Computationally Efficient Genomic Prediction From Whole Genome Sequence Data In Dairy Cattle

Submitted by

TingTing Wang (MSc)

A thesis submitted in total fulfilment of

the requirements for the degree of

Doctor of Philosophy

Department of Computer Science and Information Technology

School of Engineering and Mathematical Sciences

College of Science, Health and Engineering

La Trobe University

Bundoora, Victoria 3083

Australia

July 2016

Contents

List of Figures

List of Tables

Abbreviations

AUC – Area Under the Receiver Operating Characteristic

BD – Bipolar Disorder

BLUP – Best Linear Unbiased Prediction

BP – Base Pair

CAD – Coronary Artery Disease

CD – Crohn's Disease

CI – Calving Interval

DTD – Daughter Trait Deviations

EM – Expectation-Maximization

GBLUP – Genomic Best Linear Unbiased Prediction

GEBV – Genomic Estimated Breeding Value

GP – Genomic Prediction

GWAS – Genomic Wide Association Studies

HD – High Density

HT – Hypertension

ICE – Iterative Conditional Expectation

IGF1 – Insulin Like Growth Factor 1

LD – Linkage disequilibrium

LLPF – Longissimus Lumborum Muscle

MAF – Minor Allele Frequency

MAP – Maximum A Posterior

MAS – Marker Assisted Selection

MCMC – Markov Chain Monte Carlo

PEV – Prediction Error Variance

PWIGF – Post Weaning Weight

QC – Quality Control

QTL – Quantitative Traits Locus

RA – Rheumatoid Arthritis

SCC – Somatic Cell Count

SNP – Single Nucleotide Polymorphism

TBV – True Breeding Value

T1D – Type 1 Diabetes

T2D – Type 2 Diabetes

WTCCC – Welcome Trust Case Control Consortium

Abstract

The prediction of genetic merit for complex or quantitative traits from high-density SNP panels is increasingly used in animal and plant breeding, to select breeding individuals early in life in order to accelerate genetic gains. The genomic prediction methodology is also of interest in human disease studies for the prediction of disease risk and identification of causal mutations (quantitative trait loci, QTL, mapping). Bayesian models for genomic prediction, incorporating prior assumptions regarding the distribution of QTL effects, have been shown to give good accuracies of genomic prediction across a wide range of traits and species. These models are usually implemented with Monte Carlo Markov Chain (MCMC) sampling, which results in (impractically) long compute time with the very large genomic data sets now available. This study aimed to dramatically improve the computational efficiency of Bayesian methods, such that genomic prediction is practical in large genomic data sets, while maintaining their predictive ability. The thesis included three main areas of work

1) Develop an Expectation-Maximization (EM) algorithm to substitute for MCMC sampling in the Bayesian models for genomic prediction, in order to reduce compute times. One key improvement over existing fast Bayes approaches was the introduction of prediction error variance correction during SNP effect estimation to account for the errors generated by estimation of many thousands of SNP effects The EM algorithm was up to 30 orders faster than MCMC in large dairy cattle genomic data sets. However the predictive ability of the EM was not flexible to all trait architectures - for the traits affected by mutations of large effect, EM methods could have a reduced accuracy of prediction of up to 7% compared with MCMC implementations;

2) To overcome this limitation, the next strategy was to hybridize the EM algorithm with a limited number of MCMC sampling iterations (the “Hybrid”). The hybrid version was 17 orders of magnitude faster to run than the full MCMC, and gave the same

accuracy of genomic prediction across a wide range of traits with different genetic architectures.

3) Finally, the Hybrid genomic prediction algorithm was demonstrated with a large dairy cattle genomic data set, with a sizable subset of imputed whole genomic sequence data and phenotypes for milk production, fertility, and heat intolerance traits, for both genomic prediction and QTL mapping.

The final stage of this thesis demonstrates that the Hybrid genomic prediction algorithm developed here makes simultaneous genomic prediction and QTL mapping feasible in large genomic data sets, up to whole genome sequence data.

Statement of Authorship

I certify that the thesis entitled computationally efficient genomic prediction from whole genome sequence data in dairy cattle is the result of my own work. This thesis includes work by the author that has been published or accepted for publication as described in the text. Except where reference is made in the text of the thesis, this thesis contains no other material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgment in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

I also certify that there is the extent of collaboration with another person or persons; there is the extent and the nature of any other assistance received in the pursuit of the research and preparation of the thesis; (e.g., processing of statistical data, translation of texts, assistance with English expression, building and operation of equipment); and all research procedures reported in the thesis were approved by the relevant Ethics Committee or Safety Committee or authorized officer.

Full Name (Print) Tingting Wang

Full Name (Signed)

Date 21/10/2016

Acknowledgements

Being a PhD is a long but wonderful journey. One can be a "professional" researcher only when he climbs the mountain of science and tastes all the flavors of life. To pick up the most colorful pearl on the crown of scientific knowledge, no one can endeavor without help. There are always “somebodies” standing along this road who devote themselves to your research like a lighthouse. Here, I hope to sincerely thank them from the bottom of my heart.

I would like to thank Professor Yi-Ping Phoebe Chen, my principal supervisor, for your inspiration and instruction. Professor Chen’s weekly meeting has enriched my life of PhD study. Your great advice and faith for my research guide a bright direction for my future life.

I would express my appreciation to my supervisor, Professor Ben Hayes, whose enthusiasm of listening, discussing and encouraging has made my PhD studies all different. Without your invaluable contributions throughout the thesis, I have no confidence that I could have completed this work. I am forever grateful for that! Additionally, my sincere thanks go to Professor Mike Goddard. Any time when I was hindered by some statistical problems, you would always be there to shed light on them. You have taught me how wonderful the research could be!

Special thanks are reserved for the Dairy Futures Cooperative Research Centre, and the Department of Economic Development, Jobs, Transport and Resources who generously provided the scholarship to support my research. They have provided me comfortable working places with advanced facilities and services. Specially, I would like to thank Belinda Griffiths for offering me great opportunities to express my research to farmers

and industry partners. In return, I can hear different voices from them. Your contributions made my PhD studies much more colorful!

I also acknowledge lovely colleagues including Jennie Pryce, Hans Daetwyler, Iona Macleod, Kathryn Kemper, Bolormaa Sunduimijid, etc. There is a long list. Forgive me not to list everyone. All of you have offered me a great deal of help and a lot of patience! I would specially like to thank Phil Bowman. Much of my efficient C++ programming skills came from you! And special thanks to my PhD friends: Majid Khansefid, Ross Koufariotis, Hassan Aliloo, Min Wang, and Mary Abdelsayed for your company! Talking with you during the lunch break has become the most enjoyably part of every weekday.

Finally, I would like to dedicate this thesis for my family. I would express my deepest gratitude to my parents and parents-in-law. Without your endless love and effort, I could never have concentrated on my PhD study. Many thanks for my lovely daughter, Cynthia! Every time I see your beautiful face, I would forget any annoying problem no matter what it is. You are my angel! The most important acknowledgement goes to my Husband, Feng Chen, who has sacrificed so much to help me start and complete my PhD study. Your company gave me endless encouragement during this long journey!

Thesis Preface

This thesis is presented as a series of research articles (Chapters 3-6). Each chapter is prefaced by a short justification of the work, the publication status and the contribution of co-authors. Because each chapter is identical to the published/submitted paper (excluding format and reference style) there may be overlap particularly in the introduction and methods sections between chapters. Additionally, the expression may vary slightly between chapters (for example use of first/third person expression and American/English spelling), and some terms are defined in multiple chapters. For continuity the reference style has been changed within the papers to be constant throughout the thesis, additionally the figure and table numbers have been prefixed with the chapter number (e.g. Table 1 in Chapter 2 is now titled Table 2.1). Supplementary material associated with each research chapter (where applicable) is presented in the appendices including Appendix I (Chapter 8), Appendix II (Chapter 9), and Appendix III (Chapter 10). The author contributions have been removed from the title of each publication as this information is presented in the chapter preface.

List of Publications

Journals published:

- Wang T, Chen YPP, Goddard ME, Meuwissen TH, Kemper KE, Hayes BJ. (2015) A computationally efficient algorithm for genomic prediction using a Bayesian model. Genetics Selection Evolution, 47(1):34.
- Wang T, Chen YPP, Bowman PJ, Goddard ME, Hayes BJ. (2016) A hybrid expectation maximization and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping, BMC Genomics, 17:744.
- Wang T, Chen YPP, Hayes BJ. (2016) Review of the Accuracy and Computational Efficiency of Genomic Prediction Models with High Density Genotype Data. CAB review, accepted.

Conference:

- Wang T, Chen YPP, Kemper KE, Goddard M, Hayes BJ. (2015) Opt_emBR: Computationally efficient genomic prediction and QTL mapping in multi-breed populations. Proceeding of the Association for the Advancement of Animal Breeding and Genetics, 21: 449-452.

Chapter 1 General Introduction

1.1 Introduction

Genomic predictions use information from high-density genetic polymorphisms, such as single nucleotide polymorphisms (SNP), to predict the genetic merit of individuals for complex traits (Meuwissen *et al.* 2001). Genomic prediction methods could also be used to identify genome regions harboring causative mutations, or the mutations themselves (Moser *et al.* 2015). Genomic selection in animal and plant breeding could be implemented in two steps: 1) estimation of the effects of SNPs in a reference population given the phenotypes and SNP genotypes of reference individuals and (2) calculation of genetic values for selection candidates (or target individuals) based on their genotypes, followed by selection of the candidates with the highest genomic predictions for breeding. Prediction of genetic values for individuals from SNP genotypes, without the selection step, is often referred to as genomic prediction.

When estimating SNP effects, any genomic prediction method must overcome the $p >> n$ problem – that is there are usually many more SNPs (p, the parameters) than observations (the n). A number of genomic prediction methods, which dealt with the large-p-with-small-n problem, have been proposed.

Best linear unbiased prediction (BLUP) defined a linear combination of SNPs by assuming the normal prior distribution for SNP effects – that is SNP effects were treated as random variables derived from a normal distribution with the same variance (Meuwissen *et al.* 2001). SNP-BLUP was also referred to as ridge regression when the variance components or λ parameter were estimated by cross validation (Whittaker JC 2000; Meuwissen *et al.* 2001). A method termed genomic BLUP or GBLUP (Habier *et al.* 2007; Hayes & Goddard 2008; VanRaden

2008) was an equivalent model that made use of the genomic relationship matrix (with elements the proportion of the genome shared by each pair of individuals estimated from the markers) to estimate genetic value (breeding values) directly.

Alternative prior assumption of SNP effects could be considered. For example, there might be a mutation or mutations with moderate or large effects on the trait of interest, such that the SNPs in linkage disequilibrium (LD) with this mutation were associated with moderate to large effects. Another prior assumption might be that only a proportion of the SNPs were in LD with mutations affecting the trait, in which case the effect ascribed to these SNP should be zero. To accommodate these alternative assumptions, a series of methods (termed Bayesian regression models) proposed non-normal prior assumptions for SNP effects that a large proportion of SNPs had effects close to zero, or actually were zero, while a proportion of SNPs had moderate to large effects. These included Bayes A/B (Meuwissen *et al.* 2001), C(π), D(π) (Habier *et al.* 2011), Bayesian Lasso (Park & Casella 2008), etc. In detail, BayesA proposed a t-distribution for SNP effects, while BayesB, C(π), D(π), and Bayesian Lasso (also dubbed as Bayesian variable selection models) assumed mixture priors with the possibility of excluding some SNPs from the model. All these Bayesian models were also termed as non-linear model, as the result of non-linear combination of SNP effects. As a typical example of a Bayesian non-linear and variable selection method, BayesR (Erbe *et al.* 2012) assumed a proportion of SNPs with zero effects and that the others followed a mixture of three normal distributions with variances $0.0001\sigma_g^2$, $0.001\sigma_g^2$, and $0.01\sigma_g^2$, where σ_g^2 was the total genetic variance of the trait. Due to this prior distribution for SNP effects, BayesR could be quite flexible for a range of traits of different genetic architectures.

The accuracies of genomic prediction from BLUP and the Bayesian non-linear models had now been compared in data sets from a range of species. To

summarize these results, non-linear models coupled with Markov Chain Monte Carlo (MCMC) scheme had been demonstrated to had superior or equal prediction ability in many cases (Kemper *et al.* 2015; Moser *et al.* 2015; MacLeod *et al.* 2016), when compared with BLUP model. Increases in accuracy with the Bayesian non-linear methods had mostly been observed for traits with large effect causal mutations (e.g. Fat percent in dairy cattle; Type 1 diabetes, or rheumatoid arthritis in human diseases) (Kemper *et al.* 2015; Moser *et al.* 2015). Moreover, the advantage of Bayesian methods could further increase as selection candidates/validation set became more distant for the reference set, either in time or in genetic diversity (e.g. across-breeds). This was because in the BLUP model, a linear combination of effects of a large number of markers typically captured the effect of each causal variant. However, the long range of the association between markers and causal variants could be easily broken down by recombination over generations or when the animals to be predicted were more genetically distant from the reference set. In comparison with BLUP models, the nonlinear assumption from Bayesian models allowed a prediction driven by a limited number of markers in close association with each QTL.

