ASPECTS OF STATISTICAL MODELLING FOR GENOMIC SELECTION

Klara Verbyla

B.Sc. (Hons) Molecular Biology

The University of Adelaide

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy
June 2010

School of Agriculture and Food Systems
Melbourne School of Land and Environment
The University of Melbourne

ABSTRACT

The research reported in this thesis investigated aspects of statistical models used for genomic selection. The importance of, and, interest in genomic selection is driven by the desire to increase the rate of genetic gain for commercially important traits. Genomic selection could increase the rate of genetic gain by increasing the accuracy of selection through the inclusion of DNA markers.

Multiple methods and models have been proposed for implementing genomic selection. All methods have to overcome the problem that the number of DNA markers (*p*) is typically much larger than the number of phenotypic records (*n*) i.e. the *p*>*n* problem. One approach to this problem is to use Bayesian Inference which allows for an oversaturated model. Two simulation studies and a large data study were undertaken to gain a comprehensive understanding of what makes a robust and accurate Bayesian prediction model. Results from the simulation studies indicated that the match between the assumed QTL distribution and the true QTL distribution had an effect on the accuracy of the direct genomic values (DGV) produced by the different Bayesian models. Some of the models producing accurate DGV were computationally demanding. Subsequently, a novel Bayesian model using Stochastic Search Variable Selection (SSVS) for genomic selection was developed (Bayes SSVS). This model was demonstrated to produce accurate DGV and be computationally efficient.

In contrast to the results from simulated studies, the results from a real dairy cattle data study showed a general equality in the accuracy of prediction across the various Bayesian models including Bayes SSVS. The exception was for traits with atypical genetic architectures such as fat percentage in milk where Bayes SSVS and other model selection approaches performed better than other approaches assuming that all markers equally contributed to the total genetic variation.

The thesis also sought to explore the potential of genomic selection for improving novel traits that have been traditionally very difficult to select for. Energy Balance (EB) is a minimally recorded trait as the cost and measurement logistics mean it can only recorded on experimental farms. Using EB as a case study, it was demonstrated that genomic selection could provide the opportunity to select for EB and other minimally recorded through the accurate prediction of DGV. Additionally, selection for EB could be a valuable tool in finding a balance between production and non-production traits.

Another attractive feature of some of the Bayesian models for genomic selection is they can be used to map QTL. Consequently, the establishment of significance when using multi-locus models for genome wide association studies was explored using a permutation testing approach. Three examples demonstrated that the permutation testing approach could correctly identify QTL. Two specialised approaches, permuting within strata, are presented. One approach accounted for a structured pedigree satisfying the condition of exchangeability. The second approach enabled the identification of secondary moderate QTL in the presence of a major QTL. The effect of the number of permutations needed was also examined; confirming previous results. This method was shown to provide accurate identification of QTL when compared with current approaches.

DECLARATION

i. *The thesis comprises only my original work towards the PhD except where indicated in the Preface,*

ii. *Due acknowledgement has been made in the text to all other material used,*

iii. *The thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices*

Klara Verbyla

PREFACE

Publications arising from this thesis

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Short Note: Accuracy of Genomic Selection using Stochastic Search Variable Selection in Australian Holstein Friesian dairy cattle. Genetic Research 91:307–311 (Chapter 4)

Verbyla, K., P. Bowman, B. Hayes, and M. Goddard. 2010. Sensitivity of genomic selection to using different prior distributions. BMC Proceedings 4:S5 (Chapter 5)

Verbyla, K. L., H. A. Mulder and M. P. L. Calus. Significance testing for whole genome multilocus models using permutation tests. Genetics Submitted (Chapter 7)

Verbyla, K. L., M. P. L. Calus, H. A. Mulder, Y. de Haas and R. F. Veerkamp. Predicting energy balance for dairy cows using high density SNP information. Journal of Dairy Science. 93: 2757-2764 (Chapter 8)

This is to certify that the studies presented in Chapter 4 and 5, which have been published in Genetic Research and BMC Proceedings respectively, were carried out in full by Klara Verbyla. The papers were entirely written by Klara Verbyla. Mike Goddard, Ben Hayes and Philip Bowman were involved in discussions of the results and approved the final papers.

This is to certify that the studies presented in Chapter 7 and 8 that are submitted for publication in Genetics and published in the Journal of Dairy Science respectively, were carried out in full by Klara Verbyla. The study and analyses were conducted by Klara Verbyla and the paper was written in its entirety by Klara Verbyla. Mario Calus designed the software that was used for the analysis and calculation of the DGV. Mario Calus and Herman Mulder were involved in discussions of the study presented in Chapter 7. They, with Roel Veerkamp, Herman Mulder and Yvette de Haas, were involved in the discussions related to the study in Chapter 8. All co-authors approved the final papers.

Klara Verbyla

ACKNOWLEDGEMENTS

Financial support from the Co-operative Research Centre for Beef Genetic Technologies and The University of Melbourne is gratefully acknowledged. In addition, funding from the European Union Sabretain Marie Curie Host Fellowships for Early Stage Training scheme and Wageningen University and Research (WUR) is acknowledged for support during the 11 months spent in the Netherlands on a fellowship at Livestock Research, Wageningen University and Research, during 2009.

Appreciation goes to my supervisory panel of Prof. Mike Goddard, Dr. Ben Hayes (DPI Victoria) and Prof. Richard Huggins. Special acknowledgement goes to Mike and Ben who contributed a significant proportion of their valuable knowledge and time, and provided encouragement and assistance when it was needed. Additionally, I would like to thank Dr. Roel Veerkamp (WUR) and Dr. Mario Calus (WUR) for their contributions, encouragement and assistance.

The resources and expertise made available to me at the Biosciences Research Centre, Department of Primary Industries is gratefully acknowledged. Special thanks to Phil Bowman for all his help.

Others that have given their time, advice and support to help me complete this thesis include Han Mulder, Yvette de Haas, Iona McCloud and Dannielle Hulett. Special thanks also go to my parents, Ari and Janet, whose advice, wisdom and support has been priceless. Finally, to Tash, for your support through all the frustration and success, thank you.

Dedicated in loving memory to my grandmother
Bertha K. Keating

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABREVIATIONS

ABV	Australian Breeding Value
BCS	Body Condition Score
BIC	Bayesian Information Criterion
BLUP	Best Linear Unbiased Prediction
COR	Pearson Correlation Coefficient
DGV	Direct Genomic Value
DIM	Days In Milk
DYD	Daughter Yield Deviations
EB	Energy Balance
EBV	Estimated Breeding Value
GEBV	Genome Enhanced Breeding Value
GWAS	Genome Wide Association Study
LASSO	Least Absolute Selection and Shrinkage operator
LD	Linkage Disequilibrium
MAS	Marker Assisted Selection
MCMC	Markov Chain Monte Carlo
MMSE	Minimum Mean Square Error Estimator
MSE	Mean Square Error
NEG	Negative Energy Balance
PA	Parent Average
PCA	Principal Component Analysis
PLS	Partial Least Squares
QTL	Quantitative Trait Loci
REG	Regression Coefficient
RKHS	Reproducing Kernel Hilbert Spaces
SNP	Single Nucleotide Polymorphism
SP	Sire Pathway
SSVS	Stochastic Search Variable Selection
SVR	Support Vector Regression
TBV	True Breeding Value

CHAPTER 1

Introduction

1.1 THE DEVELOPMENT OF SELECTION TECHNIQUES

Genetic improvement of breeding stock has been the goal of livestock producers for centuries. The emphasis on the selection of superior animals is driven by the desire to both reduce economic costs, for example by reducing health problems or selecting animals with better feed efficiency, and to increase profit by producing animals with a better quality and quantity of the desired product. Over the years, it has become apparent that, in most species, a range of traits need to be selected to obtain animals with a balanced range of desirable characteristics.

The accuracy of the selection for the desired traits determines the amount of genetic gain. This desire to select the superior animals as breeding stock has lead to the refinement and development of selection methods as technology has become available. Traditionally, selection was based entirely on phenotypic characteristics. These traits include reproductive features, weight, body composition and carcass characteristics. Some of these characteristics were evaluated visually, for example for beef cattle, characteristics like anatomical soundness. Other characteristics include different production traits that were identified as significant and quantifiable. Initially, employing these characteristics, an animal's worth was determined using within herd ratios; that is, an animal's relative worth was measured as the difference from the herd average. The contemporary herd comparison, proposed in 1954 (Henderson et al., 1954), followed the use of simple averages based on differences in daughter and dam records (Lush, 1931, Lush, 1933). In each case, the aim was to estimate the unobserved genetic or breeding value.

In 1950, C.R Henderson introduced the idea of best linear unbiased prediction (BLUP) (Henderson, 1950). He referred to the BLUP estimates as "joint maximum likelihood estimates". The BLUP principle was concurrently developed in econometrics (Goldberger, 1962). Henderson then went on to refine the method to enable it to include genetic relationships between animals by including recorded pedigree information and computational methods for implementation (Henderson, 1973, Henderson, 1975a, b, 1976, 1977, 1978).

Computational advances in the 1980s allowed breed associations to start applying the animal model that Henderson had developed. The associations use the BLUP procedure to calculate and report estimated breeding values (EBVs) as a numeric representation of an animal's genetic worth. EBVs are an efficient way to combine heritability information with the performance of relatives and progeny to predict an animal's breeding value. EBVs effectively allow breeders to select and compare animals. The breeding values can now be reported to a base average of more than that of animal's herd or management group. Australian Breeding Values (ABVs) are recorded by the Australian Dairy Herd Improvement Scheme (ADHIS) in a national database containing all Australian Dairy Cattle. Similarly for beef cattle, BREEDPLAN reports EBVs for Australian beef cattle and is also utilised in other countries. Initially only like-treated cattle (same management group) were compared. Subsequently, management groups have been linked via common herd genetics (sires with progeny in more than one management group) allowing cross-herd comparisons.

While substantial genetic gain has been achieved by selection on EBV evident through its implementation across many industries and countries, in recent years the information on the bovine genome has increased dramatically. Single nucleotides polymorphisms (SNP) have become mapped in the hundreds of thousands while, in more recent years, some animals have had their genome completely sequenced (Liu et al., 2009). Consequently, with such an increase in information, considerably more research has been focused on identifying mutations or quantitative trait loci (QTL) affecting economically important traits. The identification of major QTL can be used to select animals with the QTL to increase the genetic gain for a desirable trait.

Despite substantial research on marker assisted selection (MAS) in general, the results have had minimal impact. This is because, as is becoming increasingly clear, most traits are not controlled by QTL with large effects but are affected by many QTL with each explaining only a small amount of the variation seen in the trait (e.g. Chamberlain et al. (2007)). Therefore identifying and selecting on only one or a few

QTL has not produce the results or genetic gain that was initially expected and desired.

Genomic Selection (Meuwissen et al., 2001) is an alternative approach to MAS. The main difference is that MAS seeks to use only a single or small subset of SNPs known to be linked to QTL, whereas genomic selection uses all SNPs at once and therefore can explain more of the variation for the trait of interest. Genomic selection does not attempt to quantify the number of or identify QTL affecting the trait of interest but seeks to capture and maximize the proportion of genetic variance that can be explained by the markers. Genomic prediction, the first step in genomic selection, uses linkage disequilibrium (LD) between the SNP markers and the QTL to capture the true QTL effects and thus the animal's true breeding value. Thus genomic prediction estimates SNP effects in order to approximate the true QTL effects.

The major challenge of genomic prediction is to accurately model the true QTL effects. This challenge is caused by disparity between the large number of SNP markers (p) and the number of records (n) that are available to estimate the SNP effects. This is the well documented *p>n* statistical problem. Any model to be used for genomic prediction must be able to accommodate the *p>n* problem.

Genomic selection is the selection of animals based on their direct genomic value (DGV) for a specific trait. These DGV are estimated breeding values, based only on marker effects and genotypes. The creation of the prediction equation and subsequent estimation of the DGV is termed "genomic prediction" and is the first step in genomic selection. This first step utilises a reference population of animals that have phenotypes, genotypes and a known pedigree. The second step is using the prediction equation to estimate DGV for a set of selection candidates and then to select the best animals based on these genomic breeding values.

The importance of, and, interest in genomic selection is driven by the desire to increase the rate of genetic gain for commercially important traits. Genomic selection provides an approach to increase the rate of genetic gain by through the inclusion of DNA markers. This is caused by a higher accuracy of selection. Additionally, genomic selection allows for the selection of juvenile animals without phenotypes.

This can be exceptionally useful for traits were phenotypic observations are only possible late in life e.g. sex limited traits and slaughter quality traits. This early selection leads to both, a decrease in generation interval and a decrease in the age at first mating. In addition, traits that using current selection techniques have a low accuracy of selection could achieve a higher accuracy of selection through the use of genomic selection. This potentially includes traits that have a low heritability, that are difficult or expensive to measure, and thus are minimally recorded, such as late life and slaughter quality traits, and disease resistance.

1.2 AIMS AND OUTLINE OF THE THESIS

Multiple methods and models have been proposed for implementing genomic selection. Perhaps the most popular approach is Bayesian Inference. This approach allows for an oversaturated and sparse model which is one approach to the *p>n* problem where the number of parameters (*p*) (in this case the SNP effects) to be estimated is greater than the number of records or observations (*n*) available for estimating them. The next chapter reviews the literature and possible models for genomic prediction.

Subsequent chapters (3-6) focus on aspects of and development of Bayesian models and their performance. Bayesian approaches are dependent on the specification of prior distributions. The choice of prior distributions therefore can have a significant impact on the accuracy of the predicted DGV through the prediction of the SNP effects. In the research reported in the Chapter 3, the effect on the accuracy of the predicted DGV caused by varying the prior distributions used for the SNP effects was investigated. The major finding was that the approaches that produced the higher accuracies were the most computationally demanding. Consequently, an additional method, known as Bayes SSVS, was developed. It is presented in Chapter 4. The advantage of Bayes SSVS is that it maintained the assumptions of the original accurate Bayes B (Meuwissen et al., 2001) approach while requiring a less computationally demanding MCMC algorithm. Using real dairy data, this approach was demonstrated to produce accurate results equivalent to the original Bayes B. This approach and the results of this study have been published in Genetics Research (Verbyla et al., 2009).

A second simulation study is presented in Chapter 5. This study used the simulated data from the 13th QTL-MAS workshop where the data structure closely resembled a linkage analysis and again examined the effect of different prior distributions. Bayes SSVS from Chapter 4 was one of three new methods not presented in Chapter 3. Bayes A that was applied in Chapter 3 was also used. This study showed that Bayesian methods that allowed SNP to explain different amounts of variation produced very similar sets of DGV compared to a genomic BLUP approach which assumes that all SNP explain an equal amount of variation. Despite producing a different set of DGV, the Bayesian genomic BLUP approach produced DGV equally highly correlated with the true breeding values (TBV). This indicated that despite its dissimilar assumptions to the other approaches that it may still provide a viable approach to genomic prediction under certain conditions. This second simulation study has been published in BMC Proceedings as part of the publications from 13th QTLMAS workshop (Verbyla et al., 2010a).

Chapter 6 contains a critical overview of a range of Bayesian approaches and examines the accuracy of these approaches when predicting DGV in real data. It examines the robustness of the different models across nine dairy traits with differing genetic architecture. In addition, it examines whether the pre-selection of smaller subsets of SNP is detrimental or beneficial to the accuracy of prediction. For this research, a set of proven bulls with Australian Breeding Values (ABV) for the nine dairy traits were used to predict DGV using Bayes A and Bayes BLUP with three different sets of SNP (two subsets and a set of all SNP). Bayes SSVS was also run to enable further comparison. In addition to DGV, Genomic Estimated Breeding Values (GEBV) were calculated. This enabled the accuracies of the GEBV to be discussed and compared to the outcomes of other studies using real data and conclusions drawn. This study was orally presented at the 2nd Nordic-Baltic Biometric Conference, 2009 (Tartu, Estonia) and the 60th Annual Meeting of the European Federation of Animal Science 2009 (Barcelona, Spain).

Genome wide association studies (GWAS) to identify QTL are still of importance for understanding biological pathways and identification of genes affecting traits. Identification of important biological factors may allow the modification of traits and a better understanding of traits' genetic architectures both of which may increase the accuracy of genomic prediction. The multi-locus models used for genomic selection can also be used for GWAS. Most QTL studies are performed using single marker models despite the fact that the use of multi-locus models overcomes many of the problems associated with estimating the variances, significance and effect sizes of all markers in separate models. The problem of determining significance however can still remain with the use of multi-locus models; to address this, in Chapter 7, a permutation testing approach is presented. The approach allows for the establishment of significance when using multi-locus models for genome wide association studies. Three examples demonstrate how the permutation approach can be used to produce thresholds that allow the declaration of significant major or moderate QTL. Two stratification approaches are presented. One approach was designed to allow for a structured pedigree within the data and to allow the condition of exchangeability to be satisfied. The second approach enables the identification of secondary moderate QTL when a major QTL is present. This study has been submitted for publication in Genetics.

The research covered in Chapter 8 demonstrated that genomic selection could be used to implement selection for novel traits that are typically minimally recorded due to cost or logistical difficulty in implementing recording in progeny testing schemes such as Energy Balance (EB). Energy Balance is generally only recorded on experimental or nucleus farms. Using a small number of Dutch genotyped animals, the accuracy of estimated breeding values predicted using purely polygenic information (EBV) versus utilising all available SNP and polygenic information (DGV) was examined. The use of SNP information showed an increase in the accuracy of prediction for EB over the simple polygenic model. The study showed that in the future, selection for EB could be performed using genomic selection which could provide a valuable tool in finding a balance between production and non-production traits. This result and study has been published in the Journal of Dairy Science (Verbyla et al., 2010b).

In the final chapter (Chapter 9), a summary of the key findings are presented and the implications of the research outcomes are discussed. In addition, possible future directions are considered, indicating the increasing potential of genomic selection, further enhanced by the contribution of this research.

CHAPTER 2
Literature Review

2.1 INTRODUCTION

As the information on the bovine genome has increased, much focus has been on the identification of quantitative trait loci (QTL) and the inclusion of genetic information in selection techniques. Substantial research has been carried out on Marker Assisted Selection (MAS) where selection is based on one or a set of markers. An overview of the principles of MAS and implementation in livestock is provided by Dekkers (2004). The markers can be in linkage with the QTL, in linkage disequilibrium (LD) with it, or it can be based on the causative mutation gene or can be the causative mutation. A few major QTL and genes affecting traits of interest have been identified and thus some possible markers for use in MAS have been found e.g. Grisart et al. (2002). However, it has become apparent that many traits are controlled by a large number of moderate and minor QTL all contributing to small amounts of the genotype and phenotypic variation (Chamberlain et al., 2007). Thus, there has been only limited successful implementation of MAS.

In 2001, Meuwissen et al. published a paper describing a new approach to using marker data for prediction of breeding values called Genomic Selection (Meuwissen et al., 2001). The main difference is that whilst MAS uses only a single or small subset of markers or single nucleotide polymorphisms (SNP), in contrast, genomic selection uses all SNP at once and therefore can explain more of the variation for the trait of interest. Another difference is that genomic selection assumes that the markers are always in linkage disequilibrium with QTL and therefore the markers can be assumed to be the QTL for the purpose of modelling. In contrast, some MAS schemes assume linkage equilibrium between the markers and QTL which means all QTL alleles are treated as different and have to be estimated separately.

Ideally, the true QTL effects could be estimated, however this is not possible. Consequently, genomic prediction and selection seeks to use the LD between the markers and the QTL to model as accurately as possible the true QTL effects. Thus the SNP themselves do not have a causal effect on the phenotype but an apparent effect due to being in LD with the QTL.

When the idea of genomic selection was first introduced, the markers available did not provide sufficient coverage of the genome to enable the markers to capture the QTL effects. However in 2003, sequencing of the bovine genome began as part of the collaborative project proposed in the bovine genomic sequencing initiative white paper (Gibbs et al., 2002). That paper indicated a goal to identify 100,000 SNP for use in identification and mapping of QTL regions. Recent bovine SNP chips such as those containing 10,000 (Affymetrix GnenChip® Bovine Genome Array) and 54,000 SNP (Illumina BovineSNP50 BeadChip) allow sufficient coverage of the genome to begin the process of developing strategies to utilise fully this information in genomic selection. These developments and the large reductions in the cost of the technology have made its application viable. Given the availability of the SNP data at reasonable cost, development of statistical methods to allow accurate prediction of breeding values for selection candidates in breeding programs from this data are critical. The discrepancy between the large number of SNP and the smaller number of phenotypic records provides the challenge if the modelling of the SNP effects is to fully utilize the available information. The following section describes the framework of genomic selection while the methods proposed for genomic prediction and the relevant literature are reviewed in the remaining sections in this chapter.

2.2 IMPLEMENTATION OF GENOMIC SELECTION

As illustrated in Figure 2.1, the success of Genomic Selection relies on deriving an accurate prediction equation for predicting breeding values from marker genotypes. As stated in Chapter 1, genomic breeding values based on animals' genotypes are called direct genomic values (DGV). These DGV can be used to rank and then select the best animals for breeding. Notation for estimated breeding values that utilise marker data is still evolving, but commonly DGV refers to an estimated breeding value from the SNP prediction equation only.

The simplest DGV is the sum of the all SNP effects e.g. $\mathrm{DGV} = \sum_{j=1}^{p} \mathrm{x}_j \hat{\boldsymbol{\beta}}_j$ where the DGV is the direct genomic values calculated for the i^{th} individual, x_j is the indicator variable representing the genotype of the jth marker for the i^{th} individual (x_j=0,1,2), $\hat{\beta}_j$ is the estimated effect associated with marker j. However, the term DGV can also be used when a polygenic effect (based on the pedigree) and the estimated mean are included (Figure 2.1). The term genomic estimated breeding value (GEBV) contains both DGV and traditional pedigree and phenotypic information. The extra information contained in the GEBV extracted from the traditional pedigree and phenotype data is not used in the calculation of the DGV.

The process of creating the prediction equation and then the prediction of DGV is hereafter referred to as genomic prediction. The prediction equation is estimated in a reference population where phenotypes and genotypes exist (Figure 2.1). Once the prediction equation is constructed, the DGV for the selection candidates are directly calculated requiring only an animal's genotype (SNP) information and the prediction equation.

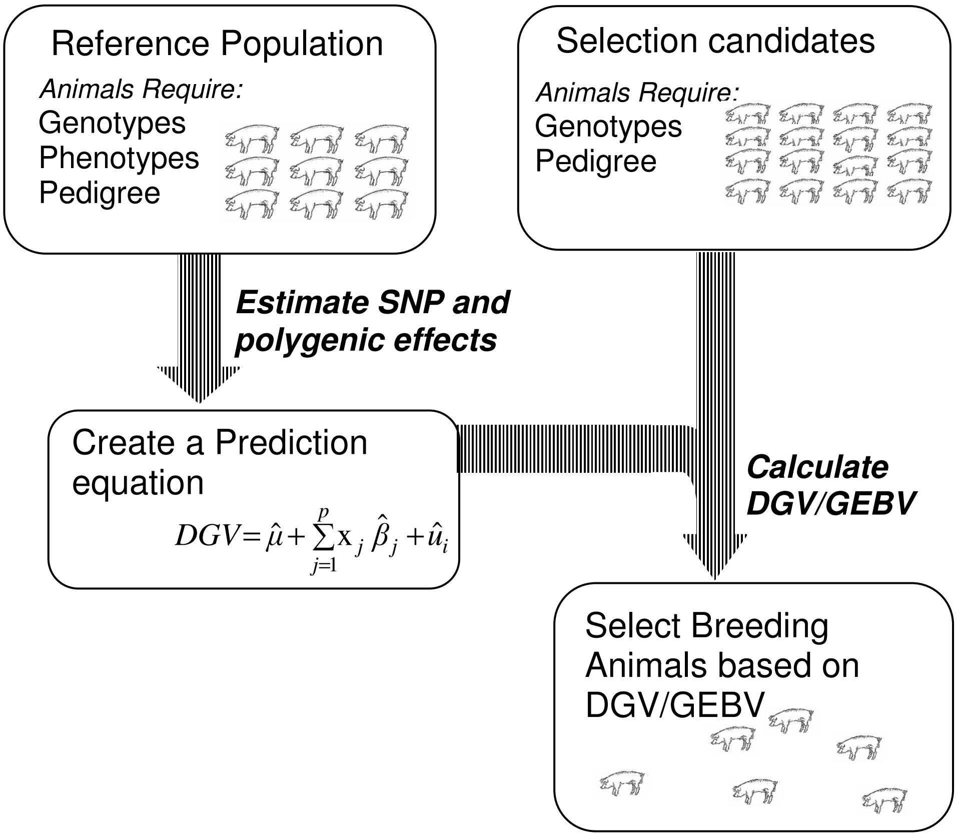
Reference Population
Animals Require:
Genotypes
Phenotypes
Pedigree
Selection candidates
Animals Require:
Genotypes
Pedigree
Estimate SNP and polygenic effects
Create a Prediction equation
$DGV=\hat{\mu}+\sum_{j=1}^{p}x_j\hat{\beta}_j+\hat{u}_i$
Calculate DGV/GEBV
Select Breeding Animals based on DGV/GEBV

Figure 2.1 - Genomic Selection Procedure

The ideal reference population has a number of requirements. Evidence suggests that the accuracy of genomic prediction increases as the number of animals within the reference population increases (Hayes et al., 2009c, Usai et al., 2009, VanRaden et al., 2009). Along with this, the type of animals that are in the reference population is equally as important. In the dairy industry, proven bulls provide an excellent option to form the reference population. This is for two reasons. Firstly reference animals need to have reliable phenotypic information; this can be in the form of recorded phenotypes but also estimated breeding values, deregressed EBVs or daughter yield deviations (DYD). Proven bulls have reliable estimated breeding values based on the phenotypes of many offspring. The second reason is because it has been demonstrated that the prediction equation produces the most accurate DGV when the animals in the reference population are related to the selection candidates (Habier et al., 2007, Habier et al., 2010b). If the prediction equation is to be used across genetically different populations, then animals from each distinct population should be present in the reference population. Additionally, Muir (2007) showed that the accuracy produced by a prediction equation persists for more generations if the reference population contains animals from multiple generations.

With a suitable reference population in place, the critical issue for genomic prediction is the method used to predict the SNP effects and establish the prediction equation. In the following section, the methods that have been proposed for genomic prediction are systematically described; they are categorized according to how they tackle the underlying statistical problem of $p>n$ where the number of markers, p, significantly exceeds the number of phenotypic records, n. The focus of part of this thesis is on aspects related to the performance of Bayesian methods for genomic prediction; it is in the other chapters of this thesis that the performance of the methods which are outlined in this chapter is evaluated.

2.3 STATISTICAL METHODS FOR GENOMIC PREDICTION

Genomic prediction relies on using SNPs located across the entire genome. This means that any statistical method implemented for genomic prediction must be able to simultaneously evaluate marker effects across the entire genome. Consequently the main difficulty to be addressed is how to handle large numbers of markers in a single model, especially when the majority of the time the number of markers, *p,* will significantly exceed the number of phenotypic records, *n* i.e. the statistical problem of *p>n.*

The classical multivariate linear regression problem assumes *p* variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\ldots, \boldsymbol{X}_p$ (where $\boldsymbol{X}_i$ is a *1 x n* vector) and a response vector $\boldsymbol{y}$, again with *n* observations. The linear relationship between the two corresponds to the simplest prediction equation and can be expressed generally as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

where $\boldsymbol{X}$ is the *n x p* design matrix, $\boldsymbol{\beta}$ is the vector of coefficients and $\boldsymbol{e}$ is the residual error also assumed to be normally distributed, $\boldsymbol{e} \sim N\left(0, \boldsymbol{I}\sigma_{\boldsymbol{e}}^2\right)$. The vector of coefficients, β, using classical regression when it is fit as a fixed effect is estimated by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$, thus requiring $p \leq n$. Unfortunately, as the number of SNP available is usually far greater than the number of phenotypic records, this approach is not viable for genomic prediction. This is because there is no unique solution to $\boldsymbol{\beta}$ but many equally good solutions (i.e. the sum of squares equals zero). For instance, one solution is $\boldsymbol{\beta} = \boldsymbol{X}'\left(\boldsymbol{XX}'\right)^{-1}\boldsymbol{y}$. Despite there being possible solutions, a problem arises when these $\boldsymbol{\beta}$ are used in a new sample and the prediction error variance is large.

Hierarchical models which overcome this *p>n* problem perceive the problem as a prediction problem where the effects of all *p* parameters are estimated (i.e. BLUP). Whilst others approach the problem as a model or variable selection problem utilising sparsity, still others use dimension reduction approaches to attempt to establish the original variable or set of variables. The models can also be distinguished by their assumptions or lack of assumptions about the distribution of QTL effects. Those models using model/variable selection and dimension reduction approaches are seeking to remove the noise and identify the important parameters explaining variation. Consequently these models can also be used in QTL studies and genome wide association studies. Another decision is whether to use a linear or non-linear predictor. This leads to the question of whether using non-linear predictors is sensible given that the SNP effects are linearly combined when calculating DGV.

A further question is whether to model the QTL effects as fixed or random. Generally the approaches proposed for genomic prediction, as discussed later in this chapter, choose to fit the effects as random. This is because it allows all SNP to be fitted in a single model. These models include both Bayesian and frequentist approaches. A Bayesian always treats all parameters as random with distributions while frequentists are able to fit both random and fixed effects. Frequentist and Bayesian perspectives differ due to the fundamental beliefs and definition that they attribute to probability. A frequentist sees probability as a long-run frequency and will calculate confidence intervals and construct significance levels. In contrast, Bayesians perceive probabilities to be a measure of the "degree of belief" and using Bayes' theorem (the rules of probability) the belief can be revised given the observed data. Chapter 3 contains an introduction to Bayesian Inference.

2.3.1 Stepwise Regression

An obvious approach to determining a prediction equation is to use the classical least-squares regression in a stepwise procedure. Two different such approaches have been proposed for genomic prediction. Meuwissen et al. (2001) present a two step procedure where each marker location is tested for significance and those that are deemed significant are included in the final fixed regression model to estimate the DGV. Alternatively, Habier et al. (2007) and Moser et al. (2009a) use a forward stepwise regression as described in Kutner et al. (2005). This approach adds and removes markers from the model based on significance until no more markers can be added or removed. Once a suitable first-order linear regression equation is developed this becomes the prediction equation and the DGV are calculated. This approach is not perfect, as ideally all markers would be included simultaneously and their significance and effect calculated concurrently. The way to be able to do this is to treat the SNP or QTL effects, $\boldsymbol{\beta}$, as random effects. All other models proposed for genomic prediction, described hereafter model, the SNP effects as random.

2.3.2 BLUP

The simplest approach to modelling the SNP effects as random is to use BLUP (Best Linear Unbiased Prediction) estimation using random regression. BLUP assumes that each SNP effect is drawn from a normal distribution with a constant variance. This assumption is actually an infinitesimal model but will be a good approximation whenever there are many QTL affecting a trait and none of them have a large individual effect. BLUP assumes the same general model and assumptions as presented earlier for a classical multivariate linear regression problem (Section 2.3) where in addition BLUP assumes that the maker effects $\boldsymbol{\beta}$ come from a normal distribution with a common variance. Thus

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e} \qquad [1]$$

$$\text{where } \boldsymbol{\beta} \sim N\left(0, \sigma_\beta^2\right)$$

The SNP effects, $\boldsymbol{\beta}$ are estimated by solving

$$\left(\boldsymbol{X}'\boldsymbol{RX} + \boldsymbol{I}\lambda\right)^{-1}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \qquad [2]$$

where $\lambda = \sigma_{\boldsymbol{e}}^2 / \sigma_\beta^2$ is constant for all markers and *R* is a diagonal matrix of weights which reflects the reliabilities of phenotypes (***y***) as predictors of breeding value. The diagonal elements of ***R*** can be set to 1 thus ***R=I***, this occurs when no information is available to weight the phenotypes or where all are highly reliable.

Assuming that the markers are dense enough so that the QTL genotypes are completely predictable from the marker genotypes, then the markers will explain the complete genetic variance ($\sigma_{\boldsymbol{a}}^2$). The genetic variance explained by the average marker will therefore be $\mathrm{H}\sigma_\beta^2$ where H is the average heterozygosity of markers. The heterozygosity of a particular marker is $\mathrm{H} = 1 - \sum p_k^2$ where p_k is the frequency of the k^{th} allele and the sum is over all alleles. Under the assumption that each β is drawn from a distribution with constant variance σ_β^2 and β is independent of the H, then $\sigma_\beta^2 = \sigma_a^2 / (H \times n_p)$.

Meuwissen et al (2001) assumed $\mathrm{H}=1$ because their markers were actually haplotypes of two multi-allelic markers. However, assuming that $\mathrm{H}=1$ leads to an underestimate of σ_{β}^{2} and so the estimates of β will tend to be shrunk excessively. When using SNP markers that had only two alleles, Habier et al (2007) used $\sigma_{\beta}^{2}=\sigma_{\boldsymbol{a}}^{2} \Big/ 2\sum_{\boldsymbol{m}} \boldsymbol{p}_{\boldsymbol{m}}\left(1-\boldsymbol{p}_{\boldsymbol{m}}\right)$ where p_{m} is the allele frequency for the m^{th} SNP. Both approaches have been applied in other studies (Moser et al., 2009a, Nielsen et al., 2009, Usai et al., 2009).

The random regression BLUP approach has been shown to be equivalent to the replacement of the additive relationship matrix (A matrix) with the genomic relationship matrix (GRM) in standard mixed models (Fernando et al., 2008, Goddard, 2008, Habier et al., 2007). For example, if the standard animal model is expressed as $\boldsymbol{y}=1_{\boldsymbol{n}}\mu+\boldsymbol{X}\boldsymbol{a}+\boldsymbol{e}$ where $\boldsymbol{a}\sim\boldsymbol{N}\left(0,\sigma_{\boldsymbol{u}}^{2}\boldsymbol{A}\right)$ and $\boldsymbol{e}\sim\boldsymbol{N}\left(0,\boldsymbol{I}\sigma_{\boldsymbol{e}}^{2}\right)$ then the variance of $\boldsymbol{y}$ is $\operatorname{var}(\boldsymbol{y})=\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}'\sigma_{\boldsymbol{a}}^{2}+\boldsymbol{I}\sigma_{\boldsymbol{e}}^{2}$. If $\boldsymbol{A}$ is replaced by $\boldsymbol{G}$ defined as $\boldsymbol{X}\boldsymbol{X}' \Big/ 2\sum_{\boldsymbol{k}}\boldsymbol{p}_{\boldsymbol{k}}\left(1-\boldsymbol{p}_{\boldsymbol{k}}\right)$ then $\boldsymbol{a}\sim\boldsymbol{N}\left(0,\sigma_{\boldsymbol{u}}^{2}\boldsymbol{G}\right)$ and $\boldsymbol{a}$ is equal to $\boldsymbol{X\beta}$ from random regression BLUP [1]. Different approaches have been suggested for creating the GRM (Hayes and Goddard, 2008b, VanRaden, 2008). Neither of these BLUP approaches are concerned with determining the major effects that may reflect the QTL affecting the trait of interest, rather they are interested only in predicting the total genetic value of each animal. Thus this approach treats genomic prediction as a pure prediction problem.

2.3.3 Ridge regression

The BLUP approach discussed in the previous section is a special form of ridge regression (Whittaker et al., 2000). The BLUP solution for the SNP effects [2], is also the ridge regression solution for the SNP effects with λ representing the penalty parameter. The penalty parameter, λ can be found in different ways such as cross validation (Draper and Smith, 1998). Whittaker et al. (2000) suggest testing a range of λ and choosing the λ that minimises the model error.

2.3.4 Bayes A

An alternative to assuming a normal distribution of SNP effects is to assume a distribution with fat tails e.g. a t-distribution. This means that like BLUP all SNP are assumed to have some effect, however, the fat tails of the assumed distribution allow for the assumption that some of the SNP are in linkage disequilibrium with QTL of moderate to large effect. This assumption can be incorporated in a Bayesian model where the prior distribution for the SNP effects has a hierarchical structure. The SNP effects sampled from a normal distribution with the variance for each SNP sampled from an inverse scaled chi square distribution (or the analogous inverse gamma) as follows:

$$\beta_i \,|v_i \sim N(0, v_i)$$
$$v_i \sim \chi^{-2}(r, s) \sim \gamma^{-1}\left(\frac{r}{2}, \frac{rs}{2}\right)$$

where r is the degrees of freedom and s is the scale parameter. This formulation means that the SNP effects are really being sampled from a student-t distribution. This is evident through the definition of the student t-distribution as probability distribution of the ratio; $Z/\sqrt{V/t}$ where Z is the normally distributed $Z \sim \boldsymbol{N}(0,1)$ and *V* has a chi square distribution with *t* degrees of freedom. The expected mean on the distribution is zero and the variance is defined as $\text{var}(\boldsymbol{\beta}) = \frac{\boldsymbol{t}}{\boldsymbol{t}-2}$.

This model and formulation was termed Bayes A by Meuwissen et al. (2001). The formulation of Bayes A means that each SNP will have some effect. The shape of the distribution that the SNP effects are sampled from is dependent on the degrees of freedom used for the inverse scaled chi square distribution (and subsequently the t-distribution). The values of the inverse scaled chi square distribution parameters, *r* and *s,* can be found for a random variable X, from the mean: $E(X) = \frac{rs}{(r-2)}$ and the variance: $\text{var}(X) = \frac{2r^2s^2}{(r-2)^2(r-4)}$. Combining the two expressions gives: $\frac{\text{var}(X)}{E(X)^2} = \frac{2}{(r-4)}$. Thus using the expected mean and variance, the hyper-parameters *r* and *s* can be set for the inverse scaled chi square distribution.

The variance for the t-distribution (inverse scaled chi-square distribution) is undefined for *t*<2 (*r*<4) and the mean is improper for *t*<1 (*r*<2). As the degrees of freedom increase for the t-distribution, the distribution resembles a normal distribution. Thus with low degrees of freedom (for the inverse scaled chi-square distribution and the resultant t-distribution), this framework allows for the majority SNP to have only minor effect with a few having larger, more major, effects. For some traits, this distribution with fatter tails (see Figure 2.1) may have a better approximation to the real distribution of QTL effects than sampling the SNP effects from a normal distribution (Schaeffer, 2006).

The t-distribution has been used more widely than other fat-tailed distributions because it is possible to sample directly from the posterior distribution when the data are normally distributed by using the hierarchical structure. The general form for the inverse scaled chi square conjugate prior and posterior distributions for the SNP effects are:

$$\pi\left(\sigma^2_{\beta_j}\right) \sim \chi^{-2}(r,s) \text{ prior}$$

$$\boldsymbol{post}\left(\sigma^2_{\beta_j}\right) \sim \chi^{-2}\left(r+n, \frac{rs+\beta_j\beta_j}{r+n}\right) \text{ posterior}$$

In the formulation of Meuwissen et al (2001), the values of r and s were calculated so that the prior distribution had the same mean and variance as an estimated distribution of QTL effects (eg. Hayes and Goddard (2001)) and *n* takes the values of 1.

Xu (2003) and ter Braak et al. (2005) applied a Bayesian method analogous to that of Bayes A with alternative prior for the SNP variances. Xu (2003) set v and s to zero to give an uninformative prior where *n* again takes the value of 1 i.e. $\boldsymbol{p}\left(\sigma^2_{\beta_i}\right) \sim \left(\sigma^2_{\beta_i}\right)^{-1}$. ter Braak et al (2005) presented an extension of the proposed prior in Xu (2003) and report that to ensure a valid posterior the prior should be $p\left(\sigma^2_{\beta_i}\right) \sim \left(\sigma^2_{\beta_i}\right)^{-1+\delta}$ where $0 < \delta \leq 0.5$ yielding a posterior where $n = 1 - 2\delta$. All these priors are conjugate in nature and thus all can be implemented using the Gibbs Sampler, a Markov Chain Monte Carlo (MCMC) sampling algorithm (see Section 3.1.2 for introduction to MCMC sampling algorithms) allowing direct sampling from the posterior distribution. The prior and posterior distributions are summarised in Table 2.1.

Table 2.1- QTL effects variances prior and posterior distributions

	Prior Distribution	*Posterior Distribution*
Meuwissen et al. (2001)	$\pi\left(\sigma^2_{\beta_i}\right) \sim \chi^{-2}(v, s)$	$post\left(\sigma^2_{\beta_j}\right) \sim \chi^{-2}\left(v+1, \frac{vs+\beta_i\beta_i}{v+1}\right)$
Xu (2003)	$\pi\left(\sigma^2_{\beta_i}\right) \sim \left(\sigma^2_{\beta_i}\right)^{-1}$	$post\left(\sigma^2_{\beta_j}\right) \sim \chi^{-2}(1, \beta_i\beta_i)$
te Braak et al. (2005)	$\pi\left(\sigma^2_{\beta_i}\right) \sim \left(\sigma^2_{\beta_i}\right)^{-1+\delta}$	$post\left(\sigma^2_{\beta_j}\right) \sim \chi^{-2}\left(1-2\delta, \frac{\beta_i\beta_i}{1-2\delta}\right)$

Xu (2003) presented results using real barley data that indicate that multiple marker Bayesian analysis analogous to Bayes A gives much clearer results compared to individual marker regression analysis in a Genome wide association study. Whittaker et al. (2000) introduced the idea of marker-assisted selection using ridge regression (this is analogous to BLUP assuming a constant variance for all SNP), but even they acknowledge that having dense markers in the model produces serious co-linearity. Xu (2003) also showed that ridge regression was not viable for entire genome scans using simulated data as it estimated small effects across the simulated genome and failed to find any large effects (as would be expected using BLUP). However, Gianola et al. (2003) present a hierarchical method using ridge regression from a Bayesian perspective. That paper however presented only the theoretical aspects and there has been no subsequent simulation or application of the model published.

2.3.5 Bayes B

Another possible assumption for the SNP effects is that many of the SNP are in genomic regions where there are no QTL and thus have zero effects, whilst a small proportion of SNP are in LD with QTL and consequently do have an effect. Reflecting this assumption, Meuwissen et al (2001) present Bayes B. This alternative approach assumes that the majority of the SNP effects are exactly zero and only a proportion (1- π) of all SNP have a non-zero effect; those that are non zero have an individual variance using the identical prior to that used in Bayes A.

Consequently the SNP effects were sampled from a normal with the variance sampled with probability π from a bulk at zero and 1- π from the inverse scaled chi square distribution, as originally formulated in Meuwissen et al (2001) is expressed as:

$$\beta_i \mid \boldsymbol{v}_i \sim \boldsymbol{N}\left(0, \sigma^2_{\beta_i}\right)$$

$$\sigma^2_{\beta_i} = 0 \text{ with probability } \pi$$

$$\sigma^2_{\beta_i} \sim \chi^{-2}(r, s) \text{ with probability 1-}\pi$$

This can alternatively be written as:

$$\beta_i \mid \sigma^2_{\beta_i} \sim (1-\pi)\boldsymbol{I}_0 + \pi N\left(0, \sigma^2_{\beta_i}\right)$$

$$\sigma^2_{\beta_i} \sim \chi^{-2}(r, S) \sim \gamma^{-1}\left(\frac{r}{2}, \frac{rS}{2}\right)$$

where $\boldsymbol{I}_0$ is a point mass at zero. This formulation is suggested as more appropriate from a Bayesian Inference perspective (Gianola et al., 2009), however both will produce the same t- distribution for the SNP effects.

This hierarchical structure means that those effects that are non-zero can be thought of as those in stronger LD with the QTL. In fact, if the number of times a SNP is included in the model (i.e. has a non-zero effect) is recorded, the posterior probability of that SNP being linked to a QTL can be calculated.

There are two issues with the use of Bayes B. The first is that π has to be predetermined. If a value which is inconsistent with the true distribution of SNP effects is chosen, the accuracy of the DGV could be negatively affected. To overcome this, a method for sampling this proportion has been presented by Fernando (2009). The proposed approach placed a uniform prior on π ($\pi = uniform(0,1)$) and it is sampled along with all other parameters during MCMC iterations. Once convergence is reached, the parameter is set to the mean of its posterior distribution and the algorithms are run again to estimate the SNP effects. In general, π is set to reflect the expected proportion of SNP in linkage disequilibrium with QTL relative to the total number of SNP.

Another potential difficulty with Bayes B is that the hierarchical priors specified means the priors are not conjugate and thus cannot be sampled using a Gibbs

Sampler. (This is unlike in Bayes A where all parameters posterior distributions can be directly sampled using the Gibbs Sampler.) Meuwissen et al (2001) use a single site updating Metropolis Hastings algorithm where each element of β (the SNP effects) and $\sigma^2_{\beta_i}$ (the SNP variance) are updated individually. The use of the mixture distribution means as the prior distribution also could mean that the dimensionality of the model is changing as the number of SNP included in the model varies. However by setting the value of π the dimensions of the model remain constant. In situations where the dimensions are dramatically changing (where π is also being sampled) the reversible jump MCMC algorithm (Green, 1995) is needed to communicate across all possible models and their differing dimensionality according to the proper acceptance ratio. The reversible jump MCMC algorithm essentially generalizes the Metropolis-Hastings algorithm and consequently the Metropolis Hastings is in fact a special form of the reversible jump MCMC algorithm.

Despite these two issues, Meuwissen et al. (2001) demonstrated in simulated data that both Bayes A and Bayes B implemented using MCMC could be applied successfully to simultaneously estimate all SNP effects across the entire genome. They also showed in simulated data that the Bayesian regressions outperformed a least squares forward stepwise approach and the genomic BLUP approach. Bayes B produced more accurate DGV than Bayes A on the simulated data which was attributed to assumed prior distribution matching the simulated distribution of QTL effects. These results are likely to be dependent on the model used to simulate the data which more closely matched the assumptions of Bayes A and B rather than the other approaches trialled.

2.3.6 LASSO

The previously mentioned Bayesian approaches; Bayes A and Bayes B, assume that any non-zero QTL effects are sampled from a t-distribution. Alternatively, the double exponential distribution is another possible distribution for the SNP effects. The double exponential distribution has long tails like the t-distribution but has a larger number of small non-zero effects (see Figure 2.2). LASSO (least absolute shrinkage and selection operator) approach (Foster et al., 2007a, Tibshirani, 1996, Vach et al., 2001) uses a double exponential for the distribution of the QTL effects when formulated in a Bayesian framework. The LASSO estimates can be derived as the Bayes posterior mode under independent double exponential priors for the QTL effects (Tibshirani, 1996). The double exponential can also be expressed as a mixture distribution of normal distribution with variance sampled from an exponential distribution.

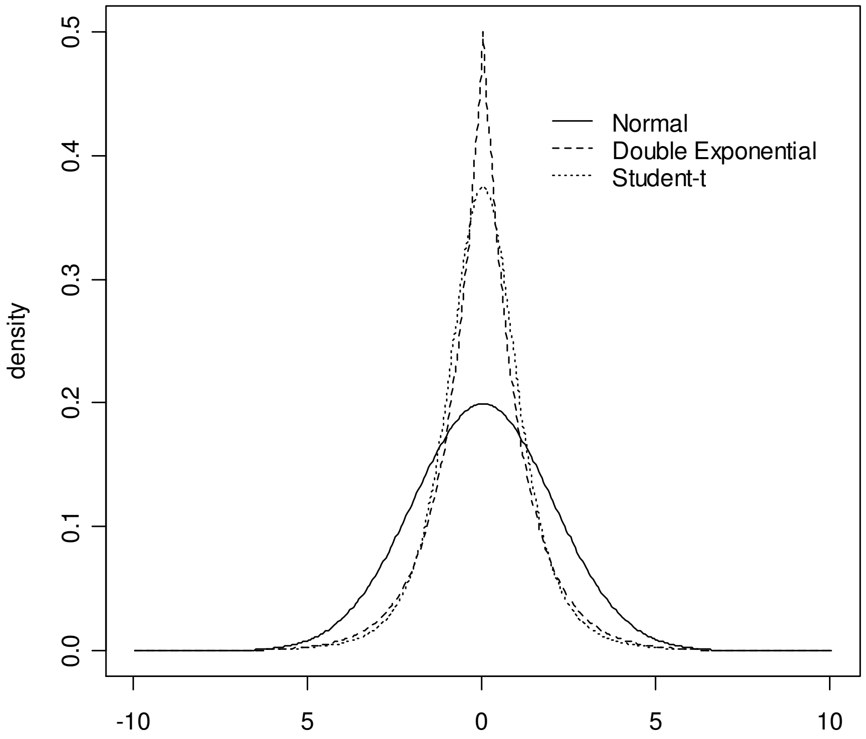

Figure 2.2- Comparison of prior distributions for the SNP effects. All distributions have a mean of 0 and variance of 2 (N(0,2) ,DE(1), t(4)). Normal used for BLUP, t-distribution used for Bayes A and B, and Double Exponential used for LASSO.

The LASSO was first proposed by Tibshirani (1996) as a technique that combines the strengths of subset reduction and ridge regression by setting some variables to zero and shrinking others. The LASSO model can be expressed as a linear random model and can also be incorporated into a linear mixed model (Foster et al., 2007a, Foster et al., 2007b). As the LASSO is a variance reduction and variable selection approach and it has the potentially advantageous feature of only including a subset of SNP in the final predictive model. This means that the LASSO is appropriate for 'sparse' models, where ridge regression is unlikely to succeed (as ridge regression forces all coefficients to be non-zero). The lasso is a form of penalized least squares that minimizes the residual sum of squares while controlling the $L1$-norm of the coefficient vector β. The estimates are the set of SNP that satisfy

$$argmin_{\beta}\left\{\sum_{i=1}^{n}\left(\boldsymbol{y}_i-\sum_{j=1}^{p}\boldsymbol{x}_{ij}\beta_j\right)^2\right\} \text{ subject to } \sum_j\left|\beta_j\right|\leq t$$

which alternatively can also be written as:

$$\hat{\beta}=argmin_{\beta}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})+\lambda\|\boldsymbol{\beta}\|_1$$

where t is the constraint parameter choose prior to estimation, $\lambda \geq 0$ is a Lagrange multiplier, which relates implicitly to the bound or shrinkage parameter *t* and controls the degree of shrinkage.

The LASSO is commonly implemented using LARS (Least Angle Regression), a model selection algorithm that allows the implementation of a stepwise approximation to LASSO (Efron et al., 2004). The use of LARS means the computationally demanding quadratic programming can be avoided (Tibshirani, 1996). Usai et al., (2009) employ LARS to estimate the SNP effects for the set of SNP deemed as significant for genomic prediction. Akin to the Bayesian methods, the LASSO's tuning or shrinkage parameter, *t*, needs to have a value determined. Tibshirani (1996) suggest using cross-validation, generalised cross validation or an analytical unbiased estimate of risk to estimate this parameter. Usai et al., (2009) used a cross-validation approach to determine the value for this parameter in order to predict genomic breeding values.

An alternative implementation of the LASSO is to use Bayesian Inference (Hans, 2009, Park and Casella, 2008). Tibshirani (1996) suggested that LASSO estimates could be interpreted as posterior mode estimates when assigning independent and identical double-exponential (Laplace) priors to each β_j. The advantage of this formulation is that a prior can be set to estimate the hyper (shrinkage) parameter or a product of this hyper-parameter. Yi and Xu (2008) like Park and Casella (2008) set a gamma prior on the squared hyper parameter and sample it directly from the posterior distribution using the Gibbs Sampler. They also both suggest using the posterior median gives the closest estimate to the LASSO estimates.

2.3.7 Non and Semi Parametric regression

Another variable selection approach to genomic prediction was presented by Gianola et al. (2006, 2003, 2008). They introduce the idea of using semi and non-parametric approaches for genomic prediction including a semi-parametric kernel mixed model, reproducing kernel Hilbert spaces (RKHS) regression and kernel regression. The main difference between these approaches and those previously presented is that they use non-linear regression. The advantages of these approaches are that no strong assumptions need to be made about the distribution of the parameters. Thus the relationship between y and x can expressed as

$$y_i = g(\boldsymbol{x}_i) + e_i \qquad i = 1,2,\ldots.,n$$

where y_i is the phenotypic value for the i^{th} individual, $\boldsymbol{x}_i$ is the vector of quantified genotypes for the i^{th} individual (also called the information set), $g(.)$ is some unknown function relating genotypes to phenotypes and e_i is the i^{th} residual term. $g(.)$ maps from the information set, in this case the SNP genotypes $\boldsymbol{x}_i = \{x_1,\ldots\ldots x_n\}$, to evaluations of the conditional expectation function, $g(\boldsymbol{x}_i) = E(y_i \mid \boldsymbol{x}_i)$. In kernel regression, a kernel is used as a weighting function in the estimation of $g(.)$. e.g. the Nadaraya-Watsib estimator

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}$$ where K is a kernel with a bandwidth h.

Kernel regression is a form of local weighted regression where given data $(\boldsymbol{X},\boldsymbol{Y})$ the aim is to find a regression function $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ such that the function best fits the original data. The idea is that the kernel is a set of identical weighted functions that assigns weight to each new data points based on distance from the original data point. The kernel functions depend only to the radius, width or variance from the data point, X_i, to a set of neighbouring locations, x. Consequently, the kernel model can be expressed as:

$$\boldsymbol{y} = 1_{\boldsymbol{n}} \mu + \sum \beta_i K(x,\ X_i)$$

where K is the kernel. The most commonly used kernel is the Gaussian (normal) kernel, defined as:

$$K(x,X)=\exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{X}\|^2}{2\alpha^2}\right]$$

where α is the bandwidth that determines the amount of smoothing. As shown, these approaches do require specification of the form of the kernel, the bandwidth (which determines the amount of smoothing) and the loss function (see below).

In RKHS regression the problem can be expressed as:

$$\hat{g}(\boldsymbol{x})=\underset{g\in H}{arg\ min}\left\{l(y,g(\boldsymbol{x}_i),\boldsymbol{x})+\lambda \parallel g \parallel_H^2\right\} \qquad [3]$$

where $l(y,g(\boldsymbol{x}_i),\boldsymbol{x})$ is the loss function, H is the Hilbert space and λ is the smoothing parameter. Due to function $g(.)$ not being given a parametric form, $\lambda \parallel g \parallel_H^2$ is included as the penalty term where $\parallel . \parallel_H$ is the norm in Hilbert space, *H*. For more details see de los Campos et al. (2009), Gianola and van Kaam (2008) and Gianola et al. (2006). Gonzalez-Recio et al. (2008) applied the non-parametric methods for genomic prediction to mortality records of broilers.

