Genomic prediction using linkage disequilibrium and co-segregation

Xiaochen Sun
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Agriculture Commons, Animal Sciences Commons, Genetics Commons, and the Statistics and Probability Commons

Recommended Citation
Sun, Xiaochen, "Genomic prediction using linkage disequilibrium and co-segregation" (2014). Graduate Theses and Dissertations. 14273.
https://lib.dr.iastate.edu/etd/14273

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Genomic prediction using linkage disequilibrium and co-segregation

by

Xiaochen Sun

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Animal Breeding and Genetics (Quantitative Genetics)

Program of Study Committee:
Jack C. M. Dekkers, Co-major Professor
Rohan L. Fernando, Co-major Professor
Alicia L. Carriquiry
Dorian J. Garrick
Daniel S. Nettleton

Iowa State University
Ames, Iowa
2014

Copyright © Xiaochen Sun, 2014. All rights reserved.
To Yisheng, whose love, patience and support made this possible.
# TABLE OF CONTENTS

**LIST OF TABLES** .......................................................... vii

**LIST OF FIGURES** ...................................................... ix

**ACKNOWLEDGMENTS** ................................................... xiii

**ABSTRACT** ............................................................... xv

**CHAPTER 1. GENERAL INTRODUCTION** ................................. 1

1.1 Introduction ......................................................... 1

1.2 Research Objectives ............................................... 5

1.3 Organization of Thesis ............................................. 6

1.4 Literature Review .................................................. 7

1.4.1 Statistical models and computation algorithms for genomic prediction of breeding values ...................................... 7

1.4.2 Sources of genetic information that contribute to accuracy of genomic prediction ............................................. 13

1.4.3 Factors that affect prediction accuracy .......................... 15

1.4.4 Problems in current genomic prediction that the thesis aims to address ........................................... 18

1.5 Bibliography .............................................................. 19

**CHAPTER 2. CONTRIBUTIONS OF LINKAGE DISEQUILIBRIUM AND CO-SEGREGATION INFORMATION TO THE ACCURACY OF GENOMIC PREDICTION IN HALF-SIB DESIGNS** ................. 30

2.1 Abstract ............................................................... 30

2.2 Introduction .......................................................... 31

2.3 Materials and Methods .............................................. 33
CHAPTER 3. GENOMIC PREDICTION OF BREEDING VALUES IN DIFFERENT MATING DESIGNS COMBINING LINKAGE DISEQUILIBRIUM AND CO-SEGREGATION

3.1 Abstract ......................................................... 59
3.2 Introduction ..................................................... 60
3.3 Materials and Methods .......................................... 62
  3.3.1 Simulation of three mating designs ......................... 62
  3.3.2 Prediction of breeding values ............................. 65
3.4 Results .......................................................... 65
3.5 Discussion ....................................................... 67
  3.5.1 Simulated mating designs ................................. 67
  3.5.2 The effect of MAF on contributions of LD and CS information ................................. 69
  3.5.3 The effect of prior distributions on prediction accuracy ................................. 70
3.6 Conclusions ...................................................... 71
3.7 Acknowledgments ................................................ 72
3.8 Bibliography ..................................................... 72
3.9 Tables ............................................................ 76
3.10 Figures .......................................................... 77
CHAPTER 4. IMPROVED ACCURACY OF GENOMIC PREDICTION
FOR TRAITS WITH RARE QTL BY FITTING HAPLOTYPES . . . . . . . . 81

4.1 Abstract ................................................. 81
4.2 Introduction ........................................... 82
4.3 Materials and Methods ................................. 83
   4.3.1 Simulated datasets ................................. 83
   4.3.2 Statistical analyses ................................. 84
4.4 Results and Discussion ................................. 85
   4.4.1 Haplotype frequencies and concordance between SNP haplotypes and
         QTL alleles ............................................ 85
   4.4.2 Prediction accuracy ................................ 86
4.5 Conclusion .............................................. 87
4.6 Acknowledgments ....................................... 88
4.7 Bibliography ............................................ 88
4.8 Tables .................................................. 90

CHAPTER 5. GENOMIC PREDICTION COMBINING LINKAGE DISEQUILIBRIUM AND CO-SEGREGATION IN POPULATIONS WITHOUT PEDIGREE ......................... 91

5.1 Abstract ................................................. 91
5.2 Introduction ........................................... 92
5.3 Materials and Methods ................................. 95
   5.3.1 Simulated datasets ................................. 95
   5.3.2 Dataset of layer chickens from a commercial breeding population . . . 96
   5.3.3 Genomic prediction using haplotypes .................... 96
5.4 Results .................................................. 98
   5.4.1 Number of haplotypes and concordance with QTL alleles ............... 98
   5.4.2 Prediction accuracy in the simulated datasets ....................... 99
   5.4.3 Prediction accuracy in layer chickens ....................... 101
5.5 Discussion ............................................. 101
5.5.1 Genetic information exploited by the SNP, haplotype, and SNP-haplotype models ........................................ 102
5.5.2 Prediction accuracy with historical LD .................................................. 103
5.5.3 Prediction accuracy without historical LD ................................................. 104
5.5.4 The effect of fitting only common haplotypes ........................................... 105
5.5.5 The effect of SNP density ................................................................. 105
5.5.6 Results from the layer chicken dataset .................................................. 106
5.6 Conclusions .............................................................................................. 106
5.7 Acknowledgments ...................................................................................... 107
5.8 Bibliography .............................................................................................. 107
5.9 Tables ......................................................................................................... 115
5.10 Figures ....................................................................................................... 117

CHAPTER 6. GENERAL DISCUSSIONS AND CONCLUSIONS .................... 129

6.1 General Discussions ................................................................................ 129
6.1.1 Research objectives ............................................................................ 129
6.1.2 Contributions of LD and CS information to prediction accuracy .... 129
6.1.3 CS information captured by the CS and haplotype models ............. 131
6.1.4 Modeling LD and CS explicitly for field datasets ......................... 131
6.1.5 Future work ......................................................................................... 133
6.2 Conclusions .............................................................................................. 134
6.3 Bibliography .............................................................................................. 135

APPENDIX A. A FAST EM ALGORITHM FOR BAYESA-LIKE PREDICTION OF GENOMIC BREEDING VALUES .......... 138

APPENDIX B. GENOMIC BREEDING VALUE PREDICTION AND QTL MAPPING OF QTLMA S2010 DATA USING BAYESIAN METHODS . 165
LIST OF TABLES

Table 2.1  Simple pedigree and SNP allele origins of 4 individuals.       51
Table 2.2  Coefficient matrix of the CS model for the pedigree in Table 2.1.  51
Table 2.3  Scenarios in the simulated half-sib designs.                   51
Table 3.1  Minor allele frequencies (MAF) of QTL and SNPs and the level of historical LD in the base population of simulated scenarios. 76
Table 4.1  Average number of unique (No. Allele) and common (No. Common) haplotype alleles, and the proportion of discordant (Discordant %) haplotype alleles across all genome segments in the scenario of common QTL and common SNPs.         90
Table 4.2  Mean prediction accuracies across 20 replicates.              90
Table 5.1  Average number of genotyped individuals with own phenotypic records in 9 generations of the layer chicken dataset.  115
Table 5.2  Average number of haplotype alleles and concordance of haplotypes with QTL alleles across 1-cM genome windows in three simulated mating designs.                                         115
Table 5.3  Average number of haplotype alleles and concordance of haplotypes with QTL alleles across 1-cM genome windows under different SNP densities and historical LD among SNPs in Pedigree 3.                                      116
Table A.1  Summary of simulated datasets and scenarios.                157
Table A.2  Accuracy of GEBV and regression coefficient of TBV on GEBV in the first validation generation of Dataset A for GBLUP, BayesA, BayesB and fastBayesA. ................................. 158

Table A.3  Computing time (in seconds) for training by BayesA, BayesB and fastBayesA. ................................................................. 158

Table B.1  Prediction accuracy of GEBV, correlation of GEBV with TBV, correlation of GEBV with genotypic value ($g$), regression of phenotype ($y$) on GEBV, regression of TBV on GEBV, and regression of genotypic value on GEBV. ................................. 180

Table B.2  Average number of SNPs (#SNP) fitted in the model, estimated variance components, and estimated heritability (Heritability). ................................................................. 181

Table B.3  Variance components estimated from datasets generated by permutation, simulation with linkage equilibrium in founders (LE simulation), and simulation with initial linkage disequilibrium (LD simulation), and thresholds for 10-SNP window variances based on 10% and 20% chromosome-wise type I error rates. ................................................................. 182
LIST OF FIGURES

Figure 2.1  Prediction accuracy with different numbers of half-sib families in training with 100 common QTL per chromosome.  . . . . . . 52
Figure 2.2  Prediction accuracy with different numbers of half-sib families in training with 100 rare QTL per chromosome.  . . . . . . . 53
Figure 2.3  Prediction accuracy with different numbers of half-sib families in training with 100 resampled QTL per chromosome.  . . . . . 54
Figure 2.4  Prediction accuracy with different numbers of half-sib families in training with SNPs and 100 QTL per chromosome in LE. . 55
Figure 3.1  Mean accuracy in three simulated pedigrees from BayesA in the Common QTL scenario with high historical LD.  . . . . . . 77
Figure 3.2  Mean accuracy in three simulated pedigrees from BayesB in the Common QTL scenario with high historical LD.  . . . . . . 77
Figure 3.3  Mean accuracy in three simulated pedigrees from BayesA in the Rare QTL scenario with high historical LD.  . . . . . . . 78
Figure 3.4  Mean accuracy in three simulated pedigrees from BayesB in the Rare QTL scenario with high historical LD.  . . . . . . . 78
Figure 3.5  Mean accuracy in three simulated pedigrees from BayesA in the Common QTL scenario with no historical LD.  . . . . . . 79
Figure 3.6  Mean accuracy in three simulated pedigrees from BayesB in the Common QTL scenario with no historical LD.  . . . . . . 79
Figure 3.7  Mean accuracy in three simulated pedigrees from BayesA in the Rare QTL scenario with no historical LD.  . . . . . . . . 80
Figure 3.8  Mean accuracy in three simulated pedigrees from BayesB in the Rare QTL scenario with no historical LD.  

Figure 5.1  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with high historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes.  

Figure 5.2  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with high historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes.  

Figure 5.3  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with high historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes.  

Figure 5.4  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with high historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes.  

Figure 5.5  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with no historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes.  

Figure 5.6  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with no historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes.
Figure 5.7  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with no historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes. ........................................ 123

Figure 5.8  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with no historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes. ........................................ 124

Figure 5.9  Mean accuracy across 8 validation generations without retraining in pedigree 3 in the Rare QTL scenario with high historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes. ........................................ 125

Figure 5.10 Mean accuracy across 8 validation generations without retraining in pedigree 3 in the Rare QTL scenario with high historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes. ........................................ 126

Figure 5.11  Prediction accuracy in 4 validation generations for three early egg quality traits in layer chicken datasets using BayesA (left) or BayesB (right) by fitting all haplotypes. ................................. 127

Figure 5.12  Prediction accuracy in 4 validation generations for two late egg quality traits in layer chicken datasets using BayesA (left) or BayesB (right) by fitting all haplotypes. ................................. 128

Figure A.1  Prediction accuracy of GEBV in five validation generations by alternative methods. ......................................................... 159

Figure A.2  Prediction accuracies of GEBV across EM iterations in the first validation generation. ......................................................... 160
Figure A.3  Shrinkage estimate of SNP effects from ridge regression (black line), BayesA (red dots), fastBayesA (blue line) and VanRaden non-linear prediction A (green line) against least squares estimate. ............................. 161

Figure A.4  Estimated SNP effects from fastBayesA (y axis) against estimates from BayesA and BayesB (x axis). ............................. 162

Figure B.1  Variances of GEBVs of 10-SNP windows across the genome. ............................. 183
ACKNOWLEDGMENTS

Foremost, I would like to express my wholehearted gratitude to my major advisors and friends, Drs. Jack Dekkers and Rohan Fernando. I am thankful to Jack for bringing me into the exciting area of genomic selection. Under his instruction in my coursework and research, I acquired rich knowledge and a broad vision that will assist me to explore my research area freely and independently. Jack also created many opportunities to sharpen my professional skills in teaching, communication, cooperation, and prioritizing multitasks, from which I will no doubt benefit in my entire career. I am greatly influenced by Rohan with his critical and meticulous attitude towards science. I have been inspired and motivated numerous times in the discussions and arguments between us. In addition, Rohan has always been encouraging and helping me to pursue my career goals by sharing with me his rich academic experiences. I greatly appreciate the tremendous contributions that Jack and Rohan made to the writing of this thesis.

My deepest gratitude also goes to the rest of my Ph.D. committee members. I am greatly thankful to Dr. Dorian Garrick, for his resourceful suggestions and incisive criticisms to many of my research topics. I would also like to thank Dr. Alicia Carriquiry, who has always been responsive and supportive despite her busy schedules. I am especially thankful to Dr. Dan Nettleton. Throughout my study of statistics I have been continuously benefited from his unique vision in linear model theory. I am also grateful to his interests and insightful discussions in many of my research topics. I feel fortunate to have such a precious opportunity to work with remarkable people at Iowa State University.

Further, I would like to thank Dr. Long Qu, for his tremendous inputs to leverage the quality of many statistical methods in my research. I would like to thank Dr. David Habier, for his detailed instruction in the development of software packages using the C++ language. I would also like to thank Dr. Ania Wolc, who has always been amazingly helpful in various
aspects of my research. Without the help from Long, David and Ania, it would be impossible for me to obtain the crucial statistical and programming skills to accomplish my research objectives.

Besides, I thank my colleagues, fellow students and officemates, Dr. Ali Toosi, Dr. Wei He, Dr. Mahdi Saatchi, Dr. Hailin Su, Dr. Nick Serao, Dr. Dawn Koltes, Dr. Dinesh Thekkoot, Jian Zeng, Hao Cheng, Ziqing Weng, Emily Waide, Melanie Hayr, etc. I am lucky and grateful to be a member of this energetic and friendly group – the Animal Breeding & Genetics.

Finally, with my whole heart and soul, I thank my wife, Dr. Yisheng Wu, for her complete and unconditioned love. I thank my parents for raising me with a love of science, and satisfying me in all of my scientific interests. I am indebted to them for their understanding and support that assisted me through every difficult moment in my life.
ABSTRACT

Genetic improvement for economically important traits in livestock populations has been revolutionized through the application of genomic selection, where the selection criterion for parents of future generations incorporates genomic estimated breeding values (GEBV). Genomic prediction is a statistical method that predicts GEBV based on high-density genotypes of single nucleotide polymorphisms (SNPs) with genome-wide coverage. The theoretical basis for genomic prediction is that the genetic variance of every quantitative trait locus (QTL) for a desired trait can be captured by SNPs due to linkage disequilibrium (LD) between QTL and SNPs. To date, most statistical models for genomic prediction are based on multiple regression of trait phenotypes on SNP genotypes. Informative prior distributions are usually specified for SNP effects that allow simultaneous estimation of all SNP effects (training). Computer simulation of genomic prediction has revealed that the accuracy of GEBV depends on the genetic basis of trait, the size of training population, and LD between QTL and SNPs, which is affected by historical and current effective population sizes ($N_e$), mutation, selection, population stratification, family structure, SNP density, and minor allele frequencies (MAF) of QTL and SNPs. With moderate to high level of LD, GEBV are expected to have significantly higher accuracy than breeding values estimated using pedigree relationships. In analyses of field datasets, higher accuracy is typically observed in populations that are closely related to the training population, whereas the accuracy in a distantly related population is often low or even zero. Further, prediction accuracy hardly improves by increasing the density of SNPs that are usually selected to have high MAF, which contradicts results from simulation studies. Evidence has been increasing that LD between QTL and SNPs in livestock populations is low because many QTL have much lower MAF than SNPs, and prediction accuracy mainly comes from co-segregation (CS) and additive relationships that are implicitly captured by SNP genotypes.
With low LD between QTL and SNPs, CS information is expected to capture QTL effects more accurately than LD information. CS refers to alleles at linked loci originating from the same parental chromosome, which is quantified by the identical grand-parental allele origins at linked loci. CS information by definition is independent from LD, but is affected by the distance between QTL and SNPs along the chromosome and current $N_e$, which is usually determined by the mating design for a specific breeding program. The objectives of this thesis were to develop a statistical method to model CS explicitly, and to study the effects of historical LD, current $N_e$, MAF of QTL and SNP density on the contributions of LD and CS information to prediction accuracy. The CS model was developed by following the transmission of QTL alleles using allele origins at SNPs. Simulated half-sib datasets were analyzed to study contributions of LD and CS information to prediction accuracy for datasets that included many unrelated families. Simulated datasets of extended pedigrees with different mating designs were analyzed to study contributions of LD and CS information to prediction accuracy across validation generations without retraining. Results from half-sib datasets showed that when LD between QTL and SNPs was low, the accuracy of the model that fits SNP genotypes (LD model) decreased when the training data size was increased by adding independent sire families, but accuracies from the CS model and a combined LD-CS model increased and plateaued rapidly with increasing the number of sire families. Results from half-sib datasets suggest that modeling CS explicitly improves prediction accuracy when LD between QTL and SNPs is low, especially when the training data size is increased by adding independent families. Results from extended pedigrees showed that the LD model resulted in high accuracy across validation generations only when LD between QTL and SNPs was high. With low LD between QTL and SNPs, modeling CS explicitly resulted in higher accuracy than the LD model across validation generations when the mating design generated a large number of close relatives. Results from extended pedigrees suggest that modeling both LD and CS explicitly is expected to improve prediction accuracy when current $N_e$ is small, and LD between QTL and SNPs is low due to distinct MAF, which is the typical situation in most livestock populations.

Application of the CS and the LD-CS models in field datasets has two major difficulties. First, obtaining allele origins for genome-wide SNPs can be computationally demanding. Sec-
ond, the application of the CS model is limited to populations with correctly recorded pedigrees. CS information in populations without pedigree can be explicitly captured by fitting SNP haplotypes. The reason is that, as shown by our previous studies, the association between 1-cM haplotypes and QTL alleles is complete with a high SNP density of 200 SNPs/cM, and therefore 1-cM haplotypes can accurately follow the transmission of QTL alleles from the most recent common ancestor. Simulated datasets of extended pedigrees with different mating designs were analyzed to study contributions of fitting SNP genotypes and haplotypes to prediction accuracy across validation generations without retraining. Results showed that fitting both SNP genotypes and haplotypes had similar accuracy as fitting only SNP genotypes when LD between QTL and SNPs was high, but had significantly higher accuracy than fitting SNP genotypes when LD between QTL and SNPs was low. In the analyses of several egg quality traits of commercial layer chickens, fitting both SNP genotypes and haplotypes improved prediction accuracy for traits for which the accuracy was almost zero by fitting only SNP genotypes. Fitting haplotypes is effective to capture CS information for genomic prediction, especially when LD between QTL and SNPs is low and LD contributes little to prediction accuracy.

In conclusion, genomic prediction models that fit SNP genotypes capture both LD and CS information. When most QTL have much lower MAF than SNPs, LD between QTL and SNPs is low, and the accuracy obtained from fitting SNP genotypes is mainly contributed by CS information that is implicitly captured by SNP genotypes. This accuracy decreases when the training data size is increased by adding independent families, and deteriorates across validation generations without retraining, because CS information captured by SNP genotypes over long chromosome distances erodes rapidly by recombination. CS information can be explicitly captured by modeling transmission of putative QTL alleles within short chromosome regions using allele origins at SNPs. Modeling CS explicitly has limited contribution to accuracy when LD between QTL and SNPs is high, but has substantial contribution to accuracy when LD between QTL and SNPs is low. CS information has greater contribution to accuracy in populations with larger current \( N_e \), because fewer haplotypes segregate in a population with a smaller current \( N_e \), and the effect of each haplotype can be estimated more accurately. Therefore, modeling CS explicitly is expected to result in high accuracy across validation generations in mating designs.
that create small current $N_e$. For populations without pedigree information, CS information can be modeled explicitly by fitting SNP haplotypes within short chromosome regions. Fitting haplotypes captures as much CS information as modeling CS by following the transmission of QTL alleles of pedigree founders, but also captures CS information from most recent common ancestors. Although fitting both SNP genotypes and haplotypes improved accuracy for several traits in layer chickens for which the SNP model had low accuracy, the potential advantage of the SNP-haplotype model in improving accuracy for livestock populations requires further study.
CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

With the establishment of the fundamental principles of quantitative genetics by R. A. Fisher, J. B. S. Haldane and S. Wright, and with the development of linear mixed model methodology by L. N. Hazel (Hazel, 1943; Hazel et al., 1994), C. R. Henderson (Henderson, 1963, 1975), S. R. Searle (Henderson, 1975), R. Thompson (Thompson et al., 2005), etc., genetic improvement for many economically important traits in livestock populations has been successful through artificial selection and breeding of animals with superior genetic merit (Hill, 2014). To improve the genetic merit of populations for desired traits, animals that have superior true breeding values (TBV) among all selection candidates should be selected as parents (Lush, 1943). In reality, TBV are not observed because they are determined by many unobservable quantitative trait loci (QTL), which are functional mutations in DNA sequence that have a direct effect on the phenotypes (Lynch and Walsh, 1998). Thus, selection decisions are based on estimates of TBV (termed estimated breeding values, EBV). The accuracy of an individual’s EBV for a specific trait is measured by the correlation coefficient between EBV and TBV, which is equivalent to the correlation coefficient between EBV and phenotype divided by the square root of trait heritability. Higher accuracies of EBV are usually observed for traits with higher narrow sense heritability. Improving accuracy of EBV is usually the major task in improving genetic gain in a population because prediction accuracy is proportional to the annual genetic gain for a specific trait (Lush, 1943).

Traditional statistical methods for estimation of breeding values are mostly based on linear mixed models that fit TBV as random effects (Henderson, 1984; Mrode and Thompson, 2005). The covariance matrix for the TBV of individuals in a population is the numerator relationship
matrix (NRM) multiplied by the additive genetic variance (Henderson, 1984). The inverse of NRM can be directly constructed using pedigree information (Henderson, 1984). The best linear unbiased prediction (BLUP) of EBV can be efficiently obtained by solving mixed model equations for the linear mixed model (Henderson, 1984). Given all the available phenotypic and pedigree data in a population, selection on BLUP EBV provides maximum genetic gain among all possible selection criteria. In most livestock species, significant genetic improvements have been achieved over decades through selection on EBV for important traits related with production, growth, reproduction and health (Hill, 2014).

Despite the success of selection on BLUP EBV estimated using numerator relationships, such EBV have several undesirable drawbacks that can limit genetic progress. First, accurate estimation of BLUP EBV requires sufficient data on the individual itself and its progeny, which are usually expensive or take long time to obtain. Second, family members tend to be ranked and selected together based on BLUP EBV unless estimated with data from a large number of progeny (progeny test), which usually results in undesirable inbreeding in progeny generations (Daetwyler et al., 2007; de Roos et al., 2011). Third, accurate estimation of BLUP EBV is difficult for traits that are difficult or expensive to measure, e.g. sex-limited traits, post-slaughter traits, diseases and longevity. (Goddard and Hayes, 2007; Fernando et al., 2007).

Since the 1980’s, the ability to genotype different types of DNA polymorphisms has enabled identification of QTL for many economically important traits (Williams et al., 1990; Vignal et al., 2002). Those polymorphisms are either causal mutations that have direct effects on traits, or molecular markers that are linked to or in linkage disequilibrium (LD) with QTL (Dekkers, 2004). In major livestock species, QTL with large effects for several economically important traits were successfully identified or mapped into target genomic regions, for example, the myostatin gene in beef cattle affecting double muscling (Kambadur et al., 1997; Fries et al., 1997; McPherron and Lee, 1997), the MC4R gene in swine affecting fatness, growth and feed-intake (Kim et al., 2000), the DGAT1 gene in dairy cattle affecting milk production (Grisart et al., 2002), and the IGF2 gene in swine affecting muscle growth (Van Laere et al., 2003). Marker-assisted selection (MAS) was introduced to enhance genetic progress by directly selecting on causal mutations or mapped genetic markers for desired traits. MAS has potential
advantage over selection on BLUP EBV because 1) MAS captures Mendelian sampling variance at major QTL; 2) MAS reduces inbreeding by relying less on pedigree relationships; and 3) MAS enables efficient selection for traits that are difficult or expensive to measure (Dekkers, 2004, 2007). Causal mutations have been successfully applied to screen Mendelian genetic diseases in commercial breeding programs (Dekkers, 2004). However, the application of MAS in improving genetic gain of complex traits was limited to few traits with identified major QTL, and the actual benefits of MAS in most commercial breeding programs were unclear (Dekkers, 2004). The limitations of MAS are mostly caused by difficulties in fine mapping of QTL, which include that 1) most QTL mapping experiments can only detect QTL with relatively large effects, while QTL with moderate or small effects lack statistical power to be detected; 2) closely linked QTL are difficult to identify separately due to low marker density (Lynch and Walsh, 1998); 3) stringent thresholds are usually used to adjust for multiple testing, resulting in many false negative signals for small QTL (Churchill and Doerge, 1994); and 4) the effects of statistically significant QTL can be biased upwards due to large random errors, known as the “Beavis effect” (Xu, 2003).

At the start of the 21st century, with the feasibility of chip-genotyping of single nucleotide polymorphisms (SNPs) with genome-wide coverage, genomic selection (GS) Meuwissen et al. (2001) emerged as a revolutionary landmark in the genetic improvement of economically important traits. In GS, all QTL underlying a quantitative trait are expected to be in high LD with SNPs, and hence the effects of all QTL can be simultaneously captured by fitting genotypes of large numbers of SNPs across the genome (Meuwissen et al., 2001; Goddard and Hayes, 2007; Fernando et al., 2007). A typical GS program includes two steps (Meuwissen et al., 2001; Goddard and Hayes, 2007; Fernando et al., 2007). First, the effects of SNPs on the desired trait are estimated using a reference or training population, in which SNP genotypes and trait phenotypes are available for all individuals. The genomic estimated breeding value (GEBV) of an individual is then calculated by the summation of the estimated SNP genotypic values carried by that individual. In the second step, accuracy of GEBV is calculated in a separate population where the individuals have genotypes for the same set of SNPs as in training (validation) (Meuwissen et al., 2001). Application of GS to a target population is expected to have similar
accuracy of GEBV as in the validation population when the structure and relationships with
the training population are similar for the validation and the target population. GS enables
selection at an early age because GEBV of selection candidates can be obtained as soon as a
DNA sample can be obtained. The accuracy of GEBV for selection candidates is expected to
be similar to that in the validation population. The accuracy of GEBV is usually higher than
that of BLUP EBV because GEBV capture the genetic variance at QTL using large numbers
of SNPs.

According to Dekkers (2007), accuracy of GEBV is determined by LD between QTL and
SNPs and accuracy of the estimated SNP effects. LD between QTL and SNPs in livestock
populations depends on historical effective population size \(N_e\) and minor allele frequencies
(MAF) of QTL and SNPs, which are not likely to change within several recent generations.
Therefore, accuracy of GEBV relies mostly on accurate estimation of SNP effects. In most
cases, the number of SNPs used for GS is much larger than the number of phenotypes, and
consequently most statistical models for the estimation of SNP effects are Bayesian hierarchical
models with informative prior distributions for SNP effects, which enable simultaneous estima-
tion of all SNP effects by shrinking small effects towards zero while keeping moderate to large
effects nearly unbiased (Gianola et al., 2009; de los Campos et al., 2013a).

To achieve high accuracy of GEBV, it is usually required that 1) high LD exists between
QTL and SNPs, and 2) the same LD is present in the training and validation populations.
High accuracies of GEBV were frequently observed in analyses of simulated datasets where the
above two conditions were satisfied. However, the accuracies in the analyses of field datasets
were usually much lower than those observed in simulation studies. Possible reasons are that 1)
LD between QTL and SNPs in real livestock populations are low because most QTL have low
minor allele frequencies (MAF) (Druet et al., 2014; MacLeod et al., 2014; Hayes et al., 2014);
and 2) LD information captured by genomic prediction models in the training population is
not conserved in the target population due to different population structures (Habier et al.,
2007; de Roos et al., 2008; Daetwyler et al., 2012; de los Campos et al., 2013b; Habier et al.,
2013). One of the objectives of this thesis is to investigate the effect of MAF of QTL and SNPs
on accuracy of GEBV.
LD between QTL and SNPs was originally regarded as the only source of information that contributed to accuracy of genomic prediction, until Habier et al. (2007) and Habier et al. (2013) showed that information from pedigree relationships and co-segregation (CS) also contribute to prediction accuracy. CS refers to alleles at linked loci originating from the same parental chromosome. Contributions of LD and pedigree relationships to prediction accuracy have been investigated by extensive simulation studies, as well as by field data analyses in many livestock species. However, few studies have been performed to investigate the contribution of CS information to prediction accuracy. The main objective of this thesis was to investigate the contributions of LD and CS information to accuracy of genomic prediction.

1.2 Research Objectives

The objectives of research presented in this thesis were 1) to develop statistical methods for modeling CS explicitly to follow the transmission of QTL alleles, either with or without using pedigree information; 2) to investigate the contributions of LD and CS information to accuracy of genomic prediction across unrelated families and across validation generations without retraining, and 3) to improve prediction accuracy by modeling both LD and CS information explicitly.

The major factors that affect contributions of LD and CS information to prediction accuracy were investigated using simulated datasets. LD is expected to have a large contribution to accuracy when historical LD between QTL and SNPs is high, which requires high SNP density and similar MAF between QTL and SNPs. Simulation scenarios were designed to study the effects of SNP density and MAF of QTL and SNPs on contributions of LD and CS information. CS is expected to have a greater contribution to accuracy in a population with a smaller current $N_e$, which is usually determined by the mating design of a specific breeding program. Datasets of extended pedigrees with different mating designs were simulated to study the effect of current $N_e$ on the contribution of CS information to prediction accuracy. The potential advantage of combined modeling LD and CS in the same model in improving accuracy across validation generations was investigated using simulated datasets and a field dataset from a commercial breeding population of layer chickens.
1.3 Organization of Thesis

The rest of this chapter provides a general review of literature on genomic prediction methods and the major factors that have effects on accuracy of genomic prediction, as revealed by both simulation studies and field data analyses.

In Chapter 2, a new statistical method was developed to explicitly model CS information using parental allele origins at SNPs for pedigreed populations. The contributions of LD and CS information to prediction accuracy were investigated by analyzing simulated half-sib datasets using the CS model, the LD model, and the combined LD-CS model.

In Chapter 3, the effects of historical LD and current $N_e$ on the contributions of LD and CS information to prediction accuracy were investigated using simulated datasets with different mating designs. The potential advantage of the LD-CS model in improving prediction accuracy across validation generations without retraining was also tested on the simulated datasets.

Chapter 4 is a paper “Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes” that was published in Proceedings of the 10th World Congress on Genetics Applied to Livestock Production (Sun et al., 2014). A haplotype model was proposed to improve prediction accuracy for traits with many rare QTL, which was used in Chapter 5 to model CS without using pedigree information.

In Chapter 5, the haplotype model in Chapter 4 was used to model CS information without using pedigree information. The ability of the haplotype model to capture CS information was investigated using the simulated datasets in Chapter 3. The potential advantage of the combined SNP-haplotype model for improving accuracy across validation generations was investigated using both the simulated datasets and a field dataset of layer chickens.

Chapter 6 provided general discussions and conclusions from research work presented in chapters 2, 3, 4 and 5. The advantages of modeling both LD and CS information explicitly in improving prediction accuracy are summarized. Issues with the application of the LD-CS methods in real livestock populations, and further improvements of the LD-CS methods in genomic prediction using sequence data are discussed.
Appendix A is a paper published in *PLoS One* (Sun et al., 2012), “A fast EM algorithm for BayesA-like prediction of genomic breeding values”. Since the SNP model was shown to have high accuracy with strong LD between QTL and SNPs, an EM algorithm for the widely-used “BayesA” method was developed as a computationally efficient alternative for conventional Markov chain Monte Carlo (MCMC) algorithms. The method “fastBayesA” in Sun et al. (2012) has been implemented in the ASReml software (version SA-4).

Appendix B is a paper published in *BMC Proceedings* (Sun et al., 2011), “Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods”. The study presented results from the comparison of several SNP models in genomic prediction and QTL detection of the common dataset provided by the 14th European QTL-MAS workshop. The BayesCπ method used by Sun et al. (2011) had highest prediction accuracy (out of 26 groups) and maximum number of correctly mapped QTL (out of 7 groups) for the simulated quantitative trait of the common dataset.

### 1.4 Literature Review

Ever since the concept of genomic selection was proposed by Meuwissen et al. (2001), numerous statistical models for genomic prediction have been developed. Meanwhile, various factors that affect accuracy of genomic prediction have been investigated using simulated and field datasets. This review of literature summarizes the statistical methods for genomic prediction, and factors that have been verified to have a large effect on accuracy of genomic prediction.

#### 1.4.1 Statistical models and computation algorithms for genomic prediction of breeding values

To date, most genomic prediction models are based on multiple regression of trait phenotypes on SNP genotypes. The main difference between SNP models is the prior distributions specified for the random SNP effects, which enable simultaneous estimation of all SNP effects by inducing shrinkage or variable selection. Comprehensive reviews of genomic prediction methods can be found in Kärkkäinen and Sillanpää (2012) and de los Campos et al. (2013a). The
effects of prior distributions on the estimation of SNP effects can be found in Gianola et al. (2009) and Gianola (2013).

The statistical models fitting SNP genotypes can be categorized into two classes. One class comprises models that fit SNP genotypes (SNP effect model), and the other class of models that fit genomic breeding values with the covariance matrix constructed from genome-wide SNP genotypes (the BV model).