In addition to the prediction for unknown phenotypes, the nonlinear assumption of SNP effects from Bayesian models, and the fact that all SNPs were fitted simultaneously, have also been demonstrated to improve the precision of identifying the causal variants (QTL mapping). The traditional methods of identifying QTL, genomic wide association studies (GWAS), analyzed a single SNP at a time, and then a stringent significance threshold was used to account for multiple testing of a very large number of SNPs. GWAS methods have been applied for detecting QTL affecting complex traits in human or animals. However there were several limitations of GWAS for detecting QTL. One was that because SNPs were fitted one by one, the LD between SNPs was not accounted for, which could result in quite imprecise QTL regions. This was overcome in the Bayesian methods, because all SNPs were fitted simultaneously, so if a small number of

SNPs captured the effect of the mutation, it did not “spillover” to other SNPs. Bayesian methods (especially for BayesR) have been demonstrated to increase the precision of causal mutation identification in human (Loh *et al.* 2015; Moser *et al.* 2015) and dairy cattle (Kemper *et al.* 2015). Another limitation with GWAS was that SNPs that exceeded the significance threshold had over-estimated effects due to the “Beavis effect”, which could greatly reduce accuracy of subsequent genomic prediction (Meuwissen *et al.* 2001). This was overcame in the Bayesian methods, by fitting all SNP simultaneously.

Beside the introduction of non-linear Bayesian models, another potential improvement for genomic prediction was the advent of whole genomic sequence data. Several researchers have investigated the benefits of using whole genome sequence data in genomic prediction, using either simulation (Clark *et al.* 2011; Druet *et al.* 2014) or real data (MacLeod *et al.* 2016). Compared with the high-density SNP panels, the key merit of whole sequence data was that it actually included the causal mutation genotypes. As a result, the potential advantage of whole sequence data included better persistence of accuracy across generations, more accurate predictions across breeds, and more precise QTL mapping (Clark *et al.* 2011; Druet *et al.* 2014; MacLeod *et al.* 2014c; MacLeod *et al.* 2016).

Though Bayesian non-linear models performed well both for genomic prediction and QTL mapping, there was one key limitation. These methods were typically employed with Monte Carlo – Markov Chain (MCMC) sampling, which resulted in very long compute times (10s of thousands of iterations were typically required before samples were actually from the posterior distributions of the parameters – that is they were not affected by starting values). As the size of genomic data dramatically increased to high-density SNP panels or even whole-genome sequence variants, the challenge of computational burden for Bayesian models became overwhelming. For example, whole genome sequence data of dairy cattle

including 28.3 million of variants (SNP and indels) have been published by 1,000 bull genomes (Daetwyler *et al.* 2014). Bayesian models coupled with MCMC had unfeasible compute times with such large data sets. To deal with the computational burden, speed-up algorithms (VanRaden 2008; Meuwissen *et al.* 2009; Hayashi & Iwata 2010; Shepherd *et al.* 2010; Yu & Meuwissen 2011; Sun *et al.* 2012) have been proposed. These algorithms introduced fast heuristic algorithms (e.g. Expectation-Maximization, Iterative Conditional Expectation) to improve the computational speed with Bayesian models. These methods have been demonstrated to speed-up Bayesian models under MCMC sampling by several orders of magnitude. However, as pointed out by (Meuwissen *et al.* 2009), a key drawback of these methods stemmed from the fact that within the methods, the effect of each SNP was estimated in turn, correcting the phenotypes for the effect of all the other SNPs. When this was done, these methods assumed the effects of all other SNPs were estimated perfectly during the estimation of the current. That is, the methods did not account for the errors generated by the estimation of all the other SNPs, which was unrealistic. This had limited the prediction accuracy of these fast methods. As a result, these methods had been rarely applied in practice. One exception is the iterative BayesA (B), which introduced standard deviation to correct SNP effects and therefore could be implemented for the practical application (Olson *et al.* 2012).

Accounting for the robust prediction accuracy from MCMC samplings and the efficient computational time of fast algorithms, one possible scheme was to hybridize the fast algorithms with MCMC. In other words, fast algorithms could be used to set up start points for the MCMC samplings, which reduced the large number of burn-in iterations. Then MCMC sampling for a limited number of iterations could be used to improve prediction accuracy.

The main body of the research in this thesis focused on developing computationally efficient algorithms for genomic prediction while maintaining

similar prediction accuracy to the MCMC methods. The ultimate objective was to be able to use whole genome sequence data in genomic prediction.

1.2 Research Objectives

The objective of this research was to develop a genomic prediction method that was both computationally efficient and which gave similar accuracy to Bayesian MCMC methods, in four steps:

1) Introduce the Expectation-Maximization (EM) algorithm to tackle the Bayesian non-linear models, in order to speed-up the computational time (termed emBayesR). One improvement of emBayesR over other fast methods was that Prediction error variance (PEV) correction was introduced to account for the errors generated by estimates from other SNP effects during the estimation of current SNP effect.

2) Extend the emBayesR model to allow different breeds (populations) and phenotype error structures (for example in dairy cattle, cows might have their own records, while a bull's "phenotype" was the average of many daughter records), as well as several speed up schemes. The prediction performance of emBayesR was then evaluated for multi-breed and across-breed predictions.

3) Improve the prediction accuracy of emBayesR by introducing Hybrid schemes, where an Expectation-Maximization algorithm was used initially, followed by a limited number of MCMC loops. In detail, we hypothesized that limited number of EM iterations until convergence promised a good starting point for MCMC, which therefore removed the need for a large number of burn-in iterations. Afterwards, a limited number of MCMC sampling could help to improve the accuracy over what could be achieved with the EM algorithm.

4) Apply the Hybrid schemes to whole genome sequence data (or large subsets of this data) to evaluate and demonstrate its ability for genomic prediction and QTL mapping in large genomic data sets.

1.3. The Outline of Thesis

The thesis included:

Chapter 2, a review of current genomic prediction methodology and the applications. The review first described the methodology of current Bayesian prediction models from the statistical viewpoint (the data model, the prior assumption(s) and the posterior features inferred from different priors). Then the performance of different prediction approaches was evaluated in terms of the prediction accuracy and computational efficiency. An important conclusion from the review of the literature was that an efficient algorithm was required to improve the computational time of Bayesian models, especially for the application to whole genome sequence data, while maintaining a similar genomic prediction accuracy. *The paper was accepted by a journal as an invited review.*

In **Chapter 3**, a novel prediction method was developed by the introduction of an Expectation-Maximization algorithm into the BayesR model (termed emBayesR). One improvement over existing fast methods was that that emBayesR introduced a prediction error variance correction from the GBLUP model to account for the prediction errors produced by the estimation of other SNP effects, when the current SNP effect was estimated. The performance of emBayesR was evaluated on the simulated and practical data. From the results, the superior computational speed of emBayesR could be clearly demonstrated but with some accuracy reductions in the traits controlled by major causal mutations with large effects. The chapter was published as:

> Wang T, Chen YPP, Goddard ME, Meuwissen TH, Kemper KE, Hayes BJ. (2015) A computationally efficient algorithm for genomic prediction using a Bayesian model. Genetics selection evolution. **47**(1):34.

The evaluation study in **Chapter 4** was conducted to investigate the prediction ability of emBayesR in more generalized situations. In detail, this study extended the emBayesR model by incorporating a polygenic breeding value (to capture genetic variation not picked up by the SNP), and weights on phenotypes to accommodate different error variances (e.g. bull phenotypes that were the average of many daughter records, versus individual cow records), and different breeds. Moreover, two novel speed-up schemes were introduced to further improve computational efficiency. The method (termed as optimized emBayesR; Opt_emBR) was evaluated for multi-breed prediction and prediction for a validation set that was genetically quite diverged from the reference set (a breed of cattle not in the reference set). From this study, the advantage and drawback of Opt_emBR were discovered when applied on the practical dairy cattle data of bulls and cows from Holstein and Jersey breeds. The results were presented in the conference paper:

Wang T, Chen YPP, Kemper KE, Goddard ME, Hayes BJ. (2015) Opt_emBR: Computationally efficient genomic prediction and QTL mapping in multi-breed populations. Proceeding of the Association for the Advancement of Animal Breeding and Genetics, **21**: 449-452.

In **Chapter 5**, A hybrid scheme (termed HyB_BR) of the EM algorithm and a limited number of MCMC loops was developed to deal with the prediction reduction of emBayesR in Chapter 3. HyB_BR took advantage of the speed up schemes and flexibility introduced in Opt_emBayesR in Chapter 4. HyB_BR was applied on both dairy cattle genomic and phenotype data and human genomic and phenotype data to evaluate accuracy of prediction, performance for QTL mapping, and ability to identify genetic architecture. The results demonstrated that HyB_BR gave identical prediction accuracy to BayesR under MCMC, while reducing computing run time by ten fold. The paper was published as:

Wang T, Chen YPP, Bowman PJ, Goddard ME, Hayes BJ. (2016) A hybrid

expectation maximization and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping, BMC Genomics, 17:744.

In **Chapter 6**, HyB_BR was applied on a large subset of whole-genome sequence data from dairy cattle. An additional speed-up scheme was introduced into HyB_BR to further improve its computational efficiency. The multi-breed and across-breed prediction ability of HyB_BR using the whole genome sequence data was also evaluated, as well as potential of the algorithm to detect causal mutations in the sequence data. The paper was in preparation for publication.

In the final chapter (**Chapter 7**), the general discussion and overall conclusions from Chapters 3-6 were presented. Several key findings from the study were described. In addition, issues for future applications of HyB_BR were discussed.

Chapter 2 Review of the Accuracy and Computational Efficiency of Genomic Prediction Models with High Density Genotype Data

2.1 Chapter preface

Justification

This chapter reviewed the methodology and application of genomic prediction. An important conclusion from the review of the literature was that a computationally efficient algorithm was required to reduce compute time of Bayesian genomic prediction models, especially for application to whole genome sequence data, while maintaining a similar genomic prediction accuracy.

Publication status:

Accepted by the journal CAB Reviews.

Submitted as

Wang T, Chen YPP, Hayes BJ. (2016) Review of the Accuracy and Computational Efficiency of Genomic Prediction Models with High Density Genotype Data. CAB review, accepted.

Statement of contributions of joint authorship

Tingting Wang (Candidate): Read and organized several hundred research papers. Tingting Wang also carried out writing up of the paper.

Yi-Ping Phoebe Chen (Principle Supervisor): Supervised the writing of the review paper.

Ben J. Hayes (Co-Supervisor): Supervised the writing of the review paper,

assisted with the logical structure of the paper and gave great contributions regarding the revision of the paper.

This chapter was an exact copy of the version submitted to CAB review, except that the reference style, table numbers and figure numbers had been reformatted.

2.2 Abstract

The prediction of complex or quantitative traits from single nucleotide polymorphism (SNP) genotypes has transformed livestock and plant breeding, and is also playing an increasingly important role in prediction of human disease. Genomic predictions are made using a prediction equation derived from regressing the phenotypes of the individuals in a reference population on all available SNPs simultaneously. Genomic selection is then selection of animals or plants for breeding based on these genomic predictions. As the rate of genetic gain that can be achieved with genomic selection is proportional to the accuracy of the genomic predictions, a key focus is now to increase the accuracy of genomic predictions. This can be achieved by increasing the size of the reference set, using denser markers, and using appropriate genomic prediction models. A wide range of genomic prediction models have been proposed, some of which use marker selection and either linear or non-linear Bayesian models for regression. The nonlinear Bayesian models under MCMC sampling give higher accuracy of genomic prediction for some traits, particularly as marker density increases, but at the cost of high computational burden. Strategies to improve computational efficiency of the nonlinear Bayesian methods are becoming more important, as the ultimate marker density is whole genome sequence, and this is increasingly affordable in many species. In this article, we review the performance of alternative models for genomic prediction. Strategies that have been proposed to improve the computational efficiency of implementing these models are

evaluated. Finally we outline what is required to enable genomic prediction from whole genome sequence data.

Keywords: Genomic prediction, Bayesian regression models, Quantitative traits locus, Linkage disequilibrium

Review Methodology: We searched the following database: CAB Abstracts, NCBI, ISI Web of Science, and Google Scholar. In addition, we used the references from the articles obtained by this method to check for additional relevant material.

2.3 Introduction

Most of the important traits in livestock and plant breeding are complex, that is, the variation in these traits is the result of mutations at many loci (Meuwissen *et al.* 2001; Hayes *et al.* 2010; Huang *et al.* 2010; de los Campos *et al.* 2013). Two approaches have been proposed to use genetic markers such as SNPs to accelerate improvement of these complex traits in livestock and crops. Genome wide association studies (GWAS), whereby SNP effects are tested one by one for an association with the trait, followed by marker assisted selection (MAS) using most significant markers from the GWAS, have been successfully implemented for some traits in crops (Huang *et al.* 2010; Li *et al.* 2010; Jiao *et al.* 2012). GWAS have also been used to discover genes and pathways involved in human diseases (Chen *et al.* 2007; Nahar *et al.* 2007; Chen & Chen 2008; de los Campos *et al.* 2012). However, for prediction of complex traits, marker assisted selection suffered from over-estimating of the effects of the most significant SNPs (Beavis 1998; Xu 2003) as well as capturing only a small proportion of the genetic variance (Yang *et al.* 2010).

In contrast, genomic prediction models use all markers simultaneously (Meuwissen *et al.* 2001). No significance threshold is set, so provided a mutation

that affects the complex trait is in linkage disequilibrium with the markers, the variance causing the mutation in the complex trait could be captured. Even if the variation caused by a single mutation is small, by summing the effect of the markers across the genome, the genetic variance captured by the markers could be a substantial proportion of the total genetic variance (Haile-Mariam *et al.* 2013; Wood *et al.* 2014). Genomic prediction is now applied widely in livestock and crops to select individuals for breeding (VanRaden *et al.* 2009; Jannink *et al.* 2010; de los Campos *et al.* 2013; Meuwissen *et al.* 2013; Lin 2014).