An alternative approach for genomic prediction implemented by Moser et al (2009) is to use support vector regression (SVR). SVR is a supervised learning method which is a machine learning technique. In fact, SVR is a specific algorithm of RKHS with an altered objective function. In [3] the loss function, $l(y,g(\boldsymbol{x}_i),\boldsymbol{x})$, uses the quadratic loss function in RKHS but replaced in SVR with the epsilon-sensitive loss function i.e.

RKHS: $l(y,g(\boldsymbol{x}_i),\boldsymbol{x})=(y-g(\boldsymbol{x}_i))'(y-g(\boldsymbol{x}_i))$

SVR: $l(y,g(\boldsymbol{x}_i),\boldsymbol{x})=|y-g(\boldsymbol{x}_i)|_\varepsilon=\begin{Bmatrix} 0 & if\ |y-g(\boldsymbol{x}_i)|\le\varepsilon \\ f\ |y-g(\boldsymbol{x}_i)|-\varepsilon & otherwise \end{Bmatrix}$

This difference changes the system from one in which the coefficients are found from a linear model to a quadratic programming problem (Moser et al., 2009a). For more information on SVR see Smola and Schölkopf (2004), Moser et al. (2009) and Vapnik (1998).

2.3.8 Principle Component Analysis and Partial Least Squares Regression

Another dimension reduction approach that has been extensively applied to *p>n* problems across many disciplines is partial least square regression (PLS) and principal component analysis (PCA). Both techniques are an extension of the multiple linear regression model and seek to reduce the dimensionality of the set of variables by finding combinations of the original predictors. Although often bracketed together PCA and PLS are actually very different. PLS tries to extract the latent factors accounting for as much of the variation as possible while modelling the responses as well. A tutorial on PLS introducing the basics of PLS is provided by Geladi and Kowalski (1986). PCA seeks to reduce the dimensions of the model by transforming a number of possibly correlated variables into a smaller number of uncorrelated variables which are the principal components. Smith (2002) provide a more thorough description of the principles of PCA. There is no physical interpretation of the principal components based on SNP. It is impossible to relate a combination of SNP on different chromosomes to the prediction of a specific QTL. Thus a DGV based on principal components produced by PCA is likely to be unreliable. However, PCA is reported to capture population structure (McVean, 2009). Solberg et al.(2009) use both PCA and PLS for genomic prediction while Moser et al. (2009) use PLS as one of five approaches tested for accurate genomic prediction. It is to be noted that PCA and PLS have no distributional assumptions only mild assumptions associated with being an extension of multiple linear regression model and the form of the input data.

2.3.9 Genetic Algorithms

Carlborg et al. (2000) present an approach to multiple QTL mapping using Genetic algorithms. Genetic algorithms are search algorithms and are a particular class of evolutionary algorithms. Genetic algorithms can be used to explore the vast set of possible models and find an approximate best model. They are non-linear predictors. Crump et al. (2007) applied a genetic algorithm to genomic prediction using a Bayesian Information Criterion (BIC) as the fitness criterion to find the best model. BIC is a criterion used for model selection seeking to find the best model with the lowest number of parameters.

2.3.10 Comparative performance of methods

The performance of many of the previously mentioned methods has been tested for both simulated and real data. In general, those methods that assume unequal variance and make reasonable assumptions about the distribution of QTL effect outperform other methods in simulated data (Meuwissen et al., 2001); this is also demonstrated in Chapter 3. However, as real data has become available, different trends have emerged. These are discussed in Chapter 5 and 8. In general, most approaches for genomic prediction (except least squares regression) that have been applied to real data have produced very similar results with the exception for traits that have a major QTL explaining a large amounts of genetic variation. When this occurs, approaches such as Bayes B that assume unequal variances for the SNP produce higher accuracies of prediction.

Much of the focus of the following chapters is on the use of Bayesian methods; however their performance is evaluated and discussed in context relative to many of the different statistical models for genomic prediction. A comprehensive comparison of the performance of the Bayesian methods is presented in Chapter 5 including discussion of how these methods perform relative to other approaches. In Chapter 8, a consolidated discussion of the performance of currently applied methods including the results from previous chapters is presented. Additionally in Chapter 8, the focus is shifted to what future methods and models may be required as the genomic information increases in the form of the number of SNP and complete sequencing of animals.

2.4 MULTI-LOCUS MODELS FOR GENOME-WIDE ASSOCIATION STUDIES

In addition to the significant focus in genomic prediction and selection, the availability of dense marker SNP panels has also lead to an increase in genome-wide association studies aiming to identify QTL (Goddard and Hayes, 2009, Hardy and Singleton, 2009, McCarthy et al., 2008). Most quantitative traits are complex traits with numerous genetic factors contributing to the genetic variation and identifying these factors could be beneficial for biomarker identification, marker assisted selection and identification of possible drug targets. Many of the statistical models highlighted for genomic prediction can be used for genome wide association studies using multi-locus models.

The advantage of fitting all markers over using the traditional single marker model is the avoidance of the problems of biased results. Results may become biased through the fitting of a single QTL (marker or interval) in a model that may be affected by the presence of other QTL not fitted in the model. This may confound the results and cause false positives (that is a significant QTL is found where there is in fact not a QTL), false negatives (that is no QTL is found where there is actually a QTL) and reporting of incorrect levels of significance and size. A further problem is caused by the multiple estimates of the residual variance leading to problems when calculating the total phenotypic variance. In addition, the total variance explained by the QTL has to be calculated from the estimates from different models which can lead to estimates of total variance that are too high.

Model or variable selection approaches such as forward stepwise least-squares regression, the LASSO, PLS and PCA already provide subsets of SNP thought to explain amounts of genetic variation and thus be linked to QTL. Other approaches like Bayes A have non-zero effects for all the SNP and thus would require some threshold to be set to enable determination of which SNP are linked to QTL. Bayes B would also require thresholds to be set to determine significance as SNP may be included in the model during only a proportion of the MCMC iterations thus resulting in a posterior probability of less than 1.

The limitation of all multi-locus approaches is that due to p>>n, the ability to distinguish significant QTL is determined by the available data. Donoho and Stodden (2006) showed that as the number of non zero parameters got closer to the number of observations, the performance of model selection methods decreased. They also found that the greater the difference between *p* and *n,* the lower the ability to recover the underlying model. They tested forward stepwise selection, the LASSO approach using the original quadratic programming (Tibshirani, 1996) and LARS, the stepwise approximation to the LASSO. As the number of non-zero parameters (*k*, where *k*<*p*) included in the model increased towards the number of records (*n*), the ability to recover the underlying sparse model was shown to be significantly reduced. In fact, the forward stepwise algorithm never improved the ability to recover the underlying model once the number of non-zero parameters exceeded twenty percent of the number of records. The addition of a False Discovery Rate threshold to the forward step selection algorithm lead to similar results to that produced by LARS and the LASSO. The ability to successfully identify QTL is therefore generally bound by not only the number of phenotypic records but the number and size of the QTL associated with the trait of interest. The ability to correctly model the true distribution of the QTL affecting the trait of interest will also have a significant impact on the ability to identify QTL.

Thus, once adequate data is available the only issue remaining is the establishment of significance. Setting of thresholds is generally arbitrary with values less than a predetermined value set to zero and deemed insignificant. One formal approach to setting significance thresholds is to use a permutation approach (Churchill and Doerge, 1994, Doerge and Churchill, 1996). A novel permutation approach for multi-locus models used in GWAS is presented in Chapter 7. This approach was demonstrated using an analogous model to that of Bayes SSVS presented in Chapter 4 with simulated and real data sets.

2.5 CONCLUSION

In this chapter, the range of methods that have been proposed for genomic prediction, have been systematically overviewed. That overview has been structured around the approach to the *p*>*n* problem, how the methods deal with the modelling of SNP effects and, then in the case of the random modelling of the SNP effects, the assumption made about the distribution of non-zero SNP effects. The overview of the methodology is summarized in Figure 2.3. Most of the approaches seek to find the best model using model selection techniques or reduce the dimensionality by variable selection. Only Bayes A and BLUP predict an effect for all SNP effects. However, the structure of the Bayes A prior distributions (Table 2.1) means that it seeks to shrink most insignificant SNP effects back to very close to zero, particularly if the degrees of freedom for the inverse scaled chi square distribution are small.

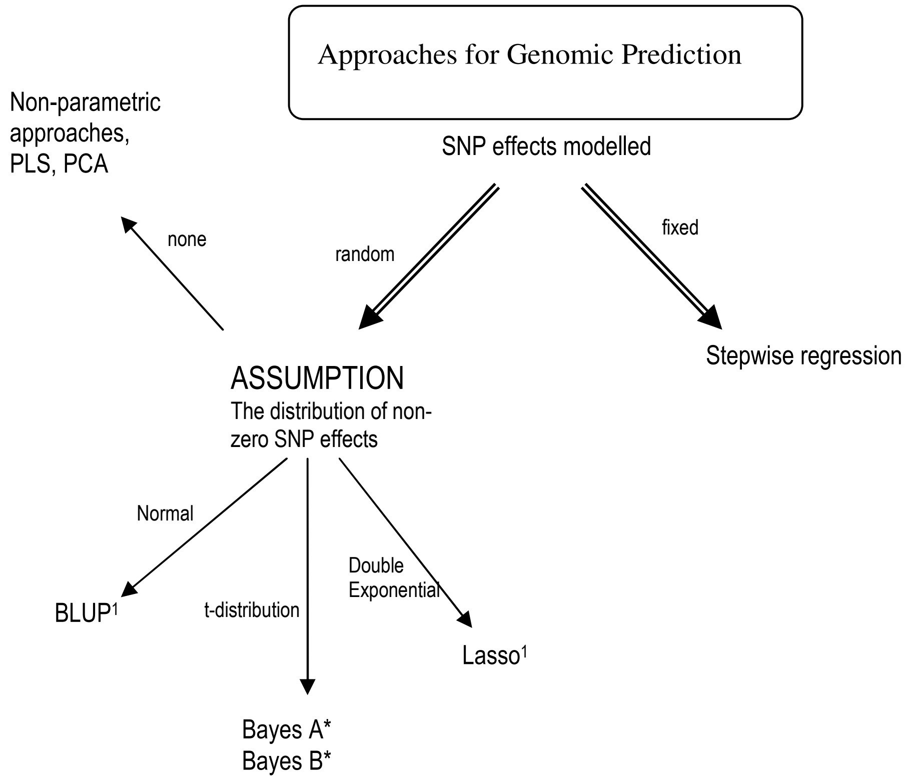

Figure 2.3- Overview of the assumptions for the different genomic prediction approaches. *Bayesian Approach, ^{1}Can be viewed as a Bayesian Approach

The most significant difference between the approaches occurs due to the modelling of the QTL effects. The non-parametric methods do not, by definition, make any assumptions about the distribution of the QTL effects. Both PLS and PCA seek to reduce the dimensions of the model and also make no assumptions about the QTL effect distribution. In stepwise regression where the QTL effects are modelled as fixed and consequently no assumptions are also made about the distribution of the SNP effects. All other approaches make assumptions about the modelling of the QTL/SNP effects. The different hierarchical models and their assumptions for the QTL effects are summarised in Table 2.2.

The final difference is between the use of linear and non-linear predictors. The question remains whether a non-linear prediction approach is appropriate when a linear model is extracted from the results and used as the prediction equation. Further discussion of this and relevant results is provided in the general discussion (Chapter 8). While some indication of the performance of the various methods has been provided, the following chapters are focussed on exploring Bayesian methods for genomic prediction, in particular the affect of different prior distributions for the SNP effects on the accuracy of the predicted DGV and different approaches to modelling these assumptions. The reasons for this focus is that previous results of Bayesian approaches to genomic prediction have been promising, but further investigation into the effect of different prior distributions on the performance of these models is needed for a better understanding of what makes a robust and accurate model.

Table 2.2- Hierarchical modelling of SNP effects

Prior Distribution	*Definition*
BLUP	$\beta_i \sim N\left(0, \sigma^2\right)$
Bayes A	$\beta_i \mid v_i \sim N(0, v_i)$ $v_i \sim \chi^{-2}(r, s) \sim \gamma^{-1}\left(\frac{r}{2}, \frac{rs}{2}\right)$
Bayes B	$\beta_i \mid v_i \sim N(0, v_i)$ $\boldsymbol{v}_i = 0$ *with probability* π $\boldsymbol{v}_i \sim \chi^{-2}(\boldsymbol{r}, \boldsymbol{s})$ *with probability* 1-π
LASSO	$\beta_i \mid v_i \sim N(0, v_i)$ $v_i \sim exp(\lambda)$

A brief introduction to the principles of Bayesian statistics, optimum selection theory and a review of possible prior distributions and relevant literature are provided in Chapter 3. In that chapter, the role and impact of different prior distributions on the accuracy of DGV are explored in data simulated under a simplistic model. The results of this Chapter lead to the development of the Bayesian approach for genomic prediction presented in Chapter 4. Stochastic Search Variable Selection is utilised in a Bayesian model to derive the accurate DGV in significantly less time than comparable methods. Chapter 5 presents a second simulation study where this new model with Bayes A and two other different models not used Chapter 3 were applied to data simulated as part of the 13th QTLMAS workshop.

A thorough comparison of different Bayesian methods for genomic prediction on real data is presented in Chapter 6. This study allowed for the comparison of different models using real data which is important for truly assessing the value and usefulness of different methods. The models were compared across a range of traits with differing genetic architecture. Also within this chapter, the impact of using pre-selected reduced sets of SNP is investigated, as well as the difference in accuracies between DGV and GEBV. The chapter ends with a discussion of the findings in the context of other published results and the implications of this comparison are discussed.

Chapter 7 presents an original permutation approach for use with Bayesian multi-locus models to enable the establishment of significance QTL in genome-wide association studies (GWAS). An example of a possible application of genomic prediction for use with minimally recorded traits is presented in Chapter 8. Energy balance in lactating dairy cattle is a minimally recorded trait but could be an important link between production and non-production traits. In Chapter 8, genomic prediction is shown even with a small data set for the difficult trait of energy balance to produce higher accuracies than a traditional pedigree BLUP approach.

Finally in Chapter 9, the major results are reviewed and impact of the increased availability of SNP information in the future on the suitability of the various methods is also considered.

CHAPTER 3
Performance of Bayesian models with Varying Prior Distributions for Genomic Prediction

3.1 INTRODUCTION

Before the introduction of genomic selection, Bayesian methods were proposed for use in QTL detection to analyse the linkage between markers and quantitative trait loci (QTL) (Hoeschele and Vanraden, 1993a, b, Satagopan et al., 1996, Thaller and Hoeschele, 1996a, b). The idea was based on the fact that other methods such as linear fixed regression, random regression and maximum likelihood all depend on the number of markers and that the QTL effects associated with the selected makers is always overestimated. Consequently, in a situation such as genomic prediction and selection, where the number of markers is large and far outweighs the number of phenotypic records, a Bayesian approach is suggested as ideal. Analogous approaches are available from a frequentist perspective fitting all effects as random (see Section 2.3 for a discussion on the difference between frequentists and Bayesian perspectives). Appreciating the advantage of Bayesian methods, Meuwissen et al. (2001) implemented two Bayesian approaches for genomic prediction in the original genomic selection paper.

The aim of this chapter was to assess the performance of Bayesian models with varying prior distributions for use in genomic prediction. The chapter begins with a brief introduction to Bayesian inference is presented in the next section, followed by a succinct review of the relevant MCMC sampling algorithms (3.1.2). Subsequently the importance of the prior distribution is explored in relation to the optimum selection criterion (3.1.3) and literature relevant to the choice of prior distribution for the SNP effects is examined (3.1.4). Then, a small data set is used to compare the performance of a range of prior distributions when using a Bayesian model for genomic prediction.

3.1.1 Bayesian Inference

A Bayesian approach is based on Bayes Theorem. Let θ be a random variable and $\boldsymbol{y}$ be the data, then Bayes Theorem states:

$$\pi(\theta \mid \boldsymbol{y}) = \frac{\mathrm{f}(\boldsymbol{y} \mid \theta)\pi(\theta)}{\mathrm{f}(\boldsymbol{y})} \qquad [3]$$

where $\pi(\theta \mid \boldsymbol{y})$ is the conditional or posterior distribution of the random variable given the data, $\mathrm{f}(\boldsymbol{y} \mid \theta)$ is the likelihood of the data given the random variable, $\pi(\theta)$ is the prior distribution of the random variable and $\mathrm{f}(\boldsymbol{y})$ is the normalising constant found by integrating out θ. The normalising constant can be left out, changing [3] into

$$\pi(\theta \mid \boldsymbol{y}) \propto \mathrm{f}(\boldsymbol{y} \mid \theta)\pi(\theta) \qquad [4]$$

In this Bayesian framework, all parameters are treated as random variables. Consequently, each variable in the model has a distribution. The parameters are divided into observables and unobservables. Observables are the parameters that can be recorded like the phenotypic data (real or simulated) and the genotypes. The unobservables are the parameters we want to estimate, for example, the SNP effects and their associated variances (Section 2.2). The distribution for the unobservables is known as the prior distribution $\pi(\theta)$ and must be specified. The distribution of the observables conditional on the random variable θ is known as the likelihood, $\mathrm{f}(\boldsymbol{y} \mid \theta)$ and is a function of the unobservables. The purpose of Bayesian analysis is to find the conditional or posterior distributions $\pi(\theta \mid \boldsymbol{y})$ of the parameters given the observed data. This distribution is dependent on the likelihood, $f(\boldsymbol{y} \mid \theta)$ and the prior distribution, $\pi(\theta)$ [4]. Thus if θ is the SNP effects, then the choice of the prior distribution for θ is important as it will effect the posterior distribution and thus the estimation of the SNP effects.

3.1.2 Markov Chain Monte Carlo sampling algorithms

It is often difficult to solve [2] directly consequently sampling approaches are used to establish the posterior distributions. Markov Chain Monte Carlo (MCMC) sampling algorithms are a class of algorithms for sampling a “parameter from approximate distributions and then correcting those draws to better approximate the target posterior distribution” (Gelman et al., 2003). The parameters are sampled sequentially where the new state is only dependent on the previous state, thus forming a Markov Chain. The quality of the approximate target posterior distribution improves as a function of the length of the chain. The ultimate aim is therefore the convergence of the chain to the target distribution.

The most popular MCMC sampling method is the Gibbs Sampler (Geman and Geman, 1984). Its popularity is due to the fact that the posterior values can be directly sampled from the conditional distribution giving at the end a set of variables that represent the target posterior and consequently it is generally quicker. However it can only be used when conjugate prior distributions are employed. A prior distribution is defined as a conjugate when, given a set likelihood, the posterior distribution and the prior distribution are from the same family of distributions and the posterior has a known form. For example, the inverse scaled chi- square distribution is a conjugate distribution for the variance under a normal likelihood, such that, the prior and posterior has the general definition (r and s as defined in Section 2.3.4):

$$\pi\left(\sigma_j^2\right) \sim \chi^{-2}(r,s) \qquad \textit{(prior)} \qquad [5]$$

$$post\left(\sigma_j^2\right) \sim \chi^{-2}\left(r+n, \frac{rs + \sum_{i=1}^{n}(x_i - \mu)^2}{r+n} \right) \qquad \textit{(posterior)} \qquad [6]$$

Thus when a conjugate prior distribution is used, the posterior distribution is known and generally it is possible to sample from it.

When a conjugate prior is not used, then the posterior distribution usually cannot be directly sampled and must be constructed. In this instance, a Metropolis Hasting algorithm can be used. The Metropolis Hastings Algorithm implemented in this study is often known as the ‘independent single site updating’ Metropolis-Hastings Algorithm (Gelman et al., 2003). In this special case, each element (within a vector such as a vector of SNP effects) is updated separately and the candidate states are generated by a distribution that is independent of the current state of the chain. The same characteristics as the traditional Metropolis-Hastings are maintained; in that, the distribution does not have to be symmetric unlike the stricter Metropolis Algorithm. The Metropolis Hastings algorithm uses multiple iterations to create the posterior; in each iteration, a new proposed state is rejected or accepted over the current state based on an acceptance ratio using their respective likelihoods.

3.1.3 Point Estimation and the Optimum Selection Criterion

Using the appropriate MCMC algorithm, the posterior distribution of the SNP effects $\boldsymbol{\beta}$ given the data, ***y***, can be established, $\pi(\boldsymbol{\beta} \mid \boldsymbol{y})$. The choice of the point estimate for each SNP effects could be the median, mean or the mode (other point estimates are also possible) of the posterior distribution. When calculating a point estimate (or making a decision), Bayesians want to minimize the expected loss of a decision rule under the prior distribution $\pi(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$. Different loss functions can be used. The mean of the posterior distribution is the most commonly used as it provides the minimum mean square error (MMSE) estimator. This means that the mean square error is used as the loss function. The use of different loss function will result in the median and mode of the posterior distribution. Due to its simplicity, the MMSE estimator is used for the point estimates of all parameters (i.e. the mean of the posterior distribution is used).

Further support of the use of MMSE estimator for the SNP effects is provided by the optimum selection criterion. The traditional estimated breeding values (EBV) produced using BLUP (Chapter 1) can be expressed as the conditional mean of the unobservable true breeding values $\boldsymbol{u}$ given the observed data $\boldsymbol{y}$, that is $\hat{\boldsymbol{u}} = E(\boldsymbol{u} \mid \boldsymbol{y})$ where $\hat{u}$ denotes the EBV. Notably in a Bayesian framework $E(\boldsymbol{u} \mid \boldsymbol{y})$ can be interpreted as the mean of the posterior distribution of $\boldsymbol{u}$ given y, thus BLUP can fit into a Bayesian formulation.

Direct genetic values (DGV- Section 2.2) for genomic selection are calculated by summing over the SNP effects β_j. Generally, $\hat{u}_i = \hat{\mu} + \sum_{j=1}^{p} \mathrm{x}_{ij} \hat{\beta}_j$ where $\hat{u}_i$ is the estimated genomic breeding value for the i$^{\text{th}}$ individual, μ is the overall mean, x_{ij} is the indicator variable representing the genotype of the jth marker for the ith individual (x_{ij}=0,1,2), βj is the size of the QTL effect associated with marker j (j=1,..p) .

Consequently, to maintain the optimum criterion for selection, the best predictor for the SNP effects is:

$$\hat{\boldsymbol{\beta}} = E(\boldsymbol{\beta} \mid \boldsymbol{y}) \qquad [7]$$

Thus, the best predictor of the SNP effects is the conditional mean of the posterior distribution of SNP effects given the phenotypic records y . Consequently, throughout this thesis where Bayesian methods are employed the MMSE estimator using the mean of the posterior distributions are used to provide point estimates of the SNP effects.

Expanding [7] gives $\hat{\boldsymbol{\beta}} = \frac{\int \boldsymbol{\beta} \times f(\boldsymbol{y} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta})d\beta}{\int f(\boldsymbol{y} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta})d\beta}$ where $\boldsymbol{\beta}$ are the true QTL effects, $\pi(\boldsymbol{\beta})$ is the prior distribution (Bayesian) or the random variable distribution (frequentist) of the QTL effects, and $f(\boldsymbol{y}|\boldsymbol{\beta})$ is the likelihood. If the QTL effects are assumed to have a normal prior with a constant variance matrix i.e. $\pi(\boldsymbol{\beta}) \sim \boldsymbol{N}(0, \sigma_{\beta}^2)$, the estimates are BLUP. Consequently, the accuracy of the estimated SNP effects and optimum selection can be seen to be dependent on the specification of the prior distribution $\pi(\boldsymbol{\beta})$.

3.1.4 Prior Distributions

The specification of prior distributions can also accommodate differing assumptions such as that a large number of markers or chromosome segments have a zero or close to zero effect or, in contrast, that all SNP have a small effect. This is done by allocating an appropriate prior distribution for the size of the SNP effects and their variances. Consequently, the distribution of QTL effects has been examined (Hayes and Goddard, 2001, Weller et al., 2005).

Hayes and Goddard (2001) performed a meta-analysis. They assumed that the QTL effects had a gamma distribution and investigated the parameters using published

QTL effect estimates for pig and dairy data. Many papers have reported the L-shaped distribution for QTL effects (see Figure 3.1) so consequently the assumption of a gamma distribution seems reasonable (Bost et al., 2001, Bost et al., 1999, Edwards et al., 1987, Jorge et al., 2005, Mackay, 2001, Wu and Li, 2000). Weller et al. (2005) examined nine traits in dairy cattle and showed that the different traits had different QTL effects distributions. They found that some traits (protein percentage and fat percentage) had L-shaped gamma distributions. Others, such as fertility and herdlife, had bell-shaped gammas while the other traits examined had left skewed bell shaped gammas (see Figure 3.1). These results suggest correctly that different traits have varying genetic architecture and thus may require different prior distributions. The robustness of different prior distributions is therefore an interesting issue; this is explored in Chapter 6.

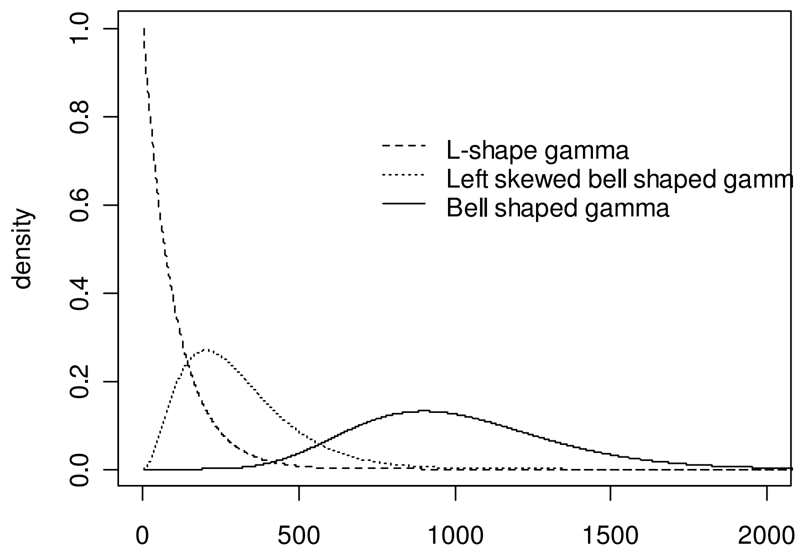

Figure 3.1- Different types of Gamma distributions found to represent QTL distribution for various traits

Xu (2003) performed QTL analyses using a Bayesian approach with markers across the entire genome and subsequently also discovered that the gene effects followed an L-shaped gamma distribution for all traits analysed. However, they also acknowledged that the gamma distribution is not a conjugate prior with a normal error distribution and consequently Gibbs sampling (Geman and Geman, 1984) cannot be used. This means the application of the Metropolis-Hastings algorithm (Gilks et al., 1996, Hastings, 1970, Metropolis et al., 1953) is needed which is less efficient than Gibbs sampling but still effective.

While the results of Hayes and Goddard (2001), Xu (2003) and Weller et al. (2005) indicate that the distribution of gene/QTL effects may be assumed to follow a gamma distribution, there is some criticism of the use of the gamma distribution as the prior in Bayesian QTL analyses and genomic prediction. Gianola et al (2003) criticise the use of the gamma prior because the effects sampled would all be strictly positive but the estimates can be positive or negative. The major problem is however the computationally demanding algorithms that would need to be used. Consequently, most Bayesian approaches to QTL analyses and genomic prediction utilize a normal prior distribution for the size of the QTL/gene effects (Gianola et al., 2003, Meuwissen et al., 2001, ter Braak, 2006, ter Braak et al., 2005, Wang et al., 2005, Xu, 2003, Yi, 2004, Yi and Xu, 2002, Yi et al., 2005). However, the variances used for the normal distribution can be sampled from different distributions. Meuwissen et al. (2001), Xu (2003), and ter Braak et al. (2005) all sample individual SNP variances from an inverse scaled chi-square distribution. This is analogous to an inverse-gamma distribution. The benefit of this formulation is an inverse scaled chi-square distribution is a conjugate distribution and the computationally efficient Gibbs Sampler can be used. This formulation also results in the SNP effects being sampled from a t-distribution. A t-distribution has fat tails allowing for a few larger effects, and unlike the gamma distribution, it no longer has the problem of being strictly positive.

Another option is presented by Kiiveri (2003) and applied in his program GeneRaVE. A normal-gamma prior is used which can be expressed as:

$$\begin{aligned} \beta_i \mid v_i &\sim N(0, v_i) \\ v_i &\sim \gamma(\lambda, \varpi) \end{aligned} \qquad [8]$$

It has a mixture distribution which has a normal distribution for size of the QTL effects and a gamma distribution for variance of the size of the QTL effects. While this offers the desired shape, it has an infinite spike at zero for $\lambda \leq 1/2$. Thus, the distribution has the normal-Jeffreys as the limiting density form. The Jeffreys prior is a non-informative prior distribution that is proportional to the square root of the determinant of the Fisher information (Gelman et al., 2003) and requires no pre-selection of a hyper-parameter. It is based on the principle that the prior density should remain constant despite re-parameterisation.

The choice of parameters for the gamma distribution [8] can have a severe impact as $\lambda \to 0$ where the limiting form has infinite mass, an infinite spike at zero and flatness for large values of $|\beta_i|$. Consequently it does not penalise large values and can strongly influence the modal behaviour of the posterior. Thus the choice of parameters for the gamma needs to be considered when using this choice of prior distribution. Additionally, under a normal likelihood the gamma is not a conjugate prior and cannot be sampled with the Gibbs Sampler.

The LASSO (Section 2.2.5) provides another option for the choice of prior for the size of the QTL effects. The estimates from this approach can be thought of as the Bayes posterior mode under independent double exponential (DE) priors for the QTL effects. Therefore, the assumption is that the QTL effects come from a double exponential distribution with mean zero. It creates the shape of the gamma (L-shaped) while no longer having the problem of being strictly positive. Its use allows the distribution to have a large frequency of SNP effects close to zero as well as being centred on zero and having both negative and positive values. The double exponential can also be expressed as a normal-exponential mixture model; thus this prior is a special case of the normal-gamma prior distribution with $\lambda = 1$.

The double exponential and normal gamma prior distributions can be expressed as mixture distributions (Griffin and Brown, 2005). Griffin and Brown (2005) also introduce the idea of a normal exponential gamma (NEG) and exponential gamma (EG) distributions.

$$\begin{aligned} \text{NEG} \quad & \beta_i \,|\, v_i \sim N(0, v_i) \\ & v_i \sim \exp(\lambda) \\ & \lambda \sim \gamma(\alpha, \beta) \\ \text{EG} \quad & \beta_i \,|\, \lambda \sim \exp(\lambda) \\ & \lambda \sim \gamma(\alpha, \beta) \end{aligned}$$

The advantage of the NEG is that, unlike other distributions, it has a finite limit at zero for all parameters in range (unlike the gamma as $\lambda \to 0$ where the limiting form has infinite mass at zero) and incorporates both of the limiting cases for the double exponential and normal-Jeffrey's cases. This is suggested by Griffin and Brown (2005) to be superior as a prior. However this highly hierarchical prior framework, means that more parameters have to be sampled. It still requires the hyper-parameters of gamma to be set correctly since choosing an inappropriate prior would negatively affect the QTL effects posterior distribution. Subsequently, this formulation also appears more computationally and time intensive.

Different prior distributions have been used in Bayesian models proposed for genomic selection. Meuwissen et al (2001) originally presented two Bayesian hierarchical models with different prior distributions assuming unequal variances across the SNP. These were called Bayes A and Bayes B and are introduced in Sections 2.2.3 and 2.2.4 respectively. The main difference between the two approaches is the assumptions about the QTL effects. The specification of Bayes A means it assumes that all SNP have some effect. Conversely, Bayes B assumes only some SNP have an effect where the other SNP have no effect and have an effect size of zero. Xu (2003) and ter Braak et al. (2005) present alternative prior distributions for the SNP effect variances for an analogous model to Bayes A. These are described in Section 2.2.3 and summarised in Table 3.1 in Section 3.2.2.3.

The objective of this study was to examine the effect that the use of different prior distributions had on the accuracy of predicted DGV for genomic prediction using Bayesian models. In this study, Bayes A and Bayes B as described originally in Meuwissen et al. (2001), and Normal-Exponential and Normal-Gamma mixture distributions are used as the hierarchical prior distributions for the SNP effects.

3.2 MATERIALS AND METHODS

3.2.1 Simulated Data

The SNP data was simulated using a mutational-drift model where the mutation rate was assumed to be 2.5 x 10^{-5} per locus per generation. The population was assumed to have an effective population size of 100 in the first 100 generations. The final set of phenotyped animals was created by crossing 50 sires and 40 dams so that each dam had fifty offspring, one with each sire. The reference and validation data sets consisted of 100 randomly selected animals (from the 2000 offspring) with 250 polymorphic markers on a single chromosome of 250cM to maintain the real world condition of $p>n$ (where the number of markers (p) was greater than the number of records/observations (n)).

Table 3.1- Summary of simulated data

	Data set 1	*Data set 2*	*Data set 3*
Number of Loci with an Effect	14	14*	0
Genetic Variance Explained	2.8% or 11.4%	2.5%-25%	0

*explaining 2% or more of the genotypic variance

The three data sets were simulated using different simulated QTL distributions. For all data sets, the size of the QTL effects were treated as fixed effects and simulated as in Xu (2003) with positive and negative effects. The SNP simulated as the QTL was assumed to be the causative mutation and was removed from the genotype data for estimation for the first simulated data set. A summary of the effects for the three data sets are presented in Table 3.1.

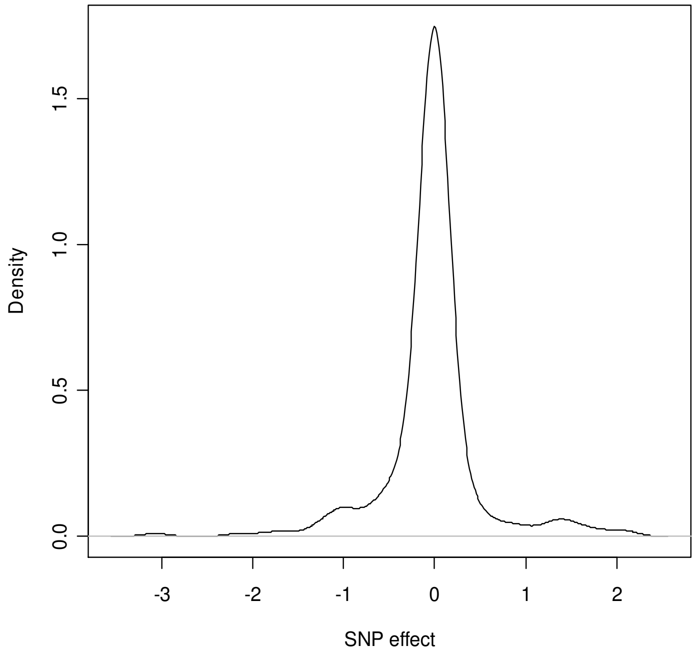

Figure 3.2 -Distribution of QTL effects for simulated dataset 2

The first data set had 14 QTL simulated with discrete values of ±1 and ±2. The effects in the second data set were sampled from a double exponential (continuous) distribution with total variance of 1. A total of 50 QTL were simulated; however only

20 had an effect over 0.5 and only 14 explained each more that 2 percent of the genetic variance (Table 3.1). This simulated data set was to reflect the results that studies such as Xu (2003) had found which indicated that most SNP effects distributions followed an L-shape gamma distribution; the double exponential produced very similar features to the L-shape gamma reflected about zero.

The third data set had no QTL effects. The phenotypic data for all data sets was obtained by adding an error term that was normally distributed with mean 0 and variance 1. This simulated data for a trait with heritability of 0.5. No polygenic, epistatic or imprinting (maternal or paternal) effects were simulated.

3.2.2 Model

At each SNP (total number of SNP, p) there are three possible combinations of two alleles (e.g. A or B), the homozygote of one allele (AA), the heterozygote (AB) and the homozygote of the other allele (BB). These are then quantitatively represented by 0, 1 and 2 respectively. Subsequently, the phenotypic models used were

$$y = \mu 1_n + \sum_{j}^{p} X_j \beta_j + e$$

where y is the vector of phenotypes of the trait being analysed for all n individuals, μ is the mean, 1_n is a vector of ones of length n, X$_j$ is a vector of indicator variables representing the genotypes of the jth marker for all individuals (x_{ij}=0,1,2), β_j is the size of the QTL effect associated with marker j and e is the residual error normally distributed as $e \sim N\left(0, \sigma_e^2 \mathbf{I}_n\right)$ where $\mathbf{I}_n$ is the n x n identity matrix .

3.2.3 Prior Specification and Iterative Algorithms

The Markov Chain Monte Carlo (MCMC) algorithms utilized during this study were the Gibbs Sampler and the Metropolis Hastings Algorithm (as described in Section 3.1.2). The Gibbs Sampler was used to directly sample the posterior distributions of the mean, the QTL effects and the error variance (Table 3.2).

Table 3.2 - Summary of Prior and Posterior Distributions for the Mean, SNP effects and residual variance.

Parameter	*Prior Distribution*	*Posterior Distribution*
Mean	$\pi(\mu) \propto 1$	$post(\mu) \sim N\left[1_n y - 1_n Xb, \sigma_e^2 / n\right]$
SNP effects	$\pi(\beta_j) \sim N(0, \sigma_{\beta_j}^2)$	$post(\beta_j) \sim N\left(\frac{[X_{ij}' y - X_{ij}' X\beta_{(ij=0)} - X_{ij}' 1_n \mu]}{[X_{ij}' X_{ij} + \lambda_i]}, \frac{\sigma_e^2}{[X_{ij}' X_{ij} + \lambda_i]}\right)$ where $\lambda_i = \sigma_e^2 / \sigma_{\beta_j}^2$ and $\beta_{(ij=0)}$ is the QTL effects with the ith individual and jth marker set to zero
Residual Variance	$\sigma_e^2 \sim \chi(-2,0)$	$post(\sigma_e^2 \mid e) \sim \chi^{-2}\left(n-2, \frac{e^T e}{n-2}\right)$

The difference between the models is based on the specification of the prior used for the variance of the SNP effects. All models used hierarchical prior distributions for the SNP effects that can be defined as normal mixture distributions. This is clearly shown in Table 3.3 where the different approaches are defined by the specification of the prior distribution for the variance of the SNP effects. The values for the respective hyper-parameters were set so that the total genetic variance equalled 1; these values are also shown in Table 3.3.

The MCMC algorithms needed to sample the variance of the SNP effects were the Gibbs Sampler for Bayes A and the Metropolis Hastings algorithm for the remaining three approaches. The inverse scaled chi square distribution used in Bayes A is conjugate under a normal likelihood and thus the posterior had a known form and could be directly sampled. (See Section 3.1.2 for the prior and posterior forms for an inverse scaled chi square distribution). The priors used by Bayes B, normal-gamma and normal-exponential mixture models involve non-conjugate prior distributions for the variance of the SNP effects. Consequently, as the posterior has no known form, the Metropolis Hastings algorithm is applied to sample from the unknown posterior distribution to enable construction of the posterior distribution.

Table 3.3- Hyper-parameter Settings

Prior *Distribution*	*Parameter specification*	*MCMC Algorithm* *(and posterior if known form)*
Bayes A	$\beta_i \mid \sigma^2_{\beta_i} \sim N\left(0, \sigma^2_{\beta_i}\right)$ $\sigma^2_{\beta_i} \sim \chi^{-2}(4.011, 0.008)$	Gibbs Sampler $post\left(\sigma^2_{\beta_i}\right) \sim \chi^{-2}\left(5.011, \frac{0.032 + \sum_{i=1}^{n}(x_i - \mu)^2}{5.011}\right)$
Bayes B	$\beta_i \mid \sigma^2_{\beta_i} \sim N\left(0, \sigma^2_{\beta_i}\right)$ $\sigma^2_{\beta_i} = 0$ with probability 0.95 $\sigma^2_{\beta_i} \sim \chi^{-2}(4.234, 0.17)$ with probability 0.05	Metropolis Hastings Algorithm
Normal-Exponential	$\beta_i \mid \sigma^2_{\beta_i} \sim N\left(0, \sigma^2_{\beta_i}\right)$ $\sigma^2_{\beta_i} \sim \exp(100)$	Metropolis Hastings Algorithm
Normal-Gamma	$\beta_i \mid \sigma^2_{\beta_i} \sim N\left(0, \sigma^2_{\beta_i}\right)$ $\sigma^2_{\beta_i} \sim \gamma(0.00176, 0.176)$	Metropolis Hastings Algorithm

3.2.4 Sampling Sequence

An MCMC sampling scheme was utilised to sample all the parameters. Thus, the sampling sequence that was implemented was as follows: (The posteriors sampled from are shown in Table 3.2 and Table 3.3.)

1. Initialize all unobservable and denote by

$$Q^{(0)} = \left[\mu^{(0)}, \beta_1^{(0)}, \beta_2^{(0)} ......\beta_p^{(0)}, \sigma_e^{2(0)}, \sigma_{\beta_1}^{2(0)}, \sigma_{\beta_2}^{2(0)} ......\sigma_{\beta_j}^{2(0)}\right]$$

2. Update each variable, namely:
   - Update the mean, μ.
   - Update the size of the QTL effects, b_j, *j=i,...p.*

     When a conjugate prior distribution is not being used, the Metropolis Hastings Algorithm is required as follows:

     - Sample $\beta_{j(new)}$ from the prior distribution $p(\beta_j)$
     - Replace current β_j by $\beta_{j(new)}$ with a probability

$$\alpha = \min\left(\frac{p(y^* \mid \beta_{j(new)})}{p(y^* \mid \beta_j)}, 1\right)$$

where $p(y^* | \beta_{j\,(new)})$ is the likelihood of the data given $\beta_{j(new)}$ and y^* is the data adjusted for the mean and all other genetic affects except β_j.

  - o Repeat

- Update the error variance σ_e^2.
- Update the variance of the size of the QTL effects, σ_j^2, $j=i,...p$.

3. Repeat step 2, until convergence to a stationary distribution has occurred.

Five separate MCMC chains with different starting seeds were used to assess variability of the methods and to confirm reliability of the results. Each was run for 10,000 cycles with the first 1000 discarded as burn in. Five replicated were run fro each data set.The programming was in R.

3.2.5 Direct Genomic Values (DGV)

The estimated DGV for each animal was found as the sum of the mean and the SNP effect i.e. $\text{DGV} = \mu \boldsymbol{1}_n + X\beta$. The accuracy of the estimated DGV was established by comparing the true breeding value (TBV) and predicted DGV using regression coefficient, Pearson correlation coefficient and mean square error (MSE). The regression coefficient was the true value regressed on the predicted. MSE was calculated as standard as $\boldsymbol{MSE} = \sum_{i=1}^{n} (\boldsymbol{u}_i - \hat{\boldsymbol{u}}_i)^2 / \boldsymbol{n}$ where n is the number of animals.

3.3 RESULTS

3.3.1 Data Set 1

The results for the accuracy of prediction for each method (and standard errors) across the five replicates are shown in Table 3.4. The figures for all models of the actual TBV versus the predicted DGV for the reference and validation populations are shown in figures 3.3-3.6. All methods were extremely accurate for the reference population as expected as this is the population where the QTL effects are estimated. The different models produce less accurate DGV in the validation population but with no clear difference in Pearson Correlation coefficient between the models. In contrast, both the normal-gamma and normal-exponential mixture prior distributions produced the lowest MSE indicating a lower bias and regression coefficients closest to one. The standard errors and figures clearly show that the Bayes A and B hierarchical priors are more variable across replicates. The model with the normal exponential hierarchical prior was the least variable between the replicates as shown in Figure 3.6 and Table 3.4 with the smallest standard errors.

Table 3.4 - Results of Data Set 1. Validation and Reference population results comparing predicted DGV and true breeding values (TBV) using regression coefficients (Reg) of true regressed onto predicted, Pearson correlation coefficient (Cor) and mean square error (MSE).

REFERENCE POPULATION

	Bayes A	Bayes B	NE	NG
Reg	1.0080 ± 0.0003	1.0096 ± 0.0004	1.0407 ± 0.000009	1.0176 ± 0.00003
Cor	0.9914 ± 0.0006	0.9955 ± 0.0002	0.9942 ± 0.000003	0.9957 ± 0.000005
MSE	0.2182 ± 0.0148	0.1179 ± 0.0053	0.1674 ± 0.0002	0.1148 ± 0.0001

VALIDATION POPULATION

	Bayes A	Bayes B	NE	NG
Reg	0.8841 ± 0.0146	0.8868 ± 0.0150	0.9617 ± 0.0002	0.9208 ± 0.00005
Cor	0.8907 ± 0.0126	0.8983 ± 0.0098	0.8978 ± 0.00007	0.9069 ± 0.00009
MSE	4.0390 ± 0.6602	4.8914 ± 0.7797	3.0489 ± 0.0089	2.8230 ± 0.0142

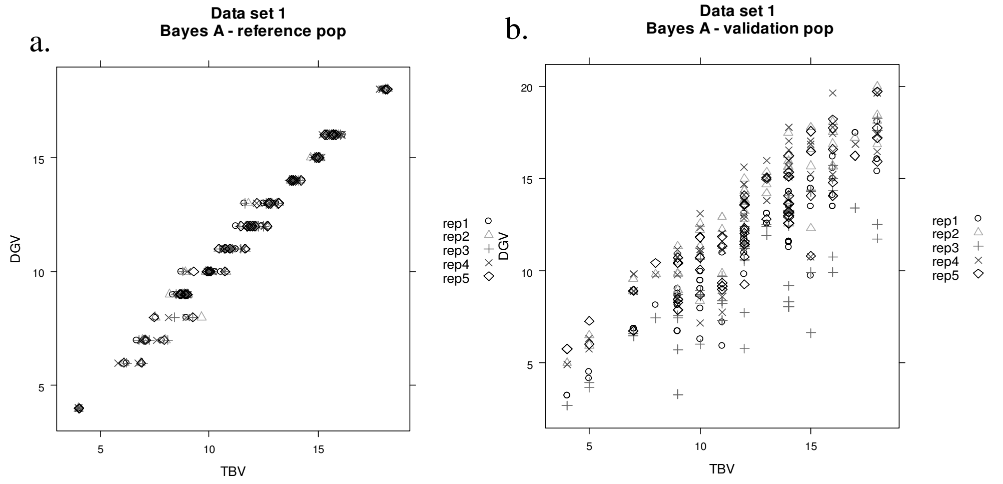

Figure 3.3- Bayes A plots- TBV vs DGV- for the reference (a.) and validation (b.) populations

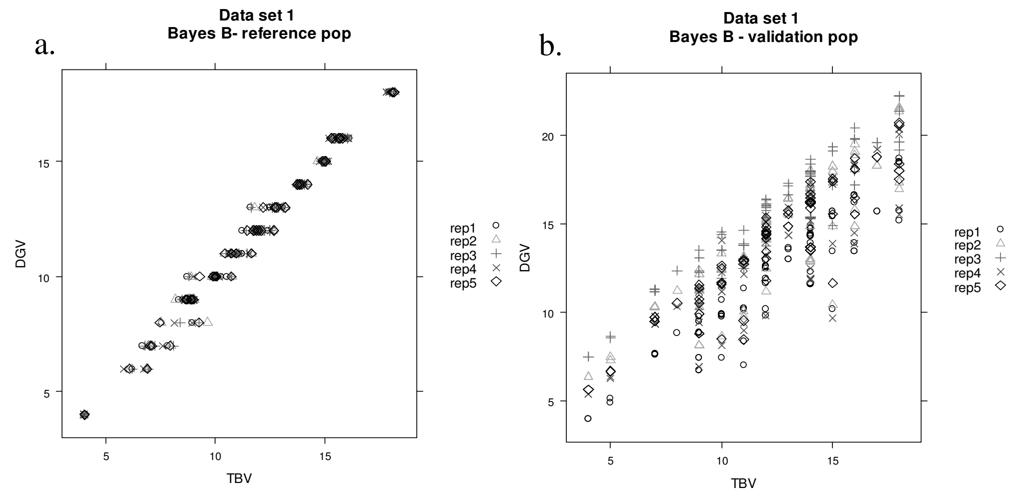

Figure 3.4- Bayes B plots- TBV vs DGV- for the reference (a.) and validation (b.) populations

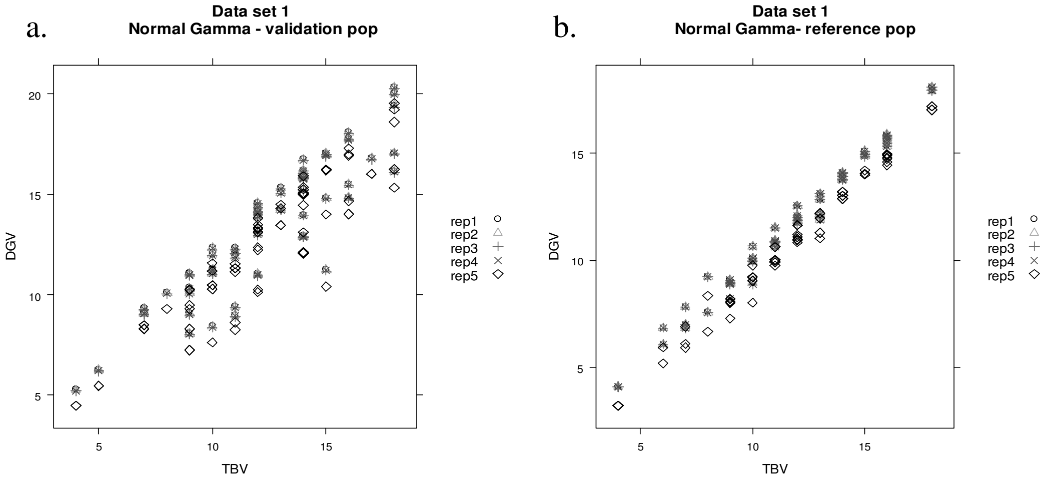

Figure 3.5- Normal Gamma plots- TBV vs DGV- for the reference (a.) and validation (b.) population

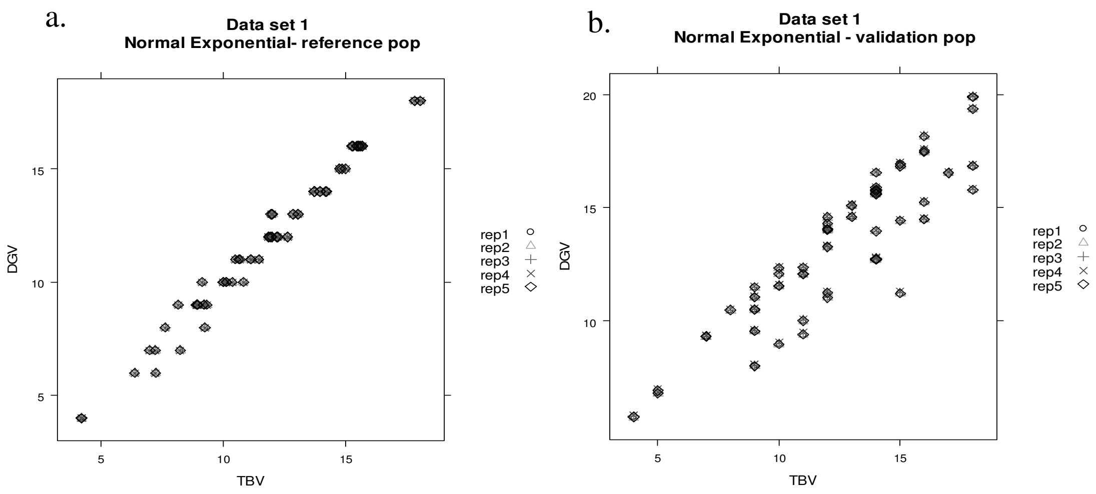

Figure 3.6 - Normal Exponential plots- True vs Predicted DGV- for the reference (a.) and validation (b.) populations

3.3.2 Data Set 2

The results for the accuracy of each method (and standard errors) using data set 2, across the five replicates are shown in Table 3.5. The table clearly shows that in both the reference and validation populations that Bayes A performed the worst. Figure 3.7 shows the reference and validation population for predicted DGV versus the TBV for Bayes A. It reveals that these lower values may be partially the result of one replicate with less accurate results caused by bias. The figures for the other different hierarchical models for the reference and validation population are shown in figures 3.7-3.10. Unlike Bayes A, Bayes B was the best performing of the models assessed using all three comparative measures. However, Bayes B, the normal-gamma and normal-exponential mixtures all produced extremely reliable results when comparing the Pearson correlation coefficient and MSE with no significance difference between their performance.

Table 3.5- Results for Data Set 2. Validation and Reference population results comparing predicted DGV and true breeding values (TBV) using regression coefficients (Reg) of true regressed onto predicted, the Pearson correlation coefficient (Cor) and mean square error (MSE).