In most SNP effect models, the residual sampling distribution for the phenotype of training individual $i$ ($y_i$) is normal:

$$y_i | \beta, \delta_j, \alpha_j, \sigma^2_e \sim \text{independent } N\left(x_i' \beta + \sum_{j=1}^{m} z_{ij} \delta_j \alpha_j, \sigma^2_e\right),$$

(1.1)

where $\beta$ is a vector of non-genetic fixed effects, $x_i$ is the design vector for fixed effects $\beta$ of individual $i$, $\alpha_j$ is the random effect of the $j$th SNP, $z_{ij}$ is the genotype at the $j$th SNP of individual $i$, coded as the number of minor allele in 0/1/2, $m$ is the total number of SNPs, $\delta_j$ is an indicator variable that equals 1 when the $j$th SNP is fitted in the model and 0 otherwise, and $\sigma^2_e$ is the residual variance.

Different SNP effect models vary in the prior distributions for $\alpha_j$ and $\delta_j$. The prior for $\alpha_j$ controls the amount of shrinkage on the estimated SNP effect $\hat{\alpha}_j$, and the prior for $\delta_j$ induces variable selection. The prior for $\alpha_j$ in most genomic prediction models is mainly derived from 1) Student-\text{t} distribution or 2) double exponential (DE) distribution.

When using a t prior for SNP effects, small effects are heavily shrunk towards zero while larger effects are only slightly or not shrunk. The amount of shrinkage is determined by the degrees of freedom of the t distribution. Models with t priors for genomic prediction were first proposed by Meuwissen et al. (2001) as “BayesA”. For computational convenience, the t distribution is usually reparameterized as a normally distributed SNP effect conditional on its variance, which follows a scaled inverse Chi-square distribution. BayesA assigns an independent normal conditional prior for the $j$th SNP effect $\alpha_j$, given its locus-specific variance $\sigma^2_j$. The locus-specific variance $\sigma^2_j$ follows a scaled inverse Chi-square distribution with degrees
of freedom $\nu_a$ and scale parameter $S^2_{\alpha}$,

$$
\alpha_j | \sigma^2_j \sim N(0, \sigma^2_j),
\sigma^2_j \sim \chi^{-2}_{\nu_a}(S^2_{\alpha}),
$$

(1.2)

where $\chi^{-2}(b)$ represents the scaled inverse Chi-square distribution with $a$ degrees of freedom and scale parameter $b$. The prior distribution (1.2) can be easily shown to be equivalent to $\alpha_j \sim t_{\nu_a}(S^2_{\alpha})$, where $t_a(b)$ is the scaled $t$ distribution with $a$ degrees of freedom and scale parameter $b$ (Gianola et al., 2009; Sun et al., 2012).

The marginal posterior distribution for $\alpha_j$ does not have a mean or mode in closed form, and the posterior mean calculated using Markov chain Monte Carlo (MCMC) samples is usually used as the point estimate of $\alpha_j$. Gibbs sampling (Casella and George, 1992) can be feasibly implemented for BayesA because sampling from full conditional distributions is convenient. Alternatively, Sun et al. (2012) (Appendix A) proposed an EM algorithm with convergence to a joint posterior mode for all SNP effects. Estimates of SNP effects using the joint posterior mode from the EM algorithm were similar to the marginal posterior mean calculated from MCMC samples (Sun et al., 2012).

The method “BayesB” (Meuwissen et al., 2001) uses a mixture prior with a point mass at zero with probability $\pi$ and a $t_{\nu_a}(S^2_{\alpha})$ elsewhere:

$$
\alpha_j | \sigma^2_j, \delta_j \sim \begin{cases} 
0, & \delta_j = 0; \\
N(0, \sigma^2_j), & \delta_j = 1,
\end{cases}
\sigma^2_j \sim \chi^{-2}_{\nu_a}(S^2_{\alpha}),
\delta_j \sim \text{Bernoulli}(1 - \pi).
$$

(1.3)

In each iteration of the MCMC, a proportion of $1 - \pi$ SNPs are assumed to have non-zero effect on the trait and are fitted in the model. SNPs with large effects are more likely to be fitted in the model among MCMC iterations than SNPs with small effects. Compared to BayesA, the posterior mean of large SNP effects in BayesB has smaller shrinkage, whereas that of small SNP effects is shrunk more heavily towards zero.

A stochastic search variable selection (SSVS) method similar to BayesB was used by Verbyla et al. (2009) for genomic prediction. The conditional prior for $\alpha_j$ in SSVS is a mixture of two
normal distributions:

\[
\alpha_j | \sigma^2_j, \delta_j \sim \begin{cases} 
N(0, \tau^{-1} \sigma^2_j), & \delta_j = 0; \\
N(0, \sigma^2_j), & \delta_j = 1,
\end{cases}
\]

\[
\sigma^2_j \sim \chi_{\nu_{\alpha}}^{-2}(S^2_{\alpha}),
\]

\[
\delta_j \sim \text{Bernoulli}(1 - \pi),
\]

(1.4)

The value of \( \tau \) is usually large to represent a normal prior with small effect variance for small \( \alpha_j \). The advantage of SSVS over BayesB is that SSVS always fits all \( m \) SNPs in the model. SSVS is computationally more convenient than BayesB but has similar shrinkage as BayesB (Verbyla et al., 2009).

The “BayesC” method proposed by (Kizilkaya et al., 2010) uses a multivariate \( t \) prior for \( m \) SNP effects:

\[
\alpha_j | \sigma^2_{\alpha}, \delta_j \sim \begin{cases} 
0, & \delta_j = 0; \\
N(0, \sigma^2_{\alpha}), & \delta_j = 1,
\end{cases}
\]

\[
\sigma^2_{\alpha} \sim \chi_{\nu_{\alpha}}^{-2}(S^2_{\alpha}),
\]

\[
\delta_j \sim \text{Bernoulli}(1 - \pi).
\]

(1.5)

The joint posterior distribution for the \( m \) SNP effects in BayesC is multivariate \( t \) with \( \nu_{\alpha} + m \) degrees of freedom. The BayesC model was further extended to “BayesC\( \pi \)” (Habier et al., 2011) where \( \pi \) was treated as unknown with a uniform prior. Method “BayesD\( \pi \)” (Habier et al., 2011) extends BayesB by using a gamma distribution as the prior for the scale parameter of the prior for effect variances.

Method “BayesR” proposed by Erbe et al. (2012) uses a mixture of four normal distributions as the prior for the \( m \) SNP effects. The four normal distributions vary in variances that correspond to the proportion of genetic variance that each SNP can potentially explain. BayesR was extended into 1) “BayesRS” (Brondum et al., 2012), where the genome is divided into small segments and SNP effects within each segment have the same mixture priors as in BayesR, and 2) “BayesRC” (Hayes et al., 2014), where SNPs are classified based on their function (non-coding/synonymous/missense/nonsense mutations) and SNP effects within each functional class have the same mixture prior as in BayesR.
LASSO (Tibshirani, 1996) regression uses DE distribution as the prior for regression coefficients. LASSO was first used for genomic prediction by (Usai et al., 2010). The conditional prior distribution for $\alpha_j$ is DE with parameter $\lambda$, 

$$\alpha_j | \lambda \propto \frac{\lambda}{2} \exp(-\lambda |\alpha_j|).$$ (1.6)

LASSO shrinks SNP effects with absolute values less than $\lambda$ exactly to zero, which is stronger than variable selection methods BayesB, BayesC or SSVS. However, the shrinkage on SNP effects with absolute values larger than $\lambda$ are the same regardless of the size of the actual SNP effects, which differs from the shrinkage by using $t$ priors where large effects are less biased.

LASSO was modified into Bayesian LASSO (Park and Casella, 2008; de los Campos et al., 2009) by using a conditional DE distribution as the prior for $\alpha_j$:

$$\alpha_j | \lambda, \sigma_e^2 \propto \frac{\lambda}{2\sqrt{\sigma_e^2}} \exp\left(-\frac{\lambda}{2\sqrt{\sigma_e^2}} |\alpha_j|\right).$$ (1.7)

The conditional prior (1.7) behaves like a mixture of an infinite number of normal distributions with variances following an exponential distribution scaled by $\sigma_e^2$. Gibbs sampling can be feasibly implemented for the Bayesian LASSO because all the full conditional distributions have closed forms. A gamma prior for $\lambda^2$ was suggested to improve convergence of the Gibbs sampler for the Bayesian LASSO (Kärkkäinen and Sillanpää, 2012; Gianola, 2013).

Most SNP models use MCMC algorithms for the estimation of SNP effects. The computation time for MCMC algorithms increases linearly with the number of SNPs and individuals fitted in the model. Alternatively, the expectation-maximization (EM) algorithm (Dempster et al., 1977) provides opportunities for fast estimation of SNP effects without MCMC. The EM algorithm searches for the posterior mode through direct maximization of the probability density functions of the posterior distributions. In addition, some generalized EM algorithms and ad hoc algorithms have been developed to deal with more complicated posterior distributions. Hayashi and Iwata (2010) formulated a generalized EM algorithm for BayesA and an ad hoc EM-like algorithm for BayesB. Sun et al. (2012) (Appendix B) developed a rigorous EM algorithm for BayesA that converged to a joint posterior mode of all SNP effects. Meuwissen et al. (2009) developed an iterative conditional expectation (ICE) algorithm for BayesB with
a different parameterization from the BayesB of Meuwissen et al. (2001), where the priors for SNP effects were mixtures of point mass at zero and a DE elsewhere. In the model of Meuwissen et al. (2009), the posterior conditional expectation of $\alpha_j$ has closed form and can be calculated analytically. Shepherd et al. (2010) improved the ICE algorithm of Meuwissen et al. (2009) by iteratively maximizing the posterior distribution for $\alpha_j$ using Gauss-Seidel iteration instead of the conditional expectation used by Meuwissen et al. (2009).

Recently, a more generic variational Bayes (VB) algorithm was introduced for genomic prediction (Li and Sillanpää, 2012; Hayashi and Iwata, 2013). In VB, posterior distributions for all parameters are approximated by independent candidate distributions, which are usually chosen from common parametric distributions for mathematical convenience. In this algorithm, the Kullback–Leibler (KL) divergence between the joint posterior distribution and the joint candidate distribution is minimized iteratively. Posterior inferences of parameters are based on the candidate distributions with updated parameter values that minimizes KL divergence, which obviate the often intractable marginal posterior distributions. Simulation studies have shown that the EM and the VB algorithm had similar accuracy as the MCMC algorithms for the same model, but were usually computationally much faster (Meuwissen et al., 2009; Shepherd et al., 2010; Hayashi and Iwata, 2010; Sun et al., 2012; Li and Sillanpää, 2012; Hayashi and Iwata, 2013).

The other class of genomic prediction models fit genomic breeding values with a covariance structure constructed from SNP genotypes, known as the GBLUP (VanRaden, 2008; Gianola et al., 2009). The GBLUP model shares the same residual sampling distribution as the SNP effect model:

$$y|\beta, g, \sigma_e^2 \sim N_n\left(\mathbf{X}\beta + g, \mathbf{I}\sigma_e^2\right),$$

where $y$ is an $n \times 1$ vector of phenotypes, $\mathbf{X}$ is the design matrix for non-genetic fixed effects $\beta$, $g$ is an $n \times 1$ vector of genomic breeding values, and $\sigma_e^2$ is the residual variance. The covariance matrix for $g$ is

$$\text{Var}(g) = \mathbf{G}\sigma_g^2,$$
where $G$ is the genomic relationship matrix (GRM) and $\sigma_g^2$ is the variance of genomic breeding values $g$. Prediction of genomic breeding values is usually achieved by solving mixed model equations for the model (1.8), similar to obtaining BLUP EBV based on pedigree relationships. Different from the inverse of NRM, the inverse of GRM, $G^{-1}$, is dense and cannot be obtained directly. Calculation of $G^{-1}$ is feasible when the number of individuals is up to several thousand, in which cases GBLUP can be computationally more efficient than SNP effect models (Strandén and Garrick, 2009; Sun et al., 2012).

Among the many strategies to construct $G$ (VanRaden, 2008; Yang et al., 2010), none of them has been generally accepted. The genomic relationship between two individual $i$ and $j$, i.e. the $ij$th element of $G$ ($G_{ij}$), is usually calculated as

$$G_{ij} \propto z_i'z_j, \quad i, j = 1, 2, \ldots, n,$$

where $z_i$ is an $m \times 1$ vector of SNP genotypes for individual $i$. Strategies to construct GRM differ in the normalizing of $z_i$ or the scaling of $z_i'z_j$ (VanRaden, 2008; Yang et al., 2010).

The GBLUP model is equivalent to the SNP effect model assuming normally distributed SNP effects with the same known effect variance (Fernando, 1998; Strandén and Garrick, 2009). The GBLUP model has also been generalized into some empirical methods that put different weights on SNPs according to their estimated effects or generic variances (Zhang et al., 2010; Sun et al., 2012; Shen et al., 2013).

Besides a variety of parametric methods, many semi/non-parametric methods from machine learning and artificial intelligence have also been introduced for genomic prediction (Gianola et al., 2006; Gianola and van Kaam, 2008; de los Campos et al., 2009; González-Camacho et al., 2012). These will not be reviewed here.

### 1.4.2 Sources of genetic information that contribute to accuracy of genomic prediction

LD, CS and pedigree relationships are the three sources of genetic information that contribute to accuracy of genomic prediction (Habier et al., 2007, 2013). Multiple regression models on SNP genotypes (SNP model) explicitly capture LD information, but also implicitly
capture information from CS and pedigree relationships. The SNP model has high accuracy when high LD exists between QTL and SNPs. CS and pedigree relationships that are implicitly captured by the SNP model have large contributions when predicting close relatives but have small contributions when predicting individuals that are distantly related with the training population. The ability of the SNP model to capture CS and pedigree relationships depends on the effective number of SNPs (Habier et al., 2013) fitted in the model. In the analyses of field datasets from livestock, accuracies were much higher for predicting close relatives than individuals that were distantly related to the training population (Habier et al., 2010; Wolc et al., 2011; Saatchi et al., 2011; Wientjes et al., 2013), which suggests that LD between QTL and SNPs in livestock populations is low, and prediction accuracy mainly comes from CS and pedigree relationships that are implicitly captured by SNP genotypes.

Fitting polygenic effects in addition to SNP genotypes was proposed to explicitly model pedigree relationships (Calus and Veerkamp, 2007). Fitting polygenic effects improved prediction accuracy when SNP density was low, but hardly improved accuracy when SNP density was high, or when SNP genotypes already explained most of the genetic variance (Meuwissen and Goddard, 2007; Calus and Veerkamp, 2007; Solberg et al., 2008).

Modeling CS explicitly was recommended to improve prediction accuracy by explicitly capture CS information (Habier et al., 2013). Several models that fit SNP haplotypes were proposed to capture identity-by-descent (IBD) among haplotypes (Calus et al., 2008; Villumsen et al., 2009; Hickey et al., 2013). Haplotype models improved prediction accuracy under very low SNP density, but had very similar accuracies as the SNP model when SNP density was greater than 20 SNPs/cM (Calus et al., 2008; Villumsen et al., 2009; Hickey et al., 2013). Fitting haplotypes significantly improved prediction accuracy when most QTL were rare (Sun et al., 2014), because high LD did not exist between loci with different MAF, but could exist between haplotypes and QTL alleles (Goddard and Hayes, 2007).

The only study to investigate the contributions of LD and CS information by modeling CS explicitly was by Luan et al. (2012). Luan et al. (2012) fitted genomic breeding values with genomic relationship matrices derived from either SNP genotypes or IBD of SNP alleles using the method of Fernando and Grossman (1989). By analyzing a dataset of Norwegian red
cattle, Luan et al. (2012) showed that prediction accuracy mainly came from CS information, and the contribution of LD information was minimal when CS was already explicitly fitted in the model. Recently, Meuwissen et al. (2014) pointed out that genomic relationships due to LD and CS and pedigree relationships are three sources of genetic relationships that exist in a pedigreed population. An ideal prediction model should account for all three types of relationships. Meuwissen et al. (2014) proposed to model LD information by fitting SNP genotypes, but model CS and pedigree relationships using relationship matrices. In this thesis, a CS model was developed that fit allelic values of pedigree founders at putative QTL, for which all statistical methods that have been applied to estimating SNP effects can be used to estimate founder allelic values.

1.4.3 Factors that affect prediction accuracy

Apart from pedigree relationships and CS information that are discussed above, the major factors that affect prediction accuracy can be categorized into 1) the genetic background of the trait, 2) LD between QTL and SNPs, and 3) the properties of the training population.

The genetic background of a trait includes its narrow sense heritability ($h^2$), the number of QTL, the distribution of QTL effects, and dominance and interactions among QTL. The accuracy is proportional to $h^2$ in some deterministic formulas to empirically calculate accuracy of genomic prediction (Goddard, 2009; Daetwyler et al., 2010). In most simulation and field data studies, prediction accuracies increased with $h^2$ (Muir, 2007; Wolc et al., 2011). Further, prediction accuracies approach to 1 when QTL were fitted in the model or in near complete LD with SNPs, but are much lower than 1 when LD between QTL and SNPs was imperfect (Daetwyler et al., 2010; Meuwissen and Goddard, 2010; de los Campos et al., 2013b). For traits controlled by a few QTL with relatively large effects (oligogenic), high accuracy can be achieved by using variable selection methods such as BayesB, BayesC$\pi$ or Bayesian LASSO, which usually have much higher accuracy than BayesA or GBLUP that always fit all SNPs in the model. However, for traits controlled by a large number of QTL with approximately normally distributed effects (polygenic), prediction accuracies are usually lower than those of oligogenic traits, and variable selection methods had similar accuracy as BayesA or GBLUP.
LD between QTL and SNPs is affected by historical effective population size, $N_e$, and the distance between QTL and SNPs on chromosome (Sved, 1971). Simulation studies have shown that prediction accuracies were lower in populations with large $N_e$ than in populations with small $N_e$, because the genetic drift that creates LD is high with small $N_e$ (Muir, 2007; Solberg et al., 2008; Meuwissen, 2009; Daetwyler et al., 2010; MacLeod et al., 2014). Increasing SNP density increases LD between QTL and SNPs by reducing the average distance between QTL and SNPs, but increased LD is only achieved between QTL and SNPs with similar MAF (de los Campos et al., 2013b; Druet et al., 2014). In simulation studies, increasing SNP density resulted in higher prediction accuracy when QTL and SNPs were simulated with similar MAF (Muir, 2007; Solberg et al., 2008; Meuwissen and Goddard, 2010; VanRaden et al., 2011). Improvement in prediction accuracy by increasing SNP density was more significant for oligogenic than polygenic traits (Meuwissen and Goddard, 2010; Daetwyler et al., 2010; Sun et al., 2012).

In practice, LD between QTL and SNPs is difficult to quantify because most QTL are not observed. The pairwise LD between SNPs is usually used to represent LD between QTL and SNPs, under the assumption that the distribution of MAF is similar for QTL and SNPs. The $r^2$ measure of LD as the correlation coefficient between SNP genotypes (Falconer and Mackay, 1996) is above 0.2 between two loci separated by less than 100 kilo-basepairs (kb), which is regarded as useful LD for genomic prediction because it can be conserved across breeds and validation generations (de Roos et al., 2008; Goddard, 2009). In the studies of LD in cattle populations, LD between adjacent SNPs was greater than 0.2 with a SNP density above 200 SNPs/cM (de Roos et al., 2008; Qanbari et al., 2010; Espigolan et al., 2013), which is equivalent to the 600K SNP chip for the bovine genome. However, in genomic prediction for cattle, increasing SNP density from the 50K to 770K SNP chip resulted in similar or limited increases in prediction accuracy (Erbe et al., 2012; Hozé et al., 2014; Gunia et al., 2014; Saatchi and Garrick, 2014). The most likely reason is that most QTL for economically important traits have low MAF (Druet et al., 2014; MacLeod et al., 2014; Hayes et al., 2014), and LD between QTL
and SNPs can not be improved by increasing the density of SNP chip due to the discrepancy in MAF of QTL and SNPs (MacLeod et al., 2014; Sun et al., 2014). Including SNPs with similar MAF as QTL improved prediction accuracy because rare SNPs could be in high LD with rare QTL (de los Campos et al., 2013b; Druet et al., 2014; Sun et al., 2014). Prediction accuracy for traits with many rare QTL was also improved by fitting SNP haplotypes, because haplotypes can be in complete association with QTL alleles (Goddard and Hayes, 2007; Sun et al., 2014) under high SNP density.

To achieve high prediction accuracy requires a training population with sufficiently large size and that has LD between QTL and SNPs consistent with the target population (Goddard and Hayes, 2009; Wray et al., 2013). The size of the training population that is required to achieve a specific accuracy increases with $N_e$, and decreases with an increase in trait heritability (Goddard and Hayes, 2009; Goddard, 2009). Prediction accuracy is high for individuals closely related with the training population (Habier et al., 2007, 2010; Saatchi et al., 2011; Wientjes et al., 2013; Hozé et al., 2014; Saatchi and Garrick, 2014), whereas accuracy was low or zero when the validation individuals were separated by many generations from the training population (Habier et al., 2007, 2010; Saatchi et al., 2011; Wolc et al., 2011; Wientjes et al., 2013; Weng et al., 2014a), or from breeds that were not included in the training population (Hayes et al., 2009a; Daetwyler et al., 2012; Kachman et al., 2013; Zhou et al., 2014; Saatchi and Garrick, 2014; Calus et al., 2014). These results suggest that accuracy of genomic prediction in real livestock populations is mainly contributed by CS and pedigree relationships provided by close relatives in the training population, rather than by LD between QTL and SNPs (Daetwyler et al., 2012).

The low LD between QTL and SNPs in livestock populations is mainly resulted from the difference in their MAF. The QTL for economically important traits are likely to have low MAF either because traits have undergone long directional selection (Hayes et al., 2010; Daetwyler et al., 2014; Druet et al., 2014), or because QTL tend to be mutations that occur more recently than SNPs (Hayes et al., 2010; Druet et al., 2014). The SNPs included on SNP chips usually have high MAF chosen from sequencing and prototype genotyping of reference samples (Matukumalli et al., 2009). Since high LD cannot exist between two loci with distinct MAF,
LD information has little contribution to prediction accuracy. Modeling CS explicitly can improve prediction accuracy when most QTL have low MAF, because CS information is hardly affected by the level of LD between QTL and SNPs.

With the reduced costs to obtain whole genome sequences for individual animals, genomic prediction using all polymorphisms in the genome could be used to improve prediction accuracy (Meuwissen and Goddard, 2010). With sequence data, all QTL for the trait can be directly fitted in the prediction model without relying on LD between QTL and SNPs. Prediction accuracies in simulated sequence datasets were significantly higher than accuracies in simulated SNP chip genotypes (Meuwissen and Goddard, 2010; Druet et al., 2014; MacLeod et al., 2014). However, accuracy using sequence data from actual cattle populations was only improved marginally over accuracy using the 770K SNP chips (Hayes et al., 2014). Possible reasons include the errors in calling SNPs from sequence and the difficulty in assembling a sufficiently large training population to accurately estimate the effects of sequence SNPs (Druet et al., 2014; Hayes et al., 2014).

### 1.4.4 Problems in current genomic prediction that the thesis aims to address

According to the above review of literature, the major problem in genomic prediction for livestock populations is that LD between QTL and SNPs is low due to most QTL having low MAF, and accuracy of SNP models mainly comes from CS and pedigree relationships that are implicitly captured by SNP genotypes. The objective of the research presented in this thesis is to improve prediction accuracy by explicitly modeling CS information when LD between QTL and SNPs is low. CS information follows the transmission of QTL alleles among relatives using parental allele origins of SNPs. In populations where pedigrees are not available, the haplotype model in a previous study (Sun et al., 2014) was used to explicitly model CS without using pedigree information. Contributions of LD and CS information to accuracy of genomic prediction were investigated using simulated datasets with different scenarios of historical LD, current \( N_e \), MAF of QTL and SNPs, and SNP density. The potential advantage of modeling CS explicitly in addition to SNP genotypes was tested on simulated and field datasets.
1.5 Bibliography


CHAPTER 2. CONTRIBUTIONS OF LINKAGE DISEQUILIBRIUM AND CO-SEGREGATION INFORMATION TO THE ACCURACY OF GENOMIC PREDICTION IN HALF-SIB DESIGNS

2.1 Abstract

Traditional genomic prediction models using multiple regression on single nucleotide polymorphism (SNP) genotypes exploit linkage disequilibrium (LD) between quantitative trait loci (QTL) and SNPs. Results from real data analyses show that prediction accuracy is usually much higher for individuals that are close relatives of individuals in the training populations than for distantly related individuals. The possible reason is that historical LD between QTL and SNPs is weak, and prediction accuracy of SNP models is mainly contributed by pedigree relationships and co-segregation (CS) of QTL with SNPs in close relatives. Information from additive relationship only contributes to within family prediction, and decreases fast over generations. Information from CS persists over generations and families, and is expected improve prediction accuracy when modeled explicitly. In this study, a method to explicitly model CS is developed by following the transmission of putative QTL alleles using allele origins at SNPs. Bayesian hierarchical models that combine information from LD and CS (LD-CS model) are developed for genomic prediction in pedigree populations. Here, CS is modeled as founder allelic values at putative QTL within every 1-cM genomic window that one individual inherits from its ancestors, while LD is modeled as the effects of genotypes at all SNPs. Contributions of LD and CS information to prediction accuracy are investigated using the LD model, the CS model and the LD-CS model, in simulated half-sib datasets. The datasets are composed of paternal half-sib families that vary in the number of sires, and within each sire family 10 half-sib progeny are used for training to predict breeding values for another set of 10 candidate
half-sibs. Results show that when historical LD between QTL and SNPs is imperfect, accuracy of the LD model decreases when the training data size is increased by adding independent sire families, but accuracies from the CS and LD-CS models increased and plateaued rapidly with increasing the number of sire families. Results suggest that modeling CS explicitly improves prediction accuracy when historical LD between QTL and SNPs is imperfect, especially when the training data include a large number of independent families.

2.2 Introduction

Currently most statistical models for genomic prediction are based on multiple regression of phenotypes on genotypic covariates of single nucleotide polymorphisms (SNPs) (Meuwissen et al., 2001; Goddard and Hayes, 2007; Habier et al., 2011). Prediction accuracy of multiple regression models relies on the assumption that all quantitative trait loci (QTL) underlying the trait are in linkage disequilibrium (LD) with SNPs, and therefore the model can explain most genetic variance at QTL by fitting SNP genotypes. Substantially higher accuracies have been observed in simulation studies for SNP models compared with prediction models using pedigree (Meuwissen et al., 2001; Habier et al., 2007; Muir, 2007), because SNP genotypes can capture more genetic variance than pedigree relationships due to high LD between QTL and SNPs in simulated datasets. In field data analyses, however, high accuracy of SNP models is observed mostly in genomic prediction of close relatives of the training population (Luan et al., 2009; VanRaden et al., 2009; Habier et al., 2010; Hayes et al., 2009b), and prediction accuracy is low or even zero in validation populations that are distantly related with the training population (Hayes et al., 2009a; Kachman et al., 2013; Zhou et al., 2014; Saatchi and Garrick, 2014). A possible reason is that historical LD between QTL and SNPs is low, and prediction accuracy of SNP models mainly comes from co-segregation (CS) and additive relationships that are implicitly captured by SNP genotypes (Habier et al., 2007, 2010; Daetwyler et al., 2012; Luan et al., 2012; Wientjes et al., 2013).

Co-segregation is an important source of information that contributes to accuracy of genomic prediction (Luan et al., 2012; Habier et al., 2013). CS of alleles at two loci is defined as non-random association between their grand-parental allele origins. For instance, the mater-
nal alleles of an individual at two loci co-segregate when both alleles originate from the same maternal chromosome (He et al., 2010; Habier et al., 2010). With high-density genotyping, the probability that a QTL allele co-segregates with its two proximal SNP alleles is high. For example, the average distance between adjacent SNPs on the Illumina Bovine SNP50 BeadChip is 50 kilo-basepairs (Matukumalli et al., 2009; Qanbari et al., 2010), and the average recombination rate between two adjacent SNPs is around 5% per meiosis. Thus, modeling CS of QTL with SNPs can follow the transmission of QTL alleles very accurately.

CS information accounts for QTL effects that cannot be fully captured by SNP genotypes under imperfect LD. LD quantifies the correlation of QTL genotypes with SNP genotypes. The effect of a QTL that is captured by SNP genotypes is the expected value of the QTL effect conditional on SNP genotypes. When LD between QTL and SNPs is imperfect, QTL effects cannot be entirely captured by SNP genotypes (He et al., 2010). In reality, however, each QTL allele is transmitted in an all-or-none form, and the deviation of QTL effect from its conditional expectation is independent of SNP genotypes. Modeling CS explicitly captures the deviation due to imperfect LD because CS follows the transmission of QTL alleles. Since CS information is not affected by the level of LD, the CS model is expected to have consistent prediction accuracy when LD between QTL and SNPs is low. Therefore, CS information has a potential advantage over LD information to improve prediction accuracy especially when the population includes many unrelated groups of individuals, e.g. unrelated nuclear families or breeds.

Simulation studies have shown that both LD and CS information contribute to prediction accuracy of the genomic best linear unbiased prediction (GBLUP) model (Habier et al., 2013). Information from historical LD was persistent and contributed to prediction accuracy across families. CS information that is captured implicitly by GBLUP was not persistent across families, and its prediction accuracy decreased with increasing number of unrelated families in training (Habier et al., 2013). Prediction accuracy from modeling CS explicitly is expected to be persistent across families, but there are no simulation studies on the contributions of LD and CS information to prediction accuracy by modeling LD and CS explicitly. In a study using datasets of Italian Brown Swiss bulls, the importance of LD and CS information in prediction
accuracy was investigated through the GBLUP model with covariance structure of genomic estimated breeding values (GEBV) constructed from either LD or CS information at SNPs (Luan et al., 2012). Prediction accuracy did not improve by fitting LD information in addition to CS, and the GBLUP model that fitted both LD and CS had similar accuracy as fitting only CS, which was slightly higher than fitting only LD (Luan et al., 2012). These results suggest that when historical LD between QTL and SNPs was low, and prediction accuracy for closely related individuals mainly came from CS instead of LD.

The method for modeling CS in Luan et al. (2012) has two limitations. First, CS is modeled at every SNP locus. Since CS signal spans long genomic distance, modeling CS across multiple SNP loci is expected to capture the same amount of CS information as modeling at every SNP locus, but can improve computation substantially. Second, the contribution of CS information at each SNP was assumed the same in Luan et al. (2012), which is not desired when the genetic variance of QTL differs across the genome. In this study, a new method is developed to model CS explicitly. The CS model follows the transmission of putative QTL within 1-cM genomic windows. A detailed description of a Bayesian hierarchical model for genomic prediction using CS is provided, and a Gibbs sampling algorithm for prediction of breeding values is derived. The LD, CS and a combined LD-CS models are used for genomic prediction in simulated half-sib datasets to investigate the contributions of LD and CS information to accuracy of genomic prediction, and to test the hypothesis that the combined LD-CS model improves prediction accuracy when historical LD between QTL and SNPs is imperfect, and the target population includes many groups of distantly related individuals.

2.3 Materials and Methods

2.3.1 Definition and statistical modeling of LD and CS

In this section, definitions of LD and CS are presented for a population with known pedigree information.

LD is defined as non-random association between allele states at two loci in the pedigree founders (Habier et al., 2010), which are the individuals that do not have parents recorded in
the pedigree. Following Meuwissen et al. (2001), the statistical model that uses LD information for prediction of GEBV of a quantitative trait is written as

\[ y = X\beta + Z\alpha + e, \]

(2.1)

where \( y \) is an \( n \times 1 \) vector of trait phenotypes of \( n \) training individuals, \( \beta \) is a vector of non-genetic fixed effects, \( X \) is the design matrix for fixed effects, \( Z \) is an \( n \times m \) matrix with each row containing genotypes at \( m \) SNPs of each training individual, \( \alpha \) is an \( m \times 1 \) vector of allele substitution effects of the \( m \) SNPs, and \( e \) is an \( n \times 1 \) vector of residuals. Informative prior distributions are usually given to \( \alpha \) to allow simultaneous estimation of all SNP effects. Details of model hierarchies are given in the following section.

In the LD model (2.1), QTL effects are not explicitly fitted in the model but SNPs are used as surrogates for QTL due to LD. QTL effects captured by SNP genotypes can be viewed as conditional expectation of QTL effects on SNP genotypes. When LD between QTL and SNPs is imperfect, true QTL effects deviate from the expected value of QTL effects that are captured by SNP genotypes. Therefore the LD model can only capture part of the genetic variance at QTL under imperfect LD.

