Improving Accuracy of Genomic Prediction in Dairy and Beef Cattle

by

Liuhong Chen

A Thesis
presented to
The University of Guelph

In partial fulfillment of requirements
for the degree of
Doctor of Philosophy
in
Animal and Poultry Science

Guelph, Ontario, Canada

© Liuhong Chen, April, 2013
ABSTRACT

IMPROVING ACCURACY OF GENOMIC PREDICTION IN DAIRY AND BEEF CATTLE

Liuhong Chen
University of Guelph, 2013

Advisor: Dr. Flavio Schenkel

The overall goal of this thesis was to improve the accuracy of genomic prediction in dairy and beef cattle by developing, evaluating and enhancing novel or existent models and approaches for genomic selection. Four studies were conducted to fulfill this goal. In the first study, the impact of using genotypes imputed from low density panels for genomic prediction was evaluated and compared between a Bayesian mixture model and the Genomic Best Linear Unbiased Prediction (GBLUP) method. Results showed that for traits affected by a few large QTL, the Bayesian mixture model resulted in greater reduction in accuracy of genomic prediction, compared to GBLUP. However, for all SNP panels, scenarios and all traits studied, the Bayesian mixture model produced greater or similar accuracy, compared to the GBLUP method. In the second study, a new computing algorithm, called right-hand side updating strategy (RHSU), was proposed and compared to the conventional Gauss-Seidel residual update algorithm (GSRU) for genomic prediction. Results showed that RHSU would outperform GSRU once the sample size exceeded a fraction of the number of the SNPs. As the sample size continued to grow, the RHSU algorithm became more efficient than GSRU. In the third study, three different strategies of forming a training population for genomic prediction, within-breed, across-breed and pooling data from different breeds, were evaluated in Angus and Charolais steers using phenotypes on residual feed intake (RFI) and genotypes on the Illumina BovineSNP50 Beadchip (50k). Results suggested that using the 50k SNP panel, within-breed genomic prediction was a safe strategy; across-breed prediction resulted in the lowest accuracy; pooling data from different breeds had a
potential to improve the accuracy but should be conducted with caution due to possible loss of accuracy. In the last study, a multi-task Bayesian learning model was proposed for multi-population genomic prediction. The performance of the multi-task model was evaluated in Holstein and Ayrshire dairy breeds. Results showed that the multi-task Bayesian learning model is effective and could be beneficial to smaller populations where only a limited number of training animals are available.
ACKNOWLEDGEMENTS

I would like to express my sincere thanks to many people who helped me during my graduate studies at the University of Guelph.

First of all, I would like to thank my advisor Dr. Flavio Schenkel for providing me with the great opportunity to pursue my PhD degree. It was my extreme luck that I could conduct my study in the newly emerging area of genomic selection. He has created a wonderful research environment that made my life very exciting and full of fun. Dr. Schenkel’s guidance, unconditional help and patience made this thesis possible.

Supports from my advisory committee members, Dr. Zeny Feng, Dr. Niel Karrow, Dr. Changxi Li, Dr. Stephen Miller, Dr. Mehdi Sargolzaei and Dr. Larry Schaeffer are greatly appreciated. In the past several years, they have provided me with invaluable comments and suggestions. When discussing with them, they can always enlighten me with fresh ideas.

Specially, I would like to thank Dr. Changxi Li at the University of Alberta, who offered me a job position during my part-time study period. Without his great support and encouragement, this work could not be completed.

I appreciate the financial support from the Natural Sciences and Engineering Research Council of Canada, L’Alliance Boviteq Inc., the AAFC Livestock Genetics & Genomics Program funds, AAFC A-base peer-reviewed research project (RBPI-1139) and the BCRC Beef Cluster project (FDE.05.09).

Last but not least, I would like to thank my wife Jinqing Zhou, my daughter Shanna Chen, my sister, my parents and my friends both in Canada and in China. Without their understanding and support, nothing would be possible.
DECLARATION OF WORK PERFORMED

I declare that all work reported in this thesis was performed by myself.
# TABLE OF CONTENTS

ABSTRACT ......................................................................................................................... ii

ACKNOWLEDGEMENTS ..................................................................................................... iv

DECLARATION OF WORK PERFORMED ........................................................................... v

LIST OF FIGURES ............................................................................................................. viii

LIST OF TABLES ................................................................................................................ x

GENERAL INTRODUCTION .............................................................................................. 1

Chapter 1 Impact of Genotype Imputation on the Performance of GBLUP and
Bayesian Methods for Genomic Prediction ................................................................. 7

Abstract ............................................................................................................................. 7

Introduction ......................................................................................................................... 8

Materials and Methods ..................................................................................................... 9

Results and Discussion ................................................................................................... 15

Conclusions ....................................................................................................................... 21

Chapter 2 An Alternative Computing Algorithm for Genomic Prediction Using
Bayesian Mixture Model and Large Training Data Sets .............................................. 30

Abstract ............................................................................................................................. 30

Introduction ......................................................................................................................... 31

Materials and Methods ..................................................................................................... 33

Results ................................................................................................................................ 39

Discussion ......................................................................................................................... 40

Conclusions ....................................................................................................................... 43
Chapter 3 Accuracy of Predicting Genomic Breeding Values for Residual Feed Intake in Angus and Charolais Beef Cattle ........................................... 48

Abstract ........................................................................................................... 48
Introduction ....................................................................................................... 49
Materials and Methods ................................................................................... 50
Results .............................................................................................................. 59
Discussion ......................................................................................................... 62
Conclusions ....................................................................................................... 66

Chapter 4 Multi-Population Genomic Prediction Using a Multi-Task Bayesian Learning Model ................................................................. 71

Abstract ........................................................................................................... 71
Introduction ....................................................................................................... 72
Methods ............................................................................................................ 74
Real data example .......................................................................................... 79
Results .............................................................................................................. 82
Discussion ......................................................................................................... 85
Conclusions ....................................................................................................... 88

GENERAL DISCUSSION .................................................................................... 94

LITERATURE CITED ......................................................................................... 100

APPENDICES ................................................................................................. 109

Appendix A ..................................................................................................... 109
Appendix B ...................................................................................................... 111
Appendix C ...................................................................................................... 115
Appendix D ...................................................................................................... 118
LIST OF FIGURES

Figure 1.1 Comparison of SNP effects estimated from Ridge-regression method and the Bayesian mixture model under scenario S0. .......................................................... 26

Figure 1.2 SNP effects for fat percentage estimated from Ridge-regression method and the Bayesian mixture model under scenario S0. ...................................................... 27

Figure 1.3 Comparison of genomic prediction accuracy reduction rate\(^1\) between the Bayesian mixture model and GBLUP method for different traits, low density panels and scenarios\(^2\). .................................................................................. 29

Figure 2.1 Relative efficiency of the RHSU computing strategy for different ratios between number of animals in the training set and number of SNP markers and different proportions of SNPs (\(\pi\)) with no effect on the trait. ..................... 44

Figure 2.2 The threshold\(^1\) value for the ratio between number of animals in the training set (\(n\)) and the number of SNP markers (\(m\)) for RHSU to outperform GSRU for different proportions of SNPs (\(\pi\)) with no effect on the trait. ......................... 45

Figure 2.3 Comparison between two computing strategies keeping the proportion of SNP (\(\pi\)) with no effect on Fat % at 0.95, using real Holstein data. ................. 46

Figure 2.4 Comparison between two computing strategies estimating the proportion of SNP (\(\pi\)) with no effect on Fat %, using real Holstein data. ................................. 47

Figure 3.1 Distributions of additive genetic relationships (estimated from pedigree) between training and validation animals. ................................................................. 69

Figure 3.2 LD phase persistence between training and validation populations. ........... 70

Figure 4.1 Accuracy of DGV predicted using training data sets of different sizes..... 91
Figure 4.2 Signals (posterior inclusion probabilities) and effects for SNPs close to the DGAT1 region estimated using 28,206 SNPs.

Figure 4.3 Signals (posterior inclusion probabilities) and effects for SNPs close to the DGAT1 region estimated using 246,668 SNPs.
LIST OF TABLES

Table 1.1 Summary of the number of SNPs on each chromosome in various SNP panels used in this study................................................................. 22

Table 1.2 Summary of the number of animals used in this study. ...................... 23

Table 1.3 Accuracy of imputed genotypes from low density panels to 50k SNP panel. ............................................................................................................. 23

Table 1.4 Accuracy of genomic prediction and posterior estimates of π using observed 50k SNP genotypes under scenario S0. ......................................................... 24

Table 1.5 Accuracy of genomic prediction using imputed 50k SNP genotypes: Accuracies from using Bayesian model are parenthesized and accuracies for GBLUP are presented outside the parenthesis. ............................................. 25

Table 3.1 Distribution of the number of animals in each year. ............................. 68

Table 3.2 Realized accuracy of DGV predicted via PBLUP, GBLUP, and BayesB methods. ................................................................. 68

Table 4.1 Number of animals used for genomic prediction. ............................. 89

Table 4.2 Performance of genomic prediction using 28,206 SNPs via the single-task Bayesian learning model, the simple data pooling method, and the multi-task Bayesian learning model. .............................................................. 89

Table 4.3 Performance of genomic prediction using 246,668 SNPs via the single-task Bayesian learning model, the simple data pooling method, and the multi-task Bayesian learning model. .............................................................. 90
Table S1 Estimated posterior model inclusion probability and effects for SNPs close to the region of DGAT1 gene using 28,206 SNPs. .......................................................... 118

Table S2 Estimated posterior model inclusion probability and effects for SNPs close to the region of DGAT1 gene using 246,668 SNPs. ......................................................... 119
GENERAL INTRODUCTION

Genomic selection has become a new tool for genetic improvement in livestock species and plants thanks to the discovery of many thousands of single nucleotide polymorphisms (SNP) spreading across the whole genome and cost-effective high throughput genotyping technology. Genomic selection refers to selection for breeding candidates based on genomic estimated breeding values (GEBV) through the use of dense markers covering the whole genome (Meuwissen et al., 2001). Prediction of GEBV usually involves two steps. First, a training population having both genotypes and phenotypes is used to derive the prediction equations for the trait of interest, where all marker effects are estimated simultaneously. Next, GEBV for selection candidates that have only genotypes are predicted by summing up the marker effects. The advantage of genomic selection is that GEBV can be obtained at an early age of a selection candidate. Therefore, the generation interval can be drastically reduced (especially in cattle) in comparison to the long waiting time in a traditional progeny-testing breeding strategy. A greater rate of genetic improvement can be achieved and the cost could be substantially reduced in a breeding program implementing genomic selection (Schaeffer, 2006).

Numerous methods have been proposed for the prediction of GEBV. The most commonly used methods are a genomic BLUP method (GBLUP), Bayesian mixture models and Bayesian shrinkage models. GBLUP uses genotypic data to calculate the genetic relationship matrix and is shown to be equivalent to Ridge-regression (Goddard, 2009; VanRaden, 2008). Bayesian mixture models include Bayesian stochastic search variable selection method (SSVS) (Calus et al., 2008; Verbyla et al., 2009), BayesB (Meuwissen et al., 2001), BayesC and BayesC\(\pi\) (Habier et al., 2011). Bayesian shrinkage
models include Bayesian adaptive shrinkage (Xu, 2003), Bayesian LASSO (Park and Casella, 2008) and extended Bayesian LASSO (Mutshinda and Sillanpaa, 2010). Among these methods, prior distributions for SNP effects are generally different. GBLUP or Ridge-regression assumes that all SNP effects are normally distributed with a common variance. Bayesian mixture models assume SNP effects follow some type of spike and slab mixture distribution, where the spike is concentrated around zero and the slab distribution has a large variance which allows the SNP effects to take a wide range of values. A larger weight is placed on the spike distribution to indicate most of the SNPs are not associated with the trait. Bayesian shrinkage models use sparse-induced priors to shrink unimportant SNP effects to zero. The performance of these methods depends on the genetic architecture underlying the studied trait (Daetwyler et al., 2010). It is widely accepted that Bayesian methods perform similarly or slightly worse than GBLUP when the target trait is affected by many quantitative trait loci (QTL) each with a small effect, and better when the trait is influenced by a few large QTL (Daetwyler et al., 2010; Hayes et al., 2009b; VanRaden et al., 2009).

In dairy cattle, genomic selection has been implemented in several countries. Applications in Holstein cattle revealed that GEBV yielded much higher accuracy for young selection candidates, compared with the accuracy of parental average (Hayes et al., 2009b; VanRaden et al., 2009). The gain in accuracy was achieved by genotyping a large number of animals, mainly proven bulls and young selection candidates, on a medium high density (HD) single nucleotide polymorphism (SNP) array (BovineSNP50 BeadChip; 50k). However, the price for the 50k SNP array has impeded it from wider applications, such as prescreening a large number of young bulls, selecting heifers for
replacement, mating or management purposes (Weigel et al., 2010). The circumstance has motivated researchers to develop alternative low-cost low-density SNP panels.

Low density panels have been extensively studied in several aspects, such as comparing different strategies to select SNPs into a low density panel, evaluating the accuracy of imputing genotypes from low density to 50k or denser panels using various imputation algorithms, evaluating the impact of using low density SNP genotypes on the accuracy of genomic prediction, and exploring alternative strategies of utilizing the low density panels (Boichard et al., 2012; Habier et al., 2009; Huang et al., 2012; Khatkar et al., 2012; Moser et al., 2010; Mukler et al., 2012; Vazquez et al., 2010; Weigel et al., 2009; Weigel et al., 2010; Zhang and Druet, 2010). A common finding from these studies is that the use of low density panels might result in some loss in genomic prediction accuracy, but possibly be a cost-effective strategy provided that a large proportion of animals are genotyped on a low-cost evenly spaced low-density (ELD) SNP panel, followed by imputing ELD genotypes into HD using a reference population genotyped with a HD panel.

Because GBLUP and Bayesian methods make different assumptions about the distributions of SNP effects, errors in imputed SNP genotypes would likely have different influences on these methods. However, little is known about how alternative genomic prediction methods react differently to errors in genotypes imputed from low density panels. Chapter 1 in this thesis was devoted to evaluate the different impact of imputation errors on the accuracy of genomic prediction using GBLUP and a Bayesian mixture model.
Genomic prediction via GBLUP can be implemented through some efficient methods and computing strategies (Legarra and Misztal, 2008; VanRaden, 2008). Application of Bayesian methods requires evaluation of the joint posterior distribution, which cannot be directly obtained. So far, most of the Bayesian methods are implemented via Markov Chain Monte Carlo (MCMC) sampling algorithms. MCMC is computationally demanding in scenarios with large number of SNPs and large training data set, which are usually the cases as in genomic selection. Alternatively, approximation methods, such as maximum a posteriori (MAP) or variational Bayes method could be used to tackle the problem. Although these methods are shown to be much faster in computing time compared to MCMC, they do not guarantee to converge to the global maxima and therefore, may result in loss of prediction accuracy compared to their MCMC counterparts (Hayashi and Iwata, 2010; Li and Sillanpaa, 2012; Meuwissen et al., 2009).

The most time-costly step in the MCMC sampling algorithm for genomic prediction is to compute the mean of the full conditional posterior distribution for each unknown SNP effect in every sampling cycle. Legarra and Misztal (2008) proposed an algorithm termed Gauss-Seidel with residual update (GSRU) which can be easily adapted to the MCMC sampling procedure. The number of computational operations with GSRU algorithm is proportional to the product of sample size and the number of SNP effects. With more and more animals being genotyped and proven, the sample size is growing accordingly, reaching several hundreds of thousands as in the case of Holstein cattle in North America. There is a need to develop an algorithm where the computing time is less dependent on the sample size. It is assumed that, in the Bayesian mixture models, such as
BayesB and BayesCπ, only a small proportion of SNPs has non-zero effects. Therefore, it is possible to take advantage of the sparsity in the vector of SNP effects to speed up the computation. In Chapter 2, an alternative computing algorithm was proposed and its performance was compared to GSRU.

The key to success of genomic selection is the accuracy of GEBV, which depends on the number of individuals in the training population, the heritability of the trait, the number of loci and the size of their additive effects on the trait, the marker density and thus, the extent of linkage disequilibrium (LD) between markers and QTL, and the LD phase persistence between the training population and the selection candidates (Goddard and Hayes, 2009). With the development of high density SNP chips and reduced costs of SNP genotyping, acquisition of a large number of animals to form a training data set still remains a challenge for small populations, such as Ayrshire and most of the beef cattle breeds. Therefore, to achieve a maximal accuracy of genome prediction, strategies for forming training populations to predict the genomic breeding value for selection candidates need to be evaluated. In Chapter 3, different strategies of forming training populations are evaluated for genomic prediction for residual feed intake (RFI) in Angus and Charolais beef populations.

One strategy to improve the accuracy of genomic prediction in a small population is to combine the data of the small population with data from other populations to increase the total number of animals in the reference set. However, simply pooling data from different populations may result in unfavorable accuracies if the marker effects are very different among the populations or if the populations have diverged for a long time (de Roos et al., 2009). Genomic prediction with multiple populations can be treated as a
multi-task problem, where each task is to derive prediction equations for each population and tasks are related through some common features among the populations. In Chapter 4, a multi-task Bayesian learning model for multi-population genomic prediction was proposed and evaluated in real data from Holstein and Ayrshire dairy cattle.

The overall goal of this thesis was to improve the accuracy of genomic prediction in dairy and beef cattle. To fulfill this goal, several existing or new methods, algorithms and strategies for genomic prediction were presented and evaluated. In Chapter 1, the performance of GBLUP and a Bayesian mixture model was evaluated for genomic prediction using 50k SNP genotypes imputed from low density SNP panels under various scenarios, where different proportions of training animals were genotyped on low density SNP panels. The two methods were compared for their sensitivity to errors in the imputed genotypes. In Chapter 2, a new computing strategy was proposed and compared to the conventional computing algorithm for the implementation of MCMC sampling steps in a Bayesian mixture model for genomic prediction. In Chapter 3, the accuracy of genomic prediction for RFI in Angus and Charolais beef populations was evaluated using three strategies to form the training population, namely within-breed, across-breed and pooling data from the two breeds. And in Chapter 4, a novel multi-task Bayesian learning model for multi-population genomic prediction was presented. The model was evaluated in Holstein and Ayrshire dairy populations, using phenotypic data on milk production traits and genotypic data from the 50k and 777k SNP panels. The accuracy of genomic prediction using the multi-task Bayesian mixture model was compared to baseline approaches of a single-task within population prediction and a simple data pooling method.
Chapter 1 Impact of Genotype Imputation on the Performance of GBLUP and Bayesian Methods for Genomic Prediction

Abstract

Genotypes derived from low density marker panels are commonly imputed to high density (HD) genotypes and subsequently used for genomic selection. Imputation errors would likely have different impacts on different genomic prediction methods, but so far have not been documented. The aim of this study was to evaluate the performance of GBLUP and a Bayesian mixture model using imputed genotypes for genomic prediction. Originally, a total of 10,309 Holstein bulls were genotyped on the BovineSNP50 BeadChip (50k). Five low density single nucleotide polymorphism (SNP) panels, containing 6,177, 2,480, 1,536, 768 and 384 SNPs, were simulated. Training and validation sets for genomic prediction, and the reference population for genotype imputation were defined according to birth year and reliability of bull proofs or de-regressed bull proofs. A fraction of 0%, 33% and 66% of the animals were randomly selected from the training and reference sets to have low density genotypes, which were then imputed into 50k genotypes. GBLUP and the Bayesian mixture model were used to predict direct genomic values (DGV) for validation animals using imputed or their actual 50k genotypes. Traits studied included milk yield, fat yield, protein yield, fat percentage, protein percentage and somatic cell score (SCS). Results showed that the impact of imputation errors on GBLUP and the Bayesian mixture model applied depends on the genetic architecture of the trait. For traits affected by a few large QTL, the Bayesian mixture model resulted in greater reduction in accuracy of genomic prediction than GBLUP, when the density of the low density SNP panel was decreased. However, for all
SNP panels, scenarios and all traits studied, the Bayesian mixture model produced higher or similar accuracy compared to the GBLUP method. When different low density panels were compared, having all animals in the training set genotyped on 50k SNP panel and animals in the validation set genotyped on the 6k SNP panel the accuracy of genomic prediction was the greatest for both GBLUP and the Bayesian mixture model. Imputation from SNP panels with a density lower than 6k was more prone to errors and resulted in lower accuracy of genomic prediction. Including genotypes imputed from the 6k panel achieved almost the same accuracy of genomic prediction as that of using the 50k panel, even when 66% of the training animals were genotyped on the 6k panel.

Introduction

Genomic selection (Meuwissen et al., 2001) has become a new tool for genetic improvement in livestock species and plants thanks to the discovery of many thousands of single nucleotide polymorphisms (SNP) spreading across the whole genome and cost-effective high throughput genotyping technology. Two common classes of methods for genomic prediction are the genomic best linear unbiased prediction (GBLUP) and Bayesian methods. Among the Bayesian methods, Bayesian mixture models are widely used (Calus et al., 2008; Habier et al., 2011; Meuwissen et al., 2001; Verbyla et al., 2009). The accuracy of genomic prediction using different methods depends on the genetic architecture underlying the studied trait (Daetwyler et al., 2010). For traits that are affected by a few large quantitative trait loci (QTL), Bayesian methods would usually outperform GBLUP, but when the trait is affected by many QTL each with a small effect, GBLUP would perform as well as Bayesian methods (Daetwyler et al., 2010; Hayes et al., 2009b; VanRaden et al., 2009).
In dairy cattle, substantial gains in accuracy of genomic estimated breeding values over traditional parental averages for young selection candidates have been observed in many countries (Hayes et al., 2009b; VanRaden et al., 2009). These gains of accuracy were achieved by using the BovineSNP50 BeadChip (50k; Illumina Inc., San Diego, USA). Recently, lower density SNP panels have drawn huge attentions due to their lower genotyping cost and possible wider applications, such as prescreening a large number of young bulls and selecting heifers for replacement, mating or management purposes (Weigel et al., 2010).