There are two types of genomic prediction models that are widely used: linear models, including best linear unbiased prediction (BLUP (Meuwissen *et al.* 2001)) and nonlinear models (for example Bayesian regression models (Meuwissen *et al.* 2001; Habier *et al.* 2011; Erbe *et al.* 2012)). As one of simplest prediction models, BLUP models assume that each and every marker has a small, but non-zero effect. BLUP models are straightforward and computationally efficient to be implemented, which have meant they are popular for the practical application. However, BLUP does result in the effect of a single causative mutation being captured by the linear combination of a large number of SNPs, typically spanning reasonably large chromosome chunks. The BLUP model could therefore have reduced accuracy in multi-breed or diverse populations, where these large chromosome chunks are not shared across breeds, or when genomic predictions are made and selected for multiple generations of breeding, as recombination could break up the chromosome chunks (Kemper *et al.* 2015). In comparison with BLUP models, Bayesian models could have flexible prior assumptions. For example, BayesA assumes many SNPs have small effect and few have moderate effect with a Student t distribution; while BayesB assumes SNP effects follow a mixture of zero effects and t distributed effects. Because these models allow for a small proportion of SNPs to had large effects, the effect of a causal mutation is not “smeared” across so many SNPs (i.e. across such large chromosome chunks), as the associations might persist better across breeds and time (Kemper *et al.* 2015).

For the same reason, these methods are also attractive for quantitative trait loci (QTL, the causative mutations) mapping (Moser *et al.* 2015).

The accuracy of genomic prediction increases with marker density in some species (Yang *et al.* 2010; Kemper *et al.* 2015). The ultimate marker density is whole genome sequence, and indeed genomic predictions have been attempted with whole genome sequence data (Ober *et al.* 2012; van Binsbergen *et al.* 2015; MacLeod *et al.* 2016). The advantage of using whole genome sequence data should be that the causative mutations are actually in the data set, compared with relying on SNPs being in LD with causative mutations, so that all the genetic variance could be captured. However what is clear from studies that have used whole genome sequence data in genomic prediction to date is that simply adding millions of additional variants from the whole genome sequence data for which effects must be estimated, while using quite small reference populations and BLUP methods, does not lead to higher accuracies of genomic prediction. Rather, large reference populations are required to take advantage of the sequence data, as well as additional biological information for variant selection (MacLeod *et al.* 2016). This strategy though does lead to very large genomic data sets, with a correspondingly large computational burden for analysis, particularly for the Bayesian methods.

Here we firstly review alternative statistical models that have been proposed for genomic prediction and the algorithms that have been used to implement them. The performance of these methods in terms of accuracy of genomic prediction is compared for a range of species and traits, and with increasing marker density up to the whole genome sequence. Then, the challenge of computational efficiency with Bayesian models will be discussed, and strategies that have been proposed to improve computational efficiency for implementing these models reviewed. Finally, we describe promising future directions that should enable genomic predictions to take advantage of whole genome sequence data in reference

populations of 100s of thousands of individuals.

2.4 Models and algorithms for genomic prediction

One of the major challenges with genomic prediction is that the number of markers (*m*) is typically much larger than the number of individuals with observations for the complex trait (*n*). An overview of the different models that have been proposed to deal with this challenge, and their major characteristics, is given in Figure 2.1. Bayesian models deal with this challenge by making *a priori* assumptions about the distribution of SNP effects, and using this information, in addition to the data (phenotypes) when the SNP effects are estimated. Non-Bayesian models for genomic prediction are outlined in Appendix I - File S1, but here our focus is on the Bayesian regression models as these are the most widely used in genomic prediction, due to their flexibility and performance in terms of the prediction accuracy.

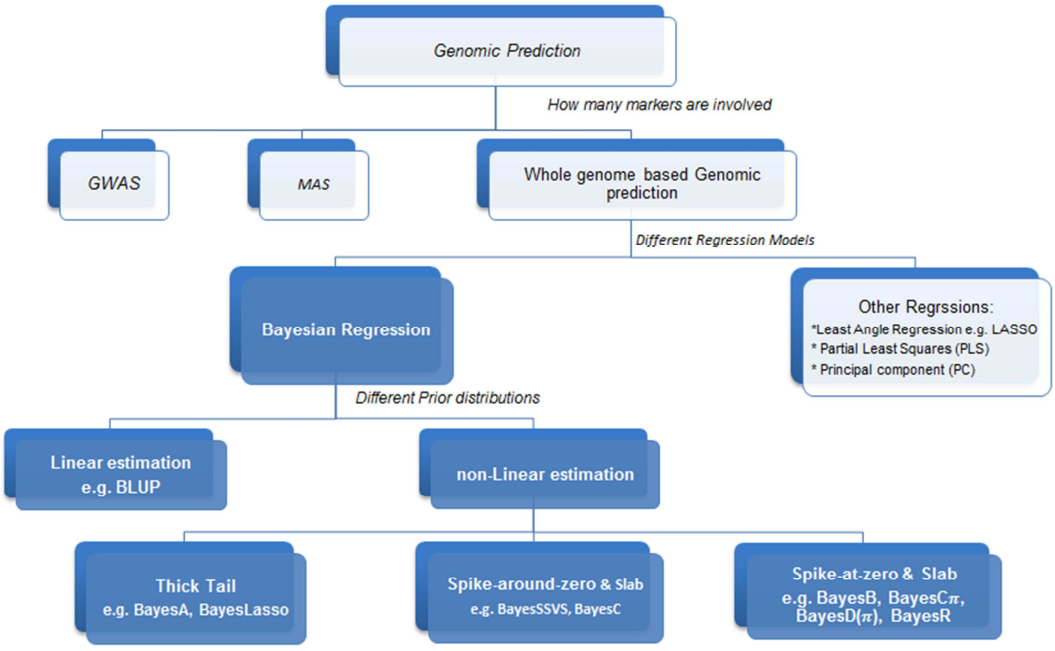


Figure 2.1. The classification of genomic prediction methods.

Under the model detailed in Appendix 1 - File S1, the true breeding value of an individual is $\mathbf{g} = \mathbf{Zu}$ with the genotype matrix $\mathbf{Z}$ and the SNP effects $\mathbf{u}$, that is

the true breeding value is the sum of the effects of the alleles at the SNPs that the individual carries. Considering only the contribution of the additive effect of the loci under the model (1), the accuracy of genomic prediction will be greatest if the estimate of the trait has the property $\hat{\mathbf{g}} = \mathrm{E}(\mathbf{g}\,|'\mathrm{data}')$**,** that is the estimates of the $\hat{\mathbf{g}}$ (the genomic estimation breeding values; **GEBV)**, match the expected values of $\mathbf{g}$ given the data (the phenotypes), as described by Goddard and Hayes (Goddard & Hayes 2007). At the level of the SNP effects, this means that the accuracy of the **GEBV** would be maximized if $\hat{\mathbf{u}} = \mathrm{E}(\mathbf{u}\,|'\mathrm{data}')$ (Goddard & Hayes 2007).

Therefore, the appropriate posterior estimation (combining information from the data and the prior) for $\hat{\mathbf{u}}$ is:

$$\hat{\mathbf{u}} = \int \mathbf{u} \times p\,('\mathrm{data}'|\mathbf{u}) \times p(\mathbf{u})du \Big/ \int p\,('\mathrm{data}'|\mathbf{u}) \times p(\mathbf{u})du \qquad (1)$$

where, $p('\mathrm{data}'|\mathbf{u})$ is the full likelihood, and $p(\mathbf{u})$ is the prior distribution for SNP effects. Equation (1) is the basis of the Bayesian regression methods to derive the posterior estimation for SNP solutions and other related parameters according to the prior density function $p(\mathbf{u})$ detailed in Appendix 1 - File S2.

One of the simplest and most widely used genomic prediction methods is best linear unbiased prediction (BLUP) (Meuwissen *et al.* 2001). BLUP assumes the effect of each marker is derived from normal distribution with the common variance across the whole markers, that is $p(\boldsymbol{u})\sim N(0, \sigma_u^2)$. In practice, the impact of this prior is that the SNP effects are shrunk (heavily) towards the mean, in proportion to the ratio of the error variance, for the complex trait phenotypes, to the variance of the normal distribution from which the SNPs are assumed to be derived (σ_u^2). When these variance components are not known (usually the case in practice!), either cross-validation, restricted maximum likelihood, or Markov Chain Monte Carlo sampling could be used to estimate them, in which case the methods are called Ridge Regression, SNP_REML and Bayesian BLUP

respectively. An equivalent model to BLUP is Genomic BLUP or GBLUP (VanRaden 2008) which predicts genetic values directly (instead of SNP effects) using a genomic relationship matrix constructed from the SNPs. GREML is the name given to the method that uses restricted maximum likelihood to estimate the variances captured by the SNPs in the genomic relationship matrix (Yang *et al.* 2010).

Due to the fact that the genomic predictions from the BLUP models are linear combinations of the SNP effects, these models are also termed linear random regression models. Before moving onto the other models, it is useful to restate the "prior" for BLUP (the same for GBLUP), that each and every SNP has a non-zero effect (regardless of how many SNPs are in the model), and these effects are very small.

Genomic Prediction methods using Bayesian regression models

Alternative prior assumptions for the distribution of SNP effects can be that there will be some SNPs with moderate effects (because these SNPs are in high linkage disequilibrium with causative mutations with moderate effects), and many SNPs with small effects, or perhaps many SNPs with zero effects, because they do not capture any of the effect of causal mutations (Meuwissen *et al.* 2001). Bayesian models with these priors include BayesA and BayesB respectively, first proposed by (Meuwissen *et al.* 2001). There are a growing number of additional Bayesian models, which differ in their prior assumptions regarding the distribution of SNP effects, Table 2.1. These models have been described collectively as the Bayesian alphabet by Gianola et al. (Gianola *et al.* 2009). To understand the ontology of Bayesian methods, it is essential to investigate the priors of these methods, and the posterior shrinkage feature inferred from these priors. In the following, we will first discuss the priors assumed by different methods that have been used widely in the literature (Table 2.1). Next, the effect of these priors on the estimates of marker effects will be demonstrated in a test data set.

Table 2.1. Summary of prior distribution feature proposed for different Genomic prediction models.

Model	**Hierarchical Prior Density**			**Prior Density conditional on the parameters** $p(\mathrm{u_i}\|\omega,\cdots)$	**Terms & description**
	Prior of $\mathbf{u_i}$ conditional on the variance $\boldsymbol{\sigma^2_{u_i}}$ ($N(\mathrm{u_i}\|0,\sigma^2_{\mathrm{u_i}})$)	**Prior of $\boldsymbol{\sigma^2_{u_i}}$ conditional on hyper-parameters ω ($p(\sigma^2_{\mathrm{u_i}}\|\omega)$)**	**Hyper-parameters ($p(\omega)$)**		
Gaussian	$u_i \sim N(0,\sigma_u^2)$	σ_u^2 was fixed	-	$u_i \sim N(0,\sigma_u^2)$	Following uniform normal distribution e.g. SNP-BLUP (Meuwissen *et al.* 2001), GBLUP (VanRaden 2008)
Thick tail	$u_i \sim N(0,\sigma_{u_i}^2)$	$\sigma_{u_i}^2 \sim \chi^{-2}(v,S)$	v, S were fixed	$u_i \sim t(0,v,S)$	Following student distribution e.g. BayesA (Meuwissen *et al.* 2001)
		$\sigma_{u_i}^2 \sim DE(0,\lambda)$	λ was fixed	$u_i \sim DE(0,\lambda)$	Following Double exponential distribution (DE) e.g. BayesLasso (Tibshirani 1996; Park & Casella 2008)
Spike-around -zero & Slab	$u_i \sim \pi N(0,\sigma_u^2+\sigma_b^2) + (1-\pi)N(0,\sigma_b^2)$	σ_u^2, σ_b^2 were fixed;	π: uniform prior	$u_i \sim \pi N(0,\sigma_u^2+\sigma_b^2) + (1-\pi)N(0,\sigma_b^2)$	The mixture of two normal distributions e.g. BSLMM (Zhou *et al.* 2013b)
	$u_i \sim \pi N(0,\sigma_{u_i}^2)+(1-$	$\sigma_{u_i}^2 \sim \chi^{-2}(v,S)$	$\pi \sim uniform(0,1)$	$u_i \sim \pi t(0,v,S)+(1-$	The mixture of t distributions e.g.

