.002	0.023	52043
C_{j2}	0.3	0.217	45604
…	…	…	…

	$A_{MAF}^{(jk)}$	$A_{\delta}^{(jk)}$	$A_{PWD}^{(jk)}$	Score T_{jk}
C_{j1}	25	34	172	231
C_{j2}	315	117	79	511
…	…	…	…	…

Figure 4.1. Work flow for identifying best matching sets

The larger the reference set, the smaller the number of candidate subsets N_j. To achieve stability of estimates, we excluded any reference sets with less than 10 SNPs or less than 50 candidate subsets C_{jk} from further analysis, since a sufficient similarity between R_j and the best matching C_j might not be assured in these cases.

For each reference set R_j and candidate subset C_{jk}, the empirical cumulative distribution functions of MAFs, pairwise differences between the MAFs, and pairwise physical distances, were calculated separately. For each of the variables the area (A) between the ECDF curves for the reference set R_j and candidate subset C_{jk}, (also called Wasserstein metric (Vaserstein, 1969; Dobrushin, 1970) was determined, which was denoted as $A_{MAF}^{(jk)}$, $A_{\delta}^{(jk)}$, and $A_{PWD}^{(jk)}$, respectively (an example is given in Fig. S4.1). For selecting a control set C_{jk} which is most similar in all characteristics, we ranked firstly all $A_{MAF}^{(jk)}$, $A_{\delta}^{(jk)}$ and $A_{PWD}^{(jk)}$ over $k = 1,\ldots,N_j$ in each characteristic separately. Finally an overall score $T_{j1},\ldots,T_{jk},\ldots,T_{jN_j}$ was built by summing up those three ranks for each C_{jk} to a total score T_{jk}. The candidate subset C_{jk} with the lowest overall score was linked as matching control set C_j to the reference set R_j.

Determining the differences in LD level and statistical significance testing. For all pairs of SNPs within each R_j and each C_j we calculated r^2 and determined their medians $\hat{m}_{R_j}$ and $\hat{m}_{C_j}$, respectively. The Wilcoxon signed rank test was then applied to compare the LD level in both regions and to test the null hypothesis that the median difference between pairs of m_{R_j} and m_{C_j} is equal to zero against the alternative hypothesis that this median difference is not equal to zero (two-sided testing). The comparisons are performed chromosome-wise as well as at the genome-wide level. Similar calculations were performed for r_S^2. In all tests we used a 5% significance level.

Data

The LD structure in genic and non-genic regions was investigated using data from three different species: *Arabidopsis thaliana*, *Homo sapiens* and *Gallus gallus domesticus* (a summary for all three data sets is given in Table 4.1).

Arabidopsis thaliana

We used an *A. thaliana* data set published by Atwell et al. (2010). Data consisted of 199 unique accessions, fully homozygous inbred lines, which had been genotyped using the Affymetrix 250 K SNP-tiling array (AtSNPtile1), and was downloaded from https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb. We removed 14 SNPs with missing genotype rate greater than or equal to 0.01 and 170 SNPs with MAF less than 0.01. All individuals passed quality control and the missing genotypes rate per individual was less than 0.0001 leaving 215,947 SNPs for downstream analysis.

Gene annotations were drawn from http://plants.ensembl.org version ‘Ensembl plant genes 21’ (Kersey, 2014), based on the current Arabidopsis Information Resource (TAIR) 2009-10-TAIR 10 assembly (http://www.arabidopsis.org). Only genes annotated from chromosome 1 to 5 were used, resulting in a total of 33,323 genes. All overlapping genes were merged to single gene regions. We selected for the analysis those genes that had at least 10 SNPs; in all 3,721 gene regions were considered.

Human (Homo sapiens)

The genotypes used for the data analysis in humans were taken from the Gene-Environment Association Studies (GENEVA, Cornelis et al., 2010, www.genevastudy.org). We used a subset of GENEVA consisting of data from the Nurses’ Health Study and the Health Professionals’ Follow-up Study. Samples had been genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 with about 780 K SNPs. The data set contained genotypic records of 5,961 individuals.

We removed all markers with a proportion of missing genotypes per SNP greater than or equal to 0.01 and all individuals with a proportion of missing genotypes per individual greater than or equal to 0.05. Furthermore, on the basis of available pedigree information, we also removed all nominally related individuals and individuals with a Hispanic genomic background so that only unrelated individuals of Caucasian origin remained in the data set. We also set a lower threshold of 0.01 for MAF. After quality control of genomic data sample size of 5,827 individuals genotyped at 684,990 SNPs loci remained.

We used gene annotations from http://ensembl.org version ‘Ensembl genes 74’ (Flicek et al., 2013). Only genes annotated from chromosome 1 to 22 were used, which resulted in a total of 54,849 genes that comprised 20,364 coding genes, 20,070 non-coding genes and 14,415 pseudogenes. After merging overlapping genes and dropping out all genic regions with less than 10 SNPs, 7,180 genic regions were retained for further analysis.

Chicken (Gallus gallus domesticus)

We used 673 individuals of a highly selected White Leghorn chicken line from a Synbreed (www.synbreed.tum.de) data set. Samples had been genotyped using the Affymetrix Axiom® Genome-Wide Chicken Genotyping Array (Kranis et al., 2013) with about 600 K SNPs. None of the individuals showed a missing genotype rate greater than or equal to 0.05, while SNPs with missing genotype rate greater than or equal to 0.01 and MAF less than0.01 were removed. After quality control a sample of size 673 individuals and 277,522 SNPs remained. We used gene annotations from http://ensembl.org version ‘Ensembl genes 74’ (Flicek et al., 2013). 17,108 genes annotated from chromosome 1 to 28 (except chromosomes 16 and 24), were used. The SNP coverage of chromosomes 16, 24 and all small chromosomes greater than 28 was not sufficient for the analysis. Upon merging the overlapping genes and removing genic regions with less than 10 SNPs, we were left with 3,033 genic regions for the analysis.

Density of markers, expressed as the number of SNPs per physical distance unit, varied across species: in *A. thaliana* the SNP density was around 3.0 – 3.6 SNPs per kilo base pair (SNPs/kbp), while in *H. sapiens* 0.20 – 0.36 SNPs/kbp were available. In *G. g. domesticus* the density of markers varied across chromosomes: for chromosomes 1 to 8 the marker density was very similar to the one in the human data set, while on chromosomes 9 to 28 the density of SNPs was about 0.4 – 1.0 SNPs/kbp. For all data sets, additional information about the distribution of allele frequencies, marker densities in genic and non-genic regions is available in supplementary Fig. S4.2-S4.7.

Table 4.1. Summary of data sets used across all species

Species	Sample size	No. of chromo-somes studied	No. of genes annota-ted	No. of genic regions studied	No. of SNPs total	No. of SNPs genic	No. of SNPs non-genic
A. thaliana	199	5	33,323	3,721	215,947	135,768	80,179
H. sapiens	5,961	22	54,849	7,180	684,990	391,576	293,414
G. g. domesticus	673	26	17,108	3,033	277,522	146,963	130,559

Data Analysis

We used the framework described above to compare LD levels in genic and non-genic regions in the human, chicken, and Arabidopsis genome. In addition, as a control, the comparison between two similar non-genic regions was performed. Imputing of missing genotypes as well as haplotype-phasing was performed using the BEAGLE software (version 3.3.2; Browning and Browning, 2009).

Before starting the analysis, some data editing was necessary: overlapping genes were observed in all species, meaning that a gene was either lying completely within another gene or two genes overlapped partially. All overlapping genes were merged to one ’genic region‘, since overlapping genes are inherited together with high probability (Normark et al., 1983; Krakauer, 2000).

All markers in-between these genic regions were assigned to non-genic regions. For each genic region *G* we selected one most similar non-genic region *IG*, using the procedure described above. In an independent procedure we chose another *IG* set, termed *IG'*, as a control, which is most similar to the *IG* but does not overlap with *IG*. In general, we searched for the best matching *IG* and *IG'* on the same chromosome as *G*. Due to the small size of chromosomes in *G. g. domesticus* from chromosome 6 onwards, we joined these chromosomes to a single chromosomal region and searched for the best matching *IG* and *IG'* in this chromosomal region.

We applied a two-sided Wilcoxon signed rank test with the null hypotheses $H_0 : \Delta_{\mathbf{G/IG}} = 0$ or $H_0 : \Delta_{\mathbf{IG/IG'}} = 0$ versus alternatives $H_1 : \Delta_{\mathbf{G/IG}} \neq 0$ and $H_1 : \Delta_{\mathbf{IG/IG'}} \neq 0$, where $\Delta_{\mathbf{G/IG}}$ refers to median differences in *G/IG* pairs and $\Delta_{\mathbf{IG/IG'}}$ described median differences in *IG/IG'* pairs. Tests are performed using chromosome- or genome-wide sets of *G, IG* and *IG'*.

Depending on the region of the genome we looked at, we expected genic and non-genic regions to differ not only in the extent of LD, but also in the haplotype frequencies. We used the haplotype diversity H to describe the variation in haplotype frequencies in a region, which is defined as (Nei and Tajima, 1981):

$$H = \frac{m}{m-1}\left(1 - \sum_{i=1}^{2^m} f_i^{\,2}\right) \in [\ 0,\ 1\ ],$$

where m is the number of SNPs in the considered region (*G, IG* or *IG'*) and f_i is the (relative) haplotype frequency of the *ith* haplotype out of the 2^m possible haplotypes. The relative haplotype frequency $f_i = \frac{n_i}{N}$ describes the proportion of the *i*th haplotype in all existing haplotypes in the considered genomic region,

We applied a two-sided Wilcoxon signed rank test with the null hypotheses $H_0 : \delta_{\mathbf{G/IG}} = 0$ and $H_0 : \delta_{\mathbf{IG/IG'}} = 0$ versus alternatives $H_1 : \delta_{\mathbf{G/IG}} \neq 0$ and $H_1 : \delta_{\mathbf{IG/IG'}} \neq 0$ for the haplotype diversities in *G/IG* and *IG/IG'* comparisons. The parameters $\delta_{\mathbf{G/IG}}$ and $\delta_{\mathbf{IG/IG'}}$ refer to median differences in haplotype diversity in *G/IG* and *IG/IG'* pairs, respectively.

The identification procedure for *G/IG* and *IG/IG'* pairs as well as all statistical analyses were implemented in R (R Core Team, 2014). The smoothing curves of pair-wise measures, based on natural cubic splines, was prepared using R-package ggplot2 (Wickham and Chang, 2013).

Results

A first comparison of the amount of the LD in genic and non-genic regions was done based on smoothed curves of r^2 against the physical distance. Here we considered SNPs comprising 99% of all SNP pairs, excluding the upper 1% of SNP pairs with large distances. At distances > 7 kbp in *A. thaliana* and distances > 400 kbp in *H. sapiens* and *G. g. domesticus*, only a few pairs of SNPs existed (see Fig. S4.8) and therefore were excluded from the analysis. A kernel smoothing of pair-wise r^2 and r_S^2 measures is displayed in Fig. **4.**2.

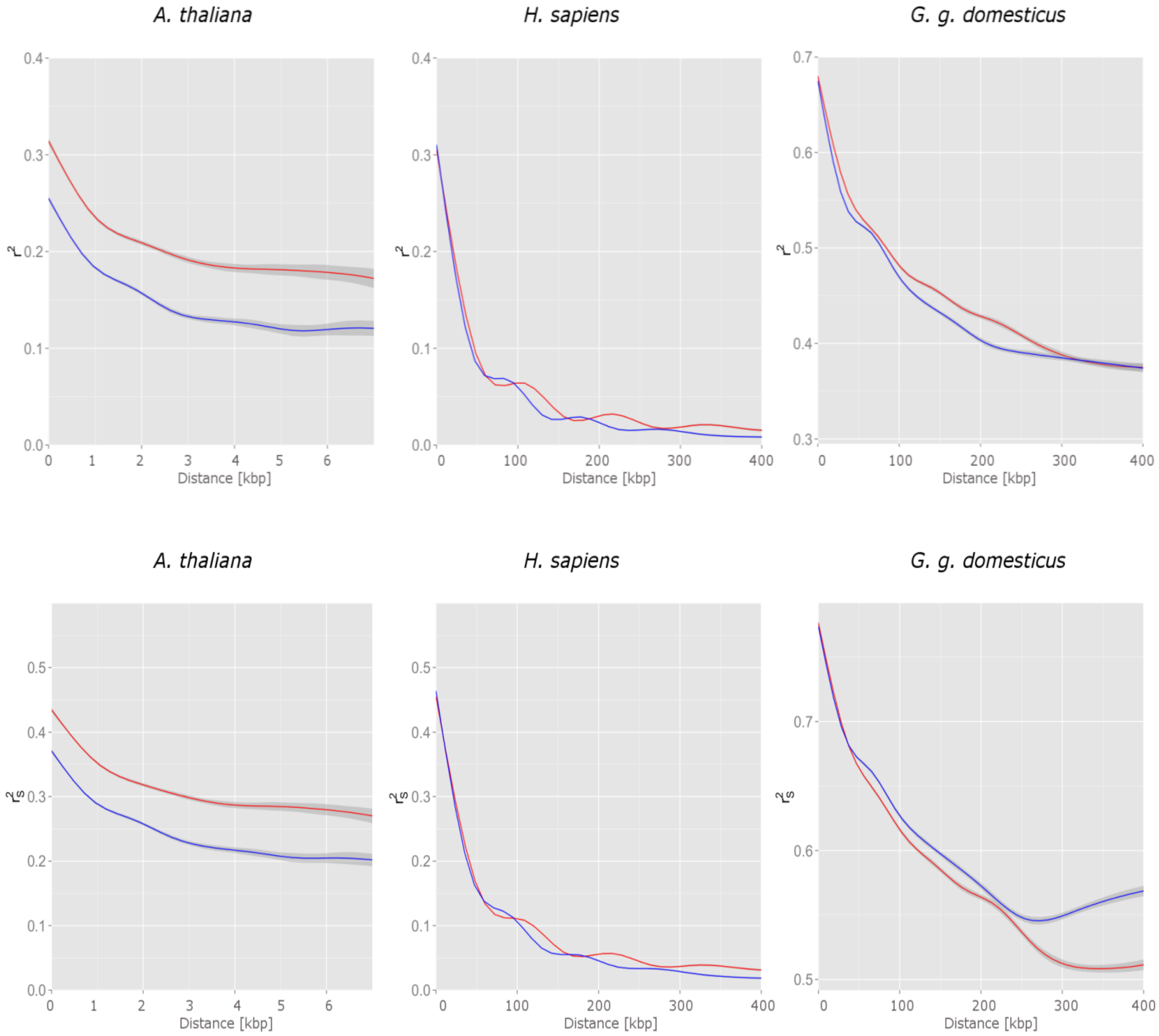


Figure 4.2. Smoothed curves of squared correlation coefficients r^2 (upper panel) and r_S^2 (lower panel), calculated for SNP pairs in genic regions (red lines) versus matching non-genic regions (blue lines) with confidence regions (shaded gray) in *A. thaliana, H. sapiens* and *G. g. domesticus*, plotted against the physical distance in kilo base pairs.

The amount of LD at very short distances in *A. thaliana* was comparable to that observed in *H. sapiens*, but the decay was much faster in *A. thaliana*: SNPs located more than 7 kbp apart have r^2 measures around 0.12 in non-genic regions and around 0.17 in genic regions, while in *H. sapiens* r^2 at this distance still is about 0.25 in both genic and non-genic regions. As expected, in the commercial chicken line we observed a high amount of LD in general, spanning over wide ranges. Regardless of the absolute levels of r^2, higher levels of LD in genic regions in contrast to non-genic regions were detected across all three species, most clearly in *A. thaliana.*

The much higher average level of LD in the highly selected White Leghorn chicken population compared to the other species is reflected by an asymmetric distribution of pair-wise r^2: the center of mass was shifted to the smaller values in *H. sapiens* and *A. thaliana*, while in *G. g. domesticus* center of mass was located in the area with high values (see Fig. S4.9). Thus we chose the median as an appropriate summary statistic to describe LD in explored genic and non-genic regions and to quantify observed differences. The significance tests for chromosome-wise *G/IG* differences ($LD_G - LD_{IG}$) in medians of r^2 and of r_S^2 yielded coherent results in most cases. Fig. 4.3 shows the averaged percentage differences $\Delta_{\mathbf{G/IG}} = (LD_G - LD_{IG}) / LD_G \cdot 100\%$ with corresponding standard errors, which are plotted against the chromosome numbers for all species (for more details see Tables S4.1 – S4.9).

In *G. g. domesticus* significant median differences in r_S^2 at 7 chromosomes (Fig. 4.3, lower panel) were positive and thus confirmed the assumption of higher LD level in genic compared to non-genic regions. This seems to be in conflict with the observation that over long distances the smoothed curve of pair-wise r_S^2 for non-genic regions is higher than that for genic regions (Fig. 4.2, lower panel). This might be due to the fact that an increased level of LD in genic regions is predominantly found in shorter chromosomes, while in some of the large chromosomes (1, 4) LD in genic regions is less than that in non-genic regions (Fig. 4.3).

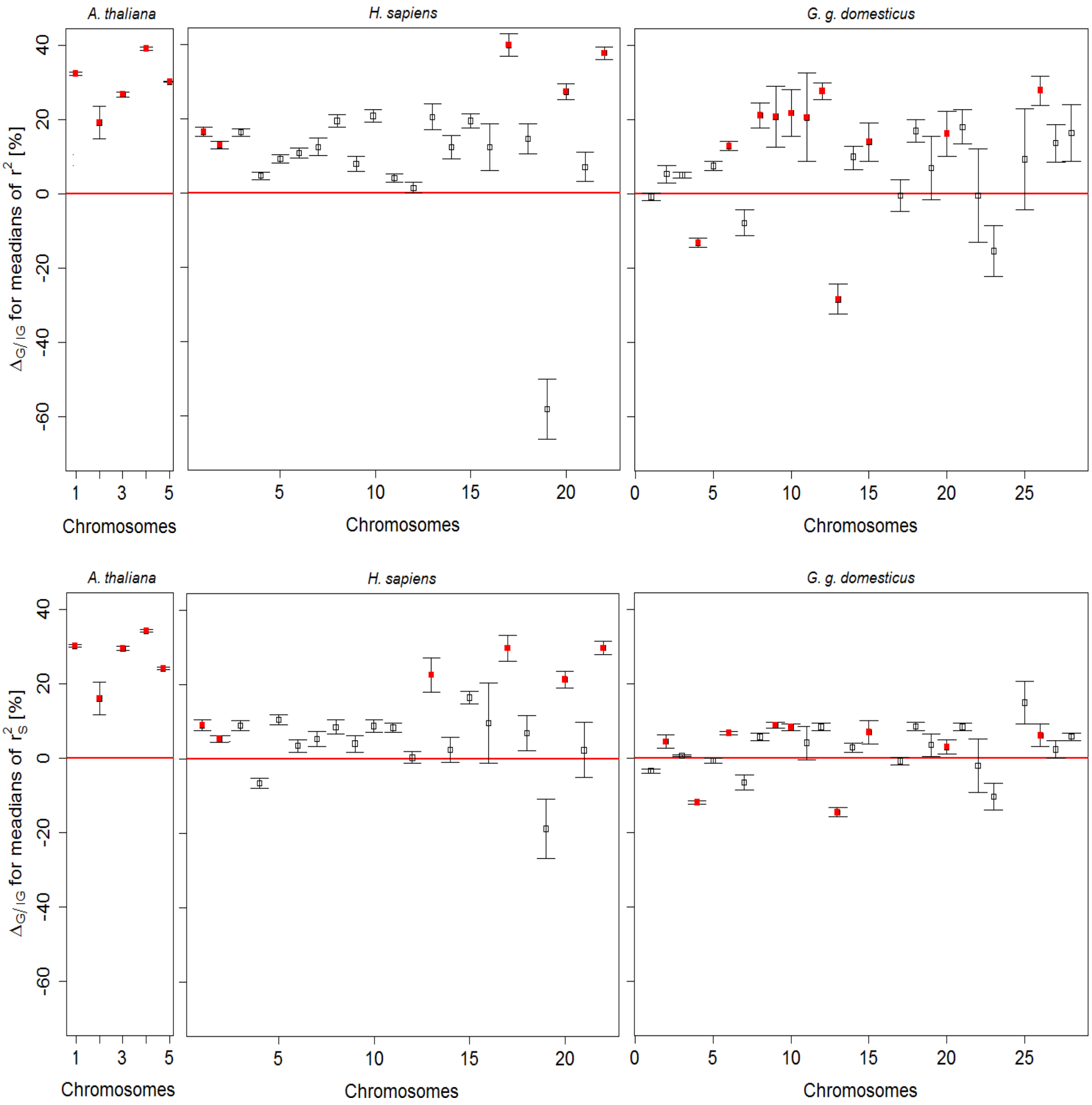


Figure 4.3. Comparison of genic (*G*) versus non-genic (*IG*) regions across chromosomes in *A. thaliana, H. sapiens and G. g. domesticus*. Chromosome-wise averaged percentage differences $\Delta_{G/IG} \pm se$ between medians of r^2 in *G* and medians in *IG* (upper panel) and chromosome-wise averaged differences $\Delta_{G/IG} \pm se$ between r_S^2 in *G* and in *IG* (lower panel), where se refer to standard errors of averages. Red filled symbols indicate significant differences in *G/IG* comparison.