Strategies of forming and utilizing low density panels have been discussed in several studies (Boichard et al., 2012; Habier et al., 2009; Huang et al., 2012; Khatkar et al., 2012). It is widely accepted that low density panels are imputed to high density (HD) panels for subsequent use in the genomic prediction (Mulder et al., 2012; Zhang and Druet, 2010). Imputed genotypes are subject to imputation errors which would likely affect the accuracy of genomic prediction (Mulder et al., 2012; Weigel et al., 2010). Different genomic prediction methods may react to imputation errors in a different way due to their different assumptions on the distribution of SNP effects. The purpose of this study was to evaluate the performance of GBLUP and a Bayesian mixture model using imputed genotypes for genomic prediction.

Materials and Methods

Genotypes

A total of 10,309 Holstein dairy bulls born between 1950 and 2007 were genotyped on the Illumina BovineSNP50 BeadChip (50k; Illumina Inc., San Diego, USA). SNPs with minor allele frequency (MAF) less than 0.05, missing rate more than
15% or P-value from Hardy-Weinberg disequilibrium test smaller than 0.0001 were removed. Only SNPs on autosomal chromosomes and with known locations on bovine genome assembly Btau4.2 were kept. After editing, 35,790 SNPs were included in this analysis.

Two commercially available evenly-spaced low density (ELD) SNP panels, the Illumina Golden Gate Bovine3K BeadChip (3k) containing 2,900 SNPs, and the BovineLD BeadChip (6k) comprising 6,909 SNPs, were considered in this study. 2,480 and 6,177 SNPs, from the 3k and 6k panel, respectively, were also included in the 35,790 SNPs from the 50k panel. Therefore, animals were not re-genotyped on the 3k and 6k panels. Instead, subsets of genotypes from the 50k panel were used to mimic genotypes from these two panels. In addition, three other ELD panels, named ELD384, ELD768 and ELD1536, with 384, 768 and 1,536 SNPs, respectively, derived from the 50k panel were simulated. These sizes were chosen according to Illumina Golden Gate technology, which allows customized genotyping with 384 to 1,536 SNPs. The markers for the three ELD panels were selected from the 50k panel using a formula compromised between uniform marker density and high MAF (Zhang and Druet, 2010). Distributions of SNPs over the chromosomes from various SNP panels are described in Table 1.1.

Phenotypes

Six traits, including milk yield, fat yield, protein yield, fat percentage, protein percentage, and somatic cell score (SCS), were used for this study. Official bull proofs and reliabilities in April 2008 and December 2011 were obtained from Canadian Dairy Network (CDN). De-regressed proofs and reliabilities derived from April 2008 bull proofs were provided by CDN. Bulls born before 2004 and with reliabilities of de-
regressed proofs greater than 0.80 in 2008 were used as the training data set (n=1,608). Bulls born in or after 2004 and with reliabilities of proofs larger than 0.80 in 2011 were used for validation (n=3,232). Bulls born before 2004, but with less reliable de-regressed proofs in 2008 were used as a reference population for genotype imputation (n=5,469). Animals in the training set were also included in the reference population for imputation as long as they have genotypes on the 50k panel in any scenarios described below. The numbers of animals used in this study are described according to birth years in Table 1.2.

**Genotype imputation**

Genotype imputation was conducted using the imputation software FImpute (version 2) developed by Sargolzaei et al. (2011a). The algorithm in the FImpute uses family imputation followed by population imputation based on a sliding window technique, which uses family-based reconstructed haplotypes (Sargolzaei et al., 2008).

**Genomic prediction**

A Bayesian mixture model, GBLUP and a Ridge-regression method were used to predict direct genomic values (DGV) for bulls in the validation set. The Ridge-regression method used in this study was equivalent to the GBLUP method (Goddard, 2009; VanRaden, 2008), and used to derive the SNP effects, which were then compared to those derived from the Bayesian mixture model.

**Bayesian mixture model**

The statistic model can be written as:

\[ y_i = \mu + \sum_{j=1}^{m} \beta_j x_{ij} + e_i \quad (i = 1, \ldots, n), \]

where \( y_i \) is the phenotypic value (de-regressed EBVs) for animal \( i \). \( \mu \) is the population mean. \( n \) is the total number of animals, and \( m \) is the total number of SNPs. \( x_{ij} \) is the
genotype of animal $i$ on SNP $j$. $x_{ij}$ was coded as 0, 1 or 2 representing the number of copies of a specified allele. $\beta_j$ is regression coefficient (allele substitution effect) for the $j^{th}$ SNP, $e_i$ is the random residual error.

The prior distribution of SNP effects follows a mixture of two normal distributions as $\left( \beta_j | \pi \right) \sim \left( 1 - \pi \right) N(0, \sigma^2_\beta) + \pi N(0, \tau \sigma^2_\beta)$, where $\tau$ is arbitrarily a very small value. $\pi$ was estimated from the data by assigning a uniform prior distribution with a support on (0,1). A scaled inverse Chi-square distribution with known degree of freedom $v_\beta$ and scale $S^2_\beta$ was assigned to $\sigma^2_\beta$, as in Habier et al. (2011). The strategy of using a group specific variance rather than a locus specific variance was to avoid the lack of Bayesian learning (Gianola et al., 2009; Habier et al., 2011; Meuwissen et al., 2001). The arbitrarily small value $\tau$ shrinks the SNP effects towards zero so that it can be effectively, but not completely removed from the model. The advantages of using such a spike and slab distribution have been discussed by Meuwissen (2009) and Calus (2010).

In short, large SNP effects are expected to be picked up by the model, and other SNP effects, each explaining a small amount of the variance, but together explaining a substantial part of the genetic variance, are also accounted for. $\mu$ follows a flat distribution, and $\left( e_i | \sigma^2_\epsilon \right) \sim N(0, \sigma^2_\epsilon)$, where $\sigma^2_\epsilon$ follows a scaled inverse Chi-square distribution with degree of freedom $v_\epsilon$ and scale $S^2_\epsilon$.

A Gibbs sampling algorithm was used to draw samples from the posterior distribution. Hyper-parameters need to be determined before running Gibbs sampling. $v_\beta$ and $v_\epsilon$ were arbitrarily set to 4 and 10, respectively. $\tau$ was chosen to be 0.0001. The scale parameter $S^2_\beta$ was derived from the expected value of a scaled inverse chi-square
distributed random variable, \( E(\sigma^2_\beta) = v_\beta S^2_\beta / (v_\beta - 2) \). So \( S^2_\beta \) was approximately determined as 
\[
S^2_\beta = \frac{(v_\beta - 2) \sigma^2_a}{v_\beta \left[ (1 - \tau) + \tau \sum_j 2p_j (1 - p_j) \right]},
\]
where \( p_j \) is the allele frequency of SNP \( j \) and \( \sigma^2_a \) is the additive genetic variance. SNP allele frequencies were estimated from the base population using Gengler’s method (Gengler et al., 2007) and a pedigree containing 73,027 animals. Similarly, \( S^2_e = (v_e - 2) \sigma^2_e / v_e \), where \( \sigma^2 \) is the environmental error variance. \( \sigma^2_a \) and \( \sigma^2 \) were estimated from preliminary analysis of de-regressed EBVs using a linear mixed model including a population mean, random animal effect and a random residual effect. Additive genetic variance and residual variance were estimated by ASReml 3.0 (Gilmour et al., 2009).

A self-developed computer program was written with ANSI C language to run the Gibbs sampling procedure. The Gibbs sampling was run for 100,000 iterations with the first 20,000 cycles discarded as burn-in. Burn-in period were determined by visually inspecting the Gibbs sampling chain. All samples were kept after the burn-in to estimate the means for the SNP effects from their posterior distribution. DGV for animals in the validation set were then calculated as the summation of the SNP effects over all loci as outlined by Meuwissen et al. (2001).

**GBLUP and Ridge-regression**

A best linear unbiased prediction method using a genomic relationship matrix formed from all SNP genotypes (GBLUP) was used to predict DGV for animals in the validation set. The genomic relationship matrix was obtained as in the first method described by VanRaden (2008). A Ridge-regression method equivalent to the GBLUP
was also used to derive the SNP effects (Goddard, 2009). GBLUP and Ridge-regression methods were implemented via the software GEBV (version 1.0), which was developed by Sargolzaei et al. (2011b).

**Scenarios**

Several scenarios were designed to mimic situations that different proportion of animals in the training and validation sets were genotyped on low density panels. In the base scenario (S0), all animals (n=10,309) were genotyped on the 50k SNP panel. In scenario 1 (S1), animals in the training and reference set (n=7,077) had genotypes from the 50k panel, and all animals in the validation set (n=3,232) had genotypes on low density panels. In scenario 2 (S2) and scenario 3 (S3), 33% and 66%, respectively, of the animals in training set were randomly selected to have their 50k genotypes replaced by low density genotypes. Same proportion of animals was of randomly selected from the reference set to have their 50k genotypes masked. Therefore, 531 and 1,061 animals for S2 and S3, respectively, in the training set had low density genotypes and were imputed to 50k genotypes. The reference population sizes used for imputation were n=7,077, n=4,741 and n=2,406 for S1, S2 and S3, respectively. Animals in the validation set remained the same as in S1 and were imputed to 50k genotypes. For genotype imputation, the reference animals that had 50k genotypes in the training set were joined into a larger reference set.

**Evaluation of accuracy for genotype imputation and genomic prediction**

Imputation of genotypes from various low density panels into 50k panel were carried out under different scenarios using FImpute. The imputed genotypes were compared to animals’ actual 50k genotypes. The percentage of SNP genotypes imputed
correctly out of total imputed SNP genotypes was calculated as a measure for imputation accuracy. Pearson’s correlation coefficient between DGV and bull proofs (December 2011) for animals in the validation set was calculated as a measure for the accuracy of genomic prediction.

**Results and Discussion**

**Accuracy of genotype imputation**

Table 1.3 shows the accuracy of genotypes that were imputed from various low density panels to the 50k SNP panel under different scenarios. The imputation accuracy was the highest (0.9841) when all animals in the training set were genotyped with 50k panel and animals in the validation set were genotyped on the BovineLD 6k panel. Accuracy was dropped rapidly when the density of the SNP panels was decreased. Imputation accuracy from 6k SNP panel were greater than that from the 3k panel by about 2, 3 and 4 percentage points, when 0%, 33% and 66% of animals in the training set, respectively, were also genotyped on the low density panel. The accuracy decreased as more training animals were genotyped on low density panels which, consequently, resulted in reduced reference set for imputation. The results on imputation accuracy were consistent with other studies (Boichard et al., 2012; Weigel et al., 2010; Zhang and Druet, 2010).

**Comparison of estimated SNP effects**

Figure 1.1 shows the size of SNP effects estimated by the Ridge-regression method against those by the Bayesian mixture model. The Bayesian mixture model tended to place more weight on a few large SNP effects and less weight on small SNP effects. This can be explained by different shrinkages between the two methods. The
ridge factor used in Ridge-regression shrinks SNP effects uniformly, while in the Bayesian mixture model, where a spike and slab prior was used, most of SNPs have small effects and are shrunk towards zero, leaving a few SNP have large effects.

To further show the difference between the two methods for areas where a large QTL exists, SNP effects for fat percentage on chromosome 14 were shown in Figure 1.2. Both GBLUP and the Bayesian mixture model identified the area harbouring the DGAT1 gene, which has a large effect on milk fat percentage in cattle (Grisart et al., 2004). The Bayesian mixture model tended to select fewer relevant SNPs, while the Ridge-regression picked many more SNPs surrounding the QTL.

**Accuracy of genomic prediction using observed 50k SNP genotypes**

Table 1.4 shows the accuracy of DGV in the validation set predicted via the GBLUP and Bayesian mixture model using the observed 50k SNP genotypes in the training and validation animals. The Bayesian method outperformed GBLUP by 3, 5, 1, 11, 5 and 0 hundredth points of accuracy, for milk yield, fat yield, protein yield, fat percentage, protein percentage and SCS, respectively, with the greatest difference for fat percentage and the least for SCS. Table 1.3 also presents the posterior estimates of π, the proportion of SNPs that were expected to have no effects on the trait.

When π is small, the trait is likely affected by many QTL with small effects, and when π is large, a few large QTL are expected to influence the trait (Habier et al., 2011). Daetwyler et al. (2010) concluded that Bayesian methods would perform similarly or slightly worse than GBLUP when the trait was affected by many QTL each with a small effect, and better if the trait was influenced by a few large QTL. The largest π was for fat percentage, for which the Bayesian mixture model showed the greatest advantage.
compared to GBLUP. Estimated $\pi$ for protein yield was 0.39, which was likely due to underestimation. And relatively smaller estimate of $\pi$ for SCS corresponded to no difference of accuracy between the Bayesian and GBLUP methods. These results agreed with other studies in terms of the performance between GBLUP and Bayesian methods on traits with different genetic architectures (Hayes et al., 2009b; VanRaden et al., 2009).

**Accuracy of genomic prediction using imputed 50k SNP genotypes**

Table 1.5 presents the accuracy of DGV for animals in the validation set predicted by the GBLUP and Bayesian mixture model, using genotypes imputed from various low density panels, under different scenarios. Accuracies achieved from the Bayesian mixture model were greater than or similar to those from GBLUP, for all low density panels, under all scenarios. Accuracy of genomic prediction decreased when the density in the SNP panel was reduced. For scenarios where genotype imputation subjected to more errors, accuracy of genomic prediction also declined more rapidly.

The trend that the accuracy changes with the density of SNP panel and with the proportion of training animals being genotyped on the low density panels agreed with results of Weigel et al. (2010). However, in their study, only up to a 3k low density panel was evaluated. Our study revealed that the 6k SNP panel performed better than the 3k panel and resulted in the least reduction of genomic prediction accuracy among all the low density panels, compared to accuracy from the 50k panel. In fact, no reduction was observed for 6k panel using GBLUP method under all scenarios, and only slight reductions (up to 1 percentage point) were observed for the Bayesian mixture model under S2 and S3.
The influence of imputation errors on the genomic prediction accuracy also depends on the trait. For example, when animals in the validation set were imputed from ELD384 SNP panel, accuracy of DGV predicted via GBLUP was reduced by a rate of 20, 24, 18, 34, 20 and 15%, for milk yield, fat yield, protein yield, fat percentage, protein percentage and SCS, respectively; and for the Bayesian mixture model, the accuracy was dropped by 22, 31, 18, 44, 25, and 15%, respectively. This was likely due to different genetic architectures underlying different traits. For traits affected by a few large QTL, such as fat percentage, accuracy of genomic prediction seemed much more sensitive to imputation errors than traits controlled by many small QTL, such as SCS. In a simulation study, it was also observed that when the density in the SNP panel was decreased, the accuracy of genomic prediction declined much more rapidly for traits with smaller number of QTL (Zhang et al., 2011). The difference in our study was that low density panels were also imputed to high densities, while in the study of (Zhang et al., 2011) low density panels were used without imputation.

Figure 1.3 compares the accuracy reduction rate between the Bayesian mixture model and GBLUP method, for different density of SNP panels, different traits and scenarios. When the density of the SNP panel was reduced, for milk yield, fat yield, fat percentage and protein percentage, greater reduction rates were observed for the Bayesian mixture model, compared with GBLUP; but for protein yield and SCS, the rate of accuracy reduction for the Bayesian mixture model was similar to or slightly less than that of GBLUP.

For chromosome regions with a large QTL, the Bayesian method tended to select the most relevant SNPs, while GBLUP or the Ridge-regression picked many more SNPs.
surrounding the QTL (Figure 1.2), and for other regions, the Bayesian method puts less weight on the SNPs than GBLUP (Figure 1.1 and Figure 1.2). So the Bayesian method could suffer more from imputation errors in large QTL regions, but the GBLUP method would suffer from imputation errors accumulated over many more SNPs. This could possibly explain why the Bayesian method had a greater reduction rate in accuracy for traits with large QTL and why the Bayesian method still resulted in higher accuracies than GBLUP for all scenarios. Relative performance of the two methods might be more related to distributions of imputation errors. If more imputation errors were distributed around the QTL, one could speculate that the Bayesian method would suffer more from these errors than GBLUP and consequently resulted in more reductions in the accuracy of genomic prediction.

Using either the GBLUP or Bayesian method, the 6k SNP panel performed better than the 3k panel. In this study, genotypes on both the 3k and 6k panel were simulated from the 50k genotypes. In reality, there are more genotyping errors in 3k genotypes than in 6k or 50k due to the Golden Gate genotyping technology used for the 3k panel (Boichard et al., 2012). So, results could be worse for the 3k panel in practice. The other lower density panels performed worse due to more inaccurate imputation of SNP genotypes. Accuracy for either genotype imputation or genomic prediction was evaluated as an average for the population. Examining imputation accuracies at an individual animal level showed that even for very low density SNP panels, there were still a substantial number of animals that achieved high imputation accuracies. These animals should have close relationships to the reference set which were used by FImpute. In results not shown above, when both parents of the target animals were genotyped on the
50k SNP panel, accuracy of imputation from ELD384 SNP panel to 50k was 0.9436; when only sires were genotyped on the 50k panel, the accuracy dropped to 0.7738; and when no parent was genotyped, the accuracy further reduced to 0.6213. Zhang and Druet (2010) also found that when the genomes of target animals were fully inherited from reference animals, imputation errors from a 384 SNP panel to 50k SNP panel can be as low as 3.2%. In this study, imputed genotypes were included in the training set without considering their imputation accuracies. In practice, only animals that have their genotypes imputed with high accuracies should be included in the training set. This scenario should be further studied to investigate the impact of using imputed genotypes on the accuracy of genomic prediction.

In the future, more and more animals might be genotyped on low density panels. One might have to decide whether to include these animals in the training population to derive genomic prediction equations. From Table 1.5, the accuracy of genomic prediction was consistently reduced when more animals in the training set were imputed, when the density of the SNP panel was lower than 6k. For 6k panel, accuracy of genomic prediction was almost not affected even when 66% of the training set was imputed, which would justify the application of this panel. This of course, was achieved by the high genotype imputation accuracy for the 6k panel. In this study, the sample size of the training set was kept constant regardless the use of imputation. The use of imputation could, however, also aid to enlarge the sample size of the training set, which might in fact increase the accuracy of the DGV. This possible scenario warrants further investigation.
Conclusions

The impact of imputation errors on GBLUP and the Bayesian mixture model depends on the genetic architecture of the traits. For traits affected by a few large QTL, the Bayesian mixture model resulted in greater reduction rate in accuracy of genomic prediction than GBLUP. However, for all SNP panels, scenarios and all traits studied, the Bayesian mixture model produced higher or similar accuracy compared to the GBLUP method. When different low density panels were compared, with all animals in the training set genotyped on 50k SNP panel, and animals in the validation set genotyped on the 6k SNP panel, the accuracy of genomic prediction was the greatest for both the GBLUP and Bayesian mixture model. Imputation from SNP panels with a density lower than 6k is more prone to errors and resulted in lower accuracy of genomic prediction. Including genotypes imputed from the 6k panel achieved almost the same accuracy of genomic prediction as that of the 50k panel, even when 66% of the training animals were genotyped on the low density panels.
Table 1.1 Summary of the number of SNPs on each chromosome in various SNP panels used in this study.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>BovineSNP50</th>
<th>BovineLD</th>
<th>Bovine3K</th>
<th>ELD1536</th>
<th>ELD768</th>
<th>ELD384</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTA1</td>
<td>2,291</td>
<td>369</td>
<td>160</td>
<td>97</td>
<td>49</td>
<td>24</td>
</tr>
<tr>
<td>BTA2</td>
<td>1,856</td>
<td>331</td>
<td>136</td>
<td>85</td>
<td>42</td>
<td>21</td>
</tr>
<tr>
<td>BTA3</td>
<td>1,792</td>
<td>284</td>
<td>120</td>
<td>77</td>
<td>38</td>
<td>19</td>
</tr>
<tr>
<td>BTA4</td>
<td>1,731</td>
<td>285</td>
<td>125</td>
<td>75</td>
<td>37</td>
<td>19</td>
</tr>
<tr>
<td>BTA5</td>
<td>1,504</td>
<td>283</td>
<td>117</td>
<td>76</td>
<td>38</td>
<td>19</td>
</tr>
<tr>
<td>BTA6</td>
<td>1,750</td>
<td>283</td>
<td>115</td>
<td>74</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>BTA7</td>
<td>1,510</td>
<td>260</td>
<td>108</td>
<td>68</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>BTA8</td>
<td>1,631</td>
<td>280</td>
<td>112</td>
<td>71</td>
<td>35</td>
<td>18</td>
</tr>
<tr>
<td>BTA9</td>
<td>1,427</td>
<td>255</td>
<td>108</td>
<td>65</td>
<td>33</td>
<td>16</td>
</tr>
<tr>
<td>BTA10</td>
<td>1,496</td>
<td>254</td>
<td>103</td>
<td>64</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>BTA11</td>
<td>1,590</td>
<td>268</td>
<td>104</td>
<td>67</td>
<td>33</td>
<td>17</td>
</tr>
<tr>
<td>BTA12</td>
<td>1,157</td>
<td>211</td>
<td>88</td>
<td>52</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>BTA13</td>
<td>1,236</td>
<td>206</td>
<td>86</td>
<td>51</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>BTA14</td>
<td>1,206</td>
<td>207</td>
<td>82</td>
<td>49</td>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>BTA15</td>
<td>1,200</td>
<td>202</td>
<td>83</td>
<td>51</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>BTA16</td>
<td>1,049</td>
<td>194</td>
<td>77</td>
<td>47</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>BTA17</td>
<td>1,137</td>
<td>185</td>
<td>72</td>
<td>46</td>
<td>23</td>
<td>12</td>
</tr>
<tr>
<td>BTA18</td>
<td>973</td>
<td>168</td>
<td>66</td>
<td>40</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>BTA19</td>
<td>977</td>
<td>163</td>
<td>57</td>
<td>39</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>BTA20</td>
<td>1,072</td>
<td>191</td>
<td>72</td>
<td>46</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>BTA21</td>
<td>947</td>
<td>174</td>
<td>73</td>
<td>42</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>BTA22</td>
<td>912</td>
<td>160</td>
<td>66</td>
<td>37</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td>BTA23</td>
<td>800</td>
<td>146</td>
<td>52</td>
<td>32</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>BTA24</td>
<td>895</td>
<td>165</td>
<td>63</td>
<td>39</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>BTA25</td>
<td>746</td>
<td>136</td>
<td>46</td>
<td>26</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>BTA26</td>
<td>752</td>
<td>139</td>
<td>46</td>
<td>31</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>BTA27</td>
<td>714</td>
<td>131</td>
<td>47</td>
<td>30</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>BTA28</td>
<td>700</td>
<td>119</td>
<td>43</td>
<td>28</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>BTA29</td>
<td>739</td>
<td>128</td>
<td>53</td>
<td>31</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>35,790</td>
<td>6,177</td>
<td>2,480</td>
<td>1,536</td>
<td>768</td>
<td>384</td>
</tr>
</tbody>
</table>
Table 1.2 Summary of the number of animals used in this study.