REFERENCE POPULATION

	Bayes A	BayesB	NE	NG
Reg	0.9455 ± 0.0019	0.9757 ± 0.0005	1.0390 ± 0.00001	1.0194 ± 0.0026
Cor	0.9856 ± 0.0012	0.9910 ± 0.0003	0.9942 ± 0.00001	0.9919 ± 0.0002
MSE	0.5520 ± 0.0374	0.3212 ± 0.0104	0.3556 ± 0.00025	0.2862 ± 0.0089

VALIDATION POPULATION

	Bayes A	BayesB	NE	NG
Reg	0.8867 ± 0.0117	0.9928 ± 0.0013	1.0775 ± 0.0001	1.0746 ± 0.0229
Cor	0.9583± 0.0070	0.981211 ± 0.0009	0.9793 ± 0.0001	0.9818 ± 0.0006
MSE	1.4463± 0.2318	0.5257 ± 0.0162	0.6482 ± 0.0008	0.5440 ± 0.0233

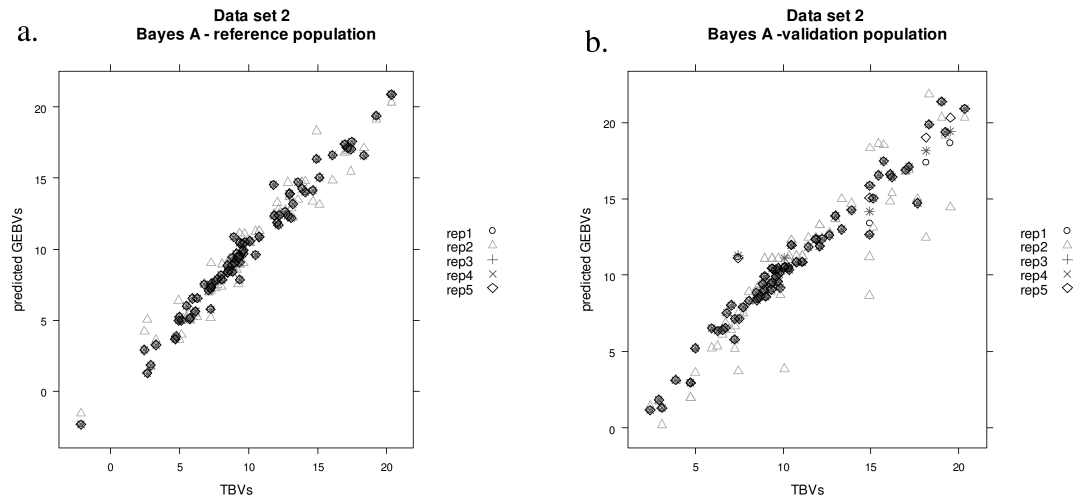

Figure 3.7- Bayes A plots- TBV vs DGV- for the reference (a.) and validation (b.) population

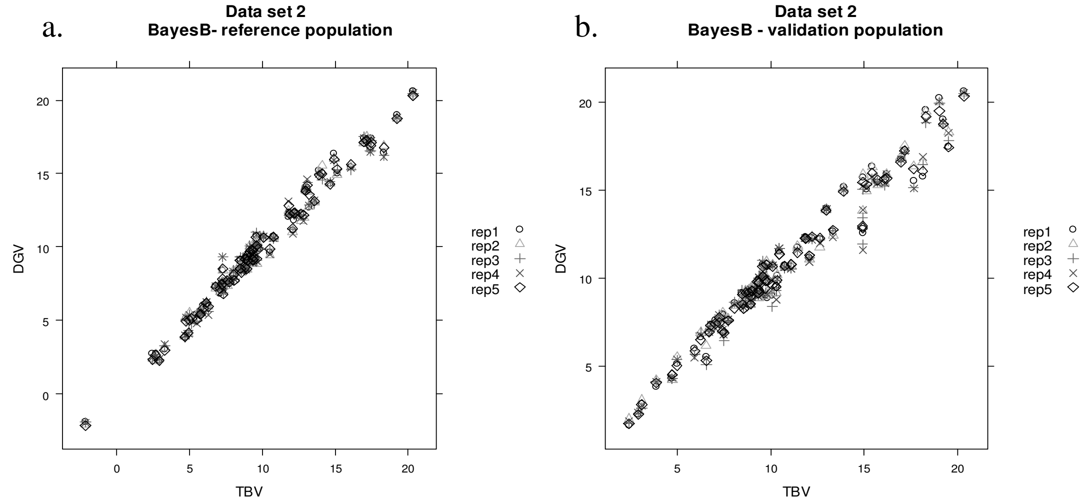

Figure 3.8 - Bayes B plots- TBV vs DGV- for the reference (a.) and validation (b.) populations

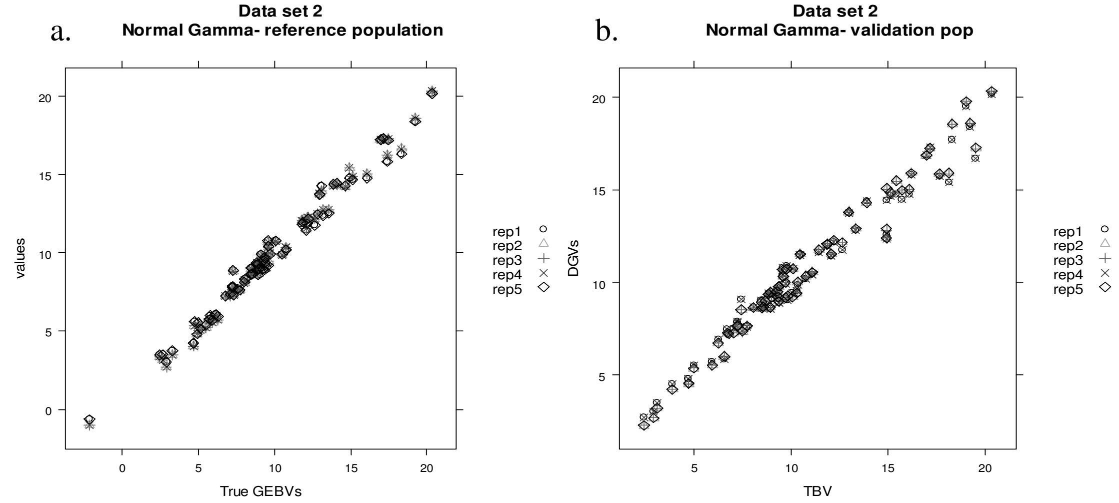

Figure 3.9 - Normal Gamma plots- TBV vs DGV- for the reference (a.) and validation (b.) populations

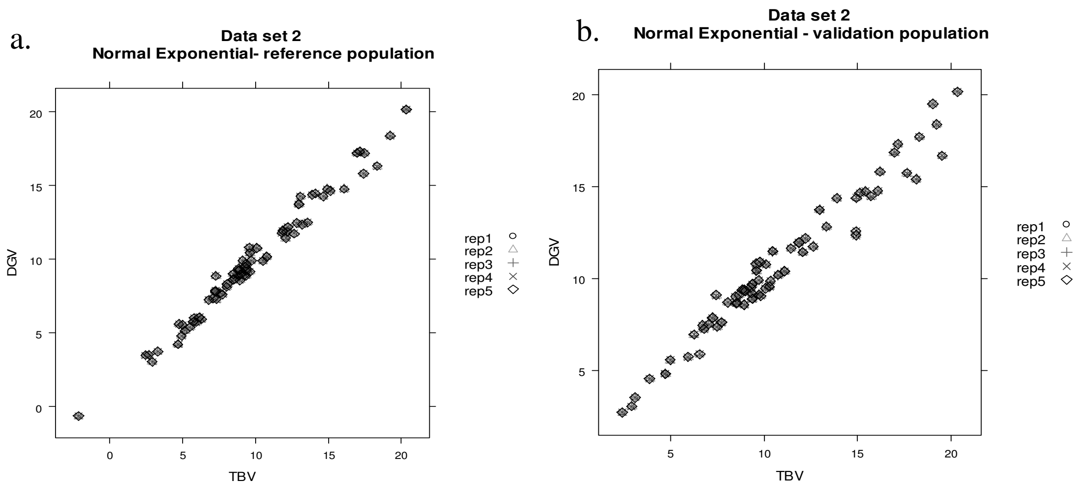

Figure 3.10 - Normal Exponential plots- TBV vs DGV- for the reference (a.) and validation (b.) populations

3.3.3 Data Set 3

Three replicates were performed for each of the hierarchical models using data set 3 where there were no SNP effects and any variation in the phenotypes was caused only by noise. The results were extremely similar across replicates. Table 3.6 presents the MSE values for the reference and validation populations and Figure 3.11 shows the results of TGV versus the predicted DGV from 1 replicate (as all replicates were extremely similar) . The most interesting result is that even though Bayes B allows SNP effects to be set to zero, Bayes A was the only model that correctly set all SNP effects to effectively zero. Nevertheless the results are still not significantly different and all methods produce relatively accurate results.

Table 3.6- Results for Data Set 3. Validation and Reference population results comparing predicted and true breeding values mean square error (MSE) with the standard error (± s.e)

Reference Population

	Bayes A	Bayes B	Normal Exponential	Normal Gamma
MSE (± s.e)	0	0.0436 ±0.0020	0.0414 ±0.0052	0.0713 ±0.0002
Validation Population				
	Bayes A	Bayes B	Normal Exponential	Normal Gamma
MSE (± s.e)	0	0.1148±0.0046	0.0858±0.0180	0.1885±0.0018

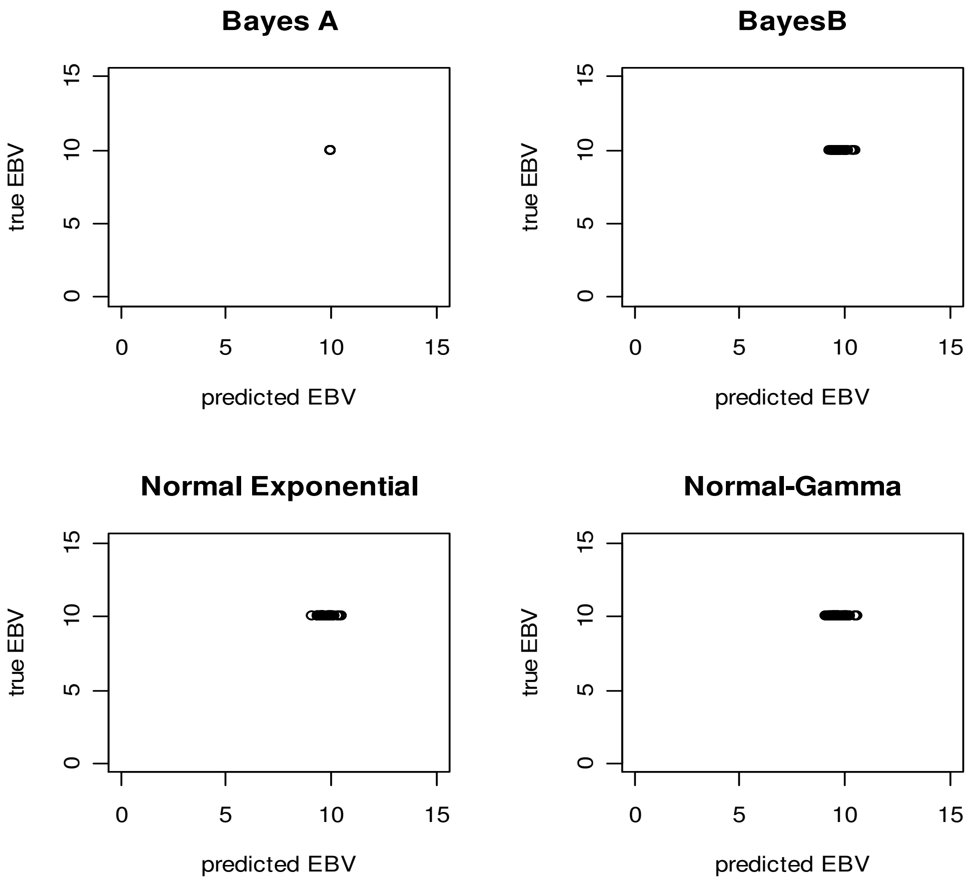

Figure 3.11- No SNP effects- predicted DGV from replicate 1 versus true DGV for the validation population.

3.3.3 DISCUSSION

3.4.1 Data Set 1

The results for the reference population for the data set 1 were as expected with all models producing regression and correlation coefficients close to 1. Bayes A had a higher MSE perhaps indicating more bias which could be accounted for by increased variation between the replicates. This variation was a result of inaccurate estimation of the mean and subsequently the SNP effects. However, in the validation population, Bayes B had the highest MSE. Both Bayes A and Bayes B had replicates that were more susceptible to variation resulting in an overall higher MSE, which is shown by the spread of predicted breeding values in Figures 3.2 and 3.3, and the greater standard errors.

On inspection of the estimated means, the values differed across replicates for both Bayes A and Bayes B. It appeared that the effect of the mean was in some replicates absorbed into the SNP effects. The reason for this is not obvious; inspection of parameter values in the MCMC chain showed that convergence had been reached, so it is most probably a reflection of the very small size of the data set and would be expected to disappear once the data set size was increased. Also contributing to the underperformance of Bayes B is that the prior probability of a SNP having a non-zero effect was set to 0.05 which relates to 12.5 (12-13) SNP being linked to a QTL .The restriction placed on the number of SNP with non-zero effects (and variances) in Bayes B may, in fact, restrict the amount of variation that can be explained by the SNP. If more than one SNP is needed to accumulate the effect of one QTL then this restriction, if too low, may negatively affect the accuracy of the DGV

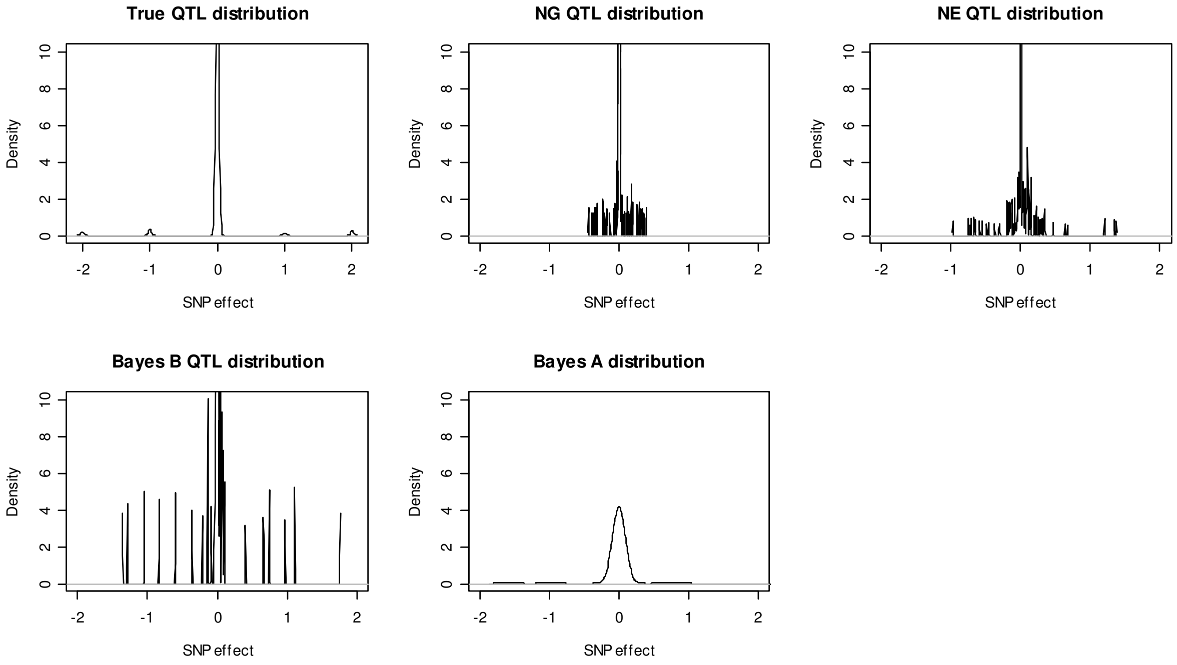

Figure 3.12- Posterior distributions and True distribution of SNP effects for all models for dataset 1. NG = normal-gamma hierarchical prior distribution, NE = normal-exponential hierarchical prior distribution.

When examining the posterior distributions of the SNP effects (Figure 3.11), the shrinkage effects of all approaches are evident. The normal-gamma hierarchical prior distribution seems to be the most affected. This may be a reflection of incorrect estimation of the genetic variance possibly resulting in increased shrinkage. Also obvious is the smooth t-distribution of the SNP effects of Bayes A.

The normal-exponential and normal-gamma prior distributions were the most consistent across the replicates, producing sets of DGV with the lowest MSE (bias and error), highest Pearson correlation coefficients and regression coefficients that were closest to one. Interestingly, while the prior specification of Bayes B sets a bulk of SNP effects to zero and thus it would seemingly be the closest prior specification to the simulated effects, the normal-exponential and normal-gamma prior distributions assumptions appear to produce DGV that are closer to the true breeding values. Insight into the cause of this is provided by Figure 3.12. The normal-gamma and normal-exponential distribution have posteriors similar to that of the simulated QTL distribution, in that they have a bulk of effects with very small effects and then some larger effects. They compensate for not having really large effects, that is, none close to the ±2 simulated, by estimating a larger proportion of smaller SNP effects. This is evident in both Figure 3.12 and Figure 3.13. Figure 3.13 shows the simulated QTL effect across the genome and the estimated SNP effects for each of the models.

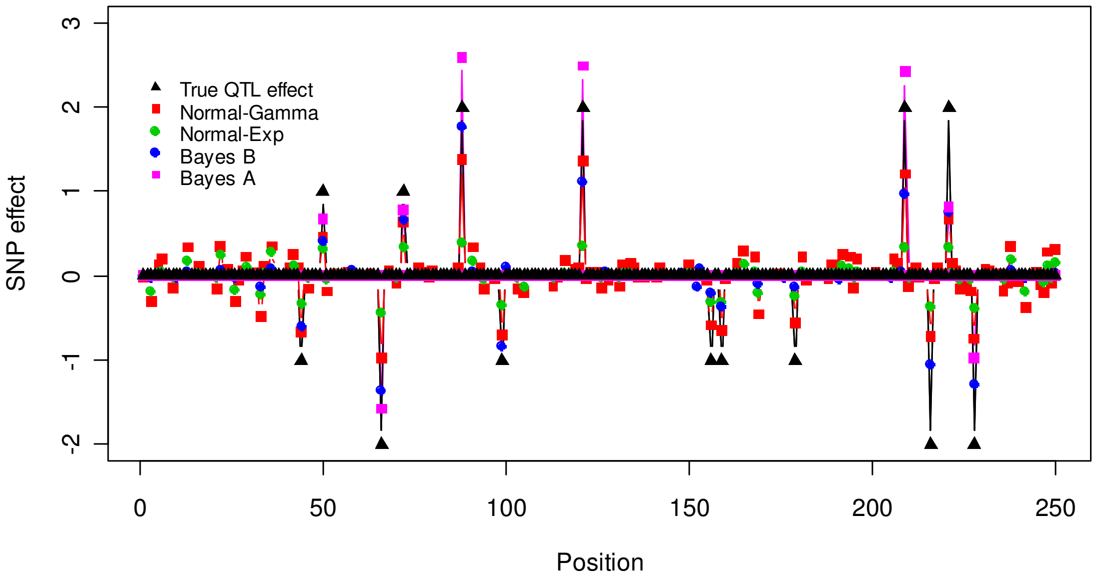

Figure 3.13– Position and effects of the simulated QTL and the estimated SNP effects for the four hierarchical models for data set 1

All models are able to correctly distinguish most QTL. However, those hierarchical models that used the Metropolis Hastings algorithm identified more QTL. Bayes A appeared unable to identify all the minor QTL with a negative effect (-1). Additionally, it overestimated the effect of the three large positive QTL effects. This is a most likely a reflection of its prior distribution with the larger effects being sampled from the fat tails and the fact that in some replicates some of the mean was absorbed by the SNP effects.

3.4.2 Data Set 2

The accuracy of the DGV produced by the different hierarchical prior distributions differed for this simulated data set. They could be split based on the hierarchical prior distribution and subsequent MCMC sampling algorithm. Bayes A produced the least accurate DGV with the lowest correlation with the TBV, the highest MSE indicating bias and a regression coefficient the furthest from one. This lack of accuracy is primarily due to the fact that its prior assumptions make it the least able to approximate the true distribution of the QTL effects. Bayes B, normal-exponential and normal-gamma prior distributions all produce very similar DGV; correlations of >0.999 and Spearman rank correlations of >0.99 between the three different sets of DGV.

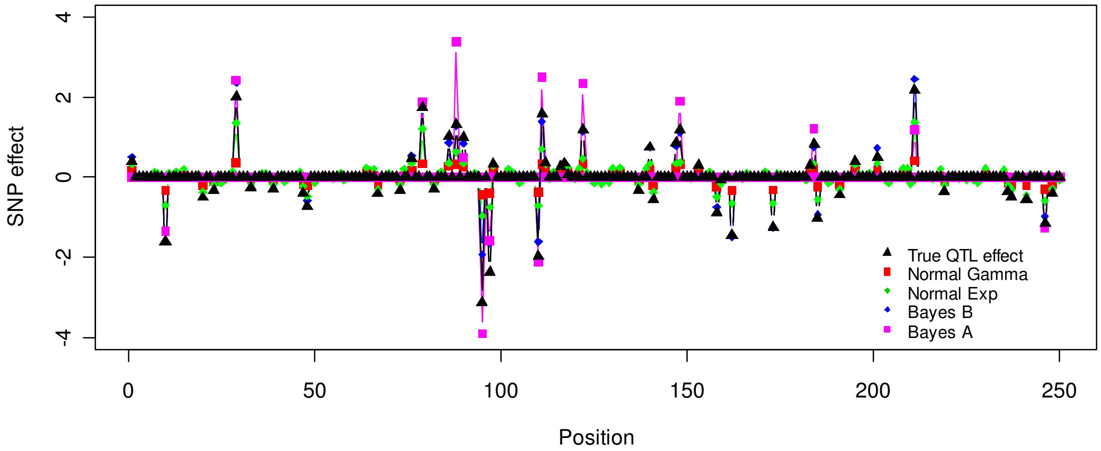

Figure 3.14– Position and effects of the simulated QTL and the estimated SNP effects for the four hierarchical models for data set 2.

It is not surprising that normal-exponential and normal-gamma prior distributions produce accurate DGV for this data set as their prior distributions are very similar to that of the simulated QTL effects, namely, double exponential distribution with many small effects and a few large effects. Additionally, Bayes B is well able to estimate the QTL effects despite the simulated data having many SNP with minor effects. In fact, Bayes B was the most accurate model with the lowest MSE and correlation and regression coefficients that are the closest to one. This indicated that the Bayes B hierarchical prior distribution provided a flexible and robust approach. Similarly to data set 1, all models were able to identify most of the major QTL (Figure 3.14). However, Bayes A again struggled to identify some of the minor QTL and this may have explained its slightly lower correlation coefficient.

3.4.3 Data Set 3

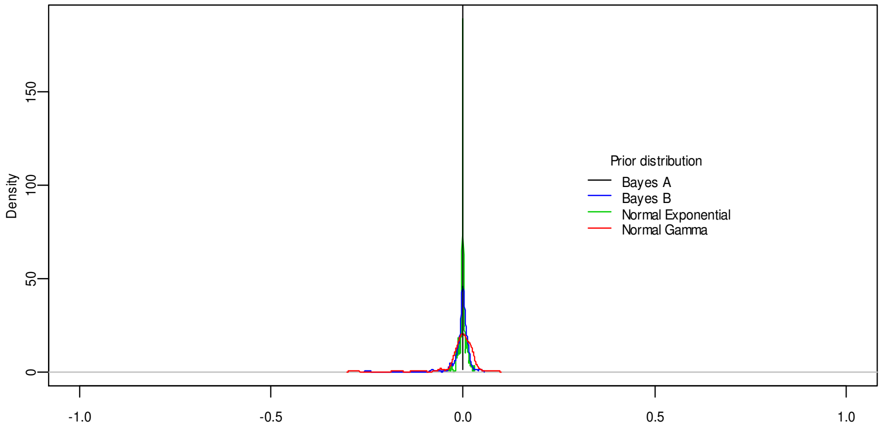

Figure 3.15 - Posterior distributions of SNP effects for data set 3

To fully explore the adaptability of the priors, through data set 3, the unrealistic case of no QTL, with any variation between animals being random, was explored. The only model able to accurately identify that there was no QTL was the model using the Bayes A hierarchical prior distribution. The reason for this result is that Bayes A has the only conjugate prior distribution that allows the use of the Gibbs sampler. This allowed the chain to converge quickly to a very accurate estimate of the mean and thus the SNP effects remained at the starting value of zero.

The Metropolis Hastings algorithm used by the other models allows, through the acceptance ratio, for SNPs by chance to take a new value. This step was designed to allow the chain to recover if it became stuck in local maxima or minima.

Consequently, if by chance a SNP may be assigned a non zero value and thus move away from the starting values of zero. Each of these accepted values make up the posterior of the SNP. The mean is then used to calculate the final SNP effect averaging all the values in the posterior. Consequently if the posterior contains even a single non-zero value then the final SNP effect will also be non-zero. The approaches do however shrink the effects back towards zero as shown in Figure 3.11 and 3.15. The normal-exponential model appears the most successful in shrinking these effects back towards zero. In contrast to Bayes A, Bayes B could never have found that there was no QTL. The specification of the prior distributions states that a certain probability of the effects must be non-zero, thus prohibiting the exact result found by Bayes A. Despite this Bayes B does adjust and produces DGV that were closer to the real mean than the normal-gamma prior distribution. Should the specification of the prior distributions have remained the same for Bayes B, the normal-exponential and normal-gamma models, an increase in the size of the data set would not be expected to change the results as all three distributions are likely to yield no-zero SNP effects.

3.4.4 Computational Time

The computational time for 10,000 iterations for each model is shown in Table 3.12. Immediately evident is that those models employing the Metropolis Hastings algorithm are significantly more computationally demanding. The use of R with the utilised code is shown to be unviable. However, the use of winBUGS may have yielded shorter times. Bayes A was viable but it generally produced the lowest accuracies. A consequence of this result was the recommendation and use of C++ and FORTRAN for all remaining computation and programming for the other studies presented in this thesis.

Table 3.7- CPU Time for 10,000 Iterations for the different hierarchical models

Prior	*Bayes A*[1]	*Bayes B*[2]	*NE*[2]	*NG*[2]
CPU Time*	0.5 hr	52 hr	337.5 hr	321.2 hr
Program	R	R	R	R

NG- normal-gamma mixture distribution, NE- normal-exponential mixture distribution, [1]Models utilising the Gibbs Sampler, [2]Models utilising the Metropolis Hastings algorithm, *hr = hours, min= minutes

3.5 CONCLUSION

The results for this small simulation study indicate that the models with the prior distribution that match to the true distribution will produce the highest accuracies. However, all models appeared generally robust, flexible and able to cope with different underlying QTL distributions. Generally Bayes A produced the DGV with the lowest correlation with the TBV and the highest MSE for the data sets containing QTL (data sets 1 and 2) indicating that it produced biased sets of DGV. This is due to the more strict assumption that the QTL effects are from a t-distribution; thus not at any stage matching the simulated QTL distribution). With such a small data set, this prior assumption may have overwhelmed the data. Bayes A does allow the more efficient Gibbs Sampler to be used reducing computational time. In contrast to Bayes A, all other hierarchical prior distributions do not have a fixed posterior rather the computationally slower Metropolis Hastings algorithm was used to construct the posterior distribution. The results and higher than normal accuracies obtained in this study are dependent on the simplistic model used to simulate the data, an increase in data set size and more realistic simulation model (or the use of real data) may change the results. These scenarios are explored in the following chapters.

The major result from this study was that those hierarchical priors that used Metropolis Hastings algorithm and that assumed unequal variances (i.e. Bayes B or normal-gamma or normal-exponential hierarchical prior distributions) produced more accurate DGV across the simulated data sets but were significantly computationally slower and thus unviable. In addition the normal –gamma and normal exponential hierarchical prior distributions over shrank the effects and found more minor QTL than were simulated. As a result, an alternative model was developed with comparable assumptions to Bayes B but with significantly less computational demands. This Bayes SSVS method is described in the next chapter (Chapter 4). It utilises stochastic search variable selection to enable the use of the Gibbs Sampler will maintaining the assumption of only a small number of significant large QTL. It equivalence to Bayes B is also proven in Chapter 4.

CHAPTER 4
Accuracy of Genomic Selection using Stochastic Search Variable Selection in Australian Holstein Friesian dairy cattle

4.1 INTRODUCTION

The results from the simulation studies in Chapter 3 suggest that more accurate DGV result from approaches that assume a majority of the SNPs have no or minor effects but have distributions with fat tails allowing a few major QTL (e.g. Bayes B, Normal-gamma and Normal-exponential prior distributions versus Bayes A). Those models with hierarchical priors such as Bayes B, the normal-exponential and normal-gamma mixtures (Chapter 3) use non-conjugate priors and thus require the use of the Metropolis Hastings algorithm which has significant time and computational demands. These demands make their use for large data sets such as those created using the Illumina BovineSNP50 beadchip (54,001 SNP) unviable.

An alternative approach is to use Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993). SSVS provides a method to maintain a constant dimensionality across all models but allows the parameters, in this case the SNPs, in the predictive set to change. It does this by not removing from the model all non-significant parameters (that is those that would be excluded from the predictive set and thus set to zero in Bayes B); instead, their effects are limited to values very close to zero.

The major advantage of this method is that, instead of using more computationally demanding algorithms, the posterior distribution of all parameters can be sampled directly using the Gibbs sampler. SSVS has been previously used for identifying multiple QTL (Yi et al., 2003), multivariate regression models (Brown et al., 1998), gene mapping (Swartz et al., 2006) and generalized linear models (George and McCulloch, 1997). It has also been utilised for analysing multi-trait QTL mapping data (Meuwissen and Goddard, 2004), and subsequently to investigate the effect that different methods for defining haplotypes and the effect of the inclusion of the polygenic effect had on the accuracy of genomic selection in simulated data (Calus et al., 2008, Calus and Veerkamp, 2007).

In this chapter, it is demonstrated that a Bayesian SSVS can be employed effectively, when compared to other methods, for genomic selection using real SNP data. The method also provides a viable alternative to more computationally demanding approaches such as Bayes B (Meuwissen et al., 2001) will maintaining nearly identical assumptions about the SNP effects. The approach used is novel and is modelled differently to the approach presented in Calus et al. (2008) where a relationship matrix is present at each SNP location describing the relationship of that SNP with the other SNP. This research presented here has been published in Genetics Research (Verbyla et al., 2009) (see Appendix A1 for the published paper).

4.2 MATERIALS AND METHODS

4.2.1 Data

The data set contained 1498 Australian Holstein-Friesian bulls genotyped for the Illumina Bovine50K array. After quality control, 39048 SNPs remained in the predictive set. The quality control applied to the SNP data is described by Hayes et al. (2009). The reference data set where the SNP effects were predicted contained 1098 proven bulls born between 1940 and 2000. The phenotypes for these bulls were Australian Breeding Values (ABV) for protein kg, fat kg, protein percentage, fat percentage and daughter fertility, all deregressed to remove any contribution from relatives (Hayes et al., 2009b). Daughter fertility is here defined as the difference between bulls for the percentage of their daughters that are pregnant 6-weeks after mating start date or 100-days after calving in year-round herds. The validation set contained 400 genotyped bulls proven in the years 2005, 2006 and 2007 with information from at least 100 milking daughters and available ABV to enable comparison with predicted DGV.

4.2.2. Model

At each locus (total number of loci, q) there are three possible combinations of two alleles (e.g. A or B), the homozygote of one allele (AA), the heterozygote (AB) and the homozygote of the other allele (BB). These are quantitatively represented by 0, 1 and 2 respectively. The model fitted to the above data was then:

$$\boldsymbol{y} = \mu \boldsymbol{1_n} + \sum_{j=1}^{q} X_j \beta_j + \boldsymbol{Zu} + \boldsymbol{e}$$

where y is the vector of phenotypes of the trait being analysed for all n individuals, μ is the mean, $\mathit{1}_n$ is a vector of ones of length n, $\mathbf{X}_j$ is a vector of indicator variables representing the genotypes of the j^th^ marker for all individuals (x_{ij}=0,1,2), β_j is the size of the SNP effect associated with marker j, **u** is the vector of random polygenic effects of length n (***Z*** is the associated design matrix) and is assumed to be normally distributed, $\boldsymbol{u} \sim N(0, \sigma_u^2 \boldsymbol{A})$ where ***A*** is the pedigree-derived additive genetic relationship matrix and ***e*** is the residual error also assumed to be normally distributed, $\boldsymbol{e} \sim N(0, \boldsymbol{I}\,\sigma_e^2)$. The polygenic effect was included to remove the effect of population structure to enable the more accurate estimation of the SNP effects. The inclusion of the polygenic effect has been shown to produce slightly better accuracies of prediction while reducing the bias of the variance components (Calus and Veerkamp, 2007).

4.2.3. Stochastic Search Variable Selection

The key feature of SSVS compared to Bayes A or B (Meuwissen et al., 2001) is the introduction of a latent or indicator variable, γ, into the hierarchical model. This enables the extraction of information relevant to variable selection. The latent variable can take either 1 or 0, representing whether the SNP is included as a significant effect in the model or not. As such, the prior distribution for each SNP effect is a normal mixture conditional on the corresponding latent variable γ and the variance which is sampled from an inverse scaled chi square distribution:

$$\beta_i \mid \gamma_i, \sigma_i^2 \sim (1-\gamma_i)\boldsymbol{N}(0, \sigma_i^2/100) + \gamma_i \boldsymbol{N}(0, \sigma_i^2)$$
$$\sigma_i^2 \sim \chi^{-2}(\boldsymbol{r}, \boldsymbol{S})$$

At the SNP effect level, this hierarchical prior distributions specification means the SNP effects are sampled from a mixture of two student t-distributions, one with a very small variance and a second larger distribution. The values of **r** and **S** were calculated as in Meuwissen et al (2001). The prior distribution of the indicator variable is chosen to reflect the belief of whether a SNP is in linkage disequilibrium with a QTL. The probability of a SNP being sampled from the smaller or larger distribution is:

$$1-p(\gamma_i=0)=p(\gamma_i=1)=p_i$$

Subsequently, the prior distribution for indicator variable is a Bernoulli distribution:

$$\gamma_i \sim \text{bernoulli}(\boldsymbol{p}_i)$$

The prior probability p_i is chosen to reflect the information available on how many QTL affect the trait of interest. It can be quantified as the number of SNP expected to be linked to a QTL divided by the total number of SNP. In genome-wide association studies or genomic selection applications, the expected proportion of QTL can sometimes be estimated based on knowledge about the trait of interest and previous QTL studies results.

The posterior distribution of the indicator variable can be sampled directly using:

$$\text{p}(\gamma_i=1 \mid \beta_j,\sigma_i^2,\gamma_{-i},\boldsymbol{u},\boldsymbol{y}) \sim \text{bernoulli}\left(\frac{\text{p}(\beta_j \mid \gamma_{-i},\gamma_i=1)\boldsymbol{p}_i}{\text{p}(\beta_j \mid \gamma_{-i},\gamma_i=1)\boldsymbol{p}_i+\text{p}(\beta_j \mid \gamma_{-i},\gamma_i=0)(1-\boldsymbol{p}_i)}\right)$$

where γ_{-i} is all terms of γ except γ_i.

The frequency that each SNP appears in the model is shown by the posterior distribution of the indicator variable. SNP that are included in the model frequently have a high posterior probability and will most likely be linked to a QTL. Consequently this approach could also be used for genome wide association studies.

4.2.4. Additional Methods

Bayes A, Bayes B and Bayes BLUP were also run on the data. Bayes A and Bayes B were as specified in Meuwissen et al (2001) and as described in Chapter 3 with the addition of a polygenic effect. A Bayesian BLUP method was also implemented. It is identical to the specification of Bayes A with the exception that all SNP have a constant equal variance that was sampled once each iteration from an inverse scaled chi square distribution.

In order to have Bayes B results for comparison to those of Bayes SSVS, a modified version of Bayes B approach was used. The modified version which consisted of running Bayes B cycles with the Metropolis Hastings Algorithm after every 100 iterations of Bayes A. If a SNP effect was found to be zero during these Metropolis Hastings algorithm iterations then it was set to zero during the subsequent Bayes A cycles. This effectively maintained the same assumptions as Bayes B, while significantly reducing the time required to reach convergence.

All methods were run for 10,000 iterations to ensure convergence. This number of iterations was shown to be sufficient for convergence with formal diagnostic methods provided in the package R, coda (Plummer et al., 2007 -b).

4.2.5. Breeding Values

Marker estimated breeding values (DGV) for bulls in the validation data set were calculated as the sum of the mean, the effects of the SNP genotypes that it carried and the polygenic effect, $\mathrm{DGV} = \hat{\mu} + \mathbf{X}\hat{\beta} + \hat{u}$. The accuracy of the methods were evaluated on the correlation, the mean square error (MSE) and the regression coefficient of the Australian breeding value (assumed to be the true breeding value) on the predicted DGV. Genomic selection aims to produce breeding values as close as possible to the true breeding value. The ABV was used for comparison as it is a most accurate predictor of the true breeding value.

4.3. RESULTS AND DISCUSSION

4.3.1. Time to Convergence

The use of the SSVS method is analogous to Bayes B in the assumption that the majority of the SNP effects are thought to be very small and insignificant. However as illustrated in Table 4.1, the fixed dimensions of the model and the conjugate nature of the prior distribution used in the Bayesian SSVS model allowed the use of the Gibbs Sampler which is significantly computationally less demanding and consequently quicker than the Metropolis Hastings algorithm used in traditional Bayes B. Given the very high computational demand of Bayes B, it was not possible to run this algorithm to convergence. The time to convergence was extrapolated from running Bayes B for 1000 iterations. The Bayes A and Bayes BLUP methods reached convergence in comparable times to Bayes SSVS.

Table 4.1 - Computational time for genomic selection methods

Method	*Computational Time* [a]
Bayes BLUP	6
Bayes A	6
Bayes B	~2440 [b]
Bayes B Modified	240
Bayes SSVS	6

[a] Processor clock hours

[b] Estimated time to convergence

4.3.2. Comparison of BAYES B and BAYES SSVS results

The correlations between the ABVs and the DGV predicted for the animals in the validation set by the modified Bayes B and Bayes SSVS for fertility and protein kg traits are shown in Table 4.2. This shows that the two methods produce almost identical correlations with the ABVs as expected. The DGV for the two methods are 99.9% and 98.0% correlated for protein and fertility respectively. This equivalence in results demonstrates that the Bayes SSVS method does maintain the SNP effect assumptions of the original Bayes B and produces near to identical results. The slightly lower result for fertility is probably due to the non-normality of the trait making it harder to estimate. The modified Bayes B produced not significantly different but slightly larger mean square errors and regression coefficients for protein (Table 4.3 and 4.4). This is most likely due to the modification to reduce the computational time to convergence. The time taken for the modified version of Bayes B was still 40 fold larger than that for the Bayes SSVS which produced identical accuracies (see Table 4.1).

Table 4.2- Correlation between predicted DGV and ABV for proven bulls (years 2005, 2006, 2007 and overall) for the modified Bayes B and Bayes SSVS.

	Bayes B (modified)	Bayes SSVS
Protein kg – 2005	0.620	0.627
– 2006	0.638	0.646
– 2007	0.502	0.490
Protein kg – Overall	0.575	0.583
Fertility – 2005	0.576	0.577
– 2006	0.430	0.429
– 2007	0.628	0.628
Fertility – Overall	0.540	0.540

Table 4.3- Mean Square Error between predicted DGV and ABV for proven bulls (years 2005, 2006, 2007 and overall) for the modified Bayes B and Bayes SSVS.

	Bayes B (modified)	Bayes SSVS
Protein kg – 2005	50.1	49.2
– 2006	53.8	52.7
– 2007	65.2	64.2
Protein kg – Overall	55.4	54.4
Fertility – 2005	5.03	5.03
– 2006	5.11	5.11
– 2007	3.02	3.02
Fertility – Overall	4.52	4.52

Table 4.4 - Regression Coefficient of predicted ABV on DGV for proven bulls (years 2005, 2006, 2007 and overall) for the modified Bayes B and Bayes SSVS.

	Bayes B (modified)	Bayes SSVS
Protein kg – 2005	1.128	1.072
– 2006	1.407	1.346
– 2007	1.435	1.274
Protein kg – Overall	1.187	1.131
Fertility – 2005	1.091	1.095
– 2006	0.783	0.781
– 2007	0.926	0.929
Fertility – Overall	0.930	0.933

4.3.3. Comparison of BLUP, BAYES A, BAYES SSVS results

The logarithm of the mean square error, regression and correlation coefficients for the predicted DGV and Australian Breeding Values (ABV) for the traits protein kg, fat kg, protein percentage and fat percentage are shown in Table 4.5. The values shown are the average values for the proven bulls in the years 2005, 2006 and 2007 from the validation data set. BLUP has the highest overall correlation and the lowest MSE between the three methods for protein kg. For the traits, fat kg and protein percentage, Bayes SSVS produces the highest correlations and has the lowest bias, but over all there are no significant differences between methods. However, there are significant differences between the performances of the methods for the trait, fat percentage.

Table 4.5- MSE, Correlation and Regression Coefficient between predicted DGV and ABV in the validation data set

Method	Measure	Bayes SSVS*	Bayes A*	Bayes BLUP*
Protein kg	$\tau_{DGV,ABV}$	0.583	0.567	0.602
	log(MSE)	4.03	4.06	3.96
	$b_{ABV,DGV}$	1.187	1.126	1.128
Fat kg	$\tau_{DGV,ABV}$	0.563	0.532	0.563
	log(MSE)	5.18	5.22	5.23
	$b_{ABV,DGV}$	0.900	0.856	0.988
Protein %	$\tau_{DGV,ABV}$	0.668	0.641	0.655
	log(MSE)	-4.94	-4.88	-4.84
	$b_{ABV,DGV}$	0.972	0.995	0.887
Fat %	$\tau_{DGV,ABV}$	0.740	0.716	0.646
	log(MSE)	-3.07	-3.24	-3.32
	$b_{ABV,DGV}$	0.874	0.864	0.925
Fertility	$\tau_{DGV,ABV}$	0.540	0.539	0.538
	log(MSE)	1.51	1.51	1.52
	$b_{ABV,DGV}$	0.933	0.942	0.905

*Average accuracies reported over validation sets from years 2005, 2006, 2007. $\tau_{DGV,ABV}$ Correlation coefficient between the ABV and predicted DGV, log(MSE) is the logarithm of the Mean square error between the ABV and predicted DGV, $b_{ABV,DGV}$ Regression coefficient of the ABV on predicted DGV.

The individual SNP variances that Bayes A and Bayes SSVS uses, allows some SNPs to have effects which are not penalised (shrunk) as severely as in BLUP. This is clearly shown in Figure 4.1, where the percentage each SNP contributes to the total

SNP effects are plotted for the three methods for the centromeric end of the bovine chromosome 14. Bayes A and Bayes SSVS have a SNP with an effect significantly greater than zero while the Bayes BLUP effects for SNP near DGAT1 are close to zero. Bayes SSVS does perform slightly better than Bayes A for fat percentage. The advantage of the Bayes SSVS over Bayes A may be the prior structure consisting of two distributions: a distribution of larger significant effects and a smaller distribution close to zero. This allows the SNP with larger effects to have values in their posterior sampled from the larger distribution, while those SNP without significance have their effects sampled from the smaller posterior distribution of values very close to zero. Traits with large effects will be more accurately predicted using SSVS than Bayes BLUP as the prior structure allows more variance to be attributed to the larger effects.

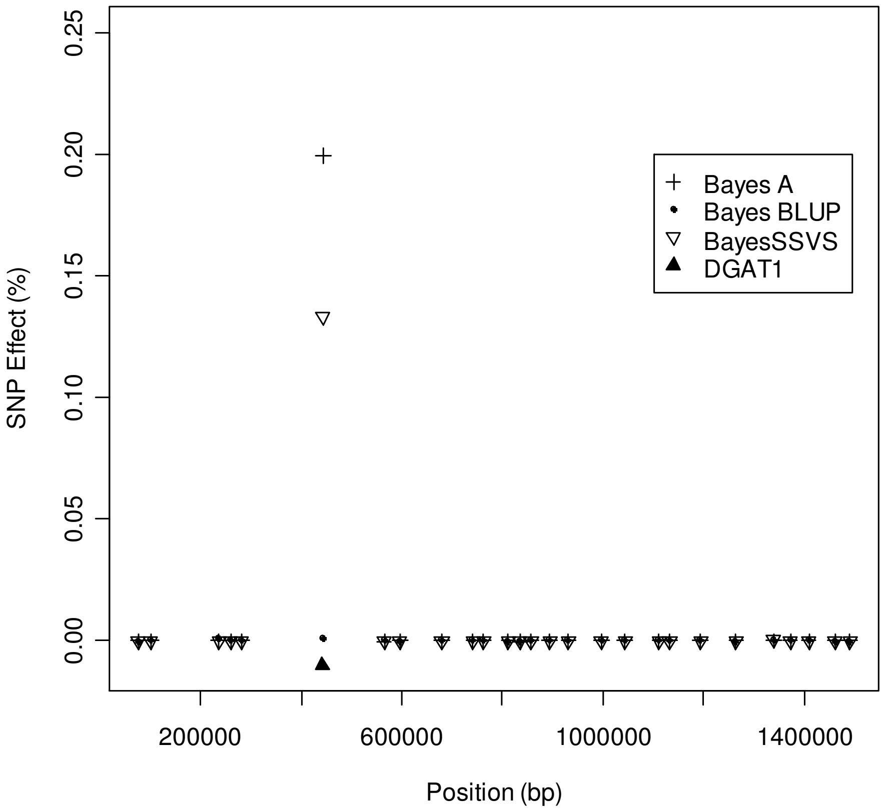

Figure 4.1- SNP effects (% of total effects) for fat percentage from Bayes A, Bayes BLUP and Bayes SSVS found on the centromeric end of chromosome 14

These differences in the method accuracies across traits or the apparent "trait by method" interactions can be explained by the distribution of QTL for the different traits. For example, Protein kg has no known genes of large effect and thus BLUP, which applies equal variances across all SNP, can be used successfully to accurately predict breeding values. In contrast, fat percentage has a mutation, DGAT1, that is common and acts additively and is known to be responsible for approximately 50% of the genetic variation for the trait (Grisart et al., 2002).

4.4. CONCLUSION

Bayesian SSVS produced more accurate DGV than the other methods for most of the dairy traits in the data set. The comparison with a modified version of Bayes B showed that it produces nearly the same results with dramatically less computational time required. For traits with a mutation of known large effect such as fat percentage, Bayes SSVS gave significantly higher accuracy of DGV than the BLUP method as expected given that its prior is closer to the real distribution of effects than that of BLUP. The use of an indicator variable in Bayes SSVS would also allow the premeditated inclusion of SNP in a model that are known to be linked to QTL of biological importance or are themselves causal mutations. Instead of using a single value for the prior probability for all SNP, a vector of probabilities could be used as prior probabilities to allow more prior information to be included should it be available.

Overall, this study had shown that Bayes SSVS method provides reduced computational time and accurate results when using real dairy data to predict genomic breeding values and provides a viable alternative to other Bayesian methods for Genomic Selection.

The performance of Bayes SSVS and Bayes B modified are further examined in simulated data produced as part of the 13th QTLMAS workshop in the next chapter. In addition, a Bayesian BLUP approach and Bayes A are applied to the data. The aim is to again examine the performance of these models and of priors (different from Chapter 3) with known QTL distributions. In that study, the performance of the models raised different issues and the implications of these are discussed

CHAPTER 5
Sensitivity of Genomic Selection to using different prior distributions

5.1 INTRODUCTION

As a response to the development of Bayes SSVS and the implementation of a modified Bayes B (Bayes A/B hybrid), a second simulation study was performed to assess the performance of these methods in simulated data. This study used the simulated data from the 13th QTL-MAS workshop. The advantage of this data set was that it allowed the testing of Bayesian genomic prediction approaches on a data with a linkage analysis structure. Consequently, it provided an opportunity to assess the performance of four Bayesian genomic prediction models on a differently structured, more realistic, simulated data set than presented in Chapter 3.

Four Bayesian models differing again through the specification of the prior distributions for the SNP effects and their respective variances were applied to estimate DGV. They included three new methods not presented in Chapter 3, Bayes SSVS (Chapter 4), Bayes BLUP and Bayes A/B along with Bayes A (used in Chapter 3). The results of this study have been published in BMC proceedings as part of the publications for 13th QTLMAS workshop (Verbyla et al., 2010a) (see Appendix A3 for the published paper).

5.2 MATERIAL AND METHODS

5.2.1 Simulated data

The data was simulated as part of the 13th QTLMAS workshop held in Wageningen, the Netherlands in 2009. The data set consisted of 2,025 individuals from two generations. All individuals had complete marker information. The first 25 individuals were the parents, 20 female and 5 male. The remaining 2000 individuals were offspring consisting of 100 full sibs (FS) families, one from each combination of a male and female parent. Each FS family has 20 offspring. Fifty FS families were phenotyped, the other 50 FS families did not have phenotypes. FS families were chosen such that each female parent has at least 40 phenotyped offspring while each male parent has 100 phenotyped offspring. The phenotypes were measured at 5 time points (0, 132, 265, 397 and 530). The phenotypes were simulated such that they could be seen as yield values representing weight during the growth of an animal or plant. The phenotypes formed points on a logistic growth curve. The true breeding values (TBV) were available for time point 600 for the animals without phenotypes.

There were 453 SNP marker loci which were randomly distributed over 5 chromosomes. Each chromosome was approximately 1 Morgan in length. 18 QTL were simulated, 6 affecting each parameter of the logistic curve with one QTL explaining 50 percent of the genetic variation for that parameter. The LD (r^2) between a marker and QTL varied between 0.16 and 1.00. The average LD (r^2) between flanking markers was 0.14. Three QTL were on chromosome 1 and 5 and four QTL were on the other chromosomes. They all acted additively and explained from 2.5% to 32% of the phenotypic variance. No polygenic, epistatic or imprinting (maternal/paternal) effects were simulated.

The TBV and the details of how the data was simulated were only revealed after the workshop and the data analysis was completed; for more details, see Coster et al. (2010).

5.2.2 Prediction of Breeding Values at Time Point 600

Due to the availability of phenotypes only at t=0, 132, 265, 397 and 530, the problem of how to model the time series data and estimate DGV at time point 600 was explored. Options included estimating the phenotype at t=600 and then deriving the DGV. This could have been done linearly or by assuming a type of growth curve. However, there was little information available to estimate any inflection points or asymptotic values. The second alternative was to estimate the DGV and then extrapolate to t=600. This approach was adopted to estimate the DGV. The predicted DGV at time points 265, 397 and 530 were found to have a linear relationship; they appear to form the linear part of the growth curve (Figure 5.1). Consequently, as there was no other information available after time point 530 to predict asymptotes etc., the

DGV at time point 600 were estimated by fitting a linear regression through the breeding values at the three linear time points; 265, 397 and 530 (Figure 5.1).

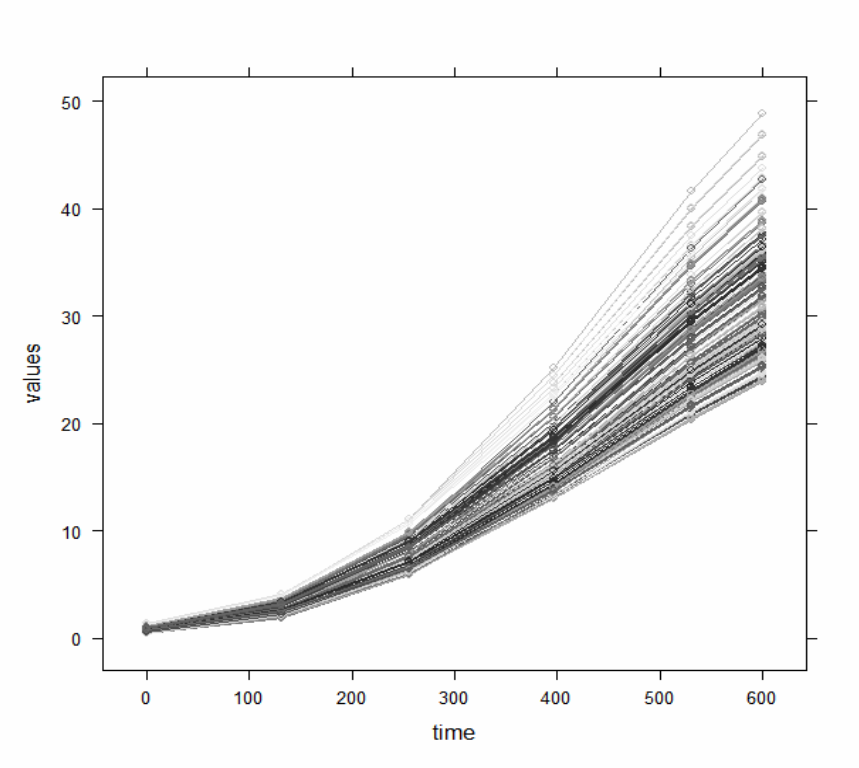

Figure 5.1– DGV predicted for t=0, 132, 265, 397 and 530 and extrapolated for t=600 using linear regression through t=265, 397 and 530.

5.2.3 Model

Unlike the model used in simulated study in Chapter 3, the model used here included the polygenic effect. This was primarily because it was unknown whether or not a polygenic effect was included in the simulated data. The model used was:

$$\boldsymbol{y} = \mu \boldsymbol{1}_n + \sum_{j}^{p} \boldsymbol{X}_j \beta_j + \boldsymbol{Zu} + \boldsymbol{e}$$

where y is the vector of phenotypes of the trait being analysed for all n individuals, μ is the mean, 1_n is a vector of ones of length n, X_j is a vector of indicator variables representing the genotypes of the jth marker for all individuals (x_{ij}=0,1,2), β_j is the size of the QTL effect associated with marker j, u is the vector of random polygenic effects of length n (***Z*** the associated design matrix) and is assumed to be normally distributed, $\boldsymbol{u} \sim \boldsymbol{N}(0, \sigma_u^2 \boldsymbol{A})$ where ***A*** is the pedigree derived additive genetic relationship matrix and **e** is the residual error also assumed to be normally distributed,

$\boldsymbol{e} \sim \boldsymbol{N}(0,\boldsymbol{I}\sigma_e^2)$ where $\boldsymbol{I}$ is the nxn identity matrix. The prior distributions for the variances of the random polygenic effects and the residual were uninformative flat priors of the form $\chi^{-2}(-2,0)$. For specification of posteriors distribution sampled for the mean, residual and QTL effects see Chapter 3 (Section 3.2.3).