Co-segregation of alleles at two loci means that these alleles share identical grand-parental allele origins, i.e. they both originate from the same chromosome of a parent. The indicator of parental allele origin at one locus is a Bernoulli variable. In this study, the allele origin indicator equals 0 if it originates from its grand-maternal allele, and 1 if from its grand-paternal allele. When the allele origins at a SNP locus of parents and offspring are known, the probability that the allele origin of a putative QTL linked to the SNP is grand-paternal (equals 1) can be calculated using recombination rates between QTL and SNPs, which is termed the probability of descent of QTL allele (PDQ). In this study, putative QTL are assumed to be located within every 1-cM genomic window. Suppose that allele origins are known for an individual’s maternal alleles at two SNPs \( M_1 \) and \( M_2 \). Then, when assuming no interference, the PDQ at the putative QTL locus is calculated as follows when the origins of both SNP alleles are the mother’s maternal allele, i.e. \( O_1^m = 0 \) and \( O_2^m = 0 \),

\[
\Pr(O_Q^m = 0 | O_1^m = 0, O_Q^m = 0) = \frac{\Pr(O_1^m = 0, O_Q^m = 0, O_2^m = 0)}{\Pr(O_1^m = 0, O_2^m = 0)} = \frac{(1 - r_1)(1 - r_2)}{1 - r_{12}},
\]

(2.2)
where \( O^m_i \) is the maternal allele origin at \( M_i \) for \( i = 1, 2 \), \( O^m_Q \) is the maternal allele origin at the QTL, \( r_1 \) is the recombination rate between \( M_1 \) and QTL, \( r_2 \) is the recombination rate between QTL and \( M_2 \), and \( r_{12} \) is the recombination rate between \( M_1 \) and \( M_2 \). Recombination rates \( r_1, r_2 \) and \( r_{12} \) can be calculated from the map distance between \( M_1 \) and \( M_2 \) using mapping functions (Haldane, 1919).

The method for modeling CS uses PDQs to follow the transmission of founder alleles at putative QTL. The true breeding value (TBV) of an individual is the summation of its maternal and paternal allelic values (denoted as \( v^m \) and \( v^p \), respectively) at putative QTL across the genome. At every putative QTL, the two allelic values of founders are assumed independent. The allelic values of non-founders are linear combinations of independent allelic values of founders with coefficients determined by the PDQs. The coefficients of maternal and paternal allelic values of a non-founder individual \( i \) (\( w'^m_i \) and \( w'^p_i \), respectively) are calculated recursively as

\[
\begin{align*}
    w'^m_i &= PDQ^m_i w'_m + (1 - PDQ^m_i) w'_p, \\
    w'^p_i &= PDQ^p_i w'_m + (1 - PDQ^p_i) w'_p,
\end{align*}
\]

where \( PDQ^m_i \) (\( PDQ^p_i \)) is the maternal (paternal) PDQ for non-founder \( i \), and \( w'^m_{dam} \) (\( w'^p_{sire} \)) is the dam’s maternal (sire’s paternal) allelic value.

Vectors \( w'^m_i \) and \( w'^p_i \) comprise the rows in the incidence matrix \( W_H \) that relates allelic values of all individual with those of founders. The incidence matrix that relates TBV with founder allelic values (\( W \)) is calculated by the summation of every two rows in \( W_H \) that correspond to the paternal and maternal allelic values of the same individual, i.e.

\[
W = (I_n \otimes [1, 1]) \times W_H,
\]

where \( I_n \) is the identity matrix with dimension \( n \). TBV of all individuals (\( g_{CS} \)) can be written as

\[
g_{CS} = Wv,
\]

where \( v \) is the vector of founder allelic values at all putative QTL.
The method for modeling CS is illustrated using a contrived pedigree of 4 individuals, in which one QTL co-segregates with a SNP with recombination rate \( r = 0.1 \) (Table 2.1). The incidence matrix \( W_H \) is given in Table 2.2. Taking the 5th row of \( W_H \) as an example, it contains coefficients for the maternal allelic value of individual 3, which is a linear combination of allelic values of its dam individual 3. According to equation (2.3),

\[
\mathbf{w}_3^m = (1 - r) \times \mathbf{w}_2^m + r \times \mathbf{w}_2^p.
\]

Allelic values for founders, individual 1 and 2, are included in \( \mathbf{v} \)

\[
\mathbf{v} = \begin{bmatrix}
\mathbf{v}_1^m \\
\mathbf{v}_1^p \\
\mathbf{v}_2^m \\
\mathbf{v}_2^p
\end{bmatrix}.
\]

In this study, putative QTL are assumed to be positioned within every non-overlapping 1-cM genome window. Transmission of putative QTL alleles is followed by PDQ calculated using allele origins of SNPs that cover each 1-cM window. When there is no recombination within a 1-cM window, PDQ is either 0 or 1, indicating certain transmission of QTL allele. When recombination occurs within a 1-cM window, PDQs are 0.5, meaning that the recombinant allelic value is the average of the two parental allelic values.

The CS model is given by

\[
y = \mathbf{X} \beta + \sum_{j=1}^{n_q} \mathbf{W}_j \mathbf{v}_j + \mathbf{e},
\]

(2.5)

where \( y \) is an \( n \times 1 \) vector of trait phenotypic values of \( n \) training individuals, \( \beta \) and \( \mathbf{X} \) are the same as in the LD model (2.1), \( \mathbf{v}_j \) is a vector of founder allelic values at the \( j \)th putative QTL, with \( n_q \) the number of putative QTL, or equivalently the length of the genome in cM, \( \mathbf{W}_j \) is the coefficient matrix for \( \mathbf{v}_j \), and \( \mathbf{e} \) is an \( n \times 1 \) vector of residuals. As in the LD model, informative prior distributions are given to \( \mathbf{v}_j \)'s to allow simultaneous estimation of founder allelic values. Details of the prior distributions are given in the following section.
The model that fits both LD and CS (LD-CS model) includes the LD and CS terms as in models (2.1) and (2.5), respectively,

\[ y = X\beta + Z\alpha + \sum_{j=1}^{n_q} W_j v_j + e. \]  

(2.6)

### 2.3.2 Bayesian inference for the LD-CS model

This section gives a description of how the Bayesian methods “BayesA” and “BayesB” in Meuwissen et al. (2001) were adapted for inference using the LD-CS model. Inference using either the LD or CS model is straightforward by excluding the CS or LD term from the LD-CS model, respectively.

In this study, \( X = 1 \) is a vector of ones with length \( n \), and \( \beta = \mu \) is the overall mean for all training individuals. Indicator variables \( \delta_l \) for \( \alpha_l \) \((l = 1, 2, \cdots, m)\) and \( \psi_{jk} \) for \( v_{jk} \) \((k = 1, 2, \cdots, n_j)\) are introduced, to indicate when the corresponding random effect is fitted in the model; \( \delta_l \) or \( \psi_{jk} \) equals 1 when the effect is included in the model and 0 otherwise.

The residual sampling distribution for \( y \) is multivariate normal:

\[ y | \mu, \alpha, v, \sigma_e^2 \sim N \left( 1\mu + Z\Delta \alpha + \sum_{j=1}^{n_q} W_j \Psi_j v_j, I\sigma_e^2 \right), \]

(2.7)

where

\[ \Delta = \text{Diagonal}\{\delta_l\}_{l=1}^m, \]

\[ \Psi_j = \text{Diagonal}\{\psi_{jk}\}_{k=1}^{n_j}, \]

where \( n_j \) is the dimension of vector \( v_j \).

The prior distribution for \( \mu \) is flat, \( \pi(\mu) \propto \text{constant} \). The prior distribution for \( \sigma_e^2 \) is a scaled inverse chi-square distribution with degrees of freedom \( \nu_e \) and scale parameter \( S_e^2 \):

\[ \pi(\sigma_e^2) = \chi_{\nu_e}^{-2}(S_e^2), \]

where \( \chi_a^{-2}(b) \) is probability density function of scaled inverse chi-square distribution with \( a \) degrees of freedom and scale parameter \( b \).

The prior distribution for \( \alpha_l \) conditional on its variance \( \sigma_l^2 \) is normal:

\[ \pi(\alpha_l | \sigma_l^2) = N(0, \sigma_l^2), \]
and the prior distribution of \( \sigma_l^2 \) is a scaled inverse chi-square distribution:

\[
\pi(\sigma_l^2) = \frac{1}{\chi_{\nu_l}^2(S_{\alpha}^2)}, \quad l = 1, 2, \ldots, m.
\]

The prior distribution for \( v_{jk} \) conditional on its variance \( \sigma_{jk}^2 \) is normal:

\[
\pi(v_{jk}|\sigma_{jk}^2) = \mathcal{N}(0, \sigma_{jk}^2),
\]

and the prior distribution of \( \sigma_{jk}^2 \) is a scaled inverse chi-square distribution:

\[
\pi(\sigma_{jk}^2) = \frac{1}{\chi_{\nu_{jk}}^2(S_{\nu}^2)}, \quad j = 1, 2, \ldots, n_q, \text{ and } k = 1, 2, \ldots, n_j.
\]

The prior distributions for \( \delta_l \) and \( \psi_{jk} \) are Bernoulli distributions:

\[
\pi(\delta_l) = (1 - \pi_{\text{SNP}}) \delta_l \pi_{\text{SNP}}^{(1-\delta_l)}, \quad l = 1, 2, \ldots, m,
\]

\[
\pi(\psi_{jk}) = (1 - \pi_{\text{CSE}}) \psi_{jk} \pi_{\text{CSE}}^{(1-\psi_{jk})}, \quad j = 1, 2, \ldots, n_q, \text{ and } k = 1, 2, \ldots, n_j,
\]

where \( \pi_{\text{SNP}} \) and \( \pi_{\text{CSE}} \) are the expected proportions of SNPs and QTL allelic values that have null effects on the trait, respectively. Method BayesB assumes that \( \pi_{\text{SNP}} \in (0, 1) \) and \( \pi_{\text{CSE}} \in (0, 1) \), while method BayesA assumes that \( \pi_{\text{SNP}} = 0 \) and \( \pi_{\text{CSE}} = 0 \).

The joint posterior distribution for model parameters \( \theta = (\mu, \alpha, v_{jk}, \delta_l, \psi_{jk}, \sigma^2_l, \sigma^2_{jk}, \sigma^2_e, \delta_l, \psi_{jk})' \) is given by

\[
p(\theta|y) \propto L(y|\mu, \alpha, v_{jk}, \delta_l, \psi_{jk}, \sigma^2_l) \pi(\alpha) \pi(\sigma^2_l) \pi(v_{jk} | \sigma^2_{jk}) \pi(\sigma^2_{jk}) \pi(\delta_l) \pi(\psi_{jk}) \pi(\sigma^2_e) \times \exp \left\{ -\frac{1}{2\sigma^2_e} \sum_{i=1}^{n} \left( y_i - \mu - \sum_{l=1}^{m} z_i \delta_l \alpha_l - \sum_{j=1}^{n_q} \sum_{k=1}^{n_j} w_{jk} \psi_{jk} v_{jk} \right)^2 \right\}
\]

\[
\times \prod_{l=1}^{m} \left( \sigma^2_l \right)^{-1/2} \exp \left\{ -\frac{\alpha^2_l}{2\sigma^2_l} \right\} \times \prod_{l=1}^{m} \left( \sigma^2_l \right)^{1+\nu_{\alpha}/2} \exp \left\{ -\frac{\nu_{\alpha} S_{\alpha}^2}{2\sigma^2_l} \right\}
\]

\[
\times \prod_{j=1}^{n_q} \prod_{k=1}^{n_j} \left( \sigma^2_{jk} \right)^{-1/2} \exp \left\{ -\frac{\psi^2_{jk}}{2\sigma^2_{jk}} \right\} \times \prod_{j=1}^{n_q} \prod_{k=1}^{n_j} \left( \sigma^2_{jk} \right)^{1+\nu_{\psi}/2} \exp \left\{ -\frac{\nu_{\psi} S_{\psi}^2}{2\sigma^2_{jk}} \right\}
\]

\[
\times \prod_{l=1}^{m} \pi_{\text{SNP}}^{(1-\delta_l)} \left( 1 - \pi_{\text{SNP}} \right)^{\delta_l} \times \prod_{j=1}^{n_q} \prod_{k=1}^{n_j} \pi_{\text{CSE}}^{(1-\psi_{jk})} \left( 1 - \pi_{\text{CSE}} \right) \psi_{jk}
\]

\[
\times \left( \sigma^2_e \right)^{1+\nu_{\psi}/2} \exp \left\{ -\frac{\nu_{\psi} S_{\psi}^2}{2\sigma^2_e} \right\},
\]

where \( L(y|\mu, \alpha, v_{jk}, \delta_l, \psi_{jk}, \sigma^2_e) \) is the probability density function of the multivariate normal distribution of (2.7).
Inference on each of the model parameters $\theta$ is based on random samples from its marginal posterior distribution, which are obtained through Gibbs sampling from the full conditional posterior distributions of each element in $\theta$. The full conditional distribution for each parameter in $\theta$ is given in the Appendix.

2.3.3 Simulation

Contributions of LD and CS information to prediction accuracy were investigated using simulated datasets of paternal half-sib designs with different number of independent sire families. The simulated genome was comprised of 2 chromosomes, each 1 Morgan in length. The number of QTL on each chromosome was 100. Initially, each chromosome was evenly covered by 2,000 SNPs, among which 5 times the desired number of QTL were randomly positioned as QTL candidates. The mutation rate was $2.5 \times 10^{-5}$ per locus per meiosis for both QTL and SNPs. The number of crossovers per chromosome was sampled from a Poisson distribution with mean 1.0, and the positions of the crossovers were sampled from a uniform distribution. QTL and SNPs were bi-allelic, with initial allele frequencies of 0.5 and genotype frequencies in Hardy-Weinberg equilibrium.

Four scenarios were simulated that differed in the minor allele frequencies (MAF) of QTL and the level of LD between QTL and SNPs (Table 2.3). The sires and dams of the half-sib families were randomly sampled from a base population with size 3,000. In the “LE” scenario, the base population was generated by randomly sampling QTL and SNP alleles of each individual with MAF equal to 0.5, which resulted in linkage equilibrium (LE) among all loci in the base population. In the “Common QTL”, “Rare QTL” and “Resampled QTL” scenarios, the base population was generated by random mating for many generations to create LD among loci, as described in the following. Following closely the simulations of Habier et al. (2007, 2010) and Sun et al. (2012), the initial generations comprised a population of effective size $N_e = 500$ that was randomly mated for 500 generations to create LD between closely linked loci; subsequently, the population was shrunk to $N_e = 200$ and randomly mated for 100 generations to create LD between loci with long genetic distances. In the next 15 generations, the population was gradually expanded to an actual size of 3,000 as the base population. From the base
population, a number of s sires and 20 × s dams were randomly sampled without replacement as the parents of half-sib offspring. Each of the s sires was mated with 20 independent dams, with each dam producing one offspring. Within each sire family, 10 random half-sib offspring were used in the training population and the other 10 in validation. Independent datasets were generated for different numbers of independent sire families \( s = 1, 2, 5, 10, 50, 100 \) and 200, corresponding to training population sizes of 10, 20, 50, 100, 500, 1,000 and 2,000, respectively.

In the base population, 1,000 SNPs with MAF larger than 0.05 and the desired numbers of QTL on each chromosome were randomly sampled according to the scenarios in Table 2.3. QTL effects were randomly sampled from the standard normal distribution. TBV of an individual was obtained as the summation of QTL effects across the genome. In the base population, QTL effects were scaled to achieve the variance of TBV equal to 50.0. Normally distributed random residuals with mean 0 and variance 50.0 were added to TBV to generate phenotypes of a quantitative trait with heritability 0.5. Thirty replicated datasets were independently simulated for each scenario. All replicated datasets used the same initial SNP positioning, but randomly differed in the position of QTL and SNPs that were selected to meet MAF requirements of each scenario and in the effects of QTL.

Values for hyperparameters \( \nu_\alpha \), \( \nu_c \) and \( \nu_e \) were 4.2 following Meuwissen et al. (2001). For the LD and CS models, values of \( S^2_\alpha \) and \( S^2_c \) were chosen such that

\[
\frac{\nu_\alpha S^2_\alpha}{\nu_\alpha - 2} = \frac{h^2 V_P}{2(1 - \pi_{\text{SNP}}) \sum_{l=1}^{m} p_l(1 - p_l)}, \quad \text{and}
\]

\[
\frac{\nu_c S^2_c}{\nu_c - 2} = \frac{h^2 V_P}{(1 - \pi_{\text{CSE}}) n_q},
\]

respectively. For the LD-CS model, values of \( S^2_\alpha \) and \( S^2_c \) were chosen such that

\[
\frac{\nu_\alpha S^2_\alpha}{\nu_\alpha - 2} = \frac{0.5 h^2 V_P}{2(1 - \pi_{\text{SNP}}) \sum_{l=1}^{m} p_l(1 - p_l)}, \quad \text{and}
\]

\[
\frac{\nu_c S^2_c}{\nu_c - 2} = \frac{0.5 h^2 V_P}{(1 - \pi_{\text{CSE}}) n_q},
\]

where \( V_P \) was the phenotypic variance in the training population, \( p_l \) was the MAF of SNP \( l \), and \( h^2 \) was the trait heritability that equaled 0.5 in the simulation. The value of \( S^2_e \) was chosen such that

\[
\frac{\nu_e S^2_e}{\nu_e - 2} = (1 - h^2) V_P.
\]
In the simulated datasets, allele origins at all SNPs are assumed known without error. Method BayesA was used to estimate SNP effects and founder allelic values at QTL. The Gibbs sampler was run for 21,000 iterations, with the first 1,000 discarded as burn-in. The estimated breeding values (EBV) were also obtained from BLUP based on pedigree relationships using ASReml (Gilmour et al., 2009).

2.4 Results

In the Common QTL scenario, accuracy of the LD model increased from below 0.2 and quickly plateaued around 0.8 after the number of half-sib families exceeded 50, which corresponds to a training size of 500. Accuracy of the LD-CS model was indistinguishable from the LD model at all training sizes. Accuracy of the CS model increased from 0.2 and plateaued around 0.45 after the training size exceeded 500, which was much lower than accuracy of the LD or LD-CS model. Accuracies from the LD, CS and LD-CS models were higher than those of pedigree BLUP, which was lower than 0.4 for all training sizes (Figure 2.1). The results suggest that when LD between QTL and SNPs is high, the LD model had high accuracy by capturing information from both LD and CS. Modeling CS explicitly in addition to LD did not improve prediction accuracy.

In the Rare QTL scenario, the level of historical LD between QTL and SNPs was low due to all QTL having much lower MAF than SNPs. Accuracy of the LD model increased with training size from zero to about 0.5, which was much lower than that in the Common QTL scenario. Accuracy of the CS model increased with training size and plateaued around 0.45, which was similar to the Common QTL scenario. Accuracy of the LD-CS model increased and became significantly higher than accuracy of both the LD and CS models when the training size exceeded 100 (10 half-sib families) (Figure 2.2). These results suggest that when LD between QTL and SNPs is imperfect, the contribution of CS information becomes more important than in situations where LD between QTL and SNPs is high, and modeling CS explicitly in addition to LD improves prediction accuracy across unrelated families.

In the Resampled QTL scenario, LD among SNPs was high but LD between QTL and SNPs was zero. The CS model had higher accuracy than the LD and LD-CS models, which increased
from 0.15 and plateaued around 0.45 after the training size exceeded 500. Accuracy of the LD-CS model was lower than that of the CS model but was higher than that of pedigree BLUP. Accuracy from the LD model increased to 0.3 when training size was below 500 (50 half-sib families), but decreased to about 0.2 when the training size exceeded 1,000, which was much lower than accuracy of pedigree BLUP (Figure 2.3). These results suggest that when there is no LD between QTL and SNPs, accuracy of the LD model comes from implicitly capturing CS information, but the ability to capture CS information decreases when a large number of unrelated families are included in the training population. The CS model had much higher accuracy than the LD model due to explicitly capturing CS information.

In the LE scenario, the ranking of accuracies of the LD, CS and LD-CS models were similar to that in the Resampled QTL scenario (Figure 2.4). Accuracy of the LD model was higher in the LE than the Resampled QTL scenario. The reason could be that when there was no LD between QTL and SNPs, accuracy of the LD model came from CS information that was implicitly captured by SNP genotypes. More CS information could be captured by a larger number of independently segregating loci (the effective number of loci) fitted in the LD model (Habier et al., 2013). SNP genotypes were independent in the LE scenario, but were highly correlated in the Resampled QTL scenario. Therefore SNP genotypes could explain more CS information in the LE than in the LD scenarios due to a larger effective number of loci in the LE scenario.

2.5 Discussion

In this study, a new method that explicitly models co-segregation information was developed for genomic prediction in pedigree populations. This method models transmission of putative QTL alleles within consecutive non-overlapping genomic window of sufficiently small size (1 cM in this study), within which the recombination rate is so small that the alleles at all polymorphic loci are expected to co-segregate for several generations. Co-segregation of QTL alleles is modeled using parental allele origins of SNPs that cover the genomic window, which are independent of the level of LD between QTL and SNPs. The method of modeling CS at putative QTL using allele origins at observable SNPs is similar to the method by Fernando
and Grossman (1989), but has an advantage that allows estimation of allelic values of founder alleles at putative QTL using single site Gibbs sampling. In Fernando and Grossman (1989), the paternal and maternal allelic values at putative QTL for every pedigree individual were fitted in the model to explain its GEBV. These values are usually correlated among individuals that are related by the pedigree, and therefore, the estimation of allelic values was achieved by solving mixed model equations that requires the inverse of covariance matrix of allelic values for all pedigree members. Computation of the Fernando and Grossman (1989) method is manageable for marker-assisted selection where the number of molecular markers is usually small. However, for genomic prediction using dense SNP panels, the CS model in this study is computationally tractable because GEBV for all pedigree members are modeled using only the allelic values of pedigree founders, which are assumed independent and MCMC methods can be feasibly implemented to estimate the allelic values.

The CS model (2.5) in this study can be written as an equivalent breeding value model

\[ y = X\beta + g_{CS} + e, \]  

(2.9)

where

\[ g_{CS} = \sum_{j=1}^{n_q} W_j v_j, \]

and

\[ \text{Var}(g_{CS}) = G_{CS} = \sum_{j=1}^{n_q} W_j D_j W'_j, \]  

with

\[ D_j = \text{diag}\{\sigma^2_{jk}\}_{k=1}^{n_j}. \]

The covariance matrix of GEBV due to CS \( g_{CS}, \ G_{CS} \), quantifies genetic covariance among individuals due to co-segregation at putative QTL. The genetic covariance between two individuals depends on the number of common founder alleles that the two individuals share through identity-by-descent, averaged across \( n_q \) QTL, with the corresponding QTL effect variances as weights. This equivalent breeding value model (2.9) was used by Luan et al. (2012) in their study of the contributions of CS and LD information to prediction accuracy in Italian Brown Swiss bulls. In Luan et al. (2012), CS is modeled at every SNP locus in the Bovine SNP
50K chip, which was assumed to be surrogates of QTL, and $G_{CS}$ was constructed by averaging across all SNPs with equal variance among allelic values of SNPs. Compared to Luan et al. (2012), the CS model in this study fits putative QTL within short genomic windows, which is much fewer than the number of SNPs in the 50K chip. Modeling CS at every SNP is not necessary because CS information is based on linkage and is conserved over long genomic distance. Furthermore, the CS model (2.5) allows different variances of QTL allelic values depending on the size of the putative QTL effects. Larger QTL effects are estimated with smaller bias, or equivalently larger weights in $G_{CS}$. A third advantage of the CS model (2.5) over Luan et al. (2012) is that the computation time for model (2.5) increases linearly with the number of individuals ($n$) times the number of allelic values ($\sum_{j=1}^{n_q} n_j$), while for the mixed model approach in Luan et al. (2012), computation time increases cubically with $n$ because it requires the inverse of a dense matrix, $G_{CS}$.

It is generally accepted that LD between QTL and SNPs, co-segregation of QTL with SNP alleles, and pedigree relationships at QTL captured by SNPs are the three main sources of information that contribute to accuracy of genomic prediction (Habier et al., 2007, 2010; Luan et al., 2012; Wientjes et al., 2013; de los Campos et al., 2013b; Habier et al., 2013). Most of the previous studies that aim to disentangle these three sources of information are based on multiple regression models on SNP genotypes (the LD model). The LD model only allows evaluating that part of CS information that is implicitly captured by SNP genotypes, which is highly variable depending on the number and density of SNPs, the level of historical LD, population structure and pedigree relationships. The CS model in this study enables complete disentanglement of CS from LD information because explicit modeling of CS information using parental allele origins does not depend on the level of LD between QTL and SNPs. As an intriguing consequence, results in this study are in contrast to some typical findings in several previous studies based on the LD model. For example, Habier et al. (2013) showed that CS information captured by SNP genotypes contributed little to prediction accuracy across half-sib families and prediction accuracy decreased rapidly with increasing training size. In this study, prediction accuracy from the CS model persisted with increasing training size regardless of historical LD. This difference is mainly because Habier et al. (2013) only considered the part
of CS information that is implicitly captured by GBLUP, while the CS model in this study captures most of CS information due to modeling CS explicitly.

In the simulated datasets with different number of unrelated half-sib families, both LD and CS information contributed to accuracy of genomic prediction. Accuracy of the LD model relies on the level of historical LD between QTL and SNPs in the base population. Accuracy of the CS model relies on accurate estimation of founder allelic values that are transmitted to half-sib offspring within the same family. Accuracy of the LD model increases rapidly with increasing training size when LD between QTL and SNPs is high, because high LD is conserved across families and increasing training size brings in more data to improve estimation of SNP effects. However, when historical LD is low or zero, accuracy of the LD model mainly comes from capturing CS information, which only exists within the same half-sib family. With more unrelated half-sib families included in the training population, accuracy of the LD model decreases and becomes lower than accuracy of the CS model. This is because only half-sibs from the same family contributes to prediction accuracy. In the CS model, the allelic values are estimated using information only within the same half-sib family, while the LD model estimates SNP effects by pooling CS information across all families, which is erroneous because linkage phase and LD is highly variable across a large number of unrelated families.

In realistic situations, such as in current livestock populations, both historical LD and recent CS information exist, and combined modeling of LD and CS explicitly is recommended to improve accuracy of genomic prediction. Simulation results in this study suggest that the LD-CS model tends to have highest prediction accuracy in almost all scenarios. Using a similar LD-CS model with (2.9), Luan et al. (2012) show that in a pedigree population of Italian Brown Swiss bulls, LD information does not contribute to accuracy beyond that due to CS information. Another study on genomic prediction of human height by de los Campos et al. (2013b) shows that the GBLUP model predicts well for close relatives, but has almost zero accuracy when predicting completely unrelated individuals. These studies imply that in livestock and human populations, prediction accuracy comes mainly from CS information, and modeling CS explicitly can achieve almost the same accuracy than fitting SNP genotypes. The simulation results in this study further suggest that modeling both LD and CS jointly improves predic-
tion accuracy from modeling either LD or CS when historical LD between QTL and SNPs is imperfect due to most QTL being rare, as represented by the Rare QTL scenario.

There are several computational problems in implementing the CS model (2.5) in field datasets. First, to obtain parental allele origins from SNP genotypes for all pedigree members can be computationally prohibitive. This is usually achieved in two steps. SNP genotypes are first phased into haplotypes, which are then used to infer parental allele origins using pedigree information (Habier et al., 2010; Meuwissen and Goddard, 2010). In addition to many efficient algorithms that are being developed, such as the method of Meuwissen and Goddard (2010), this problem can also become less demanding with the availability of increased SNP density, genome re-sequencing, and identification of multi-allelic markers such as copy number variants and insertions/deletions. Second, the computation time for the MCMC algorithm of the CS model (2.5) increases with the number of pedigree founders, because the number of allelic values at each putative QTL \( n_j \) is twice the number of founders. It is suggested that the putative QTL be modeled at every cM of the genome to reduce the total number of allelic values, as justified by the fact that recombination happens very rarely within a 1-cM genomic window in several consecutive generations. Furthermore, instead of treating founder allelic values as independent, they can be clustered according to the probability of identity-by-descent with respect to some historical common ancestors beyond pedigree founders (Meuwissen and Goddard, 2001, 2007, 2010). But when the number of founders is large, the equivalent breeding value model (2.9) is recommended since the mixed model equations have the number of genotyped individuals as dimension.

### 2.6 Conclusions

A new method that explicitly models CS information is developed for genomic prediction of breeding values. Breeding values in this CS model are modeled as the summation of independent allelic values at putative QTL among pedigree founders, which are transmitted to offspring through co-segregation with SNP alleles. When the training size is increased by adding unrelated half-sib families, accuracy of the CS model increases and plateaus, but accuracy of the LD model that fits SNP genotypes drops when historical LD between QTL and SNPs is
imperfect. Modeling both LD and CS information improves prediction accuracy compared to modeling either LD or CS, especially when historical LD is imperfect and recent CS information contributes substantially to prediction accuracy among families, which is probably the case of current genomic evaluation in most livestock populations.

2.7 Acknowledgments

Instructions from Drs. Jack Dekkers, Rohan Fernando and Dorian Garrick were greatly acknowledged. This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative, National Institute of Food and Agriculture Competitive Grant 2010-65205-20341 and by National Institutes of Health Grant R01GM099992.

2.8 Bibliography


2.9 Tables

Table 2.1  **Simple pedigree and SNP allele origins of 4 individuals.**

<table>
<thead>
<tr>
<th>Individual</th>
<th>Sire</th>
<th>Dam</th>
<th>Maternal allele origin</th>
<th>Paternal allele origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.2  **Coefficient matrix of the CS model for the pedigree in Table 2.1.**

<table>
<thead>
<tr>
<th>Parental allele</th>
<th>$v_1^m$</th>
<th>$v_1^p$</th>
<th>$v_2^m$</th>
<th>$v_2^p$</th>
<th>$\epsilon_3^m$</th>
<th>$\epsilon_3^p$</th>
<th>$\epsilon_4^m$</th>
<th>$\epsilon_4^p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^m$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$1^p$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$2^m$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$2^p$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$3^m$</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0.1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$3^p$</td>
<td>0.1</td>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$4^m$</td>
<td>0.09</td>
<td>0.81</td>
<td>0.09</td>
<td>0.01</td>
<td>0.1</td>
<td>0.9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$4^p$</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.3  **Scenarios in the simulated half-sib designs.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>SNP MAF $^1$</th>
<th>QTL MAF</th>
<th>LD ($r^2$) between SNPs</th>
<th>LD ($r^2$) between SNPs and QTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common QTL</td>
<td>0.05 ~ 0.5</td>
<td>0.05 ~ 0.5</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
</tr>
<tr>
<td>Rare QTL</td>
<td>0.05 ~ 0.5</td>
<td>0.01 ~ 0.05</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
</tr>
<tr>
<td>Resampled QTL</td>
<td>0.05 ~ 0.5</td>
<td>0.05 ~ 0.5</td>
<td>&gt; 0</td>
<td>= 0</td>
</tr>
<tr>
<td>LE</td>
<td>0.05 ~ 0.5</td>
<td>0.05 ~ 0.5</td>
<td>= 0</td>
<td>= 0</td>
</tr>
</tbody>
</table>

$^1$ Minor allele frequency.
Figure 2.1  Prediction accuracy with different numbers of half-sib families in training with 100 common QTL per chromosome. PBLUP, the pedigree BLUP model; LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
Figure 2.2  Prediction accuracy with different numbers of half-sib families in training with 100 rare QTL per chromosome. PBLUP, the pedigree BLUP model; LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
Figure 2.3  Prediction accuracy with different numbers of half-sib families in training with 100 resampled QTL per chromosome. PBLUP, the pedigree BLUP model; LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
Figure 2.4  Prediction accuracy with different numbers of half-sib families in training with SNPs and 100 QTL per chromosome in LE. PBLUP, the pedigree BLUP model; LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
2.11 Appendix

The full conditional distribution for $\mu$ is normal:

$$
\pi(\mu|\cdot, y) = N\left(\frac{1}{n}y_\mu^*, \frac{\sigma_e^2}{n}\right),
$$

where

$$
y_\mu^* = y - \sum_{l=1}^{m} z_l \delta_l \alpha_l - \sum_{j=1}^{n_q} \sum_{k=1}^{n_j} w_{jk} \psi_{jk} v_{jk},
$$

and $\cdot$ denotes all the other parameters in $\theta$ except for $\mu$. Similar rule applies in the following.

The full conditional distribution for $\alpha_l$ is normal:

$$
\pi(\alpha_l|\cdot, y) = N\left(\frac{z_l' y_l^*}{z_l' z_l + \lambda_l}, \frac{\sigma_e^2 z_l' z_l + \lambda_l}{z_l' z_l + \lambda_l}\right),
$$

where

$$
y_l^* = y - \sum_{l' \neq l} z_{l'} \delta_{l'} \alpha_{l'} - \sum_{j=1}^{n_q} \sum_{k=1}^{n_j} w_{jk} \psi_{jk} v_{jk},
$$

and

$$
\lambda_l = \frac{\sigma_e^2}{\sigma_l^2}.
$$

The full conditional distribution for $v_{jk}$ is normal distribution

$$
\pi(v_{jk}|\cdot, y) = \frac{1}{\tilde{\nu}_\alpha} \left(\tilde{S}_\alpha^2\right)^{-1},
$$

where

$$
y_{jk}^* = y - \sum_{l=1}^{m} z_l \delta_l \alpha_l - \sum_{j' \neq j} \sum_{k' \neq k} w_{j' k'} \psi_{j' k'} v_{j' k'},
$$

and

$$
\tau_{jk} = \frac{\sigma_e^2}{\sigma_{jk}^2}.
$$

The full conditional distribution for $\sigma_{l}^2$ is scaled inverse chi-square distribution

$$
\pi(\sigma_{l}^2|\cdot, y) = \chi_{-2}^{-2}(\tilde{S}_\alpha^2),
$$

where

$$
\tilde{\nu}_\alpha = \nu_\alpha + 1,
$$

$$
\tilde{S}_\alpha^2 = \frac{\nu_\alpha \sigma_{\alpha}^2 + \alpha_l^2}{\nu_\alpha + 1}.
$$
The full conditional distribution for $\sigma_{jk}^2$ is scaled inverse chi-square distribution
\[
\pi(\sigma_{jk}^2 | \cdot, y) = \chi_{\tilde{\nu}_c}^{-2}(\tilde{S}_c^2),
\]
where
\[
\tilde{\nu}_c = \nu_c + 1, \\
\tilde{S}_c^2 = \frac{\nu_c S_c^2 + v_{jk}^2}{\nu_c + 1}.
\]

The full conditional distribution for $\sigma_e^2$ is scaled inverse chi-square distribution
\[
\pi(\sigma_e^2 | \cdot, y) = \chi_{\tilde{\nu}_e}^{-2}(\tilde{S}_e^2),
\]
where
\[
\tilde{\nu}_e = \nu_e + n, \\
\tilde{S}_e^2 = \frac{(y^*)'(y^*)}{\nu_e + n},
\]
where
\[
y^* = y - 1 - \sum_{l=1}^{m} z_l \delta_l \alpha_l - \sum_{j=1}^{n_q} \sum_{k=1}^{n_j} w_{jk} \psi_{jk} v_{jk}.
\]

The full conditional distribution for $\delta_l$ is Bernoulli distribution with success probability
\[
\Pr(\delta_l = 1 | \cdot, y) = \frac{1}{1 + \exp \left( \log[h_1] - \log[h_0] \right)},
\]
where
\[
\log[h_1] = (1 - \pi_{\text{SNP}}) \exp \left\{ - \frac{1}{2 \sigma_e^2} (y_l' - z_l \alpha_l)' (y_l' - z_l \alpha_l) \right\},
\]
\[
\log[h_0] = \pi_{\text{SNP}} \exp \left\{ - \frac{1}{2 \sigma_e^2} y_l'^* y_l^* \right\},
\]
with $y_l^*$ as previously defined.