	$\pi)N\left(0,0.01\sigma_{u_i}^2\right)$		v,S were fixed	$\pi)t(0,v,0.01S)$	BayesSSVS (Verbyla *et al.* 2009; Verbyla *et al.* 2010)
Spike-at-zero & Slabs	$u_i \sim \mathrm{N}(0,\sigma_{u_i}^2)$	$\sigma_{u_i}^2 \sim \pi\left(\sigma_{u_i}^2 = 0\right) + (1-\pi)\chi^{-2}(v,S)$	$\pi \sim uniform(0,1)$ v,S were fixed	$u_i \sim \pi(u_i = 0)$ $+(1-\pi)\mathrm{t}(0,v,S)$	The mixture of point mass at zero and t distribution e.g. BayesB (Meuwissen *et al.* 2001)
	$u_i \sim \pi(u_i = 0) + (1-\pi)N\left(0,\sigma_{u_i}^2\right)$	$\sigma_{u_i}^2 \sim \chi^{-2}(v,S)$	$\pi \sim uniform(0,1)$ v,S were fixed		The mixture of point mass at zero and t distribution but with the same variance e.g. BayesC(π) (Habier *et al.* 2011)
		$\sigma_{u_i}^2 \sim \chi^{-2}(v,S)$	$S \sim gamma(1,1)$ $\pi \sim uniform(0,1)$		The mixture of point mass at zero and t distribution but with the same variance e.g. BayesD (Habier *et al.* 2011), BayesDπ(Habier *et al.* 2011)
	$u_i \sim \pi_1 N(0,0.0001\sigma_u^2) + \pi_2 N(0,0.001\sigma_u^2) + \pi_3 N(0,0.01\sigma_u^2) + \pi_4(u_i = 0)$	σ_u^2 was fixed	$\sum_{i=1}^{4} \pi_i = 1$ $\pi_i \sim Dirichlet(\alpha)$	$u_i \sim \pi_1 N(0,0.0001\sigma_u^2) + \pi_2 N(0,0.001\sigma_u^2) + \pi_3 N(0,0.01\sigma_u^2) + \pi_4(u_i = 0)$	The mixture of point mass at zero and three normal distributions e.g. BayesR (Erbe *et al.* 2012; Moser *et al.* 2015), BayesRC (MacLeod *et al.* 2016)

Models with ***thick-tailed priors*** assume all SNPs had effects, but these effects follow thick tail distributions. Compared with a normal prior assumption for the SNP effects (black dotted line), the family of thick-tailed priors assumes a large proportion of SNPs with effects close to zero (regressing these SNPs effects closer to zero than BLUP), and a small proportion of SNP with larger effects, resulting in the thick tail distribution (red curve in Figure 2.2). There are several models with thick-tailed priors:

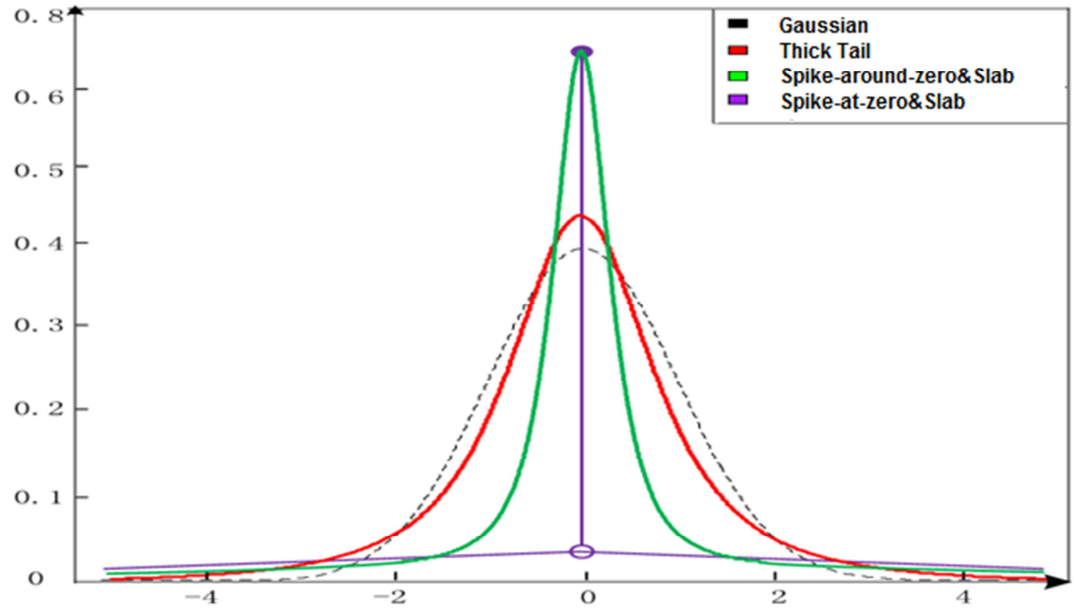


Figure 2.2. The prior density functions for BLUP and three different Bayesian models.

BayesA (Meuwissen *et al.* 2001) assumes a *t*-distribution at the level of SNP effects. Note that the strategy to make this method computationally efficient is usually achieved by allowing each SNP to have its own normal distribution (also called a SNP specific variance), and the distribution of these variances is assigned an inverted Chi-square distribution (Meuwissen *et al.* 2001), Appendix I - File S3. Many other Bayesian models also use this computational trick.

BayesLasso assumes a double exponential distribution at the level of SNP

effects, which results in greater shrinkage than in BayesA (Tibshirani 1996; Park & Casella 2008; de los Campos *et al.* 2009).

***Spike-around-zero & Slab* models** (green curve in Figure 2.2) use mixture model of SNP effects where a proportion of SNPs are derived from a distribution with almost zero variance, such that these SNP effects form a “spike around zero” , while a smaller proportion of SNPs are derived from distribution with larger variance (the “slab”). In comparison with thick-tail priors, Spike-around-zero & Slabs will regress more SNPs with small effects near zero resulting in distribution with sharper peak (spike) around zero than thick tail model. The green curve in Figure 2.2 demonstrates this. **BayesSSVS** (Verbyla *et al.* 2009; Verbyla *et al.* 2010), and **BSLMM** (Zhou *et al.* 2013b) are two popular prediction methods (Table 2.1).

BayesSSVS (Verbyla *et al.* 2009; Verbyla *et al.* 2010) implements the stochastic search variable selection (hence SSVS) scheme for markers with large effects. BayesSSVS can be considered as an extension of BayesA model, with a mixture model of two t-distributions with different variances, one 1/100 of the other.

BSLMM (Zhou *et al.* 2013b) extends the BLUP model to have one normal distribution from which many SNP effects are derived with a variance very close to zero, such that the SNP effects are very close to zero, and another normal distribution with larger variance, resulting in larger estimates of effects for the SNPs assigned to this distribution. BSLMM does not set *a priori* for either the variances of the normal distributions or the proportion of SNPs. Instead they are inferred from the data and hyper-parameters (Table 2.1), which allows this method to adapt to different underlying genetic architectures of complex traits.

Both *thick-tail* and *Spike-around-zero & Slabs* allow a large number of SNPs to have very small effects but they do not actually remove SNPs from the model.

Spike-at-zero & Slab models (purple curve in Figure 2.2) assume a large proportion of SNPs are actually zero (removed from the model), while the rest of the SNPs follow a prior distribution (e.g. t distributions or the mixtures of normal priors). Spike-at-zero & slab models include BayesB (Meuwissen *et al.* 2001), BayesC**(π)** (Habier *et al.* 2011), and BayesR (Erbe *et al.* 2012).

In the **BayesB** models (Meuwissen *et al.* 2001), a fraction (π_1) of SNPs have no effects while (1-π_1) of the SNPs share the same prior assumption as BayesA (t distribution of effects). The value of π_1 is set *a priori*. The BayesB model (as well as BayesA, LASSO, and SSVS) has been criticized by Gianola (Gianola *et al.* 2009; Gianola 2013): the strong priors in these models (the degrees of freedom for the t-distribution, or shape parameter for the BayesianLASSO) mean that when these models are applied, the posterior distribution of SNP effects is largely driven by the prior rather than the data. The reason for this is that the SNP specific variances are often estimated with more weights from the prior (the weights specified by the degrees of freedom) than from the data. To deal with these problems, Gianola (Gianola 2013) have proposed two solutions: 1) Increase the freedom by grouping the markers into different sets. 2) Define the hyper-parameters related to the variance as unknown, but derived in turn from some prior distribution.

BayesC(π) (Habier *et al.* 2011) is a method that adopts both these strategies. BayesC(π) assumes SNP effects are either zero, or follow a normal distribution, with the same variance across non-zero SNPs (i.e. a BLUP model for SNPs in the model). This assumption means the variance of the normal distribution is

estimated with large degrees of freedom from the data (with degrees of freedom the number of SNPs in the normal distrubtion-1). In contrast to BayesB, the value of π is not pre-set, rather it is estimated from the data with *a prior* assumption, Table 2.1.

Another Spike-at-zero & slab model is **BayesR** (Erbe *et al.* 2012), which also adopts both strategies suggested by Gianola (Gianola *et al.* 2009; Gianola 2013). BayesR defines a large proportion (π_1) of SNPs with zero effect and the others follows a mixture of three normal distributions (π_2, π_3, and π_4) with a series of variances $0.0001\sigma_g^2$, $0.001\sigma_g^2$, and $0.01\sigma_g^2$ (shown in Table 2.1), where σ_g^2 is the genetic variance of the trait. This allows BayesR the flexibility with respect to genetic architecture. The mixing proportion parameters π_1, π_2, π_3, and π_4 are assumed to follow Dirichlet distribution. The only problem for BayesR is that the genetic variance σ_g^2 is pre-specified according to the prior empirical knowledge, which might affect the inference of SNP effects. To deal with such problem, Moser et al. (Moser *et al.* 2015) modifies BayesR model by treating the variance σ_u^2 as unknown. The estimate of the variance is then made from the data and assuming a prior distribution for this parameter, Table 2.1.

Since *Spike-at-zero & Slabs* set a large proportion of SNPs to zero, while the remaining SNPs could have up to moderate to large effects, such priors produce SNP effect distributed with a high peak at zero point and a fat tail (the purple curve in Figure 2.2).

In Figure 2.3, the effect of the priors used by BayesR, BayesB, BayesA, and BLUP on the estimates (posterior estimates) of SNP effects is demonstrated in an example data set. In practice, the actual degree of shrinkage would depend on the size of the data and the genetic architecture of the trait (Gianola 2013). The BLUP

prior (single normal distribution) shrinks all SNP effects heavily towards zero (black line); the thick tail prior of BayesA shrinks some SNPs near zero more heavily than the SNP with moderate to large effects (green curve); BayesB shrinks many SNP to zero, while some SNPs have moderate to large effects (red curve); and similar to BayesB, BayesR shrinks many SNPs heavily towards zero but the mixture of three normal distributions rather than a single *t* distribution results in large effects being shrunk to a greater degree than in BayesB (blue curve).

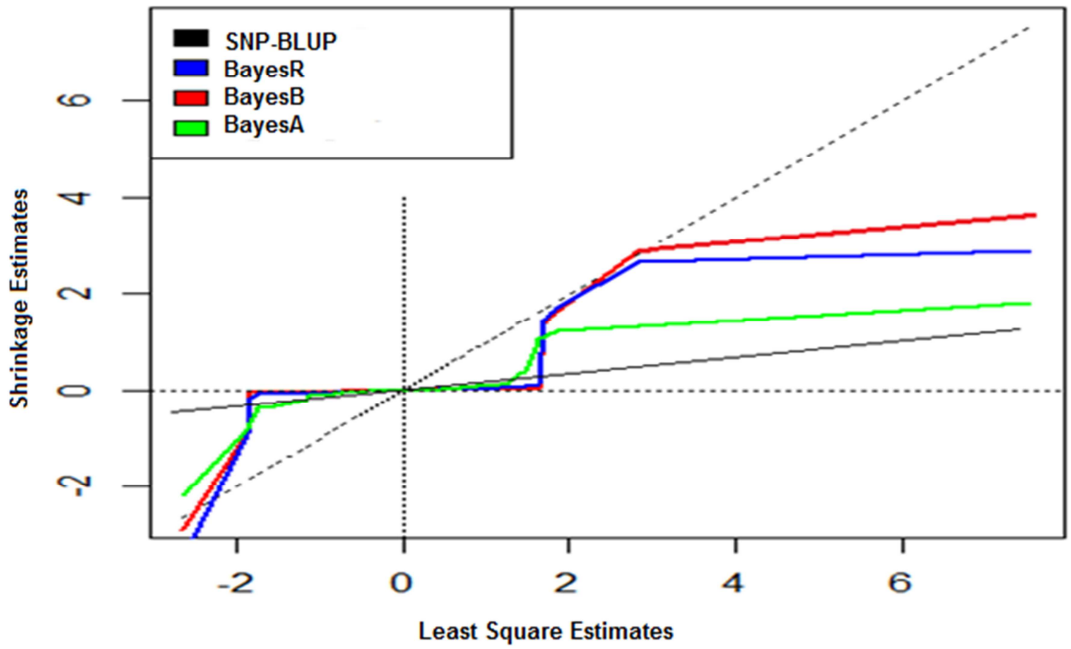


Figure 2.3. Estimated SNP effects from linear model (SNP-BLUP), thick tail model (BayesA), and mixture model (BayesB and BayesR).

2.5 Comparison of the accuracy of genomic prediction methods in simulated and real data

The genomic prediction models described above have been widely implemented in both simulated and empirical data sets. So we could use the results of these studies to investigate two questions: 1) does any prediction model outperform the

others, and what is the impact of genetic architecture on the relative performance of the models? 2) does increasing density of genomic data improve the accuracy of predictions? Here, the performance is defined in terms of prediction accuracy, which is the correlation of Genomic Estimated Breeding Value (GEBV) and True Breeding Value (TBV). In the cases of field data, where the TBV is not known, the observed phenotype has been used as the proxy for TBV to calculate the prediction accuracy.

As the genetic architecture of the simulated data is well defined, we would start with an investigation of relative performance of the models with a range of underlying genetic architectures. The impact of the genotype data with different densities on prediction accuracy will be discussed as well. Then, some theory regarding the accuracy of prediction models will be applied to explain the performance of prediction models on the simulation data. Finally, in empirical data sets from animal/plant breeding, linear and nonlinear models will be evaluated, with both high-density SNP and sequence data.