5.2.4 Prior Distributions for QTL effects and Algorithms

Four differing sets of prior distributions were assessed; the specifications are shown in Table 5.1. The Bayes BLUP model assumed the same variance for the normal distribution from which the SNP effects were assumed to be derived. The variance of the normal distribution was sampled once every MCMC iteration using a Gibbs Sampler. Bayes A (Meuwissen et al., 2001) as used in Chapter 3 and 4 assumes that the SNP effects come from a *t*-distribution. The values for the inverse scaled chi square hyper parameters (*r* and *S*) were calculated as in Meuwissen et al (2001); see Table 3.7 for values.

Table 5.1- Prior Distribution Specifications

Method	**Prior Distribution**
Bayes BLUP	$\beta_i \sim \boldsymbol{N}(0,\sigma^2)$ $\sigma^2 \sim \chi^{-2}(\boldsymbol{r}_1,\boldsymbol{s}_1)$
Bayes A	$\beta_i \mid \boldsymbol{v}_i \sim \boldsymbol{N}(0,\sigma_i^2)$ $\sigma_i^2 \sim \chi^{-2}(\boldsymbol{r}_1,\boldsymbol{s}_1)$
Bayes A/B (Hybrid)	$\beta_i \mid \boldsymbol{v}_i \sim \boldsymbol{N}(0,\sigma_i^2)$ $\sigma_i^2 = 0$ with probability 1-π $\sigma_i^2 \sim \chi^{-2}(\boldsymbol{r}_2,\boldsymbol{s}_2)$ with probability π
Bayes SSVS	$\beta_i \mid \gamma_i,\sigma_i^2 \sim (1-\gamma_i)\boldsymbol{N}(0,\sigma_i^2/100)+\gamma_i\boldsymbol{N}(0,\sigma_i^2)$ $\sigma_i^2 \sim \chi^{-2}(\boldsymbol{r}_2,\boldsymbol{S}_2)$ $\gamma_i \sim \boldsymbol{bernoulli}(p_i)$ $1-\boldsymbol{p}(\gamma_i=0)=\boldsymbol{p}(\gamma_i=1)=p_i$

Hyper-parameters were set at *p*$_i$=π=0.05, $(\boldsymbol{r}_1,\boldsymbol{s}_1)=(4.0035,0.0954)$, $(\boldsymbol{r}_2,\boldsymbol{s}_2)=(4.0692,1.8800)$.

The other two models assumed mixture distributions for the SNP effects reflecting the assumption that there is a large number of SNPs with zero or near zero effects and a second smaller set of SNPs with larger significant effects. A Bayes A/B “hybrid” method was used. This approximation to Bayes B (Meuwissen et al., 2001) was used to keep computational and time demands reasonable. In this algorithm, after every k Bayes A iterations, Bayes B via the Metropolis Hasting algorithm was employed. The Metropolis Hasting algorithm was run multiple times per SNP and then any SNP with a final state of zero in the current Bayes B iterations was set to zero for the subsequent k iterations of the Bayes A (k = 100) . The prior distributions are identical to that of the original Bayes B using a mixture prior distribution for the SNP variance allowing a proportion, π, to be set to zero. The other proportion, 1-π, is sampled from a similar mixture distribution to that used for Bayes A. See Meuwissen et al (2001) for more details of priors and conditional distributions used.

A faster alternative to both the Bayes A/B hybrid and Bayes B is to use Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993) (Bayes SSVS – Chapter 4). This avoids the problem of a non-conjugate prior and the possible changing dimensionality of the models (if p_i, the proportion of significant SNP effects, is also sampled) by providing a technique to maintain constant dimensionality across all models while still allowing the SNP in the predictive set to change. Instead of removing all non-significant parameters, their posterior distributions are limited to values close to zero. The major advantage of this method is that it can be implemented using the Gibbs Sampler instead of the more computationally demanding algorithms such as the Metropolis Hastings algorithm. The indicator variable (γ) determines whether the SNP effect is sampled from the larger distribution (i.e. significant effect) or from the small distribution with near zero effects. This model was developed as a response to the time and computational demands of Bayes B that made it unviable for use on bigger real data sets. However, the assumptions of Bayes B, in simulated data, produced higher accuracies than the faster Bayes A, thus Bayes SSVS provides a fast algorithm with similar assumptions to Bayes B. Further specification, testing and discussion of Bayes SSVS are presented in Chapter 4 and published in Verbyla et al. (2009).

The prior values of π and p_i for Bayes A\B and Bayes SSVS respectively were set to 0.05, reflecting the fact that with 435 SNP, it appeared reasonable to expect at least 21 SNP would be associated with a QTL as no additional information was available about the trait being analysed. The algorithms associated with each model were run for 30,000 iterations with the first 10,000 discarded as burn-in. The DGV at each time point were then calculated as $\mathrm{DGV} = \hat{\mu} + \mathbf{X}\hat{\beta} + \hat{\boldsymbol{u}}$.

5.3 RESULTS AND DISCUSSION

5.3.1 Breeding Values

The methods produced significantly different summary statistics (variances, means, minimums and maximums) for different sets of DGV produced for the animals without phenotypes (Table 5.2). Bayes SSVS and Bayes A/B produced very similar summary statistics which is aptly explained by the similarity in the prior distributions. Bayes BLUP produced a similar mean value but a higher variance. All approaches utilised have a much lower variance in their respective sets of DGV than the TBV. This could be caused by the extrapolation of the DGV to the 600 time point and the fact that the QTL effects were estimated linearly where they were simulated as influencing the three parameters of the logistic growth curve.

Table 5.2- Summary of DGV statistics for each model and the TBV.

	TBV	*Bayes A*	*Bayes SSVS*	*Bayes BLUP*	*BayesA/B*
Min	18.068	23.998	23.845	16.950	22.205
Max	48.711	49.994	48.877	49.314	48.206
Mean	29.569	29.601	29.609	29.512	29.572
Variance	25.346	18.441	17.913	20.721	17.124

Despite the small apparent differences in DGV produced by Bayes A, Bayes A/B and Bayes SSVS, the correlations between these sets of DGV were extremely high (>0.99). Consequently, the DGV appeared relatively insensitive to the model used when assuming unequal variances. The correlations between the predicted sets of DGV for the alternative methods, for animals without phenotypes are shown in Table 5.3.

Table 5.3- Correlations between Estimated DGV for animals with no phenotype at t=600

	Bayes SSVS	*Bayes A/B*	*Bayes BLUP*
Bayes A	0.9991	0.9901	0.8598
Bayes SSVS	1	0.9924	0.8634
Bayes A/B		1	0.8928

Table 5.4 - Comparison of True and Estimated DGV. Correlation, Mean Square Error (MSE), Rank (Spearman Rank Correlation for the first 100 animals) and regression of the true on the estimated DGV

Method	*Correlation*	*MSE*	*Rank*	*Regression*
Bayes BLUP	0.885	5.479	0.691	0.979
Bayes A	0.864	6.630	0.696	1.162
Bayes A/B	0.889	5.435	0.73	1.081
Bayes SSVS	0.869	6.232	0.71	1.024

Correlations, mean square errors, the accuracy of predicting the order of the first 100 animals (rank) and the regression coefficient between the predicted and true breeding values are shown in Table 5.4. While there is no significant difference between the methods, Bayes A/B performed the best of the methods producing the lowest MSE and the highest correlation and rank statistics. Interestingly, whilst Bayes SSVS has very similar hierarchical prior distributions, it does slightly worse than Bayes A/B. Further optimisation of the prior probability of p_i for Bayes SSVS increased the accuracy. The optimal value for *pi* was found to be 0.3; values tested were 0.05, 0.1, 0.2, 0.3, 0.4, 0.6 and 1. This was also consistent with the value used to accurately identify the QTL (Heuven and Janss, 2010). This value produced results that were similar to the results seen for Bayes A\B ($\pi = 0.05$). When p_i was set to 1, the DGV and accuracy were almost identical to Bayes A (the only difference caused by the differing hyper-parameters). This does highlight the importance of the correct assumption of the proportion assigned to the smaller and larger distributions in a mixture model. This difference between these two methods may demonstrate that Bayes SSVS is more sensitive to an incorrect assumption about this proportion. Alternatively, due to the extremely high correlation between the two sets, the slight reduction in accuracy may be a result of an introduced bias in Bayes SSVS reflected by the higher MSE in Table 5.4.

The inclusion of the polygenic effect in the model (not simulated in the data) only slightly reduced the accuracy of prediction (.01) but not significantly (Table 5.5). It was included in the model as its inclusion has been shown to produce slightly better accuracies of prediction while reducing the bias of the variance components (Calus and Veerkamp, 2007).

Table 5.5- Comparison of Bayes SSVS results with varied parameters

Proportion (p_i)	*Polygenic Effect*	*Correlation Coefficient*
0.05	Included	0.869
0.05	Not Included	0.877
0.1	Included	0.883
0.3	Included	0.891

p_i is the proportion of SNP with a significant effect for Bayes SSVS.

Bayes BLUP produced a significantly different set of DGV. This is evident by the much lower correlations with the other methods and the fact its regression coefficient is significantly different from the other approaches. These differences are caused by the very different specification of the hierarchical prior distribution assumed. Despite these differences, Bayes BLUP produces good accuracy and a low MSE (Table 5.4). The assumption of equal variance meant that the effect and variance explained by the QTL were picked up by many SNP. This meant that the effect was effectively spread across those SNP in LD with the QTL. The success of the BLUP approach in this data set can be explained by two factors. The first that the structured pedigree creates LD within families over long distances and thus allows BLUP to capture successfully and to spread the QTL effect over a number of SNP. The second factor is that the accuracy of BLUP in simulated small data studies has been shown to be inversely related to the number of SNP (Fernando et al., 2007). However, this inverse relationship was defined in the situation where the number of QTL remained constant and the accuracy of BLUP was measured as the numbers of SNP were increased. The assumptions of BLUP do predict this behaviour; as the numbers of SNP increase the amount of variance explained that can be explained by a single SNP decreases. If the proportion of SNP in LD with QTL is constant then BLUP should behave the same but if the proportion changes and less SNP are in LD with the QTL then the accuracy will decrease. In this data set, the small number of markers (453) with 18 QTL and consequently the extent of LD, allows for BLUP to be able to produce comparably accurate DGV. However if the percentage of genetic variance explained by a single QTL was to be large and the level and length of LD was low, Bayes BLUP could be expected to produce worse results as it would be difficult to spread the large effects of the QTL across a small number SNP in LD with it. This would appear to be especially true with large numbers of SNP. Thus this caveat to using Bayes BLUP should be considered when considering this method.

For all methods, the setting of the hyper parameters from the data may have increased the accuracy rather than following the method in Meuwissen et al (2001).

5.3.2 Computational requirements

The computational time for 30,000 iterations for each model is shown in Table 5.6. Evident is that that Bayes A/B, the only model employing the Metropolis Hastings algorithm is significantly more computationally demanding. However all times are viable compared with the extremely slow time when using the R code as presented in Chapter 3. Also apparent is the equivalence in CPU time required for Bayes A, Bayes BLUP and Bayes SSVS.

Table 5.6- CPU Time for 30,000 Iterations for the different hierarchical models

Prior	*Bayes A*^{1}	*Bayes A/B*^{2}	*Bayes BLUP*^{1}	*Bayes SSVS*^{1}
CPU Time*	24 min	177 min	24 min	24 min
Program	C++	C++	C++	C++

NG- normal-gamma mixture distribution, NE- normal-exponential mixture distribution , ^{1}Models utilising the Gibbs Sampler, ^{2}Models utilising the Metropolis Hastings algorithm and Gibbs Sampler, * min= minutes

5.4 CONCLUSION

All methods produced DGV that were highly correlated (greater than 0.85) with the true breeding values despite diverse assumptions and prior distributions. This indicates that the hierarchical model is relatively insensitive to the choice of prior distributions for this data set. The results cannot be seen to reflect the match between the prior and the true distribution of QTL possibly due to the mismatch between the method in which that data was simulated (using a logistic growth curve) and the approach to predicting the SNP effects (using a linear model). The results presented are also dependent on the underlying the model used to simulate the data and thus if this was to change the performance of the models may change. However, the Bayesian models do appear robust and are able to handle this lack of match between the simulated data and the model assumptions. This bodes well for real data where the genetic architecture of the trait may be unknown.

The study does show that methods assuming unequal variances (Bayes A, Bayes B and Bayes A/B) produced very similar sets of DGV in comparison to a significantly different set of DGV produced by the Bayesian genomic BLUP approach. However, Bayes BLUP produced DGV highly correlated with the true breeding values (TBV) indicating that it may provide a viable approach to genomic prediction in real data where the population is heavily structured.

A comprehensive comparison of Bayesian methods in real data is presented in the next chapter where the performance of Bayes BLUP, Bayes A and Bayes SSVS is examined across a range of traits differing genetic architecture. The results are then discussed and compared with comparable studies from different countries. In addition, the effect of pre-selecting a subset of SNP is explored for the Bayes A and Bayes BLUP methods (where all SNP included have a non-zero effect).

CHAPTER 6
Comparison of Bayesian Methods for genomic selection using real dairy data

6.1 INTRODUCTION

Real data emphasizes the *p>n* problem that simulated often data does not mimic. Currently, only a few thousand of animals with genotypes and phenotypes are available for use as reference populations and the current bovine SNP chip has 54001 SNP available and this number will continue to increase. This disparity between the number of SNP and phenotypes can be problematic. Consequently, one proposal is to first select a small number of influential SNP that are most likely to be linked to QTL affecting the trait of interest. Then in a second stage use these pre-selected SNP for more sophisticated modelling of the relationship between the SNP and the trait of interest. This approach was initially introduced as a way to pre-select markers for GWAS. Hoh et al. (2000) proposed a two stage analysis using a model-free approach to first select influential markers for further modelling in the second stage. The approach for the pre-selection step was based on a bootstrap procedure. In the context of genomic prediction, Macciotta et al.(2009) used a simple single SNP linear regression model and Long et al. (2007) developed a Machine learning classification procedure both to pre-select SNP for use in the creation of a prediction equation. In addition to reducing the dimensions of the data, the identification of a predictive subset of SNP could also reduce the cost of genotyping animals consequently making the application of genomic selection more cost effective. The raises the question of whether the use of a selected subset of SNP will produce higher accuracies will depend on the LD present between the SNP and QTL and the genetic architecture of the trait. The affect is assessed in this Chapter by preselecting SNP using single SNP linear regression with and without weights and using these reduced subsets of SNP with two genomic prediction models.

As can be seen from the systematic overview presented in Chapter 2, multiple different approaches have been proposed and implemented for genomic selection and prediction. Chapters 3 and 5 analytically considered different Bayesian approaches differentiated by their prior distribution specifications. It was demonstrated that the performance of the models in simulated data reflected the match between the prior distributions of the QTL effects in these models and the “true” or simulated distribution of QTL effects. However, the performance of approaches and the resultant accuracies of prediction are dependent on the detailed assumptions of the simulation and may not be an accurate representation of the performance of these methods in real data. This is because the true distribution of QTL effects is not known in real data. Furthermore the SNP may only be partially in LD with QTL; so it may only be possible to capture part of the effect of some QTL. Thus a good test of a genomic prediction model is its ability to reliably and accurately predict breeding values across a range of genetically diverse traits in real data. Importantly, the ideal model will be robust and able to produce highly accurate DGV for traits with differing genetic architecture. The performance of different models will be related to the equivalence between the model assumptions about the QTL distribution and the real QTL distribution. Thus, the relative performance of the different models can also be seen to give insight to the underlying genetic architecture of the traits.

This chapter presents an extension of the investigation presented in Chapter 3 and 5 to real diary data to obtain a more comprehensive comparative analysis of the performance of Bayes BLUP, Bayes A and Bayes SSVS. In addition, the affect of selecting subsets of SNP on the accuracy of genomic prediction is explored using real data. The performance of the different methods is assessed across a range of nine traits with differing genetic architectures. For example, for fat percentage, there is a well characterised mutation of large effect (DGAT1). In addition, across the nine traits, the impact on the accuracies for selection of using smaller pre-selected sets of SNP is examined for the Bayes BLUP and Bayes A models.

6.2 METHODS

6.2.1 Data

A total of 1498 Australian Holstein-Friesian bulls were genotyped for the Illumina Bovine50K array. Quality control was applied to the SNP data (see Hayes et al. (2009)) leaving a final set of 39048 SNPs. Of the 1498 animals, 1098 bulls born between 1940 and 2000 formed the reference data set. The remaining 400 genotyped bulls formed the validation set. These bulls were proven in the 2005, 2006 or 2007 with Australian Breeding Values (ABV) including information from at least 100 milking daughters to enable comparison with predicted GEBVs. The phenotypes for these bulls were ABVs for protein kg, fat kg, milk yield, protein percentage, fat percentage, daughter fertility, ASI (Australian Selection Index), APR (Australian Profit Ranking) and overall type. ASI is defined as:

$$\mathrm{ASI}_{\mathrm{ABV}} = (3.8\times \mathrm{Protein\,kg}_{\mathrm{ABV}}) + (0.9\times \mathrm{Fat\ kg}_{\mathrm{ABV}}) - (0.048\times \mathrm{Milk\ litres}_{\mathrm{ABV}})$$

APR is defined as:

$$\begin{aligned}\mathrm{APR}_{\mathrm{ABV}} = {} & \mathrm{ASI}+ (3.9\times \mathrm{Survival\,Index}) + (1.2\times \mathrm{Milking\,Speed}_{\mathrm{ABV}}) + (2.0\times \mathrm{Temperament}) \\ & - (0.26\times \mathrm{Liveweight}_{\mathrm{ABV}}) - (0.34\times \mathrm{SCC}_{\mathrm{ABV}})\end{aligned}$$

where SCC stands for somatic cell count. As in Chapter 4, daughter fertility is defined as the difference between bulls for the percentage of their daughters that are pregnant 6-weeks after mating start date or 100-days after calving in year-round herds. All ABV were deregressed to remove any contribution from relatives other than daughters (Hayes et al., 2009b).

6.2.2 Model and Prior Distributions

The model was the standard model previously presented in Chapters 4 and 5. Briefly, the model was:

$$\boldsymbol{y} = \mu \boldsymbol{1}_n + \sum_{j=1}^{q} \boldsymbol{X}_j \beta_j + \boldsymbol{Zu} + \boldsymbol{e}$$

where y is the vector of deregressed phenotypes of the trait being analysed for all n individuals, μ is the mean, $\mathit{1}_n$ is a vector of ones of length n, $\mathbf{X}_j$ is a vector of indicator variables representing the genotypes of the j^{th} marker for all individuals (x_{ij}=0,1,2), β_j is the size of the effect for marker j, **u** is the vector of random polygenic effects of length n (Z is the associated design matrix) and where $\boldsymbol{u} \sim \boldsymbol{N}(0, \sigma_u^2 \boldsymbol{A})$ and **e** is the residual error also assumed to be normally distributed, $\boldsymbol{e} \sim \boldsymbol{N}(0, \boldsymbol{I}\sigma_e^2)$.

The prior distributions used for this study are presented in Table 6.1. Three hierarchical priors distributions were used. The first, Bayes BLUP (Section 2.2.2, 4.2.4 and 5.2.4) sampled the SNP effects from a normal distribution and assumed that all SNP had an equal variance. This approach results in all SNP having some effect. Similarly, the Bayes A framework produces effects for all SNP. However in contrast to Bayes BLUP, the SNP are assumed to have unequal variance and due to the choice of conjugate priors, the SNP are sampled from a t-distribution (Section 2.2.3). The final hierarchical framework, Bayes SSVS, introduced in Chapter 4, used Stochastic Search Variable Selection (SSVS) to enable only a distinct subset of SNP to be assumed in linkage disequilibrium QTL and thus non-zero. All prior distributions were implemented using the MCMC Gibbs Sampler.

Table 6.1- Prior Distributions Specifications

Method	**Prior Distribution**
Bayes BLUP	$\beta_i \sim \boldsymbol{N}(0,\sigma^2)$ $\sigma^2 \sim \chi^{-2}(\boldsymbol{r},\boldsymbol{s})$
Bayes A	$\beta_i \mid \boldsymbol{v}_i \sim \boldsymbol{N}(0,\sigma_i^2)$ $\sigma_i^2 \sim \chi^{-2}(\boldsymbol{r},\boldsymbol{s})$
Bayes SSVS	$\beta_i \mid \gamma_i,\sigma_i^2 \sim (1-\gamma_i)N(0,\sigma_i^2/100)+\gamma_i N(0,\sigma_i^2)$ $\sigma_i^2 \sim \chi^{-2}(r,S)$ $\gamma_i \sim bernoulli(p_i)$ $1-p(\gamma_i=0)=p(\gamma_i=1)=p_i \quad (p_i=0.05)$

6.2.3 Pre-selection of SNP

To examine the effect that using selected subsets of SNP had on the accuracy of prediction, a pre-selection step was carried out to select sets of SNP that were the most likely to be linked to QTL affecting the trait of interest. These subsets were chosen using single SNP analysis carried out in ASReml (Gilmour et al., 2006a). A maternal grandsire model was implemented as follows:

$$\boldsymbol{y} = \mu\boldsymbol{1}_n + \boldsymbol{X}\beta + \boldsymbol{Z}_1u_1 + \boldsymbol{Z}_2u_2 + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector containing the phenotypic records for all individuals, $\boldsymbol{X}$ is a vector of indicator variables representing the genotypes of the i^{th} SNP marker for all individuals, β is the associated effect of the i^{th} SNP, u_1 is the random sire effect ($\boldsymbol{Z}_1$

associated design matrix) and is normally distributed, $u_1 \sim \boldsymbol{N}\left(0, \sigma^2_{u_1}\boldsymbol{I}\right)$, u_2 is the random maternal grand sire effect ($\boldsymbol{Z}_2$ associated design matrix) and is also normally distributed, $u_2 \sim \boldsymbol{N}\left(0, \sigma^2_{u_2}\boldsymbol{A}\right)$. The model was fitted with and without weights, where the weights were defined as the number of effective records (which for bulls is equal to the number of effective daughters). The model was run for each trait individually fitting all SNP separately. The results of the single SNP analyses were examined and any SNP that was reported with a *p value* < 0.1 was selected as part of the reduced set of SNP for that trait. Thus three sets of SNP were created for each trait. The first containing all SNP, a second set containing SNP selected for the trait not using weights and the final set containing SNP selected for the trait using weights.

6.2.4 Breeding Values

The direct genetic values (DGV) for bulls in the validation data set were calculated as the sum of the effects of the SNP genotypes that it carried, the mean and the estimated polygenic effect; $DGV_i = \mu + \mathbf{X}_i\hat{\beta} + \hat{u_i}$. Genomic Estimated Breeding Values (GEBV) were also calculated. They were calculated by combing the DGV and the sire maternal-grandsire pathway (SP), namely, breeding value predictions based on the sire maternal-grandsire pathway calculated at the time of the birth of the bull calves. The GEBV were calculated as follows:

$\mathrm{GEBV} = \dfrac{\mathrm{w_1 DGV + w_2 SP}}{\mathrm{w_1 + w_2}}$ where $\mathrm{w_i} = \mathrm{R_i^2}/\left(1 - \mathrm{R_i^2}\right)$ for $\mathrm{i} = 1$ for the DGV and $\mathrm{i} = 2$ for the sire pathway (Moser et al., 2009). The reliability, R^2, of the DGV was calculated as the correlation squared between the DGV and the ABV. Similarly, reliability of the sire pathway was calculated as the correlation squared between the SP and the ABV. In practice, the reliabilities would be calculated from the inverse of the Genomic relationship matrix, but this approach provided a quick approximation using optimal weights.

The accuracy of prediction for both DGV and GEBV for the methods was evaluated using the Pearson correlation coefficient.

6.3 RESULTS

6.3.1 Accuracy of Prediction

The Pearson correlation coefficient between the estimated DGV/GEBV and the ABV are shown in Table 6.1 and Table 6.2 respectively. Across all nine traits, no one method produced the highest correlation coefficient or, consequently, the highest accuracy of selection. Regardless, there are clear trends between the different hierarchical prior assumptions. For instance, Bayes BLUP which assumed equal variance across all included SNP, produced DGV and GEBV for protein and ASI with higher accuracies of selection than the other approaches which assumed unequal variances for the SNP. In contrast, the opposite is true for fat percentage. The accuracies produced for APR, fertility and Overall Type, followed very similar patterns of accuracies with only Bayes BLUP, using the SNP selected without weights, producing significantly worse results.

In general, the GEBV produced the more accurate estimated breeding values when compared to the DGV, with the exception of some accuracies for fat and protein percentage. This is highlighted in Figure 6.1 which shows a comparison of the accuracies for Bayes SSVS. Also evident is that without the addition of the sire pathway to create the GEBV, the accuracy of the DGV for fertility is poor and may be a reflection of its low heritabilty. This was true for all prediction models.

Table 6.2- Pearson correlation coefficient calculated between DGV and ABV.

Method	Data[1]	Milk	Protein	Fat	Fat %	Protein %	Overall Type	Fertility	APR	ASI
Bayes BLUP	*1*	0.640	**0.596**	0.477	0.635	0.656	0.524	0.352	0.482	**0.483**
	2	0.617	0.551	0.470	0.684	0.665	0.408	0.291	0.406	0.458
	3	**0.650**	0.581	0.496	0.648	0.650	0.481	0.371	0.452	0.487
Bayes A	*1*	0.598	0.532	0.476	0.699	0.652	0.534	0.341	0.467	0.387
	2	0.610	0.544	0.484	0.713	0.673	0.527	0.277	0.463	0.409
	3	0.605	0.551	0.476	0.705	0.655	**0.540**	**0.353**	**0.489**	0.406
Bayes SSVS	*1*	0.616	0.551	**0.505**	**0.730**	**0.675**	0.525	0.338	0.488	0.431
SP		0.383	0.336	0.363	0.436	0.425	0.311	0.326	0.369	0.284

[1] Data set used for prediction– 1 = All SNP, 2 = SNP selected without weights and 3 = SNP selected with weights. In **bold** is the highest correlation across all models for each trait. SP = Sire Pathway EBV

Table 6.3- Pearson correlation coefficient calculated between GEBV and ABV.

Method	*Data*[1]	*Milk*	*Protein*	*Fat*	*Fat %*	*Protein %*	*Overall Type*	*Fertility*	*APR*	*ASI*
Bayes BLUP	1	0.648	**0.613**	0.528	0.630	0.660	0.576	0.538	0.534	**0.521**
	2	**0.659**	0.610	0.543	0.643	0.661	0.556	0.504	0.516	0.525
	3	0.646	0.596	0.527	0.689	**0.678**	0.506	0.444	0.491	0.504
Bayes A	1	0.631	0.572	0.538	0.700	0.645	0.579	0.539	0.513	0.444
	2	0.635	0.583	0.538	0.704	0.648	**0.585**	0.536	0.531	0.456
	3	0.639	0.579	0.543	0.712	0.667	0.578	0.529	0.519	0.461
Bayes SSVS	1	0.644	0.588	**0.557**	**0.728**	0.670	0.575	**0.540**	**0.535**	0.476

[1] Data set used for prediction– 1 = All SNP, 2 = SNP selected with weights and 3 = SNP selected without weights. In **bold** is the highest correlation across all models for each trait.

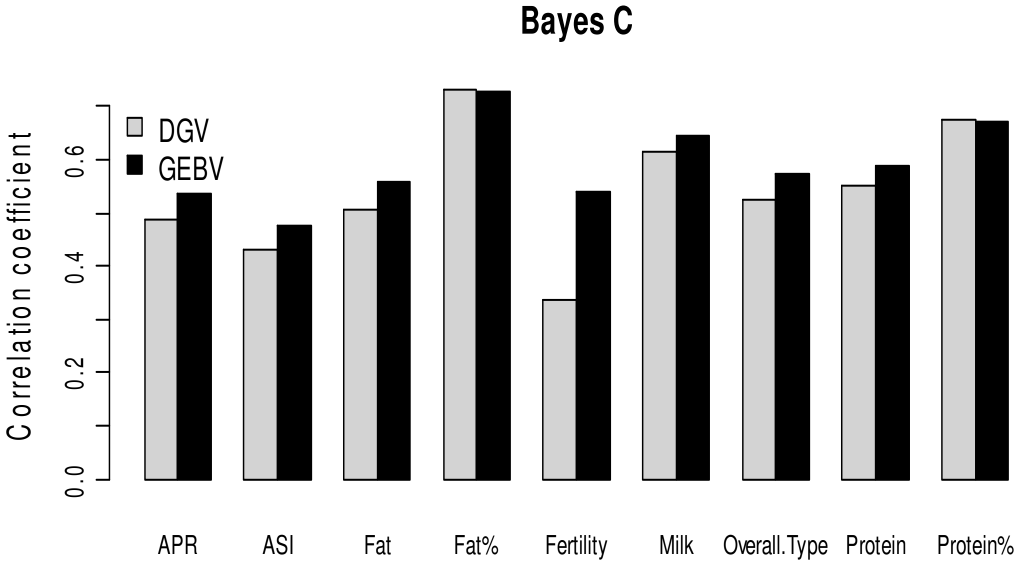

Figure 6.1- Comparison of the reliabilities for DGV and GEBV for Bayes SSVS.

6.3.2 Pre-selection of SNP

The pre-selection step produced two subsets of SNP for each trait. The number of SNP selected for each trait using the single SNP model with and without weights is shown in Table 6.4. Immediately evident is that using weights (and thus the correct error term) increased the number of SNP selected using a 0.1 significance level. In fact, it can be seen that 2 to 5 times as many SNP were selected using the weighted analyses.

Table 6.4- Numbers of SNPs pre-selected using the single SNP model with and without weights.

Trait	*Unweighted*	*Weighted*
Protein kg	5545	14489
Fat kg	4546	19280
Milk (L)	5177	18690
Protein %	5747	20318
Fat %	4518	23919
ASI	5503	16719
APR	5007	11088
Overall Type	5214	15129
Daughter Fertility	4144	10346

The difference in the number of SNP has a significant impact on the performance of the approaches. The trend across the number of SNP was not uniform across Bayes A and Bayes BLUP as seen in Figures 6.2 and 6.3 and Table 6.5 which clearly illustrate two main facts. The first is that Bayes BLUP and Bayes A perform differently when the different sets of SNP are used. Generally, Bayes A performed the best with the two smaller subsets of SNP compared with Bayes BLUP which performed better with more SNP (the larger subset and all SNP). The exceptions to this fact were for fat percentage and protein percentage where for both Bayes A and Bayes BLUP, the results show that accuracies are highest with the smallest subset (Figures 6.2 and 6.3). The second feature shown is that, other than for fertility, the addition of the sire pathway does not change the order of accuracies for the different SNP sets. However, for fertility, the ranking of accuracies for the different SNP sets change with the addition of the sire pathway. This could be a reflection of the non-normality of the trait and its low heritability.

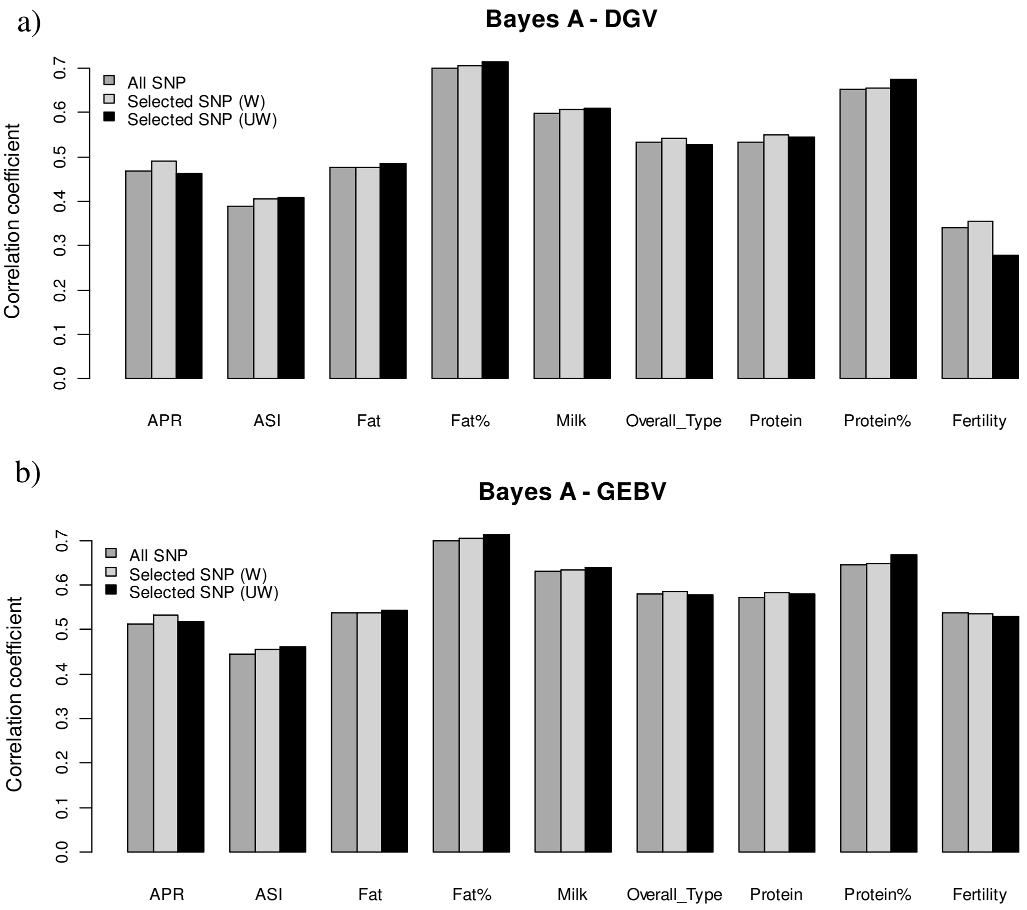

Figure 6.2- Performance of Bayes A for the three different sets of SNP for a) DGV and b) GEBV

Table 6.5- Average Correlation Coefficient across all traits for Bayes A and Bayes BLUP using different number of SNP

Number of SNP	Bayes BLUP		Bayes A	
Number of SNP	DGV[2]	GEBV[2]	DGV[2]	GEBV[2]
ALL -39048	0.538	0.583	0.521	0.573
Selected(W)- 16664[1]	0.535	0.580	0.531	0.580
Selected(UW)- 5045[1]	0.506	0.564	0.522	0.581

[1] Average Selected number of SNP across all traits, [2] Average Correlation Coefficient across all traits, Selected (W) = SNP selected in single SNP model with weights, Selected (UW) = SNP selected in single SNP model without weights.

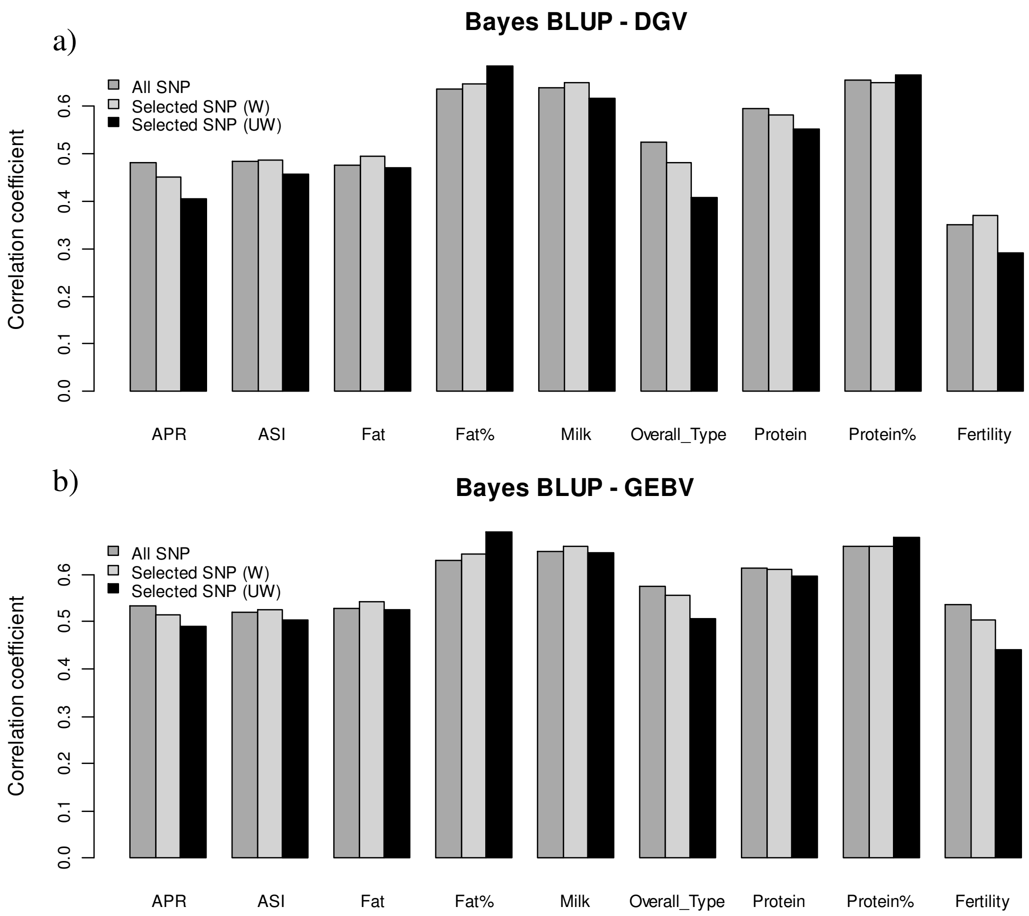

Figure 6.3 - Performance of Bayes BLUP for the three different sets of SNP for a) DGV and b) GEBV

6.4 DISCUSSION

6.4.1 Accuracy of Prediction

Over all, the accuracies produced by the different models are generally comparable. But there are observable differences between accuracies across the traits produced by the same model and differences between the accuracies across the models for the same trait. These differences can be traced, in part, to the variation in genetic architectures of traits and the subsequent match between the models' assumptions about the distribution of SNP effects and the trait's genetic architecture that is the "true" distribution of QTL effect. The most obvious difference in accuracies is for fat percentage. This can be explained largely by the fat percentage's genetic architecture has a single QTL with a much larger effect than any other traits. DGAT1 is reported to have a significant effect and thus the two models (Bayes A and Bayes SSVS) that assume SNP have unequal variances produce higher accuracies for fat percentage. This follows the trend reported in other studies. Hayes et al (2009c) reported that results from New Zealand (Harris et al., 2008), the Netherlands and United States studies (VanRaden et al., 2009) also found that Bayesian Methods performed slightly better than BLUP for traits where there is a single QTL that explain a large proportion of the genetic variance. This trend is also evident here in both DGV and GEBV for fat percentage and protein percent.

This trend of Bayesian methods to out-perform BLUP for traits with QTL explaining a large proportion of the genetic variance also occurs in breeds other than Holsteins. Gredler (2009) found a similar trend in Fleckvieh bulls when they tested Bayes A, Bayes B, linear (BLUP) (VanRaden, 2008), the LASSO and PLS for use for genomic prediction. They too found the same trend of Bayes B significantly out-performing the other method for fat percentage with a general equality of performance across the other methods.

Also apparent from the results (Table 6.2 and Table 6.3) is that BLUP actually performs slightly better than the Bayesian approaches for some traits such as Protein kg and ASI. This again may be linked to the genetic architecture of the traits where a very large number of QTL explain small amounts of variation. This feature means that the "true" distribution of QTL effects more closely matches with the assumptions made about the SNP effects distribution by Bayes BLUP than Bayes A and Bayes SSVS. Alternatively for lower heritable traits, this could be a reflection of the fact that BLUP may better capture relationships in the SNP effects than the other methods.

However, the performance of BLUP is heavily dependent on the extent of the LD. When the LD decays quickly over distance, BLUP would be unable to spread the large effects of a QTL across the one or few SNP in LD with it and thus would be inaccurate and unviable for genomic prediction. Bayes BLUP would be expected to produce worse results in situations such as multi-breed data sets. For example, within breed results have indicated that genomic selection can be successfully carried out with 50,000 SNP (assuming an $r^2 \geq 0.20$ between adjacent markers within breed). However across breeds (e.g. Jersey, Holstein-Friesian and Angus cattle) to have the same LD ($r^2 \geq 0.20$), approximately 300,000 markers are needed to obtain consistent marker effects across these breeds (de Roos et al., 2008). This need to be able to spread the effect of the QTL across multiple SNP also explains why BLUP can not produce as high accuracies for fat percentage. The large effect and variance related to DGAT1 simply does not have enough SNP in LD with it to be able to capture the entire effect and variance caused by the QTL.

For other traits such as overall type and APR, similar accuracies are produced independent of the model that is used. This indicates that there is most likely some moderate QTL and well as a large number of minor QTL. This genetic architecture allows both a BLUP approach assuming equal variances across SNP and Bayes A and Bayes SSVS approaches assuming unequal variances for the SNP to produce comparable DGV and GEBV.

Fertility was a difficult trait to predict with the ranking of models based on the accuracies changing between the DGV sets and GEBV sets. This is most likely due to the non-normality of the trait and its low heritability; consequently it was difficult to establish a stable prediction equation. Additional animals within the reference population should improve the ability to produce more accurate DGV and GEBV for traits with low heritability such as fertility.

The apparent equality of models when accounting for all traits and the robustness and ease of BLUP implementation (using a genomic relationship matrix and the traditional mixed model equations) has lead to many countries adopting a linear BLUP approach (Berry and Kearney, 2009, Reinhardt et al., 2009, Schenkel et al., 2009, VanRaden et al., 2009). In contrast, the Netherlands (de Roos et al., 2009) and the Nordic countries (Lund and Su, 2009) have implemented versions of Bayes SSVS.

The accuracies of selection (Pearson correlation coefficient) of the GEBV found in this study for the Bayesian methods using all SNP, five other approaches applied to the same data as presented here (GEBV calculated using the same approach) and the models used and results reported in other real data studies are presented in Table 6.6. This comparison of results across studies is made difficult by the disparity in the numbers of animals in the reference populations and the numbers of traits analysed.

Nevertheless it is obvious from Table 6.6, that the studies with the largest reference population generally produced higher accuracies. This outcome is expected due to the increase in information about the effects of SNP alleles provided by more phenotypic records (Hayes et al., 2009c, Usai et al., 2009, VanRaden et al., 2009). The relationship between the number of records and accuracy is determined by the number of QTL and the heritability of the trait; this has been reported elsewhere such as in Daetwyler et al. (2008) and Goddard (2008). This relationship is explored in Chapter 8 and modified for use to predict accuracies that are determined between phenotype and DGV/GEBV.

Also affecting the accuracy of the DGV and GEBV is the relationship between the reference population and the validation (or selection) population. In the dataset presented in this chapter the animals in the validation and reference sets are related. This is ideal as it has been demonstrated that the prediction equation produces the most accurate DGV when the animals in the reference population are related to the selection candidates (Habier et al., 2007, Habier et al., 2010b). Additionally, if the prediction equation is to be used across genetically different populations, then animals from each distinct population need be present in the reference population. Muir (2007) showed that the reference populations should contain animals from multiple generation in order to create a prediction equation persists for longer across generations.

Across all the Bayesian approaches and the other additional approaches applied to the data (Table 6.6), there are little differences between the approaches. The Bayes SSVS results found in this study and those reported by de Roos et al. (2009) differ. However these can again be explained by the difference in reference population size. Schenkel et al., (2009) and Berry and Kearney (2009), using comparable size reference populations to this study, report very similar accuracies. Schenkel et al., (2009) have a slightly lower average accuracy of prediction possibly caused by the larger number of traits analysed including more traits with lower heritability.

Table 6.6 - Accuracy of Selection for GEBV

Reference	Dataset (reference pop size, no. of SNP, no. of traits)	Method	Range	Average
This study	1098, 39048, 9	Bayes BLUP - All SNPs	0.52-0.66	0.59
		Bayes A - All SNPs	0.44-0.70	0.58
		Bayes SSVS	0.48-0.73	0.60
		Bayes A Haplotypes[1]	0.49-0.70	0.60
		GBLUP[1]	0.52-0.66	0.59
		LASSO[1]	0.46-0.74	0.60
		PLS – All SNPs[1]	0.53-0.67	0.59
		SVR – All SNPs[1]	0.53-0.67	0.59
De Roos et al. (2009)	3600, 48000, 12	Version of Bayes SSVS[2]	0.52-0.83	0.71
VanRaden et al (2009)	5335, 38416, 27	Linear (BLUP) [3]	0.57-0.80	0.70
		Non-linear [3]	0.59-0.88	0.71
Schenkel et al., (2009)	1179, 38416, 34	Linear (BLUP) [3]	0.35-0.73	0.56
	4127, 38416, 34	Linear (BLUP) [3]	0.36-0.77	0.59
Berry and Kearney (2009)	945, 42598, 20	Linear (BLUP) [3]	0.56-0.71	0.66
Reinhardt et al. (2009)	3684,45181,26	Linear (BLUP) [3]	0.57-0.85	0.72

[1] Methods run on the same data- Bayes A Haplotypes – Bayes A as described but run using haplotypes, GBLUP as described in Hayes et al.(2009a), LASSO as described in Usai et al (2009) , PLS and SVR as described in Moser et al (2009b).

[2]Version of Bayes SSVS based on the work of Meuwissen and Goddard (2004) and presented by Calus and Veerkamp (2007) and Calus et al., (2008)

[3]Methods described in VanRaden (2008)

The difference in the heritability of the traits analysed will also have affected the accuracies. This is due to the relationship previously mentioned between the accuracy of prediction, the number of SNP, the numbers of phenotypic records and the heritability. The higher the heritability, the less phenotypic records are needed to achieve a high accuracy of prediction. Again, this relationship is further discussed and elaborated using Energy Balance as an example in Chapter 8.

Another difference between the results could be caused by the method to construct the GEBV. Currently, most countries that have or will implement genomic selection do not select on DGV alone but combine the DGV with traditional breeding and selection information in the form of national EBV, Parent Average (PA) or predictions based on additional pedigree information i.e. sire and maternal pathways (Berry and Kearney, 2009, de Roos et al., 2009, Harris and Montgomerie, 2009, Reinhardt et al., 2009, Schenkel et al., 2009, VanRaden et al., 2009). This addition is reported to add vital parental information that is not fully contained in the DGV despite the inclusion of the polygenic effect. This information is not contained in the DGV due to the small subset of the data used in the prediction analysis. This inclusion also increases the accuracy of prediction and selection of the breeding values. For instance, this is clearly shown in Figure 6.1 using the results of Bayes SSVS as the example. Interestingly, as shown in the previous sections (6.3.2 and 6.4.2), for some models and traits, this additional information does change the ranking of the models and changes which model produces the best results. This can be explained by the proportion of genetic variance accounted for by the SNP effects for the different traits. Thus the amount of extra accuracy the EBV or PA will add to the GEBV will be trait dependent and should be able to be predicted.

Some studies use selection index theory to construct the final GEBV while others use a BLUP approach introduced by Ducrocq and Lui, (2009) (for example Reinhardt et al. (2009)). The approach used in this study incorporated predictions based on the sire maternal grandsire pathway and is weighted by the respective reliabilities. The sire maternal grandsire pathway is referred to by de Roos et al. (2009) as the sire pedigree index. de Roos et al. (2009) also include the other components that make up the national EBV, including the maternal pedigree index and the Mendelian sampling effects, to evaluate GEBV. A selection index that is more widely used to compute GEBV, combines traditional PA/EBV calculated using the traditional additive relationship matrix, a subset of PA/EBV calculated using only the subset of genotyped ancestors and the DGV (Berry and Kearney, 2009, Schenkel et al., 2009, VanRaden et al., 2009); the weights for the selection index are again calculated using the respective reliabilities. This difference in calculation may have an effect on the overall accuracies reported. For example, Reinhardt et al. (2009) report higher accuracies of prediction than those reported by VanRaden et al. (2009) for the same linear (BLUP) approach. This is despite the study reported by VanRaden et al. (2009) having 1651 more animals in their reference population while comparing an equivalent number of traits. This discrepancy could be caused by the method used to construct the GEBV with Reinhardt et al. (2009) utilising the BLUP approach of Ducrocq and Lui, (2009) while VanRaden et al. (2009) applying a slightly different selection index approach.

6.4.2 Pre-selection of SNP

The effect of the pre-selection of SNP was not considerable but the differences observed were dependent on both the model and trait with which the selected subsets were used. This is clearly shown in Figures 6.2 and 6.3 (Section 6.3.2) and in Table 6.4. For Bayes BLUP, the general relationship between the number of SNP and the accuracy of selection was positive; generally the highest correlations were found when all SNP were included in the model. The opposite was true for Bayes A where the two subsets of SNP performed better on average than when all the SNP were included. The explanation for this is that decreasing the number of SNP used has two opposing effects. Firstly, it reduces the amount of information available which tends to decrease the accuracy of DGVs; this explains why, for most traits, BLUP produces the highest accuracies with all SNP. However the reduction in the number of SNP used, may for traits with a large QTL (such as fat percentage), increase the size of the effect estimated for SNP who are in LD with the QTL whose effect is therefore larger than that of most SNP. Thus for fat and protein percentage, both Bayes A and Bayes BLUP have highest accuracies when the smallest pre-selected subset are used (Figures 6.2 and 6.3). This feature, as previously stated, can be linked to the architecture of both traits, namely, a single QTL of large effect. Even in the BLUP analysis where equal variance is assumed, decreasing the number of SNPs increases the variance ascribed to the remaining SNPs and so allows their estimated effect to be greater. The effect of using pre-selected SNPs with Bayes A for fat percentage may make the analysis more like Bayes SSVS which produced the highest accuracy.

While these results indicate that the pre-selection of a subset of SNP neither significantly benefits nor is detrimental to the accuracies of prediction produced. The method of pre-selection and the make-up of the final subset will obviously influence the results. Macciotta et al.(2009) and Gonzalez-Recio et al. (2008) both present genomic selection approaches using subsets of SNP. Macciotta et al.(2009) used an equivalent approach to this study whilst Gonzalez-Recio et al. (2008) used the machine learning procedure presented by Long et al. (2007). Neither compared the effect of the use of the subsets of SNP, versus the inclusion of all SNP, on the accuracy of prediction. Habier et al. (2009), using simulated data, examined the effect of both subsets of selected SNP and of SNP evenly spaced across the genome. They employed both a Bayesian approach (Bayes B, Section 2.2.4) and forward stepwise least squares regression (Section 2.2.1) to select the SNP. All subsets of SNP produced accuracies lower than when all SNP were used. The subset of SNP selected using Bayes B producing the lowest reduction in accuracy. Habier et al. (2009) also reported reduced accuracies when using less dense, evenly spaced, SNP. These results however are dependent on the simulated data and may not be a reflection of these approaches performance in real data.

In real data, VanRaden et al. (2009) reported that the lower SNP densities produced reduced accuracies and reliabilities. As Hayes et al. (2009c) report, the SNP must be in sufficient LD with the QTL to be able to predict the effects of all QTL. Thus a reduction in the set of SNP will also diminish the extent of LD and consequently decrease the ability of any model to predict all the QTL effects. Thus if a reduced-subset of SNP is to be used, this set of SNP should be selected based on an analysis that seeks to identify the SNP most likely to be in LD with the QTL effecting the trait.

The selection of SNP in this study did not significantly increase accuracy of prediction, but it did increase the time and computation demands. The single SNP analysis was time consuming and consequently is not recommended as it provided no convincing additional benefits. It would be of interest to see how many SNP, the set SNP and what accuracies would be produced when the SNP were selected using different models such as a Bayesian model or a machine learning procedure.

With the increase in numbers of SNP available (the next SNP chip is reported to have 850,000 SNP), the ability to pre-select the important features (SNP) related to a trait may again become an important issue as approaches and procedures seek to deal with the dramatic increase in the number of SNP. Thus it could potentially once the SNP are selected significantly reduce the time and computational demands. Additionally, using reduced number of pre-selected SNP would also provide significant economical savings by requiring selection candidates to be genotyped for only the smaller number of SNP.

6.5 CONCLUSION

The accuracies of prediction produced by all models were relatively equal with the exception of fat percentage due to the influence of DGAT1. These results agree with the results previously published (Berry and Kearney, 2009, de Roos et al., 2009, Gredler et al., 2009, Harris et al., 2008, Lund and Su, 2009, Reinhardt et al., 2009, Schenkel et al., 2009, VanRaden et al., 2009). The study shows that the Bayesian methods provide a valid approach to genomic selection however the uniformity of results means that less computationally demanding approaches are attractive. Consequently the robustness and ease of application of the genomic BLUP approach has lead to many countries adopting this approach for their genomic prediction model. The selection of subset of SNP showed neither a major increase nor decrease in accuracy, but did show different trends across models most likely as a response to the distribution of QTL effects assumed. The use of subsets and models for selection may become important as the number of SNP increases. The comparison of this study with the results of other studies reinforced the known fact that the number of animals in the reference population is a key parameter determining the accuracy of DGV. Also raised for consideration is the methodology to construct the GEBV and the affect this has on the accuracy of prediction.

CHAPTER 7
Significance testing for whole genome multi-locus models using permutation tests

7.1. INTRODUCTION

Most quantitative traits are complex traits with numerous genetic factors contributing to the genetic variation. Identifying these factors could be beneficial for biomarker identification, marker assisted selection and identification of possible drug targets. In addition to the significant focus on genomic prediction and selection in livestock, the availability of dense SNP panels has also led to an increase in genome-wide association studies aiming to identify QTL (Goddard and Hayes, 2009, Hardy and Singleton, 2009, McCarthy et al., 2008). Furthermore having increased information on a trait's genetic architecture could result in more accurate genomic prediction and selection models.

Considerable efforts have been made over the years to identify QTL (quantitative trait loci) across a range of species. Traditionally, models have included one QTL or examined one marker interval at a time (Jansen, 1993, Knott and Haley, 1992, Lander and Botstein, 1989, Luo and Kearsey, 1989, Martinez and Curnow, 1992, 1976, Weller, 1986, Zeng, 1993, 1994). However, the individual estimation of each marker or interval effect from different models can cause biased results. Results may become biased through the fitting of a single QTL in a model that may be affected by the presence of other QTL not in the model, resulting in false positives (a significant QTL is found where there is in fact not a QTL), false negatives (no QTL is found where there is actually a QTL) and reporting incorrect levels of significance and size (e.g. the Beavis effect (Beavis, 1994)).