<table>
<thead>
<tr>
<th>Year of birth</th>
<th>Reference$^1$</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950-1954</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1955-1959</td>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1960-1964</td>
<td>17</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1965-1969</td>
<td>13</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1970-1974</td>
<td>15</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>1975-1979</td>
<td>25</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>1980-1984</td>
<td>76</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>1985-1989</td>
<td>437</td>
<td>86</td>
<td>0</td>
</tr>
<tr>
<td>1990-1994</td>
<td>382</td>
<td>140</td>
<td>0</td>
</tr>
<tr>
<td>1995-1999</td>
<td>2,557</td>
<td>843</td>
<td>0</td>
</tr>
<tr>
<td>2000-2003</td>
<td>1,936</td>
<td>475</td>
<td>0</td>
</tr>
<tr>
<td>2004-2007</td>
<td>0</td>
<td>0</td>
<td>3,232</td>
</tr>
<tr>
<td>Total</td>
<td>5,469</td>
<td>1,608</td>
<td>3,232</td>
</tr>
</tbody>
</table>

$^1$Reference animals were used for genotype imputation; animals with real genotypes on the 50k SNP panel in the training set were also included into the reference set for genotype imputation.

Table 1.3 Accuracy of imputed genotypes from low density panels to 50k SNP panel.

<table>
<thead>
<tr>
<th>Scenario$^1$</th>
<th>Low density SNP panel</th>
<th>BovineLD</th>
<th>Bovine3K</th>
<th>ELD1536</th>
<th>ELD768</th>
<th>ELD384</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td></td>
<td>0.9841</td>
<td>0.9604</td>
<td>0.9430</td>
<td>0.8787</td>
<td>0.7965</td>
</tr>
<tr>
<td>S2</td>
<td></td>
<td>0.9792</td>
<td>0.9507</td>
<td>0.9300</td>
<td>0.8573</td>
<td>0.7617</td>
</tr>
<tr>
<td>S3</td>
<td></td>
<td>0.9723</td>
<td>0.9367</td>
<td>0.9120</td>
<td>0.8285</td>
<td>0.7210</td>
</tr>
</tbody>
</table>

$^1$S1: Training set had known genotypes on 50k SNP panel; S2: 33% of the training set was genotyped on one of the low density panels; S3: 66% of the training set was genotyped on the low density panels. The reference population sizes used for imputation were n=7,077, n=4,741 and n=2,406 for scenario S1, S2 and S3, respectively.
Table 1.4 Accuracy of genomic prediction and posterior estimates of $\pi$ using observed
50k SNP genotypes under scenario S0$^1$.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Accuracy</th>
<th>Posterior $\pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GBLUP</td>
<td>Bayesian</td>
</tr>
<tr>
<td>Milk</td>
<td>0.61</td>
<td>0.64</td>
</tr>
<tr>
<td>Fat</td>
<td>0.54</td>
<td>0.59</td>
</tr>
<tr>
<td>Protein</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>Fat %</td>
<td>0.64</td>
<td>0.75</td>
</tr>
<tr>
<td>Protein %</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>SCS</td>
<td>0.62</td>
<td>0.62</td>
</tr>
</tbody>
</table>

$^1$S0: All animals in the training and validation set were genotyped on 50k SNP panel.
Table 1.5 Accuracy of genomic prediction using imputed 50k SNP genotypes:

Accuracies from using Bayesian model are parenthesized and accuracies for GBLUP are presented outside the parenthesis.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Scenario</th>
<th>BovineLD</th>
<th>Bovine3K</th>
<th>ELD1536</th>
<th>ELD768</th>
<th>ELD384</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk</td>
<td>S1</td>
<td>0.61 (0.64)</td>
<td>0.60 (0.63)</td>
<td>0.59 (0.61)</td>
<td>0.55 (0.57)</td>
<td>0.49 (0.50)</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.61 (0.64)</td>
<td>0.59 (0.62)</td>
<td>0.58 (0.60)</td>
<td>0.52 (0.54)</td>
<td>0.44 (0.46)</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.61 (0.64)</td>
<td>0.59 (0.61)</td>
<td>0.57 (0.58)</td>
<td>0.50 (0.52)</td>
<td>0.39 (0.42)</td>
</tr>
<tr>
<td>Fat</td>
<td>S1</td>
<td>0.54 (0.59)</td>
<td>0.53 (0.59)</td>
<td>0.52 (0.55)</td>
<td>0.48 (0.50)</td>
<td>0.41 (0.41)</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.54 (0.59)</td>
<td>0.53 (0.59)</td>
<td>0.51 (0.54)</td>
<td>0.46 (0.49)</td>
<td>0.36 (0.37)</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.54 (0.59)</td>
<td>0.53 (0.58)</td>
<td>0.49 (0.51)</td>
<td>0.43 (0.46)</td>
<td>0.31 (0.33)</td>
</tr>
<tr>
<td>Protein</td>
<td>S1</td>
<td>0.61 (0.62)</td>
<td>0.60 (0.61)</td>
<td>0.60 (0.61)</td>
<td>0.55 (0.56)</td>
<td>0.50 (0.51)</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.61 (0.61)</td>
<td>0.60 (0.60)</td>
<td>0.59 (0.59)</td>
<td>0.52 (0.54)</td>
<td>0.44 (0.45)</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.61 (0.61)</td>
<td>0.58 (0.59)</td>
<td>0.58 (0.58)</td>
<td>0.50 (0.51)</td>
<td>0.38 (0.41)</td>
</tr>
<tr>
<td>Fat %</td>
<td>S1</td>
<td>0.64 (0.75)</td>
<td>0.63 (0.73)</td>
<td>0.59 (0.63)</td>
<td>0.54 (0.57)</td>
<td>0.42 (0.42)</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.64 (0.75)</td>
<td>0.63 (0.73)</td>
<td>0.56 (0.61)</td>
<td>0.50 (0.53)</td>
<td>0.37 (0.39)</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.64 (0.74)</td>
<td>0.62 (0.72)</td>
<td>0.53 (0.59)</td>
<td>0.45 (0.51)</td>
<td>0.32 (0.34)</td>
</tr>
<tr>
<td>Protein %</td>
<td>S1</td>
<td>0.71 (0.76)</td>
<td>0.70 (0.74)</td>
<td>0.69 (0.72)</td>
<td>0.65 (0.66)</td>
<td>0.57 (0.57)</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.71 (0.75)</td>
<td>0.69 (0.73)</td>
<td>0.68 (0.70)</td>
<td>0.62 (0.63)</td>
<td>0.53 (0.54)</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.71 (0.75)</td>
<td>0.68 (0.72)</td>
<td>0.67 (0.69)</td>
<td>0.58 (0.59)</td>
<td>0.47 (0.47)</td>
</tr>
<tr>
<td>SCS</td>
<td>S1</td>
<td>0.62 (0.62)</td>
<td>0.62 (0.62)</td>
<td>0.61 (0.61)</td>
<td>0.57 (0.57)</td>
<td>0.53 (0.53)</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>0.62 (0.62)</td>
<td>0.61 (0.61)</td>
<td>0.60 (0.60)</td>
<td>0.54 (0.55)</td>
<td>0.49 (0.50)</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>0.62 (0.61)</td>
<td>0.60 (0.60)</td>
<td>0.60 (0.60)</td>
<td>0.53 (0.54)</td>
<td>0.44 (0.46)</td>
</tr>
</tbody>
</table>

1S1: Training set had known genotypes on 50k SNP panel; S2: 33% of the training set was genotyped on one of the low density panels; S3: 66% of the training set was genotyped on the low density panels.
Figure 1.1 Comparison of SNP effects estimated from Ridge-regression method and the Bayesian mixture model under scenario S0.
Figure 1.2 SNP effects for fat percentage estimated from Ridge-regression method and the Bayesian mixture model under scenario S0.
Figure 1.3 Comparison of genomic prediction accuracy reduction rate\(^1\) between the Bayesian mixture model and GBLUP method for different traits, low density panels and scenarios\(^2\).

\(^1\)Reduction rate was calculated by comparing genomic prediction accuracy obtained using low density panels to that when all animals in the training and validation set were genotyped on the 50k SNP panel.

\(^2\)The three scenarios were S1: All animals in the training set were genotyped on the 50k SNP panel; S2: 33\% of the training set were genotyped on the low density panels; S3: 66\% of the training set were genotyped on the low density panels.
Chapter 2 An Alternative Computing Algorithm for Genomic Prediction Using Bayesian Mixture Model and Large Training Data Sets

Abstract

In genomic prediction, Bayesian methods are commonly implemented via Markov Chain Monte Carlo (MCMC) sampling scheme, which is computationally demanding in large scale scenarios. Computing time with the traditional Gauss-Seidel residual update (GSRU) algorithm is proportional to the product of the number of animals and the number of effects. As the training data grows, alternative algorithms need to be developed for efficient implementation of genomic prediction. In Bayesian mixture models, the vector of marker effects is sparse which can be exploited to speed up the computing. An alternative computing algorithm, called right-hand side updating strategy (RHSU), was proposed by taking advantage of the sparsity to calculate the most computing expensive part in the MCMC based Gibbs sampling algorithm. The RHSU was compared with GSRU by the number of floating point operations (FLOP) required in one round of Gibbs sampling. The relative efficiency (RE) of RHSU compared to GSRU was defined based on FLOP and a threshold value for the ratio between the size of the training set and the number of markers was defined for RHSU to be more efficient than GSRU in different scenarios. The two algorithms were compared in a real data example with training data sizes ranging from 1,000 to 10,000 and a marker panel containing 35,790 single nucleotide polymorphisms (SNP). Results showed that the proposed RHSU algorithm would outperform the traditional GSRU algorithm once the training set size exceeded a fraction of the number of the SNPs, which typically varied from 0.05 to 0.18 when the proportion of SNPs with no effect on the trait varied from 0.90 to 0.95. As the
sample size continued to grow, the RHSU algorithm became more efficient than GSRU. Results from a real data example agreed very well with theoretical expectations. The RE and the threshold values calculated may offer reliable tools for making comparisons and choices between the two algorithms. With adoption of the 50k SNP panel and the increasing training data size, RHSU would be very useful if Bayesian methods are preferable for genomic prediction.

Introduction

Genomic selection proposed by Meuwissen et al. (2001) is revolutionizing breeding programs in livestock and other species. This approach uses a large number of genotypic markers, mainly single nucleotide polymorphisms (SNP), to predict breeding values for selection candidate. Numerous methods have been proposed for the prediction of direct genomic values (DGV). The most common methods are Genomic Best Linear Unbiased Prediction (GBLUP) (VanRaden, 2008) and various Bayesian mixture models (Calus et al., 2008; Habier et al., 2011; Meuwissen et al., 2001; Mutshinda and Sillanpaa, 2010; Park and Casella, 2008; VanRaden, 2008; Verbyla et al., 2009; Xu, 2003). Assumptions about marker effects in these methods are usually different. For example, GBLUP assumes that all marker effects follow a normal distribution, while Bayesian mixture models assume that a large proportion of markers (π) have zero or little effects, and the others follow normal distributions or a multivariate student-t distribution, depending on which prior distributions are used. The performance of different methods depends on the genetic architecture underlying the studied trait (Daetwyler et al., 2010). Both simulation studies and real data analyses in Holstein dairy cattle have found that Bayesian methods outperformed GBLUP when the trait is controlled by a few large QTL,
and performed similarly or slightly worse than GBLUP when the trait is affected by many quantitative trait loci (QTL) with small effects (Daetwyler et al., 2010; Hayes et al., 2009b; VanRaden et al., 2009).

Unlike GBLUP, which can be implemented efficiently (VanRaden, 2008), application of Bayesian methods requires evaluation of the posterior distribution which cannot be obtained directly. Two categories of methods are proposed in literature to solve the task. The first category includes approximation methods, such as maximum a posteriori or variational Bayes method. The second category includes stochastic sampling methods, such as Markov Chain Monte Carlo (MCMC). Approximation based methods are several magnitudes faster than MCMC but they do not guarantee to converge to the global maxima and thus may result in loss of prediction accuracy compared to their MCMC counterparts (Hayashi and Iwata, 2010; Li and Sillanpaa, 2012; Meuwissen et al., 2009). On the other hand, MCMC algorithm is computationally demanding, especially when the number of SNP effects and sample size are both large, as would be the case in genomic prediction.

The most time-costly step in the MCMC sampling algorithm is to compute the mean from the full conditional distributions for each SNP effect in each sampling cycle. Several computing strategies have been compared by Legarra and Misztal (2008) and an algorithm termed Gauss-Seidel with residual update (GSRU) was proposed. Computing time with GSRU algorithm is proportional to the product of the number of records and the number of effects in the model. The sample size can grow quickly with animals being routinely genotyped and added for genomic prediction. Therefore, alternative algorithms
need to be developed to efficiently implement genomic prediction under scenarios where the large training sample size is large.

In a Bayesian mixture model, only a fraction of SNPs are assumed to have non-zero effects and the proportion of SNPs that have zero effects in two consecutive MCMC samples is high. These properties can be explored to facilitate the calculation of the full conditional means for SNP effects, and therefore have the potential to speed up the MCMC sampling procedure. The objectives of this study were to develop such an alternative computing algorithm and to compare its performance with GSRU.

Materials and Methods

Bayesian mixture model

The Bayesian mixture model as described by Habier et al. (2011) was adopted for this study, as it bears the desired feature in that irrelevant SNP effects are shrunk to zero.

Consider a training population having \( n \) animals genotyped on \( m \) SNP loci, SNP genotypes were coded 0, 1 or 2, i.e., the number of a specified allele. For the sake of simplicity, all fixed effects including general mean were assumed ignored. The additive model, by regressing phenotypes on all SNP genotypes, can be written as:

\[
y_i = \sum_{j=1}^{m} x_{ij} a_j + e_i \quad (i = 1, \cdots, n),
\]

In matrix notation, it can be written as:

\[
y = Xa + e,
\]

Where \( y_i \) is the phenotypic value for the \( i^{th} \) animal; \( x_{ij} \) is the genotype on the \( j^{th} \) SNP locus of the \( i^{th} \) animal; \( a_j \) is the allele substitution effect for the \( j^{th} \) SNP; and \( e_i \) is the random residual error.
Prior distributions

\( a_j \) was assigned a mixture of the normal distribution \( N(0, \sigma_a^2) \) and the point mass density at zero \( \delta_0(\cdot) \). The weights assigned to these two distributions were \((1 - \pi)\) and \(\pi\), respectively, where \(\pi\) can be estimated from the data with a uniform prior distribution on the support \((0, 1)\). For the convenience of computation, an indicator variable \( \gamma_j \) was introduced for each SNP so that when \( \gamma_j = 1 \), \( a_j \) was from the normal distribution \( N(0, \sigma_a^2) \), and when \( \gamma_j = 0 \), \( a_j = 0 \). Prior distribution for each \( \gamma_j \) was assigned an identically and independently distributed Bernoulli distribution with probability \((1 - \pi)\).

The prior distribution for \( \sigma_a^2 \) was a scaled inverse Chi-square distribution with a degree of freedom \( v_a \) and a scale \( s_a^2 \).

Likelihood function

The likelihood of the data given all the parameters in the model is:

\[
\begin{align*}
    f(y|\sigma_a^2, \sigma_e^2, \pi, \gamma, \alpha) &\propto (\sigma_e^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sigma_e^{-2} (y - X\alpha)'(y - X\alpha)\right].
\end{align*}
\]

Joint posterior distribution

Combining prior distributions and likelihood function, the joint posterior distribution is:

\[
\begin{align*}
    f(\sigma_a^2, \sigma_e^2, \pi, \gamma, \alpha|y) &\propto \\
(\sigma_e^2)^{-\frac{v_e+n}{2}-1} \exp\left[-\frac{(y-X\alpha)'(y-X\alpha)+v_e\sigma_e^2}{2\sigma_e^2}\right] &\left(\sigma_a^2\right)^{-\frac{v_a}{2}-1} \exp\left(-\frac{v_a\sigma_a^2}{2\sigma_a^2}\right) &\prod_j \gamma_j \left(\sigma_a^2\right)^{-\frac{1}{2}} \exp\left(-\frac{a_j^2}{2\sigma_a^2}\right) + \\
(1 - \gamma_j)\delta_0(\cdot) &\pi^{(1-\gamma_j)} (1 - \pi)^{\gamma_j}.
\end{align*}
\]
Gibbs sampling algorithm

A Gibbs sampling algorithm was designed to draw samples for unknown parameters from their full conditional posterior distributions. Full conditional posterior distributions for $\sigma_a^2$, $\sigma_e^2$ and $\pi$ can be retrieved by picking up the relevant parts from the joint posterior distribution. To avoid reducibility of Markov chain, $\gamma_j$ and $a_j$ were jointly sampled by first sampling $\gamma_j$ from $f(\gamma_j|\theta_{j-}, y)$ followed by sampling $a_j$ from $f(a_j|\gamma_j, \theta_{j-}, y)$, where $\theta_{j-}$ represents all parameters except $\gamma_j$ and $a_j$. Derivation for the density function $f(\gamma_j|\theta_{j-}, y)$ and sampling of $\gamma_j$ are described in the Appendix A. The Gibbs sampling steps were described as below:

Step 1. Initialize the parameters $\sigma_a^2$, $\sigma_e^2$, $\pi$, $\gamma$ and $a$.

Step 2. For $j = 1, \ldots, m$

a. Sample $\gamma_j$ from Bernoulli distribution with probability $\frac{1}{1+q}$,

$$q = \frac{\pi}{1-\pi} \sqrt{x_j x_j \sigma_a^2} + 1 \exp \left( -\frac{\hat{\mu}^{l+1}_{a_j}}{2\hat{\sigma}^{l+1}_{a_j}} \right),$$

where $\hat{\mu}^{l+1}_{a_j} = \frac{x_j(y-x_{1j} - \hat{\theta}_{1j-1} x_{1j} - \hat{\theta}_{1j-1} x_{j+1} m \hat{\theta}_{j+1,m})}{x_j x_j + \sigma_e^2 / \sigma_a^2}$, $\hat{\sigma}^{l+1}_{a_j} = \frac{\sigma_e^2}{x_j x_j + \sigma_e^2 / \sigma_a^2}$, $l + 1$ denotes the current iteration, and $l$ denotes the previous iteration.

b. Sample $a_j$ from $f(a_j|\gamma_j, \theta_{j-}, y)$

$$f(a_j|\gamma_j, \theta_{j-}, y) = \begin{cases} \delta_0(\cdot) & \text{if } \gamma_j = 0 \\ N(\hat{\mu}^{l+1}_{a_j}, \hat{\sigma}^{l+1}_{a_j}) & \text{if } \gamma_j = 1 \end{cases}$$

Step 3. Sample $\pi$ from Beta distribution

$$f(\pi|y) = \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1},$$

where $\alpha = m - \sum \gamma_j$ and $\beta = \sum \gamma_j$. 

35
Step 4. Sample $\sigma^a_2$ from $\chi^{-2}(v_a + \Sigma y_j, \frac{v_a^2 \Sigma a_j^2}{v_a + \Sigma y_j})$

Step 5. Sample $\sigma^e_2$ from $\chi^{-2}(v_e + n, \frac{v_e^2 + (y - Xa)(y - Xa)}{v_e + n})$

Step 6. Repeat Step 2 to Step 5 until a set number of iterations has been completed.

**Computing strategies**

The most time costly step in the above Gibbs sampling procedure is to compute $\hat{\mu}^{l+1}_{a_j}$ for each SNP $j$. Two strategies are described below. One is the GSRU algorithm, which was proposed by Legarra and Misztal (2008) and modified to accommodate the Gibbs sampling context in this study; the other is the proposed right-hand side updating strategy (RHSU), which exploits the sparsity in the vector of SNP effects. The number of floating point operations (FLOP) required in one round of Gibbs sampling was estimated for each strategy. The relative efficiency of RHSU compared to GSRU was defined based on the FLOP. The turning point of the efficiency between RHSU and GSRU was also determined.