When fitting a linear regression within species, the coefficient of determination between averages per chromosome calculated for r^2 and chromosome-wide averages calculated for r_S^2 was high for all species: 0.75 in *H. sapiens*, 0.78 in *G. g. domesticus* and 0.79 in *A. thaliana*.

So, decisions of Wilcoxon signed rank test based on the LD measure r^2 corresponded to the test decisions made for differences in a MAF independent measure r_S^2. This consistency in test results has led to the conclusion that our framework was efficient in adjusting for spatial and for MAF influences.

In case of genome-wide comparison of medians of r^2 about 31% more LD was detected in genic regions than in non-genic regions in *A. thaliana*, followed by 13.6% in *H. sapiens* and 6 % in *G. g. domesticus*. The comparisons of $\Delta_{\mathrm{IG/IG'}}$ between matching non-genic regions *IG* and *IG'* yielded no significant differences for r^2 but for r_S^2 a significant difference was found for one chromosome in *A. thaliana* and *G. g. domesticus*, respectively, which is in the expected range under the null hypothesis (Tables S4.1 – S4.9). The outcomes of chromosome-wise and genome-wide comparisons are summarized in Table 4.2.

Table 4.2. Number of chromosomes with significantly (p-value <0.05) increased LD level in the comparison of genic with matching non-genic regions ($\Delta_{\mathrm{G/IG}}$), number of chromosomes with significantly different LD levels for matching non-genic regions ($\Delta_{\mathrm{IG/IG'}}$), and the genome wide average difference in LD between genic and matching non genic regions in per cent ($\Delta_{\mathrm{G/IG}}$ [%]) for the two LD measures r^2 and r_S^2. Asterisks indicate the level of significance for the genome-wide differences.

	Chromosomes studied					Genome-wide	
	Total	$\Delta_{\mathrm{G/IG}}$		$\Delta_{\mathrm{IG/IG'}}$		$\Delta_{\mathrm{G/IG}}$ [%]	
Species		r^2	r_S^2	r^2	r_S^2	r^2	r_S^2
A. thaliana	5	5	5	0	1	31.2***	27.7***
H. sapiens	22	5	5	0	0	13.6*	8.0**
G. g. domesticus	26	10	9	0	1	6.0**	0.5

*: p-value <0.05 **: p-value <0.01 ***: p-value <0.001

We expected a higher LD in genic regions compared to non-genic regions and performed 53 chromosome-wide significance tests in total (Fig. 4.3), 18 chromosomes (34%) showed a significantly higher LD in genic regions. In two chromosomes (chromosome 4 and 13 in chicken) significantly higher LD in non-genic regions was observed. This corresponds to 3,8% of all comparisons and is below the 5% significance level. Thus the unexpected results for chromosomes 4 and 13 might be the false positive test outcomes obtained just by chance.

The Wilcoxon signed rank test, applied chromosome-wise, detected significant differences between genic and non-genic regions on all 5 chromosomes of *A. thaliana*, on about 1/4 of the human chromosomes and on about 40 per cent of the chicken chromosomes. In Fig. 4.4 chromosome-wise percentage differences in haplotype diversities $\Delta H_{G/IG} = (H_G - H_{IG}) / H_G \cdot 100\%$ for the three species are presented.

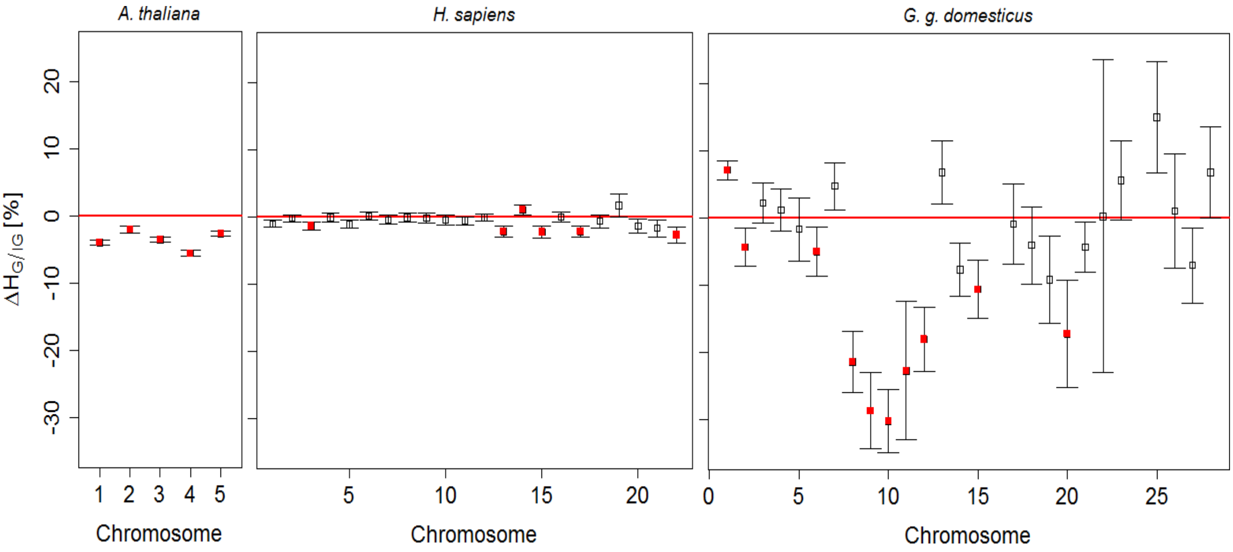


Figure 4.4. Chromosome-wise differences in haplotype diversity in *G/IG* comparisons, across species. Chromosome-wise haplotype diversity percentage differences $\Delta H_{G/IG} \pm se$ plotted against the chromosome number, where se refers to standard errors of averages. Red filled symbols indicate significant (p-value <0.05) differences in *G/IG* comparison.

The haplotype diversity in *A. thaliana* and *H. sapiens* were both relatively high, at a comparable level: chromosome-wide averages ranged between 0.85 and 0.89 in genic regions, accompanied by significantly lower haplotype diversity in *G* compared to *IG* (see Fig. S4.10 and Tables S4.10-S4.12). In *A. thaliana* we observed $\Delta H_{G/IG} = -3.5\%$ less diversity in haplotypes at the genome-wide level, while the loss of haplotype diversity in G varied between -2% and -5% at the chromosome level. In *H. sapiens*, a small significant loss $\Delta H_{G/IC} = -0.7\%$ was observed at the genome-wide level, whereas significant $\Delta H_{G/IC}$ varied between -0.7% and -2.6% at the chromosome level. In *G. g. domesticus*, haplotype diversity of -2.9% at the genome-wide level was significant, albeit smaller than that in *A. thaliana*, whereas the chromosome-wide averages in genic regions ranged between 0.40 and 0.61 and the significant $\Delta H_{G/IC}$ between -4.3% and -23.2% at the chromosome level was the largest of all three species

Discussion

Apart from the proportion of protein-coding DNA in the genome, the major question is whether the changes over generations are differently occurring in different genomic regions. We introduced a general comparison framework, which copes with difficulties arising while performing comparison of LD levels between different genomic regions, such as the impact of the extent of compared regions on the genome (spatial bias) and the impact of allele frequencies on LD (MAF caused bias). The retrieved knowledge about variation in genomic regions of interests could be used, for example, to estimate a measure for likelihood of fitness consequences of involved populations proposed by Gulko et al. (2014).

Impact of location of a region: genic versus non-genic regions

The results obtained for *A. thaliana* were in contrast to those obtained by Kim et al. (2007), who suggested that LD hot spots in arabidopsis are situated preferentially outside genic regions. On a genome-wide level, significantly more LD in genic regions was observed in all three species and thus the observation by Eberle et al. (2006) for the human genome was confirmed and quantified. The LD levels in genic regions at very short physical distances are similar in *A. thaliana* and *H. sapiens* with r^2 being about 0.3 on average (see Fig. 4.2). In *A. thaliana* a clear gap between LD amount in genic and non-genic regions is seen while in *H. sapiens* almost no G/IG difference is recorded up to a distance of about 50 kilo base pairs, while in maize, which is in contrast to *A. thaliana* an outcrossing plant, or in self-pollinating barley a comparable decay of LD (up to 3 kbp) was observed by Caldwell et al. (2006).

LD spans are so short and genic regions are more conserved in *A. thaliana* compared to humans presumably is due to the fact that *A. thaliana* is an ubiquitous plant and the sample used in our studies reflects a very large effective population size (N_e) that may explain the rapid decay of LD. Contemporary estimates of N_e of *A. thaliana*, based on sequence data of 80 strains from a wide Eurasian region indicated N_e to lie between 250,000 and 300,000 (Cao et al., 2011). The LD level observed in *G. g. domesticus* is twice as high as the LD level in *H. sapiens* and LD decays much slower than in humans. This higher LD level is observed in *G. g. domesticus* over all distances. The white layer data used originate from a commercial line, which has been intensively selected for egg laying in a closed nucleus breeding scheme. Thus the degree of relatedness among the individuals in the studied sample is relatively high: average pedigree based relatedness was 0.255 ± 0.07 and the average inbreeding coefficient was 0.10 ± 0.025. The magnitude of relatedness in the population has a strong impact on the effective population size, which is very low in commercial lines of chicken (Caldwell et al., 2006; Chao et al., 2011). For pair-wise distances ≤ 25 kbp, Qanbari at al.

(2010) reported values of r^2 between 0.60 and 0.74 in four different layer lines, which is concordant with the magnitude of LD detected in our study. Also the decay of LD observed in the white layer data set ($r^2 \approx 0.37$ for pairs of SNPs in about 400 kbp distance) was consistent with results from previous studies ($r^2 = 0.35$ for pairs of SNPs in about 200 - 500 kbp distance (Qanbari et al., 2010)). Layer breeding schemes use a small number of highly selected male individuals in each generation.

A similar monopolization of reproductive function by one or few individuals is also given in eusocial insects (like e.g. ants) causing reduced effective population size and a high degree of conservation in coding genomic regions (Romiguier et al., 2014).

Many statistical methods have been developed in the last decade to utilize high-throughput sequencing data for estimating population parameters (e.g. Quanbari et al., 2010; Li et al., 2012), among them a maximum-likelihood estimator of recombination rates based on LD patterns (Johnson and Slatkin, 2009). Thus, stronger association observed between markers in genic regions than in non-genic regions might go along with a higher recombination rate in non-genic regions. Accordingly, a lower diversity of haplotypes is expected in genic regions compared to non-genic regions. Indeed significantly less diversity of haplotypes in genic regions was noticed for all species, which confirms our results obtained for LD.

Genic regions in general appear to be more conserved than non-genic regions (e.g. Eberle et al., 2006; Nachman and Payseur, 2012; Lohmueller et al., 2011). Higher haplotype diversity in non-genic regions may be explained by the fact that recombination in these regions may affect biological cycles or pathways to a lesser extent; thus most haplotypes resulting from recombination will be neutral with respect to fitness and will not be under selection. In contrast, recombination in genic regions may affect the biological function of the respective haplotype and consequently such haplotypes with reduced fitness will be less frequently found among the progeny, resulting in a reduced haplotype diversity in genic regions. Regions with low recombination were found to contain highly conserved genes with essential cellular functions (e.g. Hussin et al., 2015). Furthermore, hitchhiking and background selection might generate a strong link between genetic diversity and recombination rate (Smith and Haigh, 2007; Gillespie, 1991; Lohmueller et al., 2011). Thus, the intensive anthropogenic selection in white layers may explain the pronounced differences between haplotype diversity in genic and non-genic regions in the white layer data.

Impact of chromosome size or size of region on LD magnitude

The suggested approach accounting for spatial and structural differences in genomic regions when comparing genic and non-genic regions provides new insights into the dependency of

LD levels on the size of chromosomes or regions. Assuming that the number of recombination events per chromosome is approximately equal, differences in recombination rates on chromosomes of different physical length are supposed (Kong et al., 2002; Smith et al., 2005; Johnson and Slatkin, 2009) with a slower decay of LD in the larger chromosomes. In contrast to the findings of Smith et al. (2005) and Uimari et al. (2005) for the human genome and Hillier et al. (2004) and Groenen et al. (2009) for the chicken genome, we do not observe weaker LD in the smaller chromosomes and stronger LD in the large chromosomes (see Fig. S4.11 and Table S4.13). Even though the chromosome-wise averaged medians scattered more in *G. g. domesticus,* there was no clear association between the size of chromosomes and the level of LD. Considering the size of genic and non-genic regions across chromosomes, a weak but significant negative association between the size and the LD of a region was detected in all species. For instance, in *G. g. domesticus* larger regions showed a slightly lower r^2 (the slope of a fitted linear regression ≈ -0.002) and also slightly lower r_S^2 (the slope of a fitted linear regression ≈ -0.001, see Fig. S4.12). This size bias is expected since physically large genic regions have more pairs of physically distant SNPs, which in turn have a lower LD (see Fig. 4.2). There was no significant size bias for the differences in medians of r^2 and of r_S^2 since we corrected for the effect of the length of the region through comparing with a region of similar size. This is exemplarily visualized for *G. g. domesticus* in Fig. S4.13.

Across all species the extent of LD measured in genic or non-genic regions did not depend on the size of the chromosome (see Table S4.13). Discrepancies between our results and results reported by Smith et al (2005) and Uimari et al. (2005) may have resulted either from the lower marker density, lower SNP call rates and smaller sample sizes in these older studies or due to bias caused by spatial differences or different distribution of allele frequencies.

Conclusions

Our study has shown that across the three considered species, the average level of LD is systematically higher in genic regions than in non-genic regions, confirming and quantifying the more qualitative result in the human genome of Eberle et al. (2006) for a wider range of species. This observed difference is not affected by other factors which might systematically differ between genic and non-genic regions, such as minor allele frequencies or SNP densities, since such differences were removed by comparing candidate sets with best matching counterparts. With this approach, it was also possible to exactly quantify the relative excess of LD on a chromosome-wise or genome-wide level. It was shown that the amount of excess LD in genic regions differs between species (with *A. thaliana* > *H. sapiens* > *G. g. domesticus*) and varies substantially between the chromosomes within the considered species.

These observations found for the widely used LD-measure r^2 in tendency were confirmed with the standardized LD-measure r_S^2 and with haplotype diversity. Based on our findings we suggest that the excess of LD in genic region is a general phenomenon resulting from evolutionary forces, since the patterns of genetic polymorphisms reflects evolutionary processes like recombination, genetic drift and selection.

The suggested approach can be varied by replacing the squared correlation r^2 by any other LD measure (e.g. D' (Lewontin, 1964), homozygosity of haplotypes (Agapow and Burt, 2001), normalized entropy difference (Zhao et al., 2005) or Kullback-Leibler distance (Gianola et al., 2012)), by accounting for more or different scaling factors or by varying the similarity score by using different weighting of those factors. The comparative assessment of the LD level in genic and non-genic regions might be used as a starting point for a more differentiated analysis of the LD structure in the genome. In our studies we applied just two categories of genomic regions: genic and non-genic regions, where genic regions were defined in accordance with annotations of known genes in Ensembl gene databases. This way of proceeding is coherent to the classification of genic regions used by Eberle et al. (2006) and provides us better comparability to their results. A promising area for improvement of our current approach is the extension of considered genetic regions by a stratification in e. g exons, introns, 5k upstream or downstream regions, 5' and 3' UTRs etc. Such analyses might require higher marker densities (up to sequence level) and considerably enlarged sample sizes, though. An especially interesting subject for further research is the contribution of purifying and positive selection across breeding populations to differences in level of LD between coding and non-coding regions of the genes. The framework described here enables comparison of LD structure in arbitrary species and any genomic regions of interests.

Acknowledgments

The authors wish to thank Heike Bickeböller for useful comments provided. Parts of the analysis were done during research stays of S. Berger at the Department of biostatistics, University of Alabama at Birmingham in Birmingham, US. This study was financially supported by RTG 1644 'Scaling problem in statistics', financed by German Research Foundation (DFG), chicken genotypes were generated in the AgroClustEr "Synbreed – Synergistic plant and animal breeding" (Funding ID: 0315528C) funded by the German Federal Ministry of Education and Research.

Competing Interests Statement. The authors have declared that no competing interests exist. Prof. Dr. Preisinger is the head of Genetic department of LOHMANN Tierzucht GMBH.

This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

Appendix

Upper Limits for Squared Correlation

Derivation of boundaries for gametic frequency. For known allele frequencies π_1 at locus 1 and π_2 at locus 2 and gametic frequency π_{12}, the Pearson's correlation coefficient is obtained by

$$r = \frac{\pi_{12} - \pi_1\pi_2}{\sqrt{\pi_1(1-\pi_1)\pi_2(1-\pi_2)}} = \frac{D}{\sqrt{\pi_1(1-\pi_1)\pi_2(1-\pi_2)}} \tag{1}$$

We consider two cases, according to the value of the numerator D:

1. Disequilibrium coefficient D is positive (denoted as D_{pos}). Consequently, a positive correlation coefficient $r = \frac{D_{pos}}{\sqrt{\pi_1(1-\pi_1)\pi_2(1-\pi_2)}} > 0$ is yielded and r becomes a maximum for the largest possible value of D_{pos}.
2. Disequilibrium coefficient D is negative (denoted as D_{neg}), this yields a negative correlation $r = \frac{D_{neg}}{\sqrt{\pi_1(1-\pi_1)\pi_2(1-\pi_2)}} < 0$, which is a minimum for smallest possible value of D_{neg}.

For given allele frequencies only the value of gametic frequency π_{12} is variable and influences the value of disequilibrium coefficient D. VanLiere and Rosenberg (2008) investigated the maximum possible of r^2 for a given pair of allele frequencies. In the following, we extend results presented by VanLiere and Rosenberg (2008) in order to obtain a general derivation of boundaries for gametic frequency π_{12}:

The largest possible value of D_{pos} and the smallest possible value of D_{neg} could be obtained by the application of Fréchet-Hoeffding bounds on the joint probability π_{12}. For this reason, some measure theoretical ideas will be presented next.

At first we define the Fréchet-Hoeffding bounds for a general case (proof is given in chapter 3.6, (Rachev and Rüschendorf, 1998)).

Theorem. *For a probability space* $(\Omega, \mathcal{F}, P)$, *where* Ω *is a non-empty sample space,* $\mathcal{F}$ *is a* σ-*algebra of subsets* $A_i \in \Omega$ *and* P *is a probability measure on* $\mathcal{F}$, *Fréchet-Hoeffding bounds are defined as*

$$max\left(0, \sum_{i=1}^{n} P_i - n + 1\right) \leq P(A_1, \dots, A_n) \leq \min(P_1, \dots, P_n) \tag{2}$$

for subsets $A_1, \dots, A_n$ *and their probabilities* $P_i = P(A_i),\ i = 1 \dots, n$

In order to apply the Fréchet-Hoeffding bounds, we treated the manifestation of different alleles at one locus as a realization of a Bernoulli random variable and we defined the appearance of one of the alleles as a success. For two loci, we have two Bernoulli distributed random variables $X_1 \sim B(\pi_1)$ and $X_2 \sim B(\pi_2)$ with success probabilities $\pi_1 = P(X_1 = 1)$ and $\pi_2 = P(X_2 = 1)$ with $0 < \pi_1, \pi_2 < 1$. Then the general form of Fréchet-Hoeffding bounds (2) applied to a two-dimensional case became

$$max(0, \pi_1 + \pi_2 - 1) \leq \pi_{12} \leq \min(\pi_1, \pi_2),$$

representing lower and upper limits for the joint distribution $\pi_{12} = P(X_1 = 1, X_2 = 1)$. Now upper and lower limits for the gametic frequency π_{12} could be used to build upper bounds for the squared correlation r^2.