A further problem is caused by the multiple estimates of the residual variance leading to problems when calculating the total phenotypic variance. Consequently, an increasingly popular alternative is to fit multiple markers in a single model (Baierl et al., 2006, Bogdan et al., 2004, Kao et al., 1999, Narita and Sasaki, 2004, Shriner, 2009, Sillanpaa and Arjas, 1998, Storey et al., 2005, Xu, 2003, Zou and Zeng, 2009).

These include both frequentist and Bayesian modelling approaches. A selection of these models (e.g. Storey et al. (2005)) do not fit all the SNP in a single model, but perform sequential tests of markers (or sets of markers) to determine the significant set of markers and then fit a final model containing the selected set of markers. In contrast, Bayesian inference models are able to fit all possible markers into a single model (Narita and Sasaki, 2004, Satagopan et al., 1996, Sen and Churchill, 2001, Shriner, 2009, Stephens, 1998, Wang et al., 2005, Xu, 2003, Yi and Xu, 2008, Yi, 2004, Yi et al., 2003, Yi et al., 2007, Yi et al., 2005, Zhang et al., 2005). These models provide two distinct advantages over single SNP analysis or a sequential scheme to select SNPs. The first is that fitting all SNP simultaneously in a single model will result in more precise locations of the QTL: a multi-locus model will identify only the SNP or set of SNPs in LD with the QTL which best explain the effect, rather than all SNPs which are in LD with the QTL. This is an advantage particularly in livestock, where low level LD can extend for more than 1 Mbp (The Bovine HapMap Consortium (2009)). The second is that the residual variance is reduced in a single analysis which will result in more power to identify QTL.

Despite the advantages of using Bayesian multi-locus models, it is still computationally demanding with the ever increasing number of markers (p) to explore the entire sampling space of all possible models (2^p) that contains all possible combinations and numbers of markers. In humans, the current SNP chip has over 900,000 SNP and the next bovine SNP chip is anticipated to have over 850,000 SNP. An inability to fully explore the total sample space and test all possible alternative hypotheses leads to its own problem of biased results.

To be able to deal with this abundance of markers juxtaposed with the smaller number of observations, Bayesian multi-locus models utilize model selection procedures or shrinkage estimation. Shrinkage estimation models include all candidate markers but their estimated effects are forced to shrink toward zero with the larger effects being shrunk the least (Meuwissen and Goddard, 2001, Wang et al., 2005, Xu, 2003). Thresholds can then be used to determine which SNP should be included in the final model. Model selection approaches seek to identify the significant parameters (Swartz et al., 2006, Yandell et al., 2007, Yi, 2004, Yi et al., 2007). One model selection approach, Bayesian Stochastic Search Variable Selection (SSVS), introduced by

George and McCulloch (1993, 1997) was discussed in Chapter 4. This model selection approach fits a hierarchical latent variable model that allows the selection of the most promising models for further investigation. Yi et al. (2003) first propose SSVS for use in QTL mapping and their approach has been widely applied and is available in the R package *qtlbim* (Yandell et al., 2007, Yi and Shriner, 2008, Yi, 2004, Yi et al., 2007, Yi and Xu, 2000, 2002, Yi et al., 2005).

As with single locus models the question of how to declare QTL significant is still an important issue when using multi-locus models. Different approaches for establishing significance have been suggested and used in multi-locus methods. These include using the false discovery rate (FDR) and variants of FDR (Conlon et al., 2006, Storey, 2003, Zou and Zeng, 2009) and Bayes Factors (Kass and Raftery, 1995, Shriner, 2009, Yi and Shriner, 2008). Storey (2003) show that positive FDR can be written as a Bayesian posterior probability under specific assumptions while Genovese and Wasserman (2002) developed a Bayesian version of FDR which has been used to declare significance (Conlon et al., 2006, Do et al., 2005, Heuven and Janss, 2010).

Another alternative to establish significance thresholds is permutation testing. This has been widely employed for identifying QTL using single locus models. Permutation testing was first suggested by Fisher (1935) and was introduced to QTL mapping by Churchill and Doerge (1994). Permutation tests work by breaking any association between two variables by permuting (shuffling) the data. This enables the establishment of a distribution of test statistics in the absence of any association. This distribution can then be used to declare if there are any real associations between the two variables. The null hypothesis for permutation testing for the multi-locus model is H_0: that no SNP is linked to a QTL. The possible alternative hypotheses are 2^p-1 in number (for p SNPs) where there is a hypothesis reflecting each possible combination of SNPs being found to be linked to each QTL. The number of the alternative hypotheses actually tested is equal to the extent of the model space explored. Thus an approach such as SSVS that explores the most likely models will test the most appropriate alternative hypotheses. Due to the assumption that with such large numbers of genome-wide SNPs, each QTL will be in LD with at least one SNP, further possible alternative hypotheses that a QTL is present but is not linked to any SNP is ignored in this study.

Nonetheless, permutation testing has been rarely used with Bayesian multi-locus models. Xu (2003) used 10 permutated data sets to examine the null distribution (that is that there is no association between genotype and phenotype). Churchill and Doerge (1994) report that when using single marker models, for a reliable p-value of 0.05, at least 1000 permutations are needed while 10000 permutations are suggested as sufficient for a reliable p-value of 0.01; for more extreme p-values, even more permutations are needed. Consequently, the 50 permutation tests used by Bauer et al. (2009) appear drastically too few in number to be able to reliably to declare significance even at a 0.05 significance level.

Permutation testing is generally a simple approach as there is only one independent variable and thus all individuals are exchangeable under the null hypothesis. This means that under the null hypothesis of no QTL effecting the trait, the observations must be able to be exchanged i.e. any order of the observations is equally likely. Permutation tests unlike other parametric tests require only this mild condition of exchangeability to be satisfied. A problem arises however if there is a second independent variable. A good example is livestock populations where phenotype, marker and structured pedigree are available. With such structured pedigrees, individuals may not be equally exchangeable. For example, within a set of individuals, some may have the same sire and consequently their genotype is not equally likely amongst all individuals but more likely amongst those with the same sire due to the shared inheritance of one allele from the common sire. Thus the condition of exchangeability may not be satisfied. Consequently, the question is what effect this pedigree-genotype relationship, or ignoring it, has on the declaration of significance.

In the study reported here, a thorough exploration of permutation testing for Bayesian multi-locus models was performed. The issue of exchangeability with a second variable (a pedigree) is examined and different procedures for accounting for this are explored. A further extension of the permutation testing using multi-locus models is demonstrated using the idea of permutation within genotypes classes presented by Doerge and Churchill (1996) based on work of Lehmann (1986). The results are evaluated and compared to previously proposed measures of significance, namely, Bayes Factors and posterior expected False Discovery Rate (*PeFDR*). In addition, the number of permutation tests needed is again explored in the context of multi-locus models. A Bayesian hierarchical latent variable multi-locus model is employed similar to that introduced in Chapter 4. The use of a Bayesian SSVS model for QTL mapping was first suggested by Yi et al. (2003) and this model is based on the work Meuwissen and Goddard (2004). The approach is demonstrated through two simulation studies and a real data example.

7.2. METHODS

7.2.1 Model

A Bayesian multi-locus hierarchical latent variable model using SSVS very similar to the model described in Chapter 4 was used to perform the genome-wide association studies. The model used here expands upon the work of Meuwissen and Goddard (2004). This approach, as in Chapter 4, uses the latent variables to indicate whether a SNP has a significant effect (i.e. is linked to a QTL) and is included in the model. The model can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{1}_n \mu + \sum_{j=1}^{m} \left(\mathbf{X}_j \left(\mathbf{q}_j v_j \right) \right) + \boldsymbol{Zu} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of phenotypes of the trait being analysed for all n individuals, $\boldsymbol{1}_{\mathrm{n}}$ is a vector of ones of length n, μ is the mean, *m* is the number of SNP markers, $\mathbf{X}_j$ is the (n x *k*) design matrix containing the information on the possible *k* alleles at the j^{th} marker for all individuals (where x_{jk}=0,1,2 having no, one or two copies of the k^{th} allele respectively), $\mathbf{q}_j$ is the vector (*kx1*) containing the effects of all *k* possible alleles at locus *j* where $q_{jk.}$ are drawn from a standard normal distribution *N*(0,1), v_j is the standard deviation of the allelic effects at locus *j* and is dependent on whether the locus effect is considered significant or not using the latent variable e.g. v_j is sampled: $\mathrm{p}(v_j \mid I_j) \sim \boldsymbol{N}(0, I_j + (1 - I_j)/100)$ where $I_j = 1$ if the SNP has a significant effect and conversely $I_j = 0$ if the SNP has a very small effect, where the prior distribution is $\mathrm{I}_j \sim \mathit{bernoulli}(p_i)$ (where $p_i = 0.05$ for all examples), **u** is the vector of random additive polygenic effects of length *n* (***Z*** is the associated design matrix) and is assumed to be normally distributed, $\boldsymbol{u} \sim \boldsymbol{N}\left(0, \sigma_u^2 \boldsymbol{A}\right)$ where $\boldsymbol{A}$ is the pedigree-derived additive genetic relationship matrix and **e** is the residual error also assumed to be normally distributed, $\boldsymbol{e} \sim \boldsymbol{N}\left(0, \boldsymbol{I}\sigma_e^2\right)$ where $\boldsymbol{I}$ is the n x n identity matrix. The allele substitution effect of a locus *j* can be calculated from the estimated effects as: $a_j = \left(q_{j1} - q_{j2}\right)v_j$ where q_{j1} (q_{j2}) is the effect of allele 1(2) at locus *j*. For the full specification of the priors used and an alternative formulation of the model see Calus et al. (2008) and Meuwissen and Goddard (2004).

The models were run for 10,000 iterations with 2000 iterations used as burn in for the simulation study and 20,000 iterations with 5000 iterations used as burn in for the real data example. No thinning was performed. This appeared sufficient for convergence and was tested using the formal diagnostic methods provided in the package R, coda (Plummer et al., 2007 -a).

Due to the Bayesian nature of the model used to analyse the data, the "test statistic" used in this study and the presented examples are the posterior probabilities of the SNPs. The model explicitly produces a posterior probability for each SNP through the inclusion of the latent variable (by calculating the number of times the SNP had a latent variable $I_j = 1$ over all iterations excluding the burn in period). The prior probability of $I_j = 1$ can be set (as presented) or estimated. Once set or estimated with the observed data, this value must be kept the same for all permuted data sets. The posterior probability can not be used directly confidently unless the value is extremely close to one. The posterior probabilities are dependent on both the data and the prior distribution and thus can be heavily affected by the choice of prior and prior parameters.

7.2.2. Permutation

Permutation tests permute the data to effectively destroy any association between the two variables to test if there is association e.g. genotype and phenotype. The un-permuted data is first analysed and test statistics for every marker are produced; the data is then permuted N times and analysed to create the null distribution. The null hypothesis is that no SNP is linked to a QTL, thus the null distribution is a distribution of test statistics created in the absence of any real association, in this case caused by permuting the data. If there is a QTL linked to a marker then breaking the relationship between the marker and phenotype will change the distribution of the test statistic; conversely, if there is no association, the distribution of the test statistic will not change. A comparison of the original test statistic and the distribution of test statistics created by the permuted data will allow the assignment of significance.

Normally when permuting the data, the phenotype data is shuffled as described in Fisher (1935). The individuals are indexed from 1,…,n and then the trait values are randomly permuted. The ith trait value after the permutation is then assigned to the ith individual. This is generally the simplest approach, as the data includes only a genotype – phenotype relationship. However, in animal breeding, the population is highly structured and consequently there is often a pedigree with relationship to the genotype and phenotypes. One option to assist in removing the effect of population structure when estimating QTL effects is to include a polygenic effect (Kennedy et al., 1992). The inclusion of a polygenic effect may also be useful in human studies where populations are genetically isolated or for some structured ethnic subgroups (Aulchenko et al., 2007). The polygenic effect is fitted to avoid the possibility that SNP effects found as significant are not linked to a QTL but caused by the population structure of the data (derived from the pedigree).

Consequently, when this triangular relationship exists there are three scenarios when permuting data. They are:

1. Permute the phenotypes breaking the phenotype-pedigree relationship, but retaining the genotype-pedigree relationship.
2. Permute the genotype/s breaking the genotype-pedigree relationship, but retaining the phenotype-pedigree relationship.
3. Permute the genotype within pedigree structures (e.g. sire families).

The first scenario for QTL mapping does not allow the estimation of a polygenic effect and QTL may be indicated that are actually artefacts caused by the population structure. The second scenario is plausible and the most simple approach when dealing with individuals with a structured pedigree but can be seen to violate the exchangeability condition for highly structured data. For example, if one sire has a *AA* genotype then all his progeny must get at least one *A* allele and a second sire has a *BB* genotype the all his progeny must get at least one *B* allele. All offspring are not equally exchangeable as any order is not equally likely.

The second scenario is explored in the first simulation study. Under this approach, the individuals are indexed from 1,…,n. The complete genotypes are then randomly permuted. The i^{th} genotype after the permutation is then assigned to the i^{th} individual. There is no modification of the genotypes themselves and all linkage between the markers is maintained. The shuffled data is analysed using the identical model to the original unshuffled data. The resulting test statistics are stored and the process is repeated N times.

The third scenario is the most correct approach as it does not violate the condition of exchangeability. To permute the data, individuals are separated into sire groups and then permuted within these groups. The genotypes of the individuals within each genotype strata are indexed from 1,…,m_k where m_k is the number of individuals in the current group k. The complete genotypes are then randomly permuted within each stratum. The i^{th} genotype after the permutation in each stratum is then assigned to the i^{th} individual within each stratum. The permuted data is then analysed. The resulting test statistics are retained and the process is repeated N times. Doerge and Churchill (1996) suggest that if there is a known major QTL, in order to establish other moderate or minor QTL, individuals can be stratified based on their genotypes (AA, AB or BB) of the conditioning marker or markers associated with the major QTL. The approach taken is therefore identical to permuting inside sire groups except that the strata are the genotype classes of the conditioning marker. This means that there will always be three strata in contrast to a variable number of strata dependent on the number of sires present in the data.

7.2.3. Thresholds

Three different thresholds can be calculated to establish significance. They are the genome-wide threshold (experimentwise), the chromosome-wide thresholds and the SNP-specific (comparisonwise or pointwise) thresholds. The experimentwise and comparisonwise thresholds are described in Churchill and Doerge (1994). The term experimentwise threshold is herein replaced with genome-wide threshold. The comparisonwise or pointwise threshold is herein called the SNP-specific threshold. The SNP-specific threshold describes the critical value used only for determining the significance of an individual SNP. The threshold is calculated from the N test statistics for the SNPs after the N permutations. The SNP specific p-value for each SNP is calculated as the probability of obtaining a test statistic that is greater than or identical to the original test statistic. The previously identified problem with this threshold is that while it provides the greatest power to detect QTL, the type I error rate can become uncontrolled through testing all SNPs where the test statistics are dependent. This is due to the type I error (false positives) rate applying only to the single SNP under consideration. More problematic and critical in a multi-locus model is that the test statistics themselves are dependent and SNP in high linkage disequilibrium may share or exchange random associations, affecting the final null distribution of test statistics created from the permutation testing. Therefore, when testing many loci and declaring significance using the SNP specific threshold, there is a high chance of false positives due to uncontrolled type I error and the use of an incorrect null distributions that was used to establish significance. Consequently, this threshold is unviable for use.

The chromosome-wide and genome-wide thresholds are both calculated by taking the maximum test statistic after each permutation, either for each chromosome or the entire genome. The distribution of those maximum values after the N permutations is then used to calculate the threshold. The genome-wide threshold is calculated from the distribution of maximum test statistics across all SNPs from each permutation. Similarly the chromosome-wide threshold is calculated by taking the maximum test statistics for each chromosome for each permutation to enable the calculation of a threshold for each chromosome. For 1000 permutations, the thresholds with a significance level of 0.05 are set as the value of the 950th test statistic when all the N test statistics are ordered in increasing size.

7.2.4. Bayes Factors and False Discovery Rate

Bayes Factors are the dominant method of Bayesian model testing. They are the Bayesian analogues of likelihood ratio tests. By taking prior probabilities into consideration, Bayes Factors can be used to compare models with and without particular markers (Kass and Raftery, 1995). To calculate a Bayes Factor, B_{12}, let $\boldsymbol{y}$ be the data, H_1 and H_2 be the two possible hypothesis that are being tested (such that H_1: the marker is linked to a QTL v H_2: the marker is not linked to a QTL). Thus $Pr(H_1)$ is the prior probability of the first hypothesis, H_1 and $Pr(H_2)$ be the prior probability of the alternative hypothesis, H_2. Similarly, $Pr(H_1 \mid \boldsymbol{y})$ is the posterior probability of H_1 and $Pr(H_2 \mid \boldsymbol{y})$, the posterior probability of the alternative H_2. Using Bayes theorem, a Bayes Factor comparing hypothesizes H_1 and H_2, for a single SNP, is defined as:

$$B_{12} = \frac{Pr(H_1 \mid \boldsymbol{y})}{Pr(H_2 \mid \boldsymbol{y})} \div \frac{Pr(H_1)}{Pr(H_2)}$$

This can be seen as the ratio of the posterior odds to the prior odds. Consequently, it can be expressed as:

$$\mathrm{B}_{12} = \frac{\mathrm{Pr}(\mathrm{H}_1 \mid y)}{1 - \mathrm{Pr}(\mathrm{H}_1 \mid y)} \div \frac{\mathrm{Pr}(\mathrm{H}_1)}{1 - \mathrm{Pr}(\mathrm{H}_1)}$$

In order to have comparable thresholds, the posterior probability $Pr(H_1 \mid \boldsymbol{y})$ relating to a Bayes Factor of 3.2 was used as the significance threshold. The value of 3.2 is the lowest value of the Bayes Factor that indicates substantial evidence in favour of the first hypothesis (Kass and Raftery, 1995). For example, if the prior probability of a SNP being linked to a QTL is 0.05, the threshold is derived by setting $\mathrm{B}_{12} = 3.2$ and $\mathrm{Pr}(\mathrm{H}_1) = 0.05$ so that:

$$\mathrm{B}_{12} = 3.2 = \frac{\mathrm{Pr}(\mathrm{H}_1 \mid y)}{1 - \mathrm{Pr}(\mathrm{H}_1 \mid y)} \div \frac{0.05}{0.95}$$

Thus $\mathrm{Pr}(\mathrm{H}_1|y)$ can be calculated by rearranging to:

$$\mathrm{Pr}(\mathrm{H}_1 \mid y) = 3.2 \times \frac{0.05}{0.95} \div \left(1 + 3.2 \times \frac{0.05}{0.95}\right)$$

It is also evident from the calculation of such a the threshold that the threshold is in fact only based on the prior probabilities and the set Bayes Factor and thus is not dependent on the data.

The false discovery rate (the posterior predicted FDR- *PeFDR*) in a Bayesian context can be expressed as (Conlon et al., 2006, Do et al., 2005, Genovese and Wasserman, 2002):

$$PeFDR = E(FDR \mid \boldsymbol{y}) = \frac{\sum_{m} \left[(1 - Pr(H_{1m} \mid \boldsymbol{y}))\delta_m \right]}{\sum_{m} \delta_m} \qquad [9]$$

where m is the number of markers, $Pr(H_{1m} \mid \boldsymbol{y})$ is the posterior probability of marker m being linked to a QTL (H_1 : that marker m is linked to a QTL) and δ_m represents the decision ($\delta_m = 0{,}1$) whether, based on the data (and posterior probability), marker m is linked to a QTL. Thus if it is decided marker m is linked to a QTL, $\delta_m = 1$, then the $Pr(H_{1m} \mid \boldsymbol{y})$ of marker m is included in the summation in [9]. For a single marker the *PeFDR* is simply the posterior probability that H_2 is correct, as follows:

$$PeFDR = 1 - Pr(H_1 \mid \boldsymbol{y}) = Pr(H_2 \mid \boldsymbol{y})$$

Thus the *PeFDR* can be used to produce a false discovery rate for each individual locus and for a group of loci thought to be linked to a QTL.

7.2.5. Simulation Study

Two data sets were simulated to demonstrate that the proposed method can be used to identify minor/moderate QTL. In both data sets, an effective population size of 100 animals was simulated for 100 generations. Each animal had a single chromosome of 1 Morgan, generated to have 1000 evenly located markers. Each consecutive generation was formed by generating 100 offspring (50 females and 50 males), their parents selected at random from the previous generation. Generation 101 and 102 comprised 500 animals each, created by crossing 50 or 20 sires and 250 dams randomly selected such that each dam had two offspring and the sires had 4 or 25 offspring respectively. The data sets analysed contained 1000 animals with phenotypes and genotypes. In the first data set, a single additive QTL was created in the centre of the chromosome explaining 20% of total genetic variation. The trait was generated to have a heritability of 0.3. Thus the QTL was moderate and explaining only 6% of the total phenotypic variation. The second data set had two additive QTL (0.43 M and 0.2M) explaining 20% and 6% of the phenotypic variance with a heritability of 0.5. The causative mutations were removed from the both data sets leaving a total of 780 and 767 polymorphic markers within the population after the 100 generations. For details of the simulation program see Mulder et al. (2009).

The first data set was used to establish that permutation could accurately identify QTL. No stratification was used when permuting this data set as each sire had only 5 offspring; a number too small to stratify within. In contrast, the permutation testing for the second data set was carried out twice once without stratification (scenario 2) and a second time permuting within sire groups (scenario 3- stratified permutation within sire families).

7.2.6. Real data example

588 dairy cows were genotyped for the Illumina BovineSNP50 bead chip (54001 SNP in total). Criteria for selecting the final set of SNPs were a call rate of over 90%, a GenCall score > 0.2 and a GenTrain score > 0.55 (Illumina genotype quality statistics), a minor allele frequency of >2.5% and a deviation from Hardy Weinberg equilibrium (χ^2 < 600). Animals with greater than 5% missing SNPs were removed. Non Mendelian error checks were used to identify genotypes of daughters that were inconsistent with their dams. A further pedigree check was performed by comparing the coefficients of the additive genetic relationship matrix and the genomic relationship matrix (VanRaden, 2008). In total, 43011 SNPs and 548 animals were retained. Of these 548 animals, 518 had phenotypes for the trait, fat percentage.

For the dairy trait fat percentage, there is a known, common mutation on centromeric end of chromosome 14. As mentioned previously, DGAT 1 (Diacylglycerol O-acyltransferase 1) is reported to explain greater than 50 percent of the total genetic variation seen in fat percentage (Grisart et al., 2002). Due to one SNP with such a large effect, it is often difficult to find other QTL with more moderate effects. A SNP located at centromeric end of chromosome 14 with a posterior probability of one can be confidently assumed to be in linkage disequilibrium with this mutation. Doerge and Churchill (1996) suggest that permutation testing using stratification within the genotype classes of a known QTL could be used to set thresholds to identify moderate or minor QTL by accounting for the known major QTL. This real data set was used to demonstrate that such an approach was also viable with multi-locus models. The SNP found with a posterior probability of 1 was used as the conditioning marker. The animals were then stratified into 3 groups based on the genotypes of this SNP (e.g. AA, AB, and BB). The permutation then occurred within these three strata and the subsequent data sets were analysed using the same procedure as the original data.

7.3. RESULTS

7.3.1. Simulation Study

The results of the analysis of the first simulated data set revealed two posterior probabilities significantly (>100 fold) greater than the others, one at 0.5M with a posterior probability of 0.503 and a second at 0.65M with a probability of 0.181 (see Figure 7.1). Both posterior probabilities are noticeably lower than the conclusive result of 1. Permutation testing was carried out with 1000 data sets created by permuting across all individuals (scenario 2- no stratification). The results of the permutation produced genome-wide threshold (significance level of 0.05) of 0.203. A Bayes Factor threshold declared both of the peaks significant in comparison to the permutation testings genome-wide threshold where only the SNP at 0.5M is significant; Figure 7.1 shows the both thresholds. If the SNP-specific threshold was incorrect to be used it would declare 58 SNP as significantly linked to a QTL giving a FDR of close to 1. Of the 58 SNP, only two are clearly evident in Figure 7.1 and the second highest peak is a false positive. Calculating the *PeFDR* only including the SNP with the largest posterior probability (so correctly deciding that only this SNP was linked to a QTL) still produces a *PeFDR* of 0.497 which is high and the inclusion of any other SNP only increases the value.

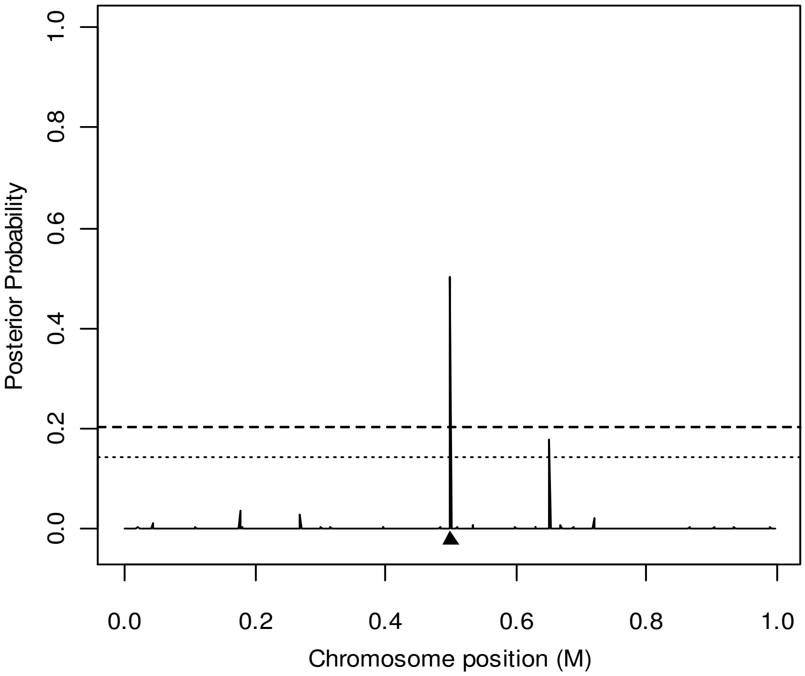

Figure 7.1- QTL analysis of the simulated data set 1. Genome-wide threshold (---) plotted for a significance level of 0.05, Bayes Factor threshold plotted for a Bayes factor of 3.2 (····) with the position of the true QTL indicated (▲).

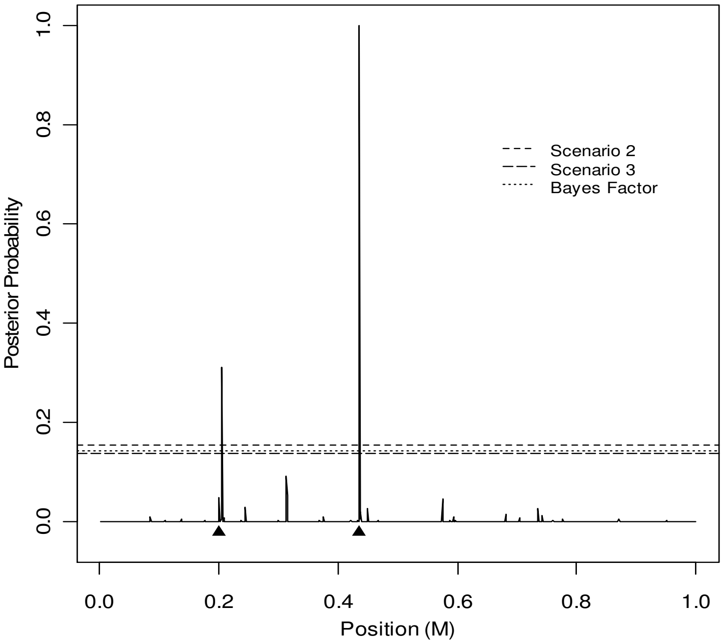

Figure 7.2- QTL analysis of the simulated data set 2. Genome-wide threshold plotted for a significance level of 0.05 for the permutations with no stratification (---) (Scenario 2) and for the permutations with stratification (—) (Scenario 3), Bayes Factor threshold plotted for a Bayes Factor of 3.2 (····) and the positions of the simulated QTL (▲).

The second simulated data set was permuted with and without stratification. The different thresholds produced using these two different approaches are shown in Figure 7.2. When the data was permuted within strata, the 0.05 significance threshold was 0.156 while when the data was permuted randomly across all individuals the threshold was 0.138. The threshold for a Bayes Factor of 3.2 had a posterior probability of 0.144. All thresholds are very similar and produce identical results correctly identifying the two QTL. The *PeFDR* is zero when only the SNP with the highest posterior probability is included and increases to 0.34 when both SNPs with the highest posterior probabilities are included.

7.3.1. Real Data Example

The results of the QTL analysis revealed a SNP with a posterior probability of one at the centrometric end of BTA chromosome 14. This SNP was used as the conditioning marker and three strata were formed using the genotypes of the individuals. Within these strata, the data was permuted 1000 times and each data set was then analysed as in the original analysis. A genome-wide threshold and chromosome-wide thresholds were calculated at a 0.05 significance level. In addition, for comparison, SNP specific thresholds were calculated. A Bayes Factors threshold was calculated for a Bayes Factor of 3.2. The results of using these thresholds including the number of SNPs found significant for each threshold and the associated (average) threshold values are shown in Table 7.1. Also shown are the results of using a Bayes Factor threshold.

Table 7.1- Estimated threshold values and numbers of significant SNP detected

	Genome-wide	*Chromosome-wide*	*SNP specific*	*Bayes Factor*
No. of Thresholds	1	30	43011	1
No. of significant SNP indicated	1*	17	1698	43
Estimated Threshold Value [a]	0.528	0.247[b]	0.002[b]	0.168

Values from 1000 permutations of the data with the posterior probability used as the test statistic, * Excluding the conditioning and surrounding SNP, [a] The posterior probability above which significance can be established at a 0.05 level, [b] mean value for all analysis points (SNP or chromosomes)

Table 7.2 presents the details for the seventeen SNPs found significant at the chromosome-wide significance levels; included are the respective SNP's Bayes

Factors and *PeFDR*. The SNP that was found linked to DGAT1 is not included in the tabulated results as it was conditioned upon and therefore already established as significant. The Bayes Factors threshold declares more than double the number of SNPs as significantly linked to QTL than the chromosome-wide thresholds. Comparatively for a false discovery rate of 0.05 no SNP or combination of SNPs, other than the single SNP associated with DGAT1, is able to produce such a rate.

Each analysis still showed a SNP or SNPs with high posterior probabilities at the centromeric end of chromosome 14. This was expected because the permutation occurred within condition marker genotype classes consequently the phenotype-genotype relationship for these SNPs and the major QTL was not broken. The posterior probabilities of SNPs located in the first 10 Mbp on BTA chromosome 14 were ignored in the calculation of the genome-wide and chromosome-wide thresholds to avoid any SNP in high LD with the conditioning marker that may have had inflated the thresholds.

The reason for conditioning on the SNP and not permuting across all individuals is due to the huge effect of DGAT1. If the permuting had been carried out across all individuals consequently by breaking the relationship between DGAT1 and the phenotypes, the large amount of variance that would have been attributed to DGAT1 would have been transferred to random associations. This would result in inflated SNP effects that would have caused a higher threshold that would most likely have only declared DGAT1 as significant. This result is not helpful as the effect of DGAT1 has already been established.

To highlight that the approach can accurately identify QTL, the focus was placed on three chromosomes (BTA 6, 19 and 26) with QTL previously reported as associated with fat percentage (Druet et al., 2006, Gautier et al., 2006, Khatkar et al., 2004). Figure 7.3 shows the posterior probabilities of all SNPs on BTA chromosome 6, 14, 19 and 26 and the genome-wide and chromosome-wide thresholds at a 0.05 significance level and the Bayes Factor threshold.

Table 7.2 - Details of the 17 Significant SNP detected using the chromosome-wide thresholds including their posterior probability, *Posterior expected FDR* and Bayes factor.

Chromosome	*Position (Mbp)*	*Posterior Probability (p_i)*	*PeFDR (1- p_i)*	*Bayes Factor*
14	0.26	1	0	∞
26	13.46	0.564	0.436	24.60
19	58.89	0.526	0.474	21.10
12	65.37	0.372	0.628	11.26
27	30.96	0.359	0.641	10.64
6	53.61	0.355	0.646	10.43
1	101.60	0.342	0.659	9.85
8	89.64	0.340	0.660	9.80
22	21.08	0.339	0.661	9.74
7	74.67	0.315	0.685	8.75
2	124.27	0.312	0.688	8.60
6	99.24	0.305	0.695	8.34
2	68.54	0.305	0.695	8.32
8	47.19	0.303	0.697	8.25
19	27.06	0.276	0.724	7.24
3	44.13	0.272	0.728	7.11
17	76.02	0.271	0.729	7.06
3	88.35	0.248	0.752	6.27
19	24.48	0.245	0.755	6.15
13	24.94	0.242	0.758	6.08

On BTA chromosome 26 there is a SNP that would be declared significant at both a genome-wide and chromosome-wide significance level (0.05). This location has been previously reported as containing a QTL affecting fat percentage (Druet et al., 2006, Gautier et al., 2006). This SNP closely links to the physical location of SCD1 (stearoyl-CoA desaturase (delta-9-desaturase)) as reported in Genbank (while not being placed on the current assembly) (Benson et al., 2007). SCD1 has been associated with carcass fatty acid composition in Japanese Black cattle (Taniguchi et al., 2004), with milk production traits in Italian Holstein (Macciotta et al., 2008) and with milk-fat composition in Dutch Holstein Friesians (Stoop et al., 2009).

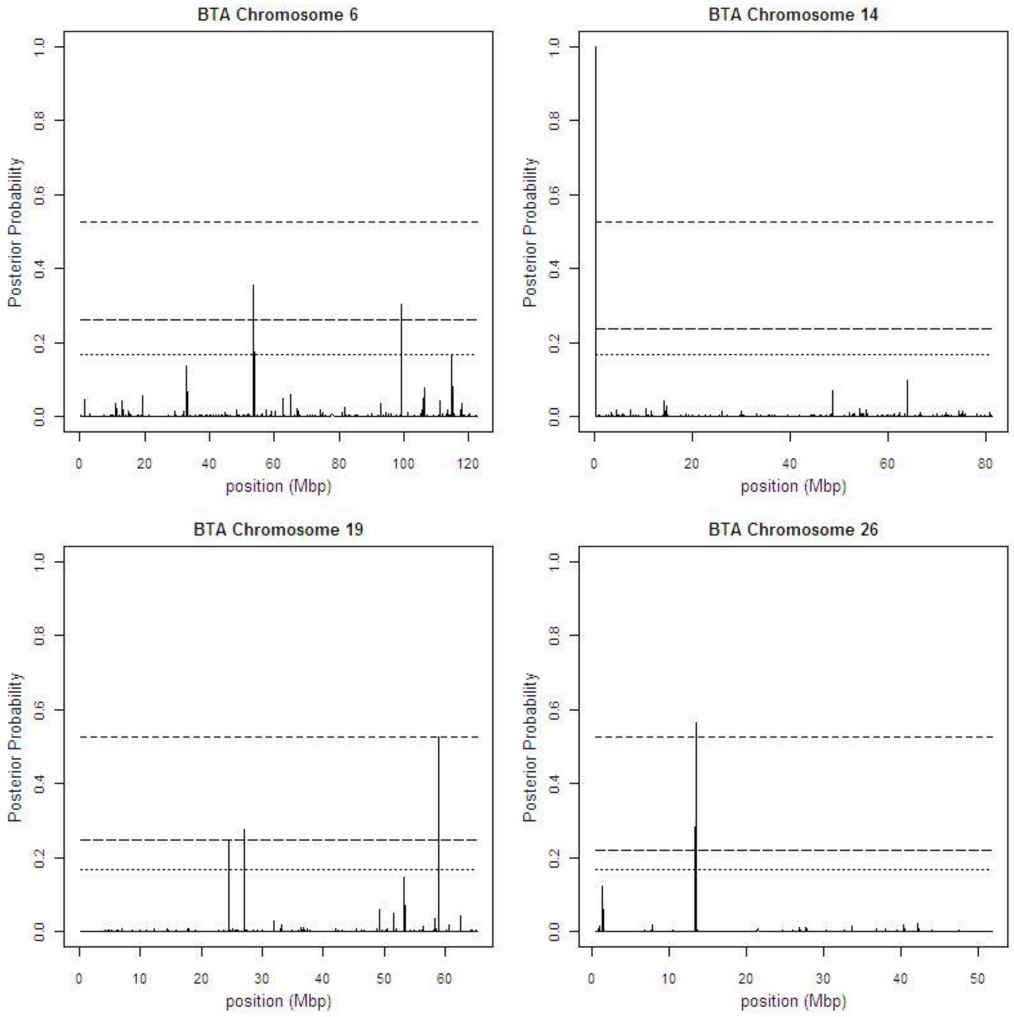

Figure 7.3 - QTL analysis of real dairy data for the trait, fat percentage. The genome-wide threshold (---), chromosome-wide thresholds (···) are plotted for a significance level of 0.05 and posterior probability for a Bayes Factor of 3.2 (-----).

On BTA 6 and BTA 19, no SNPs could be declared significant at a genome-wide level. However, 2 SNPs on each chromosome would be declared significant at the chromosomal level. On BTA 19, one SNP appears to be in association with FADS6 (Fatty acid desaturase domain family 6) while the other SNP indicates a region that has been previously suggested as associated with fat percentage (Khatkar et al., 2004). Multiple locations on BTA 6 have been previously proposed as related to fat percentage. Stoop et al (2009) report a significant association at 53 and 57cM which is consistent with the findings here. In total, across all chromosomes, there were 17 SNPs indicated as significant at the chromosome level.

7.3. DISCUSSION

Permutation tests provide an alternative robust mechanism to identify statistically significant QTL. The robustness and ease of implementation has been confirmed by the wide spread use of the approach in single locus models. The power of the tests is at least as high as unbiased parametric approaches which are highly dependent on model assumptions (Churchill and Doerge, 1994). The examples presented here show that permutation testing can be used to establish significance using multi-locus models in QTL mapping. The results are most closely matched with those that were produced using a Bayes Factor threshold of 3.2 to compare hypotheses.

Highlighted in the analysis of the first simulated data set there is a large random association in the original analysis that is incorrectly declared significant using a Bayes Factor threshold. However, using permutation testing to establish significance, this random association was correctly identified as insignificant. This suggests that permutation testing produces more robust thresholds. While the Bayes Factor threshold does not depend on the data being analysed (as demonstrated in Section 7.2.4) but only the prior probability and the Bayes Factor value; which generally set as reflecting substantial evidence (eg. it is always 3.2 or greater), thresholds established by permutation testing change reflect the empirical distribution of the data. The posterior expected false discovery rate was unviable as a method to establish significance in the examples presented here because all examples produced at maximum one SNP with a posterior probability of one or close to one. The use of *PeFDR* will only work well when there are a large number of QTL that produce posterior probabilities close to one. Of more interest here were approaches that allowed significance to be established when the posterior probabilities are low and inconclusive. Both permutation testing and Bayes Factor provide more viable and successful approaches.

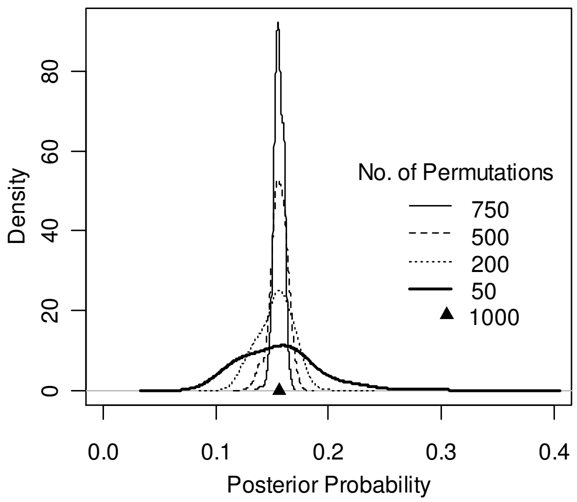

Figure 7.4- Distributions of significance thresholds at 0.05 significance level produced using 50, 200, 500 and 750 permutations

The effect of the number of permutations performed was also examined after the completion of the analyses. The results of the 1000 permuted data analyses were used with 50, 200, 500 and 750 (N) permuted data sets results (test statistics) randomly sampled. Then, a 0.05 threshold was calculated from each set; the maximum test statistic were taken from the N (50, 200, 500 or 750) permuted analyses and ordered and the N x 0.95 value was taken as the threshold. This was repeated 10,000 times for each N. Using the permutations from the second simulation study with stratification as an example, the distributions of the thresholds for each number of permutations are presented in Figure 7.3. The trend seen here is echoed in the analyses of the other data sets. What is immediately obvious is that the 50 permutation tests used in Bauer et al. (2009) are insufficient. Using only 50 permutation tests yielded a range of genome-wide thresholds for a significance level of 0.05 from a posterior probability 0.360 to 1 for the real data. This clearly shows it is possible with such a small number to get significance thresholds that are extreme and will cause incorrect interpretation of the results. As the number of permutations increase, the range of possible thresholds decreases. This result agrees with the work of Doerge and Churchill (1996) that state that 1000 permutations is the lowest number needed for a significance level of 0.05 which is generally the lowest level of significance required.

In a livestock setting, the inclusion of the polygenic effect is important to avoid declaring significance that is caused by the data structure rather than a real QTL. The inclusion of the polygenic effect means that the traditional approach of permuting just the phenotypes is infeasible, as it would also break the phenotype-pedigree relationship. Presented here is an alternative approach of essentially permuting of the entire genotypes or stratifying within sire families. The first approach allows the inclusion of the polygenic effects by breaking the relationship between phenotype and genotype as desired, but conserving the phenotype- pedigree relationship permitting the estimation of the polygenic effect; this approach violates the condition of exchangeability. In comparison, stratifying within sire families satisfies, for the most part, the condition of exchangeability. This approach of stratifying within sire families has been widely used in single QTL mapping e.g. Seaton et al. (2002). Consequently, the approach applied here could be used for data that has been designed for linkage studies. A comparison of these two approaches produced very similar thresholds and whilst the violation of exchangeability does occur, the extent of the violation appears to be minor. Consequently, stratifying within sire groups is recommended if the data supports it as this will satisfy the condition of exchangeability and adds no additional demands to the approach.

Permutation within strata was shown to be a valuable tool in establishing significance for moderate or minor QTL in the presence of a major QTL. A conditioning marker was used but any SNPs also in LD with DGAT1 or the conditioning marker were also by default conditioned upon. This increased the power to identify moderate QTL also affecting fat percentage. Due to the large amount of variance explained by DGAT1, should stratification not been used, this variance would have shifted to other SNP forming random associations with the phenotypes resulting in inflated thresholds that would have been unlikely to establish significance for any QTL other than DGAT1. Conditioning upon more than one marker i.e. there is more than one major QTL, could be problematic if the number of individuals in the strata become small. However, provided that data set is large enough and the strata are substantial, this is a further possibility.

Doerge and Churchill (1996) recommend excluding the chromosome on which the conditioning marker(s) are located, as any markers linked to the conditioning markers(s) if included will continue to show associations and inflate the thresholds. Despite this, the study excluded the first 10 Mbp when establishing the chromosome and genome-wide thresholds. This was because the chromosome-wide threshold for

BTA chromosome 14 (0.239) showed no inflation and was lower than the average chromosome-wide threshold (0.247). The genome-wide threshold was found to be lower when the entire BTA chromosome 14 was excluded (0.513 compared to 0.528). However, the genome-wide threshold was also equally low or lower when it was re-calculated after other individual chromosomes were excluded. This exclusion of other complete chromosomes meant that 14 times (out of a possible 30 times for the removal of each other chromosome) the re-calculated genome-wide threshold was lower than or as low as the threshold calculated when BTA chromosome 14 was omitted. Thus the inclusions of other chromosomes equally increase the threshold, indicating that the exclusion of only the first 10 Mbp did not inflate the final genome-wide threshold. Exclusion of the first 5 instead of 10 Mbp on BTA chromosome 14 produced identical results. However, excluding less than 5 Mbp caused some inflation of the both the genome-wide and chromosome-wide thresholds for this data set.

Consequently, it appears unnecessary to remove the entire chromosome on which the conditioning marker(s) are contained and appears sufficient to exclude the set of markers surrounding the conditioning markers. The distance from the conditioning marker is dependent on the data set and can be investigated for individual data sets; in this instance, 5 Mbp on each side of the conditioning markers seemed reasonable to avoid inflating the thresholds. There were no significant QTL on BTA chromosome 14 at either the chromosome-wide or genome-wide threshold (Figure 7.3)

Further research is needed to determine the upper and lower bounds for the size of strata that can be used effectively when stratifying on the basis of genotype classes and sire groups. The stratification within sire families produced similar results to when stratification was not performed and the data was permuted across all individuals. Interestingly, permuting within sire produced a significance threshold that was greater than the threshold produced with permutation testing without stratification. This was unexpected as permuting without stratifying removes the correlation between pedigree and genotypes. This latter relationship would generally make it harder to estimate the SNP effects, so eliminating it should cause the SNP effects to be more easily estimated, thus inflating the SNP effects and overestimating precision of the analysis. However the opposite appears true for this data set where permuting within sire families produces larger random associations, thus higher SNP effects resulting in a slightly higher threshold.

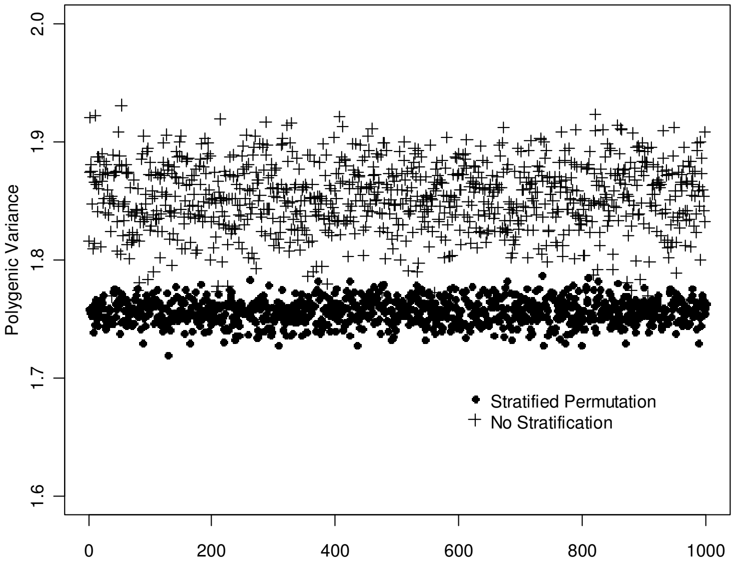

Figure 7.5- Polygenic variance for the 1000 analyses of the permuted data with and without stratification.

One explanation of this finding is that the sire effects are being attributed to the SNP effects instead of the polygenic effect. On inspection of the polygenic variance, the polygenic variance is larger when the no stratification has taken place. This indicates more variance is attributed to polygenic effect when no stratification has taken place and thus less variance is explained by the SNPs. Subsequently the SNP effects will be lower which will make it more probable for a SNP to not be included in the significant effects. This will reduce the SNP's posterior probability and thus the overall threshold. The polygenic variances from the analyses of the 1000 permuted data sets with and without stratification are shown in Figure 7.5. This observed effect may be a reflection of sire family sizes where the family groups with 25 individuals may have been small. Small sire family sizes may result in less allelic variation within families and may cause SNP effects to remain similar due to unchanged genotypes at specific loci with family groups despite permutation. More testing is required to establish a consensus on what is the effect of stratifying within family structures and what the minimum size of these groups needs to be. Obviously more relationships are present in the data than the groups stratified across, so the effect of stratification on these other relationships and consequently, the results, also requires exploration.

The advantages and disadvantages of the different thresholds have previously been discussed (Churchill and Doerge, 1994). In a multi-locus setting, the SNP specific threshold is shown to be unviable. This is due to the major problem caused by comparing the original test statistic with a distribution created from the permutation tests that due to dense linked SNPs could be incomplete or biased. Consider a (random or real) association between some linked SNPs and the phenotypes. In single locus models, the effect will be attributed to each of the linked SNPs as they are fitted in separate models; permutation testing using the single locus model therefore yields the correct distribution of the test statistic for each SNP. However in multi-locus models, the effect may be spread across all the linked SNPs or the effect may be attributed to just one of the linked SNP. Therefore, in multi-locus models the test statistic distribution of a single SNP may not include all the values that accurately indicate the true number of random associations that occurred during the permutation testing. Consequently, these distributions do not reflect the null distributions that are to be sampled from, which corresponds to the hypothesis that there is no association between the phenotypes and the SNPs. This distortion would result in an increase in the incorrect declarations of a significant SNP linked to a QTL. This increase is evident in both examples by the excessive number of SNPs established as significantly linked to QTL using this threshold. An alternative approach may be to calculate the joint posterior probabilities for overlapping intervals for SNP found in each interval (Sahara et al., 2010). This would allow region (or interval) specific threshold to be calculated.

The other two thresholds can be used as originally proposed as they use the maximum test statistic either for each chromosome or the entire genome. The genome-wide threshold has the least power to declare significant QTL affecting the trait but allows the control of the type I error rate. The chromosome-wide threshold has increased power to identify significant QTL but due to testing each chromosome separately the type I error rate could be higher than the desired value. This threshold therefore provides a balance between the control of the type I error rate and the power to identify QTL.

A possible caveat to the multi-locus method proposed, is that performing the permutations can be time consuming. The time required is dependent on the length of the time that the model and implemented approach requires to produce results. In the case of the study reported here, due to the implementation of the Bayesian latent variable model, the time demands were high. The model was run for 10,000 and 20,000 iterations for the simulated and real data respectively, to ensure convergence. Consequently, the real data example with 43011 SNPs took twelve processor hours to complete. However to run a single SNP model including a polygenic effect using ASReml (Gilmour et al., 2006b) for the 43011 SNPs took three times the amount of time needed for the multi-locus model (without the polygenic effect the time requirements were still at least twice as long). Consequently, the approach presented has lower time demands than if permutation testing was used for all 43011 SNPs using a single locus model.

Should more extreme p-values be desired, one way to reduce the computational time would be to approximate the tail of the distribution (Knijnenburg et al., 2009). This method does provide a way to test for more extreme p-values while requiring fewer permutations. However, as many or more permutations than tested here would be needed to accurately re-construct the tail of the test statistic distributions.

7.4. CONCLUSION

In summary, multi-locus models offer advantages over the traditional single locus models as they overcome the problem of multiple testing and estimation of the total variance explained by the QTL, as well as leading to more precise mapping of QTL. However the problem of establishing significance has been difficult for multi-locus models. In this chapter, a permutation approach is presented to enable the declaration of significant QTL for multi-locus models. The approach was compared to other methods to establish significance. Bayes Factors and permutation testing produced reasonable thresholds while the posterior expected FDR was shown to be unviable for use with similar data sets. The proposed approach was demonstrated to identify known QTL in both simulated and real data, and therefore provides a valuable technique to establish significance for minor and moderate QTL with or without the presence of a major QTL when using a multi-locus model. In addition two approaches to dealing with a linked third variable are presented. While promising, the results were inconclusive and more research is required.

While genomic selection and prediction do not seek to establish the QTL underlying the trait of interest, many of the models applied to predict an animal's total genetic value can be used for QTL analysis with the exception of any BLUP approaches. BLUP approaches assume equal variance across SNP and thus tend to share the effect of QTL across a range of SNP which results in very low SNP effects across most SNP. This type of approach makes it difficult to determine the exact position and effect of any QTL.

While genomic selection is currently producing results that lead to increased levels of accuracy of selection, the ability to determine correctly the biological factors and QTL underlying the trait of interest would further increase the ability to construct a robust and accurate prediction equation. Thus as the number of genotyped animals (with reliable phenotypes) increase, the ability to identify significant QTL should increase. The problem of how to identify the many minor QTL explaining only small amounts of genetic variation can be partially accounted for by the setting of thresholds and increasing the power of QTL studies.

CHAPTER 8
Predicting energy balance for dairy cows using high density SNP information

8.1 INTRODUCTION

Many countries have introduced measures of fertility into national selection indexes to try to address declining fertility rates in dairy cattle (Miglior et al., 2005, Royal et al., 2000). One explanation for these decreases is the difference between energy intake and energy usage that occurs during early lactation. This difference is defined as energy balance (EB). Energy balance provides an essential link between production and non-production traits because both depend on a common source of energy. An animal's energy must be partitioned efficiently to maintain production levels as well as an animal's ability to remain healthy and fertile. Severe negative energy balance (NEB) during early lactation has been cited as an underlying cause of the negative relationship of health and fertility with production (Butler and Smith, 1989, Jorritsma et al., 2003, Pryce et al., 2004).

Recently, there has been a major focus on trying to overcome the NEB problem by modifying the diet during the dry period (Agenas et al., 2003, Dewhurst et al., 2000, Garnsworthy et al., 2008a, b, McNamara et al., 2003). Other suggested approaches to tackling NEB include varying the length of the dry period (Watters et al., 2009) and frequency of milking (McNamara et al., 2008). However, estimates of genetic parameters suggest that NEB is not only a consequence of a poor match between nutrition and production, but there is also genetic variation (Coffey et al., 2004, Friggens et al., 2007, Veerkamp, 1998, Veerkamp et al., 2003). Veerkamp (1998) reviewed the results of different studies that reported genetic correlations for a variety of energy measures and milk yield, with values ranging from -0.05 to -0.91 and heritability for energy traits that ranged from 0.19 to 0.69. Coffey et al. (2004) demonstrated that distinct genetic lines responded differently to a range of diets and differed in the time taken to return to positive EB. Similarly, Friggens et al. (2007) concluded that variability among animals on a stable nutritional diet could not be accounted for by environmental factors and indicated a genetic basis for EB. Thus, an alternative to management approaches may be to select animals that are genetically predisposed to maintain a better EB.