**GSRU computing strategy**

Following the convention used by Legarra and Misztal (2008), $y$ corrected for all effects, except the $j^{th}$ SNP effect, is equal to the current vector of residuals, $e^{l+1,j}$, plus the estimate of the $j^{th}$ effect from the previous iteration:

$$y - X_{j,j} \hat{a}^{l+1}_{j_{j-1}} - X_{j+m} \hat{a}^{l}_{j_{j-1+m}} = e^{l+1,j} + x_j \hat{a}_j^l.$$ 

Then $\hat{\mu}^{l+1}_{a_j}$ can be calculated as:

$$\hat{\mu}^{l+1}_{a_j} = \frac{x_j e^{l+1,j} + x_j x_j \hat{a}_j^l}{x_j x_j + \sigma_e^2 / \sigma_a^2}.$$
After $\hat{a}_j^{l+1}$ was sampled, $e$ needs to be updated as below:

$$e^{l+1,j+1} = e^{l+1,j} - x_j (\hat{a}_j^{l+1} - \hat{a}_j^l).$$

However, updating $e$ is not necessary if $\hat{a}_j^{l+1} - \hat{a}_j^l = 0$. This would occur if both $\gamma_j^l$ and $\gamma_j^{l+1}$ are equal to zero. The probabilities are $P(\gamma_j^l = 0) = \pi^l$ and $P(\gamma_j^{l+1} = 0) = \pi^{l+1}$.

Because $\pi$ usually has a large value, $\gamma_j^l$ and $\gamma_j^{l+1}$ have the great chance to be both zero. The probability $p_{00} = P(\gamma_j^l = 0, \gamma_j^{l+1} = 0)$ is some value between $\pi^2$ and $\pi$. Calculation for each $\hat{\mu}_{a_j}^{l+1}$ involves a vector multiplication $x_j e^{l+1,j}$, which requires $(2n-1)$ FLOP. $e$ needs to be updated $m(1 - p_{00})$ times, and each update requires $n$ multiplications and $n$ summations. $x_j x_j$ can be pre-computed and stored in memory. So the number of FLOP in one round of iteration is about $2nm(2 - p_{00})$.

**RHSU computing strategy**

The equation for $\hat{\mu}_{a_j}^{l+1}$ can be written as below:

$$\left(x_j x_j + \sigma_e^2 / \sigma_x^2 \right) \hat{\mu}_{a_j}^{l+1} = x_j y - x_j X_{t,j-1}^t a_{l+1,j-1} - x_j^t X_{n+1} j^t a_{l+1,j+1}^t.$$ 

The right hand side (RHS) elements are sparse because most of the SNP effects are zero due to the use of Bayesian mixture model. So if $X X$ is pre-computed and stored in memory, computing RHS should involve only operations on nonzero SNP effects. However, this strategy requires for every SNP to test all the other SNP effects against zero. These logical operations could be very time consuming. Instead, a right-hand side updating strategy (RHSU) was proposed and described below.
Before running Gibbs sampling, set the right-hand side of each SNP \( j \) to \( x'_j y \). Suppose one is sampling for SNP \( k \) in cycle \( l+1 \), if the sampled SNP effect is different from its value in the previous cycle, i.e., \( \hat{a}_{k}^{l+1} \neq \hat{a}_{k}^{l} \), then update the RHS for every SNP \( j \) (\( j \neq k \)) by adding the value \( x'_j x_k (\hat{a}_{k}^{l} - \hat{a}_{k}^{l+1}) \). \( x'_j x_k \) is pre-computed and updating each SNP \( j \) requires two operations, one for multiplication and the other for addition. Updating is not necessary if \( \hat{a}_{k}^{l+1} = \hat{a}_{k}^{l} \). In each cycle of the Gibbs sampling, the expected number of SNPs whose effects are different from their previous values is \( m(1 - p_{00}) \) resulting in a total number of \( 2m^2(1 - p_{00}) \) operations. Updating \( e \) is still required for the sampling of \( \sigma_{e}^2 \). \( e \) can be calculated from \( e^{l+1} = y - X\hat{a}^{l+1} \). So if only \( \hat{a}_{k}^{l+1} \neq 0 \), then update \( e^{l+1} \). For \( n \) animals and the proportion of SNPs with no effects being \( \pi \), updating \( e \) requires \( 2nm(1 - \pi) \) operations. The total number of FLOP in one round of iteration is therefore
\[
2m^2(1 - p_{00}) + 2nm(1 - \pi).
\]

**Comparison between the two computing strategies**

The FLOP of GSRU and RHSU were used to calculate the relative efficiency (RE) of the RHSU computing strategy.

\[
RE = \frac{\text{FLOP required for GSRU}}{\text{FLOP required for RHSU}} = \frac{n(2 - p_{00})}{m(1 - p_{00}) + n(1 - \pi)}.
\]

For RE to be greater than 1, the ratio between \( n \) and \( m \) must exceed a threshold:

\[
\text{Threshold} \left( \frac{n}{m} \right) = 1 - \frac{\pi}{\pi + 1 - p_{00}}.
\]
**Real data example**

The full test data set contains 10,000 Holstein dairy bulls with genotypes on the Illumina Bovine SNP50 BeadChip comprising 54,609 SNPs. After editing, 35,790 SNPs were kept for analysis. Bull proofs for fat percentage were extracted from the Canadian Dairy Network (CDN) website and used as phenotypes. Subsets of the data including 1000, 2000, 4000, 6000 and 8000 animals were randomly selected from the full data set.

The Bayesian mixture model was used for genomic prediction implementing the GSRU and RHSU computing strategies with different training sample sizes. Parameter $\pi$ was either estimated from the data or fixed at 0.95. Computer programs were written in ANSI C language and run on a single 3.47 GHz processor in Linux system. Pseudo codes for implementation of the GSRU and RHSU computing strategies are described in Appendix B. The Gibbs sampling procedures were run for 50,000 cycles, and for each run the CPU time was recorded.

**Results**

Figure 2.1 shows the relative efficiency of RHSU under two $\pi$ values, 0.9 and 0.95, assuming $p_{00} = \pi$. RHSU outperforms GSRU when $RE > 1$. The greater the ratio $n/m$ is, the larger the $RE$ would be. When the number of animals equals the number of SNPs, $RE$ is 10.5 for the $\pi$ value of 0.95, indicating that RHSU would be 10.5 times faster than GSRU.

Figure 2.2 shows the threshold value for the ratio $n/m$ for which RHSU would outperform GSRU. For each $\pi$ value, the threshold was confined by an upper and lower bound. When $\pi$ is 0.5, $n$ needs to be larger than 0.6$m$ for RHSU to outperform GSRU in the worst case; and in the best case, $n$ needs to be larger than 0.5$m$. When $\pi$ is larger, the
requirement for \( n/m \) is less. For example, when \( \pi \) becomes 0.95, \( n \) only needs to be at least 0.09\( m \) in the worst case and 0.05\( m \) in the best case.

Figure 2.3 shows the computing hours for the two algorithms when \( \pi \) was fixed at 0.95. When sample size was small (<3000), RHSU was slower than GSRU, and when sample size was 3000 and more, RHSU outperformed GSRU. Computing time for GSRU was approximately linear to the sample size. The computing time for RHSU only changed slightly with different sample sizes. It was also observed that when sample size increased from 1000 to 4000, the computing time with RHSU decreased slightly.

Figure 2.4 shows the results of computing time (CPU hours) for the two computing strategies, using real Holstein data with training sample size ranging from 1000 to 10000. Estimated \( \pi \) values from different sample sizes were also presented in Figure 2.4. RHSU outperformed GSRU in terms of computing time under all sample sizes. Computing time for both GSRU and RHSU increased with increasing sample size, but RHSU had a slower increasing rate. Estimated \( \pi \) decreased steadily with increasing sample size, from 0.9953 when sample size was 1000, to 0.9324 when sample size was 10000.

Discussion

Figure 2.1 and Figure 2.2 compared between GSRU and RHSU based on theoretically estimated number of FLOP. The emphases in these two comparisons are different. Figure 2.1 estimates the relative efficiency of RHSU compared to GSRU, while Figure 2.2 determines the thresholds for RHSU to outperform GSRU. These two comparisons are useful if one has the data at hand and wants to decide whether RHSU
would be more efficient than GSRU and how much faster it would be by using RHSU instead of GSRU.

It can be shown that results from the real data example agreed very well with theoretical expectations. When $\pi$ was fixed at 0.95, results showed that RHSU outperformed GSRU when the sample size exceeded 3,000 (Figure 2.3). From theory, the threshold value for $n/m$ was between 0.05 and 0.093 for RHSU to be more efficient (Figure 2.2). Therefore, theoretically, the sample size needed was between 1,790 and 3,332, given that 35,790 SNPs were used. For the speed, Figure 2.3 showed that RHSU was 3.15 times faster than GSRU when the sample size reached 10,000. From theory, it was expected that RHSU would be about 2.75 to 4.59 times faster than GSRU. Results in Figure 2.4 also agreed with theoretical expectations. The agreements between actual data and theoretical expectations implied that RE and the calculated thresholds may offer reliable tools for making comparisons and choices between the two algorithms.

In Figure 2.3, it was observed that when sample size increased from 1,000 to 4,000, the computing time with RHSU decreased slightly. This might be due to the fact that the MCMC chain reached its stationary distribution faster with a larger sample size and, hence, SNP effects that were estimated to be zero from one round of Gibbs cycle are more likely to stay in zero in the next cycle. The less switches from zero’s to non-zero’s, the faster the RHSU would perform. When sample size exceeded 4,000, the effect of sample size may start to outweigh the influence of SNP effect switching.

In Figure 2.4, the inconsistent estimates of $\pi$ with different sample size was likely due to two reasons: Larger sample size has more power to detect small QTL so that more SNPs were included in the model; The priors’ effect which shrinks small effects towards
zero vanished as sample size increases. Ishwaran and Rao (2005) proposed a re-scaled spike and slab model where the effects of priors do not vanish as sample size increases. Further investigation is needed to evaluate the influence of $\pi$ on prediction accuracy as well as alternative models such as the re-scaled spike and slab model for genomic prediction.

The proposed RHSU computing strategy is less dependent on the training sample size but is more relying on the number of markers and the proportion of the marker effects being zero ($\pi$ value) compared to the traditional GSRU algorithm. In practice, for most complex traits, $\pi$ usually has a relatively large value reflecting that most of the markers have no effects. In dairy cattle, genomic prediction has been implemented using the 50k panel. A higher density SNP panel (777k) resulted in very limited improvement on the accuracy of genomic prediction (Su et al., 2012). On the other hand, more accuracy was gained by increasing the training data size than by increasing the number of SNPs (VanRaden et al., 2009). With adoption of the 50k panel and the increasing training data size, RHSU would be very useful if Bayesian methods are needed for genomic prediction.

It has to be noted that time used for pre-computing $X'X$ was not taken into account when making the comparisons. When the sample size is large, computing $X'X$ could cost a substantial amount of time. Block-wise matrix multiplication and parallel computing strategy could be used to calculate $X'X$ efficiently. Furthermore, the matrix $X'X$ only needs to be calculated once if multiple traits are to be analyzed. Calculation of $X'X$ can be conducted with iteration on animals. So, once $X'X$ is computed on some
animals, new animals that need to be added into and animals that need to be removed from the training sample can be processed very quickly.

Memory requirement is another concern when choosing computing options. RHSU requires memory storage for \( X'X \). With 50,000 markers and single precision storage, this would require 10G bytes for full storage and 5G bytes for half storage. Both RHSU and GSRU require full storage for \( X \). With 50,000 markers and 10,000 animals, and, if each element of \( X \) is stored in one byte, this requires an additional 500M bytes.

**Conclusions**

The proposed RHSU algorithm would outperform the traditional GSRU algorithm once the sample size exceeds a fraction of the number of the SNPs, which typically varied from 0.05 to 0.18 when the proportion of SNPs with no effect on the trait varied from 0.90 to 0.95. As the sample size continues to grow, the RHSU algorithm becomes more efficient than GSRU. With adoption of the 50k panel and the increasing training data size, RHSU would be very useful in practice if Bayesian methods are preferable for genomic prediction.
Figure 2.1 Relative efficiency of the RHSU computing strategy for different ratios between number of animals in the training set and number of SNP markers and different proportions of SNPs ($\pi$) with no effect on the trait.
Figure 2.2 The threshold\(^1\) value for the ratio between number of animals in the training set \(n\) and the number of SNP markers \(m\) for RHSU to outperform GSRU for different proportions of SNPs \(\pi\) with no effect on the trait.

\(^1\)The threshold is runtime dependent and its range is defined by the upper and lower bound of the shaded area.
Figure 2.3 Comparison between two computing strategies keeping the proportion of SNP (π) with no effect on Fat % at 0.95, using real Holstein data.
Figure 2.4 Comparison between two computing strategies estimating the proportion of SNP ($\pi$) with no effect on Fat %, using real Holstein data.
Chapter 3 Accuracy of Predicting Genomic Breeding Values for Residual Feed Intake in Angus and Charolais Beef Cattle


Abstract

In beef cattle, phenotypic and genotype data are usually available in a relatively small number scattered in different populations. To implement genomic selection, it is often desirable to combine information from different sources together. Therefore, strategies of forming a training population for genomic selection need to be evaluated to achieve a maximal accuracy of genome prediction. In this study, we evaluated the accuracy of predicting direct genomic values (DGV) for residual feed intake (RFI) based on 522 Angus and 395 Charolais steers genotyped with the Illumina BovineSNP50 Beadchip (50k), using three strategies of genomic prediction, including within-breed, across-breed and by pooling data from the two breeds. Two other scenarios with training and validation data split by birth year (Scenario 1) or by sire family (Scenario 2) were also considered. The results showed that the accuracy of the DGV was the highest when prediction was within breed under scenario 1, with a maximum of 0.58 for Angus and 0.64 for Charolais. The within-breed prediction accuracies dropped to 0.29 and 0.38 for Angus and Charolais, respectively, under scenario 2. Across-breed genomic prediction resulted in the lowest accuracies under both scenarios (ranging from 0.10 to 0.22) possibly due to no pedigree link and low LD phase persistence between the two breeds. Pooling data from the two breeds resulted in slightly increased accuracy under scenario 2 (0.31 for Angus and 0.43 for Charolais), but same or lower accuracy under scenario 1 (0.58 for Angus and 0.55 for Charolais). The results suggested that the genetic
relationship between validation and training population has a great impact on the accuracy of DGV. Using the 50k SNP panel, within-breed genomic prediction is a safe strategy; across-breed resulted in the lowest accuracy due to low linkage disequilibrium phase persistence between the two breeds; pooling data from different breeds has a potential to improve the accuracy but must be used with caution due to possible loss of accuracy.

Introduction

Genomic selection, which was first proposed by Meuwissen et al. (2001), offers a new tool for genetic improvement of economically important traits in livestock species and plants. In beef cattle, the production system is characterized by diversified cattle populations of pure breeds and crossbreeds. Phenotypic data of economically important traits that are difficult or costly to measure, such as feed efficiency trait residual feed intake (RFI), are usually collected in a relatively small number of animals in different populations. Therefore, it is often desirable to combine information from different sources together to implement genomic selection. Strategies of forming a training population to predict the direct genomic values (DGV) for selection candidates need to be evaluated to achieve a maximal accuracy of genomic prediction.

Simulation studies have demonstrated that for a given marker density and sample size, training in the same breed as the validation population, i.e. within-breed prediction, results in the greatest accuracy of DGV, whereas across-breed genome prediction yields a much lower prediction accuracy, especially when the breeds are more divergent (de Roos et al., 2009; Kizilkaya et al., 2010; Toosi et al., 2010). Including data from the breed or population that the prediction candidates belong to in the training population would
possibly lead to increased prediction accuracy (de Roos et al., 2009; Ibanez-Escriche et al., 2009; Kizilkaya et al., 2010; Toosi et al., 2010). With real data, the accuracies of genomic prediction were evaluated in multi-breed dairy cattle populations of Holstein, Jersey and Fleckvieh for milk production traits (Hayes et al., 2009b; Pryce et al., 2011), and the results showed a similar trend as those obtained by the simulation studies.

In this study, different strategies of forming the training population to predict DGV for RFI were evaluated in Angus and Charolais beef cattle using genotypes on the Illumina BovineSNP50 Beadchip (50k) SNP panel. A pedigree based BLUP method (PBLUP), a genomic based BLUP method (GBLUP) and a Bayesian method (BayesB) were used and compared to predict breeding values for RFI.

**Materials and Methods**

**Animal and Management**

The purebred Angus and Charolais cattle used in this study include 572 steers from 42 sires and 424 steers from 44 sires, respectively. All animals were managed according to the guidelines established by the Canadian Council of Animal Care (CCAC, 1993). The Angus and Charolais cows were bred by AI and single sire mating around July and August of each year. The steers were born at the Onefour Research Substation of the Agriculture and Agri-Food Canada Research Centre at Lethbridge (AAFC) from 2004 to 2008 and at the University of Alberta’s Kinsella Research Station in 2009 from March to May. After weaning around the age of 6 months, steers were fed a background diet of 77 or 80% barley silage, 20 or 17% steam rolled barley and 3% supplement. Prior to the feedlot test, steers were fed a transition diet for a period of approximately three weeks, with a gradual decrease of barley silage ration and increase of steam rolled barley to
introduce the steers to the finishing diet and the feeding system described below. The steers were not treated with hormonal implants as growth promotants.

**Collection of Feed Intake data**

Feed intake was measured during the finishing period for each steer from 2005 to 2010 using the GrowSafe system (GrowSafe Systems Ltd., Airdrie, Alberta, Canada). The feeding system was housed in a shed with one long side open to provide access to 24 feeding bunks in 4 pens, which was located at the AAFC Lethbridge Research Centre. Angus and Charolais steers were randomly assigned to 4 pens of 6 feed bunks each with no separation of breed types. The finishing diet consisted of 20% (2005 to 2007) or 21% (2008 to 2010) barley silage, 76% (2005 to 2007) or 75% (2008 to 2010) steam rolled barley and 4% feedlot supplement and 32% crude protein (CP) beef mineral supplement containing 440 mg/kg of monensin, trace minerals and vitamins. Fresh feed samples were taken weekly during the feedlot test and were pooled monthly and analyzed for dry matter content. The diet samples of each month were thoroughly mixed and sub-sampled at a later stage, and analyzed for energy contents by Dairyland Laboratories, Inc. (Arcadia, WI, 54612, USA).

The measurements of feed intake were conducted for a period of about 120 days, and the GrowSafe system was monitored daily to ensure the proper function of the feedlot test system. Weight measurements of all animals were taken bi-weekly. For the days of disruption on the GrowSafe system, such as power outrage, obvious mechanical problems, animals’ measurement were treated as missing for all subsequent feed intake analyses.
Calculation of Residual Feed Intake (RFI)

A customized SAS procedure (SAS Institute, Inc., Cary, NC, version 9.1.3) was developed to extract the daily feed intake for each animal based on the data collected by the GrowSafe System. Assigned feed disappearances (AFD) audit was first examined for each feed bunk and for each day. AFD was screened as described by Mao et al. (2013). Days with an average AFD < 95% of a pen (i.e., >5% of feed missing) or a minimum of AFD < 90% for any individual bunk were excluded for feed intake analyses for all bunks of the pen. The individual animal’s daily feed intake pattern was also examined along with individual animal’s health record, and days with “poor” intake were identified. When an animal had more than two consecutive days of “poor” intake, the days were excluded from further feed intake analyses for the animal. The daily DMI was further standardized to 12 MJ ME per kilogram dry matter based on the energy content of the diet.

Calculations of components used to predict expected daily DMI and subsequently RFI measures for each animal were performed separately for Angus and Charolais for each test year. The initial weight and average daily gain (ADG) were derived from a linear regression of bi-weekly body weight measurements against time (d) in SAS. The mid-point body weight (BW) was computed as the sum of the initial weight and the product of ADG times half of the days on test (DOT). Metabolic body weight (MWT) was calculated as \( BW^{0.75} \). A linear regression model by regressing DMI on the ADG and MWT can be written as:

\[
y_i = \beta_0 + \beta_1 ADG_i + \beta_2 MWT_i + e_i,
\]
Where $y_i$ is the standardized daily DMI for the $i^{th}$ steer, $\beta_0$ is the intercept, $\beta_1$ is the partial linear regression coefficient on ADG, $\beta_2$ is the partial linear regression coefficient on MWT, and $e_i$ is residual error.

The residual $e_i$, which described the difference between the standardized daily DMI and the expected DMI, was then used as a measurement for RFI for the $i^{th}$ steer.

**Genotypes**

A total of 522 Angus and 395 Charolais steers having RFI measurements were genotyped on 54,609 single nucleotide polymorphisms (SNP) using the Illumina Bovine SNP50 Beadchip. SNPs with unknown locations on bovine UMD3.1 assembly or on sex chromosomes were excluded. SNPs were further screened within each breed for minor allele frequencies (MAF) less than 0.05, missing rate larger than 0.05, significant deviation from Hardy-Weinberg equilibrium test ($P<0.001$) or highly correlated with another SNP ($r>0.95$). SNPs removed in one breed were also excluded from the other breed. After editing, 33,321 SNPs were retained for genomic prediction analyses. Missing SNP genotypes represented 0.055% and 0.065% of total number of genotypes in the Angus and Charolais population, respectively. The missing genotypes were imputed using software Beagle 3.3.2 (Browning and Browning, 2007).