Calculation of upper limits for $\mathbf{r^2}$. For all possible combinations of allele frequencies π_1 and π_2, r^2 reaches its maximum if the numerator, i.e. the squared disequilibrium coefficient D^2, is a maximum. Using D^2_{neg} as lower limit for π_{12} and D^2_{pos} as upper limit for π_{12}, the highest possible value of squared disequilibrium coefficient $D^2_{max} = \max(D^2_{neg}, D^2_{pos})$ is yielded. Thus, an upper limit for the squared correlation $r^2_{max}(D^2_{max})$ is obtained at D^2_{max}.

Two-dimensional space of success probabilities π_1 and π_2 could be divided into eight sections (see Figure S4.13), according to relation of probabilities π_1 and π_2 to each other. For each section we derived squared disequilibrium coefficient D^2_{pos} and D^2_{neg} using limiting conditions, which are pre-defined by the values of allele frequencies. By using ty $D^2_{pos} \geq D^2_{neg}$, we examined which one of two – squared positive disequilibrium coefficient D^2_{pos} or squared negative disequilibrium coefficient D^2_{neg} - is greater and achieved expressions for upper limit of squared correlation $r^2_{max}(\pi_1,\ \pi_2)$ (see Table S4.14). These calculations confirmed results reported by VanLiere and Rosenberg (2008).

As mentioned previously, in this study we use only minor allele frequencies, which take values less than 0.5. For this reason only the results from section 1 or 2 are relevant here. Without limiting the generality of foregoing, we will use the expression achieved in section 1, where $\pi_1 \leq \pi_2 \leq 0.5$ are the minor allele frequencies and are denoted as p_1 and p_2. Thus the upper limit for squared correlation is given by

$$r^2_{max}(p_1, p_2) = \frac{p_1\,(1 - p_2)}{p_2\,(1 - p_1)} \tag{3}$$

This expression is also known as odds-ratio and is used e.g. in epidemiological or in case-control studies in human medicine.

The upper limit could be rewritten by using the difference between the minor allele frequencies $\delta = p_2 - p_1 \geq 0$. Then, the upper limit can be rewritten as

$$r^2_{max}(\delta, p_2) = 1 - \frac{\delta}{p_2\,(\delta + 1 - p_2)} \tag{4}$$

In Figure A4.1 some examples of upper limit for a set of fixed values of p_2 as well as the upper limits for all combinations of p_1 and p_2 are shown.

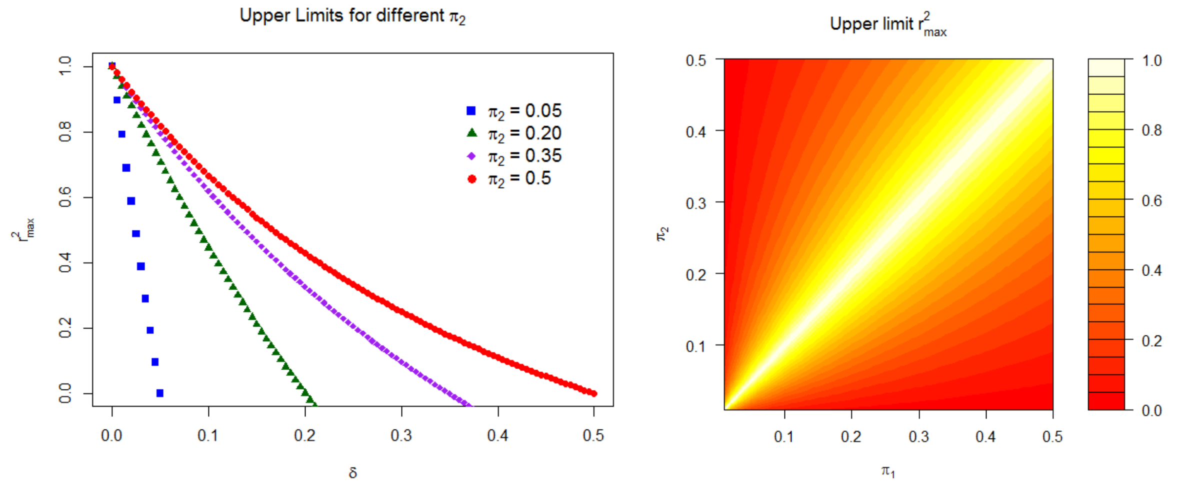


Figure A4.1. Upper limits for squared correlation. Maximal accessible squared correlation r^2_{max} between two loci against the delta MAF (δ), $p_1 \leq p_2$ are minor allele frequencies and $\delta = p_2 - p_1$ for fixed p_2 (left) and for all combinations of p_1 and p_2.

A more general result is achieved by using our knowledge about the range of minor allele frequencies: the absolute upper limit, depending only on the difference between the MAFs, is obtained by using the upper limit for MAFs $p_1 \leq p_2 \leq 0.5$:

$$r^2_{max}(\delta, p_2) = 1 - \frac{\delta}{p_2\,(\delta+1-p_2)} \leq r^2_{max}(\delta) = 1 - \frac{4\delta}{2\delta+1}$$

for all possible values of p_1 and p_2 .

Thus, a general upper limit for r^2, depending only on the differences in MAF, is given by

$$r^2_{max}(\delta) = 1 - \frac{4\delta}{2\delta + 1} \tag{5}$$

References

Agapow, P.-M., and Burt, A. (2001). Indices of multilocus linkage disequilibrium. Mol. Ecol. Notes *1*, 101–102.

Ardlie, K.G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. *3*, 299–309.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., and Hu, T.T. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature *465*, 627–631.

Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84*, 210–223.

Caldwell, K.S., Russell, J., Langridge, P., and Powell, W. (2006). Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, Hordeum vulgare. Genetics *172*, 557–567.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., et al. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. *43*, 956–963.

La Chapelle, A. De, and Wright, F.A. (1998). Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc. Natl. Acad. Sci. *95*, 12416–12423.

Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A., and Pritchard, J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. *38*, 1251–1260.

Cornelis, M.C., Agrawal, A., Cole, J.W., Hansel, N.N., Barnes, K.C., Beaty, T.H., Bennett, S.N., Bierut, L.J., Boerwinkle, E., Doheny, K.F., et al. (2010). The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet. Epidemiol. *34*, 364–372.

Devlin, B., and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics *29*, 311–322.

Dobrushin, R.L. (1970). Prescribing a system of random variables by conditional distributions. Theory Probab. Its Appl. *15*, 458–486.

Eberle, M.A., Rieder, M.J., Kruglyak, L., and Nickerson, D.A. (2006). Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet. *2*, e142.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. Nucleic Acids Res. gkt1196.

Fréchet, M. (1960). Sur les tableaux dont les marges et des bornes sont données. Rev. Inst. Int. Stat. 10–32.

Garner, C., and Slatkin, M. (2003). On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. Genet. Epidemiol. *24*, 57–67.

Gianola, D., Manfredi, E., and Simianer, H. (2012). On measures of association among genetic variables. Anim. Genet. *43*, 19–35.

Gillespie, J.H. (1991). The causes of molecular evolution (Oxford University Press).

Groenen, M.A., Wahlberg, P., Foglio, M., Cheng, H.H., Megens, H.-J., Crooijmans, R.P., Besnier, F., Lathrop, M., Muir, W.M., Wong, G.K.-S., et al. (2009). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. *19*, 510–519.

Gulko, B., Gronau, I., Hubisz, M.J., and Siepel, A. (2014). Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. bioRxiv 006825.

Hedrick, P.W. (2005). A standardized genetic differentiation measure. Evolution *59*, 1633–1638.

Hill, W.G. (1981). Estimation of effective population size from data on linkage disequilibrium. Genet. Res. *38*, 209–216.

Hill, W.G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. *38*, 226–231.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695–716.

Hussin, J.G., Hodgkinson, A., Idaghdour, Y., Grenier, J.-C., Goulet, J.-P., Gbeha, E., Hip-Ki, E., and Awadalla, P. (2015). Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat. Genet. *47*, 400–404.

Huttley, G.A., Smith, M.W., Carrington, M., and O'Brien, S.J. (1999). A scan for linkage disequilibrium across the human genome. Genetics *152*, 1711–1722.

Johnson, P.L., and Slatkin, M. (2009). Inference of microbial recombination rates from metagenomic data. PLoS Genet. *5*, e1000674.

Jorde, L.B. (1995). Linkage disequilibrium as a gene-mapping tool. Am. J. Hum. Genet. *56*, 11.

Kersey, P.J. (2014). Ensembl Plants-an Integrative Resource for Plant Genome Data. In Plant and Animal Genome XXII Conference, (Plant and Animal Genome),.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. *39*, 1151–1155.

Kim, S.Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. Genet. Epidemiol. *34*, 479–491.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat. Genet.

Krakauer, D.C. (2000). Stability and evolution of overlapping genes. Evolution *54*, 731–739.

Kranis, A., Gheyas, A.A., Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., Pirani, A., Brew, F., and Kaiser, P. (2013). Development of a high density 600K SNP genotyping array for chicken. BMC Genomics *14*, 59.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. *22*, 139–144.

Laska, E., Meisner, M., Siegel, C., Fischer, S., and Wanderling, J. (1975). Matched-pairs study of reserpine use and breast cancer. The Lancet *306*, 296–300.

Lewontin, R.C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. Genetics *49*, 49.

Lewontin, R.C. (1988). On measures of gametic disequilibrium. Genetics *120*, 849–852.

Li, D.F., Liu, W.B., Liu, J.F., Yi, G.Q., Lian, L., Qu, L.J., Li, J.Y., Xu, G.Y., and Yang, N. (2012). Whole-genome scan for signatures of recent selection reveals loci associated with important traits in White Leghorn chickens. Poult. Sci. *91*, 1804–1812.

Lin, S., Chakravarti, A., and Cutler, D.J. (2004). Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat. Genet. *36*, 1181–1188.

Lohmueller, K.E., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A.F., Grarup, N., et al. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet *7*, e1002326.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science *304*, 581–584.

Meuwissen, T.H., Karlsen, A., Lien, S., Olsaker, I., and Goddard, M.E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics *161*, 373–379.

Meyer-Lindenberg, A., Buckholtz, J.W., Kolachana, B., Hariri, A.R., Pezawas, L., Blasi, G., Wabnitz, A., Honea, R., Verchinski, B., and Callicott, J.H. (2006). Neural mechanisms of genetic risk for impulsivity and violence in humans. Proc. Natl. Acad. Sci. *103*, 6269–6274.

Morton, N.E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y., and Collins, A. (2001). The optimal measure of allelic association. Proc. Natl. Acad. Sci. *98*, 5217–5221.

Mueller, J.C. (2004). Linkage disequilibrium for different scales and applications. Brief. Bioinform. *5*, 355–364.

Mueller, J.C., Lõhmussaar, E., Mägi, R., Remm, M., Bettecken, T., Lichtner, P., Biskup, S., Illig, T., Pfeufer, A., Luedemann, J., et al. (2005). Linkage Disequilibrium Patterns and tagSNP Transferability among European Populations. Am. J. Hum. Genet. *76*, 387–398.

Nachman, M.W., and Payseur, B.A. (2012). Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. B Biol. Sci. *367*, 409–421.

Nei, M., and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. Genetics *97*, 145–163.

Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F.P., and Olsson, O. (1983). Overlapping genes. Annu. Rev. Genet. *17*, 499–525.

Qanbari, S., Hansen, M., Weigend, S., Preisinger, R., and Simianer, H. (2010). Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genet. *11*, 103.

Rachev, S.T., and Rüschendorf, L. (1998). Mass Transportation Problems: Volume I: Theory (Springer).

R Core Team (2014). R: a language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.

Romiguier, J., Lourenco, J., Gayral, P., Faivre, N., Weinert, L.A., Ravel, S., Ballenghien, M., Cahais, V., Bernard, A., Loire, E., et al. (2014). Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. J. Evol. Biol. *27*, 593–603.

Rüschendorf, L. (1981). Sharpness of Fréchet-bounds. Probab. Theory Relat. Fields *57*, 293–302.

Simianer, H., and Erbe, M. (2014). Genetics, genomics, breeding–why scale matters. J. Anim. Breed. Genet. *131*, 83–84.

Smith, J.M., and Haigh, J. (2007). The hitch-hiking effect of a favourable gene. Genet Res *89*, 391–403.

Smith, A.V., Thomas, D.J., Munro, H.M., and Abecasis, G.R. (2005). Sequence features in regions of weak and strong linkage disequilibrium. Genome Res. *15*, 1519–1534.

Smith, E.M., Wang, X., Littrell, J., Eckert, J., Cole, R., Kissebah, A.H., and Olivier, M. (2006). Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. Genomics *88*, 407–414.

Uimari, P., Kontkanen, O., Visscher, P.M., Pirskanen, M., Fuentes, R., and Salonen, J.T. (2005). Genome-wide linkage disequilibrium from 100,000 SNPs in the East Finland founder population. Twin Res. Hum. Genet. *8*, 185–197.

VanLiere, J.M., and Rosenberg, N.A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. Theor. Popul. Biol. *74*, 130–137.

Vaserstein, L.N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. Probl. Peredachi Informatsii *5*, 64–72.

Wickham, H., and Chang, W. (2013). ggplot2: An implementation of the grammar of graphics. R package version 0.9. 3.1.

Wray, N.R. (2005). Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. Twin Res. Hum. Genet. *8*, 87–94.

Xu, Y., Xie, C., Wan, J., He, Z., and Prasanna, B.M. (2013). Marker-assisted selection in cereals: platforms, strategies and examples. In Cereal Genomics II, (Springer), pp. 375–411.

Zhao, H.H., Fernando, R.L., and Dekkers, J.C.M. (2007). Power and Precision of Alternate Methods for Linkage Disequilibrium Mapping of Quantitative Trait Loci. Genetics *175*, 1975–1986.

Zhao, J., Boerwinkle, E., and Xiong, M. (2005). An entropy-based statistic for genomewide association studies. Am. J. Hum. Genet. *77*, 27–40.

Zondervan, K.T., and Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. Nat. Rev. Genet. *5*, 89–100.

Supporting Information

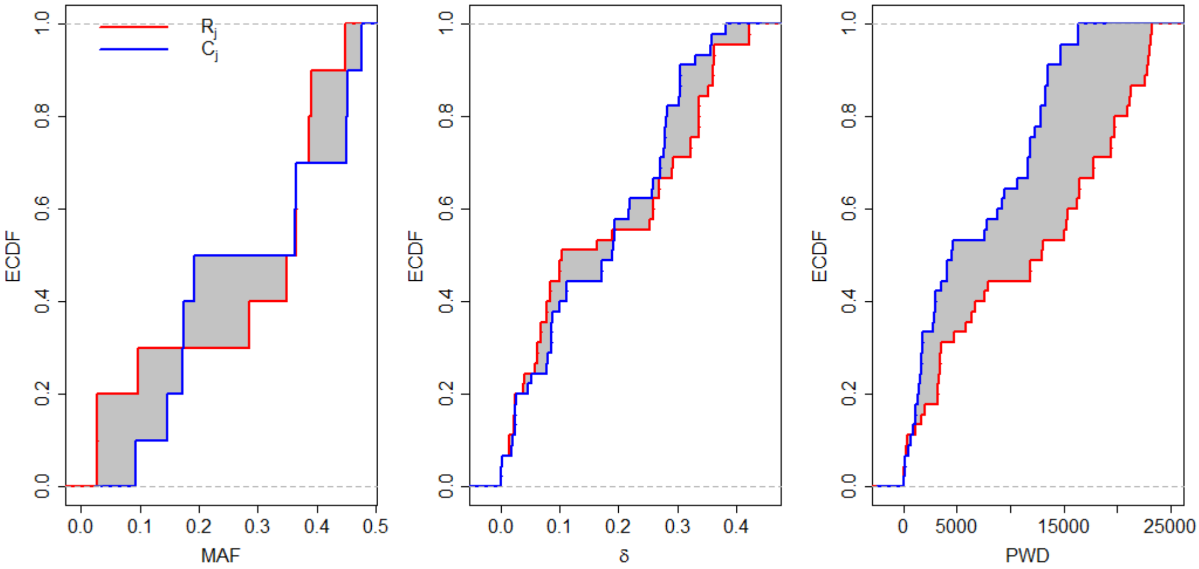


Figure S4.1. Area between the Empirical Cumulative Density Functions. ECDFs for reference set (red) and for a candidate subset (blue), the $A_{MAF}^{(jk)}$ (left), $A_{\delta}^{(jk)}$ (center), and $A_{PWD}^{(jk)}$ (right) are marked in grey.

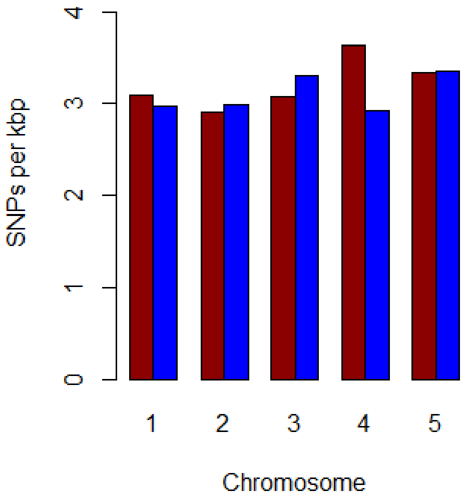


Figure S4.2. SNP-density for chromosomes 1 to 5 in A. thaliana. Red bars stand for density of SNPs in genic regions, blue bars stand for SNP-density in non-genic regions.

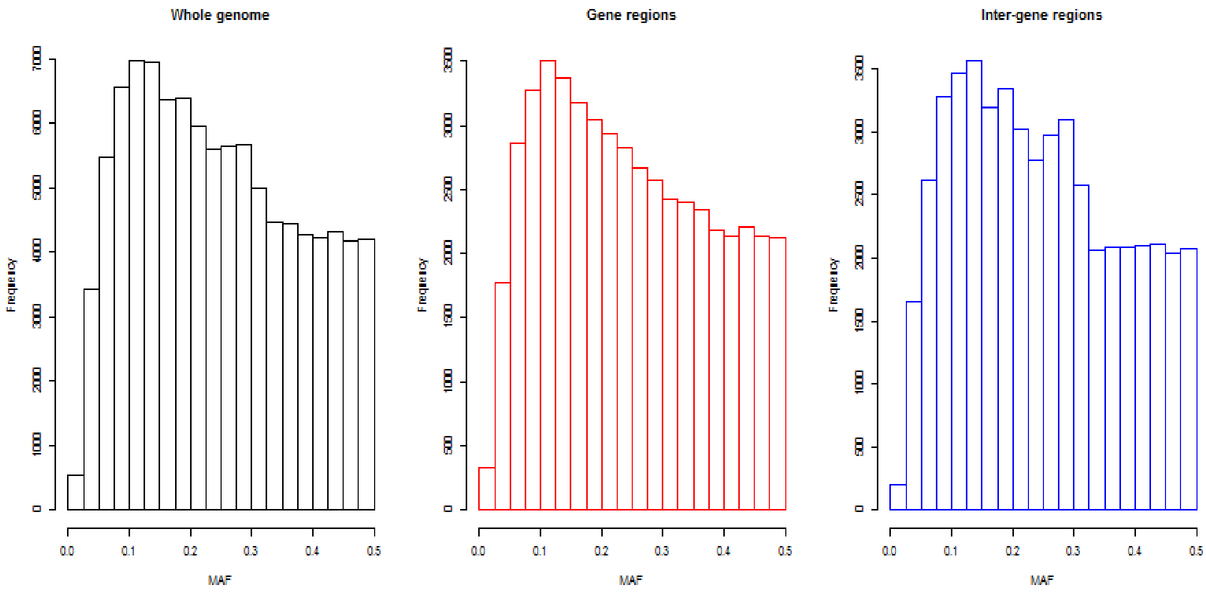


Figure S4.3 Distribution of minor allele frequencies in A. thaliana across the whole genome, in genic and in non-genic regions, respectively.