Accounting for EB in selection programs has been complicated, since measuring feed intake in progeny testing schemes is generally not practical. However, the current d evelopment of genomic prediction and selection methods as discussed in Chapter 3, 4 and 5, coupled with the increase in both selection accuracy and genetic gain that genomic selection provides over traditional selection methods (Hayes et al., 2009c) may provide one option to allow for the selection of EB.

The aim of the study presented in this chapter was to examine whether genomic prediction could be used to estimate DGV for EB using a small Dutch experimental farm data set. The objective was to demonstrate the genetic basis of EB and the potential use of genomic selection to facilitate inclusion of EB in selection programs. The study and its results have been published in the Journal of Dairy Science (Verbyla et al., 2010b) (see Appendix A3 for published version).

8.2. MATERIALS AND METHODS

8.2.1. Data

Data on 613 Holstein-Friesian heifers born between 1990 and 1997 was collected during the first 15 weeks of lactation including 450 cows participating in the breeding program of CRV (Arnhem, The Netherlands) and 163 cows originating from an experimental farm ('t Gen, the Netherlands). All animals were housed together on a single farm under the same environmental and management influences. All cows were fed ad libitum. Live weight, feed intake, and milk yield were measured on 565 of the animals. Milk samples were taken on a fixed day of the week for measurement of fat, protein, and lactose yields. Feed intake was recorded daily using automated feed intake units. Live weight was recorded once a week. Energy balance (MJ/d) was calculated, using the method described in Veerkamp et al. (2000), as the difference between energy intake and the calculated energy requirements for milk, fat and protein yields, and maintenance costs as a function of live weight. Energy balance values across weeks 2 to 15 were averaged, where possible, to give an overall EB phenotype. More comprehensive details on the data used can be found in Veerkamp et al. (2000). Raw EB phenotypes were pre-adjusted for year-season of calving and age at calving (linear, quadratic) using ASReml (Gilmour et al., 2006b), since their inclusion was not feasible in the final model due to software limitations. The residuals from this analysis were used as the EB phenotypes for the prediction of the breeding values.

588 of the 613 heifers had known pedigree and these were genotyped using the Illumina 50K SNP panel (54,001 SNP in total). The quality control criteria for selecting the final set of SNP were a call rate of over 90%, a GenCall score >0.2 and a GenTrain score >0.55 (Illunima descriptive statistics on genotype quality), a minor allele frequency of >2.5% and a lack of deviation from Hardy Weinberg equilibrium, χ^2<600 (Wiggans et al., 2009). Animals with greater than 5% missing SNP genotypes were removed. Non-Mendelian error checks identified genotypes of daughters that were inconsistent with their dams. A further, more comprehensive pedigree check was performed by comparing the coefficients of the additive genetic relationship matrix and the genomic relationship matrix (G matrix) calculated via the first method described in VanRaden (2008). This enabled inconsistencies between recorded half and full siblings to be examined. Animals with many inconsistencies between the pedigree and G matrix were removed. After all editing steps, in total, 43,011 SNP and 548 animals were retained; of these 548 animals, 527 had phenotypes for EB.

8.2.2. Statistical Analysis

8.2.2.1. Models

Two models using Gibbs Sampling were applied to estimate additive breeding values. One model included the available SNP information. This model used stochastic search variable selection (**SSVS**) (George and McCulloch, 1993), as in Chapter 4, which introduces an indicator variable, I_j, that determines whether SNP j has a large significant effect or whether the effect is insignificant and is therefore scaled back towards zero. The indicator variable for each locus j has a Bernoulli prior distribution:

$$I_j \sim \text{bernoulli}(p)$$

The prior probability p is chosen to reflect the information available on how many QTL affect the trait of interest. It can be quantified as the number of SNP expected to be linked to a QTL divided by the total number of SNP. For a complex trait such as EB, it was assumed that about 1% of the SNP were linked to a QTL ($p = 0.01$). The model extends that presented in Chapter 4 based on the model presented in Meuwissen and Goddard (2004) for multi trait QTL analysis. The SNP model can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{1}_n \mu + \sum_{j=1}^{\boldsymbol{m}} \left(\mathbf{X}_j \left(\mathbf{q}_j v_j \right) \right) + \boldsymbol{Zu} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of phenotypes corrected for fixed effects for the trait being analysed for all n individuals, $\boldsymbol{1}_{\mathrm{n}}$ is a vector of ones of length n, μ is the mean, m is the number of SNP markers, $\mathbf{X}_j$ is the (n x k) design matrix containing the information on the possible k alleles at the j^{th} marker for all individuals (where x_{jk}=0,1,2 having no, one or two copies of the k^{th} allele respectively), $\mathbf{q}_j$ is the vector ($kx1$) containing the effects of all k possible alleles at locus j where $q_{jk.}$ are drawn from a standard normal distribution N(0,1), v_j is the standard deviation of the allelic effects at locus j and is dependent on whether the locus effect is considered significant or not using the indicator variable e.g. v_j is sampled: $\mathrm{p}(v_j \mid I_j) \sim \boldsymbol{N}(0, I_j + (1 - I_j)/100)$, $\mathbf{u}$ is the vector of random additive polygenic effects of length n ($\boldsymbol{Z}$ is the associated design matrix) and is assumed to be normally distributed, $\boldsymbol{u} \sim \boldsymbol{N}\left(0, \sigma_u^2 \boldsymbol{A}\right)$ where $\boldsymbol{A}$ is the pedigree-derived additive genetic relationship matrix and $\mathbf{e}$ is the residual error also assumed to be normally distributed, $\boldsymbol{e} \sim \boldsymbol{N}\left(0, \boldsymbol{I}\sigma_e^2\right)$ where $\boldsymbol{I}$ is the n x n identity matrix. The allele substitution effect of a locus j can be calculated from the estimated effects as: $a_j = \left(q_{j1} - q_{j2}\right) v_j$ where q_{j1} (q_{j2}) is the effect of allele 1(2) at locus j. For the full specification of the priors used and an alternative formulation of the model see Calus et al. (2008) and Meuwissen and Goddard (2004). The DGV were calculated as the sum of estimated SNP effects and the polygenic effect ($\mathrm{DGV} = \sum_{j=1}^{p} \left(\mathbf{X}_{ij} \left(\hat{q}_j v_j \right) \right) + \hat{u}_i$).

The second model used was a simple additive polygenic model using pedigree-based relationships, as follows:

$$\boldsymbol{y} = \boldsymbol{1}_n \mu + \boldsymbol{Z}\,\boldsymbol{u} + \boldsymbol{e}$$

where the EBV calculated by this model were the estimated polygenic effect for each animal ($\mathrm{EBV} = \hat{u}_i$). Both models were run for 10,000 iterations to ensure convergence with the first 1000 iterations used as burn in.

8.2.2.2. Validation

Due to the small size of the data set, a 10-fold cross validation approach was carried out to assess the accuracy of predicted breeding values. The data set was randomly partitioned into 10 subsets each containing 10% of the data. Each subset was retained once as the validation data set and the remaining 9 became the reference sets. Results from the reference sets were then used to predict breeding values of animals in the validation set. Accordingly, each animal appeared only once in a validation set and had only one predicted DGV.

The DGV and EBV were assessed using accuracy, $\mathrm{r}_{\mathrm{y}\hat{\mathrm{g}}}$, defined as the Pearson correlation of the predicted breeding values (DGV or EBV) ($\hat{\mathrm{g}}$) and the phenotypes (y). The maximum achievable accuracy, $\mathrm{r}_{\mathrm{y}\hat{\mathrm{g}}}$, due to the correlation between phenotypes and predicted breeding values, was equal to the square root of the heritability of the phenotypes. The observed heritability for EB was estimated by fitting a model with year-season and age at calving (linear and quadratic regression) as the fixed effects and a random animal effect (a). The random animal effect was assumed normally distributed, $a \sim N\left(0, \sigma_a^2 \boldsymbol{G}\right)$ where σ_{a}^2 is the additive genetic variance and $\boldsymbol{G}$ was the genomic relationship matrix calculated via the first method described in VanRaden (2008). Deriving the heritability this way has been shown to produce estimates closer to the true value than using the pedigree based relationship matrix (Hayes and Goddard, 2008a).

As no daughter yield deviations (DYD) or reliable breeding values were available, the predicted breeding values (DGV and EBV) were compared to phenotypes. Most studies estimating accuracies of DGV, use DYD or reliable EBV predicted for proven bulls and consequently report accuracies of selection ($\mathrm{r}_{\mathrm{g}\hat{\mathrm{g}}}$) and reliabilities ($\mathrm{r}_{\mathrm{g}\hat{\mathrm{g}}}^2$), that compare DGV and the closest estimate of the true breeding values (g). Thus, the accuracy of selection was predicted (Daetwyler et al., 2008, Goddard, 2009) as:

$$r_{g\hat{g}} = \sqrt{\frac{\lambda h^2}{\lambda h^2 + 1}} \text{ where } \lambda = \frac{n_p}{n_G} \qquad [10]$$

where h^2 is the observed heritability, n_p is the number of phenotypic records and n_G the number of effective QTL or chromosome segments. This function therefore can be used to estimate the number of effective QTL and the number of records needed to increase the accuracy of selection. The function can be modified for use when the accuracy is calculated using the correlation between the predicted DGV and phenotypes ($\mathrm{r}_{\mathrm{y}\hat{\mathrm{g}}}$). Falconer and Mackay (1996) state that $r_{g\hat{g}} = \sigma_{\hat{g}} / \sigma_g$. The accuracy between DGV and phenotypes can be similarly expressed as $r_{y\hat{g}} = \sigma_{\hat{g}} / \sigma_y$. $\mathrm{r}_{\mathrm{y}\hat{\mathrm{g}}}$ can also be denoted as:

$$r_{y\hat{g}} = \frac{\sigma_{\hat{g}}}{\sigma_g} \times \frac{\sigma_g}{\sigma_y}$$

which can be rewritten as:

$$r_{y\hat{g}} = r_{g\hat{g}} \times \sqrt{h^2} \qquad [11]$$

when combined with [10], this gives:

$$r_{y\hat{g}} = \sqrt{\frac{\lambda h^4}{\lambda h^2 + 1}}$$

Hence, $\mathrm{r}_{\mathrm{y}\hat{\mathrm{g}}}$ can also be transformed into $\mathrm{r}_{\mathrm{g}\hat{\mathrm{g}}}$. The accuracy, $\mathrm{r}_{\mathrm{y}\hat{\mathrm{g}}}$, was subsequently used to calculate the number of QTL affecting EB and the number of records needed to improve the accuracy of the predicted DGV.

8.3. RESULTS

The pedigree check step for data quality control proved a very effectual additional measure to identify any animal that had an incorrectly recorded pedigree or where an animal may have been misidentified. It allowed checking of half-sibling and full-sibling relationships that is not possible using standard non-Mendelian checking which compares parent with offspring to establish any conflicting homozygotes. Figure 8.1 effectively illustrates the additional information contained in the SNP data about the relatedness of the animals. This is most obviously shown by the monozygotic twins that have a marker relationship of 1 (due to identical DNA) but are recorded as full siblings in the pedigree. The negative marker relationships are due to the method used to calculate the G matrix which ideally uses the allele frequencies that were present in the base population (VanRaden, 2008). However as the frequencies in the base population were unknown, the G matrix was calculated using the allele frequencies in the available highly selected population resulting in negative marker relationships.

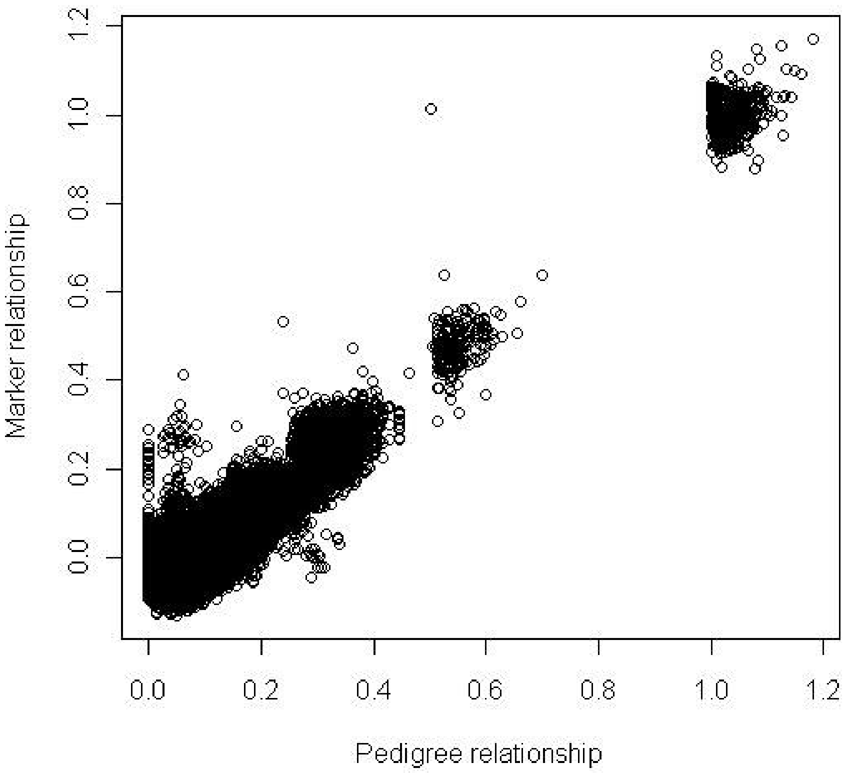

Figure 8.1- Comparison of the coefficients of the additive relationship matrix (pedigree relationship) and the coefficients of the genomic relationship matrix (markers relationship)

The accuracies ($r_{y\hat{g}}$) of predicting phenotypes for the two models and the $r^2_{y\hat{g}}$ are shown in Table 8.1. Transformed values using (1) giving the accuracies of selection ($r_{g\hat{g}}$) and the reliabilities ($r^2_{g\hat{g}}$) are also shown. The heritability for EB was estimated separately, as described earlier, with a moderate value of 0.325 (SE = 0.12). The accuracy of selection, $r_{g\hat{g}}$, was then calculated using [11]. The model including the SNP information yielded an overall accuracy of 0.29, which was higher than the overall accuracy of 0.21 produced by the polygenic model.

Table 8.1- Accuracies and reliabilities of the DGV and EBV

Model	$r_{y\hat{g}}$	$r_{g\hat{g}}$	$r^2_{y\hat{g}}$	$r^2_{g\hat{g}}$
Direct Genomic Value (DGV)	0.294 (±0.038)	0.516	0.086 (±0.025)	0.265
Estimated Breeding Values (EBV)	0.211 (±0.047)	0.370	0.044 (±0.023)	0.135

DGV is predicted using the model including both the SNP and polygenic effects and the EBV using the model including only the polygenic effect. $r_{y\hat{g}}$ is the Pearson correlation between the predicted breeding values ($\hat{g}$) and the phenotypes (y) (± standard error derived from the 10 data sets). $r_{g\hat{g}}$ is the accuracy of selection (comparing the predicted breeding values ($\hat{g}$) and the true breeding values (g)). $r^2_{y\hat{g}}$ is the reliability of the predicted phenotypes (± standard error), $r^2_{g\hat{g}}$ is the reliability of the predicted breeding values.

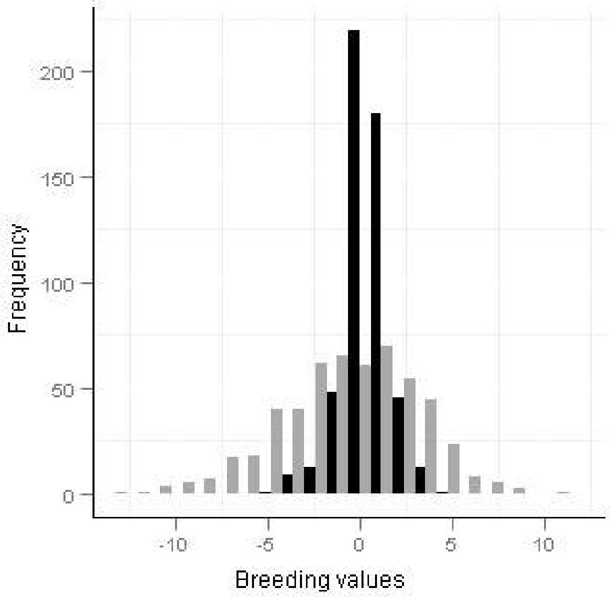

Figure 8.2- Histogram of DGV and EBV, (■) represents the estimated breeding values (EBV) predicted by the polygenic model and (▨) represents the direct genomic values (DGV) predicted by the model including the SNP information

The calculated reliability ($r^2_{g\hat{g}}$) of DGV is double that of the EBV produced by the polygenic model. This implies that the DGV explained twice as much variation as the EBV which is illustrated also by the range of the breeding values (Figure 8.2). The predicted DGV and EBV were positively correlated with a value of 0.70.

A total of 472 effective QTL for EB were predicted. A sensitivity analysis was conducted to investigate the effect of the number of effective QTL, heritability and number of records had on the expected accuracy of selection. Figure 8.3 is a plot of $r^2_{y\hat{g}}$ ($r^2_{g\hat{g}}$ is provided for comparison on the second y axis) against the number of effective QTL for differing heritability where the number of records was kept constant at the available number of 527. This shows the impact that the number of effective QTL would have on the expected accuracies and reliabilities of the DGV. It demonstrates that the greater the number of QTL affecting the trait, the lower the expected accuracy and reliability. This is due to lack of information available in the limited number of phenotypes to be able to accurately estimate large numbers of QTL effects. Figure 8.3 also illustrates that this reduction in reliability, as the number of effective QTL increases, is more gradual for higher heritabilities.

The number of total records needed to improve the accuracy was also investigated and results are shown in Figure 8.4. The heritability was set at the observed value for EB (0.325). It is evident from Figure 8.4 that the number of effective QTL has a significant impact on the number of records needed to improve the accuracy. The greater the number of effective QTL, the larger the number of phenotypic records required to reach higher accuracies and reliabilities. A total of 5,818 records with phenotype and genotype information was predicted as needed for a $r^2_{y\hat{g}}$ of 0.24 ($r^2_{g\hat{g}}$ of 0.80) for EB with the predicted 472 effective QTL.

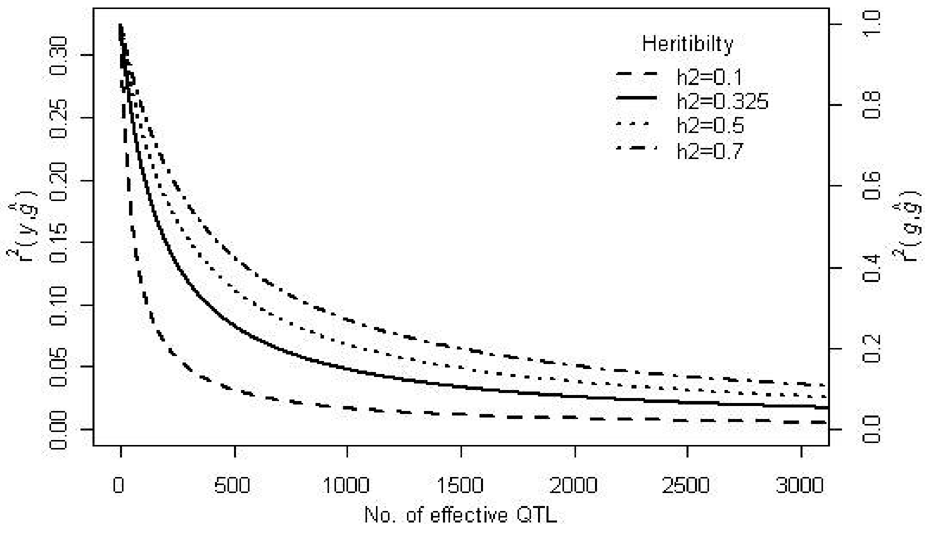

Figure 8.3- Accuracy of prediction versus the number of effective QTL where the number of records is fixed to the number used in this study (527). $r^2(y,\widehat{g})$ is the squared correlation between the phenotypes and the predicted direct genomic values, DGV (characterized in the text as $\mathrm{r}^2_{\mathrm{y}\hat{\mathrm{g}}}$). $r^2(g,\widehat{g})$ is the estimated reliability between the true breeding value and the predicted DGV (characterized in the text as $r^2_{g\hat{g}}$).

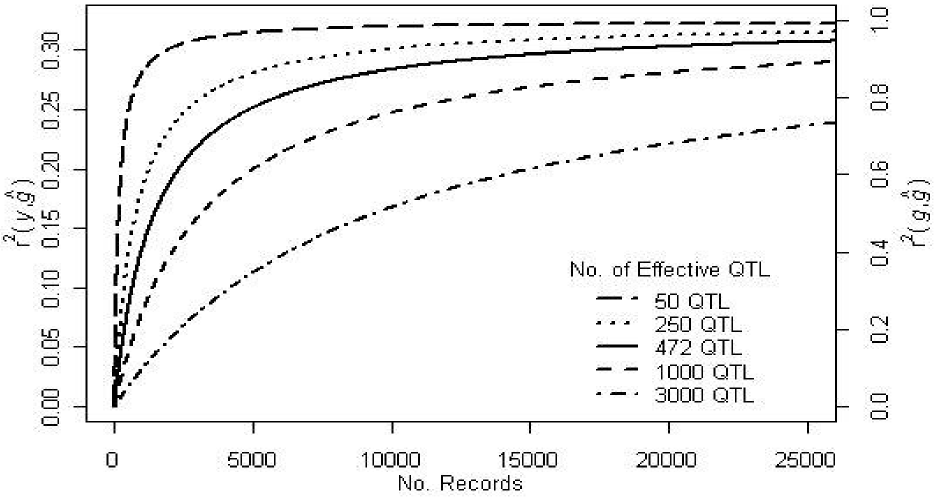

Figure 8.4- Accuracy of prediction versus the number of records for a fixed heritability of 0.325. $r^2(y,\widehat{g})$ is the squared correlation between the phenotypes and the predicted direct genomic values, DGV (characterized in the text as $\mathrm{r}^2_{\mathrm{y}\hat{\mathrm{g}}}$). $r^2(g,\widehat{g})$ is the estimated reliability between the true breeding value and the predicted DGV (characterized in the text as $r^2_{g\hat{g}}$).

8.4. DISCUSSION

The objective of this study was to demonstrate the genetic basis of energy balance and that it could potentially be incorporated into selection programs using genomic selection based on a limited reference population. EB is a minimally recorded trait and consequently only a small number of phenotypic records were available. Despite the limitation on available data, genomic prediction was able to produce accuracies greater than a traditional polygenic model. Thus the results indicate that EB can be estimated using genomic prediction. The low accuracy gained can be explained as a direct result of the small number of phenotypic records and the moderate heritability found for this trait. The heritability calculated with this data set was consistent with results of other studies (Huttmann et al., 2009, Veerkamp, 1998). In order to consider including EB in breeding schemes, higher accuracies than found here are necessary. This increase could be facilitated through an increase in the heritability of the trait or an increase in the number of phenotypic records. One way to increase the heritability would be to standardize the environmental conditions to reduce non-genetic differences between animals, but this may be difficult to do in practice. An alternative approach to increase the heritability of phenotypes would be to use deregressed breeding values or DYD of proven bulls as phenotypes, based on EB records of many daughters. This allows for an increase in the accuracy, while keeping the number of genotyped animals constant. This scenario will not lead to any additional genotyping costs since most bulls may already be genotyped as part of reference populations for other breeding goal traits. Nevertheless this may still be more costly due to the (much) higher number of recorded EB phenotypes that would be needed. An increase in the number of available records would also allow for an increase in the accuracy of predicted DGV as previously indicated in other studies (Goddard, 2009, Hayes et al., 2009c). The required increase could only occur if the measurement and recording of EB improved.

Due to the low recording of EB, a seemingly obvious solution would be to immediately select for a more widely recorded trait, like body condition score (BCS), to indirectly try to reduce NEB. The problem with using BCS is that after the first 60 days in milk (DIM), the genetic correlations between EB and BCS decrease markedly (Huttmann et al., 2009). However, until the recording of EB increases to useful levels, BCS does provide a viable option to attempt to select animals with a better EB. In the future having both EB and BCS phenotypes available will probably allow for the best prediction of energy partitioning and utilisation.

The model used to predict the DGV could also be used for whole genome association studies. Thus, the produced posterior probabilities of SNP were examined to see if there were any significant associations with EB. Due to the small number of records and large number of SNP, the power of the association study to identify QTL was very low and this was evident. There was no SNP with a high enough posterior probability to be confident that it was linked to a QTL. The prior for the expected number of QTL affecting EB was varied but results were consistently low. Although the posterior probabilities were low, there was one SNP that had 10-fold higher posterior probabilities than all the other SNP in all analyses regardless of the prior probability used. However, using a Bayes Factor (Section 7.2.4) (Kass and Raftery, 1995) with a value of 3.2 (the lowest value of the Bayes Factor that indicates substantial evidence) indicated that the result could be declared to be significant and that this SNP linked to a QTL.

The SNP is located on chromosome 21 and is in extremely close proximity to, and appears in association with, the nuclear receptor subfamily 2, group F, member 2 (NR2F2), otherwise known as chicken ovalbumin upstream promoter transcription factor II (COUP-TFII). COUP-TFII has been previously reported as playing an essential role in regulating adipogenesis, glucose homeostasis and energy metabolism (Li et al., 2009, Xu et al., 2008). It has also been reported as regulating growth hormone receptor 1A promoter activity (Xu et al., 2004) and mediating progesterone and controlling estrogen levels and thus involved in reproduction (Klinge et al., 1997, Kurihara et al., 2007, Nakshatri et al., 2000, Petit et al., 2007, Takamoto et al., 2005). Whilst the results of this association study are not conclusive and further validation is required, COUP-TFII appears to be a good candidate gene for EB.

Despite being unable to conclusively establish QTL associated with EB, results of the study allowed the estimation of the number of effective QTL influencing EB. Given the nature and complexity of EB, the number of predicted effective QTL (472) was plausible. However this prediction may be dependent just on the number of records, effective population size and the length of the genome (Goddard, 2009). The relationships with both production and non-production traits means that potentially numerous genes and pathways could be involved in the variation observed in EB. Previous whole genome association studies of residual feed intake and other traits related to EB in beef cattle, report between 4 and 120 QTL affecting the traits studied (Barendse et al., 2007, Sherman et al., 2009). These values are significantly lower than the predicted 472, but reflect the power of the studies to detect significant QTL and the number of SNP (which were 2194 and 8786 respectively), rather than the true number of effective QTL. An increase in the number of phenotypic records would also allow genome wide association studies for EB in dairy cattle to identify possible candidate genes affecting the trait and would provide a better idea of the effective number of QTL.

The ability to select and include EB in selection indexes may indirectly increase the genetic gain for fertility traits. The interval between calving and start of luteal activity (C-LA) has been demonstrated to be an indicator of fertility during later lactation (Darwash et al., 1999, Petersson et al., 2007, van der Lende et al., 2004). Veerkamp et al. (2000) reported genetic correlations between EB and C-LA of -0.60 (and -0.49 for C-LA adjusted for milk, fat and protein). A moderate to high genetic correlation similar to what was previously reported would mean that genetic gain for EB should also result in improved fertility. For example, if a bull had 25 daughters, the accuracy of selection for the bull's EBV would be 0.40 for fertility (assuming a heritability of 0.03), whereas for EB, the accuracy of selection for the bull's EBV would be 0.83. Thus, given a genetic correlation of -0.5, the accuracy of selection for fertility using EB would be 0.41. Consequently for bulls with this number of daughters or less, selection using EB will result in greater genetic gain for fertility than selecting for fertility itself. However, as number of offspring per bull increases beyond 25, the benefit of using EB rather than fertility is lost, such that, selection for fertility itself will produce better genetic gains. Thus, the use of EB in selection indexes, in addition to fertility, may prove beneficial and result in increased genetic gain for fertility.

In addition, to the possible benefits of improved fertility, EB could be used with feed intake data to select animals for feed efficiency (Veerkamp, 1998) or reduce methane emission (Hegarty et al., 2007). Improving feed efficiency could be economically desirable as feed costs contribute the greatest proportion to production costs (Simm et al., 1994). However, feed efficiency data alone cannot distinguish whether the energy is used for production or maintenance. This may result in selection of animals that have low intake and high yield but consequently have problems related to high NEB. Thus, both NEB and improved feed efficiency (or intake) data should be considered simultaneously in order to effectively reduce the feeding costs while not having detrimental effects on the animals' health and fertility.

There are many other traits including several fertility and reproduction traits such as milk progesterone profiles and milk quality trait which are difficult to record. Accounting for these traits, like EB, in selection has been complicated, since measuring them in progeny testing schemes is not practical. The study reported here demonstrates that it is possible for such traits, with similar heritabilities and expected number of QTL to produce DGV with accuracies above 0.8 when there are more than approximately 2,600 (2,581 predicted for EB) phenotypic records available for use as the reference population. This means that it is possible to select for these traits using genomic selection by combining data from experimental and nucleus herds, where individually there are a limited numbers of raw phenotypic records.

Genomic prediction is often performed using a two step procedure where the input phenotypes are pre-corrected so that the model predicting the DGV only includes the mean, polygenic and SNP effects. Since pre-correction will always introduce a new source of error, our preference would have been to include all fixed effects in the models used to predict the breeding values. There is on-going development of the genomic prediction program used, to allow the inclusion of multiple discrete and continuous fixed effects in a single model.

8.5. CONCLUSIONS

The use of SNP information to predict DGV is shown to explain variation between the EB of animals, confirming the genetic background of EB. The use of SNP information showed an increase in the accuracy of prediction for EB over the simple polygenic model for animals without an EB record. However, the number of phenotypes would need to be increased to improve the accuracy. In the future, selection for EB could be performed using genomic selection which could provide a valuable tool in finding a balance between production and non-production traits.

The potential of genomic prediction and selection to allow selection for difficult traits that would have been previously impossible or extremely difficult is demonstrated here through the use of energy balance. As discussed in chapter 5, a variety of genomic prediction methods could have been applied with the expectation that each would produce a higher accuracy of prediction than the polygenic model due to the similar accuracy of most genomic prediction approaches on real data. Bayesian approaches would be expected to produce only slightly (~1-3%) higher accuracies than genomic BLUP (Hayes et al., 2009c). The results again highlight that the accuracy of genomic prediction is heavily reliant on the availability of reliable phenotypes due to *p>n* as discussed in Chapter 2 and 5. A further discussion on genomic prediction and selection requirements for increased success and development are contained in the following chapter.

CHAPTER 9
General Discussion

9.1. INTRODUCTION

The aim of the research reported in this thesis was to investigate important aspects of genomic selection to enable a more comprehensive understanding of what makes a robust and accurate Bayesian prediction model. A further aim was to explore new possibilities introduced though genomic selection, for instance, selecting for minimally recorded traits. Results from the preliminary simulation studies in Chapter 3 indicated that the match between the assumed QTL distribution and the true QTL distribution had an effect on the accuracy of the DGV produced by the different models. Conversely in real data (Chapters 4 and 6), a general equality in the accuracy of prediction was found across the various models. The exception was for traits with atypical genetic architectures. Thus the model used to simulate may not well represent the real genetic model.

In the proceeding chapters, this thesis has presented studies and outcomes that contribute new knowledge and implications to the current abundance of research on genomic selection. Novel research undertaken for this thesis is detailed; much of these results have been published.

In this chapter, the major findings are discussed and their implications for the current and future implementation and use of genomic selection are considered.

9.2. BAYES SSVS

The development of the Bayes SSVS model presented in Chapter 4 was in response to the slow computational times produced by Bayes B and the acknowledged higher accuracies produced by models with similar assumptions to Bayes B. Bayes SSVS is a alternate formulation of a Bayesian model using Stochastic Search Variable Selection.

A similar approach using Stochastic Search Variable Selection was developed by Calus and Veerkamp (2007) using the work of Meuwissen and Goddard (2004). The model of Meuwissen and Goddard was originally presented for GWAS and was developed as a genomic prediction model by Calus and Veerkamp (2007). However, the performance of this approach was not initially demonstrated in real data until de Roos et al. (2009). Its equivalence to the original Bayes B and its performance compared with other genomic prediction models has never been presented. However, the robustness and usefulness of this type of model is highlighted by the implementation of this SSVS approach by CRV in the Netherlands for use in producing national GEBV.

The problem with the SSVS approach, as previously noted for Bayes B (Gianola et al., 2009), is setting of the value of the hyper-parameter, p_i (the proportion of SNP in the larger distribution). One solution is to set a prior distribution and also sample this parameter (Fernando, 2009). A second problem has been stated to be the way the hyper-parameters of the prior distributions for the variance of the SNP effects are determined. Gianola et al. (2009) states that the formulation of the hyper-parameters for Bayes A and consequently Bayes B and Bayes SSVS cause the prior to dominate the data as the formulation for the variance means that if r is the degrees of freedom for the inverse scaled chi-squared prior distribution, then the degrees of freedom for the conjugate posterior distribution will always be $r + 1$. This is true with respect to estimating the SNP specific variance but not with respect to estimating the SNP effect. They suggest that clusters of markers are formed such that markers in the same cluster share the same variance, resulting in shrinkage specific to individual clusters. Additionally, they propose that the clusters could be formed on the basis of biological information or a statistical procedure. Further investigation is required by the authors to demonstrate that these changes (not currently available) would actually increase the accuracy of the DGV produced by these models and that the current formulation is biasing or reducing the accuracy through the current approach. In fact, Habier et al (2010a) present alternative methods accounting for the drawbacks outlined by (Gianola et al., 2009) and determined that these statistical issues to not affect the accuracy of Bayes A and Bayes B (and thus Bayes SSVS as presented here). The major difference between the new methods presented by Habier et al (2010a) is that they used a common effect variance instead of a SNP-specific variance to avoid the suggested issue related to the formulation of the hyper-parameters and the subsequent overwhelming of the data.

As mentioned previously, the general equality of performance in real data has lead to most countries implementing a genomic BLUP approach due to the ease of implementation. Despite this, Bayes SSVS performed up to 9.8% better than the worst model especially for traits with major QTL that explain large amounts of genetic variation (Verbyla et al., 2009). As the genetic architecture of traits becomes more apparent and as the number of SNP increases, a Bayes SSVS or Bayesian model selection approach may offer advantages in selecting the best model containing only those SNP linked to the QTL. Additionally, should the genetic architecture be known, the prior distribution can to be set to match the true QTL distribution, something that is not possible with BLUP.

9.3. GENOMIC PREDICTION

The Bayesian hierarchical models were shown in this thesis to be rather robust and flexible when applied to varying genetic architectures in both simulated and real data. The results from the simulation studies presented in this thesis, and similar others, indicate that the models that assume the QTL distribution that matches fairly closely to the true distribution will produce high accuracies. In general, the accuracies of prediction for the different models were comparable. An overview of the performance, issues and important features of the Bayesian hierarchical models compared in Chapter 3-6 are presented in Table 9.1.

However, there were differences between the accuracies for specific traits. These could be explained by the variation in genetic architectures of traits and the subsequent match between the models' assumptions about the distribution of SNP effects and the trait's genetic architecture e.g. the "true" distribution of QTL effect. These differences between models were evident for only the traits with noticeably different architectures such as fat percentage which has a major QTL (mutation DGAT1 (Grisart et al., 2002)) explaining a large amount of genetic variation, or, protein kilograms where there are postulated to be thousands of QTL, each explaining only a small amount of genetic variation (eg. Chamberlain et al. (2007)). For fat percentage, Bayes SSVS and Bayes A produce significantly higher accuracies than BLUP.

Table 9.1- Overview of the Bayesian hierarchical models used for genomic prediction.

	Prior Distribution Assumptions	***Computational Demands***	***Accuracies of Prediction***	***Considerations before use***
BLUP	-Normal -Equal Variance	Low (very low[1])	High except for traits with atypical QTL distributions[2]	Relies on LD extending over long distances.
Bayes A	-t-distribution -unequal Variance	Low	Moderate/High across all traits	All SNP effects are non-zero and this single distribution can be prohibitive
Bayes B	-Mixture of t-distribution and point mass at zero -unequal Variance	High	High across all traits	Unviable due to computational and time demands
Bayes A/B	-Mixture of t-distribution and point mass at zero -unequal Variance	Moderate/High	High across all traits	Unviable due to computational and time demands
Bayes SSVS	-Mixture of two t-distribution (large and small) -unequal Variance	Low	High across all traits	The proportion of SNP sampled from the large distribution can influence the accuracies

[1]When BLUP is implemented in traditional mixed models replacing the A matrix with the GRM, [2]Those traits with QTL explaining large proportions of genetic variation e.g. fat percentage

The major result from the first simulation study (Chapter 3), where a variety of different QTL distributions were simulated, was that the hierarchical prior distributions that assumed unequal variances (i.e. Bayes B) produced more accurate DGV. However, these were significantly computationally slower than Bayes A that was able to utilise the faster Gibbs Sampler. The use of the Metropolis Hasting algorithm for the Bayes A/B hybrid approach (Chapter 4) also followed the trend of being computationally more demanding. Thus the use of the Bayes SSVS or any other approach that only requires the use of the Gibbs Sampler is significantly more efficient and given the general equality of performance with real data, more attractive.

The simulation study in Chapter 5 demonstrated that despite producing a noticeably different set of DGV, the DGV produced by a genomic BLUP approach had accuracies equivalent to or better than that of the sets of DGV estimated using the other models. This result was echoed in the real data study in Chapter 6 with Bayes BLUP producing comparable accuracies for all traits except fat percentage which has the mutation, DGAT1, explaining a large proportion of the genetic variation. For traits where there is no large QTL, the BLUP assumption of equal variance across SNP has a limited effect on accuracy of prediction. One reason for this is due to the data having a structured population (as it does in Chapter 4 and 5). This creates LD over long distances which has been reported to extend for more than 1 Mbp (The Bovine HapMap Consortium, 2009), thus when using a BLUP approach, the multiple SNP linked to a single QTL can take part of the overall QTL effect and explain small proportions of genetic variation caused by the QTL. This is also the reason that BLUP cannot produce high accuracies for fat percentage. It is unable to estimate the large effect of DGAT1 by spreading effects across the SNP linked to the mutation.

The results reported in Chapter 6 for real data studies agree with previously published results (Berry and Kearney, 2009, de Roos et al., 2009, Gredler et al., 2009, Harris et al., 2008, Lund and Su, 2009, Reinhardt et al., 2009, Schenkel et al., 2009, VanRaden et al., 2009). These indicate that while the Bayesian methods give competitive accuracies of DGV (slightly higher accuracies for traits with QTL explaining large amounts of variation), the uniformity of results across methods means that less computationally demanding approaches are attractive. For example, the robustness and ease of application of the genomic BLUP approach has lead to many countries adopting this approach for their genomic prediction model. However, it is important to note that the performance of BLUP is dependent on spreading the effects of QTL across a number of SNP. Consequently if the LD declines rapidly over short distances, for example in multi-breed data, then a BLUP approach may perform worse than other approaches such as Bayes SSVS, as it will be unable to capture the same amount of variation that approaches that allow for unequal variances between SNP.

The other approaches to genomic prediction presented in Chapter 2 including PCA, PLS, Genetic Algorithms and non-parametric approaches are yet to applied to real dairy data similar to that presented in Chapter 6 and comparable studies. Solberg et al. (2009) demonstrate that PCA and PLS produce lower accuracies and greater bias than Bayes B in simulated data, leading to the conclusion that they are unviable for genomic prediction due to the reduction in accuracy. This result for PLS was also shown in real data by Moser et al.(2009b). PLS and PCA are also reported to be less responsive to the addition of further marker information (Solberg et al., 2009), which makes these approaches less attractive with the future SNP chips which contain as many as 850,000 SNP. There is also a risk that the PCA approach will be particularly susceptible to population structure, in fact in human GWAS, variation in the first few PCA are generally removed to avoid false positive results due to population stratification (Price et al., 2006).

A more possible and rewarding alternative method, maybe the use of non-parametric models (Gianola et al., 2006, Gianola and van Kaam, 2008) (Section 2.3.6). They have been shown to produce promising results in real data (Gonzalez-Recio et al., 2008). However, further research is needed into suitable kernels and the apparent complexity of these approaches will continue to prevent many from implementing them.

The real data study also highlighted the importance of other parameters affecting the accuracy of selection including the heritability of the trait being analysed, the number of SNP (or LD between the SNP) and number of animals in the reference population. The study showed that traits with low heritability, such as fertility, produced low accuracies of prediction. However, the relationship between the number of records, heritability and accuracy of prediction means that an increase in the number of records should increase the accuracy of selection. An increase in the number of SNP would increase the LD found between the SNP and the QTL and this should also increase the accuracies of prediction (Calus et al., 2008).

Currently, most countries that have or will implement genomic selection, do not or will not select on DGV alone but combine the DGV with traditional breeding and selection information in the form of EBV, Parent Average (PA) or predictions based on additional pedigree information i.e. sire and maternal pathways (Berry and Kearney, 2009, de Roos et al., 2009, Harris and Montgomerie, 2009, Reinhardt et al., 2009, Schenkel et al., 2009, VanRaden et al., 2009). This addition is reported to add vital parental information that is not fully contained in the DGV. This information is not contained in the DGV despite the inclusion of the polygenic effect, reflecting the small subset of the total data that is used in the prediction analysis, given the limited numbers of animals genotyped to date. It is commonly accepted that in real data studies, such as those in Chapter 4 and 6, that the polygenic effect should be included to remove the effect of population structure to enable the more accurate estimation of the SNP effects. This is because the inclusion of the polygenic effect has been shown to produce slightly better accuracies of prediction while reducing the bias of the variance components (Calus and Veerkamp, 2007). Inclusion of additional pedigree information added to create the GEBV also increases the accuracy of prediction and selection as well as the reliability of the breeding values for most traits; this is demonstrated in Chapter 6. For some traits, however, such as fat percentage and protein kilograms mentioned earlier, this additional information does actually decrease the accuracy of prediction. This can be explained by the proportion of genetic variance accounted for by the SNP effects for the different traits; the amount of extra accuracy that the PA will add to the GEBV will be trait dependent and can be predicted based on the genetic architecture of the trait. For example for fat percentage almost all the genetic variation is captured by the SNPs thus the DGVs produce higher accuracies than the GEBVs.

9.4. SELECTION OF SUBSETS OF SNP

The difference between the number of SNP and phenotypes can be large. To address this, one proposal is to first select a small number of influential SNP that are most likely to be linked to QTL affecting the trait of interest; this set of SNP is used then in a second stage involving more sophisticated modelling of the relationship between the SNP and the trait of interest by simultaneously estimating the SNP effects to create the prediction equation.

The study presented in this thesis found that the pre-selection of SNP did not significantly increase accuracy of prediction, but it did increase the time and computation demands. The single SNP analysis to first select the sets of SNP was time consuming and consequently is not recommended as it provided no convincing additional benefits. Regardless, it would be interesting to examine whether pre-selecting using different models such as a Bayesian model or a machine learning procedure (that could produce different sets of SNP) would produce similar results.

In the (near) future, with the increase in the number of available SNP, the ability to pre-select the important features and possible QTL (and linked SNP) related to a trait may again become an important issue as approaches and procedures seek to deal with the dramatic increase in the dimensions of the data needed to be modelled. Thus, SNP selection could be a viable option to allow modelling of the data and potentially significantly reduce the time and computational demands. Additionally, using a reduced number of pre-selected SNP would also provide significant economical savings by requiring selection candidates to be genotyped only for the smaller selected number of SNP. To this extent, breeding companies appear likely to develop multiple low-density genotyping assays. Weigel et al. (2009) report that a set of 300 SNP selected due to having the largest effects might capture nearly half of the gain in reliability that could be achieved by using all SNP currently available through dense genotyping. They also postulate that a gain of two-thirds of the possible reliability could be achieved with 750 to 1,000 SNP.

9.5. ENERGY BALANCE

Energy balance (EB) is a minimally and difficult to record trait and, generally, it can only be recorded on nucleus or experimental farms. The study reported in this thesis achieved its objective by demonstrating the genetic basis of energy balance and that it could potentially be incorporated into selection programs using genomic selection. Despite the limitations on available data, genomic prediction was able to produce accuracies of prediction greater than a traditional polygenic model. Thus, the results indicated that EB can be estimated using genomic prediction. The low accuracy gained can be explained as a direct result of the small number of phenotypic records and the moderate heritability found for this trait. The heritability calculated with this data set was consistent with results of other studies (Huttmann et al., 2009, Veerkamp, 1998).

In order to consider including EB in breeding schemes, higher accuracies than found here would be desirable. This increase could be achieved through an increase in the heritability of the trait or an increase in the number of phenotypic records. In the future, selection for EB could be performed using genomic selection which could provide a valuable tool in finding a balance between production and non-production traits.

The study also indicated a possible candidate gene for EB. A single SNP appeared significantly related to EB. It was located on chromosome 21 and appeared to be in association with the nuclear receptor subfamily 2, group F, member 2 (NR2F2), otherwise known as chicken ovalbumin upstream promoter transcription factor II (COUP-TFII). COUP-TFII has been previously reported as playing an essential role in regulating adipogenesis, glucose homeostasis and energy metabolism (Li et al., 2009, Xu et al., 2008). It has also been reported as regulating growth hormone receptor 1A promoter activity (Xu et al., 2004) and mediating progesterone and controlling estrogen levels and thus involved in reproduction (Klinge et al., 1997, Kurihara et al., 2007, Nakshatri et al., 2000, Petit et al., 2007, Takamoto et al., 2005). Whilst the results of this association study are not conclusive and further validation is required, COUP-TFII appears to be a good candidate gene for EB.

The ability to select and include EB in selection indexes may indirectly increase the genetic gain for fertility traits and may allow selection for feed efficiency and methane emission without detrimental effects on health and fertility. Moderate to high genetic correlations have been found between EB and fertility traits. Thus, the use of EB in selection indexes, in addition to fertility, may prove beneficial and result in increased genetic gain for fertility. In addition, EB could be used with feed intake data to select animals for feed efficiency (Veerkamp, 1998) or reduce methane emission (Hegarty et al., 2007) to prevent the selection of animals that are highly productive and eat less but are therefore prone to health and fertility problems. By also selecting for a positive (or at least not an extremely negative) EB, animals selected should have a reduced number of health and fertility problems. Improving feed efficiency could be economically desirable as feed costs contribute the greatest proportion to production costs (Simm et al., 1994) while reducing methane emission has environmentally benefits.

9.6. GENOMIC SELECTION FOR DIFFICULT TO MEASURE TRAITS

The potential of genomic prediction and selection to allow selection for difficult traits, that would have been previously impossible or extremely difficult, is demonstrated in the research reported in this thesis, through the use of energy balance in Chapter 8. There are many other traits including several fertility and reproduction traits such as milk progesterone profiles and milk quality traits which are also difficult or expensive to record. Accounting for these traits, like EB, in selection has been complicated, since measuring them in progeny testing schemes is not practical. The accuracy of prediction has been shown in deterministic predictions (Daetwyler et al. 2009, Goddard 2008) to be a function of the heritability of the trait, the number of QTL and the number of records. The accuracy of prediction was defined by these authors as the correlation between the true and predicted breeding values. In this study only phenotypes were available to calculate the accuracy of prediction. Thus this original function was translated for use with phenotypes and DGV.

The study demonstrates that it is possible for such traits, with similar heritabilities and expected number of QTL to produce DGV with accuracies above 0.8 when there are more than approximately 2,600 (2,581 predicted for EB) phenotypic records available for use as the reference population. This indicates that it is possible to select for these traits using genomic selection by combining data from experimental and nucleus herds, where individually there are a limited numbers of raw phenotypic records.

9.7. PERMUTATION APPROACH FOR MULTI-LOCUS MODELS

In addition to their use for genomic prediction, genome wide SNP and multi-locus models can be used for identifying QTL affecting economically important traits. While genomic selection is useful for increasing genetic gain, understanding the biological features and pathways are equally important and may provide a way to change a trait through the identification of gene pathways to use as potential intervention targets. Also potentially useful for genomic selection, as demonstrated by the accuracies produced by fat percentage through the knowledge of the presence of DGAT1, identifying the QTL affecting the trait and understanding the genetic architecture may be useful for investigating new models and increasing accuracy.

Multi-locus models offer advantages over the traditional single locus models as they overcome the problem of multiple testing and estimation of the total variance explained by the QTL. Nevertheless, the problem of establishing significance is still important for multi-locus models. A permutation approach is presented in Chapter 7 that demonstrates that permutation testing can be used to enable the declaration of significant QTL for (Bayesian) multi-locus models. The approach is compared to other methods to establish significance. Bayes Factors and permutation testing produced useable thresholds while the posterior expected FDR was shown to be unviable for use with similar data sets.

The approach is demonstrated to identify QTL when using a multi-locus model and provide a valuable technique to establish significance for minor and moderate QTL with or without the presence of a major QTL by stratifying within genotype classes, where a known major gene exists. In addition, the problem of exchangeability when there is an additional linked independent second variable such as a structured pedigree was explored. Two approaches allowing the inclusion of the polygenic effect were presented and compared. Both approaches produced similar results and more research is required to establish the effect of stratifying within pedigree structure versus permuting across all data but potentially violating the condition of exchangeability where exchangeability means that under the null hypothesis (no association) that any order of observations is equally probable.

The effect of the number of permutations performed was also examined. The use of 50, 200, 500, 750 and 1000 permuted data sets to construct significance thresholds was investigated. Using only 50 permutation tests yielded a range of genome-wide thresholds (as a posterior probability) for a significance level of 0.05 from 0.360 to 1 for the real data. This clearly shows it is possible with such a small number to get significance thresholds that are extreme and will cause incorrect interpretation of the results. As the number of permutations increase the range of possible thresholds decreases. The use of 1000 permutations seems to be the minimum to enable confident use of a significance threshold. This result agrees with the work of Doerge and Churchill (1996) who stated that 1000 permutations is the lowest number needed for a significance level of 0.05 which is generally the lowest level of significance required.

9.8. FUTURE STUDIES

An increase in SNP information, towards whole genome re-sequence data, may lead to some approaches such as Bayesian models becoming unviable in their current implementation. This is because these approaches will require excessive computation and time to reach convergence. Due to the nature of MCMC sampling methods, multiple iterations are required for the chain to converge. If the parameters needed to be sampled increase 10 fold then the computational and memory demands are also going to increase 10 fold. The problem with this is that the computer processor speeds are no longer rapidly speeding up; instead companies are now choosing to just put more processors into a computer. Consequently, to take advantage of all available computer power, one approach would be to employ models that can utilise parallelisation or concurrency. One strategy that has been adopted is to employ blocking or local computation techniques for updating many parameters simultaneously (Boys et al., 2000, Goldstein and Wilkinson, 2000, Wilkinson and Yeung, 2002, 2004). Such techniques are very effective for improving the performance of MCMC schemes and could be used for Bayesian genomic prediction models.

Alternative faster algorithms have already been suggested, such as a faster Bayes B (Meuwissen et al., 2009) which is a non-MCMC based estimator and consequently functions faster by analytically performing the required integrations. In this approach, they maintain the original assumption that a number of SNP have a zero effect assumed by Bayes B. However, an alternative hierarchical model is used where the the non-zero SNP effects are assumed to come from a reflected exponential distribution. They use a modification of the Iterative Conditional Mode algorithm (Besag, 1986) that they call the Iterative Conditional Expectation algorithm. This algorithm used the expectation instead of the mode of the posterior and iteratively calculated $\mathrm{E}(\beta_i \mid \boldsymbol{y})$ for each SNP sequentially. Their modified fast approach produced accuracies only slightly lower than Bayes B, but not significantly. They demonstrated that their approach provided an efficient approach to modelling the same assumptions as Bayes B. In addition, Shepherd et al. (2009a, 2009b) also developed an faster method analogous to Bayes B using the much quicker EM (expectation-maximization) algorithm, emBayes B. This alternative formulation assumes the non-zero SNP effects are sampled from a double exponential distribution. Using EM theory, they use a set of E and M steps to converge to the maximum *a posteriori* parameter estimates. However, both approaches need more testing in multiple data sets.

The potential for genomic selection is currently restricted by the number of animals in the reference population. As these numbers grow, the accuracies that are achieved should increase. The increase in SNP density should lead to an increase in accuracy as more SNP located across the genome should capture more of the genetic variation by increasing the LD between markers (Hayes et al., 2009c). These increases may lead to more inequality across models, such that models that can utilise this additional information more successfully (such as model or variable selection approaches) may be able to produce more accurate DGV.

Also of interest currently is the use and sharing of genomic information and GEBV internationally (VanRaden and Sullivan, 2010), in order to build the size of reference populations. The diversity in prediction and construction of the GEBV, the difference in SNP, animals and traits means that there is significant care needed to use all the available information to provide reliable international GEBV.

9.9. CONCLUSION

Through a range of studies using both simulated and real data, the research reported in this thesis (and the associated publications) has investigated key aspects of genomic selection to enable a more comprehensive understanding of what makes a robust and accurate Bayesian prediction model and to explore the new possibilities introduced though genomic selection for selecting traits that could not be selected through alternative techniques. The comparative analysis has shown that while in real data the different models achieve an equality of prediction accuracy, however there is an exception for traits with atypical genetic architectures.

In the proceeding chapters, studies and outcomes are reported that contribute new knowledge and implications as well as an over-arching coherence to the current abundance of research on genomic selection. While many of results and trends reported here in the simulated and real data studies have been replicated in other studies, it is in this thesis that the results are drawn together into a comprehensive comparative analysis across models and methods. Additionally, novel variations and implementations have been introduced and analysed; much of the results from this innovative research have already been published.

In this chapter an overview of the main findings from the previous chapters are presented, including the implications for the current and future implementation and use of genomic selection. The future could, and should, see genomic selection become increasing effective as a selection technique as the increase in complete sequence information will further amplify the potential of genomic selection to accurately select for novel traits and increase the genetic gain across all traits.

CHAPTER 10

References

Agenas, S., E. Burstedt, and K. Holtenius. 2003. Effects of feeding intensity during the dry period. 1. Feed intake, body weight, and milk production. J. Dairy Sci. 86(3):870-882.

Aulchenko, Y. S., D.-J. de Koning, and C. Haley. 2007. Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177(1):577-585.