**Assignment of training and validation data sets**

Two scenarios were considered when splitting the data into the training and validation populations for each breed. Under scenario 1, steers were grouped according to their birth year. Steers born between 2004 and 2008 ($n=427$ for Angus and 351 for Charolais) formed the training population, and steers born in 2009 ($n=95$ for Angus and $n=44$ for Charolais) were included in the validation population. The distribution of the
number of animals in each year was shown in Table 3.1. Under scenario 2, steers were randomly selected into training and validation populations according to sire families, so that no sires overlapped between training and validation populations. The number of animals in the training and validation populations was purposely kept the same as in scenario 1 to remove the effect of different sample sizes on the accuracy of genomic prediction. By doing this way, results from the two scenarios would be comparable. This data-splitting procedure was repeated ten times and each data set was analysed separately. Data splitting under the scenario 2 was intended to mimic the situation where validation animals have lower relationship with training population.

Three strategies including within breed (training and validation in the same breed), across breeds (training in one breed, validation in the other breed), and pooling (pooling data from the two breeds together for training), were used for genomic prediction.

Adjustment for fixed effects and estimation of variance components for RFI

The linear mixed model can be written as:

\[ y = 1\mu + Xb + Za + e, \]

Where \( y \) is the vector of RFI; \( 1 \) is a vector with all elements set to 1; \( \mu \) is the overall mean; \( b \) is the vector of fixed effects, which include contemporary group and breed effect (when pooling data from the two breeds); \( a \) is the vector of animal effects; \( X \) and \( Z \) are the design matrices relating \( b \) and \( a \) to \( y \), respectively; \( e \) is the vector of random residual errors. The contemporary group was defined as the combinations of 6 test years and 4 pens per year within each breed. It was assumed that \( a \sim N(0, A\sigma_a^2) \) and
\( \mathbf{e} \sim \mathcal{N}(0, \mathbf{I} \sigma_e^2) \), where \( \mathbf{A} \) is the numerator relationship matrix formed from pedigree, \( \sigma_a^2 \) is the additive genetic variance, \( \mathbf{I} \) is an identity matrix, and \( \sigma_e^2 \) is the residual variance.

Additive genetic variance and residual variance components were estimated using ASREML (Gilmour et al., 2009). Heritability of RFI was estimated for each breed. Residual feed intake of each steer was adjusted for population mean and contemporary group effects within each breed using the outputs from ASREML. The adjusted RFI were used as phenotypic values for further analyses.

**Prediction of breeding values via PBLUP**

The pedigree based BLUP method (PBLUP) uses phenotypic values and pedigree information to estimate breeding values. It was supplies as a comparison to the GBLUP and BayesB methods, where phenotypic values and genotypes were used as sources of information. Animals in the validation population were assumed to have unknown phenotypes and were combined with animals in the training population to estimate the breeding values using numerator relationship matrix formed from the pedigree. The linear mixed model can be written in a matrix notation as:

\[
\mathbf{y}^* = \mathbf{1}\mu + \mathbf{Z}\mathbf{a} + \mathbf{e},
\]

Where \( \mathbf{y}^* \) is the vector of adjusted phenotypic value of RFI; \( \mathbf{1} \) is a vector with all elements set to 1; \( \mu \) is the overall mean; \( \mathbf{a} \) is the vector of breeding values; \( \mathbf{e} \) is the vector of random residual error; and \( \mathbf{Z} \) is the design matrix relating \( \mathbf{a} \) to \( \mathbf{y}^* \).

The mixed model equations were solved using ASREML. The previously estimated variance components were used and fixed in ASREML. The PBLUP method was used to estimate breeding values within each breed. It was not conducted across breed because there was no pedigree link between the two breeds.
**Prediction of breeding values via GBLUP**

The GBLUP method uses a genomic relationship matrix derived from the SNP markers instead of the pedigree based numerator relationship matrix. The first method described by VanRaden (2008) was used to construct the genomic relationship matrix with a modification to account for the difference between allele frequencies in the two breeds. Let \( X \) be the matrix including all genotypes from the two breeds. Without loss of generality, \( X \) is organised into two blocks \[
\begin{bmatrix}
X_1 \\
X_2
\end{bmatrix},
\]
where \( X_1 \) represents genotypes from Angus, and \( X_2 \) represents Charolais. Genotypes were coded as 0, 1 and 2 corresponding to SNP genotype aa, Aa and AA, respectively. Let \( p_i(p_j) \) be a vector of allele frequencies. The \( k \)th element \( p_{ik}(p_{jk}) \) in \( p_i(p_j) \) denotes allele frequency for SNP \( k \) in Angus (Charolais) population. Allele frequencies were estimated by simply counting the number of allele A on each locus within each breed. Then \( P = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \) is an allele frequency matrix corresponding to \( X \), with each row in \( P_1 \) (\( P_2 \)) being a replicate of the row vector \( p_1 \) (\( p_2 \)). Next, let \( W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{bmatrix} X_1 - 2P_1 \\ X_2 - 2P_2 \end{bmatrix}. \) The two-breed genomic relationship matrix \( G \) was constructed as follows:

\[
G = \begin{bmatrix}
\frac{W_1 W_1}{2 \sum p_{ik} (1 - p_{ik})} & \frac{W_1 W_2}{2 \sum \sqrt{p_{ik} (1 - p_{ik})} p_{2k} (1 - p_{2k})} \\
\frac{W_2 W_1}{2 \sum \sqrt{p_{ik} (1 - p_{ik})} p_{2k} (1 - p_{2k})} & \frac{W_2 W_2}{2 \sum p_{2k} (1 - p_{2k})}
\end{bmatrix}
\]

The \( G \) matrix was supplied to ASREML as a user defined relationship matrix. The previously estimated variance components were used and fixed when running
ASREML. The mixed model equations were solved by the preconditioned conjugate gradient (PCG) algorithm implemented in ASREML. The DGV of each individual in the validation population was predicted using training population from within breed, across breed or pooling data from the two breeds.

**Estimation of SNP effects via BayesB and prediction of the breeding values**

The BayesB method proposed by Meuwissen et al. (2001) was used to estimate the SNP effects. The statistical model can be written as follows:

\[ y_i^* = \mu + b_j + \sum_{k=1}^{m} x_{ik} \beta_k + e_i, \quad (i = 1, \cdots, n; j = 1,2) \]

Where \( y_i^* \) is the adjusted RFI for animal \( i \), \( \mu \) is the overall mean, \( b_j \) is the breed effect (Angus or Charolais when applicable). When the training population included only animals from one breed, \( b_j \) was removed; \( n \) is the total number of animals used in the training population, \( m \) equals 33,321, which is the total number of SNP effects fitted in the model; \( x_{ik} \) is the genotype of animal \( i \) for the \( k \)th SNP, which was coded the same way as described in GBLUP; \( \beta_k \) is the regression coefficient on the \( k \)th SNP and \( e_i \) is the residual effect. The priors for \( \mu \) and \( b_j \) were uniform distributions. \( e_i \) followed a normal distribution \( N(0, \sigma_e^2) \). \( \sigma_e^2 \) was assigned a scaled inverse chi-square distribution with degree of freedom \( v_e \) and a scale \( S_e^2 \). Each SNP effect \( \beta_k \) was assumed from a mixture of two distributions. The first distribution is a point mass concentrated at 0, and the second is a normal distribution \( N(0, \sigma_{\beta_k}^2) \), where \( \sigma_{\beta_k}^2 \) is a locus specific variance. The weights for the two distributions are \( \pi \) and \( 1- \pi \) correspondingly, where \( \pi \) is the proportion of
SNPs that have no effect on the trait. A scaled inverse chi-square distribution was used as the prior for $\sigma^2_\beta$ with degree of freedom $v_\beta$ and scale $S^2_\beta$.

The web-based software GenSel (Fernando and Garrick, 2008) was used to run Markov Chain Monte Carlo (MCMC) sampling algorithm for the BayesB model. Before running the BayesB procedure, a set of parameters need to be determined. At the beginning, one attempt was made to estimate $\pi$ through the method of BayesC$\pi$ as described by Habier et al. (2011). But it was found that $\pi$ could not be reliably estimated in this study (estimated $\pi$ varied from 0 to 0.92 with different sampled training populations). For this reason, $\pi$ was arbitrarily set to 0.95 under all scenarios. The degree of freedom $v_e$ ($v_\beta$) was arbitrarily set to 4 (10). The scale parameter $S^2_e$ was set to $\frac{(v_e - 2)\hat{\sigma}^2_e}{v_e}$, and the scale parameter for the locus specific variance was determined as $S^2_\beta = \frac{(v_\beta - 2)\hat{\sigma}^2_a}{v_\beta \pi m \sum 2p_k (1 - p_k)}$, where $\hat{\sigma}^2_e$ and $\hat{\sigma}^2_a$ were variance components previously estimated by ASREML, and $p_k$ is the allele frequency of SNP $k$. The MCMC chain was run for 100,000 iterations with a burn-in of 20,000 iterations. SNP effects were estimated by averaging all the samples after the burn-in period. The DGV for animal $i$ in the validation population was predicted by summing up SNP effects over all loci as follows:

$$DGV_i = \sum_{k=1}^m x_k \hat{\beta}_k,$$

where $\hat{\beta}_k$ is the estimated effect of SNP $k$.

**Evaluation of different methods**

The realised accuracy of DGV was used to evaluate the performance of different methods. It was calculated as the correlation between DGV and adjusted phenotypic values of RFI divided by square root of heritability, as described by Hayes et al. (2010).
Estimation of genetic relationship between training and validation populations

Numerator relationship matrix A was calculated within each breed from the pedigrees. All pair-wise relationships between animal i in the training population and animal j in the validation population (denoted as A_{ij} hereinafter) were used. The distribution of A_{ij} represents the genetic relationship between the training and validation populations.

Estimation of linkage disequilibrium phase persistence

Linkage disequilibrium (LD) and LD phase were estimated using the LDMAX program in the GOLD software package (Abecasis and Cookson, 2000). LD was measured as \( r^2 \), where \( r^2 = \left( p_{AB} p_{ab} - p_{Ab} p_{aB} \right)^2 / p_A p_a p_B p_b \), and LD phase was defined as \( r = \left( p_{AB} p_{ab} - p_{Ab} p_{aB} \right) / \sqrt{p_A p_a p_B p_b} \), where \( p_{AB}, p_{ab}, p_{Ab} \) and \( p_{aB} \) are the haplotype frequencies, and \( p_A, p_a, p_B \) and \( p_b \) are the allele frequencies. The LDMAX program was modified to output both \( r^2 \) and \( r \). LD and LD phase between all pairs of adjacent SNP markers were estimated in the training and validation populations under all the scenarios.

LD phase persistence is the correlation of the LD phase, for a certain marker distance, between two populations. In this study, only LD phases between adjacent SNPs were estimated. LD phase persistence between the training and validation populations was calculated as the Pearson’s correlation coefficient of the LD phase using estimates from all pairs of adjacent SNP markers.

Results

Genetic relationships between the training and validation populations

Distributions of additive genetic relationships (A_{ij}) between various training and validation populations within each breed are shown in Figure 3.1. In the Angus
population, a lower proportion (0.29%) of $A_{ij}$ was distributed in the area of $A_{ij} \geq 0.25$ under scenario 2, compared to 0.95% under scenario 1, suggesting a weaker genetic relationship between the training and validation populations under scenario 2. This is an expected result because scenario 2 was purposely designed to have no overlap of sire families between the training and validation populations. The small proportion of $A_{ij}$ distributed over the area of $A_{ij} \geq 0.25$ under scenario 2 was due to maternal half-sibs and in some rare cases, full-sibs. A similar trend was observed in Charolais, where a greater proportion (0.53%) was observed for $A_{ij} \geq 0.25$ under scenario 1, in comparison to 0.29% under scenario 2. The genetic relationships between pairs of training and validation populations of different breeds, e.g. Angus vs. Charolais or vice versa, were not considered due to no pedigree link between the Angus and Charolais populations.

**LD phase persistence between the training and validation populations**

The gap between adjacent SNPs of the used 33,321 SNPs has an average distance of 80Kb (SD=70Kb). Linkage disequilibrium (measured as $r^2$) between two adjacent SNPs were 0.23 (SD=0.25) and 0.16 (SD=0.20) for the Angus and Charolais population, respectively. The LD phase persistence between various training and validation populations are presented in Figure 3.2. Under scenario 1, the estimated LD phase persistence between Angus validation population and training populations formed within breed, across breed, and pooled data from the two breeds were 0.96, 0.70 and 0.92, respectively, and for Charolais validation population, they were 0.91, 0.66 and 0.81, respectively. The estimated LD phase persistence between the training and validation populations were very similar under scenario 2, with averages over the ten replicates being 0.96, 0.70 and 0.92 between Angus validation population and the training
population formed within breed, across breed and pooled data from the two breeds, respectively. For Charolais validation population, the same features were 0.91, 0.67 and 0.82, respectively.

Accuracy of genomic prediction

The realized accuracies of DGV for RFI are presented in Table 3.2. Under scenario 1, the accuracies of DGV were 0.54, 0.58 and 0.53 using the PBLUP, GBLUP, and BayesB methods, respectively, for the strategy of within-breed prediction, for Angus validation population, and 0.38, 0.62 and 0.64 for Charolais. For the across-breed genomic prediction strategy, the prediction accuracies were much lower, ranging from 0.10 to 0.23. With the pooled training data set used to predict the EBVs for either Angus or Charolais, a range of realized accuracies between 0.49 and 0.58 were achieved, depending on the breed to predict and method used.

Under scenario 2, the realized accuracies were lower in comparison to those obtained under scenario 1. For the within breed prediction, the average accuracies for Angus and Charolais were 0.30 and 0.31, respectively, for PBLUP, 0.29 and 0.38 for GBLUP, and 0.27 and 0.37 for BayesB. For the across breed prediction, the average accuracies were 0.12, and 0.11 for GBLUP and BayesB, respectively, when the training data of Angus were used to predict the EBVs for Charolais. When the training data of Charolais were used to predict the EBVs of Angus, the realized accuracies were 0.18 and 0.11 on average for GBLUP and BayesB, respectively. When pooling data from the two breeds as the training population, the prediction accuracies of DGV averaged 0.31 for Angus for the GBLUP and 0.28 for BayesB. For Charolais, a prediction accuracy of 0.43 was achieved for GBLUP and BayesB.
Discussion

The accuracy of genomic prediction depends on a number of factors that include the number of individuals in the training population, marker density, the heritability of the trait, the number of loci affecting the trait, the extent of linkage disequilibrium (LD) between markers and QTL, and the LD phase persistence between the training population and selection candidate. (Daetwyler et al., 2008; de Roos et al., 2009; Goddard, 2009; Meuwissen et al., 2001). In this study, we investigated the accuracy of genomic prediction for different training population forming strategies using data of two purebred beef cattle populations. The results of this study demonstrated the same trend as those reported in the simulation studies (de Roos et al., 2009; Ibanez-Escrive et al., 2009; Kizilkaya et al., 2010; Toosi et al., 2010) as well in the dairy studies with real data (Hayes et al., 2009b; Pryce et al., 2011), and indicated that the genomic prediction of breeding values for RFI within a breed or with a combined training population gave the greatest accuracy, whereas low accuracies were observed when a reference of one breed, Angus or Charolais in this study, was used to predict the DGV for selection candidates of the other breed. The results are supported by the genetic relationship and the persistence of LD phases between the training population and the validation population. The prediction accuracies of DGV for RFI were increased slightly under scenario 2 when the reference populations of both the breeds were combined to form the training population. This illustrates the increased importance of the training population size when there is less close genetic links between the training population and the selection candidates.

The lowest accuracy of across-breed genomic prediction observed in this study is likely due to two factors: genetic relationship and LD phase persistence between the
training and validation populations. The Angus and Charolais steers shared no pedigree link and therefore their pedigree genetic relationship was zero. The correlation coefficient of LD phase between the Angus and Charolais or between the validation population and training population of a different breed was between 0.67 and 0.70, which is much lower than the LD phase persistence when a training population of the same breed or combined breeds was used.

The results observed in this study suggest that the genetic relationship between the selection candidates and the training population plays a very important role in determining the accuracy of genomic prediction, which was demonstrated by the differences of prediction accuracy under the two data-splitting scenarios. In comparison to scenario 1, the accuracies of genomic prediction were generally lower under scenario 2, while LD phase persistence was not affected by different data-splitting scenarios. Under scenario 1, animals in the validation population had a larger proportion of $A_{ij} \geq 0.25$ than under scenario 2, where no sires were shared between validation and training population. Habier et al. (2010) decomposed the accuracies of genomic prediction into two parts, due to LD and due to relationship, and observed that accuracies of genomic prediction decreased with decay of LD as well as decay of relationship between training and validation animals (Habier et al., 2007; Habier et al., 2010), which is corroborated by the results of this study.

The Angus and Charolais populations shared no pedigree link. The accuracy of the across-breed genome prediction will therefore rely on the LD phase persistence between the two breeds. In this study, the estimated value of LD phase persistence for the Angus and Charolais populations were very similar to those reported by McKay et al.
(2007), at an average distance of approximate 80Kb between adjacent SNP markers. de Roos et al. (2008) evaluated the LD phase persistence between different dairy breeds, and found a similar level of LD phase persistence between breeds to the ones found in this study, at the similar distances. In this study, the lowest accuracies of DGV were achieved for the across-breed prediction, while the LD phase persistence between the Angus and Charolais populations was about 0.70. It is expected that a higher density of markers, which would give greater persistence of the LD phases between the Angus and Charolais, will likely lead to an increase in prediction accuracy for the across breed prediction.

Combining the training data from the Angus and Charolais increased training sample size, the LD phase persistence and the genetic relationship between the training population and validation population in comparison to across-breed prediction. In this study, when combining training data across the two breeds, the correlation of LD phases increased from 0.70 to 0.92 for Angus and from 0.66 to 0.81 for the Charolais. The increase in sample size, LD phase persistence and genetic relationship contribute to the increase of prediction accuracy. However, pooling training data from the two breeds resulted in larger training sample size, but lower LD phase persistence compared to within-breed prediction. Hayes et al. (2009b) and Pryce et al. (2011) reported up to a 13% increase of accuracy, depending on the traits, in comparison to the accuracies of the within-breed predictions. In this study, combining data of the two breeds to form a training population resulted in slightly higher accuracy on average in comparison to the within-breed prediction under scenario 2. However, under scenario 1, the accuracy did not improve in Angus and a reduction in accuracy was observed in Charolais. This could
possibly be due to sampling error because the sample size in the Charolais validation population was small and no replicates were done in scenario 1.

PBLUP was a traditional approach to estimate breeding values of selection candidates. Under the data splitting scenario 1, the accuracy of EBV for RFI by PBLUP was 0.54 and 0.38, for Angus and Charolais, respectively. The accuracy of PBLUP was reduced to an average of 0.30 and 0.31, for Angus and Charolais, respectively, under scenario 2, due to the decrease of genetic relationship. However, PBLUP relies only on the pedigree link between the selection candidates and the training population. Therefore, it is impossible to estimate breeding values for selection candidates if there is no genetic link between the selection candidates and the training population such as the across-breed prediction of the Angus and Charolais populations in this study.

In general, GBLUP outperformed the PBLUP method because it tracked the relationships among animals more accurately. This is not only due to possible errors in the pedigree, but also the ability of the genomic relationship matrix to capture the Mendelian sampling item. The advantage of GBLUP over PBLUP was more evident in Charolais than in Angus, which was likely due to the higher heritability observed in Charolais (0.68 in Charolais vs 0.47 in Angus). A larger additive genetic variance component (0.18) and smaller environment variance component (0.09) were estimated for Charolais, in comparison to 0.15 and 0.16 for Angus. Therefore, the RFI phenotypes used for training and validation were more reliable in Charolais than in Angus. GBLUP allowed the prediction of breeding values when there was no pedigree link between the training and validation population. This was achieved by using DNA markers which could possibly capture ancient genetic relationships. Pedigree, however, only captures
recent genetic relationships. However, low accuracies were observed for the across-breed prediction in this study, likely due to the low LD phase persistence between the Angus and Charolais breeds with the current 50k SNP panel.

In this study, BayesB gave similar or slightly lower accuracies than GBLUP method. It has been shown in the literature that Bayesian and GBLUP methods had similar accuracies for most traits in Dairy cattle (Hayes et al., 2009b; VanRaden et al., 2009). Simulation studies showed that the performance of BayesB and GBLUP methods depend on the genetic architecture of the trait (Clark et al., 2011; Daetwyler et al., 2010), and when there are few large QTL affecting the trait, BayesB outperformed GBLUP, and when the number of QTL became large, GBLUP gave similar or slightly higher accuracies than BayeB (Clark et al., 2011; Daetwyler et al., 2010). Our results implied that RFI is likely being controlled by a large number of QTL each with small effects, and GBLUP method seems to be a suitable method for genomic evaluation for RFI.

Conclusions

The results in this study suggest that the genetic relationship between validation and training population has a great impact on the accuracy of genomic prediction. Using the 50k SNP panel, within-breed genomic prediction is a safe strategy. Across-breed resulted in the lowest accuracy and therefore is not recommended with the 50k SNP panel. Pooling data from different breeds has a potential to improve the accuracy but must be used with caution due to possible loss of accuracy. GBLUP seems to be an optimum method to implement genomic prediction for RFI in beef cattle. Overall, prediction accuracies of DGV for RFI obtained in the study were still low, regardless of the training population forming strategies and statistical methods used, indicating the
need to include more animals in the training population and to increase the marker density.
Table 3.1 Distribution of the number of animals\(^1\) in each year.