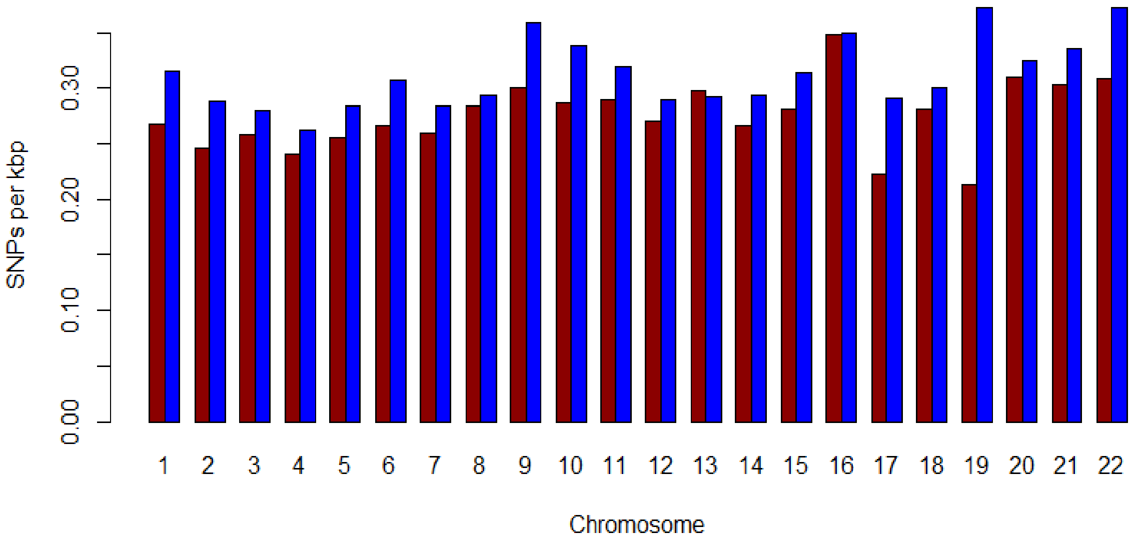


Figure S4.4. SNP-density for chromosomes 1 to 22 in H. sapiens. Red bars stand for density of SNPs in genic regions, blue bars stand for SNP-density in non-genic regions.

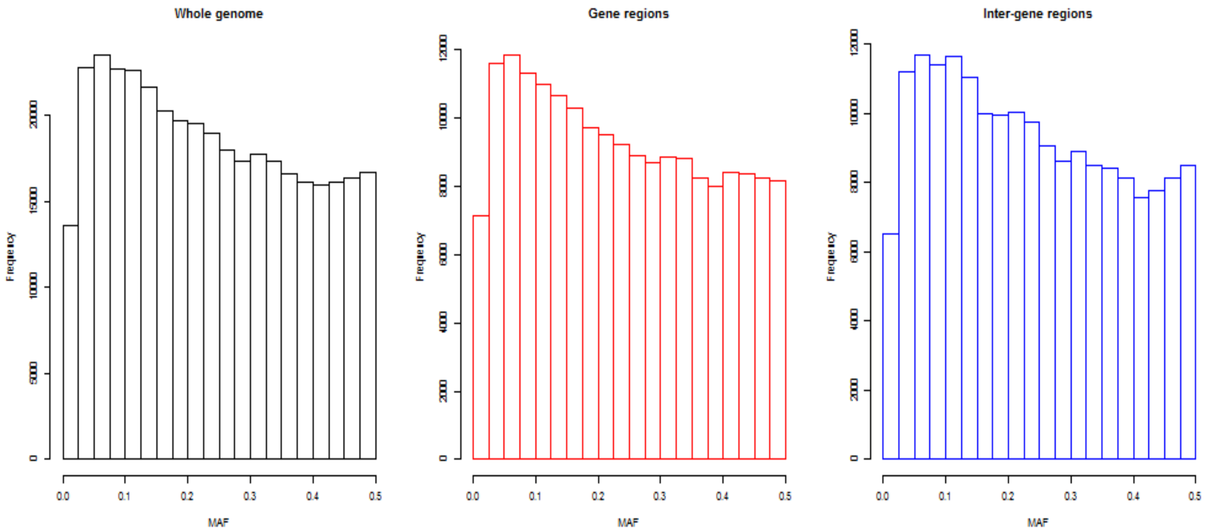


Figure S4.5. Distribution of minor allele frequencies in *H. sapiens* across the whole genome, in genic and -non-genic regions, respectively*.*

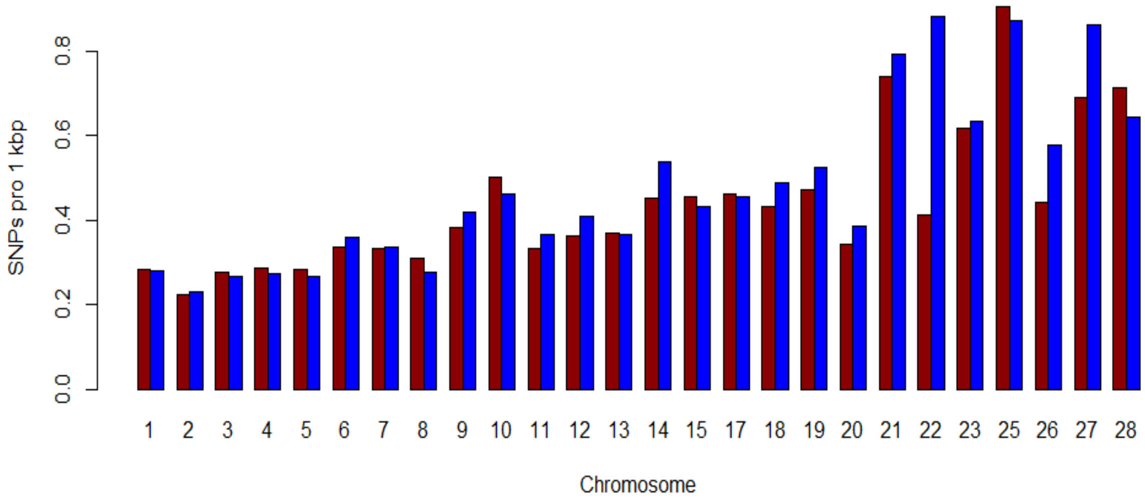


Figure S4.6. SNP-density for chromosomes 1 to 28 in G. g. domesticus. Red bars stand for density of SNPs in genic regions, blue bars stand for SNP-density in non-genic regions.

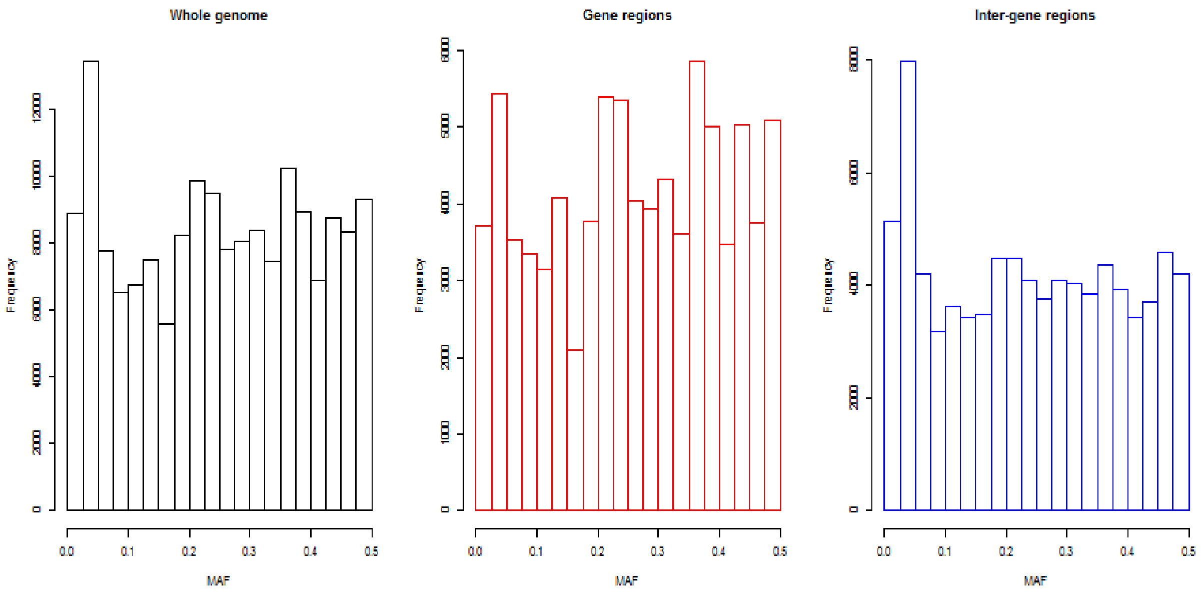


Figure S4.7. Distribution of minor allele frequencies in *G. g. domesticus* **across the whole genome, in genic and in inter-gene regions, respectively**.

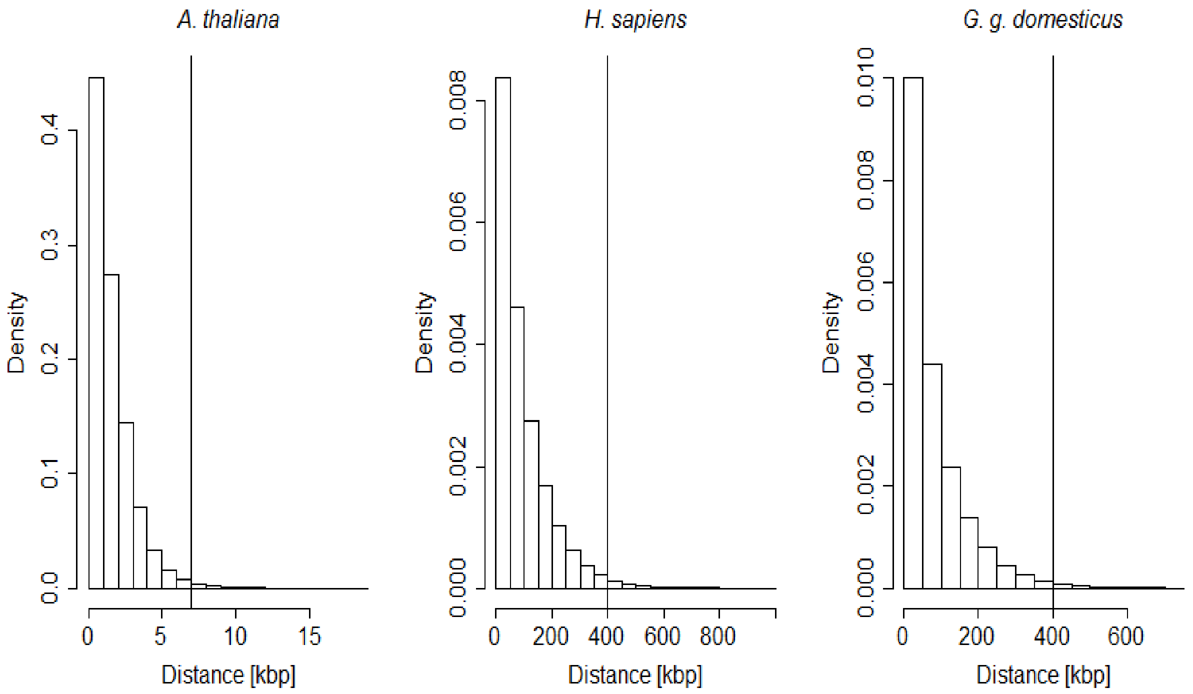


Figure S4.8. Distribution of pair-wise distances of SNP pairs in *A. thaliana, H. sapiens* and *G. g. domesticus.* The black vertical line refers to threshold cutting off the upper 1% of data points.

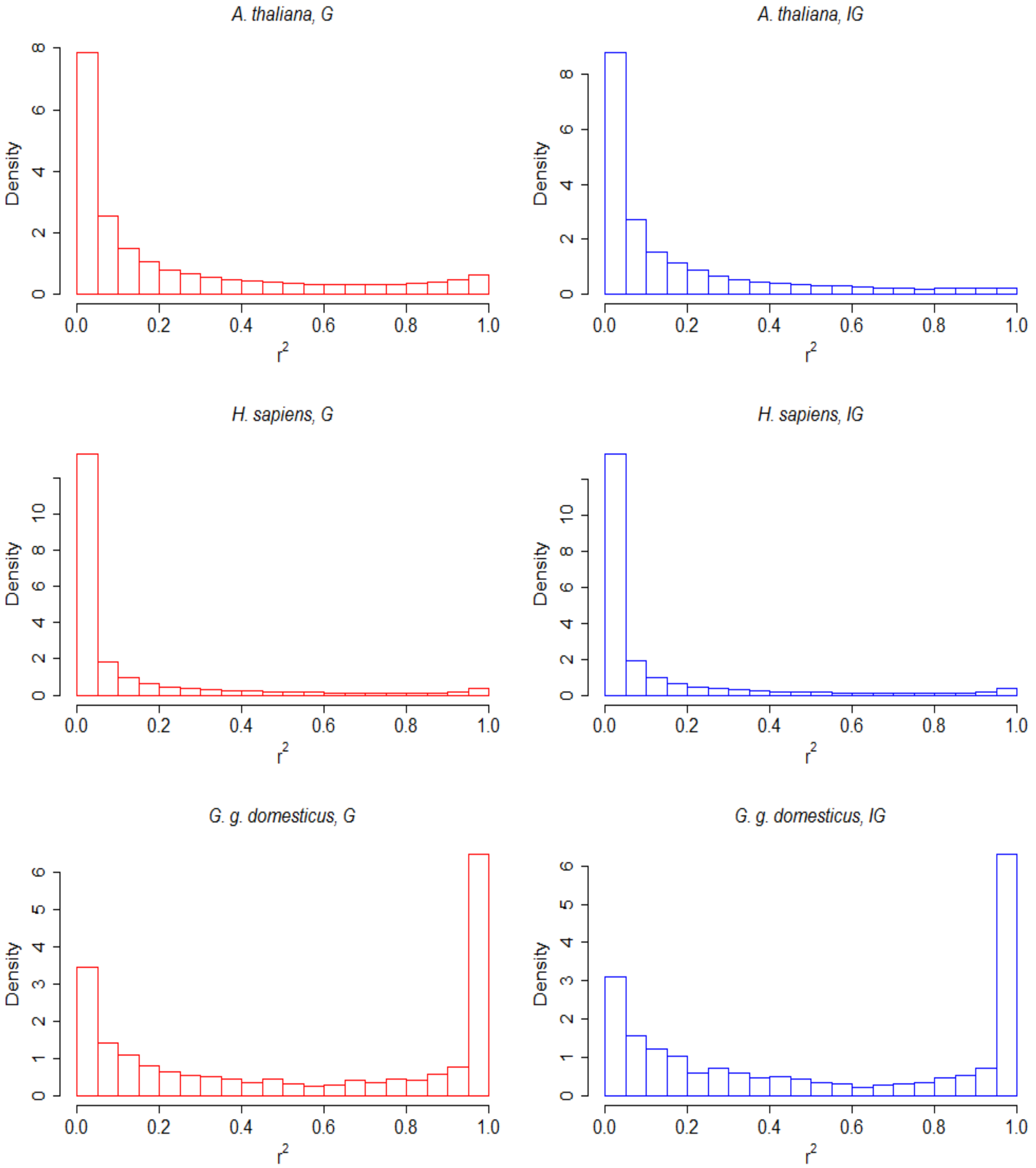


Figure S4.9. Distribution of pair-wise r^2**.** Distributions of squared correlations r^2 in A. thaliana (upper panel), H. sapiens (central panel), and G. g. domesticus (lower panel) in gene (red) and non-genic (blue) regions.

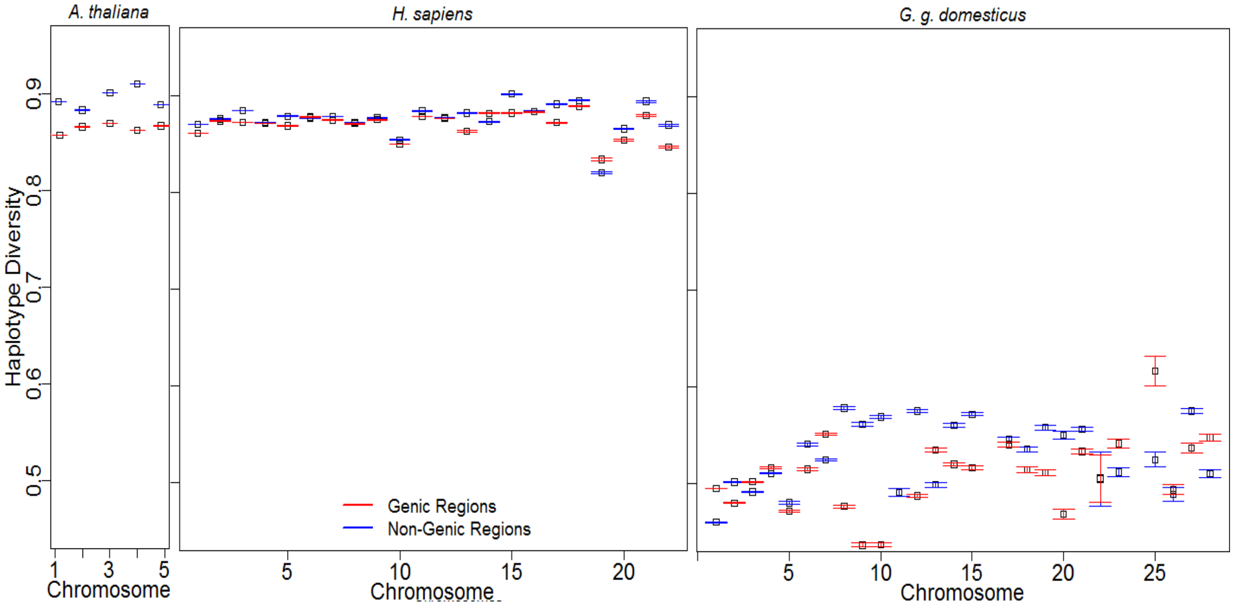


Figure S4.10. Chromosome-wise haplotype diversity in genic and non-genic regions across species. Chromosome-wise haplotype diversity in *G* (red) and *IG* (blue)

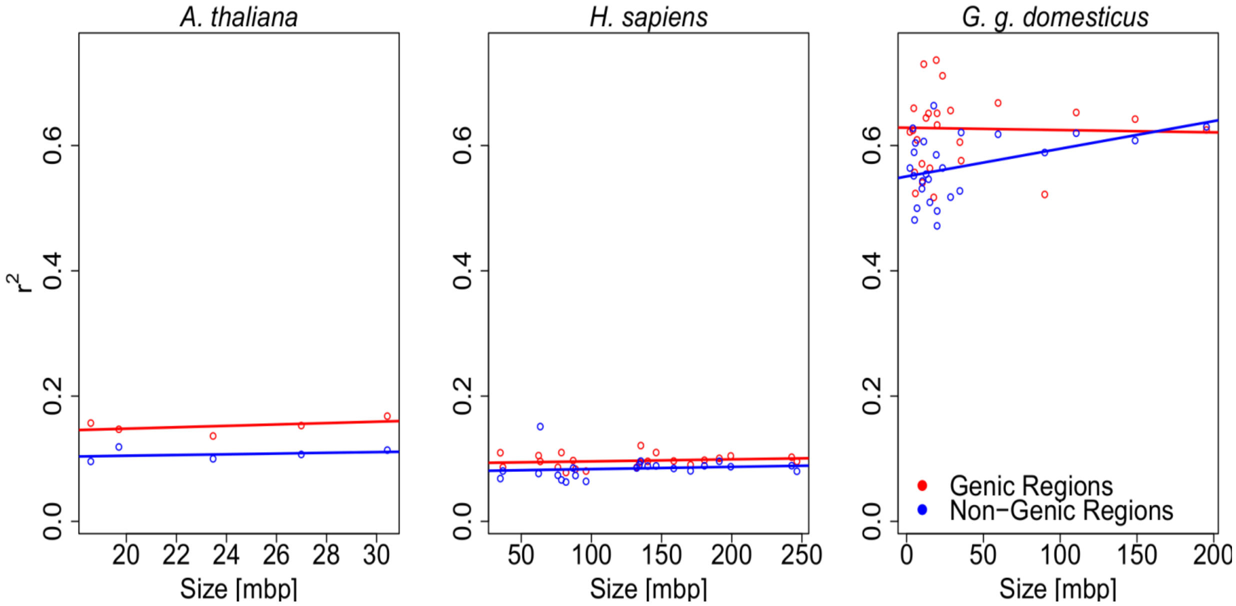


Figure S4.11. Medians of r^2 in genic and non-genic regions vs. chromosome size in *A. thaliana, H. sapiens, and G. g. domesticus*. Slope of all regression lines does not differ significantly from zero.