Baierl, A., M. Bogdan, F. Frommlet, and A. Futschik. 2006. On locating multiple interacting quantitative trait loci in intercross designs. Genetics 173(3):1693-1703.

Barendse, W., A. Reverter, R. J. Bunch, B. E. Harrison, W. Barris, and M. B. Thomas. 2007. A validated whole-genome association study of efficient food conversion in cattle. Genetics 176(3):1893-1905.

Bauer, A. M., F. Hoti, f. M. von Korf, K. Pillen, J. Leon, and M. J. Sillanpaa. 2009. Advanced backcross-QTL analysis in spring barley (H. vulgare ssp. spontaneum) comparing a REML versus a Bayesian model in multi-environmental field trials. Theoretical and Applied Genetics 119:105-123.

Beavis, W. D. 1994. The power and deceit of QTL experiments: lessons from comparitive QTL studies. Pages 250-266 in Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference. American Seed Trade Association, Washington, DC.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2007. GenBank. Nucl. Acids Res.:gkm929.

Berry, D. and F. Kearney. 2009. Genomic selection in Ireland. Proceeding of the Interbull International Workshop, Uppsala, Sweden. Bulletin no.39.

Besag, J. 1986. On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society Series B-Methodological 48:259-302.

Bogdan, M., J. K. Ghosh, and R. W. Doerge. 2004. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics 167(2):989-999.

Bost, B., D. de Vienne, F. Hospital, L. Moreau, and C. Dillmann. 2001. Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. Genetics 157(4):1773-1787.

Bost, B., C. Dillmann, and D. de Vienne. 1999. Fluxes and metabolic pools as model traits for quantitative genetics. I. The L-shaped distribution of gene effects. Genetics 153(4):2001-2012.

Boys, R. J., D. A. Henderson, and D. J. Wilkinson. 2000. Detecting homogeneous segments in DNA sequences by using hidden Markov models. Journal of the Royal Statistical Society Series C-Applied Statistics 49:269-285.

Brown, P. J., M. Vannucci, and T. Fearn. 1998. Multivariate Bayesian variable selection and prediction. Journal of the Royal Statistical Society Series B-Statistical Methodology 60:627-641.

Butler, W. R. and R. D. Smith. 1989. Interrelationships between energy-balance and postpartum reproduction function in dairy-cattle. J. Dairy Sci. 72(3):767-783.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178(1):553-561.

Calus, M. P. L. and R. F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics 124(6):362-368.

Carlborg, O., L. Andersson, and B. Kinghorn. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics 155(4):2003-2010.

Chamberlain, A. J., H. C. McPartlan, and M. E. Goddard. 2007. The Number of Loci That Affect Milk Production Traits in Dairy Cattle. Genetics 177(2):1117-1123.

Churchill, G. A. and R. W. Doerge. 1994. Empirical Threshold Values for Quantitative Trait Mapping. Genetics 138(3):963-971.

Coffey, M. P., G. Simm, J. D. Oldham, W. G. Hill, and S. Brotherstone. 2004. Genotype and diet effects on energy balance in the first three lactations of dairy cows. J. Dairy Sci. 87(12):4318-4326.

Conlon, E. M., J. J. Song, and J. S. Liu. 2006. Bayesian models for pooling microarray studies with multiple sources of replications. BMC Bioinformatics 7:247.

Coster, A., J. Bastiaansen, M. Calus, C. Maliepaard, and M. Bink. 2010. QTLMAS 2009: simulated dataset. BMC Proceedings 4(Suppl 1):S3.

Crump, R., B. Tier, G. Moser, J. S¨olkner, R. J. Kerr, A. F. Woolaston, K. R. Zenger, M. S. Khatkar, J. A. L. Cavanagh, and H. W. Raadsma. 2007. Genome-wide selection in dairy cattle: use of genetic algorithms in the estimation of molecular breeding values. in 17th Conference of the Association for the Advancement of Animal Breeding and Genetics, Armidale, NSW, Australia.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3(10):e3395.

Darwash, A. O., G. E. Lamming, and J. A. Woolliams. 1999. The potential for identifying heritable endocrine parameters associated with fertility in post-partum dairy cows. Anim. Sci. 68:333-347.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182(1):375-385.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage Disequilibrium and Persistence of Phase in Holstein-Friesian, Jersey and Angus Cattle. Genetics 179(3):1503-1512.

de Roos, A. P. W., C. Schrooten, E. Mullaart, S. van der Beek, G. de Jong, and W. Voskamp. 2009. Genomic selection at CRV. Proceeding of the Interbull International Workshop, Uppsala, Sweden. Bulletin no.39.

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim Sci. 82(13_suppl):E313-328.

Dewhurst, R. J., J. M. Moorby, M. S. Dhanoa, R. T. Evans, and W. J. Fisher. 2000. Effects of altering energy and protein supply to dairy cows during the dry period. 1. Intake, body condition, and milk production. J. Dairy Sci. 83(8):1782-1794.

Do, K. A., P. Muller, and F. Tang. 2005. A Bayesian mixture model for differential gene expression. Journal of the Royal Statistical Society Series C-Applied Statistics 54:627-644.

Doerge, R. W. and G. A. Churchill. 1996. Permutation tests for multiple loci affecting a quantitative character. Genetics 142(1):285-294.

Donoho, D. L. and V. Stodden. 2006. Breakdown point of model selection when the number of variables exceeds the number of observations. in International Joint Conference on Neural Networks.

Draper, N. R. and H. Smith. 1998. Applied regression analysis. 3rd ed. ed. John Wiley & Sons, New York.

Druet, T., S. Fritz, D. Boichard, and J. J. Colleau. 2006. Estimation of genetic parameters for quantitative trait loci for dairy traits in the French Holstein population. J. Dairy Sci. 89(10):4070-4076.

Ducrocq, V. and Z. Lui. 2009. Combining genomic and classical information in national BLUP evaluations. Proceeding of the Interbull Meeting 2009, Barcelona, Spain. Bulletin no.40.

Edwards, M. D., C. W. Stuber, and J. F. Wendel. 1987. Molecular-Marker-Facilitated Investigations of Quantitative-Trait Loci in Maize .1. Numbers, Genomic Distribution and Types of Gene-Action. Genetics 116(1):113-125.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least Angle Regression. The Annals of Statistics 32(2):407-451.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longmans Green Harlow, Essex, UK.

Fernando, R. 2009. A Mixture Model for Genomic Selection. in Statistical Genetics of Livestock for the Post-Genomic Era, Madison, WI, USA.

Fernando, R. L., D. Habier, C. Stricker, J. C. Dekkers, and L. R. Totir. 2008. Genomic selection. Acta Agriculturae Scandinavica, Section A - Animal Science 57:192–195.

Fernando, R. L., D. Habier, C. Stricker, J. C. M. Dekkers, and L. R. Totir. 2007. Genomic selection. Acta Agriculturae Scandinavica Section a-Animal Science 57(4):192-195.

Fisher, R. A. 1935. The Design of Experiments. Oliver and Boyd Ltd., London.

Foster, S. D., Verbyla A.P., and W. D. Pitchford. 2007a. Incorporating LASSO effects into a Mixed Model for Quantitative Trait Loci analysis Journal of Agriculture, Biological, and Environmental Statistics 12(2):300-314

Foster, S. D., A. P. Verbyla, and W. S. Pitchford. 2007b. A random model approach for the LASSO . Computational Statistics 23(2):217-233.

Friggens, N. C., P. Berg, P. Theilgaard, I. R. Korsgaard, K. L. Ingvartsen, P. Lovendahl, and J. Jensen. 2007. Breed and parity effects on energy balance profiles through lactation: Evidence of genetically driven body energy change. J. Dairy Sci. 90(11):5291-5305.

Garnsworthy, P. C., A. Lock, G. E. Mann, K. D. Sinclair, and R. Webb. 2008a. Nutrition, metabolism, and fertility in dairy cows: 1. Dietary energy source and ovarian function. J. Dairy Sci. 91(10):3814-3823.

Garnsworthy, P. C., A. Lock, G. E. Mann, K. D. Sinclair, and R. Webb. 2008b. Nutrition, metabolism, and fertility in dairy cows: 2. Dietary fatty acids and ovarian function. J. Dairy Sci. 91(10):3824-3833.

Gautier, M., R. R. Barcelona, S. Fritz, C. Grohs, T. Druet, D. Boichard, A. Eggen, and T. H. E. Meuwissen. 2006. Fine mapping and physical characterization of two linked quantitative trait loci affecting milk fat yield in dairy cattle on BTA26. Genetics 172(1):425-436.

Geladi, P. and B. R. Kowalski. 1986. Partial least-squares regression: a tutorial. Analytica Chimica Acta 185:1-17.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. Bayesian data analysis 2nd Ed ed. Chapman & Hall/CRC.

Geman, S. and D. Geman. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. Ieee Transactions on Pattern Analysis and Machine Intelligence 6(6):721-741.

Genovese, C. and L. Wasserman. 2002. Operating Characteristics and Extensions of the False Discovery Rate Procedure. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64(3):499-517.

George, E. I. and R. E. McCulloch. 1993. Variable Selection Via Gibbs Sampling. J. Am. Stat .Assoc. 88(423):881-889.

George, E. I. and R. E. McCulloch. 1997. Approaches for Bayesian variable selection. Statistica Sinica 7(2):339-373.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183(1):347-363.

Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3):1761-1776.

Gianola, D., M. Perez-Enciso, and M. A. Toro. 2003. On marker-assisted prediction of genetic value: Beyond the ridge. Genetics 163(1):347-365.

Gianola, D. and J. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178(4):2289-2303.

Gibbs, R., G. Weinstock, K. Steven., L. Skow, and J. Womack. 2002. Bovine Genomic Sequencing Initiative, Cattle-izing the Human Genome National Human Research Institute

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov chain Monte Carlo in practice Chapman & Hall.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2006a. ASREML. 2 ed. VSN International Ltd., Hemel Hempstead, HP1 1ES, UK.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2006b. ASREML Program user manual. 2 ed. . VSN International Ltd., Hemel Hempstead, HP1 1ES, UK.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257.

Goddard, M. E. 2008. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257.

Goddard, M. E. and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10(6):381-391.

Goldberger, A. S. 1962. Best Linear Unbiased Prediction in Generalized Linear-Regression Model. Journal of the American Statistical Association 57(298):369-&.

Goldstein, M. and D. J. Wilkinson. 2000. Bayes linear analysis for graphical models: The geometric approach to local computation and interpretive graphics. Statistics and Computing 10(4):311-324.

Gonzalez-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa, and S. Avendano. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178(4):2305-2313.

Gredler, B., K. G. Nirea, T. R. Solberg, C. Egger-Danner, T. H. E. Meuwissen, and J. Sölkner. 2009. Genomic selection in Fleckvieh/Simmental – First results. Proceeding of the Interbull Meeting 2009, Barcelona, Spain. Bulletin no.40.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711-732.

Griffin, J. E. and P. J. Brown. 2005. Alternative prior distributions for variable selection with very many more variables thans observations. Dept. of Statistics, University of Warwick.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12(2):222-231.

Habier, D., R. Fernando, K. Kizilkaya, and D. J. Garrick. 2010a. Extension of the Bayesian Alphabet for Genomic Selection. in 9th World Congress on Genetics Applied to Livestock Production. Lepzig, Germany.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389-2397.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic Selection Using Low-Density Marker Panels. Genetics 182(1):343-353.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010b. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42(1):5.

Hans, C. 2009. Bayesian lasso regression. Biometrika 96(4):835-845.

Hardy, J. and A. Singleton. 2009. Genomewide Association Studies and Human Disease. N Engl J Med 360(17):1759-1768.

Harris, B. L., D. L. Johnson, and R. J. Spelman. 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. Proceeding of theInterbull Meeting, Niagara Falls, Canada Bulletin no.38.

Harris, B. L. and W. A. Montgomerie. 2009. Current status of the use of genomic information in the national genetic evaluation in New Zealand. Proceeding of the Interbull International Workshop, Uppsala, Sweden. Bulletin no.39.

Hastings, W. K. 1970. Monte-Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57(1):97-&.

Hayes, B., P. J. Bowman, A. J. Chamberlain, Verbyla K.L., and M. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41:51.

Hayes, B. and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution 33(3):209-229.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited Review: Genomic Selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92:433-443.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009c. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92(2):433-443.

Hayes, B. J. and M. E. Goddard. 2008a. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim Sci. 86(9):2089-2092.

Hayes, B. J. and M. E. Goddard. 2008b. Technical note: Prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science 86(9):2089-2092.

Hegarty, R. S., J. P. Goopy, R. M. Herd, and B. McCorkell. 2007. Cattle selected for lower residual feed intake have reduced daily methane production. J. Anim Sci. 85(6):1479-1486.

Henderson, C. R. 1950. Estimation of Genetic Parameters. Annals of Mathematical Statistics 21(2):309-310.

Henderson, C. R. 1973. Sire evaluation and genetic trends. . Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush, American Society of Animal Science and American Dairy Science Association, Champaign, IL, :10-41.

Henderson, C. R. 1975a. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31(2):423-447.

Henderson, C. R. 1975b. Use of All Relatives in Intraherd Prediction of Breeding Values and Producing Abilities. Journal of Dairy Science 58(12):1910-1916.

Henderson, C. R. 1976. Simple Method for Computing Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics 32(1):69-83.

Henderson, C. R. 1977. Best Linear Unbiased Prediction of Breeding Values Not in Model for Records. Journal of Dairy Science 60(5):783-787.

Henderson, C. R. 1978. Undesirable Properties of Regressed Least-Squares Prediction of Breeding Values. Journal of Dairy Science 61(1):114-120.

Henderson, C. R., H. W. Carter, and J. T. Godfrey. 1954. Use of contemporary herd average in appraising progeny tests of dairy bulls. . Journal of Animal Science 13(949.).

Heuven, H. C. and L. L. Janss. 2010. Bayesian multi-QTL mapping for growth curve parameters. BMC Proceedings 4(Suppl 1):S12.

Hoeschele, I. and P. M. Vanraden. 1993a. Bayesian-Analysis of Linkage between Genetic-Markers and Quantitative Trait Loci .1. Prior Knowledge. Theoretical and Applied Genetics 85(8):953-960.

Hoeschele, I. and P. M. Vanraden. 1993b. Bayesian-Analysis of Linkage between Genetic-Markers and Quantitative Trait Loci .2. Combining Prior Knowledge with Experimental-Evidence. Theoretical and Applied Genetics 85(8):946-952.

Hoh, J., A. Wille, R. Zee, S. Cheng, R. Reynolds, K. Lindpaintner, and J. Ott. 2000. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. Annals of Human Genetics 64:413-417.

Huttmann, H., E. Stamer, W. Junge, G. Thaller, and E. Kalm. 2009. Analysis of feed intake and energy balance of high-yielding first lactating Holstein cows with fixed and random regression models. Animal 3(2):181-188.

Jansen, R. C. 1993. Interval Mapping of Multiple Quantitative Trait Loci. Genetics 135(1):205-211.

Jorge, V., A. Dowkiw, P. Faivre-Rampant, and C. Bastien. 2005. Genetic architecture of qualitative and quantitative Melampsora larici-populina leaf rust resistance in hybrid poplar: genetic mapping and QTL detection. New Phytologist 167(1):113-127.

Jorritsma, R., T. Wensing, T. A. M. Kruip, P. Vos, and J. Noordhuizen. 2003. Metabolic changes in early lactation and impaired reproductive performance in dairy cows. Vet. Res. 34(1):11-26.

Kao, C. H., Z. B. Zeng, and R. D. Teasdale. 1999. Multiple interval mapping for quantitative trait loci. Genetics 152(3):1203-1216.

Kass, R. E. and A. E. Raftery. 1995. Bayes Factors. Journal of the American Statistical Association 90(430):773-795.

Kennedy, B. W., M. Quinton, and J. A. van Arendonk. 1992. Estimation of effects of single genes on quantitative traits. J. Anim Sci. 70(7):2000-2012.

Khatkar, M. S., P. C. Thomson, I. Tammen, and H. W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. Evol. 36(2):163-190.

Klinge, C. M., B. F. Silver, M. D. Driscoll, G. Sathya, R. A. Bambara, and R. Hilf. 1997. Chicken ovalbumin upstream promoter transcription factor interacts with estrogen receptor, binds to estrogen response elements and half-sites, and inhibits estrogen-induced gene expression. J. Biol. Chem. 272(50):31465-31474.

Knijnenburg, T. A., L. F. A. Wessels, M. J. T. Reinders, and I. Shmulevich. 2009. Fewer permutations, more accurate P-values. Bioinformatics 25(12):I161-I168.

Knott, S. A. and C. S. Haley. 1992. Aspects of Maximum-Likelihood Method for the Mapping of Quantitative Trait Loci in Line Crosses. Genet. Res. 60(2):139-151.

Kurihara, I., D. K. Lee, F. G. Petit, J. Jeong, K. Lee, J. P. Lydon, F. J. DeMayo, M. J. Tsai, and S. Y. Tsai. 2007. COUP-TFII mediates progesterone regulation of uterine implantation by controlling ER activity. PLoS Genet. 3(6):1053-1064.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2005. Applied Linear Statistical Models. Vol. Ed. 5. McGraw-Hill, New York.

Lander, E. S. and D. Botstein. 1989. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps Genetics 121(1):185-199.

Lehmann, E. C. 1986. Testing Statistical Hypotheses. John Wiley and Sons, London.

Li, L. P., X. Xie, J. Qin, G. S. Jeha, P. K. Saha, J. Yan, C. M. Haueter, L. Chan, S. Y. Tsai, and M. J. Tsai. 2009. The Nuclear Orphan Receptor COUP-TFII Plays an Essential Role in Adipogenesis, Glucose Homeostasis, and Energy Metabolism. Cell Metab. 9(1):77-87.

Liu, Y., X. Qin, X.-Z. Song, H. Jiang, Y. Shen, K. J. Durbin, S. Lien, M. Kent, M. Sodeland, Y. Ren, L. Zhang, E. Sodergren, P. Havlak, K. Worley, G.

Weinstock, and R. Gibbs. 2009. Bos taurus genome assembly. BMC Genomics 10(1):180.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendano. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of Animal Breeding and Genetics 124(6):377-389.

Lund, M. S. and G. Su. 2009. Genomic selection in the Nordic countries. Proceeding of the Interbull International Workshop, Uppsala, Sweden. Bulletin no.39.

Luo, Z. W. and M. J. Kearsey. 1989. Maximum-Likelihood-Estimation of Linkage Between a Marker Gene and a Quantitative Locus. Heredity 63:401-408.

Lush, J. L. 1931. The number of daughters necessary to prove a sire. Journal of Dairy Science (14):209-220.

Lush, J. L. 1933. The bull index problem in the light of modern genetics. . Journal of Dairy Science (16):501-522.

Macciotta, N., G. Gaspa, R. Steri, C. Pieramati, P. Carnier, and C. Dimauro. 2009. Pre-selection of most significant SNPS for the estimation of genomic breeding values. BMC Proceedings 3(Suppl 1):S14.

Macciotta, N. P. P., M. Mele, G. Conte, A. Serra, M. Cassandro, R. Dal Zotto, A. C. Borlino, G. Pagnacco, and P. Secchiari. 2008. Association between a polymorphism at the stearoyl CoA desaturase locus and milk production traits in Italian Holsteins. J. Dairy Sci. 91(8):3184-3189.

Mackay, T. F. C. 2001. The genetic architecture of quantitative traits. Annual Review of Genetics 35:303-339.

Martinez, O. and R. N. Curnow. 1992. Estimating the Locations and the Sizes of the Effects of Quantitaive Trait Loci Using Flanking Markers. Theor. Appl. Genet. 85(4):480-488.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5):356-369.

McNamara, S., J. J. Murphy, F. P. O'Mara, M. Rath, and J. F. Mee. 2008. Effect of milking frequency in early lactation on energy metabolism, milk production and reproductive performance of dairy cows. Livest. Sci. 117(1):70-78.

McNamara, S., J. J. Murphy, M. Rath, and F. P. O'Mara. 2003. Effects of different transition diets on energy balance, blood metabolites and reproductive performance in dairy cows. Livest. Prod. Sci. 84(3):195-206.

McVean, G. 2009. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5(10):e1000686.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of State Calculations by Fast Computing Machines. Journal of Chemical Physics 21(6):1087-1092.

Meuwissen, T. H. E. and M. E. Goddard. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genetics Selection Evolution 33(6):605-634.

Meuwissen, T. H. E. and M. E. Goddard. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36 261–279.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829.

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genetics Selection Evolution 41:2.

Miglior, F., B. L. Muir, and B. J. Van Doormaal. 2005. Selection indices in Holstein cattle of various countries. J. Dairy Sci. 88(3):1255-1263.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009a. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetic Selection Evolution 41(1):56.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009b. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution 41(1):56.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding & Genetics 124(6):342-355.

Mulder, H. A., T. H. E. Meuwissen, M. P. L. Calus, and R. F. Veerkamp. 2009. The effect of missing marker genotypes on the accuracy of gene-assisted breeding value estimation: a comparison of methods. Animal Forthcoming(-1):1-11.

Nakshatri, H., M. S. Mendonca, P. Bhat-Nakshatri, N. M. Patel, R. J. Goulet, and K. Cornetta. 2000. The orphan receptor COUP-TFII regulates G2/M progression of breast cancer cells by modulating the expression/activity of p21(WAF1/CIP1), cyclin D1, and cdk2. Biochem. Biophys. Res. Commun. 270(3):1144-1153.

Narita, A. and Y. Sasaki. 2004. Detection of multiple QTL with epistatic effects under a mixed inheritance model in an outbred population. Genetics Selection Evolution 36(4):415-433.

Nielsen, H. M., A. K. Sonesson, H. Yazdi, and T. H. E. Meuwissen. 2009. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. Aquaculture 289(3-4):259-264.

Park, T. and G. Casella. 2008. The Bayesian Lasso. Journal of the American Statistical Association 103(482):681-686.

Petersson, K. J., B. Berglund, E. Strandberg, H. Gustafsson, A. P. F. Flint, J. A. Woolliams, and M. D. Royal. 2007. Genetic analysis of postpartum measures of luteal activity in dairy cows. J. Dairy Sci. 90(1):427-434.

Petit, F. G., S. P. Jamin, I. Kurihara, R. R. Behringer, F. J. DeMayo, M. J. Tsai, and S. Y. Tsai. 2007. Deletion of the orphan nuclear receptor COUP-THII in uterus leads to placental deficiency. Proc. Natl. Acad. Sci. USA 104(15):6293-6298.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2007 -a. coda: Output analysis and diagnostics for MCMC. R package version 0.13-1.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2007 -b. coda: Output analysis and diagnostics for MCMC. . R package version 0.13-1.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. 38(8):904-909.

Pryce, J. E., M. D. Royal, P. C. Garnsworthy, and I. L. Mao. 2004. Fertility in the high-producing dairy cow. Livest. Prod. Sci. 86(1-3):125-135.

Reinhardt, F., Z. Lui, F. Seefried, and G. Thaller. 2009. Implementation of genomic evaluation in German Holsteins. Proceeding of the Interbull Meeting 2009, Barcelona, Spain. Bulletin no.40.

Royal, M. D., A. O. Darwash, A. P. E. Flint, R. Webb, J. A. Woolliams, and G. E. Lamming. 2000. Declining fertility in dairy cattle: changes in traditional and endocrine parameters of fertility. Anim. Sci. 70:487-501.

Sahara, G., B. Guldbrandsen, L. Janss, and M. S. Lund. 2010. Comparison of Association Mapping Methods in a Complex Pedigree Population. Genetic Epidemiology 34:455-462.

Satagopan, J. M., Y. S. Yandell, M. A. Newton, and T. C. Osborn. 1996. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144(2):805-816.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123(4):218-223.

Schenkel, F. S., M. Sargolzaei, G. Kistemaker, G. B. Jansen, P. Sullivan, B. J. Van Doormaal, P. M. VanRaden, and G. R. Wiggans. 2009. Reliability of genomic evaluation of Holstein cattle in Canada. Proceeding of the Interbull International Workshop, Uppsala, Sweden. Bulletin no.39.

Seaton, G., C. S. Haley, S. A. Knott, M. Kearsey, and P. M. Visscher. 2002. QTL Express: mapping quantitative trait loci in simple and complex pedigrees. Bioinformatics 18(2):339-340.

Sen, S. and G. A. Churchill. 2001. A statistical framework for quantitative trait mapping. Genetics 159(1):371-387.

Shepherd, R. K., T. H. E. Meuwissen, and J. A. Woolliams. 2009a. Genomic Selection using a Fast EM Algorithm 1. Understanding the Methodology Pages 80-83 in Association for the Advancement of Animal Breeding and Genetics, Barossa Valley, South Australia.

Shepherd, R. K., T. H. E. Meuwissen, and J. A. Woolliams. 2009b. Genomic Selection using a Fast EM Algorithm 2.Analysis of Simulated Data Pages 84-87 in Association for the Advancement of Animal Breeding and Genetics, Barossa Valley, South Australia.

Sherman, E. L., J. D. Nkrumah, C. Li, R. Bartusiak, B. Murdoch, and S. S. Moore. 2009. Fine mapping quantitative trait loci for feed intake and feed efficiency in beef cattle. J. Anim. Sci. 87(1):37-45.

Shriner, D. 2009. Mapping multiple quantitative trait loci under Bayes error control. Genetic Research 91:147-159.

Sillanpaa, M. J. and E. Arjas. 1998. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148(3):1373-1388.

Simm, G., R. F. Veerkamp, and P. Persaud. 1994. The Economic-Performance of Dairy-Cows of Different Predicted Genetic Merit for Milk Solids Production. Anim. Prod. 58:313-320.

Smith, L. I. 2002. A tutorial on Principal Components Analysis.

Smola, A. J. and B. Schölkopf. 2004. A tutorial on support vector regression. Statistics and Computing 14:199-222. .

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genetics Selection Evolution 41.

Soller, M., T. Brody, and A. Genizi. 1976. Power for Experimental Design for Detection of Linkage Between Marker Loci and Quantitative Loci in Crosses Between Inbred Lines. Theor. Appl. Genet. 47(1):35-39.

Stephens, D. A. 1998. Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics 54(4):1334-1347.

Stoop, W. M., A. Schennink, M. Visker, E. Mullaart, J. A. M. van Arendonk, and H. Bovenhuis. 2009. Genome-wide scan for bovine milk-fat composition. I. Quantitative trait loci for short- and medium-chain fatty acids. J. Dairy Sci. 92(9):4664-4675.

Storey, J. D. 2003. The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. The Annals of Statistics 31(6):2013-2035.

Storey, J. D., J. M. Akey, and L. Kruglyak. 2005. Multiple locus linkage analysis of genomewide expression in yeast. Plos Biology 3(8):1380-1390.

Swartz, M. D., M. Kimmel, P. Mueller, and C. I. Amos. 2006. Stochastic search gene suggestion: A Bayesian hierarchical model for gene mapping. Biometrics 62(2):495-503.

Takamoto, N., I. Kurihara, K. Lee, F. J. DeMayo, M. J. Tsai, and S. Y. Tsai. 2005. Haploinsufficiency of chicken ovalbumin upstream promoter transcription factor II in female reproduction. Mol. Endocrinol. 19(9):2299-2308.

Taniguchi, M., T. Utsugi, K. Oyama, H. Mannen, M. Kobayashi, Y. Tanabe, A. Ogino, and S. Tsuji. 2004. Genotype of stearoyl-CoA desaturase is associated with fatty acid composition in Japanese Black cattle. Mamm. Genome 15(2):142-148.

ter Braak, C. J. F. 2006. Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising. Computational Statistics & Data Analysis 51(2):1232-1242.

ter Braak, C. J. F., M. P. Boer, and M. Bink. 2005. Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170(3):1435-1438.

Thaller, G. and I. Hoeschele. 1996a. A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci .1. Methodology. Theoretical and Applied Genetics 93(7):1161-1166.

Thaller, G. and I. Hoeschele. 1996b. A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci .2. A simulation study. Theoretical and Applied Genetics 93(7):1167-1174.

The Bovine HapMap Consortium. 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324(5926):528-532.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1):267-288.

Usai, M. G., M. E. Goddard, and B. J. Hayes. 2009. LASSO with cross-validation for genomic selection. Genetics Research 91(06):427-436.

Vach, K., W. Sauerbrei, and M. Schumacher. 2001. Variable selection and shrinkage: comparison of some approaches. Statistica Neerlandica 55(1):53-75.

van der Lende, T., L. Kaal, R. M. G. Roelofs, R. F. Veerkamp, C. Schrooten, and H. Bovenhuis. 2004. Infrequent milk progesterone measurements in daughters enable bull selection for cow fertility. J. Dairy Sci. 87(11):3953-3957.

VanRaden, P. and P. Sullivan. 2010. International genomic evaluation methods for dairy cattle. Genetics Selection Evolution 42(1):7.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91(11):4414-4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92(1):16-24.

Vapnik, V. N. 1998. Statistical Learning Theory. New York: John Wiley & Sons.

Veerkamp, R. F. 1998. Selection for economic efficiency of dairy cattle using information on live weight and feed intake: A review. J. Dairy Sci. 81(4):1109-1119.

Veerkamp, R. F., B. Beerda, and T. van der Lende. 2003. Effects of genetic selection for milk yield on energy balance, levels of hormones, and metabolites in lactating cattle, and possible links to reduced fertility's. Livest. Prod. Sci. 83(2-3):257-275.

Veerkamp, R. F., J. K. Oldenbroek, H. J. Van Der Gaast, and J. H. J. Van Der Werf. 2000. Genetic correlation between days until start of luteal activity and milk yield, energy balance, and live weights. J. Dairy Sci. 83(3):577-583.

Verbyla, K., P. Bowman, B. Hayes, and M. Goddard. 2010a. Sensitivity of genomic selection to using different prior distributions. BMC Proceedings 4(Suppl 1):S5.

Verbyla, K. L., M. P. Calus, H. A. Mulder, Y. de Haas, and R. F. Veerkamp. 2010b. Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. Journal of Dairy Science 93:2757-2764

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Short Note: Accuracy of Genomic Selection using Stochastic Search Variable Selection in Australian Holstein Friesian dairy cattle. Genetic Research 91:307–311.

Wang, H., Y. M. Zhang, X. M. Li, G. L. Masinde, S. Mohan, D. J. Baylink, and S. Z. Xu. 2005. Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics 170(1):465-480.

Watters, R. D., M. C. Wiltbank, J. N. Guenther, A. E. Brickner, R. R. Rastani, P. M. Fricke, and R. R. Grummer. 2009. Effect of dry period length on reproduction during the subsequent lactation. J. Dairy Sci. 92(7):3081-3090.

Weigel, K. A., G. de los Campos, O. Gonzalez-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science 92(10):5248-5257.

Weller, J. I. 1986. Maximum-Likelihodd Techniques for the Mapping and Analysis of Quantitative Trait Loci with the Aid of Genetic Markers Biometrics 42(3):627-640.

Weller, J. I., M. Shlezinger, and M. Ron. 2005. Correcting for bias in estimation of quantitative trait loci effects. Genetics Selection Evolution 37(5):501-522.

Whittaker, J. C., R. Thompson, and M. C. Denham. 2000. Marker-assisted selection using ridge regression. Genetical Research 75(2):249-252.

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431–3436.

Wilkinson, D. J. and S. K. H. Yeung. 2002. Conditional simulation from highly structured Gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models. Statistics and Computing 12(3):287-300.

Wilkinson, D. J. and S. K. H. Yeung. 2004. A sparse matrix approach to Bayesian computation in large linear models. Computational Statistics & Data Analysis 44(3):493-516.

Wu, R. L. and B. L. Li. 2000. A quantitative genetic model for analyzing species differences in outcrossing species. Biometrics 56(4):1098-1104.

Xu, Q., N. Walther, and H. Jiang. 2004. Chicken ovalbumin upstream promoter transcription factor II (COUP-TFII) and hepatocyte nuclear factor 4 gamma (HNF-4 gamma) and HNF-4 alpha regulate the bovine growth hormone receptor 1A promoter through a common DNA element. J. Mol. Endocrinol. 32(3):947-961.

Xu, S. Z. 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163(2):789-801.

Xu, Z., S. Yu, C. H. Hsu, J. Eguchi, and E. D. Rosen. 2008. The orphan nuclear receptor chicken ovalbumin upstream promoter-transcription factor II is a critical regulator of adipogenesis. Proc. Natl. Acad. Sci. USA 105(7):2421-2426.

Yandell, B. S., T. Mehta, S. Banerjee, D. Shriner, R. Venkataraman, J. Y. Moon, W. W. Neely, H. Wu, R. von Smith, and N. J. Yi. 2007. R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. Bioinformatics 23(5):641-643.

Yi, N. and D. Shriner. 2008. Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses. Heredity 100(3):240-252.

Yi, N. and S. Xu. 2008. Bayesian LASSO for Quantitative Trait Loci Mapping. Genetics 179(2):1045-1055.

Yi, N. J. 2004. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. Genetics 167(2):967-975.

Yi, N. J., V. George, and D. B. Allison. 2003. Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics 164(3):1129-1138.

Yi, N. J., D. Shriner, S. Banerjee, T. Mehta, D. Pomp, and B. S. Yandell. 2007. An efficient bayesian model selection approach for interacting QTL models with many effects Genetics:ahead of print.

Yi, N. J. and S. Z. Xu. 2000. Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155(3):1391-1403.

Yi, N. J. and S. Z. Xu. 2002. Mapping quantitative trait loci with epistatic effects. Genetical Research 79(2):185-198.

Yi, N. J., B. S. Yandell, G. A. Churchill, D. B. Allison, E. J. Eisen, and D. Pomp. 2005. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170(3):1333-1344.

Zeng, Z. B. 1993. Theoretical Basis for Separation of Multiple Linked Gene Effects in Mapping Quantitative Trait Loci. Proc. Natl. Acad. Sci. USA 90(23):10972-10976.

Zeng, Z. B. 1994. Precision Mapping of Quantitative Trait Loci. Genetics 136(4):1457-1468.

Zhang, M., K. L. Montooth, M. T. Wells, A. G. Clark, and D. B. Zhang. 2005. Mapping multiple quantitative trait loci by Bayesian classification. Genetics 169(4):2305-2318.

Zou, W. and Z. B. Zeng. 2009. Multiple interval mapping for gene expression QTL analysis. Genetica 137(2):125-134.

CHAPTER 11

Appendices

Appendix A- Publications from thesis

A1 - Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Short Note: Accuracy of Genomic Selection using Stochastic Search Variable Selection in Australian Holstein Friesian dairy cattle. Genetic Research 91:307–311

A2 - Verbyla, K., P. Bowman, B. Hayes, and M. Goddard. 2010a. Sensitivity of genomic selection to using different prior distributions. BMC Proceedings 4:S5

A3 - Verbyla, K. L., M. P. Calus, H. A. Mulder, Y. de Haas, and R. F. Veerkamp. 2010b. Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. Journal of Dairy Science 93:2757-2764

doi:10.1017/S0016672309990243

SHORT NOTE
Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle

KLARA L. VERBYLA[1,2,3]*, BEN J. HAYES[1], PHILIP J. BOWMAN[1] AND MICHAEL E. GODDARD[1,2,3]

[1] *Biosciences Research Division, Department of Primary Industries Victoria, 1 Park Drive, Bundoora 3083, Australia*
[2] *Melbourne School of Land and Environment, The University of Melbourne, Parkville 3010, Australia*
[3] *The Cooperative Research Centre for Beef Genetic Technologies, University of New England, Armidale, NSW 2351, Australia*

(*Received 21 July 2009 and in revised form 10 September 2009*)

Summary

Genomic selection describes a selection strategy based on genomic breeding values predicted from dense single nucleotide polymorphism (SNP) data. Multiple methods have been proposed but the critical issue is how to decide whether an SNP should be included in the predictive set to estimate breeding values. One major disadvantage of the traditional Bayes B approach is its high computational demands caused by the changing dimensionality of the models. The use of stochastic search variable selection (SSVS) retains the same assumptions about the distribution of SNP effects as Bayes B, while maintaining constant dimensionality. When Bayesian SSVS was used to predict genomic breeding values for real dairy data over a range of traits it produced accuracies higher or equivalent to other genomic selection methods with significantly decreased computational and time demands than Bayes B.

1. Introduction

Traditionally selection to improve profitability of livestock production has been based on phenotypic and pedigree information. However, the availability of dense single nucleotide polymorphisms (SNPs) and dramatic reduction in the cost of acquiring this information has allowed the inclusion of genome wide marker information in the prediction of animals' breeding values.

Meuwissen *et al.* (2001) introduced genomic selection as a selection strategy based on genomic breeding values predicted from dense marker data. The method implicitly recognized the fact that quantitative traits such as those affecting profit of livestock production are controlled by the segregation of large numbers of multiple quantitative trait loci (QTLs), and therefore predicts an animal's breeding value by simultaneously evaluating and summing large numbers of marker effects across the entire genome. The method makes the assumption that the markers are in linkage disequilibrium (LD) with the QTL. The higher the density of the markers is, the greater the level of LD between the markers and the QTL and thus the greater proportion of genetic variance that can be explained by the markers.

In the reference population, where the SNP effects are predicted, the number of marker effects (p) to simultaneously estimate will typically be substantially larger than the number of animals genotyped (n), which leads to the difficulty of an over-saturated model (i.e. $p > n$). Thus, a model for genomic selection must be able to overcome this problem. The other necessity is a sparse model because of the large number of SNP effects that are zero or close to zero. Subsequently, a crucial question is how to decide whether an SNP is in, or out of the set of SNPs chosen to give the most accurate prediction of breeding values in independent data sets. One potential approach is to use shrinkage methods such as the least absolute shrinkage and selection operator (LASSO), where all SNPs are included in the predictive set but the smaller effects are shrunk back towards zero (Tibshirani, 1996). Another approach is to use the reversible jump Markov chain Monte Carlo (MCMC) algorithm

* Corresponding author. e-mail: klara.verbyla@dpi.vic.gov.au

(Green, 1995), which uses a variable dimension model space approach that allows the SNPs in the predictive set to change. Stochastic search variable selection (SSVS) (George & McCulloch, 1993) provides a method to maintain a constant dimensionality across all models but allows the SNPs in the predictive set to change. It allows this by instead of removing all non-significant parameters (those that would be excluded from the predictive set using the reversible jump algorithm) from the model, their effects are limited to values very close to zero.

The major advantage of this method is that the posterior distribution of all parameters can be sampled directly using the Gibbs sampler, instead of using more computationally demanding algorithms such as the reversible jump algorithm. SSVS has been previously used for identifying multiple QTLs (Yi *et al.*, 2003), multivariate regression models (Brown *et al.*, 1998), gene mapping (Swartz *et al.*, 2006) and generalized linear models (George & McCulloch, 1997). It has also been utilized for analysing multi-trait QTL mapping data (Meuwissen & Goddard, 2004), and subsequently to investigate the effect that different methods for defining haplotypes and the effect of the inclusion of the polygenic effect had on the accuracy of genomic selection in simulated data (Calus *et al.*, 2008; Calus & Veerkamp, 2007).

In this paper, we demonstrate that a Bayesian SSVS can be used effectively when compared with other methods for genomic selection using real SNP data. It also provides an viable alternative to more computationally demanding approaches such as Bayes B (Meuwissen *et al.*, 2001).

2. Materials and methods

(i) *SNP data*

The data set contained 1498 Australian Holstein-Friesian bulls genotyped for the Illumina Bovine50K array. After quality control, 39 048 SNPs remained in the predictive set. The quality control applied to the SNP data is described by Hayes *et al.* (2009). The reference data set where the SNP effects were predicted contained 1098 bulls born between 1940 and 2000. The phenotypes for these bulls were Australian breeding values (ABV) for protein kg, fat kg, protein percentage, fat percentage and daughter fertility, all deregressed to remove any contribution from relatives (Hayes *et al.*, 2009). Daughter fertility here is defined as the difference between bulls for the percentage of their daughters pregnant 6 weeks after mating start date or 100 days after calving in year-round herds. The validation set contained 400 genotyped bulls proven from the years 2005, 2006 and 2007 with ABV which included information from at least 100 milking daughters to enable comparison with predicted marker estimated breeding value (MEBVs).

(ii) *Model*

At each locus (total number of loci, p) there are three possible combinations of two alleles (e.g. A or B), the homozygote of one allele (AA), the heterozygote (AB) and the homozygote of the other allele (BB). These are then quantitatively represented by 0, 1 and 2, respectively. The model fitted to the above data was then

$$\boldsymbol{y} = \mu \mathbf{1}_n + \sum_{j=1}^{q} \boldsymbol{X}_j \beta_j + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of phenotypes of the trait being analysed for all n individuals, μ is the mean, $\mathbf{1}_n$ is a vector of ones of length n, $\boldsymbol{X}_j$ is a vector of indicator variables representing the genotypes of the jth marker for all individuals ($x_{ij} = 0, 1, 2$), β_j is the size of the QTL effect associated with marker j, $\boldsymbol{u}$ is the vector of random polygenic effects of length n ($\boldsymbol{Z}$ is the associated design matrix) and is assumed to be normally distributed, $\boldsymbol{u} \sim \boldsymbol{N}(0, \sigma_{\boldsymbol{u}}^2 \boldsymbol{A})$ and $\boldsymbol{e}$ is the residual error also assumed to be normally distributed, $\boldsymbol{e} \sim \boldsymbol{N}(0, \mathbf{I}\sigma_{\boldsymbol{e}}^2)$. The polygenic effect was included to remove the effect of population structure to enable more accurate estimation of the SNP effects. Its inclusion has been shown to produce slightly better accuracies of prediction while reducing the bias of the variance components (Calus & Veerkamp, 2007).

(iii) *SSVS*

The key feature of SSVS compared with Bayes A or B (Meuwissen *et al.*, 2001) is the introduction of a latent or indicator variable, γ, into the hierarchical model. This enables the extraction of information relevant to variable selection. The latent variable can take either 1 or 0, representing whether the SNP is included as a significant effect in the model or not. As such, the prior distribution for each SNP effect is a normal mixture conditional on the corresponding γ and the variance that is sampled from an inverse scaled chi-square distribution:

$$\beta_i | \gamma_i, \sigma_i^2 \sim (1-\gamma_i)\boldsymbol{N}(0, \sigma_i^2/100) + \gamma_i \boldsymbol{N}(0, \sigma_i^2),$$
$$\sigma_i^2 \sim \chi^{-2}(\boldsymbol{r}, \boldsymbol{S}).$$

At the SNP effect level, this hierarchical prior distribution specification means the SNP effects are sampled from a mixture of two-student t distributions. The values of $\boldsymbol{r}$ and $\boldsymbol{S}$ were calculated as in Meuwissen *et al.* (2001). The prior distribution of the indicator variable is chosen to reflect the belief of whether an SNP is linked to a QTL. The probability of an SNP being sampled from the smaller or larger distribution is

$$1 - p(\gamma_i = 0) = p(\gamma_i = 1) = p_i.$$

Subsequently, the prior distribution for the indicator variable is a Bernoulli distribution:

$\gamma_i \sim \text{bernoulli } (\boldsymbol{p_i})$.

The prior probability $\boldsymbol{p_i}$ is chosen to reflect the information available on how many QTLs affect the trait of interest. It can be quantified as the number of SNPs expected to be linked to a QTL divided by the total number of SNPs. In genome-wide association studies or genomic selection applications, the expected proportion of QTLs can be reasonably estimated based on knowledge about the trait of interest and previous QTL studies results.

The posterior distribution of the indicator variable can be sampled directly using

$$p(\gamma_i = 1 | \beta_j, \sigma_i^2, \gamma_{-i}, \boldsymbol{u}, \boldsymbol{y}) \sim \text{bernoulli} \left(\frac{p(\beta_j | \gamma_{-i}, \gamma_i = 1)\boldsymbol{p_i}}{p(\beta_j | \gamma_{-i}, \gamma_i = 1)\boldsymbol{p_i} + p(\beta_j | \gamma_{-i}, \gamma_i = 0)(1 - \boldsymbol{p_i})} \right),$$

where γ_{-i} is all terms of γ except γ_i.

The frequency that each SNP appears in the model is shown by the posterior distribution of the indicator variable. SNPs that are included in the model frequently have a high posterior probability and will most likely be linked to a QTL.

(iv) *Additional methods*

Bayes A, Bayes B and BLUP were also run on the data. Bayes A and Bayes B were as specified in Meuwissen *et al.* (2001) with the addition of a polygenic effect. A Bayesian BLUP method was also implemented. It is identical to the specification of Bayes A with the exception that all SNPs had a constant equal variance that was sampled once each iteration from an inverse-scaled chi-square distribution.

In order to have Bayes B results for comparison with Bayes SSVS, we also used a modified version of Bayes B approach. The modified version consisted of running Bayes B cycles with the Metropolis Hastings (MH) algorithm every 100 iterations of Bayes A. (Note the Jacobian in the acceptance ratio of the reversible jump algorithm was equal to one thus identical to the MH algorithm). If an SNP effect was found to be zero during these MH iterations then it was set to zero during the subsequent Bayes A cycles. This effectively maintained the same assumptions as Bayes B, while significantly reducing the time required to reach convergence.

(v) *Breeding values*

MEBVs for bulls in the validation data set were calculated as the sum of the mean, the effects of the SNP genotypes it carried and the polygenic effect, $\text{MEBV} = \hat{\mu} + \boldsymbol{X}\hat{\beta} + \hat{u}$. The accuracy of the methods were evaluated on the correlation, the mean square error (MSE) and the regression coefficient of the ABV (assumed to be the true breeding value) on the predicted MEBV. Genomic selection aims to produce breeding values as close as possible to the true breeding value. The ABV was used for comparison as it is a most accurate predictor of the true breeding value and it is regressed according to the amount of information available.

Table 1. *Computational time for genomic selection methods*

Method	Computational time[a]
Bayes BLUP	6
Bayes A	6
Bayes B	~2440[b]
Bayes B Modified	240
Bayes SSVS	6

[a] Processor clock hours.

[b] Estimated time to convergence.

3. Results and discussion

(i) *Time to convergence*

All methods were run for 10 000 iterations to ensure convergence. This number of iterations was shown to be sufficient for convergence with formal diagnostic methods provided in the package *R*, *coda* (Plummer *et al.*, 2007). The use of the SSVS method is analogous to Bayes B in the assumption that the majority of the SNP effects are thought to be very small and insignificant. However, as illustrated in Table 1, the fixed dimensions of the model used in SSVS allow the use of the Gibbs Sampler that is significantly computationally less demanding and consequently quicker than the reversible jump MCMC algorithm or the MH algorithm used in traditional Bayes B. Given the very high computational demand of Bayes B, it was not possible to run this algorithm to convergence. The time to convergence was extrapolated from running Bayes B for 1000 iterations. The Bayes A and Bayes BLUP methods reached convergence in comparable times to Bayes SSVS.

(ii) *Comparison of Bayes B and Bayes SSVS results*

The correlations between the ABVs and the MEBV predicted for the animals in the validation set by the modified Bayes B and Bayes SSVS for fertility and protein kg traits are shown in Table 2. This shows that the two methods produce almost identical correlations with the ABVs as expected. The MEBV for the

Table 2. *Correlation between predicted MEBV and ABV for proven bulls (years 2005, 2006, 2007 and overall) for the modified Bayes B and Bayes SSVS*

	Bayes B (modified)	Bayes SSVS
Protein kg		
2005	0·620	0·627
2006	0·638	0·646
2007	0·502	0·490
Protein kg		
Overall	0·575	0·583
Fertility		
2005	0·576	0·577
2006	0·430	0·429
2007	0·628	0·628
Fertility		
Overall	0·540	0·540

two methods are 99·9 and 98·0% correlated for protein and fertility, respectively. This equivalence in results demonstrates that the Bayes SSVS method does maintain the SNP effect assumptions of the original Bayes B and produce near to identical results. The slightly lower result for fertility is probably due to the non-normality of the trait making it harder to estimate and by the modification of the original Bayes B. The modified Bayes B produced not significantly different but slightly larger MSEs and regression coefficients (results not shown). This is most likely due to the modification to reduce the computational time to convergence. The time taken for the modified version of Bayes B was still 40-fold larger than for the Bayes SSVS that produced identical accuracies (see Table 1).

Table 3. *MSE, correlation and regression coefficient between predicted MEBV and ABV in the validation data set*

Method	Measure	Bayes SSVS[a]	Bayes A[a]	Bayes BLUP[a]
Protein kg	$r_{\mathrm{EBV,ABV}}$	0·583	0·567	0·602
	log(MSE)	4·03	4·06	3·96
	$b_{\mathrm{EBV,ABV}}$	0·987	0·997	1·055
Fat kg	$r_{\mathrm{EBV,ABV}}$	0·563	0·532	0·563
	log(MSE)	5·18	5·22	5·23
	$b_{\mathrm{EBV,ABV}}$	0·9	0·856	0·988
Protein %	$r_{\mathrm{EBV,ABV}}$	0·668	0·641	0·655
	log(MSE)	−4·94	−4·88	−4·84
	$b_{\mathrm{EBV,ABV}}$	0·972	0·995	0·887
Fat %	$r_{\mathrm{EBV,ABV}}$	0·740	0·716	0·646
	log(MSE)	−3·07	−3·24	−3·32
	$b_{\mathrm{EBV,ABV}}$	0·874	0·864	0·925
Fertility	$r_{\mathrm{EBV,ABV}}$	0·540	0·539	0·538
	log(MSE)	1·51	1·51	1·52
	$b_{\mathrm{EBV,ABV}}$	0·933	0·942	0·905

[a] Average accuracies reported over validation sets from years 2005, 2006 and 2007.

$r_{\mathrm{EBV,ABV}}$, correlation coefficient between the ABV and the predicted MEBV.

log(MSE) is the logarithm of the MSE between the ABV and the predicted MEBV.

$b_{\mathrm{EBV,ABV}}$, regression coefficient of the ABV on predicted MEBV.

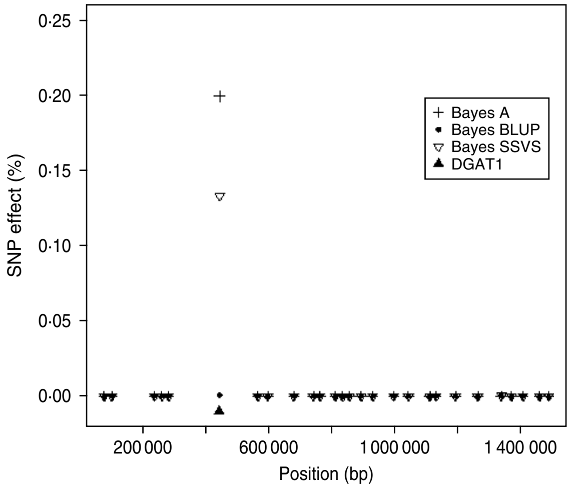

Fig. 1. SNP effects (%) for fat percentage from Bayes A, Bayes BLUP and Bayes C found on the centromeric end of chromosome 14.

(iii) *Comparison of BLUP, Bayes A, Bayes SSVS results*

The logarithm of the MSE, regression and correlation coefficients for the predicted MEBV and ABV for the traits fertility protein kg, fat kg, protein percentage and fat percentage are shown in Table 3. The values shown are the average values for the proven bulls in the years 2005, 2006 and 2007 from the validation data set. BLUP has the highest overall correlation and the lowest MSE between the three methods for protein kg. For the traits, fat kg and protein percentage, Bayes SSVS produces the highest correlations and has the lowest bias; however, there are no significant differences between methods. However, there are significant differences between the methods for fat percentage. These difference in the method accuracies across traits or the apparent 'trait by method' interactions can be explained by the distribution of QTLs for the different traits. For example, protein kg has no known genes of large effect and thus BLUP, which uses equal variances across all SNPs, can be used successfully to accurately predict breeding values. In contrast, fat percentage has a known mutation, DGAT1, that is common and acts additively and is known to be responsible for explaining a large percentage of genetic variation for the trait (Grisart *et al.*, 2002). The individual SNP variances that Bayes A and Bayes SSVS uses, allows effects of a large size not to be penalized (shrunk) as severely as in BLUP. This is clearly shown in Fig. 1, where the percentage each SNP contributes to the total SNP effects are plotted for the three methods for the centromeric end of the bovine chromosome 14. Bayes A and Bayes C have an SNP with an effect significantly greater than zero, while the Bayes BLUP effects for SNP near DGAT1 and surrounding the mutation are close to zero. Bayes SSVS does perform slightly better than Bayes A for fat percentage. The advantage of the Bayes SSVS over Bayes A may be the prior structure consisting of two distributions: a distribution of larger significant effects and a smaller distribution close to zero. This allows the SNP with larger effects to have values in their posterior sampled from the larger distribution, while those SNPs without significance have their effects sampled from the smaller posterior distribution of values very close to zero. Traits with large effects will be more accurately predicted using SSVS than Bayes A as the prior structure allows more variance to be attributed to the larger effects.

4. Conclusion

Bayesian SSVS produced more accurate MEBV for most of the dairy traits in our data set than other methods. The comparison with a modified version of Bayes B showed that it is equivalent and produces the same results with dramatically less computational time required. For traits with a mutation of known large effect such as fat percentage, Bayes SSVS gave significantly higher accuracy of MEBV than the BLUP method as expected given that its prior is closer to the real distribution of effects than that of BLUP. The use of an indicator variable in Bayes SSVS would also allow the premeditated inclusion of SNPs in a model that are known to be linked to QTL of biological importance. Instead of using a single value to set the prior probability for all SNPs a vector of probabilities could be used as prior probabilities to allow more prior information to be included should it be available. Overall, this study has shown that the Bayes SSVS method provides reduced computational time and accurate results when using real dairy data to predict genomic breeding values and provides a viable alternative to other Bayesian methods for genomic selection.

The authors would like to thank Mehar Khatkar and Herman Raadsma for provision of part of the data used in this paper.

References

Brown, P. J., Vannucci, M. & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **60**, 627–641.

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W. & Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**, 553–561.

Calus, M. P. L. & Veerkamp, R. F. (2007). Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* **124**, 362–368.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E. I. & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C. *et al.* (2002). Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**, 222–231.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009). Invited Review: Genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science* **92**, 433–443.