Genomic prediction in simulated SNP panels and sequence data

In simulated data, the genetic architecture of the simulated complex trait is defined by the number of QTLs, the distribution of QTLs effects, and the minor allele frequency (MAF) of QTLs (Harris & Johnson 2010; Meuwissen & Goddard 2010; Clark *et al.* 2011; Druet *et al.* 2014; MacLeod *et al.*).

Table 2.2. The impact of increasing density and different model on prediction accuracy for simulation data

Marker Size	No. of QTLs				Distribution of QTLs	MAF of QTLs	Results		Reference
	Small (≤100)	Moderate (100~1,000)	Large (1,000~10,000)	Infinitesimal (>10,000)			Linear VS. Non-linear	Increasing density	
One genome	30	-	-	-	Normal	≥0.05	Nonlinear had 40% increase	Sequence data had 5-10% advantage	(Meuwissen & Goddard 2010)
20K~1000K	-	-	1,500	-	Gamma	≥0.05	Similar	Increasing density had 1% increase	(Harris & Johnson 2010)
50K, 500K	-	-	10,000	-	Heavy tail	≥0.05	Nonlinear had 1.6% higher reliability	500K had 1.6% advantage than 50K in reliability	(VanRaden *et al.* 2011)
5K, 60K, sequence	100	-	-	-	Gamma	≥0.05	Nonlinear had 26% increase	Sequence had 9% increase than 60K.	(Clark *et al.* 2011)
5K, 60K, sequence	-	1,000	-	-	Gamma	≥0.05	Nonlinear had 6% increase	Sequence had 7% increase than 60K.	

5K, 60K, sequence	-	-	10,000	Infinitesimal	Gamma	≥0.05	Similar	Similar	
5K, 60K, sequence	100	1,000	-	-	Gamma	<0.01 (Rare model)	Nonlinear had 26% increase	-	
60K, 600K	15~50	-	-	-	Normal	≥0.05	Nonlinear had up to 3.6% increase	Sequence had up to 11.8% increase than other HD SNP chips	(MacLeod *et al.*)
50K, Sequence	-	1,000	-	-	Normal	≥0.05 (Neutral distribution)	-	Sequence had 1.5% advantage than 50K	(Druet *et al.* 2014)
50K, Sequence	-	1,000	-	-	Normal	<0.01 (Rare model)	-	Sequence had up to 30% increase than HD	

A large number of studies have used simulated data to assess performance of genomic prediction methods, Table 2.2. Meuwissen and Goddard (Meuwissen & Goddard 2010) simulated whole genome sequence data, with a larger number of SNPs and 3 or 30 causal mutations (QTLs) affecting a complex trait, on one chromosome. In comparison with GBLUP, BayesB gave up to a 40% accuracy improvement. They also found that prediction accuracy was improved as marker density increased. Macleod et al. (MacLeod *et al.*) used a similar approach but simulated a much larger number of QTLs (3000). GBLUP and BayesR were two genomic prediction models that were compared with marker density from 60,000, 600,000, up to whole genome sequence. The results showed there was up to 11.8% advantage from sequence data compared with other SNP panels, and BayesR could take a better advantage of sequence data than GBLUP.

In contrast, other authors have found limited advantage from increasing marker density. Harris and Johnson (Harris & Johnson 2010) reported very minimal gain (1% improvement) by increasing the density of simulated SNP panel (20K, 100K, 500K, 1000K), while different prediction models (linear and nonlinear methods) had very similar prediction accuracy on the same dataset. There were a very large number of QTLs in their study.

Clark et al. (Clark *et al.* 2011) and Druet et al. (Druet *et al.* 2014) explored the relationship between genomic prediction accuracy and genetic architecture in more details. Clark et al. simulated a range of different genetic architectures including a Rare QTL model (with QTL at low MAF), Common QTL model (with QTL at the MAF expected under a neutral model), and a pseudo-infinitesimal model. In the pseudo-infinitesimal model, there were a large number of QTLs (>10,000) with very small individual effects. When BayesB and GBLUP were implemented in the above data sets, BayesB gave more accurate GEBV under

the Rare QTLs model. However, for the pseudo-infinitesimal model, the accuracy of BayesB was very close to GBLUP. Clark et al. also investigated the effect of marker density, simulating full genome sequence with ~1.67 million of SNP, as well as 60K, and 5K SNP chips. The results showed that the advantage of sequence data over 60K in terms of genomic prediction accuracy was 5% to 15%, depending on the MAF of the QTL. Similar to the results from Clark et al., Druet et al. reported 1.5% ~30% advantage of using sequence data over 50K SNP chip data with a common QTL model and Rare QTL model respectively.

Considering the above results, is there some way of predicting which method could perform best on a particular data set? Daetwyler et al. (Daetwyler *et al.* 2010) and Goddard (Goddard 2009) presented an equation to predict the accuracy:

$$R = \sqrt{\frac{N_p h^2}{N_p h^2 + T}} \qquad (2)$$

where, R is the accuracy; N_p is the size of reference set; h^2 is the heritability; And T is the number of loci affecting the trait. In the linear models (BLUP), the appropriate value of T *is equal to* M_e (the number of effective independent chromosome segments in the population), as all the SNPs (which track the chromosome segments) are in the model. The number of effective chromosome segments is given by $M_e = \frac{2N_e L}{log(2N_e)}$, with N_e= the effective population size, L= the size of the genome in Morgans (For Holstein cattle, where N_e is 100 and the length of the genome is 30 Morgans, the M_e is 2607). For the nonlinear models (e.g. BayesB and BayesR), which can set SNP effects to zero, the appropriate $T = \min(M_e, N_{QTL})$ (where N_{QTL} is the number of QTLs). That is, the number of QTL can actually be less than the number of effective chromosome segments in the population. The formula (2) tells us that:

1) Prediction accuracy will increase as a large number of records N_p are used,

2) For the same number of records, prediction accuracy will be higher for traits with higher heritability,
3) If the number of QTL is less than the effective number of chromosome segments, prediction accuracy will be higher with BayesB and BayesR.

A further point to make is that the advantage of increasing marker density will be greater in populations with large N_e, as in those populations a higher density of markers is required to ensure at least one marker is in LD with at least one QTL.

Point 3) above explains why Meuwissen and Goddard (Meuwissen & Goddard 2010) has observed such a large advantage of BayesB over GBLUP – with 3 or even 30 QTL, the number of QTL is much smaller than the number of chromosome segments. In contrast, Harris and Johnson (Harris & Johnson 2010) simulated a much larger number of QTLs, and with $N_{QTL} \sim M_e$, so there was less advantage of BayesB. They also simulated smaller Ne, so there was less advantage of increased marker density.

To summarize non-linear Bayesian models have an advantage over BLUP models (linear models) when there are a small number of major QTLs (e.g. 30~100 in (Meuwissen & Goddard 2010; Clark *et al.* 2011; MacLeod *et al.*)). But the advantage diminishes gradually with increasing number of QTLs (Clark *et al.* 2011). In infinitesimal model with more than 10,000 QTLs, nonlinear Bayesian models have minor advantage over BLUP (Harris & Johnson 2010; Clark *et al.* 2011; VanRaden *et al.* 2011). Further, the number of QTLs with low MAF determines the impact of the density of genotype data on the accuracy. When the QTLs are at reasonable MAF, high-density (600K) data is enough in most livestock species to ensure adequate LD between the SNP and QTL for accurate genomic prediction. With common SNP architectures, the gain from using sequence data over 600K is very small as demonstrated in (Harris & Johnson

2010; Druet *et al.* 2014; MacLeod *et al.*). But for Rare QTLs models in which there exist a number of QTLs with small MAF (<0.01), there could be a substantial advantage by using sequence data (Meuwissen & Goddard 2010). An outstanding question is of course how many QTLs do actually affect complex traits, and what is the allele frequency spectrum of these QTLs. As shown below, for some traits the non-linear Bayesian methods do have some advantages over BLUP, while for others the advantage is minimal – this would imply that genetic architecture varies quite markedly between traits.

Genomic predictions in livestock from empirical HD density SNP chips to sequence data

In this section, we first review the overall application of genomic prediction in a range of livestock data sets in Table 2.3. Then we focus on comparing the performance of linear (BLUP) and nonlinear models (BayesB, BayesCπ and BayesR) in terms of the accuracy of GEBV. Afterwards, we investigate the advantage of whole genome sequence data over SNP chips.

Table 2.3. Genomic prediction on a range of HD SNP chips for livestock.

Specie	Population size	No. Of Marker	Traits	Methods	Results	Reference
Cattle	2,937 Norwegian Red bulls	25K/54K 777K	22 production and functional traits	BLUP	777K HD data had just Marginal increase for prediction accuracy than 54K medium density data	(Solberg *et al.* 2011)
Cattle	33,414 Holsteins	50K 500K	-	LinearNonlinear	1)The reliability gain by increasing the number of markers to 500K was only 1.6%; 2) Nonlinear model had 1.5~1.6% reliability increase than linear	(VanRaden *et al.* 2011)
Cattle	2,000 Holstein	624,930 36,673	Residual Feed Intake 250-day Body weight	GBLUP, BayesA, BayesSSVS	Bayesian methods had up to 10% advantage than GBLUP in Australia	(Pryce *et al.* 2012)
Cattle	4,539 Holstein 4,403 RDC	777K 54K	Protein, Fertility, Udder health	GBLUP, Bayesian mixture	1) 0.5% ~1% reliability improvement on 777K HD SNP than that based on the 54K data on two breeds; 2) Bayesian mixture had 0.5% higher reliability than GBLUP in Holstein, but similar in RDC.	(Su *et al.* 2012)

Cattle	996 Holstein, 93 Jersey	58,532 624,213	Milk, Fat, Protein	GBLUP_mod , BayesR, BayesA	1) BayesA≅BayesR BayesA(R) >GBLUP ; 2) Compared with 50K chips, The improvement of accuracy on 600K SNPs was very limited.	(Erbe *et al.* 2012)
Cattle	10,181 Holstein	729,068	Residual Feed Intake Carcass and meat quality	GWAS, GBLUP, BayesR	BayesR was Up to 0.04 greater that GBLUP	(Bolormaa *et al.* 2013)
Cattle	17,925 Holstein	632,003	Milk production in dairy cattle	Multibreed GWAS	identified and confirmed a large number of QTL with more accurate locations information.	(Raven *et al.* 2014)
Cattle	161,341 Holstein	50K 300K	28 traits including yield traits and functional traits	GBLUP, fastBayesA	1) nonlinear model only improved an average 0.8% reliability of prediction than GBLUP; 2) Prediction Reliability on 50K SNP chip was only 0.2% less than the one on 800K SNP.	(VanRaden *et al.* 2013)
Cattle	11,527 Holstein, 4,687 Jersey	632,002	5 milk yield traits Composition traits across breed	GBLUP, BayesR	BayesR had the average of 7% increase compared with GBLUP	(Kemper *et al.* 2015)
Sheep	2,812	43,929	Milk yield, fat%, Somatic cell count	BayesCπ,	BayeCπ had around 2%~5%	(Duchemin *et*

			(SCC)	GBLUP, PLS	accuracy advantage than others	*al.* 2012)
Sheep	8075~10,772 sheep	48,599	Carcass and novel meat quality, Greasy fleece weight, Eye muscle depth	GBLUP, BayesSSVS, BayesR	BayesSSVS had 20% accuracy advantage than GBLUP in crossbreeding population.	(Daetwyler *et al.* 2012a; Daetwyler *et al.* 2012b)
Pig	351	34,000~40,000 from PorcineSNP60	Four US breeds	-	High LD gave the promise of the probability of genomic selection	(Badke *et al.* 2012)
Pig	3534	52,842	Five purebred traits	BayesB, ssGBLUP	BayesB≅GBLUP	(Cleveland *et al.* 2012)
Pig	4,763	450, 3k, 6k from PorcineSNP60	Total number born	ssGBLUP	Strategies to optimize development of low-density panels could improve GEBV accuracy	(Cleveland & Hickey 2013)
Pig	8,187	38,453	Growth rate Lean meat percentage Weight at three weeks of age, number of teat	GBLUP	GBLUP yielded higher prediction accuracies than based on pedigree.	(Meuwissen *et al.* 2014)
Chicken	2,708	23,356	egg production, egg weight, egg color, shell strength, age at sexual maturity, body weight,	BLUP family, BayesC	For the accuracy of genomic prediction, BLUP ≅BayesC	(Wolc *et al.* 2011)

			albumen height, yolk weight			
Chicken	1,351	580,954	Three traits, body weight at 35 days, ultrasound area of breast meat and hen house production	RKHS	Whole-genome based genomic selections were the promising tool for the genomic prediction of complex traits.	(Kranis *et al.* 2013), (Morota *et al.* 2014)

*The population size is the total number of animals including reference sets and validation sets, with HD SNP chips (actual genotyped or imputed data.

The nonlinear models outperform BLUP for many, but not all traits in beef and dairy cattle, Table 2.4. The advantage is generally more apparent for traits with higher heritability including fat%, milk yield, peak shear force measured in longissimus Lumborum muscle (LLPF) related to RFI, and also post weaning weight (PWIGF). The advantage is also greater for traits with known genes of large effect, including fat% in milk productions of dairy cattle, where a mutation in the DGAT1 gene explains up to 30% of the variance (Grisart *et al.* 2002), and traits that could be assumed to have a simpler genetic architecture, such as insulin like growth factor 1 (IGF1) levels. For these traits, the nonlinear models have up to 20% higher GEBV accuracies. In contrast, for traits with low heritability such as fertility, nonlinear methods just have a minor advantage ($\leq 1\%$) over linear models (Su *et al.* 2012; VanRaden *et al.* 2013), Table 2.4.