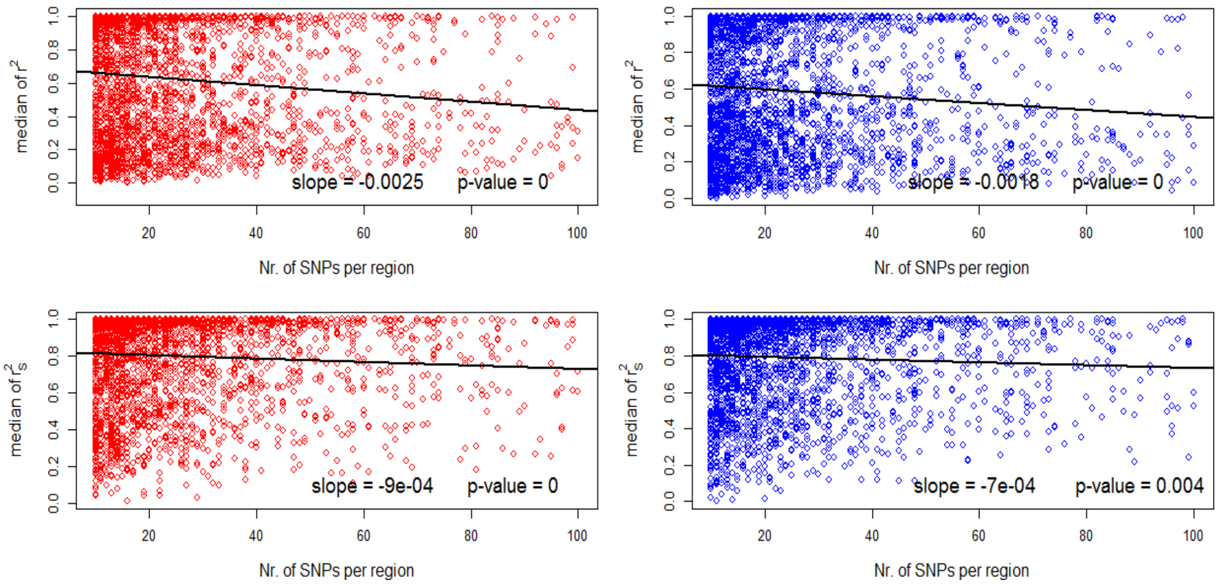


Figure S4.12. Relationship between magnitude of LD and the size of regions measured in number of SNPs, across chromosomes in chicken. Genic regions are drawn in red and non-genic regions in blue, X-axis reflects number of SNPs per region, Y-Axis reflects medians of r^2 (upper panel) or medians of r_S^2 (lower panel). The slope of the linear regression and its corresponding p-value are drown in each panel

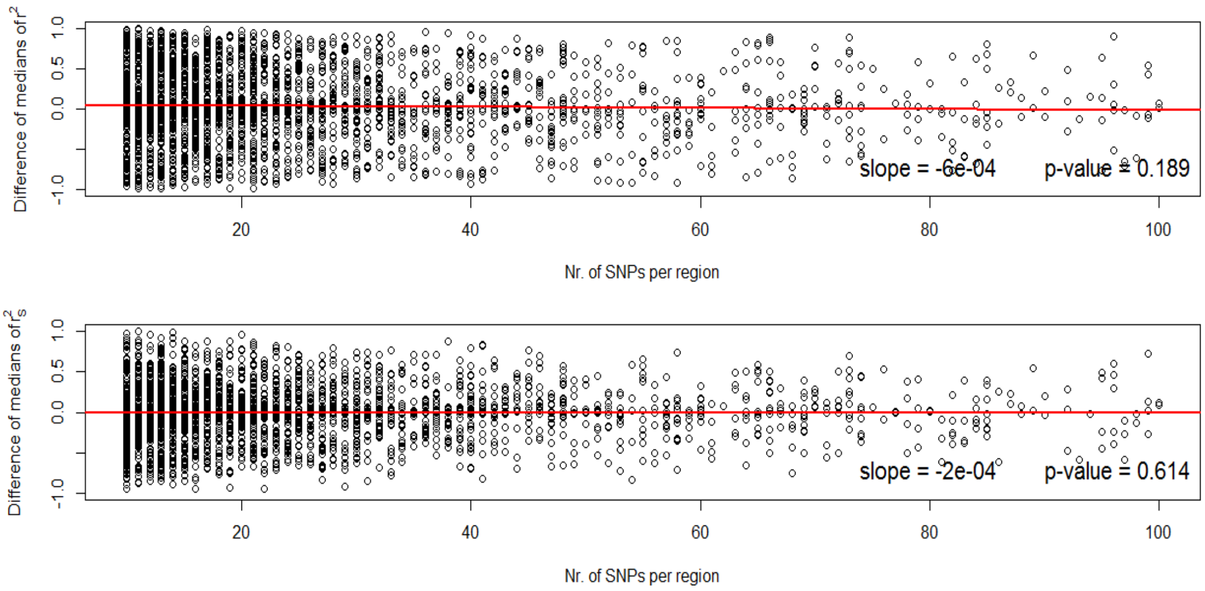


Figure S4.13. G/IG differences in medians of r^2 (upper panel) or medians of r_S^2 (lower panel), against the size of regions (in number of SNPs) across chromosomes in chicken.

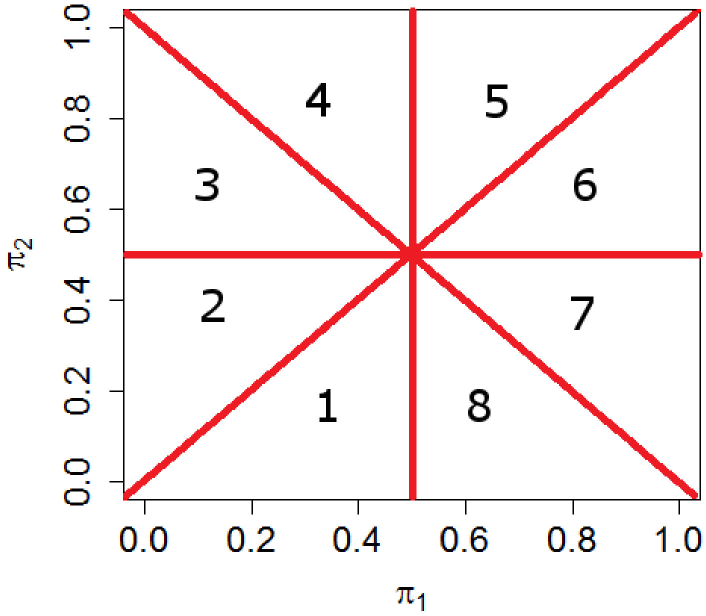


Figure S4.13 Two-dimensional probability space, divided in eight sections. X-axis und Y-axis describe the probabilities π_1 and π_2 .

Table S4.1. Chromosome-wise averaged medians of pair-wise r^2, calculated in each *G, IG* or *IG'* region for chromosome 1 to 5 in *A.thaliana*. D*ifference abs* is the absolute deviation of median in *IG* from median in *G* (or median in *IG'* from median in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

		Median		Difference		p-Val	Median		Difference		p-Val
chr	#genes	G	IG	abs	%		IG	IG‘	abs	%	
1	858	0.167	0.111	0.055	49.7	0	0.114	0.103	0.011	9.7	0.094
2	348	0.147	0.118	0.029	24.6	0.016	0.119	0.094	0.025	21.0	0.200
3	695	0.136	0.100	0.035	35.4	0	0.100	0.089	0.011	11.0	0.529
4	669	0.155	0.096	0.059	61.6	0	0.096	0.092	0.003	4.2	0.746
5	943	0.153	0.106	0.046	43.5	0	0.107	0.111	-0.004	-3.7	0.254
Genome-wide		0.154	0.106	0.048	31.2	$2 \cdot 10^{-16}$	0.106	0.099	0.007	6.6	0.2814

Table S4.2. Chromosome-wise averaged medians of pair-wise r^2, calculated in each *G*, *IG* or *IG'* region for chromosome 1 to 22 in *H.sapiens*. D*ifference abs* is the absolute deviation of median in *IG* from median in *G* (or median in *IG'* from median in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Median		Difference		p-Val	Median		Difference		p-Val
		G	IG	abs	%		IG	IG‘	abs	%	
1	661	0.096	0.080	0.016	16.7	0.038	0.080	0.083	-0.003	-3.7	0.661
2	571	0.103	0.089	0.014	13.6	0.037	0.089	0.089	0	0.0	0.657
3	437	0.105	0.087	0.018	17.1	0.181	0.087	0.084	0.003	3.4	0.223
4	410	0.101	0.096	0.005	4.9	0.372	0.096	0.092	0.004	4.2	0.195
5	405	0.098	0.089	0.009	9.2	0.433	0.089	0.090	-0.001	-1.1	0.612
6	406	0.090	0.081	0.009	10.0	0.991	0.081	0.083	-0.002	-2.5	0.103
7	318	0.096	0.085	0.011	11.4	0.888	0.085	0.085	0	0.0	0.956
8	322	0.110	0.089	0.021	19.1	0.064	0.089	0.082	0.007	7.9	0.497
9	298	0.096	0.088	0.008	8.3	0.471	0.088	0.090	-0.002	-2.3	0.996
10	344	0.121	0.096	0.025	20.7	0.070	0.096	0.092	0.004	4.2	0.553
11	344	0.094	0.091	0.003	3.2	0.857	0.091	0.082	0.009	9.9	0.674
12	395	0.086	0.085	0.001	1.2	0.930	0.085	0.075	0.010	11.8	0.192
13	188	0.080	0.064	0.016	20.0	0.130	0.064	0.067	-0.003	-4.7	0.954
14	244	0.097	0.085	0.012	12.4	0.134	0.085	0.078	0.007	8.2	0.196
15	226	0.078	0.063	0.015	19.2	0.125	0.063	0.057	0.006	9.5	0.372
16	206	0.083	0.073	0.01	12.0	0.867	0.073	0.077	-0.004	-5.5	0.856
17	253	0.110	0.066	0.044	40.0	0.000	0.066	0.062	0.004	6.1	0.214
18	178	0.086	0.074	0.012	14.0	0.468	0.074	0.075	-0.001	-1.4	0.511
19	90	0.096	0.151	-0.055	57.3	0.097	0.151	0.119	0.032	21.2	0.378
20	177	0.105	0.076	0.029	27.7	0.004	0.076	0.075	0.001	1.3	0.682
21	89	0.086	0.080	0.006	7.0	0.584	0.080	0.088	-0.008	-10.0	0.743
22	108	0.110	0.068	0.042	38.2	0.013	0.068	0.073	-0.005	-7.4	0.437
Genome-wide		0.098	0.084	0.013	13.6	$2\cdot10^{-5}$	0.0844	0.0824	0.002	2.4	0.378

Table S4.3. Chromosome-wise averaged medians of pair-wise r^2, calculated in each *G, IG* or *IG'* region for chromosome 1 to 26 in *G. g. domesticus*. D*ifference abs* is the absolute deviation of median in *IG* from median in *G* (or median in *IG'* from median in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

		Median		Difference		p-Val	Median		Difference		p-Val
chr	#genes	G	IG	abs	%		IG	IG‘	abs	%	
1	531	0.625	0.630	-0.005	-0.8	0.532	0.630	0.630	0	0	0.711
2	346	0.642	0.608	0.034	5.3	0.185	0.608	0.609	-0.001	-0.2	0.738
3	310	0.651	0.620	0.031	4.8	0.176	0.620	0.623	-0.003	-0.5	0.525
4	255	0.522	0.589	-0.067	-12.8	0.010	0.589	0.565	0.024	4.1	0.293
5	183	0.664	0.618	0.046	6.9	0.185	0.618	0.644	-0.026	-4.2	0.669
6	140	0.605	0.528	0.077	12.7	0.010	0.528	0.563	-0.035	-6.6	0.204
7	141	0.576	0.621	-0.045	-7.8	0.195	0.621	0.574	0.047	7.6	0.082
8	95	0.656	0.518	0.138	21.0	0.005	0.518	0.566	-0.048	-9.3	0.239
9	83	0.711	0.564	0.147	20.7	0.002	0.564	0.551	0.013	2.3	0.772
10	110	0.633	0.496	0.137	21.6	0.003	0.496	0.511	-0.015	-3.0	0.827
11	52	0.701	0.585	0.116	16.6	0.007	0.585	0.604	-0.019	-3.3	0.797
12	94	0.651	0.472	0.179	27.5	0.000	0.472	0.546	-0.074	-15.7	0.174
13	72	0.517	0.664	-0.147	-28.4	0.022	0.664	0.722	-0.058	-8.7	0.350
14	101	0.564	0.509	0.055	9.8	0.301	0.509	0.587	-0.078	-15.3	0.075
15	75	0.644	0.554	0.090	14.0	0.098	0.554	0.551	0.003	0.5	0.790
17	68	0.541	0.543	-0.002	-0.4	0.815	0.543	0.554	-0.011	-2.0	0.502
18	57	0.730	0.606	0.124	17.0	0.024	0.606	0.587	0.019	3.1	0.757
19	60	0.571	0.531	0.040	7.0	0.553	0.531	0.561	-0.030	-5.7	0.340
20	39	0.651	0.546	0.105	16.1	0.324	0.546	0.492	0.054	9.9	0.831
21	63	0.609	0.500	0.109	17.9	0.051	0.500	0.564	-0.064	12.8	0.174
22	7	0.624	0.628	-0.004	-0.6	1.000	0.628	0.685	-0.057	-9.1	1.000
23	39	0.524	0.604	-0.080	-15.3	0.277	0.604	0.562	0.042	6.9	0.438
25	10	0.622	0.564	0.058	9.3	0.846	0.564	0.509	0.055	9.8	0.770
26	26	0.814	0.589	0.225	27.6	0.012	0.589	0.631	-0.042	-7.1	0.354
27	36	0.557	0.481	0.076	13.6	0.346	0.481	0.373	0.108	22.5	0.058
28	39	0.660	0.552	0.108	16.4	0.121	0.552	0.520	0.032	5.8	0.805
Genome-wide		0.621	0.584	0.037	6.0	0.008	0.584	0.591	-0.007	-1.2	0.57

Table S4.4. Chromosome-wise averaged medians of pair-wise r_S^2, calculated in each *G, IG* or *IG’* region for chromosome 1 to 5 in *A.thaliana*. *Difference abs* is the absolute deviation of median in *IG* from median in *G* (or median in *IG’* from median in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

		Median		Difference		p-Val	Median		Difference		p-Val
chr	#genes	G	IG	abs	%		IG	IG‘	abs	%	
1	858	0.311	0.218	0.093	29.9	10^{-6}	0.218	0.201	0.017	7.8	0.017
2	348	0.278	0.233	0.045	16.2	0.0018	0.233	0.203	0.030	12.9	0.130
3	695	0.275	0.194	0.081	29.5	10^{-6}	0.194	0.185	0.009	4.6	0.411
4	669	0.296	0.195	0.101	34.1	10^{-6}	0.195	0.196	-0.001	-0.5	0.941
5	943	0.290	0.221	0.069	23.8	10^{-6}	0.221	0.225	-0.004	-1.8	0.284
Genome-wide		0.292	0.211	0.081	27.7	$2\cdot10^{-16}$	0.211	0.203	0.008	3.7	0.1454

Table S4.5. Chromosome-wise averaged medians of pair-wise r_S^2, calculated in each *G, IG* or *IG'* region for chromosome 1 to 22 in *H.sapiens*. D*ifference abs* is the absolute deviation of median in *IG* from median in *G* (or median in *IG'* from median in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Median		Difference		p-Val	Median		Difference		p-Val
		G	IG	Abs	%		IG	IG'	abs	%	
1	661	0.208	0.189	0.019	9.1	0.038	0.189	0.195	-0.006	-3.2	0.998
2	571	0.213	0.201	0.012	5.6	0.037	0.201	0.193	0.008	3.9	0.168
3	437	0.217	0.198	0.019	8.8	0.181	0.198	0.190	0.008	4.0	0.406
4	410	0.202	0.216	-0.014	-6.9	0.372	0.216	0.202	0.014	6.5	0.084
5	405	0.226	0.203	0.023	10.2	0.433	0.203	0.205	-0.002	-1.0	0.982
6	406	0.200	0.193	0.007	3.5	0.991	0.193	0.201	-0.008	-4.2	0.136
7	318	0.213	0.202	0.011	5.2	0.888	0.202	0.197	0.005	2.5	0.636
8	322	0.231	0.211	0.020	8.7	0.064	0.211	0.192	0.019	9.0	0.116
9	298	0.214	0.205	0.009	4.2	0.471	0.205	0.208	-0.003	-1.5	0.880
10	344	0.243	0.221	0.022	9.1	0.070	0.221	0.220	0.001	0.5	0.966
11	344	0.216	0.198	0.018	8.3	0.857	0.198	0.197	0.001	0.6	0.645
12	395	0.196	0.195	0.001	0.5	0.930	0.195	0.189	0.006	3.1	0.830
13	188	0.209	0.162	0.047	22.5	0.130	0.162	0.177	-0.015	-9.3	0.809
14	244	0.213	0.208	0.005	2.3	0.134	0.208	0.190	0.018	8.7	0.382
15	226	0.179	0.150	0.029	16.2	0.125	0.150	0.137	0.013	8.7	0.272
16	206	0.183	0.165	0.018	9.8	0.867	0.165	0.170	-0.005	-3.0	0.771
17	253	0.225	0.158	0.067	29.8	0.000	0.158	0.148	0.010	6.3	0.350
18	178	0.182	0.169	0.013	7.1	0.468	0.169	0.169	0	0	0.690
19	90	0.232	0.276	-0.044	-19.0	0.097	0.276	0.265	0.011	4.0	0.872
20	177	0.224	0.177	0.047	20.9	0.004	0.177	0.179	-0.002	-1.1	0.642
21	89	0.200	0.196	0.004	2.0	0.584	0.196	0.217	-0.021	-10.7	0.479
22	108	0.237	0.166	0.071	29.9	0.013	0.166	0.187	-0.021	-12.7	0.260
Genome-wide		0.2119	0.1949	0.017	8.0	$3 \cdot 10^{-6}$	0.1949	0.1923	0.0026	1.3	0.188

Table S4.6. Chromosome-wise averaged medians of pair-wise r_S^2, calculated in each *G, IG* or *IG'* region for chromosome 1 to 26 in *G. g. domesticus*. *Difference abs* is the absolute deviation of median in *IG* from median in *G* (or median in *IG'* from median in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