<table>
<thead>
<tr>
<th>Birth year</th>
<th>Angus</th>
<th>Charolais</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>86</td>
<td>85</td>
</tr>
<tr>
<td>2005</td>
<td>95</td>
<td>92</td>
</tr>
<tr>
<td>2006</td>
<td>102</td>
<td>64</td>
</tr>
<tr>
<td>2007</td>
<td>78</td>
<td>70</td>
</tr>
<tr>
<td>2008</td>
<td>66</td>
<td>40</td>
</tr>
<tr>
<td>2009</td>
<td>95</td>
<td>44</td>
</tr>
<tr>
<td>Total</td>
<td>522</td>
<td>395</td>
</tr>
</tbody>
</table>

\(^1\)The animals were those have successful measurements on RFI and genotypes.

Table 3.2 Realized accuracy of DGV predicted via PBLUP, GBLUP, and BayesB methods.

<table>
<thead>
<tr>
<th>Scenario(^1)</th>
<th>Training</th>
<th>Validation</th>
<th>PBLUP</th>
<th>GBLUP</th>
<th>BayesB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>Angus</td>
<td>Angus</td>
<td>0.54</td>
<td>0.58</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Charolais</td>
<td>Angus</td>
<td>-</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Pooling</td>
<td>Angus</td>
<td>-</td>
<td>0.58</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Charolais</td>
<td>Charolais</td>
<td>0.38</td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Angus</td>
<td>Charolais</td>
<td>-</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Pooling</td>
<td>Charolais</td>
<td>-</td>
<td>0.55</td>
<td>0.49</td>
</tr>
<tr>
<td>Scenario 2(^2)</td>
<td>Angus</td>
<td>Angus</td>
<td>0.30 ± 0.06</td>
<td>0.29 ± 0.06</td>
<td>0.27 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>Charolais</td>
<td>Angus</td>
<td>-</td>
<td>0.18 ± 0.02</td>
<td>0.11 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>Pooling</td>
<td>Angus</td>
<td>-</td>
<td>0.31 ± 0.06</td>
<td>0.28 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>Charolais</td>
<td>Charolais</td>
<td>0.31 ± 0.08</td>
<td>0.38 ± 0.10</td>
<td>0.37 ± 0.10</td>
</tr>
<tr>
<td></td>
<td>Angus</td>
<td>Charolais</td>
<td>-</td>
<td>0.12 ± 0.02</td>
<td>0.11 ± 0.02</td>
</tr>
<tr>
<td></td>
<td>Pooling</td>
<td>Charolais</td>
<td>-</td>
<td>0.43 ± 0.09</td>
<td>0.43 ± 0.08</td>
</tr>
</tbody>
</table>

\(^1\)Scenario 1: Training and validation populations formed base on birth year; Scenario 2:

Training and validation populations formed preventing sires to overlap between populations;

\(^2\)Each training and validation was replicated 10 times; Accuracy was presented as average ± standard error.
Figure 3.1 Distributions of additive genetic relationships (estimated from pedigree) between training and validation animals.
Figure 3.2 LD phase persistence between training and validation populations.
Chapter 4 Multi-Population Genomic Prediction Using a Multi-Task Bayesian Learning Model

Submitted for publication in PLOS ONE (March, 2013).

Abstract

The goal of this study was to develop a multi-task learning model for multi-population genomic prediction. A multi-task Bayesian learning model was proposed for different populations to share information through a common set of latent indicator variables. The multi-task model was evaluated in real data and compared to baseline approaches of a single-task model and a simple data pooling method. The data are from Holstein and Ayrshire dairy populations with phenotypes on five milk production traits, milk yield, fat yield, protein yield, fat percentage and protein percentage, and genotypes on two different densities of single nucleotide polymorphisms (SNP) panels, the Illumina BovineSNP50 (50k) and BovineHD (777k) BeadChips. Genotypic data were edited so that 28,206 and 246,668 SNPs from the two panels were kept for analyses. 2,084 Holstein and 393 Ayshire bulls born before 2004 were used to derive genomic prediction equations. A set of 214 animals from Holstein and 65 from Ayrshire born in or after 2004 were used for validation. The single-task Bayesian learning, the simple data pooling method and the multi-task Bayesian learning model were implemented in a self-developed computer program. Results showed that, when 28,206 SNPs were used, the multi-task model produced a gain of accuracy from 0 to 7 percentage points (corresponding to 0 – 14.5%) in the Ayrshire validation set, compared to the single-task model, while the simple data pooling method resulted in a reduction of accuracy ranging from 8 to 10 percentage points (corresponding to 6.0 – 15.4%) except for protein
percentage, for which there was an increase by 3 percentage points (corresponding to 6.3%). When 246,668 SNPs were used, the accuracy achieved from the multi-task model increased from 0 to 3 percentage points (corresponding to 0 – 4.5%), compared to the single-task model, while using the pooling method resulted in a reduction in accuracy from 1 to 9 percentage points (corresponding to 1.5 – 17.6%). The accuracy in the Holstein validation population was similar using the three methods. In conclusion, the proposed multi-task Bayesian learning model is effective and could be beneficial to populations with small number of animals.

Introduction

Genomic prediction has become a new tool for selection of candidates based on genomic estimated breeding values (GEBV) through the use of dense markers covering the whole genome (Meuwissen et al., 2001). To predict GEBV, a training data set with genotypes and phenotypes is used to derive the prediction equations, where all marker effects are estimated simultaneously. GEBV for selection candidates that have genotypes are then predicted by summing up all the marker effects. The accuracy of GEBV is affected by several factors, of which the number of individuals in the training data set and the marker density are of crucial importance (Goddard, 2009; Hayes and Goddard, 2008).

In Holstein dairy cattle, genomic prediction has been successfully applied using the Illumina BovineSNP50 single nucleotide polymorphism (SNP) panel (Hayes et al., 2009b; VanRaden et al., 2009). For smaller populations such as Ayrshire in dairy cattle, acquisition a large number of animals to be included in the training data set for genomic prediction still remains a challenge. One strategy is to combine data of the small population with data of other populations to increase the size of the training set. However,
simply pooling data from different populations may result in unfavorable accuracies if the marker density is low or the populations have diverged for a long time (de Roos et al., 2009; Hayes et al., 2009a; Pryce et al., 2011). Increasing the marker density is one of the possible solutions because the linkage disequilibrium (LD) phase persistence between markers and quantitative trait loci (QTL) among different populations would likely to improve. However, a recent study in Jersey and Holstein dairy cattle reported a very limited advantage by using a very high density SNP panel (Erbe et al., 2012). Few studies have been dedicated to research new methods or strategies other than simply pooling data together for genomic prediction. Brondum et al. (2012) proposed an approach called BayesRS for multi-population genomic prediction, where location specific genetic variances derived in one population were used as priors for another population. They concluded that for some traits, BayesRS might be advantageous compared to pooling training data sets for distantly-related populations; but for closely related populations the method did not perform better than simply pooling data together.

Multi-task learning, the term first coined by Caruana (1997), aims to improve learning performance by simultaneously learning from related tasks. In speech recognition, image vision and many other areas where data are collected from multiple sources, multi-task learning has been successfully applied (Li and Bilmes, 2006; Lu et al., 2008; Yuan and Yan, 2010). Recently, the multi-task learning has attracted a growing interest in biological science for sequence analysis (Jacob and Vert, 2008; Widmer et al., 2010), gene expression analysis (Yang et al., 2011), genome-wide association study (GWAS) (Puniyani et al., 2010) and other purposes (see (Xu and Yang, 2011) for a
detailed review). To our knowledge, multi-task learning has not been introduced to genomic prediction yet.

Bayesian learning models using Bayesian stochastic search variable selection (BSSVS) has been widely used and proven effective for genomic prediction in a single population (Calus et al., 2008; Habier et al., 2011; Meuwissen et al., 2001; Verbyla et al., 2009). BSSVS uses some types of spike and slab distributions as priors for SNP effects. A latent indicator variable (0 or 1) is associated with each SNP, with 0 indicates the SNP is irrelevant and excluded from the model, and 1 indicates the SNP is related to the trait phenotype. In this study, it was assumed that multiple populations share the same set of latent indicator variables and can be learned simultaneously. The goal was to develop a multi-task Bayesian learning model for multi-population genomic prediction.

Methods

In this section, a Bayesian learning model using a spike and slab mixture distribution as prior for the SNP effects was firstly described for genomic prediction in a single homogeneous population (single-task). Then, a simple data pooling method by forming a training data set from multiple populations for genomic prediction was described. Next, the single-task model was extended to a multi-task Bayesian learning model. In addition, a new Gibbs sampling algorithm was designed to infer the unknown parameters in the multi-task Bayesian learning model. A computer program was also developed to implement the multi-task Bayesian learning model, single-task Bayesian learning and the simple data pooling method.
Single-task Bayesian learning model

In a single reference population of $n$ animals with phenotypes and genotypes on $m$ SNP markers, the statistical model can be written as:

$$y_i = \mu + \sum_{j=1}^{m} x_{ij}a_j + e_i$$

Where $y_i$ is the phenotypic value for the $i^{th}$ animal ($i = 1, n$), $\mu$ is the intercept, $x_{ij}$ is the genotype for the $j^{th}$ SNP locus ($j = 1, m$) of the $i^{th}$ animal, which is coded 0, 1 or 2, depending on the number of copies from a specified allele, $a_j$ is the regression coefficient for the $j^{th}$ SNP (allele substitution effect), and $e_i$ is the random residual error.

A flat prior distribution is assigned to $\mu$. $a_j$ is assumed a mixture of a normal distribution $N(0,\sigma_a^2)$ and a point mass density at zero (denoted by a Dirac delta function $\delta_0(a_j)$ hereinafter). The weights for the two distributions are $(1 - w)$ and $w$, respectively, so that $(a_j|w,\sigma_a^2) \sim (1 - w)N(0,\sigma_a^2) + w\delta_0(a_j)$. $w$ follows an uniform prior distribution. A latent indicator variable $\gamma_j$ is introduced for each SNP, so that when $\gamma_j = 1$, $a_j \sim N(0,\sigma_a^2)$, and when $\gamma_j = 0$, $a_j = 0$. Prior distribution for each $\gamma_j$ is assumed i.i.d. and follows Bernoulli distribution with probability $(1 - w)$. So the joint prior distribution for $\gamma$ is $f(\gamma|w) = \prod_j w^{(1-\gamma_j)}(1 - w)\gamma_j$. And residual errors are from a multivariate normal distribution $N(0,1\sigma_e^2)$. The prior distribution for $\sigma_a^2 (\sigma_e^2)$ is a scaled inverse Chi-square distribution with degree of freedom $\nu_a (\nu_e)$ and scale factor $s_a^2 (s_e^2)$.

The single-task Bayesian mixture model described above is similar to the BayesC$\pi$ model as described by Habier et al. (2011). Instead of using $\pi$ as in their model, $w$ is used in this paper to denote the proportion of SNPs having zero effects, and $\pi$ is used to stand for the mathematical constant.
**Simple data pooling method**

Suppose animals are from a number of \( c \) different populations. In a simple data pooling method, animals from multiple populations are pooled together to form a single training data set. It is assumed that the population origin for each individual is known prior to the analysis. Population origin is included as a fixed effect. The effect of each SNP is assumed to be the same across populations.

**Multi-task Bayesian learning model**

For \( c \) populations with \( n_k \) animals in each population \( k \), the statistic model can be written as:

\[
y_{ik} = \mu_k + \sum_{j=1}^{m} x_{ijk} a_{jk} + e_{ik} \quad (i = 1, \ldots, n_k \text{ and } k = 1, \ldots, c),
\]

In matrix notation, this can be written as:

\[
y_k = \mu_k + X_k a_k + e_k,
\]

where \( y_{ik} \) is the phenotypic value for the \( i^{th} \) animal in the \( k^{th} \) population, \( \mu_k \) is the general mean of population \( k \), \( x_{ijk} \) is the genotype for the \( j^{th} \) SNP locus of the \( i^{th} \) animal in the \( k^{th} \) population, \( a_{jk} \) is the \( j^{th} \) SNP effect in population \( k \), and \( e_{ik} \) is the random residual effect.

In the model, \( a_{jk} \) allows the \( j^{th} \) SNP effect to have a unique size and direction in population \( k \). To share information among different populations, a common latent indicator variable indicating whether SNP \( j \) is associated with a QTL is used across populations. Accommodating these features into a Bayesian model produces the multi-task Bayesian learning model.

The following prior distributions for the unknown parameters and hyper-parameters are assumed in the multi-task Bayesian learning model:
\( \mu_k \sim \text{flat distribution}, \)

\[
(a_{jk} \mid \gamma_j, \sigma^2_{ak}) \sim \gamma_j N(0, \sigma^2_{ak}) + (1 - \gamma_j) \delta_0(a_{jk}), \]

\( (\gamma_j \mid w) \sim w^{1 - \gamma_j} (1 - w)^{\gamma_j}, \)

\( w \sim \text{Uniform}(0, 1), \)

\[
(\sigma^2_{ak} \mid v_{ak}, s^2_{ak}) \sim \frac{\left(\frac{v_{ak} s^2_{ak}}{2}\right)^{\frac{v_{ak}}{2}}}{\Gamma\left(\frac{v_{ak}}{2}\right)} \left(\sigma^2_{ak}\right)^{-(l + v_{ak})/2} \exp\left(-\frac{v_{ak} s^2_{ak}}{2\sigma^2_{ak}}\right),
\]

\[
(e_k \mid \sigma^2_{ek}) \sim N(0, I \sigma^2_{ek}),
\]

\[
(\sigma^2_{ek} \mid v_{ek}, s^2_{ek}) \sim \frac{\left(\frac{v_{ek} s^2_{ek}}{2}\right)^{\frac{v_{ek}}{2}}}{\Gamma\left(\frac{v_{ek}}{2}\right)} \left(\sigma^2_{ek}\right)^{-(l + v_{ek})/2} \exp\left(-\frac{v_{ek} s^2_{ek}}{2\sigma^2_{ek}}\right).
\]

The likelihood function of the whole data given all the parameters in the model is:

\[
\prod_{k=1}^{c} \frac{1}{(2\pi \sigma^2_{ek})^{\frac{n_k}{2}}} \exp\left(-\frac{1}{2} \sigma^2_{ek} (y_k - \mu_k - X_k a_k) (y_k - \mu_k - X_k a_k)\right).
\]

So the joint posterior distribution is:

\[
f(\sigma^2_a, \sigma^2_e, w, \gamma, a, \mu \mid y) \propto \prod_{k=1}^{c} \left(\frac{1}{\sigma^2_{ek}}\right)^{v_{ek} + n_k - 1} \exp\left[-\frac{(y_k - \mu_k - X_k a_k) (y_k - \mu_k - X_k a_k) + v_{ek} s^2_{ek}}{2\sigma^2_{ek}}\right] \left(\sigma^2_{ek}\right)^{-\frac{v_{ek}}{2}} \exp\left(-\frac{v_{ek} s^2_{ek}}{2\sigma^2_{ek}}\right) \prod_{k=1}^{c} \prod_{j=1}^{m} \gamma_j \left(\sigma^2_{ak}\right)^{-\frac{v_{ak}}{2}} \exp\left(-\frac{a_{jk}^2}{2\sigma^2_{ak}}\right) \left(1 - \gamma_j\right) \delta_0(a_{jk})^w (1 - w)^{1 - w}.
\]

**Gibbs sampling algorithm**

A Gibbs sampling algorithm was designed to draw samples for unknown (hyper-) parameters from their full conditional posterior distributions. To avoid reducibility of Markov chain, \( \gamma_j \) and \( a_{jk} \) are jointly sampled by first sampling \( \gamma_j \) from \( f(\gamma_j \mid \theta_j, y) \)
followed by sampling $a_{jk}$ from $f\left(a_{jk} \mid \gamma_j, \theta_j, y\right)$, where $\theta_j$ represents all parameters except $\gamma_j$ and $a_{jk}$. Full conditional posterior distributions for $\mu_k, w, \sigma_{ak}^2$ and $\sigma_{ek}^2$ can be derived by picking up the relevant parts from the joint posterior distribution. Derivation for the density function $f\left(\gamma_j \mid \theta_j, y\right)$ and sampling of $\gamma_j$ are described in the Appendix.

The Gibbs sampling steps are described as below:

Step 1. Initialize the parameters $w, \gamma, \sigma_{ak}^2, \sigma_{ek}^2, \mu_k$ and $a_k$.

Step 2. For $j = 1, \ldots, m$

a. Sample $\gamma_j$ from Bernoulli distribution with probability $1/(1 + q_j)$, where

$$q_j = \frac{w}{1 - w} \sqrt{\prod_k \left(\frac{x'_{jk} x_{jk} \sigma_{ak}^2}{\sigma_{ek}^2} + 1\right) \exp\left(-\frac{1}{2} \sum_k \frac{\hat{\mu}_{a_{jk}}^2}{\hat{\sigma}_{a_{jk}}^2}\right)}$$

for detailed expression of $\hat{\mu}_{a_{jk}}$ and $\hat{\sigma}_{a_{jk}}^2$.

b. For $k = 1, \ldots, c$, sample $a_{jk}$ from

$$f\left(a_{jk} \mid \gamma_j, \theta_j, y\right) = \begin{cases} \delta_0(a_{jk}) & \text{if } \gamma_j = 0 \\ N(\hat{\mu}_{a_{jk}}, \hat{\sigma}_{a_{jk}}^2) & \text{if } \gamma_j = 1 \end{cases}$$

Step 3. Sample $w$ from Beta distribution

$$f\left(w \mid \gamma\right) = \frac{1}{B(\alpha, \beta)} w^{\alpha-1} (1-w)^{\beta-1}, \text{ where } \alpha = m - \sum \gamma_j \text{ and } \beta = \sum \gamma_j.$$ 

Step 4. For $k = 1, \ldots, c$, sample $\sigma_{ak}^2$ from scaled inverse Chi-square distribution

$$\chi^2\left(v_{ak} + \sum \gamma_j, \frac{v_{ak} \sigma_{ak}^2 + \sum \gamma_j a_{jk}^2}{v_{ak} + \sum \gamma_j}\right)$$

Step 5. For $k = 1, \ldots, c$, sample $\sigma_{ek}^2$ from scaled inverse Chi-square distribution
\[ \chi^2 \left( \frac{v_{ek} + n_k, \frac{v_{ek}}{v_{ek} + n_k} + \left( y_k - \mu_k - X_k a_k \right) \left( y_k - \mu_k - X_k a_k \right)^\prime}{v_{ek} + n_k} \right) \]

Step 6. For \( k = 1, \ldots, c \)

Sample \( \mu_k \) from \( N \left( \frac{\left( y_k - X_k \hat{\alpha}_k \right) \left( y_k - X_k \hat{\alpha}_k \right)^\prime}{n_k}, \frac{\hat{\sigma}_{ek}^2}{n_k} \right) \)

Repeat Step 2 to 6 until a set number of iterations are reached.

**Computer program**

A computer program called MTBL was written with ANSI standard C language to implement the multi-task Bayesian learning model, single-task Bayesian learning and the simple data pooling method. The Gibbs sampling algorithms used for the single-task Bayesian learning model and simple data pooling method were similar to that used in the multi-task Bayesian learning model.

**Real data example**

**Data**

Data from Holstein and Ayrshire dairy breeds were used to evaluate the performance of the Bayesian multi-task learning model and to compare with the single-task and simple data pooling methods. Two SNP panels were used for the analysis. One was the Illumina BovineSNP50 BeadChip (50k) and the other was the Illumina BovineHD BeadChip (777k). Five traits were considered, including milk yield, fat yield, protein yield, fat percentage and protein percentage.

**Genotypes**

1,659 animals from Holstein breed (1,496 males and 163 females) and 520 Ayrshire bulls were genotyped on the 777k panel. SNPs meeting one of the following
criteria were excluded, minor allele frequency (MAF) lower than 0.05, missing genotype rate greater than 0.10, highly correlated with any other SNP genotype (with 95% of genotypes identical) and not located on autosomal chromosomes (or unknown location). SNPs filtered in one breed were also discarded in the other breed. After editing, 246,668 SNPs were kept for analyses.

1,608 Holstein dairy bulls were genotyped on the 50k panel. 126 of these 1,608 bulls also had genotypes on the 777k panel. Same criteria as used for the 777k data were used to edit the 50k genotypic data. Most of the SNPs in the 50k panel are also present in the 777k panel. So, if an animal did not have 50k genotypes but was genotyped on the 777k panel, then the subset genotypes for the 50k SNPs were extracted from its 777k genotypes. After editing, 520 Ayrshire bulls and 3,141 Holstein animals (2,978 males and 163 females) had genotypes on 28,206 SNPs from the 50k panel.

**Phenotypes**

Bull proofs (Estimated breeding values, EBV) from April 2008 and December 2011 were provided by Canadian Dairy Network (CDN). Proofs with reliability less than 0.65 were excluded.

**Assigning Training and validation animals**

Animals born before 2004 and with bull proofs in 2008 were used to derive prediction equations (training). Animals born in or after 2004 and with bull proofs in 2011 were used for validation.

Originally, the number of Holstein animals in the training set was 602 for the 777k data. The set of 1,482 bulls that had genotypes only on 50k panel was included into the training set by imputing their 28,206 SNP genotypes into 246,668 SNP genotypes.
Genotype imputation was conducted using the imputation software FImpute, which was developed by (Sargolzaei et al., 2011a). The set of 2,084 animals that had genotypes on the 777k panel (actual plus imputed genotypes) was used as reference population. A pedigree containing 15,731 Holstein animals was used. After imputation, the numbers of animals used for training and validation were the same for the 50k and 777k panel. The number of animals for genomic prediction is shown in Table 4.1.