In general, the advantage of non-linear models also becomes clearer as the number of phenotypes increases. For example, for the protein production in dairy cattle, nonlinear models have very similar prediction accuracy as GBLUP when there are relatively small number of individuals (Verbyla *et al.* 2010). However, once the reference size increases to 16,000 individuals, the advantage of BayesR over GBLUP is clearly observed by (Kemper *et al.* 2015). When the reference population size is increased by combining animals of multiple breeds (Erbe *et al.* 2012; Bolormaa *et al.* 2013; Kemper *et al.* 2015), the advantage of nonlinear models over BLUP model is even more obvious, Table 2.4. This is likely a reflection of the fact that the linear combination of SNP effects from the BLUP model to capture each mutation effect means that the causative mutations often “smear” across many markers encompassing long chromosome segments. Such association might be broken down due to the recombination in less closely related individuals from different breeds (Kemper *et al.* 2015), which therefore makes BLUP model perform worse than BayesR for multi breeds prediction.

Table 2.4. Prediction ability of Linear and Nonlinear model on a range of important traits with different heritability in Cattle.

Breed	**Markers**	**Traits**	h^2	**Results**			**Reference**
				The accuracy advantage of nonlinear over linear	+Multibreeds	+Poly	
Beef Cattle (Holstein)	600K	Residual Feed Intake	0.22	+0.12	-	-	(Pryce *et al.* 2012)
		250-day Body weight	0.28	+0.03	-	-	
Beef Cattle (Bos Taurus, Bos indicus & composite)	729K	Residual Feed Intake	0.36~0.56	+0.04 (LLPF +0.16)	+0.04	-	(Bolormaa *et al.* 2013)
		Carcass and meat quality	0.23~0.52	+0.01	+0.04	-	
		Growth traits	0.24~0.53	(PWIGF, EIGF +0.23)	+0.04	-	
Dairy Cattle (Holstein & Jersey)	600K	Milk, Fat, Protein	0.33	+0.05	+0.03	-	(Erbe *et al.* 2012)
Dairy Cattle (Nordic Holstein & Red Dairy Cattle)	777K	Protein	0.39	+1.1% in reliability	-	-	(Su *et al.* 2012)
		Fertility	0.04	+0.3% in reliability	-	-	
		Udder health	0.04	+0.6% in reliability	-	-	

Dairy Cattle (Holstein & Brown Swiss from four countries)	777K	Milk, Fat, Protein, Productive life, calving	0.3 0.08~0.09	+0.3%~1.9% in reliability	+2.6%~3.2%	-	(Vanraden *et al.* 2012)
		Stature	0.45	+1.3%			
		SCC	0.11	+0.4%			
Dairy Cattle (Holstein from five countries)	300K	Fat%, Protein%	0.55	+3%~6% in reliability	-	-	(VanRaden *et al.* 2013)
		Milk, Fat, body depth, Productive life, calving	0.3 0.08~0.09	+1%~3% in reliability	-	-	
		Other functional traits	0.04~0.20	0~1% or minus in reliability	-	-	
Dairy Cattle (Holstein, Montbeliarde, Normande)	777K	Milk production trait, SCS	0.3	+0.193	+0.016	-	(Hozé *et al.* 2014)
Dairy Cattle (Hostein & Jersey)	600K	Milk production	0.33	+0.16(fat%)	+0.08	+0.03	(Kemper *et al.* 2015)
		Stature	0.45	+0.01	+0.01	-	
		Fertility, Survival	0.03	+0.01	+0.03	-	
Dairy Cattle (Holstein)	Sequence	Milk, fat, protein	0.33	+0.05	-	-	(MacLeod *et al.* 2016)

However, when the size of markers is increased to a very large number (e.g. 30 millions), the performance between Bayesian and BLUP model might need further investigation.

The difference between sequence data and SNP array data is twofold: (1) there are a larger number of rare variants in the sequence data, as the SNPs on the SNP arrays are nearly always selected because they have high MAF; 2) The actual causative mutations (QTLs) are in the sequence data, which means the LD between SNPs and QTLs is now less important. There are relatively few examples of the use of whole genome sequence data in genomic predictions. Ober et al. (Ober *et al.* 2012) implemented both GBLUP and BayesB on 157 inbred lines of *Drosophila melanogaster* with ~2.5 million of SNPs. The *Drosophila* lines had a range of phenotypes including startle response and resistance to desiccation. The results showed that the accuracy gain from using sequence data compared with 150K HD SNP chip was very limited, although it did have to be pointed out that the reference set was very small in size. Both of two other studies, in dairy cattle (van Binsbergen *et al.* 2015), and chickens (Heidaritabar *et al.* 2016), using 12 million and 4.6 million sequence variants respectively, reported very little advantage from using sequence data. Particularly in the chicken example, the number of individuals with whole genome sequence was very small (24 animals) (Heidaritabar *et al.* 2016), which might have resulted in poor imputation of sequence into the reference population. The study in dairy cattle was based on more animals with imputed sequence data (sequence data from (Daetwyler *et al.* 2014)) and a much larger reference set, with the BayesSSVS method used to predict SNP effects. The result for this study (van Binsbergen *et al.* 2015) indicated that to take advantage of the sequence data, additional biology information was needed to identify a more predictive subset, prior to running the genomic prediction methods.

In contrary, two other studies have reported an advantage of using sequencing data in genomic predictions, both in dairy cattle. Brondum et al. (Brøndum *et al.* 2015) used sequence data from the 1000 bull genomes project (Daetwyler *et al.* 2014) to impute sequence data into a large multi-breed reference population genotyped with high density SNP. These authors then conducted GWAS for each target trait, to identify putative causative mutations. The putative causative mutations (8-10 per trait) were then added to a 54K SNP panel for genomic predictions. Up to a 4% improvement in accuracy was achieved, over the 54K SNP panel alone.

Macleod et al. (MacLeod *et al.* 2016) applied a model, which extended BayesR to take in additional biological information (BayesRC) model to derive genomic predictions from imputed sequence data in dairy cattle. Gene expression information from mammary gland was used to classify sequence variants. In total, the genomic data consisted of 16,214 bulls and cows from two breeds (Holstein and Jersey), with imputed sequence data for 1,674,245 sequence variants (imputed from the 1000 bull genomes project (Daetwyler *et al.* 2014)). The results showed a 2-5% increase in accuracy of prediction as a result of using sequence data, depending on traits. Gains in accuracy were even larger for across breed predictions (where the predicted breed was not in the reference set).

2.6 Implementation of Bayesian regression models and computational performance

As the size of genomic data increases dramatically, the running time of genomic prediction algorithms has aroused attentions as well. In the following, we will discuss the implementation of these algorithms, and then the computational performance of Bayesian regression methods will be evaluated.

To date, Bayesian models coupled with the random walking scheme of Markov Chain Monte Carlo (MCMC) have been investigated to be the perfect match to conduct posterior estimation for parameters (e.g. SNP effects) with no closed form. Two typical MCMC algorithms termed Metropolis-Hasting algorithm (MH) and Gibbs sampling are implemented for genomic prediction. As an easier and faster implementation scheme, Gibbs sampling is usually used when all the parameters can be sampled with the known distributions. Compared with Gibbs sampling, MH aims at drawing random samples from a probability distribution for which direct sampling is difficult. Therefore, since the effects and other parameters defined by BayesA, BayesC, and BayesR follow the forms of a known distribution, Gibbs sampling can be implemented on BayesA, BayesC and BayesR. On the contrary, due to unknown posteriors for SNP effects and other random parameters, BayesB, BayesD, and BayesDπ apply MH algorithm. When comparing the computational time between BayesA, B, C, D(π) and R, we can easily conclude that both BayesC and BayesR are faster than others. The assumption for the variance decides this. In detail, the marker-specific variances of BayesA, BayesB, and BayesD(π) require to be sampled repeatedly for each iteration. However, BayesC defines the common variance across all the SNPs, which therefore needs to be updated once; BayesR selects one out of four variances for each SNP instead of sampling them. To our best knowledge, as much easier and faster implement algorithms, BayesC and BayesR with Gibbs sampling have become more popular than BayesB.

The time complexity of MCMC methods is $O(mn)$ for each MCMC cycles. Usually, to obtain the best solutions for effects, the choice of the number of the cycles is dependent on the size of data. On 800K SNP chips with 16,000 individuals, around 40,000 iterations are required with first 20,000 loops removed out (Kemper *et al.* 2015; MacLeod *et al.* 2016). When faced up with millions of markers, MCMC scheme of Bayesian regression models could lead to huge

computational burden, which is the main limitation of Bayesian prediction methods to be applied to practical applications.

2.7 Improving the computationally efficiency of implementing Bayesian regression models

The nonlinear Bayesian models are attractive, resulting in higher prediction accuracy for some traits. However, these models are usually implemented with Markov Chain Monte Carlo (MCMC) sampling to obtain posterior estimates of the SNP effects. This is computationally intensive, and would result in extremely long run times if implemented with large reference populations with imputed whole genome sequence data.

A series of fast versions of the Bayesian methods have been developed, all of which introduce more efficient algorithms to replace MCMC sampling. Heuristic algorithms, including Iterative Conditional Expectation (ICE) and Expectation-Maximization (EM) are the most popular substitutions. These proposals have the same hierarchical models of Bayesian methods (prior assumptions, etc.) but with different implementations. Figure 2.4 lists seven fast algorithms that have been described including nonlinear BayesA(B) (VanRaden 2008), fastBayesB (Meuwissen *et al.* 2009), emBayesB (Shepherd *et al.* 2010), and fastBayesA (Sun *et al.* 2012) (and see Appendix I - File S4 for more detail).

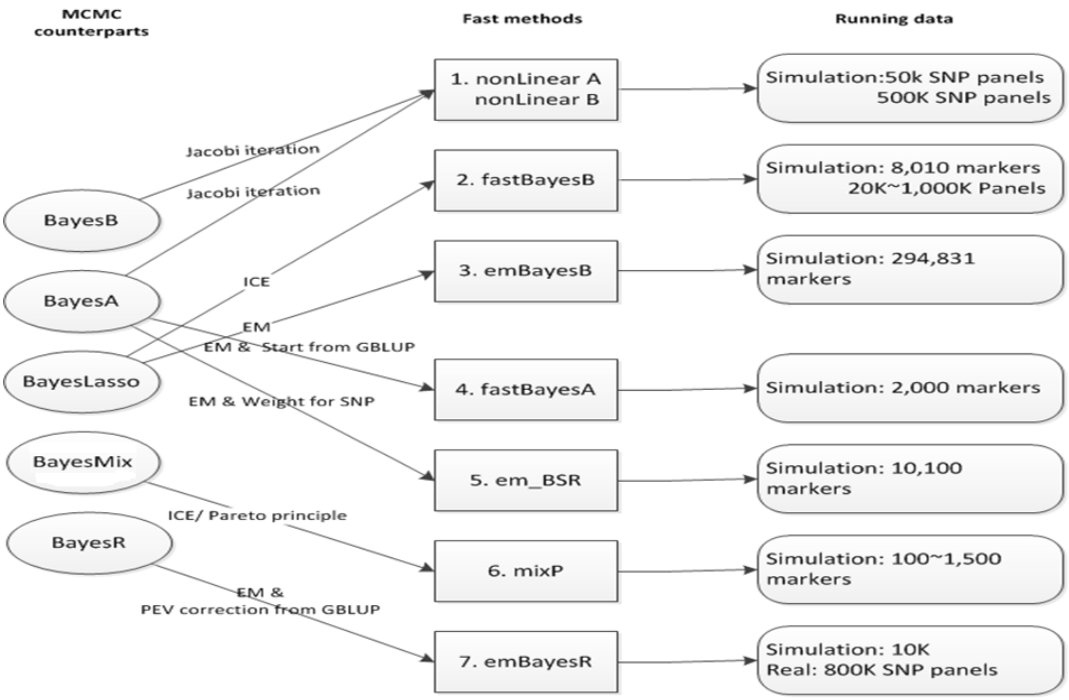
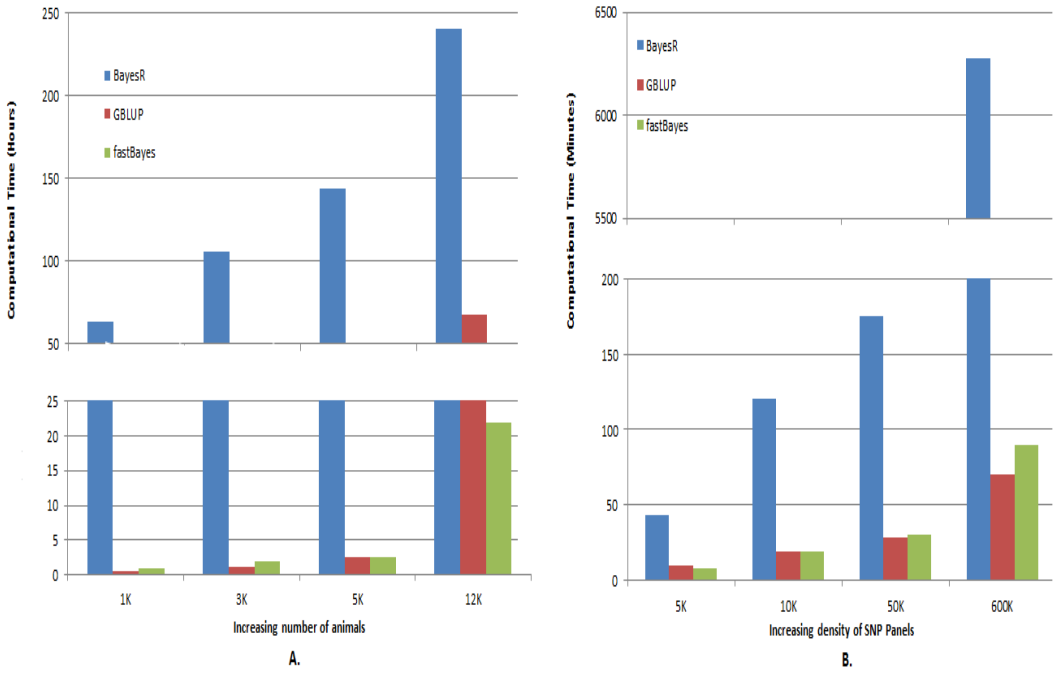


Figure 2.4. Fast Bayesian methods from the MCMC counterparts and their application on simulated and real data.