The full conditional distribution for $\psi_{jk}$ is Bernoulli distribution with success probability
\[
\Pr(\psi_{jk} = 1 | \cdot, y) = \frac{1}{1 + \exp \left( \log[g_0] - \log[g_1] \right)},
\]
where

\[
\log[g_1] = (1 - \pi_{\text{CSE}}) \exp \left\{ -\frac{1}{2\sigma^2_e} (y^*_j - w_{jk}v_{jk})' (y^*_j - w_{jk}v_{jk}) \right\},
\]

\[
\log[g_0] = \pi_{\text{CSE}} \exp \left\{ -\frac{1}{2\sigma^2_e} y^*_j y^*_j \right\},
\]

with \( y^*_j \) as previously defined.
CHAPTER 3. GENOMIC PREDICTION OF BREEDING VALUES IN DIFFERENT MATING DESIGNS COMBINING LINKAGE DISEQUILIBRIUM AND CO-SEGREGATION

3.1 Abstract

Co-segregation (CS) information can have substantial contribution to accuracy of genomic prediction when modeled explicitly. Previous studies have shown that modeling CS explicitly improves prediction accuracy across unrelated families, and modeling linkage disequilibrium (LD) in addition to CS does not further improve prediction accuracy when historical LD between quantitative trait loci (QTL) and single nucleotide polymorphisms (SNPs) is low. CS information is expected to be more important in populations with smaller current effective population sizes, $N_e$, which are usually resulted from mating designs that generates many closely related individuals in each generation. The effect of current $N_e$ on accuracy from modeling CS explicitly has not been studied. In this study, datasets of deep pedigrees with different current $N_e$ were simulated that varied in the number of parents and family sizes in each generation. The effect of current $N_e$ on the importance of CS versus LD information was investigated by accuracy of genomic prediction when LD and CS were modeled explicitly. Results showed that the LD model had persistently high accuracy across validation generations only when historical LD between QTL and SNPs was high. Modeling CS explicitly resulted in higher accuracy than the LD model across validation generations when the mating design generated many close relatives. These results suggested that modeling both LD and CS explicitly is expected to improve prediction accuracy when current $N_e$ is small, and LD between QTL and SNPs is low, which is the typical situation in most livestock populations.
3.2 Introduction

The feasibility in obtaining genotypes of dense single nucleotide polymorphisms (SNPs) with genome-wide coverage has improved accuracy of estimated breeding values by genomic prediction (Hayes et al., 2009b; VanRaden et al., 2009; Daetwyler et al., 2010; Garrick, 2011; Wolc et al., 2011; Ostersen et al., 2011). To date, most statistical models for genomic prediction are based on multiple regression of phenotypes on SNP genotype covariates (SNP models). The estimated SNP effects are used to predict genomic estimated breeding values (GEBV) for selection candidates, which are usually progenies of the individuals in the training population (Meuwissen et al., 2001). Linkage disequilibrium (LD) between quantitative trait loci (QTL) and SNPs was initially thought to be the only source of genetic information that contributed to accuracy of genomic prediction using SNP models, until Habier et al. (2007) and Habier et al. (2013) showed that co-segregation (CS) of QTL with SNPs and pedigree relationships that were implicitly captured by SNP genotypes also contributed to prediction accuracy. According to Habier et al. (2013), prediction accuracy due to high historical LD stayed almost constant across unrelated families or multiple validation generations. CS information contributes to prediction accuracy among related individuals, because CS exists between linked QTL and SNPs that are transmitted together from parents to offspring. Habier et al. (2013) showed that the genomic best linear unbiased prediction (GBLUP) model can capture part of CS information, but accuracy of GBLUP due to CS information decreases across unrelated families and multiple validation generations. Prediction accuracy of GBLUP due to pedigree relationships was least persistent across validation generations compared to that due to LD or CS, because pedigree relationships decrease by half per generation (Habier et al., 2007, 2013).

In analyses of field datasets using SNP models, high accuracy of genomic prediction has mainly been observed among close relatives (Luan et al., 2009; VanRaden et al., 2009; Habier et al., 2010; Hayes et al., 2009b), and prediction accuracy decreases rapidly when the validation individuals are separated from training by many generations (Habier et al., 2010; Wolc et al., 2011; Wientjes et al., 2013; Weng et al., 2014a). The latter does not agree with results from simulation studies in which LD was high between QTL and SNPs (Meuwissen et al., 2001;
Habier et al., 2007; Muir, 2007; Habier et al., 2013). These results suggest that LD between QTL and SNPs is low in current livestock populations, and prediction accuracy of the SNP model mainly comes from CS and pedigree relationships that are implicitly captured by SNP genotypes (Habier et al., 2007, 2010; Daetwyler et al., 2012; Luan et al., 2012; Wientjes et al., 2013).

The low LD between QTL and SNPs in livestock populations is probably resulted from the difference in their minor allele frequencies (MAF). QTL for economically important traits are likely to have low MAF either because the traits have undergone long directional selection (Hayes et al., 2010; Daetwyler et al., 2014; Druet et al., 2014), or because some of QTL are mutations that occur more recently than SNPs (Hayes et al., 2010; Druet et al., 2014). SNPs included in SNP chips usually have high MAF that are chosen from sequencing and prototype genotyping of reference samples (Matukumalli et al., 2009). Since LD between two loci with different MAF is low, LD information has little contribution to prediction accuracy when most QTL have much lower MAF than SNPs. Modeling CS explicitly can increase accuracy under low historical LD because CS information follows transmission of QTL alleles among related individuals, which is independent of the level of LD between QTL and SNPs.

Modeling CS explicitly has been shown to improve prediction accuracy over SNP models in validation populations that are closely related with the training population. Luan et al. (2012) proposed a CS model where the covariance matrix of breeding values was constructed using CS information at all genotyped SNPs across the genome. In their analysis of Italian Brown Swiss bulls, Luan et al. (2012) showed that LD information did not increase prediction accuracy in several offspring generations of the training individuals when CS was already explicitly modeled for genomic prediction. In a previous study by Sun et al. (2014c), a new method was developed to explicitly model CS information by following transmission of putative QTL alleles. Prediction accuracy of the CS model increased and plateaued as the number of unrelated half-sib families in the training population increased, whereas prediction accuracy of the LD model dropped. Results of Luan et al. (2012) and Sun et al. (2014c) suggest that CS can have substantial contribution to prediction accuracy when modeled explicitly, especially when historical LD between QTL and SNPs is low.
Persistence of accuracy across validation generations without retraining (long-term accuracy) is an important criterion to evaluate contributions of different sources of genetic information to prediction accuracy. Habier et al. (2013) showed that LD information was more persistent than CS information implicitly captured by SNP genotypes because CS information decayed across generations due to recombinations within large chromosome segments. Modeling CS explicitly at small putative QTL regions is expected to improve long-term accuracy because recombinations are less likely to happen within small chromosome segments. The contribution of CS information to long-term accuracy by modeling explicitly has not been studied.

Current effective population size \((N_e)\) is another important factor that affects the contribution of CS information to long-term accuracy. For a given size of the training and validation populations, the individuals are more closely related with a smaller current \(N_e\), and CS information is expected to contribute more to long-term accuracy because fewer founder alleles are inherited, each by relatively more offspring. The long-term accuracy is expected to be higher under smaller \(N_e\) because founder allelic values can be estimated more accurately due to more data available for each allele. In livestock populations, current \(N_e\) are affected by mating designs and can vary greatly among different breeding programs. Therefore it is important to study the effect of current \(N_e\) on the contribution of CS information to long-term accuracy before the final application of modeling CS to the improvement of accuracy in livestock breeding programs.

The objectives of this study are 1) to investigate contributions of LD and CS information to long-term accuracy under different current \(N_e\) created by realistic mating designs, and 2) to investigate the effects of historical LD and MAF of QTL on the advantage of modeling LD and CS explicitly in improving long-term accuracy.

### 3.3 Materials and Methods

#### 3.3.1 Simulation of three mating designs

To study the effect of current \(N_e\) on long-term accuracy due to LD or CS information, three mating designs were simulated. The mating design were represented by three pedigrees with
13 non-overlapping generations but different numbers of parents and numbers of offspring per mating. The founders (1st generation) of all three pedigrees were comprised of 5 sires, each mated with 10 dams. Sires and dams of the 1st generation were randomly sampled from a base population of size 2,000. Every mating in the 1st generation produced 6 male and 6 female progenies (2nd generation). The three different mating designs started from the mating of the selected individuals from the 2nd generation, which were used as parents for the 3rd generation.

In pedigree 1, 5 sires and 50 dams were randomly selected in each generation starting from the 2nd generation. Each sire was mated with 10 different dams, each producing 6 male and 6 female progenies. Pedigree 1 represented a balanced nested design where a few sires were selected in each generation and each sire contributed equally to the next generation. The current $N_e$ for pedigree 1 is calculated as $N_e = \frac{4 \times 5 \times 50}{5 + 50} = 18.2$, following Falconer and Mackay (1996).

In pedigree 2, all 300 sires and 300 dams from generation 2 were used as parents. Each sire was mated with 1 dam, producing 1 male and 1 female progenies. Pedigree 2 represented an outbred population where all individuals survived, but each individual had relatively limited contribution to future generations. The current $N_e$ of pedigree 2 is

$$N_e = \left( \frac{1}{13} \sum_{i=1}^{13} \frac{1}{N_{ei}} \right)^{-1} = 173.3,$$

where $N_{e1} = 18.2$ and $N_{ei} = 600$ for $i = 2, 3, \ldots , 13$ (Falconer and Mackay, 1996).

In pedigree 3, 5 sires and 70 dams were randomly selected in each generation starting from the 2nd generation. One out of the five sires was mated with 50 dams, each dam producing 5 male and 5 female progenies, representing an influential sire family. Each of the other 4 sires was mated with 5 dams, each dam producing 2 male and 3 female progenies, representing 4 small sire families. Pedigree 3 represented an unbalanced nested design where the genetics of one individual dominated the future generations. The current $N_e$ of pedigree 3 is much less than 18.2.

The simulated genome comprised 2 chromosomes each 1 Morgan in length. Each chromosome was evenly covered by 4,000 SNPs. 50 candidate QTL were randomly positioned within each cM of the genome. The mutation rate for QTL and SNPs was $2.5 \times 10^{-5}$ per meiosis.
per locus. The number of crossovers per chromosome was sampled from a Poisson distribution with mean 1.0, and the positions of the crossovers were sampled from a uniform distribution.

For each pedigree, two scenarios were simulated for LD between SNPs in the base population, from which the founders are randomly sampled. In the scenario of high LD between SNPs, the base population was generated as follows. The initial generations comprised a population with $N_e = 500$ that was randomly mated for 500 generations to generate LD between closely linked loci, after which the population was shrunk to $N_e = 200$ and randomly mated for another 100 generations to create LD between loci over long genetic distances. In the next 10 generations, the population was gradually expanded to an actual size of 2,000 as the base population. In the scenario of no LD between SNPs, a population of actual size 2,000 was generated as base population with SNP and QTL alleles randomly sampled with frequency 0.5. This results in a population that is both in linkage equilibrium and in Hardy-Weinberg equilibrium.

In the base population, 2,000 segregating SNPs on each chromosome and one segregating QTL within each cM of genome were sampled according to MAF in each scenario (Table 3.1). QTL effects were randomly sampled from a standard Normal distribution. The true breeding value (TBV) was obtained as the summation of all QTL genotypic values for a given individual. Allele substitution effects of QTL were scaled in the base population to achieve genetic variance 4.29. Normally distributed random errors with mean 0 and variance 10.0 were added to TBV to generate phenotypes of a quantitative trait with narrow sense heritability 0.3. For each pedigree, 50 replicated datasets were independently simulated for each scenario in Table 3.1. All replicated datasets used the same initial SNP positioning but varied in QTL effects and, after selection of loci based on MAF, in the positions of QTL and SNPs.

The first 5 pedigree generations, with size 2,455 (pedigree 1 and 2) and 2,475 (pedigree 3), were used for training. Each of the following 8 generations, with size 600, were used for validation. Prediction accuracy were calculated as the correlation coefficient between genomic estimated breeding values (GEBV) and TBV in each validation generation.
3.3.2 Prediction of breeding values

The LD, CS and LD-CS models described in Sun et al. (2014c) were used for prediction of GEBV. Putative QTL were located within every 1-cM genome. Parental allele origins were assumed known without error for all individuals. True allele origins were used to calculate the probabilities of descent of QTL alleles. Method BayesA and BayesB were used. The value of $\pi_{\text{SNP}}$ for BayesB in the LD and LD-CS models was calculated as $1 - \frac{\text{Number of QTL}}{\text{Number of SNPs}}$, indicating the proportion of SNPs that are not in LD with any QTL. The value of $\pi_{\text{CSE}}$ for BayesB in the CS and LD-CS models was 0.95, indicating the proportion of founder alleles without QTL allele. For each replicated dataset, the Gibbs sampler was run for 21,000 iterations with the first 1,000 discarded as burn-in. Point estimates of SNP effects and of founder allelic values at putative QTL were posterior means calculated from the MCMC samples.

3.4 Results

In the Common QTL scenario with high historical LD, the LD model had higher accuracy than either the CS or LD-CS model. Modeling CS in addition to LD resulted in slightly lower accuracy, but accuracies from the LD and LD-CS models only dropped marginally across the 8 validation generations without retraining (Figure 3.1). These results suggest that when historical LD between QTL and SNPs in high, the LD model has persistently high accuracy across validation generations by accurately capturing QTL effects. Accuracy of the CS model was much lower than that of the LD or LD-CS model, and decreased rapidly across validation generations (Figure 3.1). This is because recombinant haplotypes accumulate and dissipate with generations, and the allelic values of recombinant alleles can not be accurately estimated due to limited data available for each allele. Accuracies of the LD and LD-CS models were similar for all three pedigrees, because accuracy was mostly contributed by LD that was generated by historical population. Decreases in accuracy across validation generations were less severe in pedigrees 1 and 3 compared with pedigree 2 (Figure 3.1). The reason is that the number of founder alleles was much smaller in pedigrees 1 and 3 than in pedigree 2 and, thus, the allelic values could be estimated more accurately due to more data available per founder allele.
Similar trends in prediction accuracy were observed for BayesB compared with BayesA, except that the difference in accuracy between the LD and LD-CS models was smaller for BayesB, especially in pedigrees 1 and 3 (Figures 3.1 and 3.2).

In the Rare QTL scenario with high historical LD, the CS and LD-CS models had higher accuracy than the LD model (Figures 3.3 and 3.4). The decrease of accuracy across validation generations was larger for the LD model than for the CS and LD-CS models, especially in pedigrees 1 and 3 (Figures 3.3 and 3.4). These results suggest that when LD between QTL and SNPs is low, accuracy of the LD model mostly comes from capturing CS information, and CS information implicitly captured by the LD model decreases across validation generations due to recombination. In pedigrees 1 and 3, the LD-CS model had slightly higher accuracy than the CS model when using BayesA, but accuracies of the CS and LD-CS models were almost the same when using BayesB (Figures 3.3 and 3.4). In pedigree 2, the LD-CS model had significantly higher accuracy than the CS model for both BayesA and BayesB. This is because when current \( N_e \) is large, as in pedigree 2, the CS model has a disadvantage due to a large number of segregating alleles, each with relatively few data that contribute to estimation of its allelic value. Modeling LD in addition to CS compensates for this disadvantage by implicitly capturing CS information. In conclusion, the contribution of CS information to prediction accuracy is more pronounced in pedigrees with few parents than in pedigrees with many parents, because with few parents, founder allelic values can be estimated more accurately due to more data available.

In the Rare QTL scenario with high historical LD, accuracy of BayesB was higher than that of BayesA for the CS model (Figures 3.3 and 3.4). This is because when QTL alleles have low MAF, the proportion of founder alleles that carry QTL is low. BayesB is more effective than BayesA to accurately estimate founder allelic values with QTL.

In the Common QTL scenario without historical LD, the LD, CS and LD-CS models had almost the same accuracy in either pedigree 1 or 3; while in pedigree 2, the LD model had much lower accuracy than the CS and LD-CS models (Figures 3.5 and 3.6). When there is no historical LD, only CS information contributes to prediction accuracy. Recent LD between linked QTL and SNPs can be created quickly within several generations in pedigrees 1 and 3.
due to high genetic drift, in which case the LD model can capture as much CS information as the CS model. The creation of recent LD is slow in pedigree 2 due to much less drift compared with pedigrees 1 and 3, and hence the LD model can only capture part of CS information.

In the Rare QTL scenario without historical LD, the CS and LD-CS models had similar accuracy, which was higher than accuracy of the LD model (Figures 3.7 and 3.8). This is because high LD cannot be created within several recent generations due to the difference in MAF between QTL and SNPs, and the LD model can only capture limited CS information. For all 3 pedigrees, accuracy of BayesB was higher than that of BayesA for the CS and LD-CS models, but was lower for the LD model. The reason is that the allelic values at rare QTL can be estimated more accurately for the CS model with higher shrinkage of BayesB than BayesA, whereas BayesA has higher accuracy than BayesB for the LD model due to fitting a larger effective number of SNPs that capture more CS information implicitly (Habier et al., 2013).

3.5 Discussion

The main objective of this research was to study contributions of LD and CS information to long-term accuracy with different current $N_e$. The level of historical LD and MAF of QTL are two major factors that affect long-term accuracy from modeling LD and CS explicitly, which were also investigated by simulated datasets of extended pedigrees. Results from the LD model are in agreement with those of Habier et al. (2013), in that the ability of the LD model to capture CS information decreases rapidly with generations. Results from the CS and LD-CS models are firstly presented by this study. Specific reasons for the results in long-term accuracy in different scenarios are discussed in the previous section. In this section, the mechanism by which LD and CS information contribute to long-term accuracy, as well as the effects of historical LD and MAF of QTL on long-term accuracy are discussed.

3.5.1 Simulated mating designs

Three mating designs were simulated that differed in the number of parents per generation and the number of progenies per mating. Pedigrees 1 and 3 resemble the breeding program where a few sires are selected and intensively used for breeding in each generation. CS informa-
tion had significant contribution to prediction accuracy in pedigrees 1 and 3, because a limited number of sire alleles segregate among a large number of their progenies, and the allelic values can be estimated accurately based on the amount of data available. Pedigree 1 is a balanced nested design with identical family sizes, which is similar to the structure of nucleus herds in swine (Cleveland et al., 2012) or poultry (Wolc et al., 2011) breeding programs. Pedigree 3 is an unbalanced design with an influential sire in each generation that has more than 80% of the total progenies, which resembles a dairy cattle population, where artificial insemination is widely used (Schaeffer, 2006). CS information had a larger contribution in pedigree 3 than in pedigree 1, because most progenies in a cohort inherit alleles from only one sire in pedigree 3. In contrast, pedigree 2 resembles an outbred population where all individuals survive and each mating has very few progenies. The number of unique parental alleles is large but each is transmitted only to very few progenies. The allelic values in pedigree 2 cannot be estimated accurately because each allele has only limited data available.

The difference between three mating designs can be quantified by current $N_e$. The $N_e$ of pedigrees 1 and 3 is less than 20, while that of pedigree 2 was close to 200. CS information has a larger contribution to prediction accuracy in a population with a smaller current $N_e$ because individuals tend to be more closely related and share more founder alleles at QTL. The importance of CS information in three mating designs is clearly illustrated in the scenario without historical LD among founders, where the long-term accuracy only stems from CS information. As shown in Figures 3.5 and 3.7, the long-term accuracy by modeling CS explicitly was most persistent in pedigree 3, followed by pedigree 1, and least persistent in pedigree 2. A similar trend was also observed for the CS model when both LD and CS information contributed to prediction accuracy (Figures 3.1 and 3.3).

The contribution of LD information should not depend on the mating design because high historical LD between QTL and SNPs is mostly between closely linked loci and hardly erodes within several recent generations. Since the LD model also implicitly captures information from CS and pedigree relationships (Habier et al., 2013), accuracy of LD model is also affected by the mating design. For example, in the scenario with high historical LD, accuracy of the LD model was higher in pedigrees 1 and 3 than in pedigree 2 (Figures 3.1 and 3.2). These results
agree with Muir (2007), who found that prediction accuracy of the GBLUP model decreased when current $N_e$ increased, and the amount of decrease was larger when QTL and SNPs were in LE than when they were in LD.

In general, when historical LD is high between QTL and SNPs, long-term accuracy is mostly contributed by LD information, and CS information has little contribution regardless of current $N_e$. However, when historical LD is low, CS information contributes most to long-term accuracy, especially when the mating design creates very small current $N_e$.

### 3.5.2 The effect of MAF on contributions of LD and CS information

LD quantifies the correlation between allele states at QTL and SNPs. The LD model captures this correlation using multiple regression on SNP genotypes. The strength of correlation depends on MAF of QTL and SNPs. Strong correlation exists only when QTL and SNPs have similar MAF, as in the Common QTL scenario. The correlation is low when most QTL have low MAF as in the Rare QTL scenario. In the simulated datasets, the correlation between allele states exists in the forms of historical LD and recent CS. Historical LD between closely linked loci is hardly eroded by recombination. The correlation generated by CS can exist between loci over long chromosome regions, which erodes fast with recombination. Both the two forms of correlation can be captured by the LD model. The LD model has persistent long-term accuracy only when historical LD between QTL and SNPs is high, which requires similar MAF between QTL and SNPs. When historical LD is low due to most QTL having low MAF, prediction accuracy of the LD model mainly comes from implicitly capturing CS information, which decreases rapidly across validation generations because CS information across long chromosome regions erodes fast with recombination. Similar results have been observed by (Habier et al., 2013) on the GBLUP model.

CS follows the transmission of QTL alleles among related individuals, which is independent of LD. As a result, prediction accuracy from the CS model is not affected by the level of historical LD. Accuracy due to CS information depends on 1) the size of founder haplotype alleles that are used to follow transmission of putative QTL alleles, which determines the rate of erosion of CS due to recombination; and 2) accuracy in estimating founder allelic values, which depends
on the amount of phenotype data for the progenies that inherit the same founder haplotype. Using the CS model that fit founder alleles of length 1 cM, the number of recombination within alleles is small, and therefore CS information contributes to long-term accuracy provided that the allelic values can be estimated correctly with sufficient data. Persistence of accuracy across validation generations is expected to improve when smaller sized alleles are fitted in the CS model, because the proportion of recombinant alleles across validation generations will decrease.

In real livestock populations, persistently high accuracy across validation generations using the LD model has rarely been observed (Habier et al., 2010; Wolc et al., 2011; Wientjes et al., 2013; Weng et al., 2014a), which suggest that historical LD between QTL and SNPs is low, and prediction accuracy relies mostly on CS information (Luan et al., 2012; Daetwyler et al., 2012). The LD-CS model is recommended to improve long-term accuracy in livestock populations due to capturing both LD and CS information explicitly.

3.5.3 The effect of prior distributions on prediction accuracy

Prior distributions in BayesA and BayesB Meuwissen et al. (2001) are used to allow simultaneous estimation of SNP effects $\alpha$ in the LD model and of founder allelic values $v_j$ in the CS model. BayesA represents a method of shrinkage regression without variable selection. In BayesA, independent $t$ prior distributions are given to $\alpha_l$ and $v_{jk}$. When using the posterior mode as point estimate of a parameter $\beta$, the amount of shrinkage imposed by a scaled $t$ prior distribution with degrees of freedom $\nu$ and scale parameter $S^2$, $t(0, \nu, S^2)$, is proportional to $\log \left(1 + \frac{\beta^2}{\nu S^2}\right)$ (Gianola, 2012, personal communication). The posterior mean used in this study is expected to be close to the mode due to the almost symmetric posterior distribution at convergence of the MCMC (Sun et al., 2012). This means that the estimates of small $\beta$ are heavily shrunk towards zero but large $\beta$ are less shrunk. BayesB represents a variable selection method. In BayesB, the prior for each $\alpha_l$ and $v_{jk}$ is a mixture of point mass at zero and $t$ distribution elsewhere. BayesB results in much heavier shrinkage towards zero than BayesA, and consequently the effective number of loci is larger in BayesA than in BayesB (Habier et al., 2013).
The effect of prior distributions on the LD model is two fold. When historical LD is high between QTL and SNPs, high prediction accuracy is usually achieved by the LD model, with nearly unbiased estimates of large SNP effects, while effectively shrinking small SNP effects towards zero. In the simulated datasets of this study, BayesB had higher accuracy than BayesA because the number of QTL was much less than the number of SNPs. When historical LD is low, prediction accuracy mainly comes from implicitly capturing CS information, which depends on the effective number of SNPs fitted in the model. Then, BayesA had higher accuracy than BayesB due to fitting relatively more SNPs, which can capture more CS information than BayesB (Habier et al., 2013).

The effect of prior distributions on the CS model is insensitive to historical LD, but depends on MAF of QTL. When MAF of QTL is high, BayesA tends to have higher accuracy than BayesB because of fitting more founder allelic values that co-segregate with common QTL alleles. When MAF of QTL is low, BayesB tends to have higher accuracy than BayesA, because only a small proportion of founder alleles carry QTL and their allelic values can be estimated accurately by variable selection of BayesB.

### 3.6 Conclusions

In this study, the effects of current \( N_e \), historical LD, and MAF of QTL on persistence of accuracy across validation generations without retraining were investigated for explicitly modeling LD and CS information. The LD model had persistently high accuracy across validation generations only when historical LD between QTL and SNPs was high, which requires similar MAF between QTL and SNPs. With high historical LD, accuracy of the LD model was much higher than that of the CS model due to capturing both LD and CS information. Accuracy due to LD information persisted across validation generations, whereas accuracy due to implicitly capturing CS information decreased fast with recombination. When historical LD between QTL and SNPs was low, accuracy of the LD model came mostly from capturing CS information, which was much lower and less persistent than that of the CS and LD-CS models. The contribution of CS information increased with smaller current \( N_e \), because there were fewer founder alleles, each was inherited by many progenies, and their allelic values could be esti-
mated more accurately with sufficient data. Since current $N_e$ of most of livestock populations is small and historical LD between QTL and SNPs tend to be low, modeling CS explicitly in addition to LD is recommend to improve long-term accuracy.

### 3.7 Acknowledgments

Instructions from Drs. Jack Dekkers, Rohan Fernando and Dorian Garrick were greatly acknowledged. This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative, National Institute of Food and Agriculture Competitive Grant 2010-65205-20341 and by National Institutes of Health Grant R01GM099992.

### 3.8 Bibliography


### 3.9 Tables

Table 3.1  **Minor allele frequencies (MAF) of QTL and SNPs and the level of historical LD in the base population of simulated scenarios.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Common QTL&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Rare QTL&lt;sup&gt;2&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High LD&lt;sup&gt;3&lt;/sup&gt;</td>
<td>No LD&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td>MAF of QTL</td>
<td>0.01 ~ 0.5</td>
<td>0.01 ~ 0.06</td>
</tr>
<tr>
<td>MAF of SNPs</td>
<td>0.06 ~ 0.5</td>
<td></td>
</tr>
<tr>
<td>LD between QTL and SNPs</td>
<td>&gt; 0</td>
<td>= 0</td>
</tr>
<tr>
<td>LD between SNPs</td>
<td>&gt; 0</td>
<td>= 0</td>
</tr>
</tbody>
</table>

<sup>1</sup> The scenario with MAF of QTL between 0.06 ~ 0.50 and MAF of SNPs between 0.06 ~ 0.50.<br><sup>2</sup> The scenario with MAF of QTL between 0.01 ~ 0.06 and MAF of SNPs between 0.06 ~ 0.50.<br><sup>3</sup> The scenario with high LD in the base population created by historical generations.<br><sup>4</sup> The scenario with LE in the base population by independently sampling genotypes of QTL and SNPs.
3.10 Figures

Figure 3.1 Mean accuracy in three simulated pedigrees from BayesA in the Common QTL scenario with high historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.

Figure 3.2 Mean accuracy in three simulated pedigrees from BayesB in the Common QTL scenario with high historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
Figure 3.3  Mean accuracy in three simulated pedigrees from BayesA in the Rare QTL scenario with high historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.

Figure 3.4  Mean accuracy in three simulated pedigrees from BayesB in the Rare QTL scenario with high historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
Figure 3.5  Mean accuracy in three simulated pedigrees from BayesA in the Common QTL scenario with no historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.

Figure 3.6  Mean accuracy in three simulated pedigrees from BayesB in the Common QTL scenario with no historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
Figure 3.7  Mean accuracy in three simulated pedigrees from BayesA in the Rare QTL scenario with no historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.

Figure 3.8  Mean accuracy in three simulated pedigrees from BayesB in the Rare QTL scenario with no historical LD. LD, the LD model; CS, the CS model; LD-CS, the combined LD-CS model.
CHAPTER 4. IMPROVED ACCURACY OF GENOMIC PREDICTION FOR TRAITS WITH RARE QTL BY FITTING HAPLOTYPES

A paper published in the *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*

X. Sun¹, R. L. Fernando¹, D. J. Garrick¹, J. C. M. Dekkers¹

¹ Iowa State University, Ames, IA, United States

4.1 Abstract

Genomic prediction estimates QTL effects by exploiting LD. High LD can only occur when SNPs and QTL have similar minor allele frequencies (MAF). Marker panels tend to use SNPs with high MAF and will have limited ability to predict rare QTL. In practice, increasing SNP density has not improved prediction accuracy. This might be explained if a trait had many rare QTL. In such cases, linear models fitting haplotypes could have an advantage because haplotypes could be in complete LD with QTL alleles. SNP genotypes were simulated with 200 SNPs per cM. Genomic breeding values were predicted using either SNP genotypes or non-overlapping haplotypes. When QTL had low MAF, prediction accuracy from haplotype models were significantly higher than for SNP models. Results suggest that haplotype models can be an efficient alternative to SNP models especially when traits are controlled by many rare QTL.

Keywords: prediction accuracy, haplotype, rare variant
Implementation of genomic evaluation into breeding programs has been successful because genomic prediction of breeding values is more accurate than pedigree-based parent average for many economically valuable traits. With the rapid progress in genotyping and next-generation sequencing technologies, high-density SNP genotypes have been collected for increasing numbers of animals through chip genotyping, genotyping-by-sequencing or imputation. Accuracy of genomic prediction is expected to increase with increasing SNP density due to the assumption that SNPs in high linkage disequilibrium (LD) with quantitative trait loci (QTL) or even the QTL themselves could be included in the panel, and hence can explain most of the additive genetic variance. However, results from both simulation and field data analyses show limited advantage in prediction accuracy of using 770K or sequencing SNPs over 50K SNPs (VanRaden et al., 2011; Erbe et al., 2012).

Given that most traits in breeding objectives have comprised survival, growth, or reproduction of the individual, they have undergone long natural and intense artificial selection, and QTL affecting such traits are likely to have low minor allele frequencies (MAF). SNPs that are included on SNP chips are usually chosen from sequencing and prototype genotyping of reference samples and have generally been chosen to have high MAF. Since high or complete LD can only exist between two loci that have similar MAF, prediction accuracy for traits controlled by rare QTL is difficult to improve by increasing density of the SNP panel if the additional SNPs have high MAF. Moreover, increasing SNP density exacerbates statistical and computational difficulties for linear models when fitting increasingly large numbers of SNPs.

Both the problems of incomplete LD and expensive computation could be addressed by fitting haplotypes constructed from phased SNP genotypes. First, although rare QTL cannot be in high LD with common SNPs, they can be in high LD with haplotypes (Goddard and Hayes, 2007). Second, with increasing SNP density, the number of observable unique haplotypes eventually asymptotes due to finite population size and becomes less than the number of SNPs, at which point haplotype models will have lower dimension than SNP models.
Previous studies on haplotype models for genomic prediction were based on haplotypes constructed from low density SNP genotypes, in which the LD between haplotype and QTL was incomplete (Calus et al., 2008; Villumsen et al., 2009; Hickey et al., 2013). Although these studies reported advantages in prediction accuracy of haplotype over SNP models for specific haplotype sizes and with modeling of similarity among haplotype alleles, the potential advantage of haplotype models in prediction accuracy and computational efficiency when SNP density approaches sequence data, where there is almost complete LD between haplotype and QTL alleles regardless of the MAF of QTL, has not been studied.