Meuwissen, T. H. E. & Goddard, M. E. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution* **36**, 261–279.

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

Plummer, M., Best, N., Cowles, K. & Vines, K. (2007). coda: Output analysis and diagnostics for MCMC. R package version 0.13-1.

Swartz, M. D., Kimmel, M., Mueller, P. & Amos, C. I. (2006). Stochastic search gene suggestion: a Bayesian hierarchical model for gene mapping. *Biometrics* **62**, 495–503.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B* (*Methodological*) **58**, 267–288.

Yi, N. J., George, V. & Allison, D. B. (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**, 1129–1138.

BMC Proceedings

PROCEEDINGS Open Access

Sensitivity of genomic selection to using different prior distributions

Klara L Verbyla[1,2,3,4*], Philip J Bowman[2], Ben J Hayes[2], Michael E Goddard[2,3,4]

From 13th European workshop on QTL mapping and marker assisted selection
Wageningen, The Netherlands. 20-21 April 2009

Abstract

Genomic selection describes a selection strategy based on genomic estimated breeding values (GEBV) predicted from dense genetic markers such as single nucleotide polymorphism (SNP) data. Different Bayesian models have been suggested to derive the prediction equation, with the main difference centred around the specification of the prior distributions.

Methods: The simulated dataset of the 13th QTL-MAS workshop was analysed using four Bayesian approaches to predict GEBV for animals without phenotypic information. Different prior distributions were assumed to assess their affect on the accuracy of the predicted GEBV.

Conclusion: All methods produced GEBV that were highly correlated with the true breeding values. The models appear relatively insensitive to the choice of prior distributions for QTL-MAS data set and this is consistent with uniformity of performance of different methods found in real data.

Background

Genomic selection describes a technique for evaluating an animal's breeding value by simultaneously evaluating and summing marker effects across the genome. It uses panels of SNPs covering the whole genome so that ideally all QTL are in linkage disequilibrium with at least one marker, thereby maximizing the proportion of genetic variance explained by the SNPs.

Meuwissen et al (2001) [1] presented three models to produce GEBV. The first invoked the infinitesimal model assumption such that all SNPs had effects derived from the same normal distribution. The other approaches used a Bayesian framework to apply hierarchical models with different prior distributions assuming unequal variances across the SNP, resulting in a *t* distribution for prior distribution for the QTL effects. The specification of the prior distributions of the QTL effects has been reported to be important to the accurate prediction of breeding values and when mapping multiple QTL across the entire genome [2].

The aim of this study was to assess the effect that different prior distributions and subsequently the models using these priors, had on the accuracy of estimated GEBV using the 13th QTL-MAS simulated data set where we had no prior knowledge of the trait's distribution of QTL effects.

Methods

Model

At each loci (total number of locus, p) there are three possible combinations of two alleles (e.g. A or B), the homozygote of one allele (AA), the heterozygote (AB) and the homozygote of the other allele (BB). These are then quantitatively represented by 0, 1 and 2 respectively. Subsequently, phenotypic records at each time point were modelled as:

$$y = \mu 1_{\mathrm{n}} + \sum_{j=1}^{q} X_j \beta_j + Zu = e$$

where y is the vector of phenotypes of the trait being analysed for all n individuals, μ is the mean, $\boldsymbol{1_n}$ is a vector of ones of length n, $\mathbf{X_j}$ is a vector of indicator

[*] Correspondence: klara.verbyla@dpi.vic.gov.au
[1]Animal Breeding and Genomics Centre, ASG Wageningen UR, PO Box 65, 8200 AB Lelystad, The Netherlands

variables representing the genotypes of the j^{th} marker for all individuals (x_{ij}=0,1,2), β_j is the size of the QTL effect associated with marker *j*, **u** is the vector of random polygenic effects of length *n* (*Z* is the associated design matrix) and is assumed to be normally distributed, $\boldsymbol{u} \sim \boldsymbol{N}$ (0, $\sigma_u^2 \boldsymbol{A}$) where ***A*** is the pedigree derived additive genetic relationship matrix and **e** is the residual error also assumed to be normally distributed, $\boldsymbol{e} \sim \boldsymbol{N}$(0, $\boldsymbol{I}\sigma_e^2$) where ***I*** is the *nxn* identity matrix. The prior distributions for the variances of the random polygenic effects and the residual were uninformative flat priors of the form X^{-2}(- 2,0). The GEBV at each time point were calculated as $\mathrm{GEBV} = \hat{\mu} + \mathbf{X}\hat{\beta} + \hat{u}$.

Prior distributions for SNP effects and algorithms

Four differing sets of prior distributions were assessed and the specifications are shown in Table 1. The Bayes BLUP model assumed the same variance for the normal distribution from which the SNP effects were assumed to be derived (maintaining the infinitesimal assumptions for traditional BLUP). The variance of the normal distribution was sampled once every MCMC iteration using a Gibbs Sampler. The SNP effects were subsequently sampled from this normal distribution. The model termed Bayes A [1] assumes that the SNP effects come from a *t*-distribution. This is because an efficient Gibbs sampling scheme to sample the SNP effects from their posterior distributions is to a sample SNP specific variance from an inverse chi-square distribution, then use this variance to define the normal distribution from which the SNP effect is sampled [1]. The values for the inverse scaled chi square hyper parameters(*r* and *S*) were calculated as in Meuwissen et al (2001) [1].

The other two models assumed mixture distributions for the SNP effects reflecting the assumption that there is a large number of SNPs with zero or near zero effects and a second smaller set of SNPs with larger significant effects. A Bayes A/B "hybrid" method was used. This approximation to Bayes B [1] was used to keep computational and time demands reasonable. In this algorithm, after every k Bayes A iterations, Bayes B via the reverse jump algorithm is employed. The Reverse Jump algorithm [3] is run multiple times per SNP and then any SNP with a final state of zero in the current Bayes B iterations is set to zero for the subsequent k iterations of the Bayes A. This maintains the correct transitions between models of differing dimensionality. The prior distributions are identical to that of the original Bayes B using a mixture prior distribution for the SNP variance allowing a proportion, 1-π, to be set to zero. The other proportion π is sampled from the same mixture distribution as Bayes A. See Meuwissen et al (2001) for more details of priors and conditional distributions used.

Table 1 Prior Distribution Specifications

Method	Prior Distribution
Bayes BLUP	$\beta_i \mid \sigma^2 \quad N\left(0,\sigma^2\right)$ $\sigma^2 \quad \chi^{-2}(r,s)$
Bayes A	$\beta_i \mid \sigma_i^2 \quad N\left(0,\sigma_i^2\right)$ $\sigma_i^2 \quad \chi^{-2}(r,s)$
Bayes A/B (Hybrid)	$\beta_i \mid \sigma_i^2 \quad N\left(0,\sigma_i^2\right)$ $\sigma_i^2 = 0$ with probability 1- π $\sigma_i^2 \quad \chi^{-2}(r,s)$ with probability π
Bayes C	$\beta_i \mid \gamma_i,\sigma_i^2 \quad (1-\gamma_i)N\left(0,\sigma_i^2/100\right)+\gamma_i N(0,\sigma_i^2)$ $\sigma_i^2 \quad \chi^{-2}(r,s)$ γ_i ~*bernoulli*(π) 1 - p(γ_i = 0) = *p*(γ_i = 1) = π

β_i is the effect for the i^th^ SNP and γ_i is the indicator variable for the i^th^ SNP.

A faster alternative to both the Bayes A/B hybrid and Bayes B is to use Stochastic Search Variable Selection (SSVS) [4] (Bayes C [5,6]). This avoids the problem of the changing dimensionally of the models by providing a technique to maintain constant dimensionality across all models while still allowing the SNP in the predictive set to change. Instead of removing all non-significant parameters, their posterior distributions are limited to values close to zero. The major advantage of this method is that it can be implemented using the Gibbs sampler instead of the more computationally demanding algorithms such as the reverse jump algorithm. The indicator variable (γ_i) determines whether the i^th^ SNP effect is sampled from the larger distribution (i.e. significant effect) or from the small distribution with near zero effects (see Table 1). The prior values of π (the proportion sampled from the non-zero distribution or the larger distribution respectively) for both Bayes A\B and Bayes C was set to 0.05, reflecting the fact that with 435 SNP, it appeared reasonable to expect at least 21 SNP would be associated with a QTL.

The algorithms associated with each model were run for 30,000 iterations with the first 10,000 discarded as burn-in.

Results and Discussion

Prediction of breeding values at time point 600

The problem of how to model the time series data and estimate GEBV at time point 600 was explored. However, there was little information available to estimate any inflection points or asymptotic values. The GEBV estimated at time points 265, 397 and 530 were found to have a linear relationship (eg. appeared to form the linear part of the growth curve). Consequently, as there was no other information available after time point 530 to predict asymptotes etc., the GEBV at time point 600 were estimated by fitting a linear regression through the breeding values at the three linear time points (265, 397 and 530).

Breeding values

The correlations between the GEBV (t=600) predicted by the alternative methods for the validation population containing the 50 full sib families without phenotypes are shown in Table 2. Correlations were extremely high between all methods other than BLUP and consequently GEBV appeared relatively insensitive to the model used when assuming unequal variances. Correlations, mean square errors, the accuracy of predicting the first 100 animals (rank) and the bias (regression coefficient) between the predicted and true breeding values are shown in Table 3. While there is no significant difference between the methods, Bayes A/B performed the best of the methods producing the lowest MSE, highest correlation and rank but was slightly more biased than Bayes C and Bayes BLUP, but not significantly. Interestingly while Bayes C has very similar hierarchical prior distributions it does worse than Bayes A/B. Further optimisation of the prior probability of π for Bayes C increased the accuracy (results not shown). The optimal value for π was 0.3 (values tested were 0.05, 0.1, 0.2, 0.3, 0.4, 0.6 and 1). This produced results more similar to the results seen for Bayes A\B. This does highlight the importance of the correct assumption of the proportion assigned to the smaller and larger distributions in a mixture model. This difference between these two methods may demonstrate that Bayes C is more sensitive to an incorrect assumption about this proportion.

The inclusion of the polygenic effect in the model (not simulated in the data) only slightly reduced the accuracy of prediction (.01) but not significantly (results not shown). It was included in the model as its inclusion has been shown to produce slightly better accuracies of prediction while reducing the bias of the variance components[7].

Table 2 Correlations Between Estimated GEBV for unphenotyped animals at t=600

	Bayes C	Bayes A/B	Bayes BLUP
Bayes A	0.999	0.991	0.860
Bayes C	1	0.993	0.863
Bayes A/B		1	0.893

Table 3 Comparison of True and Estimated GEBV

Method	*Correlation*	*MSE*	*Rank*	*Regression*
Bayes.BLUP	0.885	5.479	0.691	0.979
BayesA	0.857	7.092	0.696	1.162
BayesA/B	0.889	5.435	0.73	1.081
BayesC	0.861	6.561	0.71	1.024

Correlation coefficient between the true and predicted GEBV, Mean Square Error (MSE), Rank (Accuracy of the predicting the best 100 animals) and the Regression Coefficient of the true breeding value on the estimated GEBV.

Bayes BLUP produced a significantly different set of GEBV. This is evident by the much lower correlations with the other methods and difference in regression coefficients between BLUP and the other methods. Despite these differences Bayes BLUP produces good accuracy and a low MSE (Table 3). Hayes et al (2009) [8] reports that New Zealand, Australian, the Netherlands and United States studies all found that BLUP gave lower accuracy of GEBV than Bayesian Methods for traits where there is a single QTL that explains a large proportion of the genetic variance e.g. DGAT1 for Fat Percentage. In the current dataset a finite number of QTL were simulated where the largest amount of genetic variance explained by a single QTL was 10.5%. Despite this, Bayes BLUP is still able to produce very accurate GEBV compared to the other methods. One reason this occurs may be that a number of SNPs are required to pick up the effect of a single QTL, resulting in large numbers of SNPs with small effects, which matches the prior distribution of BLUP. However if the percentage of genetic variance explained by a single QTL was to be larger, Bayes BLUP could be expected to produce worse results. Thus this caveat to using Bayes BLUP should be considered when using this method.

Conclusion

All methods produced GEBV that were highly correlated (greater than 0.85) with the true breeding values despite diverse assumptions and prior distributions. This indicates that the hierarchical model is relatively insensitive to the choice of prior distributions for this data set. Thus all models perform well and this is consistent with the general uniformity of performance found across methods in real data. [8]. Despite the general equality in the performance of the different methods, it is still recommended that any information about a trait's QTL effect distribution and phenotypic data should be used to determine the choice of model, prior distributions and setting of the hyper parameters. This will maximise the likelihood of calculating the most accurate GEBV possible.

Acknowledgements
KV was funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission. This Publication represents the views of the Authors, not the European Commission, and the Commission is not liable for any use that may be made of the information.
This article has been published as part of BMC Proceedings Volume 4 Supplement 1, 2009: Proceedings of 13th European workshop on QTL mapping and marker assisted selection.
The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/4?issue=S1.

Author details
[1]Animal Breeding and Genomics Centre, ASG Wageningen UR, PO Box 65, 8200 AB Lelystad, The Netherlands. [2]Biosciences Research Division,

Department of Primary Industries Victoria, 1 Park Drive, Bundoora 3083, Australia. [3]Melbourne School of Land and Environment, The University of Melbourne, Parkville 3010, Australia. [4]The Cooperative Research Centre for Beef Genetic Technologies, University of New England, Armidale, NSW 2351, Australia.

Authors' contributions

KV carried out the analyses and drafted the manuscript. PB developed the Bayes A and Bayes BLUP software. KV created the Bayes C and Hybrid software using the Bayes A software. BH and MG read and suggested improvements to the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 31 March 2010

References

1. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157(4)**:1819-1829.
2. Yi NJ: **A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci.** *Genetics* 2004, **167(2)**:967-975.
3. Green PJ: **Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.** *Biometrika* 1995, **82(4)**:711-732.
4. George EI, McCulloch RE: **Variable Selection Via Gibbs Sampling.** *Journal of the American Statistical Association* 1993, **88(423)**:881-889.
5. Meuwissen THE, Goddard ME: **Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data.** *Genetics Selection Evolution* 2004, **36**:261-279.
6. Verbyla KL, Hayes BJ, Bowman PJ, Goddard ME: **Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle.** *Genetics Research* 2009, **91(05)**:307-311.
7. Calus MPL, Veerkamp RF: **Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM.** *Journal of Animal Breeding and Genetics* 2007, **124(6)**:362-368.
8. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: **Invited review: Genomic selection in dairy cattle: Progress and challenges.** *J. Dairy Sci* 2009, **92(2)**:433-443.

doi:10.1186/1753-6561-4-S1-S5
Cite this article as: Verbyla *et al.*: Sensitivity of genomic selection to using different prior distributions. *BMC Proceedings* 2010 **4**(Suppl 1):S5.

J. Dairy Sci. 93:2757–2764
doi:10.3168/jds.2009-2928

Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information

K. L. Verbyla,*†‡§[1] M. P. L. Calus,* H. A. Mulder,* Y. de Haas,* and R. F. Veerkamp*
*Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad 8200AB, the Netherlands
†Melbourne School of Land and Environment, The University of Melbourne, Parkville, Victoria 3001, Australia
‡Department of Primary Industries, 1 Park Drive, Bundoora, Victoria 3083, Australia
§The Cooperative Research Centre for Beef Genetic Technologies, University of New England, Armidale, NSW 2351, Australia

ABSTRACT

The objective of this study was to investigate the genetic basis of energy balance (EB) and the potential use of genomic selection to enable EB to be incorporated into selection programs. Energy balance provides an essential link between production and nonproduction traits because both depend on a common source of energy. A small number (527) of Dutch Holstein-Friesian heifers with phenotypes for EB were genotyped. Direct genomic values were predicted for these heifers using a model that included the genotypic information. A polygenic model was also applied to predict estimated breeding values using only pedigree information. A 10-fold cross-validation approach was employed to assess the accuracies of the 2 sets of predicted breeding values by correlating them with phenotypes. Because of the small number of phenotypes, accuracies were relatively low (0.29 for the direct genomic values and 0.21 for the estimated breeding values), where the maximum possible accuracy was the square root of heritability (0.57). Despite this, the genomic model produced breeding values with reliability double that of the breeding values produced by the polygenic model. To increase the accuracy of the genomic breeding values and make it possible to select for EB, measurement and recording of EB would need to improve. The study suggests that it may be possible to select for minimally recorded traits; for instance, those measured on experimental farms, using genomic selection. Overall, the study demonstrated that genomic selection could be used to select for EB, confirming its genetic background.

Key words: energy balance, genomic selection, dairy cow, genetic variation

INTRODUCTION

Due to declining calving performance and conception rates at first service (Royal et al., 2000) many countries have introduced measures of fertility into national selection indices to address declining fertility rates in dairy cattle (Miglior et al., 2005). One explanation for these declining rates is the difference between energy intake and energy usage that occurs during early lactation. This difference is defined as energy balance (**EB**). Energy balance provides an essential link between production and nonproduction traits because both depend on a common source of energy. This energy must be partitioned efficiently to maintain production levels as well as the animal's ability to remain healthy and fertile. Severe negative energy balance (**NEB**) during early lactation has been cited as an underlying cause of the negative relationship of health and fertility with production (Butler and Smith, 1989; Jorritsma et al., 2003; Pryce et al., 2004).

Recently, the major focus had been on trying to overcome the NEB problem by modifying the diet during the dry period (Dewhurst et al., 2000; Agenäs et al., 2003; McNamara et al., 2003; Garnsworthy et al., 2008a,b). Other suggested approaches to overcome NEB include varying the length of the dry period (Watters et al., 2009) and the frequency of milking (McNamara et al., 2008). However, estimates of genetic parameters suggest that EB is not only a consequence of a poor match between nutrition and production, but is also genetically induced (Veerkamp, 1998; Veerkamp et al., 2003; Coffey et al., 2004; Friggens et al., 2007). Veerkamp (1998) reviewed the results of different studies that reported genetic correlations for a variety of energy measures and milk yield, with values ranging from −0.05 to −0.91 and heritabilities for energy traits that ranged from 0.19 to 0.69. Coffey et al. (2004) demonstrated that distinct genetic lines responded differently to a range of diets and differed in the time taken to return to positive EB. Similarly, Friggens et al. (2007) concluded that variability among animals on a stable nutritional diet could not be accounted for by

Received November 20, 2009.
Accepted February 5, 2010.
[1]Corresponding author: klara.verbyla@dpi.vic.gov.au

environmental factors and indicated a genetic basis for EB. Thus, an alternative to management approaches may be to select animals that are genetically predisposed to maintain a better EB.

Accounting for EB in selection programs is complicated, because measuring feed intake in progeny testing schemes is not practical. Currently, much attention has been placed on the implementation of genomic selection. Genomic selection uses genomic information to predict and select animals based on their direct genomic values (**DGV**), predicted directly from SNP information, or their genomically enhanced breeding values, which are calculated by blending the DGV with conventional proofs. Genomic prediction simultaneously estimates the marker effects and creates an equation to predict DGV for genotyped selection candidates, including (young) animals that do not have phenotypic records. The recent implementation of genomic selection has been shown to increase both selection accuracy and genetic gain over traditional selection methods (Hayes et al., 2009).

In this study, we examined whether genomic prediction could be used to estimate DGV for EB using a small Dutch experimental farm data set. Our objective was to demonstrate the genetic basis of EB and the potential use of genomic selection to facilitate inclusion of EB in selection programs.

MATERIALS AND METHODS

Data

Data on 613 Holstein-Friesian heifers born between 1990 and 1997 were collected during the first 15 wk of lactation; 450 cows participated in the breeding program of CRV (Arnhem, the Netherlands) and 163 cows originated from the experimental farm (t'Gen, the Netherlands). All animals were housed together on a single farm under the same environmental and management influences. All cows were fed ad libitum. Live weight, feed intake, and milk yield were measured on 565 of the animals. Milk samples were taken on a fixed day of the week for measurement of fat, protein, and lactose yields. Feed intake was recorded daily using automated feed intake units. Live weight was recorded once a week. Energy balance (MJ/d) was calculated using the method described in Veerkamp et al. (2000) as the difference between energy intake and the calculated energy requirements for milk, fat, and protein yields, and maintenance costs as a function of live weight. Energy balance values across wk 2 to 15 were averaged, where possible, to give an overall EB phenotype. Comprehensive details on the data used can be found in Veerkamp et al. (2000). Raw EB phenotypes were preadjusted for year-season of calving and age at calving (linear, quadratic) using ASReml (Gilmour et al., 2006), because their inclusion was not feasible in the final model because of software limitations. The residuals from this analysis were used as the EB phenotypes for the prediction of breeding values.

In total, 588 of the 613 heifers had known pedigrees and these were genotyped using the Illumina 50K SNP panel (54,001 SNP in total; Illumina, San Diego, CA). The quality control criteria for selecting the final set of SNP were a call rate of >90%, a GenCall score >0.2, and a GenTrain score >0.55 (Illumina descriptive statistics relating to genotype quality), a minor allele frequency of >2.5%, and a lack of deviation from Hardy-Weinberg equilibrium, χ^2 <600 (Wiggans et al., 2009). Animals with greater than 5% missing SNP genotypes were removed. Non-Mendelian error checks identified genotypes of daughters that were inconsistent with their dams. A further, more comprehensive pedigree check was performed by comparing the coefficients of the additive genetic relationship matrix and the genomic relationship matrix (G matrix) calculated via the first method described in VanRaden (2008). This enabled inconsistencies between recorded half and full siblings to be examined. Animals with many inconsistencies between the pedigree and G matrix were removed. After all editing steps, 43,011 SNP and 548 animals were retained. Of these 548 animals, 527 had phenotypes for EB.

Statistical Analysis

Models. Two models using Gibbs sampling were applied to estimate additive breeding values. One model included the available SNP information. This model used stochastic search variable selection (SSVS; George and McCulloch, 1993), which introduces an indicator variable I_j that determines whether SNP j has a large significant effect or whether the effect is insignificant and is therefore scaled back toward zero. The indicator variable for each locus j has a Bernoulli prior distribution:

$$I_j \sim \mathit{Bernoulli}\ (p).$$

The prior probability p is chosen to reflect the information available on how many QTL affect the trait of interest. It can be quantified as the number of SNP expected to be linked to a QTL divided by the total number of SNP. For a complex trait such as EB, it was assumed that about 1% of the SNP were linked to a QTL ($p = 0.01$).

The SNP model can be expressed as follows:

$$\mathbf{y} = \mathbf{1}_n \mu + \sum_{j=1}^{m}\left(\mathbf{X}_j\left(\mathbf{q}_j v_j\right)\right) + \mathbf{Zu} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypes of the trait being analyzed for all n individuals, $\mathbf{1}_n$ is a vector of ones of length n, μ is the mean, m is the number of SNP markers, $\mathbf{X}_j$ is the ($n \times k$) design matrix containing the information on the possible k alleles at the jth marker for all individuals (where x_{jk} = 0, 1, 2, having 0, 1, or 2 copies of the kth allele, respectively), $\mathbf{q}_j$ is the vector ($k \times 1$) containing the effects of all k possible alleles at locus j where q_{jk} are drawn from a standard normal distribution N(0,1), v_j is the standard deviation of the allelic effects at locus j and is dependent on whether the locus effect is considered significant using the indicator variable, $\mathbf{u}$ is the vector of random additive polygenic effects of length n ($\mathbf{Z}$ is the associated design matrix) and is assumed to be normally distributed, $\mathbf{u} \sim N\left(0, \sigma_u^2 \mathbf{A}\right)$, where $\mathbf{A}$ is the pedigree-derived additive genetic relationship matrix, and $\mathbf{e}$ is the residual error also assumed to be normally distributed, $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$, where $\mathbf{I}$ is the $n \times n$ identity matrix. Note that the allele substitution effect of a locus j can be calculated from the estimated effects as $a_j = (q_{j1} - q_{j2})v_j$, where q_{j1} (q_{j2}) is the effect of allele 1 (2) at locus j. For the full specification of the priors used and an alternative formulation of the model, see Calus et al. (2008) and Meuwissen and Goddard (2004). The DGV were calculated as the sum of estimated SNP effects and the polygenic effect:

$$DGV = \sum_{j=1}^{p}\left(\mathbf{X}_{ij}\left(\hat{q}_j v_j\right)\right) + \hat{u}_i.$$

The second model used was a simple additive polygenic model: $\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Zu} + \mathbf{e}$, where the EBV calculated by this model were the estimated polygenic effects for each animal $\left(\mathrm{EBV} = \hat{u}_i\right)$. Both models were run for 10,000 iterations to ensure convergence, with the first 1,000 iterations used as burn in.

Validation. Because of the small size of the data set, a 10-fold cross validation approach was carried out to assess the accuracy of predicted breeding values. The data set was randomly partitioned into 10 subsets each containing 10% of the data. Each subset was retained once as the validation data set and the remaining 9 became the reference sets. Results from the reference sets were then used to predict breeding values of animals in the validation set. Accordingly, each animal appeared only once in a validation set and had only one predicted DGV.

The DGV and EBV were assessed using accuracy, $\mathbf{r}_{y\hat{g}}$, defined as the Pearson correlation of the predicted breeding values (DGV or EBV) ($\hat{g}$) and the phenotypes (y). The maximum achievable accuracy due to the correlation being between phenotypes and predicted breeding values was equal to the square root of the heritability of the phenotypes. The observed heritability for EB was estimated by fitting a model with year-season and age at calving (linear and quadratic regression) as the fixed effects and a random animal effect (a). The random animal effect was assumed normally distributed, $a \sim N\left(0, \sigma_a^2 \mathbf{G}\right)$, where σ_a^2 was the additive genetic variance and $\mathbf{G}$ was the genomic relationship matrix calculated via the first method described in VanRaden (2008). Deriving the heritability this way has been shown to produce estimates much closer to the true value than using the pedigree-based relationship matrix (Hayes and Goddard, 2008).

Because no daughter yield deviations (**DYD**) or reliable breeding values were available, the predicted breeding values (DGV and EBV) were compared with phenotypes. Most studies estimating accuracies of DGV use DYD or reliable EBV predicted for proven bulls and consequently report accuracies of selection $\left(\mathbf{r}_{g\hat{g}}\right)$ and reliabilities $\left(\mathbf{r}_{g\hat{g}}^2\right)$ that compare DGV and the closest estimate of the true breeding values (g). Thus, for these studies the accuracy of selection was calculated (Daetwyler et al., 2008; Goddard, 2009) as

$$\mathbf{r}_{g\hat{g}} = \sqrt{\frac{\lambda h^2}{\lambda h^2 + 1}},$$

$$\text{and } \lambda = \frac{n_p}{n_G}, \qquad [1]$$

where h^2 is the observed heritability, n_p is the number of phenotypic records, and n_G is the number of effective QTL or chromosome segments. This function can be used when the accuracy is calculated using the correlation between the predicted DGV and phenotypes $\left(\mathbf{r}_{y\hat{g}}\right)$. Falconer and Mackay (1996) state that $\mathbf{r}_{g\hat{g}} = \sigma_{\hat{g}} / \sigma_g$. The accuracy between DGV and phenotypes can be similarly expressed as $\mathbf{r}_{y\hat{g}} = \sigma_{\hat{g}} / \sigma_y$. And $\mathbf{r}_{y\hat{g}}$ can be denoted as:

$$\mathbf{r}_{y\hat{g}} = \frac{\sigma_{\hat{g}}}{\sigma_g} \times \frac{\sigma_g}{\sigma_y},$$

which can be rewritten as $\mathbf{r}_{y\hat{g}} = \mathbf{r}_{g\hat{g}} \times \sqrt{h^2}$, and when combined with [1] gives

$$\mathbf{r}_{y\hat{g}} = \sqrt{\frac{\lambda h^4}{\lambda h^2 + 1}}. \qquad [2]$$

Hence, $\mathbf{r}_{y\hat{g}}$ can also be transformed into $\mathbf{r}_{g\hat{g}}$. The accuracy, $\mathbf{r}_{y\hat{g}}$, was used to calculate the number of QTL affecting EB and the number of records needed to improve the accuracy of the predicted DGV.

RESULTS

The pedigree check step for data quality control proved a very effective additional measure to identify any animal that had an incorrectly recorded pedigree or where an animal may have been misidentified. It allowed checking of half-sibling and full-sibling relationships, which is not possible using non-Mendelian checking. Figure 1 effectively illustrates the additional information contained in the SNP data about the relatedness of the animals. This is most obviously shown by the monozygotic twins that have a marker relationship of 1 (because of identical DNA) but are recorded as full sibs in the pedigree. The negative marker relationships are due to the method used to calculate the G matrix, which ideally uses the allele frequencies that were present in the base population (VanRaden, 2008). However, because the frequencies in the base population were unknown, the G matrix was calculated using the allele frequencies in the available highly selected population resulting in negative marker relationships.

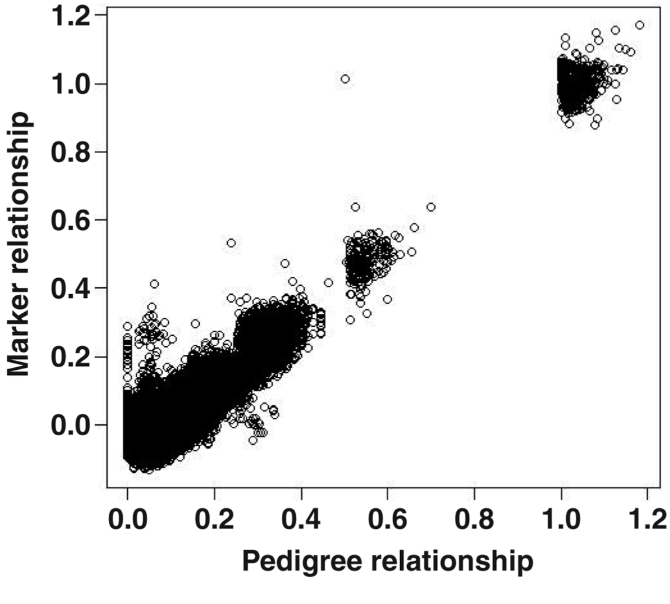

Figure 1. Comparison of the coefficients of the additive relationship matrix (pedigree relationship) and the coefficients of the genomic relationship matrix (markers relationship).

Table 1. Accuracies and reliabilities[1] of direct genomic values (DGV) and EBV

Model[2]	$\mathbf{r}_{y\hat{g}}$	$\mathbf{r}_{g\hat{g}}$	$\mathbf{r}^2_{y\hat{g}}$	$\mathbf{r}^2_{g\hat{g}}$
DGV	0.294	0.516	0.086	0.265
EBV	0.211	0.370	0.044	0.135

[1] $\mathbf{r}_{y\hat{g}}$ = Pearson correlation between the predicted breeding values ($\hat{g}$) and the phenotypes (y); $\mathbf{r}_{g\hat{g}}$ = accuracy of selection [comparing the predicted breeding values ($\hat{g}$) and the true breeding values (g)]; $\mathbf{r}^2_{y\hat{g}}$ = reliability of the predicted phenotypes; and $\mathbf{r}^2_{g\hat{g}}$ = reliability of the predicted breeding values.

[2] DGV was predicted using the model that included both the SNP and polygenic effects, and EBV was predicted using the model that included only the polygenic effect.

The accuracies ($\mathbf{r}_{y\hat{g}}$) of predicting phenotypes for the 2 models and the $\mathbf{r}^2_{y\hat{g}}$ are shown in Table 1. Transformed values, using [1] to give the accuracies of selection ($\mathbf{r}_{g\hat{g}}$) and reliabilities ($\mathbf{r}^2_{g\hat{g}}$), are also shown. The model that included the SNP information yielded an overall accuracy of 0.29, which was higher than the overall accuracy of 0.21 produced by the polygenic model.

The calculated reliability ($\mathbf{r}^2_{g\hat{g}}$) of the DGV is double that of the EBV produced by the polygenic model. This implies that the DGV explained twice as much variation as the EBV, which is also illustrated by the range of breeding values (see Figure 2). The predicted DGV and EBV were positively correlated with a value of 0.70.

The heritability for EB was estimated separately, as described earlier, with a moderate value of 0.325 (SE = 0.12). This value was then used with the accuracy $\mathbf{r}_{y\hat{g}}$ and number of phenotypic records to predict the number of effective QTL for EB. A total of 472 effective QTL were predicted. Figure 3 shows a plot of $\mathbf{r}^2_{y\hat{g}}$ ($\mathbf{r}^2_{g\hat{g}}$ is provided for comparison on the second y-axis) against the number of effective QTL for differing heritabilities where the number of records was kept constant at the available number of 527. This shows the effect that the number of effective QTL would have on the expected accuracies and reliabilities of the DGV. It demonstrated

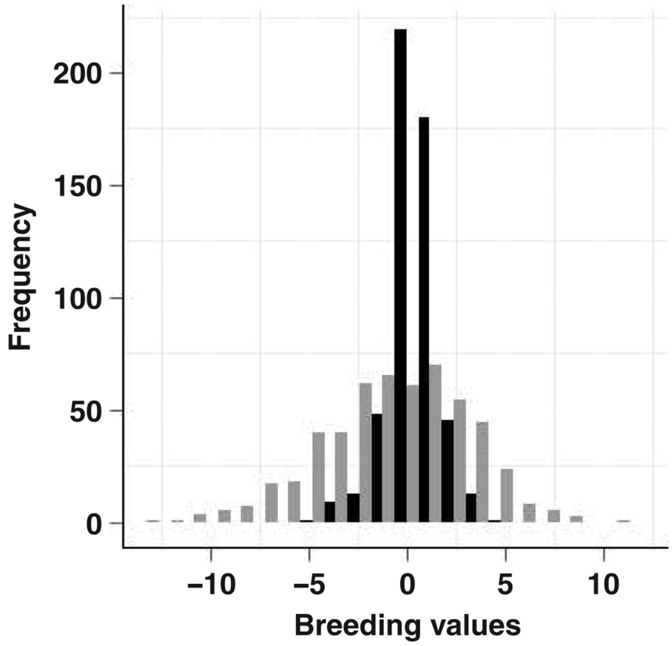

Figure 2. Histogram of direct genomic value (DGV) and EBV; black bars represent the EBV predicted by the polygenic model, and gray bars represent the DGV predicted by the model including the SNP information.

that the greater the number of QTL affecting the trait, the lower the expected accuracy and reliability. This is because of a lack of information available in the limited number of phenotypes to be able to accurately estimate large numbers of QTL effects. Figure 3 also illustrates that this reduction in reliability, as the number of effective QTL increases, is more gradual for higher heritabilities.

The number of total records needed to improve the accuracy was also investigated and results are shown in Figure 4. The heritability was set at the observed value for EB (0.325). It is evident from Figure 4 that the number of effective QTL has a significant effect on the number of records needed to improve the accuracy. The greater the number of effective QTL, the larger the number of phenotypic records required to reach higher accuracies and reliabilities. We predicted that 5,818 records with phenotype and genotype information would be needed for an $\mathbf{r}^2_{y\hat{g}}$ of 0.24 ($\mathbf{r}^2_{g\hat{g}}$ of 0.80) for EB with the predicted 472 effective QTL.

DISCUSSION

The objective of this study was to demonstrate the genetic basis of EB and show that it could be incorporated into selection programs using genomic selection based on a limited reference population. Energy balance is a minimally recorded trait and consequently only a small number of phenotypic records was available. Despite the limitation on available data, genomic prediction was able to produce accuracies greater than a traditional polygenic model. Thus, the results indicated that EB could be estimated using genomic prediction. The low accuracy gained can be explained as a direct result of the small number of phenotypic records and the moderate heritability found for this trait. The heritability calculated with this data set was consistent with results of other studies (Veerkamp,

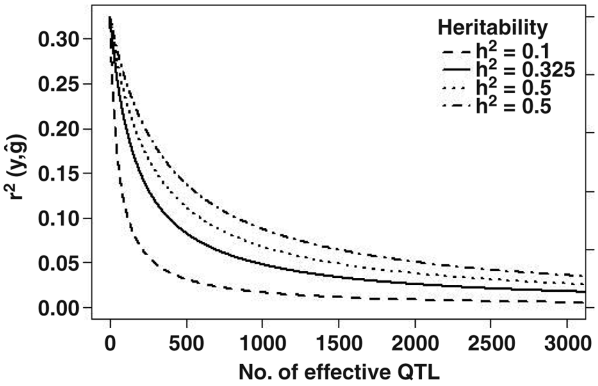

Figure 3. Accuracy of prediction versus the number of effective QTL, where the number of records is fixed to the number used in this study (527); $r^2\left(y,\hat{g}\right)$ is the squared correlation between the phenotypes and the predicted direct genomic values (DGV, characterized in the text as $\mathbf{r}^2_{y\hat{g}}$); $r^2\left(g,\hat{g}\right)$ is the estimated reliability between the true breeding value and the predicted DGV (characterized in the text as $\mathbf{r}^2_{g\hat{g}}$).

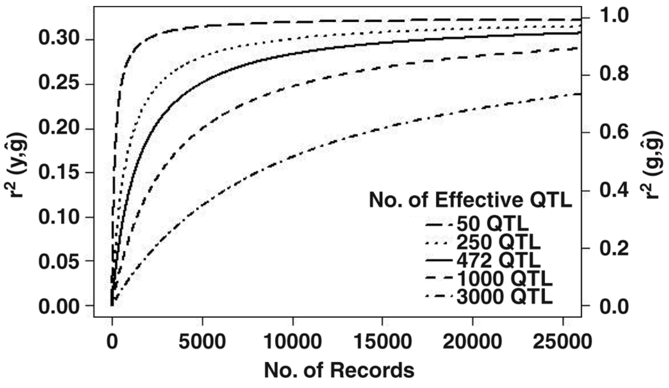

Figure 4. Accuracy of prediction versus the number of records for a fixed heritability of 0.325; $r^2\left(y,\hat{g}\right)$ is the squared correlation between the phenotypes and the predicted direct genomic values (DGV, characterized in the text as $\mathbf{r}^2_{y\hat{g}}$); $r^2\left(g,\hat{g}\right)$ is the estimated reliability between the true breeding value and the predicted DGV (characterized in the text as $\mathbf{r}^2_{g\hat{g}}$).

1998; Huttmann et al., 2009). To consider including EB in breeding schemes, higher accuracies than found here would be necessary. This increase in accuracy could be facilitated through an increase in the heritability of the trait or an increase in the number of phenotypic records. One way to increase the heritability would be to standardize the environmental conditions to reduce nongenetic differences between animals, but this may be difficult to do in practice. An alternative approach to increase the heritability of phenotypes would be to use deregressed breeding values or DYD of proven bulls as phenotypes, based on EB records of many daughters. This allows for an increase in the accuracy while keeping the number of genotyped animals constant. This scenario would not lead to any additional genotyping costs because most bulls may already be genotyped as part of reference populations for other breeding goal traits. Note, however, that this approach may still be more costly because of the (much) higher number of recorded EB phenotypes that would be needed. An increase in the number of available records would also allow for an increase in the accuracy of predicted DGV as indicated in other studies (Hayes and Goddard, 2008; Goddard, 2009). The required increase could occur only if the measurement and recording of EB improved.

Because of infrequent recording of EB, a seemingly obvious solution would be to immediately select for a widely recorded trait such as BCS to reduce NEB indirectly. The problem with using BCS is that after the first 60 DIM, the genetic correlations between EB and BCS decrease markedly (Huttmann et al., 2009). However, until the recording of EB increases to useful levels, BCS does provide a viable option to attempt to select animals with a better EB. In the future having both EB and BCS phenotypes available should allow for the best prediction of energy partitioning and utilization.

The model used to predict the DGV could also be used for whole-genome association studies. Thus, the produced posterior probabilities of SNP were examined to see if there were any significant associations with EB. Because of the small number of records and large number of SNP, the power of the association study to identify QTL was very low and this was evident. There was no SNP with a high enough posterior probability to be confident that it was linked to a QTL. The prior for the expected number of QTL affecting EB varied but results were consistently low (results not presented). Although the posterior probabilities were low, one SNP had 10-fold higher posterior probabilities than all the other SNP in all analyses regardless of the prior probability used. This SNP is located on BTA21 and is in extremely close proximity to, and appears in association with, the nuclear receptor subfamily 2, group F, member 2 (NR2F2), otherwise known as chicken ovalbumin upstream promoter transcription factor II (COUP-TFII); COUP-TFII has been previously reported as playing an essential role in regulating adipogenesis, glucose homeostasis and energy metabolism (Xu et al., 2008; Li et al., 2009). It has also been reported as regulating growth hormone receptor 1A promoter activity (Xu et al., 2004), mediating progesterone, and controlling estrogen levels and thus involved in reproduction (Klinge et al., 1997; Nakshatri et al., 2000; Takamoto et al., 2005; Kurihara et al., 2007; Petit et al., 2007). Although the results of this association study are not conclusive and further validation is required, COUP-TFII appears to be a good candidate gene for EB.

Despite being unable to establish QTL conclusively associated with EB, results of the study allowed an estimation of the number of effective QTL influencing EB. Given the nature and complexity of EB, the number of predicted effective QTL (472) was plausible. The relationships with both production and nonproduction traits mean that numerous genes and pathways could be involved in the variation observed in EB. Previous whole-genome association studies of residual feed intake and other traits related to EB in beef cattle identify between 4 and 120 QTL affecting the traits studied (Barendse et al., 2007; Sherman et al., 2009). These values are significantly lower than the predicted 472, but reflect the power of the studies to detect significant QTL and the number of SNP (which were 2,194 and 8,786 respectively), rather than the true number of effective QTL. An increase in the number of phenotypic records would also allow genome-wide association studies for EB in dairy cattle to identify possible candidate genes affecting the trait and would provide a better idea of the effective number of QTL.

The ability to select and include EB in selection indices may indirectly increase the genetic gain for fertility traits. The interval between calving and start of luteal activity (**C-LA**) has been demonstrated to be an indicator of fertility during later lactation (Darwash et al., 1999; van der Lende et al., 2004; Petersson et al., 2007). Veerkamp et al. (2000) reported genetic correlations between EB and C-LA of −0.60 (and −0.49 for C-LA adjusted for milk, fat, and protein). A moderate to high genetic correlation similar to what was previously reported would mean that genetic gain for EB should also result in improved fertility. For example, if a bull had 25 daughters, the accuracy of selection for the bull's EBV would be 0.40 for fertility (assuming a heritability of 0.03), whereas the accuracy of selection for the bull's EBV would be 0.83 for EB. Thus, given a genetic correlation of −0.5, the accuracy of selection for fertility using EB would be 0.41. Consequently, for bulls with this number of daughters or fewer, selection using EB would result in greater genetic gain for fertility compared with selecting for fertility itself. However, as the number of offspring per bull increases beyond 25, the benefit of using EB rather than fertility is lost, such that selection for fertility itself will produce better genetic gains. Thus, the use of EB in selection indices, in addition to fertility, may prove beneficial and result in increased genetic gain for fertility.

In addition to the possible benefits of improved fertility, EB could be used with feed intake data to select animals for feed efficiency (Veerkamp, 1998) or to reduce methane emission (Hegarty et al., 2007). Improving feed efficiency could be economically desirable because feed costs contribute the greatest proportion to production costs (Simm et al., 1994). However, feed efficiency data alone cannot distinguish whether the energy is used for production or maintenance. This may result in selection of animals with low intake and high yield that, consequently, have problems related to high NEB. Thus, NEB and improved feed efficiency (or intake) data should be considered simultaneously to effectively reduce the feeding costs while not having detrimental effects on the animals' health and fertility.

Many other traits including several fertility and reproduction traits such as milk progesterone profiles and milk quality traits are difficult to record. Accounting for these traits, like EB, in selection has been complicated, because measuring them in progeny testing schemes is not practical. This study demonstrates that it is possible for such traits with similar heritabilities and expected number of QTL to produce DGV with accuracies >0.8 when there are more than approximately 2,600 (2,581 predicted for EB) phenotypic records available for use as the reference population. This demonstrates that it is possible to select for these traits using genomic selection by combining data from experimental and nucleus herds, where individually there are a limited numbers of raw phenotypic records.

The statistical approaches used in this study are generally accepted as appropriate for genomic prediction. Genomic prediction is often performed using a 2-step procedure in which the input phenotypes are precorrected so that the model predicting the DGV includes only the mean, polygenic, and SNP effects. Because precorrection will always introduce a new source of error, our preference would be to include all fixed effects in the models used to predict the breeding values. There is ongoing development of the genomic prediction program used, to allow the inclusion of multiple discrete and continuous fixed effects in a single model.

CONCLUSIONS

The use of SNP information to predict DGV is shown to explain variation among the EB of animals, confirming the genetic background of EB. The use of SNP information showed an increase in the accuracy of selection for EB over the simple polygenic model. However, the extent of recording would need to be improved to increase the accuracy. In the future, selection for EB could be performed using genomic selection, which could provide a valuable tool in finding a balance between production and nonproduction traits.

ACKNOWLEDGMENTS

The authors acknowledge funding for the study from the Dutch Dairy Board (PZ; Zoetermeer, the Netherlands) and Senter Novem (The Hague, the Netherlands). The study was also performed as part of the RobustMilk and SABRETRAIN projects. The RobustMilk project is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. The first author was funded by the SABRETRAIN Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission. This publication represents the views of the authors, not the European Commission, and the Commission is not liable for any use that may be made of the information

REFERENCES

Agenäs, S., E. Burstedt, and K. Holtenius. 2003. Effects of feeding intensity during the dry period. 1. Feed intake, body weight, and milk production. J. Dairy Sci. 86:870–882.

Barendse, W., A. Reverter, R. J. Bunch, B. E. Harrison, W. Barris, and M. B. Thomas. 2007. A validated whole-genome association study of efficient food conversion in cattle. Genetics 176:1893–1905.

Butler, W. R., and R. D. Smith. 1989. Interrelationships between energy-balance and postpartum reproduction function in dairy cattle. J. Dairy Sci. 72:767–783.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561.

Coffey, M. P., G. Simm, J. D. Oldham, W. G. Hill, and S. Brotherstone. 2004. Genotype and diet effects on energy balance in the first three lactations of dairy cows. J. Dairy Sci. 87:4318–4326.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395.

Darwash, A. O., G. E. Lamming, and J. A. Woolliams. 1999. The potential for identifying heritable endocrine parameters associated with fertility in post-partum dairy cows. Anim. Sci. 68:333–347.

Dewhurst, R. J., J. M. Moorby, M. S. Dhanoa, R. T. Evans, and W. J. Fisher. 2000. Effects of altering energy and protein supply to dairy cows during the dry period. 1. Intake, body condition, and milk production. J. Dairy Sci. 83:1782–1794.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longmans Green, Harlow, UK.

Friggens, N. C., P. Berg, P. Theilgaard, I. R. Korsgaard, K. L. Ingvartsen, P. Lovendahl, and J. Jensen. 2007. Breed and parity effects on energy balance profiles through lactation: Evidence of genetically driven body energy change. J. Dairy Sci. 90:5291–5305.

Garnsworthy, P. C., A. Lock, G. E. Mann, K. D. Sinclair, and R. Webb. 2008a. Nutrition, metabolism, and fertility in dairy cows: 1. Dietary energy source and ovarian function. J. Dairy Sci. 91:3814–3823.

Garnsworthy, P. C., A. Lock, G. E. Mann, K. D. Sinclair, and R. Webb. 2008b. Nutrition, metabolism, and fertility in dairy cows: 2. Dietary fatty acids and ovarian function. J. Dairy Sci. 91:3824–3833.

George, E. I., and R. E. McCulloch. 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88:881–889.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2006. ASREML Program User Manual. 2nd ed. VSN International Ltd., Hemel Hempstead, UK.

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

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

Hayes, B. J., and M. E. Goddard. 2008. Technical note: Prediction of breeding values using marker-derived relationship matrices. J. Anim. Sci. 86:2089–2092.

Hegarty, R. S., J. P. Goopy, R. M. Herd, and B. McCorkell. 2007. Cattle selected for lower residual feed intake have reduced daily methane production. J. Anim. Sci. 85:1479–1486.

Huttmann, H., E. Stamer, W. Junge, G. Thaller, and E. Kalm. 2009. Analysis of feed intake and energy balance of high-yielding first lactating Holstein cows with fixed and random regression models. Animal 3:181–188.

Jorritsma, R., T. Wensing, T. A. M. Kruip, P. Vos, and J. Noordhuizen. 2003. Metabolic changes in early lactation and impaired reproductive performance in dairy cows. Vet. Res. 34:11–26.

Klinge, C. M., B. F. Silver, M. D. Driscoll, G. Sathya, R. A. Bambara, and R. Hilf. 1997. Chicken ovalbumin upstream promoter transcription factor interacts with estrogen receptor, binds to estrogen response elements and half-sites, and inhibits estrogen-induced gene expression. J. Biol. Chem. 272:31465–31474.

Kurihara, I., D. K. Lee, F. G. Petit, J. Jeong, K. Lee, J. P. Lydon, F. J. DeMayo, M. J. Tsai, and S. Y. Tsai. 2007. COUP-TFII mediates progesterone regulation of uterine implantation by controlling ER activity. PLoS Genet. 3:1053–1064.

Li, L. P., X. Xie, J. Qin, G. S. Jeha, P. K. Saha, J. Yan, C. M. Haueter, L. Chan, S. Y. Tsai, and M. J. Tsai. 2009. The nuclear orphan receptor COUP-TFII plays an essential role in adipogenesis, glucose homeostasis, and energy metabolism. Cell Metab. 9:77–87.

McNamara, S., J. J. Murphy, F. P. O'Mara, M. Rath, and J. F. Mee. 2008. Effect of milking frequency in early lactation on energy metabolism, milk production and reproductive performance of dairy cows. Livest. Sci. 117:70–78.

McNamara, S., J. J. Murphy, M. Rath, and F. P. O'Mara. 2003. Effects of different transition diets on energy balance, blood metabolites and reproductive performance in dairy cows. Livest. Prod. Sci. 84:195–206.

Meuwissen, T. H. E., and M. E. Goddard. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261–279.

Miglior, F., B. L. Muir, and B. J. Van Doormaal. 2005. Selection indices in Holstein cattle of various countries. J. Dairy Sci. 88:1255–1263.

Nakshatri, H., M. S. Mendonca, P. Bhat-Nakshatri, N. M. Patel, R. J. Goulet, and K. Cornetta. 2000. The orphan receptor COUP-TFII regulates G2/M progression of breast cancer cells by modulating the expression/activity of p21(WAF1/CIP1), cyclin D1, and cdk2. Biochem. Biophys. Res. Commun. 270:1144–1153.

Petersson, K. J., B. Berglund, E. Strandberg, H. Gustafsson, A. P. F. Flint, J. A. Woolliams, and M. D. Royal. 2007. Genetic analysis of postpartum measures of luteal activity in dairy cows. J. Dairy Sci. 90:427–434.

Petit, F. G., S. P. Jamin, I. Kurihara, R. R. Behringer, F. J. DeMayo, M. J. Tsai, and S. Y. Tsai. 2007. Deletion of the orphan nuclear receptor COUP-THII in uterus leads to placental deficiency. Proc. Natl. Acad. Sci. USA 104:6293–6298.

Pryce, J. E., M. D. Royal, P. C. Garnsworthy, and I. L. Mao. 2004. Fertility in the high-producing dairy cow. Livest. Prod. Sci. 86:125–135.

Royal, M. D., A. O. Darwash, A. P. E. Flint, R. Webb, J. A. Woolliams, and G. E. Lamming. 2000. Declining fertility in dairy cattle: Changes in traditional and endocrine parameters of fertility. Anim. Sci. 70:487–501.

Sherman, E. L., J. D. Nkrumah, C. Li, R. Bartusiak, B. Murdoch, and S. S. Moore. 2009. Fine mapping quantitative trait loci for feed intake and feed efficiency in beef cattle. J. Anim. Sci. 87:37–45.

Simm, G., R. F. Veerkamp, and P. Persaud. 1994. The economic performance of dairy cows of different predicted genetic merit for milk solids production. Anim. Prod. 58:313–320.

Takamoto, N., I. Kurihara, K. Lee, F. J. DeMayo, M. J. Tsai, and S. Y. Tsai. 2005. Haploinsufficiency of chicken ovalbumin upstream promoter transcription factor II in female reproduction. Mol. Endocrinol. 19:2299–2308.

van der Lende, T., L. Kaal, R. M. G. Roelofs, R. F. Veerkamp, C. Schrooten, and H. Bovenhuis. 2004. Infrequent milk progesterone measurements in daughters enable bull selection for cow fertility. J. Dairy Sci. 87:3953–3957.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

Veerkamp, R. F. 1998. Selection for economic efficiency of dairy cattle using information on live weight and feed intake: A review. J. Dairy Sci. 81:1109–1119.

Veerkamp, R. F., B. Beerda, and T. van der Lende. 2003. Effects of genetic selection for milk yield on energy balance, levels of hormones, and metabolites in lactating cattle, and possible links to reduced fertility. Livest. Prod. Sci. 83:257–275.

Veerkamp, R. F., J. K. Oldenbroek, H. J. Van Der Gaast, and J. H. J. Van Der Werf. 2000. Genetic correlation between days until start of luteal activity and milk yield, energy balance, and live weights. J. Dairy Sci. 83:577–583.

Watters, R. D., M. C. Wiltbank, J. N. Guenther, A. E. Brickner, R. R. Rastani, P. M. Fricke, and R. R. Grummer. 2009. Effect of dry period length on reproduction during the subsequent lactation. J. Dairy Sci. 92:3081–3090.

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431–3436.

Xu, Q., N. Walther, and H. Jiang. 2004. Chicken ovalbumin upstream promoter transcription factor II (COUP-TFII) and hepatocyte nuclear factor 4 gamma (HNF-4 gamma) and HNF-4 alpha regulate the bovine growth hormone receptor 1A promoter through a common DNA element. J. Mol. Endocrinol. 32:947–961.

Xu, Z., S. Yu, C. H. Hsu, J. Eguchi, and E. D. Rosen. 2008. The orphan nuclear receptor chicken ovalbumin upstream promoter-transcription factor II is a critical regulator of adipogenesis. Proc. Natl. Acad. Sci. USA 105:2421–2426.