		Median		Difference		p-Val	Median		Difference		p-Val
chr	#genes	G	IG	abs	%		IG	IG‘	abs	%	
1	531	0.794	0.821	-0.027	-3.4	0.075	0.821	0.817	0.004	0.5	0.987
2	346	0.799	0.770	0.029	3.6	0.050	0.770	0.773	-0.003	-0.4	0.933
3	310	0.827	0.822	0.005	0.6	0.688	0.822	0.830	-0.008	-1.0	0.809
4	255	0.723	0.808	-0.085	-11.8	0.000	0.808	0.783	0.025	3.1	0.241
5	183	0.804	0.811	-0.007	-0.9	0.777	0.811	0.819	-0.008	-1.0	0.985
6	140	0.809	0.755	0.054	6.7	0.037	0.755	0.767	-0.012	-1.6	0.470
7	141	0.771	0.821	-0.050	-6.5	0.056	0.821	0.801	0.020	2.4	0.498
8	95	0.803	0.758	0.045	5.6	0.109	0.758	0.782	-0.024	-3.2	0.527
9	83	0.853	0.777	0.076	8.9	0.023	0.777	0.779	-0.002	-0.3	0.471
10	110	0.791	0.726	0.065	8.2	0.022	0.726	0.740	-0.014	-1.9	0.457
11	52	0.808	0.782	0.026	3.2	0.137	0.782	0.823	-0.041	-5.2	0.318
12	94	0.800	0.731	0.069	8.6	0.067	0.731	0.768	-0.037	-5.1	0.226
13	72	0.745	0.852	-0.107	-14.4	0.015	0.852	0.879	-0.027	-3.2	0.148
14	101	0.764	0.742	0.022	2.9	0.533	0.742	0.792	-0.050	-6.7	0.122
15	75	0.841	0.783	0.058	6.9	0.042	0.783	0.765	0.018	2.3	0.603
17	68	0.768	0.774	-0.006	-0.8	0.724	0.774	0.777	-0.003	-0.4	0.949
18	57	0.861	0.788	0.073	8.5	0.038	0.788	0.770	0.018	2.3	0.408
19	60	0.786	0.759	0.027	3.4	0.271	0.759	0.805	-0.046	-6.1	0.348
20	39	0.800	0.776	0.024	3.0	0.572	0.776	0.702	0.074	9.5	0.225
21	63	0.809	0.741	0.068	8.4	0.094	0.741	0.818	-0.077	-10.4	0.126
22	7	0.827	0.844	-0.017	-2.1	0.402	0.844	0.898	-0.054	-6.4	1.000
23	39	0.718	0.792	-0.074	-10.3	0.225	0.792	0.761	0.031	3.9	0.380
25	10	0.871	0.741	0.130	14.9	0.375	0.741	0.768	-0.027	-3.6	1.000
26	26	0.895	0.840	0.055	6.2	0.034	0.840	0.851	-0.011	-1.3	0.681
27	36	0.776	0.758	0.018	2.3	0.883	0.758	0.686	0.072	9.5	0.046
28	39	0.852	0.803	0.049	5.8	0.395	0.803	0.771	0.032	4.0	0.674
Genome-wide		0.795	0.791	0.004	0.5	0.059	0.791	0.794	-0.003	-0.4	0.438

Table S4.7. Chromosome-wise averaged means of pair-wise r^2, calculated in each *G*, *IG* or *IG'* region for chromosome 1 to 5 in *A.thaliana*. D*ifference abs* is the absolute deviation of mean in *IG* from mean in *G* (or mean in *IG'* from mean in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Mean		Difference		p-Val	Mean		Difference		p-Val
		G	IG	abs	%		IG	IG'	abs	%	
1	858	0.256	0.196	0.060	23.4	10^{-6}	0.196	0.183	0.013	6.6	0.005
2	348	0.235	0.207	0.028	11.9	0.003	0.207	0.190	0.017	8.2	0.049
3	695	0.231	0.179	0.052	22.5	10^{-6}	0.179	0.172	0.007	3.9	0.423
4	669	0.240	0.166	0.074	30.8	10^{-6}	0.166	0.170	-0.004	-2.0	0.437
5	943	0.243	0.195	0.048	19.8	10^{-6}	0.195	0.203	-0.008	-4.0	0.026
Genome-wide		0.242	0.188	0.054	22.3	$2 \cdot 10^{-16}$	0.188	0.185	0.003	1.6	0.339

Table S4.8. Chromosome-wise averaged means of pair-wise r^2, calculated in each *G*, *IG* or *IG'* region for chromosome 1 to 22 in *H.sapiens*. D*ifference abs* is the absolute deviation of mean in *IG* from mean in *G* (or mean in *IG'* from mean in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Mean		Difference		p-Val	Mean		Difference		p-Val
		G	IG	abs	%		IG	IG‘	abs	%	
1	661	0.203	0.190	0.013	6.4	0.005	0.190	0.187	0.003	1.6	0.308
2	571	0.199	0.191	0.008	4.0	0.150	0.191	0.186	0.005	2.6	0.308
3	437	0.203	0.187	0.016	7.9	0.017	0.187	0.181	0.006	3.2	0.366
4	410	0.199	0.202	-0.003	-1.5	0.206	0.202	0.195	0.007	3.5	0.175
5	405	0.206	0.190	0.016	7.8	0.007	0.190	0.191	-0.001	-0.5	0.984
6	406	0.188	0.186	0.002	1.1	0.646	0.186	0.192	-0.006	3.2	0.144
7	318	0.197	0.194	0.003	1.5	0.580	0.194	0.188	0.006	3.1	0.138
8	322	0.209	0.191	0.018	8.6	0.080	0.191	0.186	0.005	2.6	0.607
9	298	0.198	0.192	0.006	3.0	0.765	0.192	0.191	0.001	0.5	0.534
10	344	0.217	0.203	0.014	6.5	0.235	0.203	0.202	0.001	0.5	0.675
11	344	0.201	0.193	0.008	3.9	0.393	0.193	0.189	0.004	2.1	0.564
12	395	0.191	0.187	0.004	2.1	0.328	0.187	0.181	0.006	3.2	0.517
13	188	0.193	0.169	0.024	12.4	0.001	0.169	0.175	-0.006	-3.6	0.953
14	244	0.192	0.188	0.004	2.1	0.374	0.188	0.181	0.007	3.7	0.277
15	226	0.179	0.163	0.016	8.9	0.128	0.163	0.153	0.010	6.1	0.051
16	206	0.185	0.176	0.009	4.9	0.406	0.176	0.171	0.005	2.8	0.373
17	253	0.204	0.166	0.038	18.6	0.000	0.166	0.158	0.008	4.8	0.136
18	178	0.175	0.174	0.001	0.6	0.975	0.174	0.175	-0.001	-0.6	0.670
19	90	0.206	0.230	-0.024	-11.7	0.351	0.230	0.223	0.007	3.0	0.636
20	177	0.210	0.191	0.019	9.1	0.050	0.191	0.183	0.008	4.2	0.547
21	89	0.195	0.188	0.007	3.6	0.740	0.188	0.188	0.000	0.0	0.825
22	108	0.212	0.173	0.039	18.4	0.006	0.173	0.178	-0.005	-2.9	0.392
Genome-wide		0.199	0.188	0.011	5.3	$6\cdot10^{-8}$	0.188	0.185	0.004	1.9	0.012

Table S4.9. Chromosome-wise averaged means of pair-wise r^2, calculated in each *G, IG* or *IG'* region for chromosome 1 to 26 in *G. g. domesticus*. D*ifference abs* is the absolute deviation of mean in *IG* from mean in *G* (or mean in *IG'* from mean in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Mean		Difference		p-Val	Mean		Difference		p-Val
		G	IG	abs	%		IG	IG‘	abs	%	
1	531	0.645	0.644	0.001	0.2	0.850	0.644	0.643	0.001	0.2	0.891
2	346	0.648	0.622	0.026	4.0	0.046	0.622	0.627	-0.005	-0.8	0.615
3	310	0.668	0.625	0.043	6.4	0.022	0.625	0.637	-0.012	-1.9	0.170
4	255	0.559	0.602	-0.040	-7.7	0.013	0.602	0.586	0.016	2.7	0.177
5	183	0.678	0.626	0.052	7.7	0.031	0.626	0.661	-0.035	-5.6	0.011
6	140	0.629	0.593	0.036	5.7	0.095	0.593	0.600	-0.007	-1.2	0.542
7	141	0.615	0.632	-0.020	-2.7	0.381	0.632	0.617	0.015	2.43	0.322
8	95	0.687	0.570	0.117	17.0	0.000	0.570	0.568	0.002	0.4	0.825
9	83	0.669	0.596	0.073	10.9	0.012	0.596	0.599	-0.003	-0.5	0.958
10	110	0.660	0.545	0.115	17.4	0.000	0.545	0.550	-0.005	-0.9	0.732
11	52	0.709	0.595	0.114	16.1	0.001	0.595	0.601	-0.006	-1.0	0.788
12	94	0.677	0.552	0.125	18.5	0.000	0.552	0.572	-0.020	-3.6	0.205
13	72	0.563	0.660	-0.100	-17.2	0.011	0.660	0.686	-0.026	-3.9	0.130
14	101	0.609	0.569	0.040	6.6	0.227	0.569	0.604	-0.035	-6.2	0.015
15	75	0.658	0.581	0.077	11.7	0.049	0.581	0.576	0.005	0.9	0.835
17	68	0.598	0.590	0.008	1.4	0.939	0.590	0.590	0	0	0.959
18	57	0.719	0.631	0.088	12.2	0.013	0.631	0.615	0.016	2.5	0.328
19	60	0.598	0.581	0.017	2.8	0.800	0.581	0.6	-0.019	-3.3	0.473
20	39	0.686	0.602	0.084	12.2	0.171	0.602	0.567	0.035	5.8	0.117
21	63	0.639	0.554	0.085	13.3	0.040	0.554	0.562	-0.008	-1.4	0.649
22	7	0.619	0.65	-0.030	-5.0	0.578	0.650	0.653	-0.003	-0.5	0.578
23	39	0.582	0.624	-0.040	-7.2	0.435	0.624	0.577	0.047	7.5	0.019
25	10	0.616	0.543	0.073	11.9	0.557	0.543	0.560	-0.017	-3.1	1.000
26	26	0.810	0.613	0.197	24.3	0.002	0.613	0.632	-0.019	-3.1	0.745
27	36	0.567	0.511	0.056	9.9	0.279	0.511	0.476	0.035	6.9	0.131
28	39	0.679	0.57	0.109	16.1	0.036	0.570	0.560	0.010	1.8	0.664
Genome-wide		0.642	0.609	0.033	5.2	$8 \cdot 10^{-7}$	0.6091	0.6124	-0.003	-0.5	0.290

Table S4.10. Chromosome-wise averaged haplotype diversity, calculated in each *G, IG* or *IG’* region for chromosome 1 to 5 in *A.thaliana*. D*ifference abs* is the absolute deviation of mean in *IG* from mean in *G* (or mean in *IG’* from mean in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Mean		Difference		p-Val	Mean		Difference		p-Val
		G	IG	abs	%		IG	IG’	abs	%	
1	858	0.857	0.892	-0.034	-3.8	0	0.892	0.898	-0.006	-0.7	0.012
2	348	0.865	0.883	-0.018	-2.0	0.007	0.883	0.891	-0.008	-0.9	0.083
3	695	0.869	0.901	-0.031	-3.5	0	0.901	0.901	0	-0.1	0.832
4	669	0.862	0.910	-0.048	-5.3	0	0.910	0.904	0.006	0.6	0.049
5	943	0.866	0.889	-0.023	-2.6	0	0.889	0.886	0.003	0.4	0.268
Genome-wide		0.864	0.895	-0.032	-3.5	0	0.895	0.896	-0.005	-0.1	0.747

Table S4.11. Chromosome-wise averaged haplotype diversity, calculated in each *G, IG* or *IG'* region for chromosome 1 to 22 in *H.sapiens*. D*ifference abs* is the absolute deviation of mean in *IG* from mean in *G* (or mean in *IG'* from mean in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Mean		Difference		p-Val	Mean		Difference		p-Val
		G	IG	abs	%		IG	IG‘	abs	%	
1	661	0.861	0.870	-0.009	-1.1	0.059	0.870	0.868	0.002	0.2	0.930
2	571	0.874	0.876	-0.002	-0.2	0.867	0.876	0.879	-0.004	-0.4	0.358
3	437	0.872	0.884	-0.012	-1.4	0.036	0.884	0.889	-0.005	-0.5	0.611
4	410	0.871	0.872	-0.001	-0.1	0.585	0.872	0.881	-0.009	-0.9	0.112
5	405	0.868	0.878	-0.010	-1.1	0.122	0.878	0.872	0.007	0.8	0.304
6	406	0.878	0.877	0.001	0.1	0.567	0.877	0.873	0.004	0.5	0.311
7	318	0.874	0.878	-0.004	-0.5	0.577	0.878	0.880	-0.002	-0.2	0.283
8	322	0.871	0.872	-0.001	-0.1	0.980	0.872	0.883	-0.011	-1.2	0.136
9	298	0.875	0.877	-0.002	-0.2	0.464	0.877	0.881	-0.005	-0.5	0.550
10	344	0.850	0.854	-0.004	-0.5	0.778	0.854	0.856	-0.002	-0.2	0.677
11	344	0.878	0.884	-0.005	-0.6	0.325	0.884	0.876	0.008	0.9	0.347
12	395	0.876	0.877	-0.001	-0.1	0.503	0.877	0.881	-0.004	-0.4	0.519
13	188	0.863	0.882	-0.019	-2.2	0.013	0.882	0.874	0.008	0.9	0.839
14	244	0.881	0.873	0.009	1.0	0.021	0.873	0.879	-0.006	-0.7	0.432
15	226	0.882	0.902	-0.020	-2.2	0.006	0.902	0.903	-0.002	-0.2	0.676
16	206	0.883	0.883	0	-0.1	0.760	0.883	0.891	-0.007	-0.8	0.531
17	253	0.872	0.891	-0.019	-2.1	0.003	0.891	0.898	-0.007	-0.8	0.378
18	178	0.889	0.895	-0.006	-0.7	0.940	0.895	0.893	0.002	0.2	0.474
19	90	0.834	0.820	0.014	1.7	0.412	0.820	0.825	-0.005	-0.6	0.906
20	177	0.854	0.865	-0.012	-1.4	0.100	0.865	0.873	-0.008	-0.9	0.756
21	89	0.879	0.894	-0.015	-1.7	0.171	0.894	0.882	0.012	1.3	0.338
22	108	0.847	0.869	-0.023	-2.6	0.007	0.869	0.859	0.01	1.2	0.398
Genome-wide		0.871	0.876	-0.006	-0.7	0.001	0.876	0.878	-0.001	-0.2	0.264

Table S4.12. Chromosome-wise averaged haplotype diversity, calculated in each *G, IG* or *IG’* region for chromosome 1 to 26 in *G. g. domesticus*. D*ifference abs* is the absolute deviation of mean in *IG* from mean in *G* (or mean in *IG’* from mean in *IG*) in corresponding regions, *Difference %* gives the percentage of deviation. *p-Val* is the p-value based on Wilcoxon signed rank test. Significant differences (p < 0.05) are marked in red.

chr	#genes	Mean		Difference		p-Val	Mean		Difference		p-Val
		G	IG	abs	%		IG	IG‘	abs	%	
1	531	0.495	0.460	0.035	7.1	0.000	0.460	0.456	0.005	1.1	0.918
2	346	0.480	0.501	-0.021	-4.4	0.028	0.501	0.474	0.027	5.4	0.006
3	308	0.502	0.491	0.011	2.2	0.573	0.491	0.470	0.022	4.5	0.002
4	255	0.516	0.511	0.006	1.2	0.914	0.511	0.501	0.010	2.0	0.343
5	181	0.472	0.480	-0.008	-1.7	0.537	0.480	0.433	0.047	9.8	0.000
6	140	0.515	0.541	-0.026	-5.0	0.014	0.541	0.509	0.032	5.9	0.034
7	141	0.551	0.525	0.026	4.7	0.494	0.525	0.491	0.034	6.5	0.124
8	95	0.476	0.578	-0.102	-21.4	0.000	0.578	0.560	0.018	3.1	0.217
9	83	0.436	0.561	-0.125	-28.7	0.000	0.561	0.529	0.032	5.7	0.128
10	110	0.437	0.569	-0.132	-30.2	0.000	0.569	0.558	0.011	1.9	0.969
11	45	0.400	0.491	-0.091	-22.8	0.012	0.491	0.488	0.003	0.6	0.858
12	94	0.487	0.575	-0.088	-18.1	0.000	0.575	0.563	0.012	2.1	0.371
13	72	0.535	0.499	0.036	6.7	0.125	0.499	0.490	0.009	1.8	0.656
14	101	0.520	0.560	-0.040	-7.7	0.062	0.560	0.512	0.048	8.6	0.001
15	75	0.516	0.572	-0.055	-10.7	0.013	0.572	0.571	0.000	0.0	0.851
17	68	0.540	0.545	-0.005	-0.9	0.934	0.545	0.532	0.014	2.6	0.345
18	57	0.514	0.535	-0.021	-4.1	0.249	0.535	0.542	-0.006	-1.1	0.639
19	60	0.511	0.558	-0.047	-9.2	0.121	0.558	0.540	0.018	3.2	0.420
20	39	0.469	0.550	-0.081	-17.3	0.032	0.550	0.548	0.002	0.4	0.704
21	63	0.533	0.556	-0.023	-4.3	0.491	0.556	0.548	0.008	1.4	0.719
22	7	0.506	0.505	0.001	0.2	1.000	0.505	0.532	-0.028	-5.5	0.578
23	39	0.541	0.512	0.030	5.5	0.403	0.512	0.510	0.002	0.4	0.845
25	10	0.616	0.525	0.092	14.9	0.106	0.525	0.586	-0.062	-11.8	0.160
26	26	0.494	0.489	0.005	1.0	0.980	0.489	0.500	-0.011	-2.2	0.269
27	36	0.537	0.575	-0.038	-7.1	0.293	0.575	0.579	-0.004	-0.7	0.379
28	39	0.547	0.510	0.037	6.8	0.521	0.510	0.508	0.002	0.4	0.841
Genome-wide		0.499	0.513	-0.015	-3.0	10^{-6}	0.513	0.496	0.017	3.4	0.001

Table S4.13. Slopes and in regressions of chromosome-wise averaged r^2 and r_S^2 medians on size of the chromosomes.

Species		Genic regions		Non-genic regions	
		slope	p-value	slope	p-value
A. thaliana	r^2	0.00111	0.4254	0.00058	0.6199
	r_S^2	0.00162	0.3280	0.00074	0.7249
H. sapiens	r^2	0,00003	0.4210	0.00004	0.5870
	r_S^2	0.00001	0.9290	0.00011	0.2980
G. g. domesticus	r^2	-0.00004	0.9030	0.00044	0.0360
	r_S^2	-0.00014	0.4460	0.00019	0.2190

Table S4.14. Upper Limit r^2_{max} under different limiting conditions

Section	Limiting conditions			r^2_{max}
1	$0 \leq \pi_2 \leq \pi_1 \leq 0.5$	and	$\pi_2 \leq 1 - \pi_1$	$\frac{\pi_2\,(1-\pi_1)}{\pi_1\,(1-\pi_2)}$
2	$0 \leq \pi_1 \leq \pi_2 \leq 0.5$	and	$\pi_2 \leq 1 - \pi_1$	$\frac{\pi_1\,(1-\pi_2)}{\pi_2\,(1-\pi_1)}$
3	$0 \leq \pi_1 \leq 0.5 \leq \pi_2 \leq 1$	and	$\pi_2 \leq 1 - \pi_1$	$\frac{\pi_1\,\pi_2}{(1-\pi_1)(1-\pi_2)}$
4	$0 \leq \pi_1 \leq 0.5 \leq \pi_2 \leq 1$	and	$\pi_2 \geq 1 - \pi_1$	$\frac{(1-\pi_1)(1-\pi_2)}{\pi_1\,\pi_2}$
5	$0.5 \leq \pi_1 \leq \pi_2 \leq 1$	and	$\pi_2 \geq 1 - \pi_1$	$\frac{\pi_1\,(1-\pi_2)}{\pi_2\,(1-\pi_1)}$
6	$0.5 \leq \pi_2 \leq \pi_1 \leq 1$	and	$\pi_2 \geq 1 - \pi_1$	$\frac{\pi_2\,(1-\pi_1)}{\pi_1\,(1-\pi_2)}$
7	$0 \leq \pi_2 \leq 0.5 \leq \pi_1 \leq 1$	and	$\pi_2 \geq 1 - \pi_1$	$\frac{(1-\pi_1)(1-\pi_2)}{\pi_1\,\pi_2}$
8	$0 \leq \pi_2 \leq 0.5 \leq \pi_1 \leq 1$	and	$\pi_2 \leq 1 - \pi_1$	$\frac{\pi_1\,\pi_2}{(1-\pi_1)(1-\pi_2)}$

5TH CHAPTER

General Discussion

Ever since Meuwissen et al. (2001) proposed use of genomic selection for improvement in marker-assisted selection in animal breeding programs, it has established itself in many areas of breeding. Whole-genome data of important breeding species like cattle, chicken or pig is available for predicting breeding values and association analyses in animal breeding (Stock and Reents, 2013). With currently available genotyping methods, SNP arrays with up to one million genomic markers are used in animal breeding, while those with about 3.000 markers are still used in plant breeding. Steady progress in gene sequencing technologies that enable cost effective identification of millions of DNA sequence reads in a single run, has led to an increase in the usage of genomic data for prediction of genetic merit. In the last ten years the genome sequencing costs have reduced from about $1,100 per mega base pair in July 2004 to $0.05 in July 2014 (http://www.genome.gov/sequencingcosts). The reduced genotyping costs allow increasing the sample size and consequently improving the power of the association analyses. For instance, in cattle, more than 90% of young dairy bulls from Holstein, Jersey and Brown Swiss breeds are genotyped (Schefers and Weigel, 2012). Also in the swine industry, the use of markers considerably improved the estimation of breeding values, even though the genotyping cost benefits are much lower as compared to dairy cattle (Van Eenennaam et al., 2014). In addition, the growing number of sequenced genomes across other species has opened opportunities to get fresh insights into the inheritance of traits and diseases (e.g. Fan et al., 2010; Daetwyler et al., 2012; Erbe et al., 2012). This explosion of information begs the question of whether the performance of genomic models will change given the increase in marker density. High-density data provided by modern methods of genomic sequencing are characterized by the high degree of non-random association between the markers (e.g. de los Campos at al., 2009), called linkage disequilibrium (LD), a quantity that tends to decay with growing physical distance. The investigation of the magnitude and the patterns of non-random association between loci has been a central question in genomic research (Georges, 2007; Amaral et al., 2008; Goddard and Hayes, 2009; Megens et al., 2009), mostly in the context of mapping genes causative for traits or diseases. In population genetics, the knowledge of LD structure helps to trace back the phylogenetic development of different species and offers fresh perspectives on evolutionary processes leading up to their development (Ardlie et al., 2002; Flint-Garcia et al., 2003; Wade et al., 2009; Qanbari et al., 2010).