Thus, the objectives of this study were to investigate the effect of MAF of QTL on prediction accuracy and to test the hypothesis that prediction accuracy can be improved with less computational burden by fitting haplotypes.

4.3 Materials and Methods

4.3.1 Simulated datasets

The initial generations comprised a population with effective size 500 that was randomly mated for 500 generations to reach mutation-drift equilibrium, before being reduced to effective size 100 and randomly mated for another 100 generations to generate LD spanning longer genomic distances. The population was then expanded to 2,000 individuals in the following 20 generations to represent the base population. A random sample of 1,500 individuals from this population was sampled, of which 1,000 individuals were used for training and the remaining 500 for validation.

The genome comprised two chromosomes, each with length 100 cM. Initially, 80,000 SNPs were evenly positioned on each chromosome and a sufficient number of QTL candidate loci were randomly positioned within every 1-cM chromosomal segment. All SNPs and QTL were bi-allelic with initial allele frequencies 0.5. QTL effects were randomly sampled from a Gamma distribution with scale 0.4 and shape parameter 1.66, and had equal chance to be positive or negative. Mutation rate was $2.5 \times 10^{-6}$ per locus per meiosis.
In the base population, 20,000 SNPs per chromosome and 1 QTL in each 1-cM segment were randomly sampled according to different assumptions on MAF of SNPs and QTL. Two scenarios were simulated for the MAF of QTL: 1) all QTL had MAF > 0.06 (common QTL), and 2) all QTL had MAF between 0.01 and 0.06 (rare QTL). For both common and rare QTL scenarios, datasets were generated where all 40,000 SNPs had MAF > 0.06 (common SNP). Specifically for the rare QTL scenario, an additional dataset with all 40,000 SNPs having MAF > 0.01 was generated.

In the base population of size 2,000, the effects of the selected QTL were scaled to achieve a total genetic variance of 4.29. True breeding values (TBV) were calculated by summing up all QTL effects for a given individual. Normal random variables with mean zero and variance 10.0 to represent residual effects were added to TBV to generate phenotypic values for a trait with heritability 0.3. Twenty random replicates were simulated for each combination of scenarios of MAF of QTL and MAF of SNPs.

4.3.2 Statistical analyses

Genomic estimated breeding values (GEBV) for validation individuals were predicted using linear mixed models fitting SNP genotypes or haplotypes. Models BayesA and BayesB (Meuwissen et al., 2001) were used to estimate SNP allele substitution or haplotype effects.

In the analyses with models fitting haplotypes, the linkage phase of the 40,000 SNPs was assumed known without error for both training and validation individuals. This assumption is justified because high phasing accuracy could be achieved under simulated SNP density, e.g. Browning and Browning (2007). The haploid genome was divided into non-overlapping segments of 1.0 or 0.2 cM. Unique SNP haplotypes for each segment that had a frequency > 0.01 in the combined training and validation population with size 1,500 were defined as common haplotypes. Either all unique or only common SNP haplotypes were fitted in the model for genomic prediction. Those haplotypes only present in validation population had zero estimated effects, and they didn’t contribute to prediction of GEBV.

Formulation of models BayesA and BayesB based on haplotypes (termed “BayesA\textsubscript{H}” and “BayesB\textsubscript{H}”, respectively) was similar to Meuwissen et al. (2001), except every unique haplotype
allele was considered to have a random effect with an independent $t$ distribution as prior. Value of $\pi$ in BayesB_H was defined as the proportion of unique SNP haplotypes that were not in LD with any QTL alleles, which was set to 0.97 and 0.95 when segment sizes were 1.0 and 0.2 cM, respectively.

Point estimates for SNP allele substitution effects and haplotype effects were their posterior means estimated from Markov chain Monte Carlo samples with chain length 11,000 and the first 1,000 discarded as burn-in. Prediction accuracy of GEBV was represented by the Pearson correlation coefficient between GEBV and TBV in validation individuals.

4.4 Results and Discussion

4.4.1 Haplotype frequencies and concordance between SNP haplotypes and QTL alleles

Frequencies of unique haplotype alleles were calculated for one dataset with MAF of QTL and SNP > 0.06. With SNP haplotype size 1.0 cM, the total number of unique haplotype alleles across all 1.0-cM segments was 10,559, of which 1,628 were common haplotypes (Table 4.1). When haplotype size was 0.2 cM, the numbers of all and common haplotype alleles were 11,069 and 3,722, respectively. Under mutation and random drift, only 15% and one third of haplotype alleles were common when haplotype size was 1.0 and 0.2 cM, respectively (Table 4.1). The dimension of the haplotype model was one quarter of the dimension of the SNP model, and models fitting only common haplotype alleles had less than one tenth dimension of the SNP model, resulting in a potential 10-fold greater computational efficiency for haplotype models. Table 1 shows results from one simulated dataset where SNP density was 20 per cM, 10 times less dense than the aforementioned scenario, which was similar to Villumsen et al. (2009) and Hickey et al. (2013). When SNP density increased 10 fold, the number of unique haplotype alleles increased less than two fold and the number of common haplotype alleles stayed the same, which suggested that the dimension of haplotype model would not increase much with increased SNP density.
Linkage disequilibrium exploited by the haplotype model was investigated by the concordance between haplotype alleles and QTL alleles. Discordant haplotypes were defined as the haplotypes that carried both the major and minor QTL allele within the haplotype region, which meant that the LD between haplotype and the QTL allele was incomplete. The proportion of discordant haplotypes among all unique haplotypes within the population is given in Table 1. When SNP density was 200 per cM, there were no discordant haplotypes, suggesting complete LD between haplotype and QTL alleles, while a small proportion of discordant haplotypes existed when SNP density was 20 per cM.

4.4.2 Prediction accuracy

Prediction accuracy from SNP and haplotype models with different MAF of QTL. When SNP MAF > 0.06, prediction accuracies of SNP models were much higher for traits that were controlled by common QTL than for traits controlled by rare QTL (see first two columns of Table 4.2). This suggests that prediction accuracies from the same SNP panel can vary between traits, depending on the MAF of QTL for the trait, and that traits for which the QTL have similar MAF as SNPs on the panel are expected to have relatively high accuracy. Including SNPs with MAF < 0.06 into the model could increase prediction accuracy of SNP models for traits controlled by rare QTL (third columns of Table 4.2). These results are in agreement with Druet et al. (2014), who found that prediction accuracy could be increased up to 30% using sequencing data when the trait was controlled by many rare QTL, because many more rare SNPs can be captured by sequencing data than SNP chips.

Prediction accuracies from haplotype models generally followed a similar trend as accuracies from SNP models, but were less affected by the MAF of QTL. This could be explained by the fact that haplotypes tended to be in higher or complete LD with QTL than single SNPs, regardless of the MAF of QTL. Haplotype models had no advantage over SNP models when QTL were common variants, but had significant advantage when QTL were rare variants (Table 4.2). Results suggest that for traits where the prediction accuracy hardly improves by increasing chip SNP density, haplotype models may give higher prediction accuracy due to capture of QTL alleles by complete LD.
Models that fitted 0.2-cM haplotypes generally had higher prediction accuracy than models that fitted 1.0-cM haplotypes. There are two possible reasons for this. First, smaller size genome segments had fewer unique haplotype alleles and hence a smaller number of effects to be estimated within one segment, resulting in more accurate estimates of unique haplotype effects because more data is available to estimate their effects. Second, compared with large size segments, recombinations happened less often within small size segments and hence the proportion of discordant haplotypes tended to be smaller. On the other hand, the size of segments needed be large enough to allow enough segregating alleles to be in high or complete LD with QTL alleles. One critical question for haplotype models is the optimal segment size to achieve complete LD while to keep the overall number of haplotypes small. Villumsen et al. (2009) reported that fitting 10-SNP haplotypes of length 1.0-cM gave highest prediction accuracy with a simulated marker density of 10 SNPs per cM. The optimal segment size for haplotype models largely depended on SNP density, level of LD and effective population size, and hence needs to be determined for specific datasets.

In most scenarios, prediction accuracy only decreased marginally when rare haplotypes were excluded from the model. Since few data were available to estimate effects of rare haplotypes, the estimated effects would be shrunk to zero and, thus, excluding rare haplotypes from the model had only minimal effect on prediction accuracy. The advantage of excluding rare haplotypes is the significant improvement in computational efficiency since a large proportion of haplotypes is rare, thus could result in an up to 10-fold reduction in the dimensionality of the model.

4.5 Conclusion

Under SNP density similar to genotyping by a 770K SNP chip or sequencing, haplotype models were shown to have significantly higher prediction accuracy than SNP models for traits controlled by rare QTL, with much less computation effort required. Thus, haplotype models can be efficient alternatives to SNP models when SNP density is high because they result in prediction accuracies that are less sensitive to the MAF of the underlying QTL and are computationally more efficient.
4.6 Acknowledgments

This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative, National Institute of Food and Agriculture Competitive Grant No. 2012-67015-19420 and by National Institutes of Health Grant R01GM099992.

4.7 Bibliography


4.8 Tables

Table 4.1  **Average number of unique (No. Allele) and common (No. Common) haplotype alleles, and the proportion of discordant (Discordant %) haplotype alleles across all genome segments in the scenario of common QTL and common SNPs.**

<table>
<thead>
<tr>
<th>Segment length</th>
<th>No. Alleles</th>
<th>No. Common</th>
<th>Discordant %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>200 SNPs per cM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0cM</td>
<td>52.8</td>
<td>8.1</td>
<td>0</td>
</tr>
<tr>
<td>0.2cM</td>
<td>11.1</td>
<td>3.7</td>
<td>0</td>
</tr>
<tr>
<td><strong>20 SNPs per cM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0cM</td>
<td>36.9</td>
<td>7.4</td>
<td>1.7%</td>
</tr>
<tr>
<td>0.2cM</td>
<td>5.8</td>
<td>3.1</td>
<td>2.7%</td>
</tr>
</tbody>
</table>

Table 4.2  **Mean prediction accuracies\(^1\)** across 20 replicates.

<table>
<thead>
<tr>
<th>MAF QTL</th>
<th>&gt; 0.06</th>
<th>0.01 ~ 0.06</th>
<th>0.01 ~ 0.06</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF SNP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BayesA</td>
<td>0.778</td>
<td>0.491</td>
<td>0.647</td>
</tr>
<tr>
<td>BayesB</td>
<td>0.829</td>
<td>0.613</td>
<td>0.788</td>
</tr>
<tr>
<td>BayesA(_H), 1.0cM(^2), a(^3)</td>
<td>0.729</td>
<td>0.652</td>
<td>0.665</td>
</tr>
<tr>
<td>BayesA(_H), 1.0cM, c(^3)</td>
<td>0.728</td>
<td>0.646</td>
<td>0.659</td>
</tr>
<tr>
<td>BayesA(_H), 0.2cM(^2), a</td>
<td>0.776</td>
<td>0.657</td>
<td>0.685</td>
</tr>
<tr>
<td>BayesA(_H), 0.2cM, c</td>
<td>0.767</td>
<td>0.643</td>
<td>0.674</td>
</tr>
<tr>
<td>BayesB(_H), 1.0cM, a</td>
<td>0.721</td>
<td>0.774</td>
<td>0.792</td>
</tr>
<tr>
<td>BayesB(_H), 1.0cM, c</td>
<td>0.736</td>
<td>0.756</td>
<td>0.774</td>
</tr>
<tr>
<td>BayesB(_H), 0.2cM, a</td>
<td>0.811</td>
<td>0.769</td>
<td>0.798</td>
</tr>
<tr>
<td>BayesB(_H), 0.2cM, c</td>
<td>0.806</td>
<td>0.743</td>
<td>0.772</td>
</tr>
</tbody>
</table>

\(^1\) Standard errors of mean were less than 0.025.
\(^2\) 1.0cM, haplotype models with segment size 1.0cM; 0.2cM, haplotype models with segment size 0.2cM.
\(^3\) a, haplotype models fitting all unique SNP haplotypes; c, haplotype models fitting only common SNP haplotypes.
CHAPTER 5. GENOMIC PREDICTION COMBINING LINKAGE DISEQUILIBRIUM AND CO-SEGREGATION IN POPULATIONS WITHOUT PEDIGREE

5.1 Abstract

In genomic prediction of complex traits, prediction models that fit SNP genotypes have high accuracy only when prediction candidates are closely related with the training population, whereas prediction accuracy is low or zero for individuals that are distantly related with the training population. Further, prediction accuracy hardly improves by increasing the density of SNP chip. Results from genomic prediction in livestock populations suggest that historical LD between quantitative trait loci (QTL) and single nucleotide polymorphisms (SNP) is low, and prediction accuracy comes mainly from co-segregation (CS) information that is implicitly captured by SNP genotypes. Fitting 1-cM haplotypes across the genome to explicitly capture CS information, in addition to fitting SNP genotypes to capture historical LD information, is proposed to improve prediction accuracy. In this study, simulation results show that the combined SNP-haplotype model had similar accuracy as the SNP model when historical LD was high, but had significantly higher accuracy than the SNP model when historical LD was low. In the analyses of several egg quality traits, the SNP-haplotype model improved prediction accuracy for traits for which the SNP model had low accuracy. Fitting haplotypes in addition to SNP genotypes is an effective approach to capture CS information for genomic prediction, especially when LD between QTL and SNPs is low and LD contributes little to prediction accuracy.
5.2 Introduction

Genomic prediction of quantitative traits using high-density single nucleotide polymorphism (SNP) genotypes (Meuwissen et al., 2001) has been widely used in animal (Goddard and Hayes, 2009; Georges, 2014) and plant (Heffner et al., 2009; Jannink et al., 2010) breeding programs, as well as in humans for the prediction of complex traits such as height and disease risk (Wray et al., 2007, 2013). To date, most genomic prediction models are based on multiple regression of trait phenotypes on SNP genotypes (SNP models) (Meuwissen et al., 2001; Kärkkäinen and Sillanpää, 2012; de los Campos et al., 2013a). The main difference among the SNP models that have been used is the prior distribution specified for random SNP effects, which allows simultaneous estimation of all SNP effects by inducing shrinkage or variable selection (Gianola, 2013; de los Campos et al., 2013a). The information that is used by SNP models is the association between allele states of quantitative trait loci (QTL) and SNPs, which usually exists in three forms of genetic information: linkage disequilibrium (LD), co-segregation (CS) and additive relationships (Habier et al., 2013).

In a pedigree population, LD is defined as the non-random association between allele states of QTL and SNPs among pedigree founders. The level of LD is determined by mutation, selection, drift, which depends on historical effective population size ($N_e$), and recombination (Sved, 1971). In human and livestock populations, high historical LD usually exists only between loci with very short genomic distance (Reich et al., 2001; de Roos et al., 2008; Qanbari et al., 2010). Simulation studies have shown that, with high historical LD, accuracy of the SNP model is consistently high across unrelated families, and persistent across multiple validation generations without retraining (Habier et al., 2007; Sun et al., 2014b). LD was regarded as the only source of information that contributed to accuracy of genomic prediction, before Habier et al. (2007) and Habier et al. (2013) showed that SNP models also implicitly captured information from CS and pedigree relationships.

Co-segregation between QTL and SNP alleles refers to alleles at linked loci originating from the same parental chromosome. CS creates associations between allele states of linked loci among relatives, which can span long distances along a chromosome. For example, all
SNPs and QTL on a chromosome co-segregate when there is no recombination. Habier et al. (2013) showed that prediction accuracy of the SNP model that is contributed by implicitly capturing CS information was low for the prediction of unrelated families; within the same family, this accuracy decreased rapidly across validation generations without retraining. The reason is that CS information only exists within families, while CS information in other families is also captured by the SNP model but adds noise to within-family prediction. Further, the SNP model picks up CS information regardless of the genomic distance between loci, but CS information over long genomic distances erodes rapidly due to recombination. Modeling CS information explicitly within short genomic regions improves persistence of accuracy across unrelated families or validation generations. Using a CS model that followed transmission of 1-cM founder alleles at putative QTL, Sun et al. (2013) showed that prediction accuracy of the CS model increased and plateaued as the number of unrelated half-sib families increased in the training population, whereas prediction accuracy of the SNP model dropped. In another study, Sun et al. (2014b) showed that modeling CS in addition to LD resulted in a significant increase in accuracy over eight validation generations without retraining when historical LD was low and recent $N_e$ was small.

Current livestock populations usually include many close relatives, and both LD and CS information contribute to the prediction accuracy of the SNP model. In analyses of human, livestock and plant datasets using the SNP model, high prediction accuracy is often observed for the prediction of individuals that are closely related to the training population, while accuracy was low for the prediction of individuals that were distantly related to the training population (Habier et al., 2010; Albrecht et al., 2011; Saatchi et al., 2011; Clark et al., 2012; Windhausen et al., 2012; Wientjes et al., 2013; de los Campos et al., 2013b; Crossa et al., 2013). Wolc et al. (2011) and Weng et al. (2014a) showed that prediction accuracy decreased rapidly over eight validation generations without retraining in layer chickens. In multiple breed genomic prediction, improved accuracy using a multi-breed over a single breed reference population has been mainly observed in target breeds that were also included in the training population (Brøndum et al., 2011; Zhou et al., 2014), whereas accuracy for breeds that were distantly related to the training population remained low or even zero (Hayes et al., 2009a; Erbe et al.,
Furthermore, increasing SNP density has not significantly improved prediction accuracy. In cattle populations, similar accuracy was observed when using the 770K SNP chip or sequence SNPs compared with using the 50K chip (Erbe et al., 2012; Su et al., 2012; Gunia et al., 2014; Hayes et al., 2014), in contrast to results from simulation studies (Meuwissen and Goddard, 2010; VanRaden et al., 2011; Druet et al., 2014; MacLeod et al., 2014). Results from field datasets suggest that historical LD between QTL and SNPs is low, and prediction accuracy of the SNP model comes mainly from capturing CS information among close relatives, rather than historical LD that is conserved among distantly related individuals (Daetwyler et al., 2012; Luan et al., 2012). Therefore, prediction accuracy is expected to increase when CS is modeled explicitly in addition to LD (Luan et al., 2012; Sun et al., 2013, 2014b), especially when historical LD between QTL and SNPs is low.

Low LD between QTL and SNPs in livestock populations is mainly resulted from the difference in their minor allele frequencies (MAF). QTL for economically important traits are likely to have low MAF either because the traits have undergone long directional selection (Hayes et al., 2010; Daetwyler et al., 2014; Druet et al., 2014), or because some of QTL are mutations that occur more recently than SNPs (Hayes et al., 2010; Druet et al., 2014). The SNPs included on SNP chips usually have high MAF chosen from sequencing and prototype genotyping of reference samples (Matukumalli et al., 2009). Since LD between two loci with different MAF is low, LD information has little contribution to prediction accuracy. Modeling CS explicitly using haplotypes can improve prediction accuracy when most QTL have low MAF, because CS information is hardly affected by the level of LD between QTL and SNPs.

CS models described in Luan et al. (2012) and Sun et al. (2013) use parental allele origin information among relatives, which limits the application of the CS model to populations with pedigree information. When pedigree information is not available for genotyped individuals, SNP haplotypes can be used to capture CS information. With sufficient SNP density, haplotypes with the same allele state are likely to have originated from the same common ancestor without recombination (Goddard and Hayes, 2007), and therefore individuals that inherit the same haplotype are expected to inherit the same QTL allele. In Sun et al. (2014), the haplotype...
that carries only one of QTL alleles is defined complete concordance, and the haplotype that carries either the major or the minor QTL allele is defined discordance. With a SNP density of about 10 SNPs/cM, genomic prediction based on haplotype models have been shown to have similar accuracy to prediction based on SNP models (Calus et al., 2008; Villumsen et al., 2009; Hickey et al., 2013), because the concordance between haplotypes with QTL alleles was low due to low SNP density. The potential advantage of the haplotype model in improving prediction accuracy has been shown by Sun et al. (2014), where complete concordance was achieved between 1-cM haplotypes and QTL alleles with a SNP density of 200 SNPs/cM. Then, the haplotype model had significantly higher prediction accuracy than the SNP model when historical LD was low.

The main objective of this research was 1) to study the ability of fitting haplotypes Sun et al. (2014) to capture CS information without using pedigree information, 2) to study the potential advantage of fitting both SNP genotypes and haplotypes in improving persistence of accuracy across validation generations without retraining (long-term accuracy), and 3) to study the effects of historical LD, current $N_e$, and MAF of QTL on contributions of LD and CS information to long-term accuracy. Contributions of LD and CS information was first investigated by analyzing simulated datasets of Sun et al. (2014b), and then, the potential advantage of combined modeling SNP genotypes and haplotypes was tested on several egg quality traits in a commercial breeding population of layer chickens.

### 5.3 Materials and Methods

#### 5.3.1 Simulated datasets

Simulated datasets from Sun et al. (2014b) were used. To study the effects of current $N_e$ on contributions of LD and CS information to long-term accuracy, Sun et al. (2014b) simulated three pedigrees with different mating designs. The simulated pedigrees included 13 non-overlapping generations. In each progeny generation with size 600, the three pedigrees differed in the number of parents per generation and the number of progenies per mating. Pedigree 1 was a balanced nested design with 5 sires per generation. Pedigree 2 was an outbred
population in which all individuals survived as parents in each generation. Pedigree 3 was an unbalanced nested design with 5 sires per generation, one of which had more than 80% of progenies that composed the next generation while the other 4 together had less than 20% of progenies that composed the next generation. Details of the simulated datasets are given by Sun et al. (2014b).

For the simulated datasets, real linkage phase in simulation of all QTL and SNP loci were used in the analyses. The phased genome was divided into $n_q$ non-overlapping windows of length 1.0 cM. The number and percentage of discordant haplotypes within every cM was calculated.

5.3.2 Dataset of layer chickens from a commercial breeding population

Details of the layer chicken dataset were given by Weng et al. (2014b) and Weng et al. (2014a). In this study, only birds with both genotypes and phenotype were used, and came from 9 generations. The average number of birds with both genotypes and phenotype in each generation is summarized in Table 5.1.

Three early egg quality traits, early egg weight (eew), early punches score (eps) and early yolk weight (eyw), and two late egg quality traits, late egg weight (lew) and late yolk weight (lyw), were analyzed. Genotypes of 23,043 SNPs were used after the quality control, as in Weng et al. (2014a). The first 5 generations with approximately 1,500 birds in total, were used for training, and each of the following 4 generations with approximately 300 birds per generation, were used for validation. Beagle (Browning and Browning, 2007) was used to phase the SNP genotypes of all birds in the training and validation generations.

5.3.3 Genomic prediction using haplotypes

The haplotype model was the same as that used by Sun et al. (2014) with details given as follows. The allelic value of each unique haplotype within a 1-cM genome window in the training population was fitted as an independent random effect. Let $n_j$ be number of unique haplotypes at the $j$th cM. At the $j$th cM, the covariate for the $k$th unique haplotype of individual $i$, $h_{ijk}$, is the number of copies of that allele carried by individual $i$, coded as 0/1/2. The haplotype
model for phenotype is given by

\[
y = \mathbf{1}\mu + \sum_{j=1}^{n_q} \mathbf{H}_j\mathbf{\Omega}_j\mathbf{\gamma}_j + \mathbf{e},
\]  

(5.1)

where \( \mathbf{y} \) is an \( n \times 1 \) vector of phenotypes of the \( n \) training individuals, \( \mathbf{\gamma}_j \) is an \( n_j \times 1 \) vector of the \( n_j \) allelic values at the \( j \)th cM, \( \mathbf{H}_j \) is an \( n \times n_j \) matrix of covariates for \( \mathbf{\gamma}_j \) at the \( j \)th cM, \( \mathbf{\Omega}_j \) is a diagonal matrix of indicator variables with the \( k \)th diagonal element \( \omega_{jk} = 1 \) if the \( k \)th allelic value \( (\mathbf{\gamma}_{jk}) \) has an effect on the trait and \( \omega_{jk} = 0 \) otherwise, and \( \mathbf{e} \) is an \( n \times 1 \) vector of residuals. The genomic estimated breeding value (GEBV) for individual \( i \) was the summation of estimated allelic values carried by that individual across all \( n_q \) cM.

The haplotype model (5.1) is different from the co-segregation (CS) model of Sun et al. (2014c). In the CS model for a pedigreed population, the allelic values of pedigree founders are assumed independent. The number of independent allelic values equals twice the number of pedigree founders within every cM of genome. In the haplotype model (5.1), the allelic values of 1-cM haplotypes with unique allele state are assumed independent. Within the \( j \)th cM, the number of unique haplotypes \( n_j \) in the training population is expected to be much smaller than the number of all possible alleles, \( 2^{s_j} \), where \( s_j \) is the number of SNPs in the \( j \)th cM. The haplotype model and the CS model are similar in that, with high SNP density, the same haplotype alleles can be thought as being inherited from the same common ancestor without recombination, and hence carry the same QTL allele without mutation at QTL.

Bayesian methods “BayesA” and “BayesB” (Meuwissen et al., 2001) were adapted for the estimation of allelic values \( \mathbf{\gamma} \) (Sun et al., 2014,c). Each allelic value was assumed to have an independent \( t \) prior distribution with degrees of freedom \( \nu_{\gamma} \) and scale parameter \( S_\gamma^2 \). The prior for \( \omega_{jk} \) is Bernoulli with probability \( \operatorname{Pr}(\omega_{jk} = 1) = 1 - \pi_{\text{Hap}} \) in BayesB, and \( \omega_{jk} = 1 \) in BayesA. \( \pi_{\text{Hap}} \) is the proportion of haplotype alleles that have zero effect on the trait.

A combined SNP-haplotype model was constructed to explicitly model both LD and CS information without using pedigree information. The SNP-haplotype model expanded the model (5.1) to include a term for modeling SNP genotypes (Sun et al., 2014c).

\[
y = \mathbf{1}\mu + \mathbf{Z}\mathbf{\alpha} + \sum_{j=1}^{n_q} \mathbf{H}_j\mathbf{\Omega}_j\mathbf{\gamma}_j + \mathbf{e},
\]  

(5.2)
where \( Z \) is an \( n \times m \) matrix with rows containing genotype covariates at \( m \) SNPs for the training individuals, and \( \alpha \) is an \( m \times 1 \) vector of allele substitution effects of the \( m \) SNPs. Details of Bayesian inference of models 5.1 and 5.2 are similar to Sun et al. (2014c).

In the analyses of the layer chicken dataset, the length of haplotypes was 1 million base pairs (Mbp) instead of 1 cM. The order of SNP was based on build 2 of the chicken genome (WASHUC2, May 2006). Either all the unique haplotypes or only haplotypes with frequency larger than 1\% (common haplotypes) were fitted in the model. In BayesB, the value of \( \pi_{\text{Hap}} \) was 0.97 when fitting all haplotypes, and 0.95 when fitting common haplotypes. The value of \( \nu_\gamma \) was 4.2 following Meuwissen et al. (2001). For the haplotype model (5.1), the value of \( S_\gamma^2 \) was chosen such that

\[
\frac{\nu_\gamma S_\gamma^2}{\nu_\gamma - 2} = \frac{h^2 V_P}{(1 - \pi_{\text{Hap}}) n_q}
\]

For the SNP-haplotype model 5.2, values of \( S_\alpha^2 \) and \( S_\gamma^2 \) were chosen such that

\[
\frac{\nu_\alpha S_\alpha^2}{\nu_\alpha - 2} = \frac{0.5 h^2 V_P}{2 (1 - \pi_{\text{SNP}}) \sum_{l=1}^m p_l (1 - p_l)}, \quad \text{and} \quad \frac{\nu_\gamma S_\gamma^2}{\nu_\gamma - 2} = \frac{0.5 h^2 V_P}{(1 - \pi_{\text{Hap}}) n_q},
\]

where \( V_P \) is the phenotypic variance, \( p_l \) is the MAF of the \( l \)th SNP, and \( h^2 \) is trait heritability that equaled 0.5 in the simulated datasets, and equaled the trait heritability estimated using pedigree relationships in the layer chicken dataset (Weng et al., 2014a). A Gibbs sampler similar to that of Sun et al. (2014c) was constructed to generate random samples from the marginal posterior distribution of each parameter. Point estimates of \( \hat{\gamma} \) and \( \hat{\alpha} \) were the posterior means calculated from the random samples of the Gibbs sampler with 21,000 iterations, in which the first 1,000 were discarded as burn-in.

## 5.4 Results

### 5.4.1 Number of haplotypes and concordance with QTL alleles

At a SNP density of 20 SNPs/cM and high historical LD among SNPs, the average number of haplotypes across 1-cM genome windows was similar for all three pedigrees, with the largest
number observed in pedigree 2 (26.8) and smallest in pedigree 3 (22.9) (Table 5.2). The average number of common haplotypes was around 4 in all three pedigrees. These results suggest that more rare haplotypes exist in populations with a large current \( N_e \) than those with a small \( N_e \), while the number of common haplotypes was similar for different \( N_e \). When there was no LD among SNPs, the number of haplotypes was 7 times larger than that with high LD among SNPs, but the number of common haplotypes was only twice of that with high LD among SNPs. The number of haplotypes doubled by increasing SNP density from 20 to 200 SNPs/cM with high LD, but the number of common haplotypes increased only from about 4 to 6, resulting in a slight decrease in the proportion of common haplotypes (Table 5.3). These results suggest that historical LD, current \( N_e \), and SNP density have large effects on the number of 1-cM unique haplotypes, but much less effects on the number of common haplotypes.

At a SNP density of 20 SNPs/cM, the average percentage of discordant haplotypes across 1-cM genome windows was less than 2% in all three pedigrees. The percentage of discordant haplotypes reduced to 0.4% by increasing SNP density from 20 to 200 SNPs/cM, suggesting that the concordance between haplotypes and QTL alleles was near complete with a SNP density of 200 SNPs/cM (Tables 5.2 and 5.3). These results are very similar to Sun et al. (2014), except that there were no discordant haplotypes in Sun et al. (2014) with 200 SNPs/cM, while a small proportion of discordant haplotypes were observed in this study. A possible reason is that Sun et al. (2014) simulated an outbred population with size 2,000, while in this study the pedigree populations had limited current \( N_e \) across 13 generations of more than 7,000 individuals, which had a larger chance to observe discordant haplotypes.

### 5.4.2 Prediction accuracy in the simulated datasets

In the Common QTL scenario with high historical LD, genomic predictions using the SNP model and the SNP-haplotype model had persistently high accuracy across the 8 validation generations without retraining (Figures 5.1 and 5.2). The haplotype model had slightly lower accuracy than the SNP and SNP-haplotype models. Fitting only common haplotypes had similar accuracy as fitting all haplotypes in both the haplotype and SNP-haplotype models (Figures 5.1 and 5.2). Accuracies of genomic predictions using BayesA and BayesB were very
similar for the SNP and SNP-haplotype models, whereas for the haplotype model BayesB had lower and less persistent accuracy than BayesA. Accuracies were very similar for pedigrees 1 and 3 and were slightly higher than for pedigree 2 (Figures 5.1 and 5.2).

In the Rare QTL scenario with high historical LD, the decrease in accuracy across validation generations was much more significant than in the Common QTL scenario (Figures 5.3 and 5.4). The haplotype model and the SNP-haplotype model had similar accuracies for all three pedigrees, which were significantly higher than accuracy of the SNP model (Figures 5.3 and 5.4). Accuracies in the first few validation generations were much higher for pedigrees 1 and 3 than for pedigree 2, but decreased to a similar accuracy for all three pedigrees in the 8th validation generation. BayesB had significantly higher accuracy than BayesA in all three pedigrees (Figures 5.3 and 5.4). Again, fitting only common haplotypes resulted in similar accuracy as fitting all haplotypes for both the haplotype and SNP-haplotype models (Figures 5.3 and 5.4).

In the Rare QTL scenario with high historical LD, increasing SNP density from 20 to 200 SNPs/cM improved accuracies for all models by a magnitude of 0.05. Accuracies of the haplotype and SNP-haplotype models using BayesB decreased from 0.9 to 0.8 across validation generations, which was less than the decrease of accuracy for BayesA (from 0.8 to 0.65) (Figures 5.9 and 5.10). Fitting only common haplotypes resulted in a slightly lower accuracy than fitting all haplotypes, especially when using BayesB (Figures 5.9 and 5.10).

In the Common QTL scenario with no historical LD, when using BayesA, prediction accuracies for pedigrees 1 and 3 dropped from above 0.80 to about 0.65 across validation generations, whereas accuracy for pedigree 2 dropped from about 0.7 to below 0.55 (Figures 5.5 and 5.6). When fitting all haplotypes, the haplotype and SNP-haplotype models had similar accuracies as the SNP model for pedigrees 1 and 3, but had slightly higher accuracy than the SNP model for pedigree 2 (Figures 5.5 and 5.6). In contrast to the scenarios with high historical LD, accuracy of the haplotype model dropped significantly when fitting only common haplotypes, and was much lower than accuracy of the SNP model (Figures 5.5 and 5.6). Accuracies of all models were lower for BayesB compared to BayesA, in particular for the haplotype model (Figures 5.5 and 5.6).
In the Rare QTL scenario with no historical LD, the ranking of accuracies of the three models was similar to that in the Common QTL scenario with no historical LD. The difference between accuracies of the SNP model and the SNP-haplotype model was much larger for the Rare than for the Common QTL scenario (Figures 5.7 and 5.8).

5.4.3 Prediction accuracy in layer chickens

For the early egg quality traits eew and eyw, which had moderately high accuracy using the SNP model, the haplotype and SNP-haplotype models had similar or slightly lower accuracies. However, for the trait eps, for which the SNP model had very low accuracy, fitting haplotypes improved accuracy by 0.05 (Figure 5.11). For the late traits lew and lyw, the SNP-haplotype model had similar or higher accuracy than the SNP model (Figure 5.12).