**Evaluation**

The single-task Bayesian learning, the simple data pooling method and the multi-task Bayesian learning model were implemented in the self-developed computer program MTBL. When 28,206 SNPs were used in the analyses, the Gibbs sampling procedure was run for 50,000 iterations with the first 10,000 discarded as burn-in. And when 246,668 SNPs were used, the Gibbs sampling procedure was run for 100,000 iterations with the first 50,000 discarded as burn-in. Burn-in period was determined by visually inspecting the Gibbs chain. All samples were kept after the burn-in period.

SNP effects were estimated by averaging across all samples after the burn-in period. After the estimation of SNP effects, the direct genomic values (DGV) was calculated for animals in the validation set by summing up all the SNP effects. The performance of different methods was measured using two criteria. The first is the realized accuracy, which is measured as Pearson’s correlation coefficient between DGV and 2011 bull proofs for validation animals ($r_{(DGV, EBV)}$). The second is the standardized mean squared error (SMSE), which is calculated as the mean squared prediction error between DGV and 2011 EBV for validation animals, normalized by the sample variance of trait phenotypes.
Posterior model inclusion probability for each SNP was estimated from the frequency of 1’s in the samples of the latent indicator variable. The posterior model inclusion probability was viewed as a signal for detecting a nearby QTL as being used in GWAS (Guan and Stephens, 2011).

Results

Performance using 28,206 SNPs

The performance of the single-task Bayesian learning model, the simple data pooling method and the multi-task Bayesian learning model using 28,206 SNP genotypes for genomic prediction is shown in Table 4.2. Depending on the trait, the multi-task model produced a gain of accuracy from 0 to 7 percentage points (corresponding to 0 – 14.5%) in the Ayrshire validation set compared to the single-task model, while using the simple data pooling method resulted in a reduction of accuracy from 8 to 10 percentage points (corresponding to 6 – 15.4%) except for protein percentage, for which there was an increase of 3 percentage points (corresponding to 6.3%). In the Holstein population, the single-task model, simple data pooling method, and the multi-task model showed a similar accuracy for all the traits. The multi-task model yielded the lowest SMSE for all the traits, in both the Ayrshire and Holstein validation set, while the simple data pooling method showed the largest SMSE.

To show the consequence of using different methods on the estimation of SNP effects, Figure 4.2 presents the posterior model inclusion probabilities (signal) and the estimated SNP effects for fat percentage for 4 SNPs in the region of DGAT1 gene. The SNP ID and their locations on the chromosome, the estimated effects and posterior model inclusion probabilities from different models are shown in Table S1 in Appendix D.
Using the single-task Bayesian learning model, the signals were stronger in the Holstein population than in the Ayrshire population, and, consequently, resulted in larger estimated SNP effects. The closest SNP to the DGAT1 gene is located at 1.7Mb. In the Holstein population, this SNP received the strongest signal and the largest SNP effects for all the three methods. In the Ayrshire population, only the last SNP close to 2.0Mb showed a weak signal and effect. For this SNP, the signal was enhanced in both populations when the simple data pooling method or the multi-task model was used, compared to using the single-task model. In the Ayrshire population, the estimated SNP effects by the simple data pooling method deviated from those estimated by the single-task model much more than that from the multi-task model. For the Holstein population, the estimated SNP effects from the three methods were close to each other except for the SNP at 2.0Mb. It seemed that when data from the two populations were combined as in the simple data pooling or multi-task methods, the SNP effects shifted a little from the left side of the DGAT1 gene to the right side. This was possibly due to the stronger signal for the SNP at 2.0Mb when data from the two populations were combined, compared to the signal from using the single-task method.

**Performance using 246,668 SNPs**

Genomic prediction using genotypes on the 246,668 SNPs involved 1,482 Holstein training animals whose genotypes were imputed from the 28,206 SNPs. To evaluate the imputation accuracy, 126 animals that were genotyped on both the SNP panels were used. These animals were removed from the reference population and their genotypes were imputed from 28,206 to 246,668 SNPs. Their imputed genotypes were
compared to the real genotypes. The ratio of the genotypes that were correctly imputed was 0.9930.

To further evaluate whether using the imputed genotypes to enlarge the training sample size can increase the accuracy of genomic prediction, training set using the 602 Holstein animals that had real 246,668 SNP genotypes and using 2,084 animals (containing 1,482 animals with imputed genotypes) were compared using the single-task Bayesian learning model to predict DGV for the validation animals. The accuracy using these two training sets was shown in Figure 4.1. As can be seen from Figure 4.1, using the larger training set with imputed genotypes resulted in increased accuracy for all the traits.

Table 4.3 shows the results of genomic prediction using genotypes on 246,668 SNPs via the single-task Bayesian learning model, the simple data pooling method, and the multi-task Bayesian learning model. In the Ayrshire validation population, depending on the trait, the accuracy achieved from the multi-task model increased by 0 to 3 percentage points (corresponding to 0 – 4.5%), compared to the single-task model, while the simple data pooling method resulted in a reduction by 1 to 9 percentage points (corresponding to 1.5 – 17.6%). The accuracy in the Holstein validation population was similar using the three methods. For all the traits and in both the Ayrshire and Holstein validation set, the multi-task model yielded the lowest SMSE while the simple data pooling method showed the largest SMSE.

Figure 4.3 presents the posterior model inclusion probabilities and estimated SNP effects for fat percentage for seven SNPs in the region of DGAT1 gene. The SNP ID and their locations on BTA14, the estimated effects and posterior model inclusion
probabilities from different models are shown in Table S2 in Appendix D. The three SNPs at 1.51Mb, 1.70Mb and 1.92Mb were the same as the first three SNPs in the DGAT1 gene region as used in the lower density (28,206 SNPs) data analyses. The SNP at 2.00Mb was also close to the last one of the four SNPs in the DGAT1 region used in the lower density. Three other SNPs at 1.59Mb, 1.86Mb and 1.97Mb were also included in the DGAT1 region in this higher density SNP data. In the Holstein population, the SNP at 1.51Mb had no signal as opposed to the previous results in the lower density. The SNP at 1.70Mb received a weakened signal. The SNP at 1.92Mb was a little complicated, because there were two additional SNPs close to it, one on each side. In the single-task model, the signal for this SNP increased, but when data were combined as in the simple data pooling method and the multi-task model, the signal was weakened and shared by its two flanking SNPs. The SNP at 2.00Mb behaved as the last SNP in the lower density. So combining data from the two populations enhanced the signal for this SNP and increased the estimated effect size. The effects for SNPs at 1.5Mb and 1.7Mb diminished compared to the effects in the lower density results. This could be due to the grouping effect, where closely linked SNPs may split the QTL effect (Kemper et al., 2012; Xu, 2003). In the Ayrshire population, the simple data pooling method also resulted in larger differences in the estimated SNP effects than the multi-task model, compared to the single-task model.

Discussion

Traditionally, genomic prediction with data from multiple populations were implemented either by running genomic prediction within each population (single-task) or by simply pooling data together. Single-task genomic prediction cannot utilize information from other populations and therefore, accuracy can be limited by the sample
size. On the other hand, it is difficult to effectively account for the differences of SNP effects among different populations by simply pooling data together. If the marker density is low or the populations are divergent from each other, simply pooling data together may result in unfavorable accuracies (de Roos et al., 2009). The multi-task Bayesian learning model proposed in this study uses information from all populations simultaneously while allowing the SNP effects to vary in different populations. Different populations share information through a common set of latent indicator variables. When the target trait has a similar genetic background in related populations, it is reasonable to assume some shared QTL affecting a common trait in different populations. However, the linkage phase between SNP markers and QTL are likely to be inconsistent, especially when the marker density is low. Therefore, the multi-task Bayesian learning model is more flexible about the SNP effects and is likely to have better performance than a simple data pooling method.

Results from real data analyses in this study showed that the multi-task Bayesian learning model produced a higher accuracy and lower standardized mean squared error for DGV in general, compared to the single-task model. However, the gain of accuracy from the multi-task model was reduced when higher density of SNP genotypes was used. On the other hand, simply pooling data together resulted in a reduced accuracy compared to the single-task model, especially when the marker density was lower. Although less reduction of accuracy was observed using the simple data pooling method under the higher density markers, the SMSE was still high, suggesting biases in the estimated SNP effects. The SNP effects could be different across populations, especially when lower density markers were used (de Roos et al., 2009). The SNP effects for fat percentage in
the DGAT1 gene region (Figure 4.2 and Figure 4.3) also suggested that the SNP effects estimated by the simple data pooling method deviated from single-task estimates drastically in the smaller population. The SNP effect estimates in the multi-task model were more consistent to that in the single-task model for each population. For SNPs that were only detected by combining the two populations, multi-task model also captured their effects (Figure 4.2 and Figure 4.3). These results supported the use of multi-task Bayesian learning model for multi-population genomic prediction.

The advantage of using the multi-task Bayesian learning model over using of the single-task model varies among different traits. Gains of accuracy by using the multi-task model were higher for fat percentage and protein percentage traits than for other traits (Table 4.2 and Table 4.3). This is likely due to that large QTL or genes such as DGAT1, have larger influence on the percentage traits than on the yield traits (Grisart et al., 2004; Pimentel Eda et al., 2011). The statistical power of detecting a QTL largely depends on the sample size (Stranger et al., 2011). Therefore, small populations could benefit more from the multi-task model by sharing QTL signals with larger populations (Figure 4.2 and Figure 4.3).

From our limited validation study, where only two dairy cattle breeds and only milk production traits were considered, the multi-task Bayesian learning model was shown to be effective and more beneficial to populations with smaller sample size. For the larger populations, the results didn’t show any improvement. But it was shown that by combining the two populations, the SNPs selected and their estimated effects could vary between the multi-task model and the single-task model. The no improvement in the larger population could be due to that the smaller population is too small to be able to
have a significant impact on the larger population. In practice, as in beef cattle, there are situations that data are scattered in different small populations and there is a need to join the data together for analysis. It would be interesting to test the performance of the multi-task model under such scenarios.

The proposed multi-task Bayesian learning model can be viewed as a general framework for multi-population genomic prediction. The spike and slab distribution for the SNP effects is a special case and is similar to the one as employed in BayesC\(\pi\) (Habier et al., 2011). Other types of mixture distribution currently being used for genomic prediction (Calus et al., 2008; Meuwissen et al., 2001; Verbyla et al., 2009), can also be easily accommodated into the multi-task system. The strategy in the multi-task Bayesian learning model allows different populations to share information by using a common latent variable for each SNP. However, other strategies could be used, for example, by modeling the joint distribution of the SNP effects among different populations. Further investigations are required to evaluate alternative strategies for sharing information across populations.

**Conclusions**

A multi-task Bayesian learning model was proposed for multi-population genomic prediction, where information was shared across populations through a common set of latent indicator variables and the SNP effects are allowed to vary in different populations. Accuracy of genomic prediction achieve from the multi-task Bayesian learning model was generally higher than that of the single-task model and the simple data pooling method for animals in the smaller population. Larger gains were achieved from the multi-task model when a lower density SNP panel was used. The multi-task model is effective
and could be beneficial to populations where a small number of training set animals are available.

**Table 4.1** Number of animals used for genomic prediction.

<table>
<thead>
<tr>
<th></th>
<th>Holstein</th>
<th>Ayrshire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>2084</td>
<td>393</td>
</tr>
<tr>
<td>Validation set</td>
<td>214</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td>2298</td>
<td>458</td>
</tr>
</tbody>
</table>

**Table 4.2** Performance of genomic prediction using 28,206 SNPs via the single-task Bayesian learning model, the simple data pooling method, and the multi-task Bayesian learning model.

<table>
<thead>
<tr>
<th>Breed</th>
<th>Trait</th>
<th>Single-task</th>
<th>Pooling</th>
<th>Multi-task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$r_{(DGV,EBV)}$</td>
<td>SMSE</td>
<td>$r_{(DGV,EBV)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ayrshir e</td>
<td>Milk yield</td>
<td>0.52</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>Ayrshir e</td>
<td>Fat yield</td>
<td>0.64</td>
<td>0.35</td>
<td>0.55</td>
</tr>
<tr>
<td>Ayrshir e</td>
<td>Protein yield</td>
<td>0.70</td>
<td>0.29</td>
<td>0.60</td>
</tr>
<tr>
<td>Ayrshir e</td>
<td>Fat %</td>
<td>0.66</td>
<td>0.53</td>
<td>0.58</td>
</tr>
<tr>
<td>Ayrshir e</td>
<td>Protein %</td>
<td>0.48</td>
<td>0.69</td>
<td>0.51</td>
</tr>
<tr>
<td>Ayrshir e</td>
<td>Milk yield</td>
<td>0.66</td>
<td>0.26</td>
<td>0.65</td>
</tr>
<tr>
<td>Holstein</td>
<td>Fat yield</td>
<td>0.63</td>
<td>0.22</td>
<td>0.64</td>
</tr>
<tr>
<td>Holstein</td>
<td>Protein yield</td>
<td>0.69</td>
<td>0.18</td>
<td>0.69</td>
</tr>
<tr>
<td>Holstein</td>
<td>Fat %</td>
<td>0.74</td>
<td>0.45</td>
<td>0.74</td>
</tr>
<tr>
<td>Holstein</td>
<td>Protein %</td>
<td>0.67</td>
<td>0.46</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Table 4.3 Performance of genomic prediction using 246,668 SNPs via the single-task Bayesian learning model, the simple data pooling method, and the multi-task Bayesian learning model.

<table>
<thead>
<tr>
<th>Breed</th>
<th>Trait</th>
<th>Single-task</th>
<th></th>
<th></th>
<th>Pooling</th>
<th></th>
<th></th>
<th>Multi-task</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$r_{(DGV,EBV)}$</td>
<td>$SMSE$</td>
<td>$r_{(DGV,EBV)}$</td>
<td>$SMSE$</td>
<td>$r_{(DGV,EBV)}$</td>
<td>$SMSE$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ayrshire</td>
<td>Milk yield</td>
<td>0.54</td>
<td>0.47</td>
<td>0.53</td>
<td>7.33</td>
<td>0.55</td>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fat yield</td>
<td>0.67</td>
<td>0.32</td>
<td>0.62</td>
<td>11.67</td>
<td>0.67</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein yield</td>
<td>0.72</td>
<td>0.28</td>
<td>0.68</td>
<td>0.32</td>
<td>0.72</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fat %</td>
<td>0.66</td>
<td>0.52</td>
<td>0.65</td>
<td>0.77</td>
<td>0.69</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein %</td>
<td>0.51</td>
<td>0.68</td>
<td>0.42</td>
<td>0.94</td>
<td>0.53</td>
<td>0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holstein</td>
<td>Milk yield</td>
<td>0.64</td>
<td>0.27</td>
<td>0.64</td>
<td>2.13</td>
<td>0.64</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fat yield</td>
<td>0.63</td>
<td>0.22</td>
<td>0.64</td>
<td>3.93</td>
<td>0.63</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein yield</td>
<td>0.66</td>
<td>0.19</td>
<td>0.66</td>
<td>0.19</td>
<td>0.66</td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fat %</td>
<td>0.77</td>
<td>0.41</td>
<td>0.77</td>
<td>1.61</td>
<td>0.78</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein %</td>
<td>0.71</td>
<td>0.41</td>
<td>0.68</td>
<td>0.43</td>
<td>0.70</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure 4.1** Accuracy of DGV predicted using training data sets of different sizes.

DGV was predicted for animals in the Holstein validation set (n = 214) using genotypes on 246,668 SNPs via a single-task Bayesian learning model. In one training data set (n = 602), all animals have actual genotypes on the 246,668 SNPs; in the other training data set (n = 2,084), 1,482 animals have their high density genotypes imputed from 28,206 SNPs.
Figure 4.2 Signals (posterior inclusion probabilities) and effects for SNPs close to the DGAT1 region estimated using 28,206 SNPs.

Signals and effects were estimated for fat percentage. Location of the DGAT1 gene (dashed line) was determined from the bovine assembly UMD3.1. A) SNP signals in Holstein; B) SNP effects in Holstein; C) SNP signals in Ayrshire; D) SNP effects in Ayrshire.
**Figure 4.3** Signals (posterior inclusion probabilities) and effects for SNPs close to the DGAT1 region estimated using 246,668 SNPs. Signals and effects were estimated for fat percentage. Location of the DGAT1 gene (dashed line) was determined from the bovine assembly UMD3.1. A) SNP signals in Holstein; B) SNP effects in Holstein; C) SNP signals in Ayrshire; D) SNP effects in Ayrshire.
GENERAL DISCUSSION

Genotype imputation errors impact both GBLUP and Bayesian methods. The difference of the impact depends on the genetic architecture of the traits. For traits affected by a few large QTL, the Bayesian mixture model used in Chapter 1 resulted in greater reduction rates in accuracy of genomic prediction than GBLUP. However, for all SNP panels, scenarios and all traits studied in Chapter 1, the Bayesian mixture model produced higher or similar accuracy compared to the GBLUP method. For chromosome regions with a large QTL, the Bayesian method tended to select the most relevant SNPs, while GBLUP or the Ridge-regression picked many more SNPs surrounding the QTL, and for other regions, the Bayesian method puts less weight on the SNPs than GBLUP. So the Bayesian method could suffer more from imputation errors in large QTL regions, but the GBLUP method would suffer from imputation errors accumulated over many more SNPs, which would tend to cancel out if the imputation errors are random across the genome. This could possibly explain why the Bayesian method had a greater reduction rate in accuracy of genomic prediction for traits with large QTL and why the Bayesian method still resulted in higher accuracies than GBLUP for all scenarios. The relative performance of the two methods might be more related to distributions of imputation errors. If more imputation errors were distributed around the QTL, one could speculate that the Bayesian method would suffer more from these errors than GBLUP and consequently more reduction in the accuracy of genomic prediction would be observed. Further investigations are needed to test this hypothesis.

In the future, more and more animals would be genotyped on low density panels. One might be interested to include these animals in the training population to derive
genomic prediction equations. Before putting a low density panel into application, it is important to know whether, as well as how much including imputed genotypes would result in loss of accuracy in genomic prediction. Results in Chapter 1 showed that for a SNP panel with a density lower than 6k, the accuracy of genomic prediction was reduced when a proportion of animals in the training set were imputed. However, for the 6k panel, accuracy of genomic prediction was almost not affected even when 66% of the training set was imputed, which would justify the application of this panel.

However, some limitations of the study in Chapter 1 should be addressed. First, the sample size of the training set was kept constant regardless of the imputation. The use of imputation could, however, also aid to enlarge the sample size of the training set, which might in fact increase the accuracy of the GEBVs. Second, imputed genotypes were included in the training set without considering their imputation accuracies. In practice, only animals that have their genotypes imputed with high accuracies should be included in the training set. These possible scenarios warrant further investigation and the conclusions from the current study should be used with caution.

Genomic prediction using Bayesian methods are mostly implemented using MCMC sampling algorithms which are computationally demanding. In practice, Bayesian methods may not be favored due to their high computational demands. For most complex traits, GBLUP can perform as well as Bayesian methods. However, for traits where a few large QTL exist, Bayesian methods have superior performance and may be preferable. The new computing algorithm RHSU in Chapter 2 can implement Bayesian methods more efficiently than the traditional algorithm GSRU when the training data size becomes large. The relative efficiency of RHSU compared to GSRU depends on two
factors, the ratio between the number of animals in the training data and the number of SNPs, and the proportion of SNPs having no effect on the trait ($\pi$). For most complex traits, $\pi$ usually has a relatively large value reflecting that most of the markers have no effect. Therefore, in practice where a 50k SNP panel was adopted, only a few thousands of animals would be sufficient for RHSU to outperform GSRU, which suggests the RHSU would be very useful if Bayesian methods are needed for genomic prediction.

Two tools were developed in Chapter 2 for the theoretical comparison between the RHSU and GSRU algorithms. One is the relative efficiency (RE) defined based on FLOP, and the other is the threshold value for the ratio between number of animals in the training set and number of SNP markers defined for RHSU to be more efficient than GSRU in different scenarios. Results from a real data example agreed very well with theoretical expectations from these tools. Therefore, RE and the calculated thresholds may offer reliable tools for making comparisons and choices between the two computing algorithms in practice.

However, it has to be noted that time used for pre-computing $X'X$ was not taken into account when making comparisons between GSRU and RHSU. When the sample size is large, computing $X'X$ could cost a substantial amount of time. Block-wise matrix multiplication and a parallel computing strategy could be used to calculate $X'X$ efficiently. Furthermore, the matrix $X'X$ only needs to be calculated once if multiple traits are to be analyzed. Calculation of $X'X$ can be conducted with iteration on animals. So, once $X'X$ is computed on some animals, new animals that need to be added into and animals that need to be removed from the training sample can be processed very quickly. Efficient algorithms for pre-computing $X'X$ need to be developed in future.
Accuracy of genomic prediction is largely dependent on the number of animals in the training data set. Assembling a large training set is a challenge for small breeds or populations, such as Ayrshire in dairy cattle and most beef cattle populations. In Chapter 3, different strategies of forming a training population were evaluated for genomic prediction for RFI in Angus and Charolais beef populations with 50k SNP genotypes. Results suggested that using the 50k SNP panel, within-breed genomic prediction is a safe strategy; across-breed prediction resulted in the lowest accuracy and therefore is not recommended with the 50k SNP panel; pooling data from different breeds has a potential to improve the accuracy but must be used with caution due to possible loss of accuracy. The prediction accuracies of DGV for RFI obtained in the study of Chapter 3 were still low, regardless of the training population forming strategies, indicating the need to include more animals in the training population and to increase the marker density.