In genomic models, the manifestation of a trait of interest is explained as the observed manifestation of genomic markers, while plenty of markers may be located in regions that do not contribute to genetic variance. Only markers that are in LD with an unknown quantitative trait locus (QTL) can capture the effects of causal loci. Adverse as well as beneficial effects of variation of LD level were investigated in the present work. The precision of estimation procedures of linear regression models was the subject of **chapter 2**, while **chapter 3** raised the issue of the predictive ability of commonly used quantitative methods applied to data from unrelated individuals. In **chapter 4** the comparison of LD structure in genic and non-genic regions was made by using a new scale-corrected comparison method.

Does too much LD in marker data affect the performance of genomic models?

The instability of marker effect estimations due to the degree of multicollinearity in the marker data was examined in the present thesis. The performances of three linear regression models – Single Marker Regression (SMR), Multiple Marker Regression (MMR) and Linear Mixed Models (LMM) were compared after varying the magnitude of LD in the marker data.

Simulation studies were used to examine the precision of effect estimates in models under comparison for traits with different genetic architectures (different heritability and minor allele frequency (MAF) distribution), using marker data with a predefined LD structure. To quantify the differences between the models, correlations between the estimates from SMR and MMR ($\mathrm{Cor}(\hat{\boldsymbol{\beta}})$), between the predictions ($\mathrm{Cor}(\hat{\mathbf{u}})$) and between predictive errors ($\mathrm{Cor}(\hat{\mathbf{u}}-\mathbf{u})$) in LMM were used. These correlations were derived analytically using the model assumptions and known variance structure of simulated data sets. Additionally, sample correlations were derived from 2500 replications in each scenario.

The LD structure of marker data seemed to be reflected by correlations between estimates from SMR and LMM. Even more interesting was the observation about the error in estimates from MMR and LMM: for weak LD the values of correlation between the estimation errors scattered around zero and an increase in LD led to an increase in negative correlation between the errors in estimates at both loci. Thus, the reduction of error in the estimated effects $\hat{\beta}_j - \beta_j$ as well as that in the predictions $\hat{u}_j - u_j$ at first locus may increase the error at the second. In contrast to MMR, predictions of marker effects in LMM seemed to be more sensitive to the LD in the data and were affected noticeably when LD in the data exceeded $r^2 \approx 0.6$. The results of MMR and LMM in simulations scenarios with heritability fixed at 0.3, 0.5 or 0.7 for LD varying between 0.01 and 0.81 and MAF varying between 0.05 and 0.5 are shown in Figure 5.1.

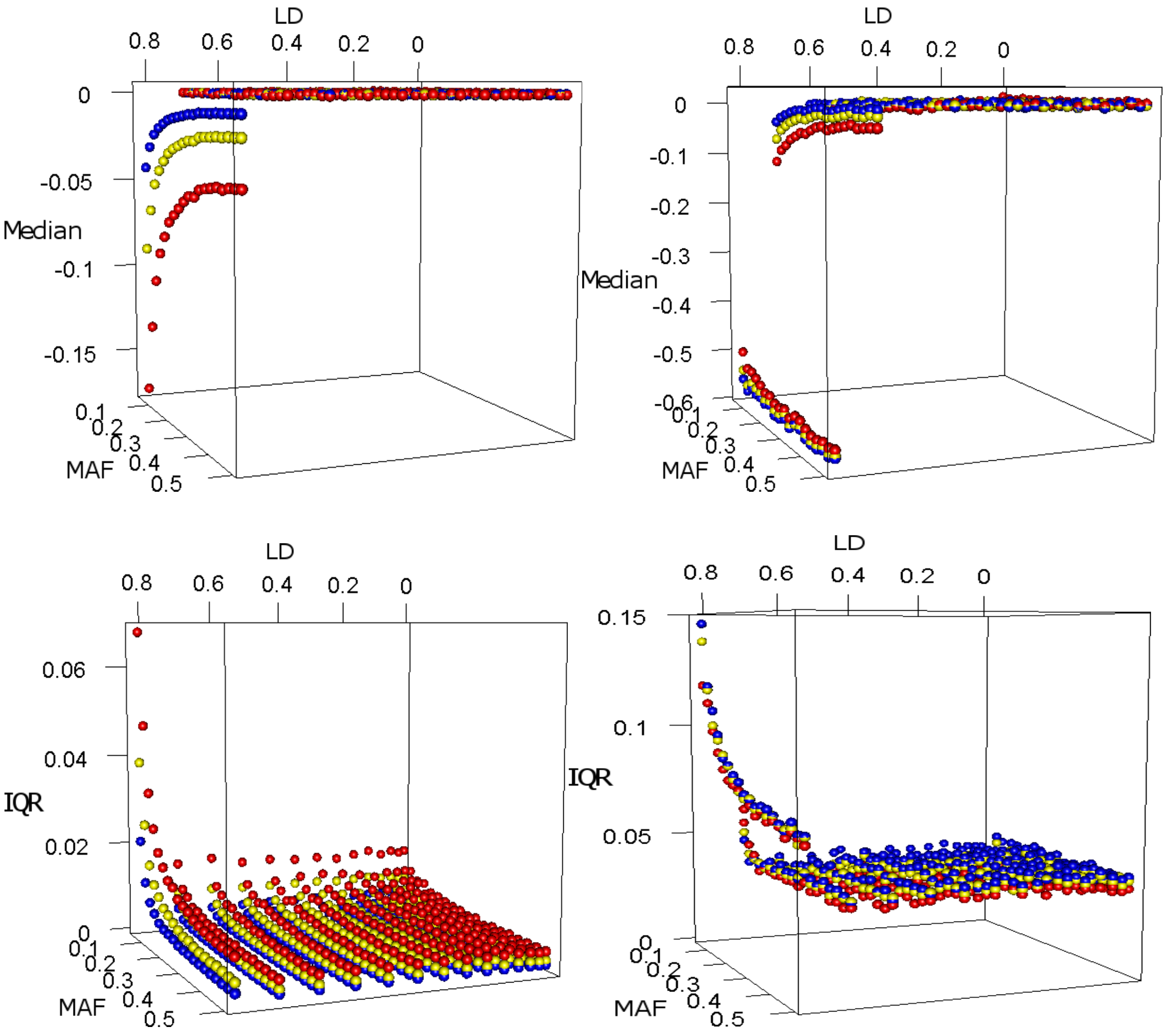


Figure 5.1. Medians of correlation of estimation errors (upper panel) in MMR (left), correlation in predictive errors in LMM (right) and corresponding IQR (lower panel). Red filled points refer to scenarios with heritability equal to 0.3, yellow filled points refers to scenarios with heritability equal to 0.5 and blue filled points refers to scenarios with heritability equal to 0.7.

In the upper panel are the medians of correlation of errors in prediction and in the lower panel are the corresponding interquartile range (IQR) that help visualize the dispersion of the data points. Clearly, the MAF in simulated marker sets influences the medians and the IQR of correlations among errors: larger values were observed for smaller MAF. About 95% of the correlation coefficients ranged from -0.03 to -0.18 in the MMR model, and from -0.25 to -0.8 in the LMM. Thus, LMM is strongly influenced by the high amount of LD in the marker data. Wang et al. (1998) reported the ability of LMM to capture not only the main effect QTLs, but rather estimates for epistatic and the gene-environment interaction effects are obtained. However, the marker data set used for these studies consist of a few hundreds markers and the amount of LD and related difficulty based on redundant information from markers was not relevant.

In all models, no impact of LD was detected on the estimates and predictions of marker effects as long as the amount of LD did not exceed $r^2 \approx 0.6$ level. Depending on the model, LD above a model specific limit had a noticeable adverse effect on estimates and predictions and led to a loss in precision. In MMR this negative impact was more pronounced for traits with moderate to low heritability, like the productive or fitness traits (e.g. milk yield, litter size or hatchability). Obviously, the extent of LD influenced the precision of estimates much more strongly in the lower MAF scenarios in all three models; also the threshold for the extent of harmful LD increased with MAF. The impact of allele frequencies in the MMR, and in the LMM was in the same range, level of LD in the data influenced estimates less severely for common variants (threshold for harmful LD at $r^2 \approx 0.8$) and more severely for MAF=0.05 (threshold for harmful LD at $r^2 \approx 0.6$).

The intensity of dispersion was also clearly lower for common variants compared to low MAF data sets. In MMR the averaged IQR was larger for traits with moderate to low heritability, while in LMM the dispersion was in general larger than in MMR, albeit the heritability of the trait had no clear impact on IQR.

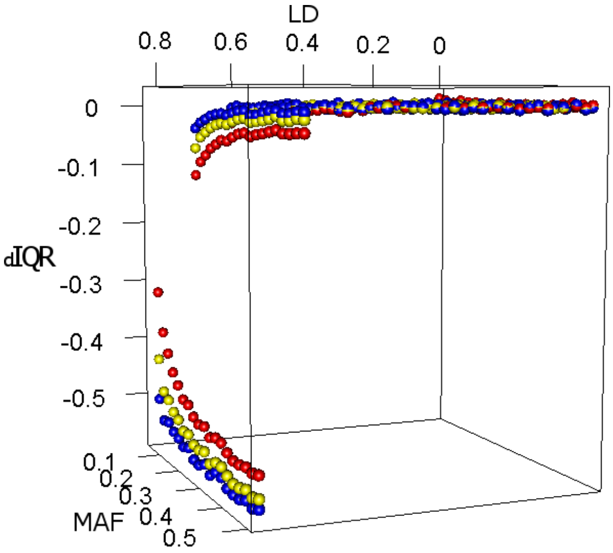


Figure 5.2. Comparison of performance in MMR and LMM. Averaged differences $dIQR = IQR\left(\mathrm{Cor}(\hat{u}_1 - u_1, \hat{u}_j - u_j) - \mathrm{Cor}(\hat{\beta}_1, \hat{\beta}_j)\right)$ in inter quartile ranges (IQR) of correlations of predictive errors in LMM and estimates from MMR. Red filled points refers to scenarios with heritability equal to 0.3, yellow filled points refers to scenarios with heritability equal to 0.5 and blue filled points refers to scenarios with heritability equal to 0.7.

A strong impact of allele frequency of markers on goodness of fit was observed with all considered models. Generally, the magnitude of MSE in LMM and MMR models was comparable, whilst the amount of MSE in the SMR model was up to ten times higher.

MMR provided more reliable results compared to LMM and SMR and seems to be an appropriate approach for performing analysis in dense marker data sets. However, the main limitation of MMR that inhibits its application as a QTL mapping tool still remains the restriction that the number of explanatory variables must be smaller than the sample size.

Is too little LD in marker data adverse for performance of genomic models?

Whole-Genome Regression (WGR) methods (Meuwissen et al., 2001), where phenotypes are regressed on all markers simultaneously, are widely used for prediction of traits of interest. The predictive performance of WGR methods when used for the prediction of phenotypes in distantly related individuals was investigated in our studies. The factors influencing prediction accuracy of WGR, such as trait heritability, marker density, the genetic architecture of the trait, the extent of LD between markers and QTL, the sample size and the method used have been intensively investigated and described in literature (Crossa et al., 2010; Erbe et al., 2013; Wimmer et al., 2013;Gusev et al., 2013; Speed et al., 2012). In most of the available studies, family data from populations with intensive history of recent selection have been used. The accuracy of prediction depends on degree of relatedness between the individuals in the training data set and the new individual (Gao et al., 2013), especially if the method is able to capture the relatedness patterns in the sample. Gao et al. (2013) evaluated performance of five Bayesian methods and GBLUP for genomic predictions of milk, fat, protein, fertility and mastitis, applied to a Nordic Holstein high density marker data set. Four training data sets were considered, which differ in the degree of relatedness between the training and testing data sets. The influence of different methods and degree of relatedness was investigated, however the impact of different genetic architectures was not considered in these studies.

In data from less related individuals, there is a lack of within-family disequilibrium (Muir and Aggrey, 2003) due to lack of relatedness in the training data set. We examined the factors that affect the prediction accuracy of WGRs using human data from distantly related individuals, considering the impact on missing heritability and on prediction accuracy of: (a) the extent of LD between markers and QTL, (b) the complexity of the trait architecture, and (c) the statistical model used (Bayes A, Spike-Slab and two Genomic Best Linear Unbiased Predictor (GBLUP) methods).

In case only QTLs are used for the analysis, thereby without disturbing noise from numerous markers without effects, the prediction accuracy with the GBLUP was as good as those with Bayes A and Spike-Slab and the correlation between the true and predicted phenotype was on average, about 0.45. In the remaining scenarios, GBLUP performed the poorest and its performance was not affected by the architecture of the trait. Bayes A and Spike-Slab performed clearly better than the GBLUP, when the trait complexity decreases and a small number of QTL explained the genetic variance. However, as the trait architecture became more complex, no differences between the methods were detected: all methods performed equally poorly.

The results achieved in this study have several implications. Firstly, estimates of missing heritability derived from data sets consisting of unrelated individuals using WGR methods need to be treated with caution. Although those estimates are indicative of how imperfect LD between markers and QTL can limit the ability of a model to capture genetic signals, they also indicate that under some circumstances estimates can have a sizeable bias. Additionally, we observed that in some scenarios these estimates of heritability can vary significantly between methods. This is not surprising because the proportion of variance explained by a model depends both on the input information (markers/QTL, etc.) and on the statistical model used. This inter-dependency between model used and present genetic architecture a trait has been over-looked so far. For instance, Krag et al. (2013) evaluated estimation of heritability of two Bayesian and one restricted maximum likelihood methods, performing extensive simulation studies. Simulation scenarios, reflecting different marker densities and population structures, for heritability varying between 0.05 and 0.5 were performed in this study, whereas the number of QTL was fixed across all scenarios. Importantly, the model that yielded highest estimated genomic heritability is not necessarily the one that yielded the best prediction accuracy. Thus, none of genomic methods is generally applicable, however a suitable method might be chosen for each specific question, depending on the type of genomic data available for the analysis.

The prediction accuracy of Spike-Slab model and Bayes A was significantly higher than the GBLUP; the superiority of the Spike-Slab over Bayes A was also systematic, but very small in magnitude, which suggests that this implementation should be the approach of choice for quantitative genetic analysis, particularly for the traits with unknown genetic architecture.

Furthermore, the computational time of the Spike-Slab implementation used in our studies (Zhou et al., 2013) was about 10 - 12 hours, which is four times faster than that for Bayes A (computational time of 2 days). The main limitation of this implementation is the restriction on the size of data. In our case the software was not able to cope with more than 400K markers for 5,758 individuals.

One way to improve prediction accuracy using data from less related individuals, is the utilization of sequence data. In this way, some two-step estimation procedures, where a subset of influential markers is chosen in the first step and used as weights in the second step, estimates of marker effects are obtained (e.g. de los Campos et al., 2013a; Zhang et al., 2014). Apart from that, the key aspect of the next-generation sequencing is the ability to simultaneously sequence millions of DNA fragments. The large amount of additional genomic information can be used not only as a source of a larger number of SNPs, but also as a source of insertions or deletions. For the present study, this novel source of genomic information was not available. In general, sequence data are still very expensive and are not available in all species. A further difficulty of using sequence data for the estimation of effects and predictions is the small sample sizes; this is expected to affect the factors investigated in the present work to quite an extent.

Real analysis for an additional data set: hopes and reality

The results achieved for human height using GENEVA data set were very close to the results from the simulation for infinitesimal model scenarios with different distribution of MAF in markers and QTLs. Human height is believed to be a trait affected by a very large number of small-effect QTL (e.g. Allen et al., 2010; Yang et al., 2010). We estimated a sizeable proportion of missing heritability and obtained very similar, albeit poor, prediction accuracies across methods (correlation of about 0.16 - 0.17). Thus, for very complex traits such as human height, all the evaluated methods yielded low prediction accuracy.

Real analysis of a trait with a simple genetic architecture may confirm the results from simulation studies for scenarios where a small number of variants have impact on the trait. For this reason we were looking for a data set with phenotypic records for traits which may be influenced by a small number of genes. In the GENEVA data set most records are ordinal or nominal variables, based on questionnaires, thus not suitable for performing quantitative analysis with WGR. However, some appropriate traits seem to be included in the British Cohort 1958 data set (BC58), which consists of records of unrelated individuals born in one week in March 1958. Between September 2002 and December 2003 a follow-up biomedical survey of 9,377 individuals was undertaken (Power and Elliott, 2006). To a large extent, the traits recorded in the biomedical survey are nominal or ordinal variables, achieved using questionnaires. Thus, these records are less appropriate for the genomic estimation and prediction when applying GBLUP, Bayes A or Spike-Slab. After a thorough search, five metric variables were chosen that are available in BC58 data set: the growth factor 1 (IGF1), total cholesterol (CHOL), high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL) and triglyceride (TRIG) as phenotypes for an additional analysis. We used a subset of n = 2,997 individuals, genotyped with Affymetrix Genome-Wide Human SNP v6.0

DNA Array, after quality control p=737,837 SNPs remained for the analysis. The analysis was performed using the Spike-Slab implementation of (Zhou et al., 2013), which has been shown to be the best and fastest approach. Figure 5.3 shows the correlations between the true and estimated phenotypes and the estimates of heritability for the above-mentioned traits, averaged over 30 training-testing partitions.