5.5 Discussion

In this study, the haplotype model of Sun et al. (2014) was used to model CS without using pedigree information. Sun et al. (2014) showed that when the SNP density was 200 SNPs/cM, the concordance of QTL alleles with 1-cM haplotypes was complete. With complete concordance, haplotypes can accurately capture the CS of QTL alleles among individuals that inherit the same haplotype allele from the most recent common ancestor. A model fitting both SNP genotypes and haplotypes was proposed to capture both LD and CS information. The potential advantage of modeling haplotypes to improve prediction accuracy across validation generations without retraining was tested on simulated datasets with different levels of historical LD, current \( N_e \), and SNP density. Genomic prediction across four validation generations without retraining was also performed on several egg quality traits in a commercial breeding population of layer chickens, to test the advantage of the SNP-haplotype model that was observed in simulation studies. This section gives detailed discussion on the ability of modeling haplotypes to capture CS information, as well as the effects of historical LD, current \( N_e \) and SNP density on accuracy across validation generations without retraining.
5.5.1 Genetic information exploited by the SNP, haplotype, and SNP-haplotype models

Genetic information captured by the SNP model was investigated by Sun et al. (2014b). The ability of the haplotype model to capture CS information was similar to that of the CS model that followed the transmission of QTL alleles in a pedigreed population (Sun et al., 2014b). When historical LD between QTL and SNPs was high, accuracy due to LD was persistently high across multiple validation generations without retraining, in agreement with (Habier et al., 2007, 2013; Sun et al., 2013, 2014c,b). CS information had a significant contribution to prediction accuracy only when historical LD was low, in agreement with Sun et al. (2013, 2014c,b). Accuracy of the haplotype model persisted across validation generations without retraining. Persistence of accuracy due to CS was stronger in populations with smaller current $N_e$. Increasing SNP density improved accuracy of the haplotype model by reducing the proportion of discordant haplotypes. With high SNP density, fitting haplotypes smaller than 1 cM can further improve persistence of prediction accuracy across validation generations, because smaller haplotypes can still be in complete concordance with QTL alleles, but are less likely to be broken by recombination.

In addition to CS information, the haplotype model also captures historical LD information. In the Common QTL scenario with high historical LD, accuracy of the haplotype model dropped from 0.9 to 0.8 across the 8 validation generations, whereas accuracy of the CS model of (Sun et al., 2014b) dropped from below 0.85 to 0.65. Compared to the CS model in Sun et al. (2014) that only captures CS information, the increased accuracy of the haplotype model comes from historical LD information. The reason is that the haplotype model captures CS starting from the most recent common ancestors, while the CS model for a pedigree population captures CS starting from the founders of the pedigree, which are usually included in the training population. In the CS model, the allelic value of the same haplotype allele from different founders are treated as independent, while in the haplotype model they are treated as the same. The haplotype model captures historical LD information because the same haplotype allele in different founders contributes to the estimation of the allelic value of the same QTL.
The SNP-haplotype model captures both LD and CS information by modeling them explicitly. The SNP-haplotype model had similar accuracy with the SNP model when historical LD was high, but had significantly higher accuracy than the SNP model when historical LD was low. The SNP-haplotype model can be feasibly applied to achieve high accuracy without using pedigree information, provided that the linkage phases of SNPs are available. In field datasets, phasing accuracies higher than 0.99 have been frequently reported when SNP density is high (Browning and Browning, 2011).

### 5.5.2 Prediction accuracy with historical LD

When historical LD is high, prediction accuracy of the SNP model is driven by capturing LD between QTL and SNPs. The level of LD between two loci highly depends on their MAF. In the Common QTL scenario, LD was high between QTL and SNPs, and therefore the SNP model had much higher accuracy across validation generations than the haplotype model. Although the haplotype alleles were in close to complete concordance with QTL alleles under the simulated SNP density of 20 SNPs/cM, fitting haplotypes in addition to SNP genotypes did not further improve prediction accuracy. However, in the Rare QTL scenario, LD between QTL and SNPs was low due to much different MAF of QTL and SNPs, and accuracy of the SNP model mainly came from capturing CS information over long chromosome regions. This information was broken down by recombination, and did not contribute much to accuracy across validation generations. Modeling haplotypes increased prediction accuracy compared to the SNP model due to explicitly following co-segregation of haplotypes with QTL alleles.

In real data analyses, high prediction accuracy of SNP models was often observed for individuals that were closely related to the training population, and prediction accuracy was low for individuals distantly related to the training population (Habier et al., 2010; Albrecht et al., 2011; Saatchi et al., 2011; Clark et al., 2012; Windhausen et al., 2012; Wientjes et al., 2013; de los Campos et al., 2013b; Crossa et al., 2013). Moreover, in dairy cattle, the improvement of prediction accuracy was limited when increasing SNP density from the 50K chip to the 770K chip or sequence (Erbe et al., 2012; Su et al., 2012; Gunia et al., 2014; Hayes et al., 2014), which was shown to achieve consistent LD among distantly related subpopulations by
reducing the average distance between pairwise SNPs (de Roos et al., 2008; Qanbari et al., 2010). These results suggest that, in field datasets, LD between QTL and SNPs is low and is not greatly increased by the increased SNP density, and therefore prediction accuracy mainly comes from CS information that is implicitly captured by SNP genotypes. The main reason for low LD between QTL and SNPs is because most QTL tend to have low MAF (Yang et al., 2010; Daetwyler et al., 2014; MacLeod et al., 2014), while SNPs chips are developed by including SNPs with high MAF in prototype genotyping (Matukumalli et al., 2009). Prediction accuracy for traits with many rare QTL by fitting common SNPs remains low regardless of the SNP density (de los Campos et al., 2013b; Druet et al., 2014; Sun et al., 2014), but can be improved by either including rare SNPs in the panel (Druet et al., 2014; Sun et al., 2014) or fitting haplotypes (Goddard and Hayes, 2007; Sun et al., 2014). The SNP-haplotype model is expected to improve prediction accuracy because it captures the effects of common QTL by modeling LD, and also the effects of rare QTL by modeling CS through complete concordance between haplotypes and QTL alleles.

5.5.3 Prediction accuracy without historical LD

The simulated datasets with no historical LD were contrived since many studies showed that LD between closely linked SNPs was high in livestock populations (de Roos et al., 2008; Qanbari et al., 2010). The simulated datasets without historical LD were designed to investigate the ability of the SNP model and the haplotype model to capture CS information. CS information captured by the SNP model is the association between alleles states of linked QTL and SNPs, which depends on the level of random genetic drift. In pedigrees 1 and 3 with strong genetic drift, the number of unique haplotypes was limited. The SNP model could capture as much CS information as the haplotype model by fitting a large number of SNP genotypes. However, the number of unique haplotypes was large in pedigree 2 due to the near absence of drift. The SNP model had much lower accuracy than the haplotype model for pedigree 2 because CS information of different haplotypes was confounded across multiple SNPs.

The ability of the SNP model to capture CS information is also affected by MAF of QTL. With strong genetic drift in pedigrees 1 and 3, the effect of MAF of QTL on accuracy was small
because rare QTL alleles in the base population could become common alleles within family, whereas in pedigree 2, MAF did not change much due to the near absence of drift. As a result, in the Rare QTL scenarios, the SNP model had much lower accuracy in pedigree 2 than in pedigrees 1 and 3.

5.5.4 The effect of fitting only common haplotypes

In a population that is in equilibrium between mutation, drift and recombination, the distribution of frequencies of multiple alleles is skewed, with only a few common alleles and most alleles being rare (Ewens, 1972). This was observed in the simulated datasets with high historical LD (Tables 5.2 and 5.3). Since the allelic values of rare haplotypes (frequency less than 1%) are difficult to estimate accurately due to lack of sufficient data, excluding them from the haplotype model has limited effect on prediction accuracy, but improves computation substantially (Sun et al., 2014). In contrast, when there is no historical LD, the unique haplotypes in the training and validation population are random samples from all possible $2^{n_j}$ haplotypes in the $j$th cM of genome, where $n_j$ is the number of SNPs within the $j$th cM. The allele frequencies of haplotypes tend to be uniformly distributed, where each allele has a low probability to be sampled close to $\frac{1}{2^{n_j}}$.

5.5.5 The effect of SNP density

Increasing SNP density significantly improves concordance between 1-cM haplotypes and QTL alleles (Tables 5.2 and 5.3). Prediction accuracy of the haplotype model increased with a higher SNP density due to capturing CS information more accurately. In the Rare QTL scenario, accuracy of the SNP model also increased with SNP density. The reason is that the ability of the SNP model to capture CS information got improved with more SNPs fitted within each cM of genome (Habier et al., 2013), but LD between QTL and SNPs remained low due to their different MAF. With a higher SNP density, haplotypes of a smaller size can still be in complete concordance, but the proportion of recombinant haplotypes decreases across validation generations. The haplotype model can therefore achieve higher accuracy by fitting haplotypes of less than 1cM in length.
5.5.6 Results from the layer chicken dataset

For traits with moderate to high accuracy of the SNP model, the SNP-haplotype model had either similar or higher accuracy than the SNP model. A possible explanation is that QTL of traits are in high LD with SNPs (Wolc et al., 2014) and, thus, fitting haplotypes does not have extra advantage. However, for the trait eps, where the SNP model had accuracy close to zero, fitting haplotypes alone or in addition to SNP genotypes improved accuracy significantly. A reasonable explanation may be that the QTL for the trait eps undergo stronger directional selection and hence have much lower MAF than the QTL for traits related with egg weight (Wolc et al., 2014). Results from the layer chicken dataset are in agreement with results from simulated datasets, where the SNP-haplotype model can have high accuracy across different traits, and can potentially improve accuracy for traits for which the SNP model has low accuracy.

The available SNP density for the layer chicken dataset is much lower than that in the simulation. The advantage of the SNP-haplotype model is expected to be more obvious under a SNP density of around 200 SNPs/cM. Reasons for the similar accuracy of the SNP-haplotype model to that of the SNP model include 1) incomplete concordance between haplotypes and QTL alleles due to limited SNP density, 2) low phasing accuracy with limited SNP density (Browning and Browning, 2007), and 3) errors in the build 2 of the chicken genome (Wang et al., 2013).

5.6 Conclusions

In this study, modeling haplotypes was firstly proposed to explicitly capture CS without pedigree information. With a SNP density of 200 SNPs/cM, 1-cM haplotypes were in near complete concordance with QTL alleles, and therefore could accurately capture CS of QTL alleles from most recent common ancestors. The SNP-haplotype model had similar accuracy as the SNP model when historical LD was high between QTL and SNPs, but had significantly higher accuracy than the SNP model when historical LD was low. When applied to several egg quality traits, the SNP-haplotype model had similar accuracy as the SNP model for traits
for which accuracy of the SNP model was moderate or high, but had significantly higher accuracy than the SNP model for traits for which accuracy of the SNP model was almost zero. The SNP-haplotype model is recommended for traits for which accuracy is limited when fitting SNP genotypes, and accuracy does not improve by increasing SNP density. Prediction accuracy of the haplotype model is expected to improve by increasing SNP density, because complete concordance can be achieved among haplotypes with much smaller sizes, for which the proportion of recombinant haplotypes will be much smaller across generations.

5.7 Acknowledgments

Instructions from Drs. Jack Dekkers, Rohan Fernando and Dorian Garrick were greatly acknowledged. This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative, National Institute of Food and Agriculture Competitive Grant 2010-65205-20341 and by National Institutes of Health Grant R01GM099992.

5.8 Bibliography


5.9 Tables

Table 5.1  Average number of genotyped individuals with own phenotypic records in 9 generations of the layer chicken dataset.

<table>
<thead>
<tr>
<th>Generation</th>
<th>Early traits</th>
<th>Late traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>295</td>
<td>295</td>
</tr>
<tr>
<td>2</td>
<td>322</td>
<td>323</td>
</tr>
<tr>
<td>3</td>
<td>295</td>
<td>295</td>
</tr>
<tr>
<td>4</td>
<td>360</td>
<td>357</td>
</tr>
<tr>
<td>5</td>
<td>287</td>
<td>278</td>
</tr>
<tr>
<td>6</td>
<td>260</td>
<td>268</td>
</tr>
<tr>
<td>7</td>
<td>300</td>
<td>291</td>
</tr>
<tr>
<td>8</td>
<td>240</td>
<td>277</td>
</tr>
<tr>
<td>9</td>
<td>300</td>
<td>290</td>
</tr>
</tbody>
</table>

Table 5.2  Average number of haplotype alleles and concordance of haplotypes with QTL alleles across 1-cM genome windows in three simulated mating designs. SNP density is 20 SNPs/cM.

<table>
<thead>
<tr>
<th>Mating design</th>
<th>Pedigree 1</th>
<th>Pedigree 2</th>
<th>Pedigree 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Allele(^1)</td>
<td>23.7</td>
<td>26.8</td>
<td>22.9</td>
</tr>
<tr>
<td>No. Common(^2)</td>
<td>4.1</td>
<td>4.8</td>
<td>3.7</td>
</tr>
<tr>
<td>Common(^3)%</td>
<td>17.5%</td>
<td>17.9%</td>
<td>16.1%</td>
</tr>
<tr>
<td>No. Discordant(^4)</td>
<td>0.38</td>
<td>0.46</td>
<td>0.41</td>
</tr>
<tr>
<td>Discordant(^5)%</td>
<td>1.6%</td>
<td>1.7%</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

\(^1\) Average number of unique haplotypes across 1-cM genome windows.
\(^2\) Average number of haplotypes with frequency larger than 1% (common haplotypes) across 1-cM genome windows.
\(^3\) Average percentage of common haplotypes across 1-cM genome windows.
\(^4\) Average number of haplotypes that carry either the major or the minor QTL allele (discordant haplotypes) across 1-cM genome windows.
\(^5\) Average percentage of discordant haplotypes across 1-cM genome windows.
Table 5.3  **Average number of haplotype alleles and concordance of haplotypes with QTL alleles across 1-cM genome windows under different SNP densities and historical LD among SNPs in Pedigree 3.**

<table>
<thead>
<tr>
<th>SNP density</th>
<th>20 SNPs/cM</th>
<th>200 SNPs/cM</th>
<th>20 SNPs/cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical LD</td>
<td>&gt; 0</td>
<td>&gt; 0</td>
<td>= 0</td>
</tr>
<tr>
<td>No. Allele$^2$</td>
<td>22.9</td>
<td>45.4</td>
<td>168.6</td>
</tr>
<tr>
<td>No. Common$^3$</td>
<td>3.7</td>
<td>5.9</td>
<td>8.3</td>
</tr>
<tr>
<td>Common%$^4$</td>
<td>16.1%</td>
<td>13.0%</td>
<td>4.9%</td>
</tr>
<tr>
<td>No. Discordant$^5$</td>
<td>0.41</td>
<td>0.18</td>
<td>1.7</td>
</tr>
<tr>
<td>Discordant%$^6$</td>
<td>1.8%</td>
<td>0.4%</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

1 Average number of unique haplotypes across 1-cM genome windows.
2 Average number of haplotypes with frequency larger than 1% (common haplotypes) across 1-cM genome windows.
3 Average percentage of common haplotypes across 1-cM genome windows.
4 Average number of haplotypes that carry either the major or the minor QTL allele (discordant haplotypes) across 1-cM genome windows.
5 Average percentage of discordant haplotypes across 1-cM genome windows.
Figure 5.1  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with high historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.2  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with high historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.3  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with high historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.4  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with high historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.5  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with no historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.6 Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Common QTL scenario with no historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.7  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with no historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.8  Mean accuracy across 8 validation generations without retraining in three simulated pedigrees in the Rare QTL scenario with no historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes. The simulated SNP density is 20 SNPs/cM.
Figure 5.9  Mean accuracy across 8 validation generations without retraining in pedigree 3 in the Rare QTL scenario with high historical LD using BayesA, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes.
Mean accuracy across 8 validation generations without retraining in pedigree 3 in the Rare QTL scenario with high historical LD using BayesB, fitting all (top panel) or common (bottom panel) haplotypes. SNP, the SNP model; Hap, the haplotype model fitting all haplotypes; SNP-Hap, the SNP-haplotype model fitting all haplotypes; HapC, the haplotype model fitting only common haplotypes; SNP-HapC, the SNP-haplotype model fitting only common haplotypes.
Figure 5.11  Prediction accuracy in 4 validation generations for three early egg quality traits in layer chicken datasets using BayesA (left) or BayesB (right) by fitting all haplotypes. eew, early egg weight; eps, early punches score; eyw, early yolk weight. 1.0Mbp, models fitting 1.0Mbp haplotypes; All, models fitting all haplotypes; SNP, the SNP model; Hap, the haplotype model; SNP-Hap, the SNP-haplotype model.
Figure 5.12  Prediction accuracy in 4 validation generations for two late egg quality traits in layer chicken datasets using BayesA (left) or BayesB (right) by fitting all haplotypes. lew, late egg weight; lyw, late yolk weight. 1.0Mbp, models fitting 1.0Mbp haplotypes; All, models fitting all haplotypes; SNP, the SNP model; Hap, the haplotype model; SNP-Hap, the SNP-haplotype model.
CHAPTER 6. GENERAL DISCUSSIONS AND CONCLUSIONS

6.1 General Discussions

6.1.1 Research objectives

The objectives of this thesis were 1) to develop statistical methods that model CS information explicitly, 2) to study contributions of LD and CS information to prediction accuracy using simulated datasets that differ in mating designs and historical LD in pedigree founders, and 3) to study the ability to improve prediction accuracy by explicitly modeling CS in addition to LD information. A CS model that uses pedigree information was developed in Chapter 2. In the CS model, the transmission of 1-cM haplotype alleles at putative QTL from pedigree founders to offspring is followed using parental allele origins of SNPs. In Chapter 4, a haplotype model was developed to improve prediction accuracy for traits for which QTL had much lower minor allele frequency (MAF) than SNPs (Sun et al., 2014). The haplotype model was used to model CS information without using pedigree information in Chapter 5. Instead of fitting haplotype alleles of pedigree founders in the CS model, the haplotype model fits 1-cM unique haplotype alleles that are segregating in the training population. With high SNP density, haplotypes with the same allele state have high probability to be in identity-by-descent (Rosenberg and Nordborg, 2002), and therefore can explicitly capture CS of QTL alleles with SNPs from most recent common ancestors.

6.1.2 Contributions of LD and CS information to prediction accuracy

In this thesis, the definitions of LD and CS that contribute to prediction accuracy followed Habier et al. (2010) and Habier et al. (2013). LD information was defined as the association between allele states of QTL and SNPs in pedigree founders. CS information was defined as
the association between grand-parental allele origins of linked QTL and SNPs, i.e. alleles of QTL and SNPs originating from the same chromosome of a pedigree founder. The level of LD in founders is determined by mutation, recombination and random genetic drift in historical generations before pedigree founders, as well as by genomic distances and MAF of QTL and SNPs. In general, LD between QTL and SNPs with smaller distances is higher when MAF of QTL and SNPs are similar, whereas LD is low between QTL and SNPs with much different MAF, regardless of the genomic distance between them. As a result, increasing SNP density only improves LD between QTL and SNPs with similar MAF. The amount of CS information in a pedigree is affected by the average relationship between individuals due to mating designs, which can be represented by current effective population sizes ($N_e$). The allelic values of founder haplotypes can be estimated more accurately in populations with smaller current $N_e$, because more phenotypic data are available for each allelic value. In summary, LD is expected to have a large contribution to accuracy when historical LD between QTL and SNPs is high, which requires high SNP density and similar MAF between QTL and SNPs; CS is expected to have a greater contribution to accuracy in populations with smaller current $N_e$. Simulation scenarios were designed to study the effects of current $N_e$, MAF of QTL and SNPs, and SNP density on contributions of LD and CS information to prediction accuracy.

Results from simulation studies showed that CS information had little contribution to accuracy when LD between QTL and SNPs was high, but had a substantial contribution when LD was low. When LD between QTL and SNPs was low, accuracy of the LD model mostly came from CS information implicitly captured by SNP genotypes, which decreased rapidly when training population included more independent families, or when validation population was separated by more generations from the training population. A possible explanation is that, when LD between QTL and SNPs is low, the necessary phenotypic data for estimation of QTL effects only come from relatives that inherit the same haplotype allele co-segregating with QTL. The CS and haplotype models can correctly distinguish unrelated families among training individuals. However, phenotypes of members of all families in training are used for estimation of QTL effects in the LD model. Phenotypes of members of unrelated families provide no information but add errors to estimation of QTL effects that segregate within families.
6.1.3 CS information captured by the CS and haplotype models

The CS model and the haplotype model are similar in that they both model CS of QTL alleles with 1-cM haplotypes, but CS information captured by these two models are substantially different. The CS model uses parental allele origins at SNPs to follow the transmission of QTL alleles in a pedigree. CS information captured by the CS model starts from pedigree founders, where all haplotype allelic values at putative QTL are assumed independent. In every cM of genome, the number of haplotypes fitted by the CS model equals twice the number of founders. Accuracy of the CS model is independent from LD between QTL and SNPs because no information from allele states is used in the CS model. In contrast, the haplotype model uses haplotype allele states to capture CS of QTL alleles among individuals that share the same unique haplotypes. With high SNP density, haplotypes with the same allele state have high probability to be identical-by-descent to the most recent common ancestor (Rosenberg and Nordborg, 2002). Fitting haplotypes in recent generations also captures historical LD information, because the same haplotype allele shared by different pedigree founders has the same allelic value in the haplotype model, and therefore generates genetic covariance among pedigree founders. The difference between the CS and haplotype models is significant when the current \( N_e \) is large, because the number of unique haplotypes fitted by the haplotype model can be much smaller than the number of founder haplotypes fitted in the CS model. However, this difference is small when current \( N_e \) is small, because many founder haplotypes will be lost by drift, and haplotypes with the same allele state are likely to have originated from the same founder.

6.1.4 Modeling LD and CS explicitly for field datasets

In the simulated datasets, modeling both LD and CS explicitly in the combined LD-CS or SNP-haplotype model had higher accuracy than the LD model only when LD between QTL and SNPs was low, but had similar accuracy as the LD model when LD was high. The effects of QTL with low MAF (rare QTL) could not be accurately estimated by fitting SNPs with large MAF (common SNPs) because of low LD of SNPs with such QTL, but could be estimated
accurately when QTL effects were captured by modeling CS information. In the analyses of egg quality traits of a commercial breeding population, the SNP-haplotype model did not have higher accuracy than the LD model for most traits, but had significantly higher accuracy than the LD model for a few traits where accuracy of the LD model was almost zero. QTL for traits with low accuracy when using the SNP model could have much lower MAF than SNPs, and therefore LD between QTL and SNPs was low. The lack of superiority of the SNP-haplotype model for most traits was probably due to low SNP density in the chicken dataset, which was much lower than that in simulation studies where the SNP-haplotype model had an advantage. When SNP density was low, the concordance between haplotypes and QTL alleles was low, and consequently fitting haplotypes could not accurately capture CS information.

In most livestock populations, the LD-CS and SNP-haplotype models are expected to have higher accuracy than the LD model for the following reasons. First, the current $N_e$ of most livestock populations is small due to intensive selection and artificial insemination (Hayes et al., 2003; de Roos et al., 2008; Qanbari et al., 2010), where CS information can have significant contribution to accuracy. Second, LD between QTL and SNPs is incomplete in most livestock populations, as shown by field data results that 1) accuracy for validation populations that are distantly related to the training population was low or zero (Daetwyler et al., 2012; Kachman et al., 2013), and 2) accuracy has not been improved by increasing SNP density (Erbe et al., 2012; Su et al., 2012; Hayes et al., 2014). Third, results from field datasets showed that prediction accuracy was mainly contributed by CS instead of LD information (Daetwyler et al., 2012; Luan et al., 2012). Therefore, further analyses of field datasets with high SNP density are necessary to investigate the potential advantage of the LD-CS and SNP-haplotype models.

Limitations in the application of the CS model in field datasets include difficulties to obtain allele origin at SNPs and computational challenges due to large numbers of pedigree founders. Possible solutions to these difficulties were given in Chapter 2. Application of the haplotype model in field datasets is straightforward with the availability of linkage phases at all SNPs. Phasing accuracies greater than 0.99 have been frequently reported with dense SNP genotypes (Browning and Browning, 2011), which facilitates the application of the haplotype model in field datasets.
6.1.5 Future work

In livestock populations, modeling CS information explicitly using allele origins of SNPs or SNP haplotypes is expected to benefit from increasing SNP densities. Higher accuracy in haplotype phasing and inference of allele origins can be achieved by increasing SNP densities. Prediction accuracy of the haplotype model is expected to increase with a higher SNP density because 1) smaller haplotypes can still be in complete concordance with QTL alleles but are less likely to be broken by recombination, and 2) the number of unique haplotypes within one genome region will decrease with size. Further, increasing SNP density allows more accurate identification of recombination sites on chromosomes. In both the CS and haplotype models, the number of recombinant haplotypes can be minimized by fitting haplotypes with dynamic sizes that avoid coverage of recombination sites, instead of arbitrarily fitting 1-cM haplotypes. The optimal haplotype sizes and different approaches to construct haplotypes requires further studies.

In the near future, whole genome sequences will be available for a large number of individuals (Hayes et al., 2014; Daetwyler et al., 2014). Accuracy of genomic prediction is expected to be greatly improved with the possibility in directly fitting genotypes at all QTL for desired traits, without relying on LD between QTL and SNPs (Hayes et al., 2014). However, accurate estimation of SNP effects will remain difficult using prediction models that fit millions of SNPs, unless a sufficiently large training population can be assembled (Goddard and Hayes, 2009). To make things worse, the necessary amount of phenotypic data to achieve the same accuracy of estimated QTL effects is much larger for rare than common QTL (Wray et al., 2013; Yang et al., 2014; Lee et al., 2014). As a result, genomic prediction using SNPs from whole genome sequencing has been reported to have similar accuracy as using SNP chips in cattle (Hayes et al., 2014) and maize (Crossa et al., 2013). Accuracy of the haplotype model is expected to improve with an increase in SNP density up to sequence data, because 1) complete concordance can be achieved among haplotypes with much smaller size than 1 cM, which can be inherited across multiple generations without recombination, 2) with increasing SNP density, the number of unique haplotypes across genome will plateau and become much less than the
number of SNPs to be fitted in the prediction model, and therefore haplotype allelic values can be estimated more accurately than SNP effects given the available training data, and 3) haplotypes can capture effects of both common and rare QTL, whereas fitting sequence SNPs after quality-control of MAF only captures effects of common QTL. The haplotype model can be further extended to a QTL model, where QTL effects are directly fitted in the model, with QTL genotypes inferred from haplotypes due to complete concordance, similar to the model of Habier et al. (2010). Future studies can be focused on improving prediction accuracy by fitting haplotypes, and to develop haplotype-based QTL models by inferring QTL genotypes.

6.2 Conclusions

In conclusion, LD and CS are two important sources of genetic information that contribute to accuracy of genomic prediction. Genomic prediction models that fit SNP genotypes capture both LD and CS information. When most QTL have much lower MAF than SNPs, LD between QTL and SNPs is low, and accuracy from the SNP model is mainly contributed by CS information that is implicitly captured by SNP genotypes. This accuracy decreases when the training data size is increased by including a larger number of independent families, and deteriorates across validation generations without retraining, because CS information captured by SNP genotypes over long chromosome distances erodes rapidly by recombination. CS information can be explicitly captured by modeling transmission of putative QTL alleles within short chromosome regions (e.g. 1.0 cM) using allele origins at SNPs. Modeling CS explicitly has limited contribution to accuracy when LD between QTL and SNPs is high, but has substantial contribution to accuracy when LD between QTL and SNPs is low. CS information has greater contribution to accuracy in populations with larger current $N_e$, because for a given population size, fewer haplotypes segregate when the current $N_e$ is small, and the effect of each haplotype can be estimated more accurately because more phenotypic data is available for each haplotype. Therefore, modeling CS explicitly is expected to increase prediction accuracy across validation generations in mating designs that create small current $N_e$. For populations without pedigree information, CS information can be modeled explicitly by fitting SNP haplotypes within short chromosome regions. Fitting haplotypes captures as much CS information as modeling
CS by following the transmission of QTL alleles of pedigree founders, but also captures CS information from most recent common ancestors. Fitting both haplotypes and SNP genotypes increased accuracy for several egg quality traits for which the SNP model had low accuracy, but potential advantage of the SNP-haplotype model in improving prediction accuracy for livestock populations requires further study. The haplotype model can be a computationally efficient alternative for the SNP model with sequence data, because the number of unique haplotypes across genome will be much less than the number of sequence SNPs, and therefore haplotype allelic values can be estimated more accurately than SNP effects.

6.3 Bibliography


APPENDIX A. A FAST EM ALGORITHM FOR BAYESA-LIKE PREDICTION OF GENOMIC BREEDING VALUES


Xiaochen Sun¹, Long Qu¹², Dorian J. Garrick¹, Jack C. M. Dekkers¹, Rohan L. Fernando¹,*

¹ Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, Iowa 50011, U. S. A.

² Department of Statistics, Iowa State University, Ames, Iowa 50011, U. S. A.

* E-mail: Corresponding rohan@iastate.edu

Abstract

Prediction accuracies of estimated breeding values for economically important traits are expected to benefit from genomic information. Single nucleotide polymorphism (SNP) panels used in genomic prediction are increasing in density, but the Markov Chain Monte Carlo (MCMC) estimation of SNP effects can be quite time consuming or slow to converge when a large number of SNPs are fitted simultaneously in a linear mixed model. Here we present an EM algorithm (termed “fastBayesA”) without MCMC. This fastBayesA approach treats the variances of SNP effects as missing data and uses a joint posterior mode of effects compared to the commonly used BayesA which bases predictions on posterior means of effects. In each EM iteration, SNP effects are predicted as a linear combination of best linear unbiased predictions of breeding values from a mixed linear animal model that incorporates a weighted marker-based realized relationship matrix. Method fastBayesA converges after a few iterations
to a joint posterior mode of SNP effects under the BayesA model. When applied to simulated quantitative traits with a range of genetic architectures, fastBayesA is shown to predict GEBV as accurately as BayesA but with less computing effort per SNP than BayesA. Method fastBayesA can be used as a computationally efficient substitute for BayesA, especially when an increasing number of markers bring unreasonable computational burden or slow convergence to MCMC approaches.

**Introduction**

Genomic prediction of breeding values for economically important traits of farm animals based on high-density genome-wide SNP genotypes is typically performed in two steps (Meuwissen et al., 2001). First, allele substitution effects of SNPs are estimated from a reference population with both trait phenotypes and SNP genotypes (training); then, the genomic estimated breeding values (GEBV) for selection candidates, often the genotyped progeny of the training population, are obtained by summing the estimated SNP effects across the genome (Meuwissen et al., 2001; Calus, 2010). In this second step, which in a research context we refer to as validation, the prediction accuracy of GEBV can be assessed by the correlation of GEBV with either true breeding values (TBV) or phenotypes. Comparative studies on both simulated and field data have shown that GEBV tend to have higher accuracy than breeding values estimated using pedigree relationships (Habier et al., 2007; Calus, 2010), depending on the genetic architecture of the trait (Daetwyler et al., 2010), the nature of the SNP panel (Meuwissen et al., 2001; Solberg et al., 2008; Meuwissen, 2009), the size of the training data (Meuwissen, 2009; VanRaden et al., 2009; Hayes et al., 2009b), the population structure (Hayes et al., 2009a) and the relationship between training and validation individuals (Habier et al., 2007, 2010).