As shown in Figure 2.5, the results demonstrate that fast methods are up to ten orders faster than their MCMC counterparts. However, due to the heavy shrinkage for SNP effects, both EM and ICE versions of Bayesian models could lead to a reduction in accuracy of genomic prediction. One of the reasons for this might be the assumption that when the effect of a current SNP is estimated, the effects of all the other SNP are assumed to be estimated without error, which is obviously not the case (Sun *et al.* 2012).



A.) The computational time in hours requires for BayesR, GBLUP, and fastBayes according to different number of animas (1,000, 3,000, 5,000, 12,000) on the same density of SNP panels (600K).

B.) The computational time in minutes requires for BayesR, GBLUP, and fast Bayes according to different density of markers (5K, 10K, 50K, 600K) on the same number of animals (3,049). (All of the methods were running on eight threads for one trait (milk yield))

Figure 2.5. Computational time of GBLUP, BayesR and fast Bayesian methods according to increasing size of animals (A.) and increasing density of SNP panels (B.)

Other approaches to speed up the implementation of nonlinear Bayesian models have focused on making the MCMC sampling more efficient. A block sampling approach (where blocks of SNP were sampled as one) was described in Calus et al. (Calus 2014), which demonstrated computational time could be reduced by 74.5-93%, and memory usage by 13.1-66.4%. Another scheme was proposed by Moser et al. (Moser *et al.* 2015), whereby only the 500 SNPs with the largest effects on the trait continued to be sampled in the MCMC chain after a sufficient number of chains – this reduced computation time for BayesR by a factor of 3 to 6, depending on the size of the data set.

2.8 Conclusion

The availability of whole genome sequence data in the major livestock species has potential to improve the accuracy of genomic selection, leading to accelerate genetic gains for target traits. A major challenge however is developing genomic prediction methods that are both computationally efficient enough to derive predictions from this data in reasonable timeframes, and make best use of the data, in order to maximize prediction accuracy. In this review, two hot topics are investigated:

1) Do nonlinear models (e.g. Bayesian models) outperform the linear model (BLUP), with marker densities up to whole genome sequence data?

2) Does increasing density of genomic data, up to whole genome sequence, improve the accuracy of genomic predictions?

A superficial glance of the literature is far from conclusive on both these points, with results in different papers seeming to contradict each other. Deeper investigation however leads to the following conclusions.

The genetic architecture of the trait is very important. In cases where there are mutations of moderate to large effects, the nonlinear methods have a clear advantage. In fact, what is defined as moderate to large actually changes as the data set size increases, as the power to accurately identify causative mutations increases – the nonlinear methods have greater advantage in larger data sets, where a QTL explaining as little as 1% of the genetic variance may be considered moderate, and if a QTL as large as 1% of the variance does exist, the nonlinear methods will have an advantage. Many traits of interest in livestock breeding appear to have this architecture. However, if the trait really is pseudo infinitesimal (controlled by an extremely large number of loci all with very small effects), the advantage of BayesB, BayesR etc. over BLUP is small or non-existent. And, when

the size of markers is increased to a very large number, the performance comparison between Bayesian and BLUP model might need further investigation.

Increasing marker density to whole genome sequence is only an advantage (in terms of more accurate genomic predictions), if a nonlinear genomic prediction is used, if the data set is large, and if some external information is used to assist in the identification of sequence variants that are more likely to affect the trait, and if the MAF of the QTL is such that the QTL has more extreme allele frequencies than the SNP on the SNP chips. Further, the advantage of the sequence data will be greatest when the effective population size is large, including in multi-breed populations.

Given these conclusions, we suggest the following as useful areas of research, with the ultimate goal of using whole genome sequence data routinely in genomic predictions:

1) Improve the prediction ability of current fast versions of Bayesian regression models. Fast versions (non-MCMC) Bayesian models are necessary for their computational efficiency, but published algorithms do give a reduction in accuracy relative to MCMC implementations. One possible scheme is to hybridize EM/ICE schemes with limited number of MCMC iterations.

2) Make use of external and biological information when estimating effects of sequence variants to improve the accuracy. For example McLeod et al. (MacLeod *et al.* 2016) grouped markers into two or more clusters according to various sets of biological information, including gene expression and annotations in BayesRC.

3) Include non-genotyped animals in one-step methods, to further improve accuracy of genomic predictions and reduce any biases associated with

genotyping certain sets of animals (e.g. a bias might be induced if animals are heavily selected on genomic estimated breeding values, and only the selected animals receive phenotypic records, and this is not accounted for in the genomic prediction method). These methods are well developed for GBLUP (Christensen *et al.* 2012; Legarra & Ducrocq 2012; Wang *et al.* 2012a; Misztal *et al.* 2014) with great advances in computational efficiency being made, but less attention has been given to nonlinear one step methods (Fernando *et al.* 2014a; Liu *et al.* 2014) to improve the accuracy.

2.9 Supporting information

All the supporting files are located in Appendix I (Chapter 8) as follows:

File S1 - The introduction of non-Bayesian models including Penalized regression and orthogonal linear models (the theory and differences).

File S2 - The description of the model and prior density function for Bayesian regression models.

File S3 - The example of deriving the conditional prior density function according to Bayesian theory (BayesA was chosen as the example).

File S4 - The detailed review of previous fast algorithms under Bayesian models.

2.10 Acknowledgements

The authors acknowledge the support and fund from Dairy Future CRC.

Chapter 3 A computationally efficient algorithm for genomic prediction using a Bayesian model

3.1 Chapter preface

Justification

This chapter introduced the Expectation-Maximization (EM) (termed emBayesR) algorithm to reduce the computational time required to implement Bayesian non-linear models for genomic prediction. One improvement of emBayesR over previous methods was that Prediction error variance (PEV) correction was introduced to account for the errors generated by estimation of other SNP effects during the estimation of the current SNP effect.

Publication status:

Published in the journal *Genetics Selection Evolution*.

Published as

Wang T, Chen YPP, Goddard ME, Meuwissen TH, Kemper KE, Hayes BJ. (2015) A computationally efficient algorithm for genomic prediction using a Bayesian model. *Genetics Selection Evolution*, 47(1):34.

Statement of contributions of joint authorship

Tingting Wang (Candidate): implemented EM methods to the BayesR model, analyzed the data and drafted the manuscript

Yi-Ping Phoebe Chen (Principle Supervisor): supervised the study.

Michael E Goddard (Collaborator) and Theo HE Meuwissen (Collaborator)

contributed the valuable idea of the PEV correction.

Kathryn E Kemper (Collaborator) implemented BayesR on 630 K high-density SNP panel.

Ben J. Hayes (Co-Supervisor): supervised this project, instructed the implementation of the algorithm on dairy cattle data, and gave a great contribution regarding the organizing/revising of the paper.

This chapter is an exact copy of the version published to Genetics Selection Evolution, except that the reference style, table numbers and figure numbers have been reformatted.

3.2 Abstract

Background

Genomic prediction of breeding values from dense single nucleotide polymorphisms (SNP) genotypes is used for livestock and crop breeding, and can also be used to predict disease risk in humans. For some traits, the most accurate genomic predictions are achieved with non-linear estimates of SNP effects from Bayesian methods that treat SNP effects as random effects from a heavy tailed prior distribution. These Bayesian methods are usually implemented via Markov chain Monte Carlo (MCMC) schemes to sample from the posterior distribution of SNP effects, which is computationally expensive. Our aim was to develop an efficient expectation–maximization algorithm (emBayesR) that gives similar estimates of SNP effects and accuracies of genomic prediction than the MCMC implementation of BayesR (a Bayesian method for genomic prediction), but with greatly reduced computation time.

Methods

EmBayesR is an approximate EM algorithm that retains the BayesR model assumption with SNP effects sampled from a mixture of normal distributions with increasing variance. emBayesR differs from other proposed non-MCMC implementations of Bayesian methods for genomic prediction in that it estimates the effect of each SNP while allowing for the error associated with estimation of all other SNP effects. emBayesR was compared to BayesR using simulated data, and real dairy cattle data with 632 003 SNPs genotyped, to determine if the MCMC and the expectation-maximization approaches give similar accuracies of genomic prediction.

Results

We were able to demonstrate that allowing for the error associated with estimation of other SNP effects when estimating the effect of each SNP in emBayesR improved the accuracy of genomic prediction over emBayesR without including this error correction, with both simulated and real data. When averaged over nine dairy traits, the accuracy of genomic prediction with emBayesR was only 0.5% lower than that from BayesR. However, emBayesR reduced computing time up to 8-fold compared to BayesR.

Conclusions

The emBayesR algorithm described here achieved similar accuracies of genomic prediction to BayesR for a range of simulated and real 630 K dairy SNP data. EmBayesR needs less computing time than BayesR, which will allow it to be applied to larger datasets.

3.3 Background

Genomic prediction uses information from high-density genetic polymorphisms, such as single nucleotide polymorphisms (SNP) panels, to predict the genetic merit of individuals for quantitative traits. Selection based on these estimated breeding values could substantially increase the rates of genetic improvement for quantitative traits in animal and plant species (Meuwissen *et al.* 2001).

Implementation of genomic selection is a two-step process: (1) estimation of the effects of SNPs in a reference population given the phenotypes and SNP genotypes of reference individuals and (2) calculation of genomic estimated breeding values (GEBV) for selection candidates based on their genotypes (Meuwissen *et al.* 2001). If the SNP effects are random variables drawn from a prior distribution, the accuracy of GEBV is maximized if, in step (1), SNP effects are estimated by their expected value conditional on the data.

Several methods, which differ in the assumed prior distribution of SNP effects, have been proposed to estimate SNP effects for genomic prediction. The prior assumption that SNP effects are all drawn from the same normal distribution results in the statistical method called best linear unbiased prediction (BLUP). BLUP for genomic prediction can be implemented using two equivalent models (VanRaden 2008). Either the SNP effects are estimated directly, termed SNP_BLUP (e.g. (Meuwissen *et al.* 2001)), or a genomic relationship matrix is calculated from SNP genotypes, termed genomic BLUP (GBLUP) (VanRaden 2008; Yang *et al.* 2010). Other models assume that the SNP effects follow a non-normal distribution. For example, in the model called BayesA, the SNP effects follow a Student's t distribution (Meuwissen *et al.* 2001), while mixture distributions are used in BayesB (Meuwissen *et al.* 2001), BayesC, BayesCπ (Habier *et al.* 2011) and BayesR (Erbe *et al.* 2012), and exponential distributions are used in BayesLASSO (Park & Casella 2008). With real data and for some traits, GBLUP methods achieve levels of accuracy of genomic prediction similar to non-normal distributions methods such as BayesA, BayesB, and BayesR when moderate SNP densities (e.g. 50K in dairy cattle; less in some crop species with extensive linkage disequilibrium) were used (Habier *et al.* 2007; Daetwyler *et al.* 2012b; Pryce *et al.* 2012; Gao *et al.* 2013; Wimmer *et al.* 2013). As described by several authors, GBLUP has the advantage that it is computationally efficient (Strandén & Garrick 2009; Aguilar *et al.* 2011; Misztal *et al.* 2014). However, for

traits with quantitative traits loci (QTL) of large to moderate effect, the Bayesian methods can give higher accuracies of prediction than GBLUP (Hayes *et al.* 2010; Verbyla *et al.* 2010; Riedelsheimer *et al.* 2012). Moreover, genomic prediction models that assume non-normal distributions of effects in some cases give higher accuracies than GBLUP when very large numbers of SNPs (e.g. 630K or whole-genome sequence data) are used, particularly for multi-breed and across-breed predictions (Erbe *et al.* 2012; Bolormaa *et al.* 2013; Daetwyler *et al.* 2013; MacLeod *et al.* 2014a; MacLeod *et al.* ; Kemper *et al.* 2015). A disadvantage of these methods, however, is that it is difficult, if not impossible, to write closed form solutions for estimates of SNP effects or other parameters, so Markov chain Monte Carlo (MCMC) sampling is used to derive posterior distributions for these effects (e.g. (Mäntysaari)). However, this is computationally expensive, particularly when the number of SNPs is large. For example, the BayesB method can result in the highest accuracy of genomic prediction in some situations, but, since it uses a Metropolis Hastings algorithm, computing time with large numbers of SNPs (e.g. 800 000 SNPs) is very long. Other methods, such as BayesA, BayesLASSO, and BayesR, are usually implemented using Gibbs sampling. While Gibbs sampling is faster than the Metropolis Hasting algorithm, it is still slow with very large numbers of SNPs genotyped in large numbers of individuals.