Genomic prediction with multiple populations can be treated as a multi-task learning problem, where each task is to derive prediction equations for each population and tasks are related through some common features among the populations. In Chapter 4, a multi-task Bayesian learning model for multi-population genomic prediction is proposed. In the multi-task Bayesian learning model, information was shared across populations through a common set of latent indicator variables, and the SNP effects are allowed to vary in different populations. Results from real data in Holstein and Ayrshire dairy populations suggested that, the multi-task model was effective and more beneficial to the smaller Ayrshire population, where a small number of animals were available, than for the larger Holstein population, for which the results did not show any improvement. This was possibly due to the fact that Ayrshire population was too small to be able to
have a significant impact on the larger Holstein population. In practice, as in beef cattle, there are situations where data are scattered in different small populations and there is a need to join the data together for analysis. It would be interesting to test the performance of the multi-task model under such scenarios.

The proposed multi-task Bayesian learning model in Chapter 4 can be viewed as a general framework for multi-population genomic prediction. The spike and slab distribution for the SNP effects used in the study was a special case. Other types of mixture distribution currently being used for genomic prediction can also be easily accommodated into the multi-task system. The strategy in the multi-task Bayesian learning model allows different populations to share information by using a common latent variable for each SNP. However, other strategies could be used, for example, by modeling the joint distribution of the SNP effects among different populations. Further investigations are required to evaluate alternative strategies for sharing information across populations.

In summary, four studies have been conducted in this thesis. In the first study, the impact of genotype imputation errors on GBLUP and Bayesian methods was evaluated. It was found that imputation errors have greater impact on the Bayesian method than on GBLUP for traits that are influenced by a few large QTL; in the second study, an alternative computing strategy was proposed to implement the MCMC sampling steps in a Bayesian mixture model. The proposed algorithm was shown to be computationally more efficient than the conventional one once the sample size exceeded a fraction of the number of SNPs; in the third study, accuracy of genomic prediction for residual feed intake in Angus and Charolais beef cattle was investigated and different strategies of
forming the training population were compared. It was concluded that across-breed prediction resulted in the lowest accuracy and combining data by pooling different breeds in the training set had potential to improve genomic prediction accuracy; in the last study, a multi-task Bayesian learning model was proposed for multi-population genomic prediction. It was demonstrated that the multi-task model was effective and beneficial to the small genotyped Ayrshire population. This thesis is expected to provide significant contributions to the field of genomic prediction and a basis for future studies, which would include extending the multi-task model to accommodate crossbred animals and investigating models and methods using sequence genotypes for genomic prediction.


Huang, Y. J., J. M. Hickey, M. A. Cleveland, and C. Maltecca. 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genetics Selection Evolution 44.


Su, G. et al. 2012. Comparison of genomic predictions using medium-density (similar to 54,000) and high-density (similar to 777,000) single nucleotide polymorphism marker panels in nordic holstein and red dairy cattle populations. J Dairy Sci 95: 4657-4665.


APPENDICES

Appendix A.

Derivation of \( f(y_j | \theta_j, y) \).

With the assumption of independence between \( \theta_j \) and \( r_j \), by Bayes’ theorem, one has

\[
f(r_j = 1 | \theta_j, y) = \frac{f(y | r_j = 1, \theta_j, r_j) f(r_j = 1)}{f(y | r_j = 1, \theta_j) f(r_j = 1) + f(y | r_j = 0, \theta_j) f(r_j = 0)} = \frac{1}{1 + q}, \tag{A.1}
\]

Where

\[
q = \frac{f(y | r_j = 0, \theta_j, r_j) f(r_j = 0)}{f(y | r_j = 1, \theta_j) f(r_j = 1)}. \tag{A.2}
\]

Similarly,

\[
f(r_j = 0 | \theta_j, y) = \frac{q}{1 + q}.
\]

Now the conditional likelihoods \( f(y | y_j = 0, \theta_j) \) and \( f(y | y_j = 1, \theta_j) \) are derived.

Suppose one is at the \( (l + 1) \)th Gibbs sampling iteration, and wants to sample \( y_j \) and \( a_j \),

the linear model given all parameters except \( y_j \) and \( a_j \), can be written as:

\[ y^* = x_j a_j + e, \]

Where \( y^* = y - X_{1:j-1} \hat{a}_{1:j-1}^l - X_{j+1:m} \hat{a}_{j+1:m}^l \).

Given the priors that \( (a_j | y_j) \sim y_j N(0, \sigma_a^2) + (1 - y_j) \delta_0(\cdot) \) and \( (e | \sigma_e^2) \sim N(0, I \sigma_e^2) \), the conditional likelihood of \( y \) can be written as:

\[
f(y | r_j = 0, \theta_j) = (2\pi \sigma_e^2)^{-n/2} \exp\left[-\frac{(y^* - y^*)^2}{2\sigma_e^2}\right]. \tag{A.3}
\]

And
Where $V_{y^*}$ is the co-variance matrix of $y^*$, $p$ is the mathematical constant.

Next, $|V_{y^*}|$ and $V_{y^*}^{-1}$ are derived.

$$ |V_{y^*}| = |x_jx'_j\sigma_a^2 + 1\sigma_e^2| = \sigma_e^{2n} |x_j\sigma_a x'_j\sigma_a + 1| $$

By applying Sylvester's determinant theorem, one has

$$ |V_{y^*}| = \sigma_e^{2n} \left(x'_jx_j\sigma_a^2 + 1\right) \quad \text{[A.5]} $$

It can be easily proven that $V_{y^*}^{-1}$ is:

$$ V_{y^*}^{-1} = \sigma_e^{-2} \left(I - \frac{x_jx'_j}{x'_jx_j + \sigma_e^2 / \sigma_a^2}\right) \quad \text{[A.6]} $$

Substituting [A.5] and [A.6] into [A.4], one has

$$ f(y | r_j = 1, \theta_j) = (2\pi \sigma_e^2)^{-n/2} (x'_jx_j\sigma_a^2 + 1)^{-1/2} \exp\left\{-\frac{y^* - y^* y^* (x'_jx_j + \sigma_a^2 / \sigma_e^2)}{2\sigma_e^2}\right\} \quad \text{[A.7]} $$

Denoting $\hat{\mu}_{a_j} = \frac{x'_j(y - x_{1:j-1}a_{1:j-1} + x_{j+1:m}a_{j+1:m})}{x_jx'_j + \sigma_e^2 / \sigma_a^2}$ and $\hat{\sigma}_a^2 = \frac{\sigma_e^2}{x_jx'_j + \sigma_e^2 / \sigma_a^2}$, and substituting [A.7] and [A.3] into [A.2], one gets

$$ q = \frac{f(y | r_j = 0, \theta_j) f(r_j = 0)}{f(y | r_j = 1, \theta_j) f(r_j = 1)} = \frac{\pi}{1 - \pi} \sqrt{x'_jx_j\sigma_a^2 + 1} \exp\left\{-\frac{\hat{\mu}_{\hat{a}_j}^2}{2\hat{\sigma}_{\hat{a}_j}^2}\right\} \quad \text{[A.8]} $$

Therefore, $\gamma_j$ can be drawn from a Bernoulli distribution with probability $\frac{1}{1 + q}$. 

110
Appendix B.

Pseudo code for implementing GSRU

//X and y were stored in memory.

//initialise
PI = USER_INPUT_PI;
varE = USER_INPUT_varE;
varA = USER_INPUT_varA;
ve = USER_INPUT_ve;
vg = USER_INPUT_vg;
varG = varA / (1 - PI) / mean2pq /NSNP;
Se = (ve - 2) * varE / ve;
Sg = (vg - 2) * varG / vg;

for(i=0; i<NSNP; i++)
{
    r[i] = 0;
    g[i] = 0;
    xpx[i] = 0;
    for(j = 0; j<NANIM; j++)
        xpx[i] += X[i][j] * X[i][j];
}

for(i=0; i<NANIM; i++)
{
    e[i] = y[i];
}

//Gibbs sampling
for(nround = 0; nround < MAXROUND; nround++)
{
    //sample varE
    scaleE = ve * Se;
    for(i=0; i<NANIM; i++)
    {
        scaleE += e[i] * e[i];
    }
    varE = INVCHI(NANIM + ve, scaleE);

    //sample latent variable r[i] and SNP effect g[i]
    nr = 0;
    scaleG = vg * Sg;
    lamda = varE / varG;
    for(i=0; i<NSNP; i++)
    {
        lhs = lamda + xpx[i];
    }
rhs = xpx[i] * g[i];
for(j=0; j<NANIM; j++)
    rhs += X[i][j] * e[j];
gmean = rhs / lhs;
vari = varE/lhs;
logq = log(PI/(1-PI)) - (gmean*gmean/vari - log(xpx[i]/lamda +1))/2;
pr = 1 / (1+exp(logq));
rp[i] = r[i];
r[i] = Bernoulli(pr);
old = g[i];
if(r[i] == 1)
{
    g[i] = Normal(gmean, sqrt vari);
    scaleG += g[i] * g[i];
    nr++;
}
else
{
    g[i]=0;
}
if(rp[i] == 1 || r[i]==1)
{
    old -= g[i];
    for(j = 0; j < NANIM; j++)
        e[j] += X[i][j] * old;
}

//sample varG
varG = INVCHI(vg + nr, scaleG);
//sample PI
PI = 1- Beta(1+ nr, 1+ NSNP - nr);
//adjust Sg
Sg = (vg - 2) * varA / (1-PI) / mean2pq /NSNP / vg;
Pseudo code for implementing RHSU

//X, y and X'X were stored in memory.

//initialise
PI = USER_INPUT_PI;
varE = USER_INPUT_varE;
varA = USER_INPUT_varA;
ve = USER_INPUT_ve;
vg = USER_INPUT_vg;
varG = varA / (1-PI) / mean2pq /NSNP;
Se = (ve - 2) * varE / ve;
Sg = (vg - 2) * varG / vg;
for(i=0; i<NSNP; i++)
{
    r[i] = 0;
    g[i] = 0;
    rhs[i] = 0;
    for(j=0; j<NANIM; j++)
    {
        rhs[i] += X[i][j] * y[j];
    }
}
for(i=0; i<NANIM; i++)
{
    e[i] = y[i];
}
Load_Precomputed_XX();

//Gibbs sampling
for(nround = 0; nround < MAXROUND; nround++)
{
//sample varE
    scaleE = ve * Se;
    for(i=0; i<NANIM; i++)
    {
        scaleE += e[i] * e[i];
        e[i] = y[i];
    }
    varE = INVCHI(NANIM + ve, scaleE);

//sample latent variable r[i] and SNP effect g[i]
    nr = 0;
    scaleG = vg * Sg;
    lamda = varE / varG;
    for(i=0; i<NSNP; i++)
\[
\begin{align*}
\text{lhs} &= \lambda + XX[i][i]; \\
gmean &= \text{rhs}[i] / \text{lhs}; \\
\text{vari} &= \text{varE}/\text{lhs}; \\
\text{logq} &= \log(\text{PI}/(1-\text{PI})) - (gmean \cdot \text{gmean}/\text{vari} - \log(XX[i][i]/\lambda + 1))/2; \\
pr &= 1 / (1+\exp(\text{logq})); \\
\text{rp}[i] &= r[i]; \\
r[i] &= \text{Bernoulli}(pr); \\
\text{old} &= g[i]; \\
\text{if}(r[i] == 1) \\
&\{ \\
\hspace{1em} g[i] &= \text{Normal}(\text{gmean}, \sqrt{\text{vari}}); \\
\hspace{1em} \text{scaleG} +\text{=} g[i] \cdot g[i]; \\
\hspace{1em} nr++; \\
\hspace{1em} \text{for}(j=0; j<NANIM; j++) \\
\hspace{2em} &\{ \\
\hspace{3em} e[j] &= X[i][j] \cdot g[i]; \\
\hspace{2em} &\} \\
&\} \\
\text{else} \\
&\{ \\
\hspace{1em} g[i]=0; \\
&\} \\
\text{if}(\text{rp}[i] == 1 || r[i]==1) \\
&\{ \\
\hspace{1em} \text{old} &= g[i]; \\
\hspace{1em} \text{for}(j = 0; j < i; j++) \\
\hspace{2em} \text{rhs}[j] &= XX[i][j] \cdot \text{old}; \\
\hspace{1em} \text{for}(j= i+1; j < \text{NSNP}; j++) \\
\hspace{2em} \text{rhs}[j] &= XX[i][j] \cdot \text{old}; \\
&\} \\
\text{// sample varG} \\
\text{varG} &= \text{INVCHI}(\text{vg} + \text{nr}, \text{scaleG}); \\
\text{// sample PI} \\
\text{PI} &= 1 - \text{Beta}(1+ \text{nr}, 1+ \text{NSNP} - \text{nr}); \\
\text{// adjust Sg} \\
\text{Sg} &= (\text{vg} - 2) \cdot \text{varA} / (1-\text{PI}) / \text{mean2pq} / \text{NSNP} / \text{vg}; \\
\end{align*}
\]
Appendix C.

Sampling of \( r_j \) in a Multi-task Bayesian learning model.

With the assumption of independence between \( \theta_j \) and \( r_j \), from Bayes’ theorem, one has

\[
f(r_j = 1 | \theta_j, y) = \frac{f(y | r_j = 1, \theta_j) f(r_j = 1)}{f(y | r_j = 1, \theta_j) f(r_j = 1) + f(y | r_j = 0, \theta_j) f(r_j = 0)} = \frac{1}{1 + q_j}, \tag{C.1}
\]

where

\[
q_j = \frac{f(y | r_j = 0, \theta_j) f(r_j = 0)}{f(y | r_j = 1, \theta_j) f(r_j = 1)} = \frac{f(r_j = 0)}{f(r_j = 1)} \prod_{k=1}^{c} \frac{f(y_k | r_j = 0, \theta_{j,k})}{f(y_k | r_j = 1, \theta_{j,k})}. \tag{C.2}
\]

Similarly,

\[
f(r_j = 0 | \theta_j, y) = \frac{q_j}{1 + q_j}.
\]

Next, the conditional likelihoods \( f(y_k | r_j = 0, \theta_{j,k}) \) and \( f(y_k | r_j = 1, \theta_{j,k}) \) will be derived.

Suppose one is at the \((l + 1)\)th Gibbs sampling iteration, and wants to sample \( \gamma_j \) and \( a_{jk} \), the linear regression model given all parameters except \( \gamma_j \) and \( a_{jk} \) can be written as:

\[
y_{k}^* = x_{jk} a_{jk} + e_k,
\]

Where \( y_{k}^* = y_k - X_{1k;jk} \hat{\alpha}_{1k;jk}^{l+1} - X_{jk+1:mk} \hat{\alpha}_{jk+1:mk}^l \).

Given the priors that \( (a_{jk} | \gamma_j) \sim \gamma_j N(0, \sigma_{\alpha k}^2) + (1 - \gamma_j) \delta_0 (a_{jk}) \) and \( (e_k | \sigma_{ek}^2) \sim N(0, \sigma_{ek}^2) \), the conditional likelihood of \( y_k \) can be written as:

\[
f(y_k | r_j = 0, \theta_{j,k}) = (2\pi \sigma_{ek}^2)^{-n_k/2} \exp \left( -\frac{y_k^* y_k^*}{2\sigma_{ek}^2} \right) \tag{C.3}
\]
And

\[
   f(y_k \mid r_j = 1, \theta_{ - r_j}) = (2\pi)^{-n_k/2} \left| V_{y_k} \right|^{1/2} \exp \left( -\frac{y_k^* V_{y_k}^{-1} y_k^*}{2} \right)
\]  

[C.4]

Where \( V_{y_k} = x_{jk} x_{jk}^\prime \sigma_{ak}^2 + I \sigma_{ek}^2 \) is the (co-)variance matrix of \( y_k^* \) given that \( r_j = 1 \).

Then, \( |V_{y_k}| \) and \( V_{y_k}^{-1} \) are derived.

\[
   |V_{y_k}| = |x_{jk} x_{jk}^\prime \sigma_{ak}^2 + I \sigma_{ek}^2| = \sigma_{ek}^{2n_k} \left| x_{jk} \frac{\sigma_{ak}}{\sigma_{ek}} x_{jk}^\prime \frac{\sigma_{ak}}{\sigma_{ek}} + I \right|
\]

By applying Sylvester’s determinant theorem, one has

\[
   V_{y_k}^{-1} = \sigma_{ek}^{-2} (I - \frac{x_{jk} x_{jk}^\prime}{x_{jk}^\prime x_{jk} + \sigma_{ek}^2 / \sigma_{ak}^2})
\]  

[C.5]

It can easily verify that \( V_{y_k}^{-1} \) is:

\[
   V_{y_k}^{-1} = \sigma_{ek}^{-2} (I - \frac{x_{jk} x_{jk}^\prime}{x_{jk}^\prime x_{jk} + \sigma_{ek}^2 / \sigma_{ak}^2})
\]  

[C.6]

Substituting [C.5] and [C.6] into [C.4], one has

\[
   f(y \mid r_j = 1, \theta_{ - r_j}) = (2\pi\sigma_{ek}^2)^{-n_k/2} (x_{jk}^\prime x_{jk} \frac{\sigma_{ek}^2}{\sigma_{ak}^2} + 1)^{-1/2} \exp \left( -\frac{y_k^* y_k^* - (x_{jk} y_k^*)^2 / (x_{jk}^\prime x_{jk} + \sigma_{ek}^2 / \sigma_{ak}^2)}{2\sigma_{ek}^2} \right)
\]  

[C.7]

Denote \( \hat{\mu}_{ak} = \frac{x_{jk}^\prime \left( y_k - X_{1k, jk-1} \hat{\alpha}_{l+1}^{l+1} - X_{jk, jk+1mk} \hat{\alpha}_{jk+1}^l \right)}{x_{jk}^\prime x_{jk} + \sigma_{ek}^2 / \sigma_{ak}^2} \), \( \hat{\sigma}_{ak}^2 = \frac{\sigma_{ek}^2}{x_{jk}^\prime x_{jk} + \sigma_{ek}^2 / \sigma_{ak}^2} \), and substitute [C.7] and [C.3] into [C.2], one gets

116
\[ q_j = \frac{w}{1-w} \sqrt{\prod_k (x'_{jk} x_{jk} \frac{\sigma^2_{jk}}{\sigma^2_{ek}} + 1)} \exp \left( -\frac{1}{2} \sum_k \frac{\mu^2_{jk}}{\sigma^2_{ajk}} \right) \]  

Finally, \( \gamma_j \) can be drawn from a Bernoulli distribution with probability \( \frac{1}{1+q_j} \).
### Appendix D.

Table S1 Estimated posterior model inclusion probability and effects for SNPs close to the region of DGAT1 gene using 28,206 SNPs.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>BTA</th>
<th>Location</th>
<th>Estimated SNP Effect</th>
<th>Posterior Model Inclusion Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Single-task</td>
<td>Multi-task</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Holstein</td>
<td>Ayrshire</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Single-task</td>
<td>Multi-task</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Holstein</td>
<td>Ayrshire</td>
</tr>
<tr>
<td>BTA-34956-no-rs</td>
<td>14</td>
<td>1514056</td>
<td>0.056217</td>
<td>0.000014</td>
</tr>
<tr>
<td>ARS-BFGL-NGS-94706</td>
<td>14</td>
<td>1696470</td>
<td>0.084615</td>
<td>0.000016</td>
</tr>
<tr>
<td>Hapmap52798-ss46526455</td>
<td>14</td>
<td>1923292</td>
<td>0.00548</td>
<td>0.000014</td>
</tr>
<tr>
<td>UA-IFASA-6878</td>
<td>14</td>
<td>2002873</td>
<td>0.010926</td>
<td>0.002001</td>
</tr>
</tbody>
</table>
Table S2: Estimated posterior model inclusion probability and effects for SNPs close to the region of DGAT1 gene using 246,668 SNPs.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>BTA</th>
<th>Location</th>
<th>Estimated SNP Effect</th>
<th>Posterior Model Inclusion Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Single-task</td>
<td>Multi-task</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Holstein  Ayrshire</td>
<td>Holstein  Ayrshire</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pooling</td>
<td>Pooling</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Single-task</td>
<td>Multi-task</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Holstein  Ayrshire</td>
<td>Holstein  Ayrshire</td>
</tr>
<tr>
<td>BTA-34956-no-rs</td>
<td>14</td>
<td>1514056</td>
<td>0.000002 0.000006</td>
<td>0.000002 0.000002 0</td>
</tr>
<tr>
<td>BovineHD1400000187</td>
<td>14</td>
<td>1585385</td>
<td>0.006190 0.000002</td>
<td>0.000005 0.000001 0.000001 0.13416 0.00068</td>
</tr>
<tr>
<td>ARS-BFGL-NGS-94706</td>
<td>14</td>
<td>1696470</td>
<td>0.033223 0.000001</td>
<td>0.019057 0.005697 0.000216 0.67228 0.0007 0.43808 0.14924</td>
</tr>
<tr>
<td>BovineHD1400000239</td>
<td>14</td>
<td>1855090</td>
<td>0.000004 0.000003</td>
<td>0.013883 0.000141 0.000092 0.0007 0.00052 0.2928 0.0059</td>
</tr>
<tr>
<td>Hapmap52798-ss46526455</td>
<td>14</td>
<td>1923292</td>
<td>0.029155 0.000011</td>
<td>0.000185 0.008877 0.00102 0.5541 0.00118 0.00858 0.16784</td>
</tr>
<tr>
<td>BovineHD1400000262</td>
<td>14</td>
<td>1967325</td>
<td>0.028532 0.000004</td>
<td>0.024003 0.029725 0.001134 0.44882 0.00076 0.54828 0.55402</td>
</tr>
<tr>
<td>BovineHD1400000271</td>
<td>14</td>
<td>2002126</td>
<td>0.000359 0.00001</td>
<td>0.055125 0.047684 0.008568 0.01624 0.00106 1 0.8594</td>
</tr>
</tbody>
</table>