Currently, two classes of methods are used to overcome the over-parameterization problem of linear models used for genomic prediction when relating a lesser number of phenotypes to a larger number of SNP genotypes. The first is best linear unbiased prediction of SNP effects from a linear mixed model in which random SNP effects are assumed to be independently and identically distributed as zero-mean normal random variables with a common effect variance (ridge regression) (Meuwissen et al., 2001; Habier et al., 2007). This corresponds to an assumed
genetic architecture characterized by a large number of loci contributing equally to the overall genetic variance of the trait. The model for ridge regression is equivalent to an animal model in which a marker-derived realized relationship matrix is used as the variance-covariance structure of random genomic breeding values (GBLUP) (Fernando, 1998; Habier et al., 2007; VanRaden et al., 2009). Equation (3) of Habier et al. (2007) showed that the expected covariance between marker genotypes of two individuals is proportional to the additive relationship coefficient among them. Assuming variance components known, solving for SNP effects as linear combination of best linear unbiased predicted breeding values from GBLUP can be efficient because the dimension of mixed model equations for GBLUP is the number of individuals, which is usually much smaller than the number of SNPs (Strandén and Garrick, 2009). The second class of methods for genomic prediction do not necessarily result in prediction rules that are linear in the observed phenotypes. These methods are often based on Bayesian hierarchical models and are implemented through Markov chain Monte Carlo (MCMC) sampling, for instance, BayesA (Meuwissen et al., 2001), BayesB (Meuwissen et al., 2001), Bayesian LASSO (Park and Casella, 2008; de los Campos et al., 2009), BayesC_π (Habier et al., 2011), etc. Prior distributions for SNP effects are chosen to shrink ignorable small effects towards zero. Sampled SNP effects are averaged over MCMC iterations to obtain posterior means of SNP effects. Depending on the choice of priors, most Bayesian hierarchical methods impose stronger shrinkage towards zero on small SNP effects and less shrinkage on relatively large effects by allowing each SNP to have a distinct effect variance (e.g. BayesA) and/or by fitting a mixture distribution that assumes any SNP might come from a continuous distribution or a distribution degenerate at zero (e.g. BayesB). The mixture fraction is influenced through a hyperparameter \( \pi \), which specifies the prior proportion of SNPs that have zero effects. At the cost of higher computing effort, Bayesian methods tend to achieve higher prediction accuracy than GBLUP for simulated datasets (Meuwissen et al., 2001; Calus, 2010; Habier et al., 2007; Sun et al., 2011). Further, results from real data often show that methods that fit all SNPs in the model (GBLUP and BayesA) tend to give similar accuracy as methods with variable selection, suggesting that most economically important traits might be controlled by a large number of loci with relatively small effects (Hayes et al., 2009b; Luan et al., 2009; Habier et al., 2010; Wolc et al., 2011).
Several non-MCMC algorithms have been proposed to improve computational efficiency for linear models with differential shrinkage of SNP effects and/or with variable selection. VanRaden et al. (2009) presented two non-linear predictions A and B that are analogous to BayesA and BayesB in Meuwissen et al. (2001), respectively. The ratio of residual variance over common effect variance in ridge regression, which controls the amount of shrinkage of SNP effects, is modified depending on the size of estimated SNP effects to allow differential shrinkage. Estimates of SNP effects are calculated efficiently using Jacobi iteration. Both simulation (VanRaden et al., 2009) and real data (VanRaden et al., 2011) showed that VanRaden et al. (2009) non-linear predictions were fast and accurate for large datasets. Moreover, Expectation-Maximization (EM) algorithms (Dempster et al., 1977) can in some cases be computationally more efficient than MCMC approaches. Bayesian LASSO, which uses a double exponential (DE) prior distribution for SNP effects, and BayesA, which assumes $t$ prior distribution for SNP effects, have been adapted to fast non-MCMC deterministic or EM algorithms. Meuwissen et al. (2009) presented a fast heuristic iterative conditional expectation (ICE) algorithm, where the posterior expectation of SNP effects was calculated analytically, assuming a fixed known DE parameter and dispersion parameters. Shepherd et al. (2010) formulated an EM algorithm which they called emBayesB, based on the same model as ICE, which used an indicator variable for each SNP that is in linkage disequilibrium (LD) with QTL as missing data, and estimated SNP effects and the DE parameter in the M-step. Yi and Banerjee (2009) derived an EM algorithm for a BayesA model for QTL detection by treating the unknown SNP effect variances as missing data. Hayashi and Iwata (2010) developed a generalized EM algorithm (EM-BSR) with a slightly different M-step and further extended it to a heuristic algorithm for the BayesB model. BayesA modeling of SNP effects can be more appealing than LASSO, in that the estimated SNP effects are nearly unbiased for large effects, while in LASSO the bias does not diminish even when SNP effects are large (Fan and Li, 2001).

In this study we formulate a principled EM algorithm (termed “fastBayesA”) that converges to a joint posterior mode of SNP effects under the BayesA model. By applying the method to simulated datasets with contrasting sizes and genetic architectures, fastBayesA is shown to predict GEBV as accurately as BayesA but with less computing effort per SNP than BayesA.
The latter will become more important as SNP densities increase to that provided by individual DNA sequence.

Materials and Methods

Statistical Model

The linear mixed model for phenotypes based on GBLUP is

$$y = X\beta + Z\gamma + e,$$

where \( y \) is an \( n \times 1 \) vector of phenotypes, with \( n \) equal to the number of individuals in the training dataset; \( \beta \) is a vector of fixed effect parameters and \( X \) is a known design matrix relating fixed effects to phenotypes; \( Z \) is an \( n \times m \) matrix of SNP genotypes in the “0/1/2” allele dosage coding, with row \( i \) containing genotypes of \( m \) SNPs for individual \( i \); \( \gamma \) is an \( m \times 1 \) zero-mean random vector of allele substitution effects with \( \text{Var}(\gamma|\sigma^2) = \text{diag}\{\sigma_j^2\}_{j=1}^m \), where \( \sigma^2 \) is an \( m \times 1 \) vector with the \( j \)th element \( \sigma_j^2 \) being the effect variance of SNP \( j \); and \( e \) is an \( n \times 1 \) vector of independently and normally distributed random errors with mean 0 and variance \( \sigma_e^2 \).

In Meuwissen et al. (2001), GBLUP assumes that effect variances \( \sigma_j^2 \) are known and the same for all SNPs and that the SNP effects are marginally normally distributed, whereas BayesA assumes a scaled inverse Chi-square prior distribution for effect variances with scale parameter \( S^2_\gamma \) and degrees of freedom \( \nu_\gamma \), and a normal distribution for the effect of SNP \( j \) conditional on its variance, i.e.,

$$\gamma_j|\sigma^2 \sim \text{independent } N(0, \sigma_j^2),$$

where \( \gamma_j \) is the \( j \)th element of \( \gamma \), and

$$\sigma_j^2 \sim \text{i.i.d. } \frac{\nu_\gamma S^2_\gamma}{\chi^2_{\nu_\gamma}}$$

for all \( j = 1, 2, \cdots, m \). It can be shown that in BayesA the marginal distribution of the SNP effect is scaled univariate-t with degrees of freedom \( \nu_\gamma \) and scale parameter \( S^2_\gamma \) (Gianola et al., 2009).
Efficient solving of SNP effects using an equivalent animal model

The calculation strategy to develop fastBayesA follows Strandén and Garrick (Strandén and Garrick, 2009) and is generalized here. The phenotype can be modeled by the following animal model (Henderson, 1984):

\[ y = X\beta + g + e, \]

where \( y, X, \beta \) and \( e \) are as previously defined, \( g \) is an \( n \times 1 \) vector of genomic breeding values of the individuals, which can be modeled as the sum of the \( m \) SNP effects, as described above, i.e., \( g = Z\gamma \). This genomic animal model is equivalent to the GBLUP model given normality of SNP effects. The (co)variance matrix of genomic breeding values is

\[ \text{Var}(g|\sigma^2) = \text{Var}(Z\gamma|\sigma^2) = ZDZ' = G\sigma^2_g, \]

where \( D = \text{Var}(\gamma|\sigma^2) \), \( G \) is the realized relationship matrix derived from the SNP genotypes and \( \sigma^2_g \) is the variance of genomic breeding values. Element \( G_{vw} \) of \( G \) is the proportion of SNPs that are IBD between individuals \( v \) and \( w \) (Hayes et al., 2009c; VanRaden, 2008). For GBLUP, the common effect variance of SNPs is equal to \( \sigma^2_g = \sum_{j=1}^{m} \frac{\sigma^2_j g_j^2}{p_j(1-p_j)} \) in which \( p_j \) is the minor allele frequency of SNP \( j \) (Habier et al., 2007). Given \( D \), the BLUP \( \hat{\gamma} \) of SNP effects \( \gamma \) can be efficiently computed in two steps using the animal model (Strandén and Garrick, 2009). First the BLUP of genomic breeding values \( \hat{g} \) is obtained by solving the mixed model equations of the animal model, then \( \hat{\gamma} \) can be solved following Strandén and Garrick (2009) as:

\[ \hat{\gamma} = DZ'G^{-1}\hat{g}. \]

EM algorithm for estimating SNP effects

We use the above relationships to develop an EM algorithm for BayesA by treating the effect variance of each SNP as missing data. In the E-step, the conditional expectation of the logarithm of the joint probability of \( y, \gamma \) and \( \sigma^2 \), with expectation taken over the distribution of \( \sigma^2 \) conditional on the observed phenotypic data \( y \) and the current estimate (the \( k \)th step)
\( \hat{\gamma}^{(k)} \) of SNP effects, is calculated:

\[
\mathbb{E}_{\sigma^2|y, \gamma = \hat{\gamma}^{(k)}} \left[ \log \left\{ p(y, \gamma, \sigma^2) \right\} \right] = \mathbb{E}_{\sigma^2|y, \gamma = \hat{\gamma}^{(k)}} \left[ \log \left\{ p(y|\gamma)p(\gamma|\sigma^2)p(\sigma^2) \right\} \right]
\]

\[
= \mathbb{E}_{\sigma^2|y, \gamma = \hat{\gamma}^{(k)}} \left[ \log \left\{ p(y|\gamma) \right\} + \log \left\{ p(\gamma|\sigma^2) \right\} + \log \left\{ p(\sigma^2) \right\} \right]
\]

where we use the shorthand notation \( p(\cdot) \) to represent the marginal density of \( \cdot \) and \( p(\alpha|\theta) \) notation represents the conditional density of \( \alpha \) given \( \theta \). The first term of this expectation is free of \( \sigma^2 \). The second term of the expectation is over the sum of the logarithms of normal densities for \( \gamma_j \) and can be calculated element-wise. And the third term is free of \( \gamma \). Hence

\[
\mathbb{E}_{\sigma^2|y, \gamma = \hat{\gamma}^{(k)}} \left[ \log \left\{ p(y, \gamma, \sigma^2) \right\} \right] = \log \left\{ p(y|\gamma) \right\} + \frac{1}{2} \sum_{j=1}^{m} \left\{ \frac{\gamma_j^2}{\sigma_j^2} \right\} + \log(2\pi\sigma_j^2) + R
\]

\[
= \log \left\{ p(y|\gamma) \right\} - \frac{1}{2} \sum_{j=1}^{m} \gamma_j^2 \mathbb{E}_{\sigma^2|y, \gamma = \hat{\gamma}^{(k)}} \left( \frac{1}{\sigma_j^2} \right) + R',
\]

where \( R \) and \( R' \) are the remaining terms that are free of \( \gamma \). As shown in Appendix S1, the conditional distribution of \( \sigma_j^2 \) given \( \gamma \) is a scaled inverse Chi-square distribution with degrees of freedom \( \nu_\gamma + 1 \) and scale parameter \( \frac{\gamma_j^2 + \nu_\gamma S_\gamma^2}{\nu_\gamma + 1} \), and

\[
\mathbb{E}_{\sigma^2|y, \gamma = \hat{\gamma}^{(k)}} \left( \frac{1}{\sigma_j^2} \right) = \left( \frac{\gamma_j^{(k)} + \nu_\gamma S_\gamma^2}{\nu_\gamma + 1} \right)^{-1}
\]

and

\[
\mathbb{E}_{\sigma^2|y, \gamma = \hat{\gamma}^{(k)}} \left[ \log \left\{ p(y, \gamma, \sigma^2) \right\} \right] = \log \left\{ p(y|\gamma) \right\} - \frac{1}{2} \sum_{j=1}^{m} \frac{\gamma_j^2}{\gamma_j^{(k)} + \nu_\gamma S_\gamma^2} + R'.
\]

The M-step of the algorithm is to maximize the above expectation with respect to \( \gamma \), which is equivalent to finding the BLUP of SNP effects as described in the previous section, using \( \frac{\gamma_j^{(k)} + \nu_\gamma S_\gamma^2}{\nu_\gamma + 1} \) as effect variance for SNP \( j \), i.e., the \( j \)th diagonal element of \( D \). After iterating between the E-step and the M-step until convergence, a local posterior mode of \( \gamma \) will be obtained. Details of the maximization and the estimation equations are shown in Appendix S2. Because of the success of GBLUP in traditional breeding methods, we choose the starting values for \( \sigma_j^2 \) to be the variance under the GBLUP method, i.e. \( \frac{\sigma^2}{2\sum_{j=1}^{m} \sigma_j^2} \), where \( \sigma^2 \) is the genetic variance, which will be assumed known in simulation.
Simulation

Prediction of breeding values and computational efficiency of fastBayesA were compared to other methods by applying to simulated phenotypes and SNP genotypes of pedigreed populations. The initial generation comprised a population of effective size 500 that was randomly mated for 1,000 generations to reach mutation-drift equilibrium and then gradually expanded to an actual size of 2,000 in the next 4 generations. In the 1,004th generation, 20 sires and 200 dams were randomly sampled without replacement from the 2,000 individuals in generation 1,004 to represent the founders of the pedigree. Each of the 20 sires in these and subsequent generations was randomly mated to 10 different dams, with each dam producing 1 male and 1 female offspring. That scheme continued for several generations at a constant size of 400 (200 male and 200 female offspring).

Two datasets were generated for the comparison of alternative methods in terms of prediction accuracy of GEBV (Dataset A) and computing time (Dataset B). Dataset A includes four scenarios of different genetic architectures and Dataset B varies in training size and genome length. The scenarios used in each dataset are summarized in Table A.1. The standard scenario was a training group of 1,020 individuals from the first three pedigree generations, two chromosomes with ~ 1,000 SNPs each, and a total number of \(0.1M_e\) QTL, (A1 and B2 of Table A.1), where \(M_e\) is the number of independently segregating loci across the genome, computed following Goddard (2009) and Hayes et al. (2009c) and is given in Table A.1 for the different scenarios. SNP loci and QTL were sampled among simulated loci to have minor allele frequency larger than 0.05. For scenario B1, B2 and B4, the first 2, 3 and 6 pedigree generations were used for training, respectively, and the five generations following training were used for validation.

Each chromosome was 1 Morgan in length and initially evenly covered by 2,000 SNPs, among which 5 times the desired number of QTL were randomly positioned as candidate QTL to guarantee enough QTL segregating at mutation-drift equilibrium. The SNPs and QTL were biallelic, with initial allele frequencies 0.5 and in Hardy-Weinberg equilibrium. Mutation rate was \(2.5 \times 10^{-5}\) per meiosis per locus for both QTL and SNPs. The number of crossovers
per chromosome was sampled from a Poisson distribution with mean 1. Recombination rates were modeled by the Haldane mapping function (Haldane, 1919). At generation 1,004, all SNPs with minor allele frequency less than 0.05 were eliminated and the desired number of QTL were randomly selected from candidate QTL with minor allele frequency larger than 0.05. QTL effects were generated according to different scenarios and scaled to achieve a total genetic variance of 1.0 in generation 1,005. In scenarios where QTL variances were heterogeneous, QTL effects were randomly sampled from a Gamma distribution with shape parameter 0.4 and scale parameter 1.66 (Meuwissen et al., 2001), while in scenarios with constant QTL variances, the effect of the \( j \)th QTL was backsolved as the square root of \( \frac{1}{2p_j(1-p_j)} \), with equal probability of being positive or negative, where \( p_j \) is the minor allele frequency at generation 1,004.

True breeding values were obtained by summing up all QTL effects for a given individual. In Dataset A, normal random errors with mean 0 and variance 1.0 or 9.0 were added to true breeding values to generate phenotypes of traits with heritability 0.5 or 0.1, respectively. The simulated heritability for all scenarios in Dataset B was 0.5. For each scenario, these activities were repeated to provide 50 replicates. All replicates used the same initial SNP positioning but varied in the position of QTL and SNPs and in the effects of QTL after selecting loci with minor allele frequencies larger than 0.05.

For the analysis of the simulated datasets using the Bayesian methods, the degrees of freedom of the prior distribution for effect variance and residual variance was 4.2, following Meuwissen et al. (2001). BayesA and BayesB were implemented in genomic selection software GenSel (Fernando and Garrick, 2010). Formulation of BayesA and BayesB was almost identical with Meuwissen et al. (2001) except that the effect of each SNP instead of haplotype was sampled by MCMC in GenSel. The proportion of the number of QTL over the total number of SNPs was used for \( \pi \) in BayesB. Simulated variance components were provided to the mixed model equations in fastBayesA and used to estimate hyperparameters of prior distributions for variance components.

For Bayesian methods, the MCMC was run for 21,000 iterations, with the first 1,000 discarded as burn in. The fastBayesA algorithm stopped when the change of estimated SNP
effects became small, i.e.

\[
\frac{\left[\hat{\gamma}^{(k)} - \hat{\gamma}^{(k-1)}\right] \left[\hat{\gamma}^{(k)} - \hat{\gamma}^{(k-1)}\right]}{\hat{\gamma}^{(k)} \hat{\gamma}^{(k)}} < 1 \times 10^{-4}.
\]

**Results**

**Prediction accuracy and bias of GEBV under alternative genetic architectures**

Eight scenarios of contrasting heritability, number of QTL and distribution of QTL variance were simulated to represent a range of genetic architectures. The average correlation and regression coefficient of TBV on GEBV in the first validation generation from 50 replicates are shown in Table A.2. Method fastBayesA had similar accuracy to BayesA and was much more accurate than GBLUP but less accurate than BayesB, regardless of genetic architecture or heritability. The results are as expected, in that fastBayesA predicts GEBV with similar accuracy as BayesA.

As the number of QTL increased from 0.1 to 2.0\(M_e\), the accuracy of (fast)BayesA and BayesB decreased by up to 0.08, while that of GBLUP did not drop as much. This result is in accordance with Daetwyler et al. (2010), in that the accuracy of GBLUP was not affected by the number of QTL. However, even when the number of QTL was 2.0\(M_e\), the accuracy of the Bayesian methods remained higher than that of GBLUP, which contradicts Daetwyler et al. (2010), who found that the advantage of BayesB over GBLUP diminished as the number of QTL increased up to 1.0\(M_e\). The contradiction was probably due to the fact that the training size relative to genome length was much larger in our study than in Daetwyler et al. (2010).

Bias in the prediction of GEBV is shown by the deviation of regression coefficients from 1.0 in Table A.2. Except for BayesB, which had regression coefficients close to 1.0, regression coefficients were substantially below 1.0 for the other methods, as low as 0.75. In all scenarios, the regression coefficients for fastBayesA were smaller than those for BayesA, indicating larger bias of fastBayesA than BayesA in predicting TBV. This suggests that the estimated SNP effects and hence GEBV are not shrunk enough. The reason might be that the joint posterior mode of SNP effects, which is obtained as the estimate in fastBayesA, can deviate substantially
from the posterior means used in BayesA due to the asymmetry of the posterior densities. An
improper scale of the genomic relationship matrix could also result in biased GEBV.

Decline of accuracy over generations

Figure A.1 shows the mean prediction accuracy of GEBV in five consecutive generations
after training in the scenario with heritability 0.5 and 0.1 M_e QTL with equal variance. For all
four methods, accuracy decreased with generations, in agreement with Habier et al. (2007). The
accuracies of fastBayesA and BayesA were very similar in all five generations and were higher
than accuracies of GBLUP and lower than accuracies of BayesB. The decrease in accuracy over
the five generations was largest for GBLUP and smallest for BayesB, with (fast)BayesA in be-
tween. Similar trends were also observed in other scenarios with different genetic architectures
(results not shown).

Accuracies across EM iterations

To study the optimizing property of fastBayesA, accuracies of GEBV in the five validation
generations were calculated at each EM iteration until the convergence criterion was achieved.
Figure A.2 shows the accuracy at each iteration in the first validation generation from one
random replicate of each scenario in Dataset A (heritability was 0.5). The accuracy of GEBV
from fastBayesA increased gradually with iteration and stabilized at a higher accuracy than
GBLUP, which is the accuracy achieved in the first iteration. In Figure A.2, the accuracy
stabilized within 10 steps but the algorithm continued for several more steps before reaching
the convergence criterion, which was based on changes in estimated SNP effects rather than
estimated breeding values. This indicates that the accuracy of GEBV is insensitive to small
changes in SNP effects.

Computational efficiency of EM

Computational efficiency of different methods was compared in relation to training pop-
ulation size and size of SNP panels. Results are in Table A.3. Method fastBayesA has less
computing effort per SNP than BayesA. The increase in computation time is likely to be be-
tween quadratic to cubic with the number of individuals, depending upon the actual algorithm used for solving the mixed model equations.

**Discussion**

In this study, a fast EM algorithm fastBayesA was developed for genomic selection without MCMC. The method is non-stochastic, but only approximates BayesA estimates of marker effects and GEBV because it uses a joint posterior mode of effects rather than the posterior means used in BayesA. Compared with MCMC-based Bayesian methods on the simulated datasets, fastBayesA was shown to have similar prediction accuracy to BayesA but less computational effort per SNP than BayesA.

An EM algorithm with the marginal distribution of SNP effects modeled as a $t$ distribution was first proposed by Yi and Banerjee (2009) for mapping QTL with epistatic and genotype-by-environment interaction effects. Since their main objective was to map major QTL, they used few degrees of freedom and a small scale parameter for the inverse Chi-square prior for the effect variance, which imposed heavy shrinkage on small effects such that only large effects would be detected. This is not ideal for genomic prediction for which many SNPs with small effects can usefully contribute to predictions in models influenced by polygenic gene action. Based on the same EM formulation as Yi and Banerjee (2009), Hayashi and Iwata (2010) presented a generalized EM algorithm (EM-BSR) for genomic prediction, but in the M-step only partial maximization is performed. The method fastBayesA that was developed in this study, following Yi and Banerjee (2009), was also designed for predicting breeding values but has a different formulation than EM-BSR in the maximization step. In fastBayesA, the posterior distribution of SNP effects was jointly maximized using BLUP, which is more efficient and requires fewer EM iterations to converge. The advantage of the M-step of fastBayesA is that all SNP effects can be estimated simultaneously and computational efficiency is insensitive to the number of SNPs.

The computational efficiency of fastBayesA is sensitive to the number of individuals in training since construction and inversion of the realized relationship matrix is computationally expensive. For datasets with a large number of training individuals, the faster Jacobi iteration...
as in VanRaden et al. (2009) can be used to obtain the BLUPs of SNP effects in fastBayesA. Since computing time of the Bayesian MCMC methods is expected to increase linearly with the number of markers, fastBayesA can be advantageous over MCMC-based methods as marker density increases, as it will until all polymorphisms available from whole genome resequencing are used as candidates.

Both in BayesA and fastBayesA, inferences are based on the same posterior distribution that may not be unimodal, and both methods have to be used with caution. In BayesA the posterior mean is used to estimate SNP effects, and when the marginal posterior distribution for SNP effect is multimodal, the MCMC sampler will tend to stay within the neighborhood of a local mode and fail to visit other modes that are distant from this one (Celeux et al., 2000). Therefore, the empirical distribution from the MCMC samples may be different from the true posterior distribution and the posterior mean estimated by MCMC samples may not be accurate. In fastBayesA a joint posterior mode is used to estimate SNP effects, and the mode that the EM algorithm finds may not be the global mode. The GBLUP estimates of SNP effects provide a reasonable starting point that guarantees fastBayesA estimates will at least be no worse than GBLUP estimates.

Method fastBayesA results in similar prediction accuracy as BayesA because of their identical modeling of SNP effects. Any differences in accuracy are due to the fact that the joint posterior mode of SNP effects used in fastBayesA can be quite different from the posterior means used in BayesA. In Figure A.3, shrinkage estimation of SNP effects from ridge regression, BayesA, fastBayesA and VanRaden non-linear prediction A (VanRaden A) (VanRaden et al., 2009) are plotted against least squares estimates. Comparing with ridge regression, BayesA, fastBayesA and VanRaden A shrink small effects towards zero more than large effects. The estimates from fastBayesA are indistinguishable to that from BayesA for those effects larger than a certain value around 0.1 standard deviation and they are close to least squares estimates, but smaller effects are shrunk more heavily toward zero by fastBayesA than BayesA. The reason may be that the local modes of small effects that fastBayesA finds tend to be closer to zero than the mean. This suggests that calculating the mean like VanRaden A instead of mode can be an advantage in some cases since the maximization is over all possible effect values.
without getting stuck at local modes. Figure A.4 shows that in scenarios with 0.1\(M_e\) QTL, most of the large effects from fastBayesA tend to be bigger than those from BayesA but similar to those from BayesB, which indicates that with few QTL, the joint mode that fastBayesA finds tend to be larger than BayesA posterior means but close to BayesB posterior means, and that the shrinkage of large effects with fastBayesA is less than with BayesA but similar to BayesB. Furthermore, in scenarios with 2.0\(M_e\) QTL, most of the large effects from fastBayesA are bigger than those from either BayesA or BayesB, indicating that with a large number of QTL, the posterior mode that fastBayesA finds are even larger than posterior means of BayesB. However, Figure A.4 also shows that in all four scenarios of genetic architectures, there are subsets of estimated SNP effects that are almost zero with fastBayesA but are large with BayesA and BayesB. The reason might be that for these subsets of SNP effects, fastBayesA chose a mode that is close to zero and is far from the posterior means. This explains the lower accuracy of fastBayesA than BayesB, since some moderately large effects in BayesB are over-shrunk to zero by fastBayesA due to the convergence to a local mode. The above observations suggest that the shrinkage behavior of fastBayesA and the shape of the posterior distribution of SNP effects under the BayesA model require further study.

The regression coefficient of TBV on GEBV was smaller than 1.0 in most scenarios of Dataset A for both fastBayesA and BayesA, which means the variance of GEBV was inflated and GEBV should be shrunk more to make prediction of TBV unbiased (Meuwissen et al., 2001). Biases were greater for fastBayesA than BayesA, likely because of insufficient shrinkage of large effects, as shown in Figure A.3. Another reason might be that for BayesA residual variance was sampled by MCMC iteration while the simulated real residual variance was used for fastBayesA. The bias for fastBayesA is expected to become smaller than observed here when the residual variance is also updated as mean square error in each step of EM iteration (Appendix S2). This modified algorithm was applied to the 50 replicates of scenario A1. The average regression coefficient became 0.996 with no change in prediction accuracy.

Each single step of fastBayesA can be regarded as BLUP of breeding values based on a weighted marker-derived relationship matrix. The realized relationship between each pair of individuals not only incorporates information of genome fragments that are IBS or IBD given
high density SNP genotypes but also incorporates information about genetic architecture by allowing differing sizes of contributions of each SNP to the overall genetic variance. The relationship matrix used here is similar to the trait-specific relationship matrix in the heuristic TA-BLUP of Zhang et al. (2010) but differs in that TA-BLUP used genetic variance as weights for different SNPs. Method fastBayesA and TA-BLUP share the idea that SNPs that are in LD with QTL contribute more to the genetic covariance between individuals for a specific trait than SNPs that are in linkage equilibrium with QTL, but the maximizing behavior of TA-BLUP is not clear. Approximately, TA-BLUP could be regarded as one step of fastBayesA with an improper prior for effect variance, with degrees of freedom and scale parameter close to zero. Yi and Banerjee (2009) used degrees of freedom equal to 0.01 and scale parameter equal to $1 \times 10^{-4}$ for the prior of effect variance, which resulted in strong shrinkage of small effects. With this choice of hyperparameters, the effect variance of each SNP is dominated by the squared estimated effect and hence for small effects, the effect variance diminishes with EM iteration and the estimated effect is shrunk to zero. Method fastBayesA with such an improper prior was tested on datasets with $0.1M_e$ QTL with heterogeneous variance and heritability 0.5, and resulted in much lower prediction accuracy at convergence than in the first several iterations for several replicates (result not shown). This, however, suggests that improper priors, as in Yi and Banerjee (2009), can be used to identify the largest effects in genome wide QTL mapping studies but at the risk of decreased predictability for breeding values due to ignoring many small effects.

Method fastBayesA inherits the main advantages that GBLUP possesses and which MCMC-based methods lack. First, animals that have not been genotyped can be included in the model through pedigree relationship using single-step approach by Legarra et al. (2009) and Misztal et al. (2009), in which phenotypes from ungenotyped animals contribute to the estimates of breeding values and hence marker effects. For MCMC-based methods, genotypes of ungenotyped animals must be imputed in order to include them into the analysis since genotype is indispensable. Second, prediction error variance and hence reliability or accuracy of the GEBV of each animal (especially validation animals) could be obtained using methods by Strandén and Garrick (2009). For MCMC methods, the reliability of GEBV is available only when
the posterior distribution of GEBV is known. This requires interim validation during Markov Chain using the sampled SNP effects to calculate the prediction error variance of GEBV.

In conclusion, a fast EM algorithm fastBayesA is shown to approach BayesA estimates of marker effects without requiring MCMC. Simulation studies showed that fastBayesA has similar accuracy to BayesA under a range of genetic architectures. Method fastBayesA can be an appropriate substitute for BayesA for datasets with large numbers of markers or for pedigreed population with ungenotyped animals.

Acknowledgments

Critics from three reviewers were greatly acknowledged.

Bibliography


Fernando, R. L. and Garrick, D. J. (2010). *GenSel - User manual for a portfolio of genomic selection related analyses*. Animal Breeding and Genetics, Iowa State University, Ames, IA, USA.


Table A.1 **Summary of simulated datasets and scenarios.** Scenarios differed in training data size, number of chromosomes, number of QTL, and whether the genetic variance contributed by QTL was constant (const) or heterogeneous (hetero).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset A</th>
<th>Dataset B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario</td>
<td>A1</td>
<td>A2</td>
</tr>
<tr>
<td>Training size</td>
<td>1,020</td>
<td>620</td>
</tr>
<tr>
<td>No. chromosomes</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$M_e$</td>
<td>241</td>
<td>241</td>
</tr>
<tr>
<td>No. QTL</td>
<td>$0.1M_e$</td>
<td>$0.1M_e$</td>
</tr>
<tr>
<td>QTL variance</td>
<td>hetero</td>
<td>const</td>
</tr>
</tbody>
</table>
Table A.2  **Accuracy of GEBV and regression coefficient of TBV on GEBV in the first validation generation of Dataset A for GBLUP, BayesA, BayesB and fastBayesA.**

<table>
<thead>
<tr>
<th>Heritability</th>
<th>0.5</th>
<th>0.1</th>
<th>0.5</th>
<th>0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. QTL</td>
<td>0.1$M_e$</td>
<td>2.0$M_e$</td>
<td>0.1$M_e$</td>
<td>2.0$M_e$</td>
</tr>
<tr>
<td>QTL Variance</td>
<td>Hetero$^1$</td>
<td>Const$^2$</td>
<td>Hetero</td>
<td>Const</td>
</tr>
<tr>
<td>GBLUP</td>
<td>0.777$^3$</td>
<td>0.777</td>
<td>0.765</td>
<td>0.749</td>
</tr>
<tr>
<td>BayesA</td>
<td>0.832</td>
<td>0.834</td>
<td>0.778</td>
<td>0.764</td>
</tr>
<tr>
<td>BayesB</td>
<td>0.869</td>
<td>0.866</td>
<td>0.789</td>
<td>0.777</td>
</tr>
<tr>
<td>fastBayesA</td>
<td>0.839</td>
<td>0.841</td>
<td>0.777</td>
<td>0.763</td>
</tr>
</tbody>
</table>

**Regression coefficient of TBV on GEBV**

| GBLUP        | 0.979$^4$ | 0.981 | 0.984 | 0.968 | 0.953 | 0.949 | 0.954 | 0.888 |
| BayesA       | 0.947 | 0.955 | 0.985 | 0.976 | 0.942 | 0.952 | 0.956 | 0.901 |
| BayesB       | 1.019 | 1.009 | 0.996 | 0.991 | 1.050 | 1.083 | 0.964 | 0.932 |
| fastBayesA   | 0.902 | 0.905 | 0.887 | 0.873 | 0.887 | 0.891 | 0.906 | 0.867 |

1. Heterogeneous genetic variance of QTL.
2. Constant genetic variance of QTL.
3. Mean of correlation of TBV with GEBV over 50 replicates. Standard errors were less than 0.006 for all scenarios with heritability 0.5 and less than 0.015 for scenarios with heritability 0.1.
4. Mean of regression coefficient of TBV on GEBV over 50 replicates. Standard errors were less than 0.012 for all scenarios with heritability 0.5 and less than 0.036 for scenarios with heritability 0.1.

Table A.3  **Computing time (in seconds) for training by BayesA, BayesB and fastBayesA.**

<table>
<thead>
<tr>
<th>Training size</th>
<th>620</th>
<th>1,020</th>
<th>1,020</th>
<th>1,020</th>
<th>2,220</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. chromosomes</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>BayesA</td>
<td>321.7</td>
<td>479.8</td>
<td>1,215.2</td>
<td>2,492.8</td>
<td>928.8</td>
</tr>
<tr>
<td>BayesB</td>
<td>376.8</td>
<td>473.7</td>
<td>1,194.0</td>
<td>2,384.5</td>
<td>687.9</td>
</tr>
<tr>
<td>fastBayesA</td>
<td>25.3</td>
<td>63.0</td>
<td>114.6</td>
<td>168.2</td>
<td>350.5</td>
</tr>
</tbody>
</table>
Figure A.1  Prediction accuracy of GEBV in five validation generations by alternative methods. The scenario is $0.1M_e$ QTL with heterogeneous variance, heritability 0.5. Results are averaged over 50 replicates.
Figure A.2  Prediction accuracies of GEBV across EM iterations in the first validation generation. The four scenarios are $0.1 M_e$ QTL with constant variance ($0.1 M_e$, Const), $0.1 M_e$ QTL with heterogeneous variance ($0.1 M_e$, Hetero), $2.0 M_e$ QTL with heterogeneous variance ($2.0 M_e$, Hetero) and $2.0 M_e$ QTL with constant variance ($2.0 M_e$, Const). Results for each scenario are averaged over 50 replicates.
Figure A.3  
Shrinkage estimate of SNP effects from ridge regression (black line), BayesA (red dots), fastBayesA (blue line) and VanRaden non-linear prediction A (green line) against least squares estimate. SNP effects are measured in standard deviation units.
Figure A.4  Estimated SNP effects from fastBayesA ($y$ axis) against estimates from BayesA and BayesB ($x$ axis). All SNPs across 50 replicates are pooled for each scenario. Red dots show estimated SNP effects, and the blue line represents $y = x$. 
Supporting Information

Appendix S1 Expectation of the reciprocal of a scaled inverse Chi-square random variable.