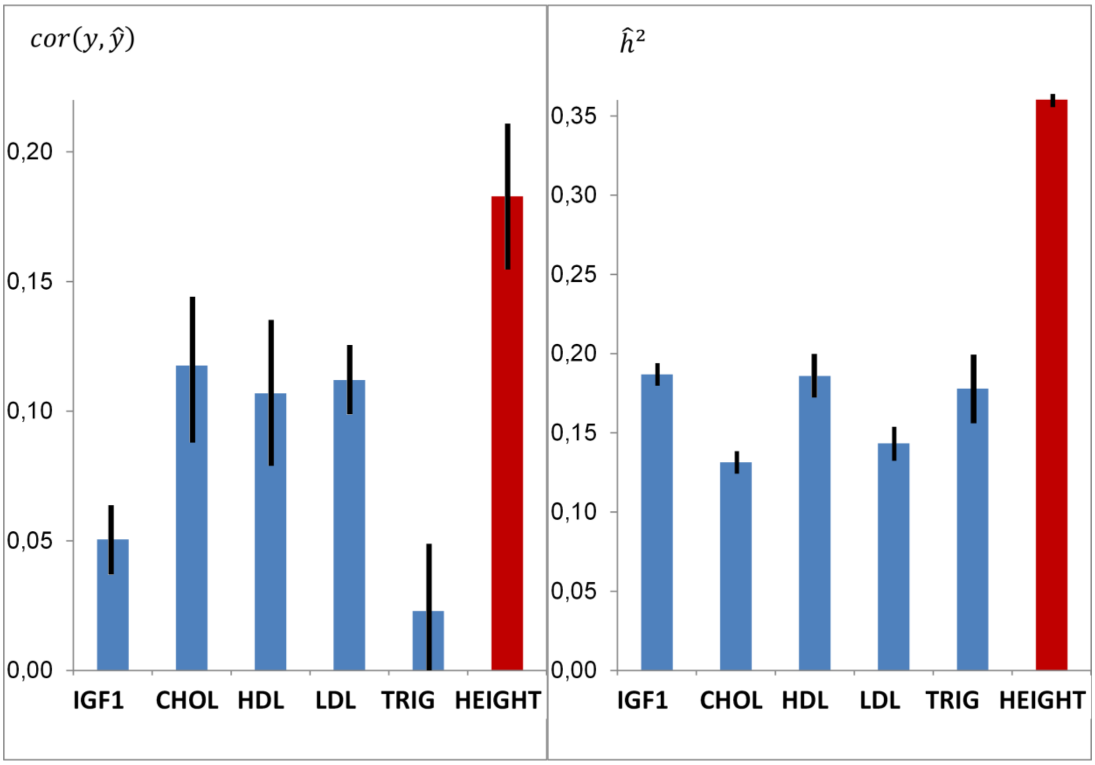


Figure 5.3. Correlation between phenotypes and genomic predictions in the BC58 data set (blue) and in the GENEVA data set (red). Correlation (averages over 30 replicates and corresponding standard errors) between phenotypes and genomic predictions using the Spike-Slab method.

The hopes to confirm the observations from the simulation scenarios with traits of a less complex architecture were not fulfilled: the accuracy of predictions for new traits ranged from 0.02 to 0.12, which was on average weaker than that for human height (average at 0.17) for individuals from GENEVA. This does not necessarily mean that these traits do not have the desired genetic architecture. We know from our studies presented in the second chapter that heritability of a trait has a strong impact on the performance of genomic approaches. The heritability estimates of CHOL and LDL were the lowest of all BC58 traits and were on average between 0.13 and 0.14, while the heritability estimates of IGF1, HDL and TRIG were very similar with values around 0.18. Thus the heritabilities of traits from BC58 are at least two times smaller than the heritability estimates of human height in GENEVA data set and also smaller than the heritability of the phenotypes ($h^2 = 0.5$) in simulation scenarios. Even if these traits might be influenced by a small number of genes, it would be difficult to determine the differences in performance of methods due to general low prediction accuracy.

Does the parameter choice or length of MCMC chains in Bayesian analyses bring our results into question?

The Bayesian methods applied in our studies on prediction accuracy in unrelated individuals are widely used in animal and plant breeding. The crucial point in the application of Bayesian WGR methods is the choice of priors and specification of hyperparameters. This point is intensively discussed in the scientific literature (e.g. Gianola, 2013). Lehermeier et al. (2013) reported a strong impact of the choice of hyperparameters in Bayesian methods, although the impact of chosen prior is reduced by increasing sample size. Thus we decided to perform sensitivity analysis in order to examine how the change in the prior parameters influences the predictive ability.

In the BGLR-package used for the analysis of simulated and real data, GBLUP is implemented as a Bayesian Reproducing Kernel Hilbert Spaces Regressions (RKHS) with a Gaussian kernel, where a scaled-inverse χ^2 density is assigned to the variance parameters. The default degree of freedom is set to *df=5*, which gives a relatively un-informative prior and should guarantee a finite prior variance. We performed analysis with *df=15*, predicted for the same testing-training data sets (TST-TRN) partitioning and calculated correlation between predictions from both setting: the correlations in both training and testing data sets were >0.99, showing that predictions were not sensitive to the choice of the degrees of freedom in the RKHS implementation in BGLR.

For BGLR-implementations of Bayes A and GBLUP, we performed 50,000 MCMC iterations, whereby the first 10,000 iterations were considered as a “burn in” phase of the sampling algorithm and consequently discarded from the posterior distribution sampling. In the GEMMA software, used for performing the Spike-Slab model, default number of MCMC iterations is set to 1,000,000 which seems to be much too high. Thus, we reduced the number of iterations to 100,000. A convergence diagnostic carried out for all methods using the R-package *coda* (Plummer et al., 2010), which deliver detailed summary statistics of all marginal posterior distributions as well as traceplots and kernel density plots of all variables enabling the visual control of convergence behaviour. Furthermore, we performed sensitivity analyses to examine the convergence behavior of the algorithms for the different numbers of iteration (*nIter*): GBLUP, Bayes A and Spike-Slab predictions in different simulation scenarios were obtained and visualized in Figure 5.4.

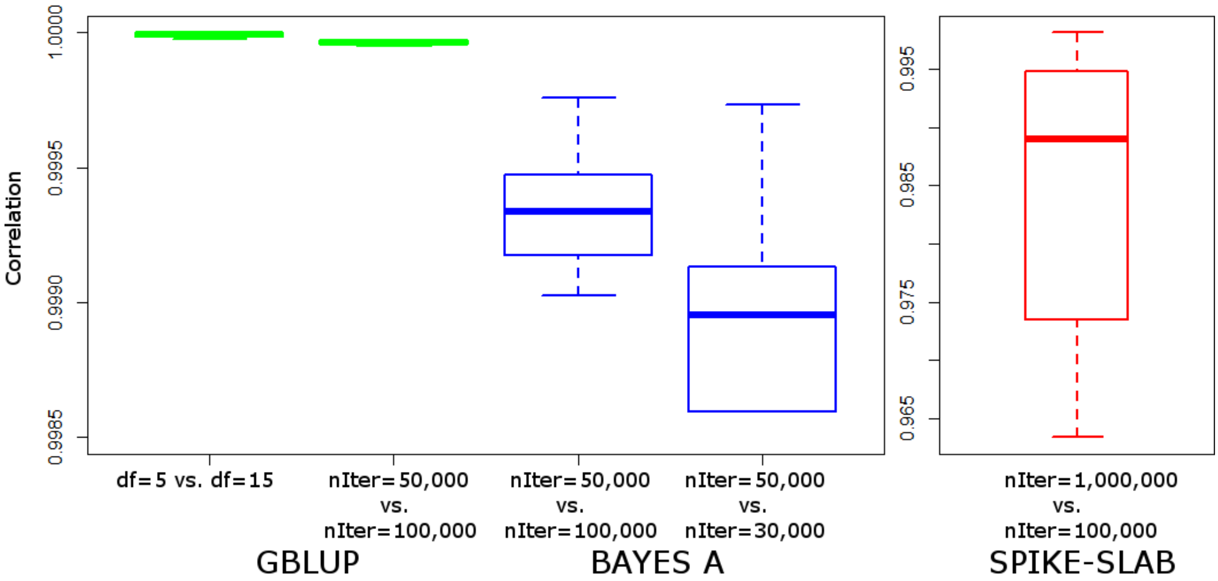


Figure 5.4. Correlation between predictions in BC58 data set and in GENEVA data set for different hyperparameter. Averaged correlation between genomic predictions obtained in GBLUP (green), Bayes A (blue) and Spike-Slab (red) with default and deviating values of hyper parameters: degree of freedom (df) and number of iterations (nIter).

For all methods, the correlations achieved from sampling algorithms with different numbers of iterations were relatively high and did not vary across simulation scenarios. In Bayes A the correlation between the predictions performed using 100,000 or 50,000 iterations was $Cor(\hat{y}_{100,000}, \hat{y}_{50,000}) = 0.9994 \pm 0.0003$ and in GBLUP for the same settings, $Cor(\hat{y}_{100,000}, \hat{y}_{50,000}) = 0.9993 \pm 0.0001$. In Spike-Slab the correlation between predictions achieved using default *nIter*=100,000 or *nIter*=1,000,000 was $Cor(\hat{y}_{1,000,000}, \hat{y}_{100,000}) = 0.984 \pm 0.013$ and thus the lowest of all. Nevertheless, the concordance in predictions was high and we decided to keep the chosen parameters.

To what extent does the degree of association between loci differ between genic and non-genic regions?

In Chapter 4 a comparison method was developed which copes with difficulties arising while performing comparison of LD levels between different genomic regions such as the impact of the extent of compared regions on the genome (spatial bias) and the impact of allele frequencies on LD (MAF caused bias). The differences in LD structure between genic and non-genic regions in human, chicken and arabidopsis were examined using this method. In the first step, similar pairs from the genic and non-genic regions (G/IG) were identified. Applying the Wilcoxon signed rank test, we detected significant higher LD level in genic regions on about 30% of chromosomes in human (*H. sapiens*) and in chicken (*G. g. domesticus*), while in arabidopsis (*A. thaliana*) about 20% higher LD in genic regions was observed on all chromosomes. As control, comparisons of pairs of similar non-genic regions (IG/IG') were performed and, as expected, no significant differences between those regions were discovered. Even on a genome-wide level, significantly more LD was observed in genic regions from all three species; thus the observations of higher LD in genic regions by Eberle et al. (2006) were confirmed and quantified.

The LD levels at very short physical distances were similar in *A. thaliana* and *H. sapiens* with r^2 being about 0.25 in average. However in *A. thaliana* a clear gap between LD amount in genic and non-genic regions was registered in that region while in *H. sapiens* almost no G/IG difference was recorded up to a distance of about 50 kilo base pairs. Why are the LD spans so short and why are genic regions more conserved in *A. thaliana* compared to humans? *A. thaliana* is a globally distributed plant and the sample used in our studies consists of inbred lines. This sample has a complex population structure and a very large effective population size which may explain the rapid decay of LD (Kim et al., 2007). In general, LD in plants vary depending on the choice of a population (Flint-Garcia et al., 2003): for instance, in barley Caldwell et al. (2006) reported $r^2 = 0.2$ at a distance of about 212 kbp.

The LD level observed in *G. g. domesticus* was twice as high as the LD level in *H. sapiens* and decay was much slower than in humans. This higher LD level was observed in *G. g. domesticus* over all distances: the white layer data originated from a commercial breed that has been intensively selected for egg laying. Thus the degree of relatedness among those individuals was relatively high. The magnitude of relatedness in the population had a strong impact on the effective population size, which is very low in commercial lines of chicken (Qanbari et al., 2010; Li et al., 2012). Thus, it is not surprising that the individuals share long sequences of chromosomes and the total amount of LD in populations from breeding programs is relatively high. The natural decay of LD occurs at slower rate due to stronger and directed selection pressure.

A framework that accounts for spatial and structural differences in genomic regions for comparing genic and non-genic regions gave us new insights into the dependency of LD levels on size of chromosomes or regions. In contrast to findings of Smith et al. (2005) and Uimari et al. (2005), we did not observe weaker LD in the small chromosomes and stronger

LD in the large chromosomes. Across all species, the extent of LD measured in genic or non-genic regions does not depend on the size of the chromosome. These discrepancies from previous studies may be caused by lower marker density, lower SNP call rates (>80%) or smaller sample sizes in older studies. Differences detected in studies of Smith et al. (2005) and Uimari et al. (2005) may also be caused only by spatial differences or different distribution of allele frequencies. In order to gain a deeper insight into the relationship between LD and size of genomic regions, a detailed analysis in the chicken data set was performed: linear regression of the medians of both considered LD measures was performed against the size of genic and non-genic regions. Although for both LD measures the slopes of regression curves were negative and differed significantly from zero, all absolute values were very tiny and could be ignored. The differences in G/IG comparison did not depend on the size of regions at all.

The results of significance tests of haplotype diversity confirmed our observations of differences in LD levels: significantly less diversity of haplotypes in genic regions was noticed for all species. One possible reason may be the interferences of the molecular mechanisms responsible for survival of an organism and the resulting damage of vital processes. Another reason for more conserved variants in genic regions might be connection to the fertility disrupters (e.g. Naz, 1999; Anway et al., 2005) in case of recombination in genic regions, which affect productivity capacity of living organisms. In such cases affected individuals are no longer available in the parental gene pools.

Main Conclusions

The presence of LD complicates modelling of genomic data, since in many models the assumption of independence of explanatory variables plays a central role. A unique solution for effect estimates is impossible if this restriction to the data is violated and the reliability of the marker effect estimates in different models is reduced. An increase in estimation errors was recorded if the LD level between the loci increased. According to Günther et al. (2011), SNPs located in genes and in particular in introns are significantly more frequently detected by GWAS. In combination with higher LD in genic regions, the precision of marker effect estimates for markers in those regions is seriously affected.

The assessment of prediction accuracy suggests that for traits in which a limited number of regions explain a sizeable proportion of genetic variance, the use of WGR methods that perform variable selection or differential shrinkage of estimates of effects is strongly recommended over ridge-regression type methods such as the GBLUP. On the other hand, for very complex traits such as human height all the methods evaluated yielded low prediction accuracy. It remains to be determined whether significant increases in sample size (which likely should be by orders of magnitude) will also yield substantial gains in prediction accuracy.

The strategy we proposed to account for scale effects in LD comparisons of different genomic regions proved to be efficient: using a haplotype based measure r^2 we determined significantly higher extent of LD in genic regions compared to non-genic regions. In all probability, this is a general phenomenon since it was observed in the human, animal (chicken) and plant (arabidopsis) data sets we studied. Additional studies, especially the comparisons of different regions of the genome (coding, non-coding), are needed to confirm and refine our results. However, some issues pertaining to the nature of LD were identified and need further discussion. In particular, simulation studies based on related individuals for investigating the impact of LD level on single SNP effect might give new insights.

The results of our studies indicate a strong impact of high LD between the markers on estimates of random marker effects in linear models. These results are especially relevant for the estimation of marker effects in animal and plant breeding, where the populations consist of closely related individuals and consequently the LD amount in the data is very high. In our studies we observed that 30% of SNP pairs $r^2 \geq 0.60$ and about 10% of SNP pairs $r^2 \geq 0.80$ in a data set of a highly selected White Leghorn chicken, which might be crucial for the precision of estimates for a substantial part of markers. The degree of relatedness between the individuals in the sample, have been shown to have a strong impact on prediction accuracy in particular for such methods as GBLUP, which is able to capture the relatedness patterns in the sample. Thus, the differential shrinkage methods like Bayes A and variable selection methods like the Spike-Slab model have proven to be more robust and reliable if there is a lack of within-family disequilibrium due to lack of relatedness in the training data set.

Availability of high-density marker data set in many species and related increase of LD amount in data, which is an advantage on the one hand, is an inconvenience on the other: the prediction accuracy in samples of less related individuals could be improved, while the estimates of maker effects would lose their precision. In this context, we provide a powerful tool for comparison of LD in different genomic regions, taking into account scale differences.

REFERENCES

Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., and Raychaudhuri, S. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

Amaral, A.J., Megens, H.-J., Crooijmans, R.P., Heuven, H.C., and Groenen, M.A. (2008). Linkage disequilibrium decay and haplotype block structure in the pig. Genetics *179*, 569–579.

Anway, M.D., Cupp, A.S., Uzumcu, M., and Skinner, M.K. (2005). Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science *308*, 1466–1469.

Ardlie, K.G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. *3*, 299–309.

Caldwell, K.S., Russell, J., Langridge, P., and Powell, W. (2006). Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, Hordeum vulgare. Genetics *172*, 557–567.

Crossa, J., de los Campos, G., Pérez-Rodrigues, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., and Yan, J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics *186*, 713–724.

Daetwyler, H.D., Swan, A.A., van Der Werf, J.H., and Hayes, B.J. (2012). Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genet. Sel. Evol. *44*, 33.

Eberle, M.A., Rieder, M.J., Kruglyak, L., and Nickerson, D.A. (2006). Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet. *2*, e142.

Van Eenennaam, A.L., Weigel, K.A., Young, A.E., Cleveland, M.A., and Dekkers, J.C. (2014). Applied Animal Genomics: Results from the Field. Annu Rev Anim Biosci *2*, 105–139.

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

Fan, B., Du, Z.-Q., Gorbach, D.M., and Rothschild, M.F. (2010). Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Aust J Anim Sci *23*, 833–847.

Flint-Garcia, S.A., Thornsberry, J.M., and IV, B. (2003). Structure of Linkage Disequilibrium in Plants*. Annu. Rev. Plant Biol. *54*, 357–374.

Gao, H., Su, G., Janss, L., Zhang, Y., and Lund, M.S. (2013). Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. J. Dairy Sci. *96*, 4678–4687.

Georges, M. (2007). Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. Annu Rev Genomics Hum Genet *8*, 131–162.

Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics. 194.3 (2013): 573-596

Goddard, M.E., and Hayes, B.J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. *10*, 381–391.

Günther, T., Schmitt, A.O., Bortfeldt, R.H., Hinney, A., Hebebrand, J., and Brockmann, G.A. (2011). Where in the genome are significant single nucleotide polymorphisms from genome-wide association studies located? Omics J. Integr. Biol. *15*, 507–512.

Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., et al. (2013). Quantifying missing heritability at known GWAS loci. PLoS Genet. *9*, e1003993.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. *39*, 1151–1155.

Krag, K., Janss, L.L., Shariati, M.M., Berg, P., and Buitenhuis, A.J. (2013). SNP-based heritability estimation using a Bayesian approach. Animal *7*, 531–539.

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

Li, D.F., Liu, W.B., Liu, J.F., Yi, G.Q., Lian, L., Qu, L.J., Li, J.Y., Xu, G.Y., and Yang, N. (2012). Whole-genome scan for signatures of recent selection reveals loci associated with important traits in White Leghorn chickens. Poult. Sci. *91*, 1804–1812.

Megens, H.-J., Crooijmans, R.P., Bastiaansen, J.W., Kerstens, H.H., Coster, A., Jalving, R., Vereijken, A., Silva, P., Muir, W.M., Cheng, H.H., et al. (2009). Comparison of linkage disequilibrium and haplotype diversity on macro-and microchromosomes in chicken. BMC Genet. *10*, 86.

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

Muir, W.M., and Aggrey, S.E. (2003). Poultry Genetics, Breeding, and Biotechnology (CABI).

Naz, R.K. (1999). Endocrine disruptors: effects on male and female reproductive systems (CRC Press).

Plummer, M., Best, N., Cowles, K., and Vines, K. (2010). Coda: output analysis and diagnostics for MCMC. R package version 0.13-4.

Power, C., and Elliott, J. (2006). Cohort profile: 1958 British birth cohort (national child development study). Int. J. Epidemiol. *35*, 34–41.

Qanbari, S., Hansen, M., Weigend, S., Preisinger, R., and Simianer, H. (2010). Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genet. *11*, 103.

Schefers, J.M., and Weigel, K.A. (2012). Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Anim. Front. *2*, 4–9.

Smith, A.V., Thomas, D.J., Munro, H.M., and Abecasis, G.R. (2005). Sequence features in regions of weak and strong linkage disequilibrium. Genome Res. *15*, 1519–1534.

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.

Stock, K.F., and Reents, R. (2013). Genomic Selection: Status in Different Species and Challenges for Breeding. Reprod. Domest. Anim. *48*, 2–10.

Uimari, P., Kontkanen, O., Visscher, P.M., Pirskanen, M., Fuentes, R., and Salonen, J.T. (2005). Genome-wide linkage disequilibrium from 100,000 SNPs in the East Finland founder population. Twin Res. Hum. Genet. *8*, 185–197.

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. Science *326*, 865–867.

Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science *280*, 1077–1082.

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H.-J., Wang, Y., and Schön, C.-C. (2013). Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. Genetics *195*, 573–587.

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

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.