Given that
\[ \gamma_j | \sigma_j^2 \sim N(0, \sigma_j^2) \]
and
\[ \sigma_j^2 \sim \frac{\nu \gamma S^2_\gamma}{\chi^2_{\nu \gamma}}, \]
The joint distribution of \( \gamma_j \) and \( \sigma_j^2 \) is
\[ p(\gamma_j, \sigma_j^2) \propto \exp\left(-\frac{\nu \gamma S^2_\gamma}{2\sigma_j^2}\right)(\sigma_j^2)^{-1+\nu \gamma 2} \cdot \exp\left(-\frac{\gamma_j^2}{2\sigma_j^2}\right) \]
\[ \propto \exp\left(-\frac{\nu \gamma S^2_\gamma + \gamma_j^2}{2\sigma_j^2}\right)(\sigma_j^2)^{-1+\frac{\nu \gamma 2+1}{2}}. \]

This is the kernel of the conditional distribution of \( \sigma_j^2 \) given \( \gamma_j \), which is a scaled inverse Chi-square distribution with degrees of freedom \( \nu \gamma + 1 \) and scale parameter \( \frac{\nu \gamma S^2_\gamma}{\nu \gamma + 1} \).

To show that
\[ E_{\sigma^2 | y = \gamma^{(k)}} \left( \frac{1}{\sigma_j^2} \right) = \left( \frac{\gamma_j^{(k)} + \nu \gamma S^2_\gamma}{\nu \gamma + 1} \right)^{-1}, \]
it suffices to show that the expectation of the reciprocal of a scaled inverse Chi-square variable is the reciprocal of its scale parameter. Suppose \( X \) is a scaled inverse Chi-square random variable with degrees of freedom \( \nu \) and scale parameter \( S^2 \), the probability density function for \( X \) is given by
\[ p(x | \nu, S^2) = \frac{(\nu S^2)^{\nu \gamma 2}}{\Gamma(\nu \gamma 2)} \cdot \frac{\exp\left(-\frac{\nu S^2}{2\nu \gamma 2}\right)}{x^{1+\nu \gamma 2}}. \]

It follows that the probability density function for \( Y = \frac{1}{X} \) is
\[ p(y | \nu, S^2) = \frac{(\nu S^2)^{\nu \gamma 2}}{\Gamma(\nu \gamma 2)} \cdot \frac{\exp\left(-\frac{\nu S^2}{2\nu \gamma 2}y\right)}{y^{1+\nu \gamma 2}} \cdot \left| \frac{1}{y^2} \right| \]
\[ = \frac{(\nu S^2)^{\nu \gamma 2}}{\Gamma(\nu \gamma 2)} \cdot \frac{\exp\left(-\frac{\nu S^2}{2\nu \gamma 2}y\right)}{y^{1+\nu \gamma 2}} \cdot y^{\nu \gamma 2-1}. \]
This is the probability density function of Gamma distribution with shape parameter \( \frac{\nu}{2} \) and rate parameter \( \frac{\nu S^2}{2} \). The expectation of Gamma distribution is the shape over rate and therefore the expectation of \( \frac{1}{X} \) is \( \frac{1}{\nu} \).

**Appendix S2 Estimation equations for parameters from fastBayesA.**

At convergence of the EM algorithm when \( \hat{\gamma}^{(k-1)} \approx \hat{\gamma}^{(k)} \), the fastBayesA estimates of SNP effects (\( \hat{\gamma} \)) and fixed effects (\( \hat{\beta} \)) satisfy

\[
\hat{\gamma} = \left( Z'Z + \hat{D}^{-1}\sigma^2_e \right)^{-1} Z' \left( y - X\hat{\beta} \right),
\]

and

\[
\hat{\beta} = \left( X'\hat{V}^{-1}X \right)^{-1} X'\hat{V}^{-1}y,
\]

in which

\[
\hat{D} = \text{diag} \left\{ \frac{\hat{\gamma}_j^2 + \nu_\gamma S^2_\gamma}{\nu_\gamma + 1} \right\}_{j=1}^m,
\]

and

\[
\hat{V} = Z\hat{D}Z' + I \sigma^2_e.
\]

In this study the residual variance is assumed known from simulation. In most cases where the residual variance is unknown, the estimate in the \( k \)th step of EM iteration is calculated as

\[
\{ \hat{\sigma}_e^2 \}^{(k)} = \frac{\left[ y - X\hat{\beta}^{(k)} - Z\hat{\gamma}^{(k)} \right]' \left[ y - X\hat{\beta}^{(k)} - Z\hat{\gamma}^{(k)} \right]}{n - \text{rank}(X)}.
\]
APPENDIX B. GENOMIC BREEDING VALUE PREDICTION AND QTL MAPPING OF QTLMAS2010 DATA USING BAYESIAN METHODS


Xiaochen Sun1, David Habier1, Rohan L. Fernando1,∗, Dorian J. Garrick1, and Jack C. M. Dekkers1§

1 Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, Iowa 50011, U. S. A.

§ Corresponding author

Abstract

Background

Bayesian methods allow prediction of genomic breeding values (GEBVs) using high-density single nucleotide polymorphisms (SNPs) covering the whole genome with effective shrinkage of SNP effects using appropriate priors. In this study we applied a modification of the well-known BayesA and BayesB methods to estimate the proportion of SNPs with zero effects (π) and a common variance for non-zero effects. The method, termed BayesCπ, was used to predict the GEBVs of the last generation of the QTLMAS2010 data. The accuracy of GEBVs from various methods was estimated by the correlation with phenotypes in the last generation. The methods were BayesCπ and BayesB with different π values, both with and without polygenic effects, and best linear unbiased prediction using an animal model with a genomic or numerator relationship matrix. Positions of quantitative trait loci (QTLs) were identified based on the
variances of GEBVs for windows of 10 consecutive SNPs. We also proposed a novel approach to set significance thresholds for claiming QTL in this specific case by using pedigree-based simulation of genotypes. All analyses were focused on detecting and evaluating QTL with additive effects.

Results

The accuracy of GEBVs was highest for BayesC\(\pi\), but the accuracy of BayesB with \(\pi\) equal to 0.99 was similar to that of BayesC\(\pi\). The accuracy of BayesB dropped with a decrease in \(\pi\). Including polygenic effects into the model only had marginal effects on accuracy and bias of predictions. The number of QTL identified was 15 when based on a stringent 10\% chromosome-wise threshold and increased to 21 when a 20\% chromosome-wise threshold was used.

Conclusions

The BayesC\(\pi\) method without polygenic effects was identified to be the best method for the QTLMAS2010 dataset, because it had highest accuracy and least bias. The significance criterion based on variance of 10-SNP windows allowed detection of more than half of the QTL, with few false positives.

Background

Genomic prediction of breeding values of individuals is based on a large number of SNPs across the whole genome giving high-density coverage. Each QTL is expected to be in linkage disequilibrium (LD) with at least one SNP because of the high marker density, hence the effects of all QTL are expected to be captured by SNPs (Meuwissen et al., 2001). Bayesian methods enable prediction of the effects of high-density SNPs covering the whole genome, even when the number of SNPs is much larger than the number of individuals with phenotypic and genotypic records. By specifying proper prior distributions for SNP effects, the ignorable small SNP effects are coerced to zero and only SNPs with larger effects on phenotype are fitted in the model. In BayesB, as proposed by Meuwissen et al. (2001), the prior specification for a SNP
effect is zero with fixed probability $\pi$, and normally distributed with a locus-specific variance with probability $(1 - \pi)$. The variance has an inverted Chi-square distribution with known degrees of freedom and scale parameter derived from the assumed known additive genetic variance. In this study, we applied a modification of BayesB, BayesC$\pi$ (Habier et al., 2011), where a single effect variance is common to all SNPs with non-zero effects, and the probability that a SNP has zero effect, $\pi$, is treated as unknown. This modification aims at overcoming the drawbacks of BayesB pointed out by Gianola et al. (2009), that the full-conditional posterior distribution is dominated by the prior and not by the data.

The availability of genome-wide SNP panels enables detection of statistical associations between a trait and any SNP in terms of a genome-wide association study (GWAS), enhancing the possibility of mapping QTL across the genome (Goddard and Hayes, 2009). Bayesian methods, such as those described above, are useful for GWAS QTL mapping because the inferences are based on the joint posterior distribution, which takes full account of all unknown parameters (Hoeschele et al., 1997; Zou and Zeng, 2008). The posterior probability of inclusion of each SNP into the model (we will refer to this as the model frequency) is mostly used as the criterion to detect QTL (Yi et al., 2005), as well as its derivatives, such as the Bayes factor (Yi et al., 2005), estimated QTL intensity (Sillanpää and Arjas, 1999), and the Bayes information criterion (Ball, 2001). Theoretically, within a class of SNPs that have the same model frequency, the model frequency indicates the proportion of the SNPs among them that are associated with QTL. This is, however, not always the true, especially when QTL and SNPs are in high LD and the effect of a single QTL could be spread over multiple SNPs. Therefore, to address the problem of model frequency with high density SNP panels, new criteria are needed to claim presence of a QTL in a frequentist way. Permutation tests, such as those used in least squares or maximum likelihood QTL interval mapping (Churchill and Doerge, 1994) for cross or family designs are not possible when data are from complex pedigrees, as was the case for the QTLMAS 2010 data.

Against this background, in this study we aimed to: (i) identify the Bayesian approach that most accurately predicts GEBV for the QTLMAS2010 data; (ii) develop a new criterion based on the 10-SNP window variance for QTL detection to concentrate signals from high density
SNP panels; and (iii) set significance thresholds for the window variance criterion to claim QTL when pedigree relationships exist among individuals.

Methods

Dataset

The simulated dataset was provided in advance of the 14th European QTL-MAS Workshop (Szydlowski and Paczynska, 2011). The population consisted of individuals in 5 generations (including founders) from 20 founders. Individuals from the first four generations had phenotypes for a quantitative trait. Full pedigree and gender were known. The genome contained 5 chromosomes, each 100 million base-pairs in length. All individuals were genotyped for 10,031 SNPs that were evenly spaced across the genome.

Predicting GEBVs

Four methods were used and compared for estimation of the marker effects and GEBV: BayesB (Meuwissen et al., 2001), BayesCπ (Habier et al., 2011), an animal model using the genomic relationship matrix (G-BLUP), and an animal model using the numerator-relationship matrix (P-BLUP). The latter P-BLUP results in the standard pedigree-based BLUP EBVs (Henderson, 1975). The effect of including polygenic effects was also investigated for the marker-based methods (BayesB, BayesCπ and G-BLUP). The statistical model for the marker-based methods with polygenic effects was

\[ y = 1\mu + Ws + u + \sum_j X_j \alpha_j \delta_j + e \]

where \( y \) is an \( N \times 1 \) vector of phenotypes with \( N \) being the numbers of individuals, \( \mu \) is the overall mean, \( W \) is the incidence matrix for gender, \( s \) is a \( 2 \times 1 \) vector with fixed gender effects, \( u \) is a vector with random polygenic effects of all individuals with \( \text{var}(u) = A\sigma_u^2 \), \( A \) is the numerator relationship matrix and \( \sigma_u^2 \) is the polygenic variance), \( X_j \) is an \( N \times 1 \) vector of genotypes at SNP \( j \), coded 0/1/2, \( \alpha_j \) is the random allele substitution effect for SNP \( j \), \( \delta_j \) is a 0/1-indicator variable which equals 1 if SNP \( j \) is included in the model and zero otherwise, and


\[ e \] is a vector of random residuals. Given the estimated marker effects and marker genotypes of an individual, its GEBV was calculated by

\[
\text{GEBV}_i = \hat{u}_i + \sum_j X_{ij} \hat{\alpha}_j
\]

where \( \hat{u}_i \) is the estimated polygenic effect of individual \( i \), \( X_{ij} \) is the marker genotype at SNP \( j \) of individual \( i \), and \( \hat{\alpha}_j \) is the estimated effect of SNP \( j \).

Method G-BLUP fitted all SNPs in the model, assuming that every SNP explained an equal proportion of the total genetic variance. Model BayesC used a modification of model BayesB of Meuwissen et al. (2001), and was described in detail by Habier et al. (2011). Model BayesC differs from BayesA and BayesB in its specification of the probability that a SNP has zero effect (\( \pi \)) and the variance of SNP effects \( \sigma^2_{\alpha_j} \). In BayesA and BayesB, each SNP has a locus-specific effect variance and this variance has a scaled inverted Chi-square distribution with degrees of freedom \( \nu_a \) and scale \( S^2_a \), which are functions of the assumed known additive genetic variance (Meuwissen et al., 2001). In BayesC, all SNP effects \( \alpha_j \) have a common variance, i.e. \( \sigma^2_{\alpha_j} = \sigma^2_{\alpha} \), which has a scaled inverse Chi-square prior distribution with degrees of freedom \( \nu_a \) and scale \( S^2_a \). As a result, the marginal distribution of all SNP effects in BayesC is a multivariate student’s t-distribution, \( t(0, \nu_a, S^2_a) \) (Habier et al., 2011). Furthermore, in BayesC the probability that a SNP has zero effect (\( \pi \)) was treated as unknown with uniform (0,1) prior. The prior for residuals \( e \) was a normal distribution with mean 0 and variance \( \sigma^2_e \). Gibbs sampling was applied to calculate the posterior means of model parameters \( \mu, s, \alpha, \sigma^2_{\alpha}, \sigma^2_{e}, \) and \( \pi \). The MCMC algorithms were run for 50,000 samples, with the first 20,000 samples discarded as burn in.

Effects of SNPs were estimated using the phenotypes and genotypes of individuals in the first three generations (training), which were then used to predict GEBVs of individuals in the fourth generation (validation) to evaluate the accuracy of GEBVs of the marker-based methods. The method giving the highest correlation of GEBVs with phenotypes in the validation population was used to predict the GEBVs of the fifth generation, for which only SNP genotypes but no phenotypes were available. For the fifth generation predictions, the first four generations were used to estimate SNP effects.
Detecting QTL

The parameter that was used for QTL detection was the variance of the GEBV of chromosome segments comprised of 10 adjacent SNPs, which we termed windows. First, SNP effects and variances were estimated using individuals in the first four generations by BayesC$\pi$, as described above. The GEBV for the 10-SNP window $l$ of individual $i$ ($\hat{\alpha}_{il}$) was computed as

$$\hat{\alpha}_{il} = \sum_{j=1}^{t+9} X_{ij} \hat{\alpha}_j, \quad j = 1, 2, \ldots, 10022,$$

and the variance of this prediction was calculated across individuals in the first four generations. For 1-SNP windows, this method is equivalent to calculating SNP variance as $2p_j(1 - p_j)\hat{\alpha}_j^2$ (Falconer and Mackay, 1996) for SNP $j$. Windows with variance of GEBVs above a predefined threshold were identified as QTL regions. Significant windows that overlapped were considered to identify the same QTL if there was only one variance peak among the SNPs covered by them. The variance for each window was graphically presented against genomic location of the SNP on the x-axis. Within each selected region, the SNP with the largest variance was used to quantify the position and variance of the QTL.

The threshold for the window variance for declaring presence of a QTL was determined by deriving the distribution of the window variance in data simulated under the null hypothesis of no LD between QTL and SNPs. Three strategies were used to generate data sets without LD between QTL and SNPs but using the original phenotypes, so as to maintain the distribution of phenotypes. The first strategy was to simply permute phenotypes against SNP genotypes across individuals in the training data. This strategy maintains LD relationships among SNPs in the original data but breaks all pedigree relationships and prevents SNPs to account for polygenic effects in the permuted data, in contrast to what happens in real data from pedigree populations (Habier et al., 2007). The second strategy was to randomly simulate SNP genotypes of individuals in the first 4 generations using the pedigree and SNP placement from the QTLMAS2010 data. The SNPs were assumed to be in linkage equilibrium (LE) in the founder generation. In this case, when estimating the simulated SNP effects using real phenotypes, the SNPs are expected to only capture polygenic effects through the pedigree but not the effects of QTL that underly the existing phenotypes. The simulation assumed that
1 million base-pairs mapped to 1 centimorgan and hence each chromosome was 1 Morgan in length. The third strategy was to simulate LD between the simulated SNPs in the founder generation at a level similar to that found in the QTLMAS2010 data, which was estimated using \( E(r^2) = (1 + 4cN_e)^{-1} \) (Sved, 1971). Multiple historical generations prior to the founder generation of the pedigree were simulated to create this LD. The effective size \((N_e)\) of the base population was set to 500, and randomly mated for 1,000 discrete generations, then reduced to an effective size of 100, and then increased over the next 10 generations to a size of 1,500, from which the 20 founders of the pedigree were randomly sampled. For all three datasets simulated under the null hypothesis, SNP effects and window variances were estimated using the simulated marker genotypes and real phenotypes by BayesC without polygenic effects and \(\pi\) set equal to the posterior mean of \(\pi\) from BayesC\(\pi\) when training on the first four generations using the original genotypes, i.e. the method used to obtain GEBVs for the final generation. The latter was done because estimates of \(\pi\) in the simulated null data set were much lower and resulted in very low significance thresholds because variances explained by each SNP were very low. The variances of GEBVs of all 10-SNP windows were calculated using the estimated SNP effects from the simulated data to obtain the distribution of the window variance under the null hypothesis.

To account for multiple testing across a chromosome, significance levels for the window variance were adjusted by dividing desired comparison-wise type I error rates by the effective number of loci \((M_e)\) in the genome, which was calculated by \(M_e = (2N_eL)/\ln(4N_eL)\), where \(N_e\) is the effective population size and \(L\) is the length of a chromosome, which was set to 1 Morgan (Goddard, 2009). This can be referred to as a Bonferroni adjustment for multiple testing across each chromosome. To set the thresholds, 10% (primary list) and 20% (secondary list) chromosome-wise type-I error rates were used, where the former was stringent and the latter more liberal.
Results and discussion

Accuracy of GEBV prediction

The accuracy of GEBVs was estimated in three ways: (i) the correlation of GEBVs with phenotypes divided by the square root of heritability (estimated from the full dataset with pedigree relationships using ASReml (Gilmour et al., 2009), which resulted in $\hat{h}^2 = 0.54$), (ii) the correlation of GEBVs with true breeding values (TBV), and (iii) the correlation of GEBVs with genotypic values. All these three accuracies were based on training on the first 3 generations and validation in generation 4. Results are in Table B.1.

The simulated QTLMAS2010 dataset had 30 biallelic additive QTL, 2 pairs of epistatic QTL and 3 paternally imprinted QTL. The QTL from each pair of epistatic QTL were close together and behaved as a single multi-allelic additive QTL. Each of the epistatic QTL-pairs and the imprinted QTL had the same effect as the largest additive QTL. The genotypic value of an individual was the sum of the genotypic value expressed in the phenotype at each of the QTL but the TBV also accounted for the imprinting effects that the individual had on its progeny. Thus, the TBV could deviate considerably from the genotypic values because the imprinted QTLs had large effects. In this study, all marker-based methods only fitted additive effects of SNPs derived based on the regression of SNP genotype on phenotype, which includes the effect of the imprinted QTL. As a result, as shown in Table B.1, the accuracy of prediction estimated from the correlation of GEBV with phenotype in the validation population was similar to the correlation of GEBV with genotypic values and the correlation of GEBV with TBV was much lower since the GEBV did not account for imprinting effects of parents on progeny.

The accuracy of P-BLUP was lowest among all methods, as expected. Method G-BLUP, which always fitted all SNPs in the model, had lower accuracy than BayesB and BayesC$\pi$. The Bayesian methods had quite similar accuracies, but BayesC$\pi$ tended to be the most accurate. Methods that fitted fewer SNPs performed better than those that fitted more. This might be explained by the fact that under the marker density of QTLMAS2010 data (measured as average $r^2 = 0.22$ between adjacent markers on chromosome 1, following Calus and Veerkamp
there were up to 100 SNPs in strong LD with the QTL and fitting more SNPs in the model resulted in underestimation of the effects of those SNPs.

The posterior mean of \( \pi \) in BayesC\( \pi \) was 0.988, that is, on average 124 SNPs were fitted in the model, which was similar to that of BayesB when \( \pi = 0.99 \) (Table B.2). Also BayesB with \( \pi = 0.99 \) and BayesC\( \pi \) fitted almost the same subset of SNPs when looking at the model frequencies of the SNPs fitted in the model. This explains the similar accuracy of BayesB with \( \pi = 0.99 \) and BayesC\( \pi \).

The bias of GEBV was evaluated based on the departure from unity of the regression coefficients of phenotype, TBV, and genotypic value on GEBV in the validation data (Table B.1). In general, all regression coefficients were very close to 1, showing that biases were small for all methods. For the marker-based methods, the regression coefficients of phenotype on GEBV were closest to 1; regression coefficients for TBV and genetic value were less than 1 and tended to be smallest for genotypic value. All regression coefficients dropped when the model included polygenic effects.

Model BayesC\( \pi \) without polygenic effects was applied to obtain the GEBVs of the final generation (5), with training on the first four generations because it resulted in high accuracy and small bias of GEBV based on training in the first three generations. Results at the bottom of Table B.1 show that the GEBVs from training on the first four generations were more accurate and less biased compared with training on the first three generations, because the training population size increased by 977 individuals and the SNP effects were more accurately estimated.

**Estimated variances**

Variance components estimated by the different models are shown in Table 2. Including polygenic effects in the model resulted in a larger estimated genetic variance and a smaller residual variance, and the estimated heritability was closer to the true value of 0.5, in accord with Calus and Veerkamp (2007). However, since no polygenic effects were simulated in the QTLMAS2010 dataset, including polygenic effects underestimated the variance explained by the SNPs, because some genetic variance due to relationships captured by the SNPs was taken
over by polygenic effects. Furthermore, the estimated variance components were not sensitive to the average number of SNPs included in the model, showing that around 100 SNPs were sufficient to capture most genetic variance.

**QTL mapping**

Several parameters estimated by BayesC\(\pi\) can be used to identify QTL regions, for instance, the absolute estimated effects of SNPs, the posterior inclusion probabilities (model frequencies) of SNPs, and the genetic variances explained by SNPs. Many Bayesian QTL mapping studies have applied model frequency or its derivatives as criteria to detect QTL (Yi et al., 2005; Sillanpää and Arjas, 1999; Ball, 2001). In those studies the markers were less dense and QTL were expected to be in LD with only one or several adjacent markers. However, for the high density SNP panel of the QTLMAS2010 data, the QTL and markers are expected to be in high LD (average \(r^2 = 0.22\) between adjacent markers on chromosome 1) and the effect of a single QTL could be spread over multiple SNPs. This results in too many signals in model frequency which could increase the probability of false positives and false negatives. To address this problem, we accumulated the effects of adjacent SNPs together into a genomic window. A window size of 10 was used in this study and the variance of GEBV of each 10-SNP window was used as the criterion to detect QTL. Several windows that shared the same SNP with a large effect were considered to identify the same QTL region. Within each region, because windows were overlapping, the window with the highest variance of GEBV was used and the SNP within this window that explained the largest proportion of genetic variance was used to denote the position of the QTL (Figure B.1).

Results of the three strategies to set significant thresholds are summarized in Table B.3. Plots of window variances against the identity of the first SNP of each window are shown in Figure B.1. With permutation of phenotypes against genotypes, the SNPs did not capture much genetic variance because pedigree relationships were destroyed and the variances of all 10-SNP windows were close to zero. Consequently the thresholds set by permutation were extremely low. The threshold determined by simulation of SNP genotypes was more reasonable than that from permutation because the relationships between individuals remained unchanged. Because
no QTL existed in the simulated genotypes, the SNPs only captured pedigree relationships. Genotypes of SNPs simulated without and with linkage disequilibrium in the founders captured similar proportions of total variance, but different subsets of SNPs were fitted in the model. The average number of SNPs in the model in the MCMC iterations was similar with 124 SNPs for simulated data sets due to the strong fixed prior $\pi$, but the model frequencies of the fitted SNPs were higher when LD was simulated and each of these SNPs explained a more genetic variance. As a result, the window variance thresholds were much higher for the data set with LD among founders. Therefore, for QTLMAS2010 data, where the genetic relationship among individuals were known, pedigree-based simulation of genotypes with initial LD was used to obtain the distribution of window variances under the null hypothesis of no intrinsic relationships between marker genotypes and phenotypes.

The threshold allowing a 10% chromosome-wise type-I error rate detected 13 QTLs of which 2 were false positives (Figure B.1). Each of the epistatic QTL-pairs was detected as one large QTL. A total of 20 small additive QTLs and 2 imprinted QTLs were missed. The threshold allowing a 20% chromosome-wise type-I error rate identified 6 more QTLs but 4 of these were false positives.

Adjustment for multiple testing was based on a Bonferroni-type of adjustment based on an estimate of the effective number of independent tests conducted. A more appropriate adjustment for multiple testing would be replicating the simulation multiple times and picking the highest window variance within each simulation. This replication procedure would resemble the method based on permutation tests proposed by Churchill and Doerge (1994), but would be more expensive computationally.

The window variance calculated using the sum of model-averaged SNP effects within a specific window will always underestimate the true QTL variance because of the shrinkage of SNP effects by BayesC$\pi$ and the incomplete LD between SNPs and QTL. Estimation of the variance of a window can be improved by computing the variance based on the sampled window effects from each sample of the MCMC chain, which is less shrunk than the posterior mean of the window effects.
Although grouping SNPs into windows is effective to concentrate signals, it also has several drawbacks. First, if say two QTL fall into the same region, by window variance they would likely be detected as one QTL; for example, additive QTL11 and QTL12 were detected as a single QTL (Figure B.1). Second, the effect of a single QTL may spread over more markers than the window length, especially in regions with weak LD between QTL and SNP; in this case windows over a wide region may show high variance, giving rise to the detection of multiple QTL for a region in which there is only one QTL. This is very likely to be the reason for the false positives reported around QTL1 (Figure B.1), whose effect spread over more than 40 SNPs when estimated by BayesC$\pi$. Third, window variance works well for relatively large QTL, but may shrink signals for small QTL, such as the eight undiscovered QTL on chromosome 4. Most of these eight QTL had detectable signals of SNP model frequency, but the window variances were below the thresholds that were set. All these drawbacks need to be further investigated, including the optimal size of windows to use.

The use of windows in this study is fundamentally different from the use of haplotypes to detect QTL, although both use combinations of adjacent markers. An alternative method may well be constructing haplotypes using two or more adjacent SNP alleles and estimating haplotype effects using Bayesian methods. Villumsen et al. (2009) showed that there is an optimal haplotype length for the accuracy of GEBV prediction depending on the population, LD, and marker spacing. Using haplotypes allows combining linkage disequilibrium and linkage analysis information by including the probability of identity-by-decent between haplotypes at the same locus, and the improved accuracies of LD-mapping (Grapes et al., 2004) and genomic selection (Calus et al., 2008) have already been reported. It is hence worthwhile to investigate the use of haplotype on the precision of Bayesian QTL mapping.

Conclusions

In this simulated dataset, BayesC$\pi$ slightly outperformed BayesB in the accuracy of predicting GEBV, but the accuracy of BayesB was similar to BayesC$\pi$ when its $\pi$ was set equal to the posterior mean of $\pi$ from BayesC$\pi$. The prediction accuracy of TBV was lower than that of genotypic values. Window variance allowed detection of most large QTLs but had insufficient
power to detect the small QTLs. Since the model only captured additive effects of QTLs, each epistatic QTL-pair was detected as one multi-allelic additive QTL and the two imprinted QTLs were not detected. The results expose the need for advanced statistical approaches to address more complicated patterns of genetic effects that exist in real data.

**List of abbreviations used**

GEBV: Genomic Estimated Breeding Value; SNP: Single Nucleotide Polymorphism; QTL: Quantitative Trait locus; LD: Linkage Disequilibrium; GWAS: Genome-wide Association Study; BLUP: Best Linear Unbiased Prediction; EBV: Estimated Breeding Value; MCMC: Markov Chain Monte Carlo; LE: Linkage Equilibrium; TBV: True Breeding Value

**Authors’ contributions**

XS carried out the analysis and drafted the manuscript. DH contributed to programming BayesC\(\pi\), developed the program for pedigree-based simulation, and helped to interpret the results. RLF contributed to programming BayesC\(\pi\) and helped to interpret the results. DJG contributed to programming BayesC\(\pi\) and helped to interpret the results. JCMD was the overall coordinator of the project, developed the method to set thresholds, and helped to interpret the results and draft the manuscript.

**Competing interests**

The authors declare that they have no competing interests.

**Acknowledgments**

Support for this research was provided by USDA AFRI Competitive Grants No. 2007-35205-17862 and 2010-65205-20341 from the National Institute of Food and Agriculture. Organizers of the 2010 QTLMAS workshop are acknowledged for providing the data.
Bibliography


Tables

Table B.1  **Prediction accuracy of GEBV, correlation of GEBV with TBV, correlation of GEBV with genotypic value (g), regression of phenotype (y) on GEBV, regression of TBV on GEBV, and regression of genotypic value on GEBV.** Results are based on training on the first three generations and validation on generation 4 using P-BLUP, G-BLUP, BayesB with different π’s, and BayesCπ, and without (No Poly) and with (Poly) polygenic effects.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Correlation of GEBV with</th>
<th>Regression coefficient on GEBV of</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>y&lt;sup&gt;1°&lt;/sup&gt;</td>
<td>TBV</td>
</tr>
<tr>
<td>P-BLUP</td>
<td>0.545</td>
<td>0.410</td>
</tr>
<tr>
<td>G-BLUP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Poly</td>
<td>0.746</td>
<td>0.610</td>
</tr>
<tr>
<td>Poly</td>
<td>0.737</td>
<td>0.597</td>
</tr>
<tr>
<td>BayesB, π = 0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Poly</td>
<td>0.781</td>
<td>0.632</td>
</tr>
<tr>
<td>Poly</td>
<td>0.778</td>
<td>0.628</td>
</tr>
<tr>
<td>BayesB, π = 0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Poly</td>
<td>0.788</td>
<td>0.640</td>
</tr>
<tr>
<td>Poly</td>
<td>0.784</td>
<td>0.634</td>
</tr>
<tr>
<td>BayesB, π = 0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Poly</td>
<td>0.793</td>
<td>0.646</td>
</tr>
<tr>
<td>Poly</td>
<td>0.790</td>
<td>0.636</td>
</tr>
<tr>
<td>BayesCπ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Poly</td>
<td>0.796</td>
<td>0.650</td>
</tr>
<tr>
<td>Poly</td>
<td>0.796</td>
<td>0.642</td>
</tr>
<tr>
<td>BayesCπ gen 5²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Poly</td>
<td>–</td>
<td>0.679</td>
</tr>
</tbody>
</table>

<sup>1</sup>Calculated as correlation of phenotype (y) with GEBV, divided by the square root of estimated heritability.

<sup>2</sup>Training on the first 4 generations and predicting generation 5.
Table B.2  **Average number of SNPs (#SNP) fitted in the model, estimated variance components, and estimated heritability (Heritability).** Results are based on training on the first three generations and validation on generation 4 using P-BLUP, G-BLUP, BayesB with different $\pi$’s, and BayesC$\pi$, and without (No Poly) and with (Poly) polygenic effects.

<table>
<thead>
<tr>
<th>Methods</th>
<th>#SNP</th>
<th>Estimated variance components</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Marker</td>
<td>Polygenic</td>
</tr>
<tr>
<td>True value$^2$</td>
<td></td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>P-BLUP</td>
<td>10031</td>
<td>–</td>
<td>54.44</td>
</tr>
<tr>
<td>G-BLUP</td>
<td></td>
<td>38.53</td>
<td>12.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44.54</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38.53</td>
<td>12.09</td>
</tr>
<tr>
<td>BayesB, $\pi = 0.75$</td>
<td>2508</td>
<td>44.28</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39.05</td>
<td>11.06</td>
</tr>
<tr>
<td>BayesB, $\pi = 0.95$</td>
<td>502</td>
<td>43.96</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>38.05</td>
<td>12.80</td>
</tr>
<tr>
<td>BayesB, $\pi = 0.99$</td>
<td>100</td>
<td>43.44</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>37.43</td>
<td>12.35</td>
</tr>
<tr>
<td>BayesC$\pi$</td>
<td></td>
<td>124</td>
<td>45.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>40.21</td>
</tr>
<tr>
<td>BayesC$\pi$ gen 5$^3$</td>
<td></td>
<td>92</td>
<td>47.13</td>
</tr>
</tbody>
</table>

$^1$Total genetic variance = marker variance + polygenic variance.

$^2$Total QTL variance = residual variance = 51.76 in the QTLMAS2010 dataset.

$^3$Training on the first 4 generations.
Table B.3  Variance components estimated from datasets generated by permutation, simulation with linkage equilibrium in founders (LE simulation), and simulation with initial linkage disequilibrium (LD simulation), and thresholds for 10-SNP window variances based on 10% and 20% chromosome-wise type I error rates.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Variance Components</th>
<th>Window variance threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genotypic</td>
<td>Residual</td>
</tr>
<tr>
<td>Permutation</td>
<td>3.15</td>
<td>98.59</td>
</tr>
<tr>
<td>LE simulation</td>
<td>20.83</td>
<td>79.84</td>
</tr>
<tr>
<td>LD simulation</td>
<td>17.14</td>
<td>83.40</td>
</tr>
<tr>
<td>Original$^1$</td>
<td>47.13</td>
<td>53.48</td>
</tr>
</tbody>
</table>

$^1$Estimated from the original QTLMA$^2$010 dataset using BayesC$^\pi$, training on the first 4 generations.
Figure B.1  **Variances of GEBVs of 10-SNP windows across the genome.** Data sets were generated by permutation (Permuted dataset), simulation with linkage equilibrium in founders (LE simulation dataset), and simulation with initial linkage disequilibrium (LD simulation dataset). The bottom panel show window variances obtained for the original QTLMAS 2010 dataset (Original dataset), as well as the location and variances of true QTLs, along with their mode of inheritance (Additive = additive QTL, Epistatic = epistatic QTL, Imprinted = imprinted QTL). Horizontal lines show the 10% (solid) and 20% (dash) chromosome-wise thresholds for window variance derived from the LD simulation.