In dairy cattle routine genomic evaluations, different genomic prediction methods have been implemented by different countries and organizations (Mäntysaari 2014). According to Mantysaari (Mäntysaari 2014), GBLUP, or its single-step implementation (Aguilar *et al.* 2010; Christensen & Lund 2010), is one of the most popular genomic prediction methods implemented for official genomic evaluation in many countries, including Canada, New Zealand, Australia, Germany and Ireland. By contrast, only two countries, i.e. The Netherlands and Switzerland have implemented MCMC non-linear models (BayesA and BayesC) for genomic

prediction. In addition, non MCMC versions of BayesA (also termed nonlinear A (VanRaden 2008)) are used for genomic prediction in the USA. In the future, genomic evaluations may be based on whole-genome sequence data and Bayesian methods may be required to take advantage of this data (Meuwissen & Goddard 2010; Clark *et al.* 2011). Therefore, a way to implement Bayesian models that is faster to compute than the MCMC methods is desirable.

There have been a number of proposals to reduce the computing time required to arrive at satisfactory estimates of the SNP effects from Bayesian methods (e.g. (VanRaden 2007; Meuwissen *et al.* 2009; Gianola 2013)). These proposals use algorithms other than Gibbs sampling. For instance, VanRaden (VanRaden 2008) described an iterative method to implement approximations of both BayesA and BayesB. Meuwissen (Meuwissen *et al.* 2009) described a method termed fastBayesB by using iterative conditional expectation (ICE) in the BayesLASSO model. FastBayesB iteratively calculated each SNP's posterior mean, conditioning on current estimates of all other SNPs as if they were true effects. FastBayesB greatly reduces computing time but several parameters required to describe the prior distribution of SNP effects are assumed to be known. This issue was dealt with in a later publication by an expectation- maximization (EM) algorithm that estimated those parameters by maximizing a joint posterior probability based on the prior distribution of SNP effects, in a method called EmBayesB (Shepherd *et al.* 2010). Lower prediction accuracies were observed for these methods compared with MCMC implementations (Meuwissen *et al.* 2009; Shepherd *et al.* 2010). Two potential reasons for this are: (1) the errors in the estimates of SNP effects other than the SNP for which the effect is being estimated were ignored (Meuwissen *et al.* 2009), and (2) the prior distribution of SNP effects that they assume (a double exponential) may not match the true distribution of SNP effects as well as the mixture distribution assumed by BayesB and BayesR.

Our aim in this paper was to develop a fast EM counterpart to MCMC BayesR (emBayesR). BayesR assumes that SNP effects are drawn from a mixture of normal distributions, one with zero variance (and hence zero effects). BayesR shares some of the advantages of BayesB, in that SNP effects can be zero, moderate, or large, but is more computationally efficient since it can be implemented with Gibbs sampling (Erbe *et al.* 2012). In BayesR, the proportion of SNPs in each normal distribution is estimated from the data, instead of being pre-set as a constant value in BayesB. Consequently, BayesR is able to approximate a wide range of possible true distributions of SNP effects. With real data, BayesR achieves accuracies comparable to BayesA (Erbe *et al.* 2012) and BayesB (Goddard and Meuwissen, unpublished data).

Our EM algorithm retains the BayesR model assumption that SNP effects are assumed to be derived from four different normal distributions, but requires much less computing time than BayesR. It also differs from other EM methods by estimating the effect of each SNP while accounting for the errors in the estimates of all other SNPs. It does this by treating the combined effect of the other SNPs as a residual breeding value, and approximating its prediction error variance from a GBLUP prediction. To compare speed and accuracy of prediction of emBayesR with that from BayesR, we used both a simulated dataset and a real dataset on 630K SNPs for dairy cattle.

3.4 Methods

In this section, we first describe the model of BayesR (here also named MCMC_BayesR) for genomic prediction and second, an EM algorithm named emBayesR. Finally, the 10K simulated data and 630K real dairy data that were used to evaluate the performance of emBayesR, are described.

Statistical model for emBayesR and prior distributions of parameters

The linear model for phenotypes is:

$$\mathbf{y} = \mathbf{1}_{\boldsymbol{n}}\mu + \mathbf{Zg} + \mathbf{e}, \quad (1)$$

where, **y** is a $n \times 1$ vector of phenotypic records (*n* is the number of animals); $\mathbf{1}_{\boldsymbol{n}}$ is a *n* x *1* vector of 1s, *μ* is the population mean; $\mathbf{Z}$ is a $n \times m$ design matrix with elements $\mathbf{Z}_{\mathrm{i}} = (\mathbf{x}_{\mathrm{i}} - 2p_i)/\sqrt[2]{2p_i(1-p_i)}$, in which $\mathbf{x}_{\mathrm{i}}$ is the $n \times 1$ vector of genotypes for the *i*th SNP (0, 1 or 2 copies of the second allele), and p_i is the allele frequency of each SNP *i* (m is the number of SNPs); $\mathbf{e}$ is a $n \times 1$ vector of random normal deviates, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$; $\mathbf{g}$ is a $m \times 1$ vector of SNP effects.

For convenience, polygenic effects were not included in the model but they can be readily added (and have been added in the MCMC version of BayesR, e.g. (Erbe *et al.* 2012)).

BayesR (Erbe *et al.* 2012) assumes that SNP effects (**g**) are drawn from a mixture of four normal distributions N(0, $\boldsymbol{\sigma}_{\boldsymbol{k}}^{\mathbf{2}}$) according to the proportion vector $\mathbf{Pr} = \{Pr_k | k = 1,2,3,4\}$. Variances used were $\boldsymbol{\sigma}_{\mathbf{k}}^{\mathbf{2}} = \{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2\}$ for the analysis of the real dairy data and $\boldsymbol{\sigma}_{\mathbf{k}}^{\mathbf{2}} = \{0, 0.0006 * \sigma_g^2, 0.006 * \sigma_g^2, 0.06 * \sigma_g^2\}$ for the analysis of the simulated data, where σ_g^2 is total genetic variance (Erbe *et al.* 2012). Here, the coefficients of σ_g^2 used to define $\boldsymbol{\sigma}_{\mathbf{k}}^{\mathbf{2}}$ for the simulated data were different to those used for the real data because of the criterion that the sum of the variance across all SNPs approaches the overall genetic variance explained by SNPs. In the simulation data, with 10 050 SNPs, there were only 50 QTL (17 QTL in $\sigma_k^2[2]$, 16 QTL in $\sigma_k^2[3]$ and 17 QTL in $\sigma_k^2[4]$). To make the overall variance summed over all the SNPs approximately equal to σ_g^2, vector $\boldsymbol{\sigma}_{\mathbf{k}}^{\mathbf{2}}$ for the simulated data was set to $\{0, 0.0006 * \sigma_g^2, 0.006 * \sigma_g^2, 0.06 * \sigma_g^2\}$. For the real data (with high-density SNP panels), the value of $\boldsymbol{\sigma}_{\mathbf{k}}^{\mathbf{2}}$ that is

$\{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2\}$ was assumed as in Erbe et al. (Erbe *et al.* 2012). In addition, the proportion of SNPs in each normal distribution (Pr_k; $\sum_{k=1}^{4} Pr_k = 1$) was assumed to follow a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (1,1,1,1)^T$, which is a 4 x 1 vector of the pseudo-counts of the number of SNPs in each distribution. Therefore, the BayesR model has two fixed parameters as input: $\boldsymbol{\sigma_k^2}$ and $\boldsymbol{\alpha}$ (the prior for **Pr**).

For each SNP *i*, there is a latent binary variable b_{ik} (b_{ik} = 0 or 1) that indicates whether or not the effect of SNP *i* follows the normal distribution with variance $\boldsymbol{\sigma_k^2}$ (*k* = 1, 2 ,3 ,4). Therefore:

$$p(b_{ik} = 1|Pr_k) = Pr_k. \tag{2}$$

Then, the prior distribution of each SNP effect (g_i) conditional on variable b_{ik} is:

$$p(g_i|b_{ik}) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{g_i^2}{2\sigma_k^2}\right), & \text{if } b_{ik} = 1 \ (k = 2,3,4) \\ \delta(g_i), & \text{if } b_{i1} = 1 \end{cases}, \tag{3}$$

where $\delta(g_i)$ denotes the Dirac delta function with all probability mass at $g_i = 0$.

Then, the joint distribution $p(g_i, \mathbf{b_i})$ conditional on **Pr** is:

$$p(g_i, \mathbf{b_i}|\mathbf{Pr}) = \prod_{k=1}^{4} p(g_i|b_{ik}) \times p(b_{ik}|Pr_k)$$

$$= (\delta(g_i)Pr_1)^{b_{i1}} \prod_{k=2}^{4} \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{g_i^2}{2\sigma_k^2}\right) Pr_k\right)^{b_{ik}}. \tag{4}$$

Expectation- maximization steps for emBayesR

An EM algorithm is applied to BayesR to obtain estimates of parameters, including SNP effects ($\hat{\mathbf{g}}$) and the proportion of SNP effects in each distribution ($\widehat{\mathbf{Pr}}$). The aim of emBayesR is to predict $\mathbf{Zg}$ by $\mathbf{Z}\hat{\mathbf{g}}$ as accurately as possible.

The best predictor for g_i would be $\hat{g}_i = \mathrm{E}(g_i|\mathbf{y})$, but we approximated this by estimating $\hat{g}_i$ by the value of g_i that maximizes the posterior probability $\mathrm{P}(g_i|\mathbf{y}, \widehat{\mathbf{Pr}}, \hat{\mu}, \widehat{\sigma_e^2})$, where $\widehat{\mathbf{Pr}}$, $\hat{\mu}$ and $\widehat{\sigma_e^2}$ are the MAP (Maximum A Posterior) estimator of $\mathbf{Pr}$, μ, and σ_e^2, conditional on $\mathbf{y}$. In the following, we first deal with estimating $\hat{g}_i$ and then return to $\widehat{\mathbf{Pr}}$.

For estimation of g_i, we maximized the marginal posterior of g_i rather than the joint posterior of all $\mathbf{g}$. To do this, we first introduce two vectors of missing data $(\mathbf{u}, \mathbf{b_i})$, and use the EM algorithm to integrate them out of the posterior distributions. Here, $\mathbf{u}$ is the combined effects of all other SNPs except the current SNP, i.e. $\mathbf{u} = \mathbf{Zg} - \mathbf{Z_i}g_i$, and the other vector $\mathbf{b_i} = \{b_{ik}|k = 1,2,3,4\}$ is for indicator variables that determine which normal distribution each SNP effect is derived from, as described above. Then Equation (1) can be re-written as:

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Z}_\mathrm{i}g_i + \mathbf{u} + \mathbf{e}. \quad (5)$$

The full posterior distribution with the missing data, $p(g_i, \mathbf{u}, \mu, \mathbf{b_i}|\mathbf{y}, \widehat{\mathbf{Pr}})$ is (following Bayes' theorem):

$$p(g_i, \mathbf{u}, \mathbf{b_i}|\mathbf{y}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}) = \frac{f(\mathbf{y}|g_i, \mathbf{u}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}})p(g_i, \mathbf{b_i}|\widehat{\mathbf{Pr}})}{p(\mathbf{y}, \mathbf{u})} \propto f(\mathbf{y}|g_i, \mathbf{u}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}})p(g_i, \mathbf{b_i}|\widehat{\mathbf{Pr}}) \quad (6)$$

where the likelihood of the data $f(\mathbf{y}|g_i, \mathbf{u}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}})$ can be expressed as:

$$f(\mathbf{y}|g_i, \mathbf{u}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}) = \frac{1}{(2\pi\widehat{\sigma_e^2})^{\frac{n}{2}}} exp\left[-\frac{1}{2\widehat{\sigma_e^2}}(\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i}g_i)'(\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i}g_i)\right],$$

with $\mathbf{y}^* = \mathbf{y} - \mathbf{1_n}\hat{\mu}$. Then, the log of the posterior is:

$$logp(g_i, \mathbf{u}, \mathbf{b_i}|\mathbf{y}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}) = logf(\mathbf{y}|g_i, \mathbf{u}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}) + logp(g_i, \mathbf{b_i}|\widehat{\mathbf{Pr}}) + constant.$$

This can be re-written as:

$$logf(\mathbf{y}|g_i, \mathbf{u}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}) = -0.5nlog\widehat{\sigma_e^2} - \frac{1}{2\widehat{\sigma_e^2}}(\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i}g_i)'(\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i}g_i) \quad (6a)$$

$$logp\left(g_i, \mathbf{b_i}|\widehat{\mathbf{Pr}}\right) = b_{i1} log\left(\delta(g_i)\widehat{Pr}_1\right) + \sum_{k=2}^{4} b_{ik}\left(-\frac{1}{2} log\sigma_k^2 - \frac{g_i^2}{2\sigma_k^2} + log\widehat{Pr}_k\right). \quad (6b)$$

In the E-step of emBayesR, we will take expectation of the log posterior function of Equation (6) over the missing data $(\mathbf{u}, \mathbf{b})$. Only the second term (6b) in the equation $logp\left(\mathrm{g_i}, \mathbf{u}, \mathbf{b_i}|\mathbf{y}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}\right)$ involves $\mathbf{b_i}$. Therefore:

$$\mathrm{E}_{\mathbf{b_i}} logp\left(\mathrm{g_i}, \mathbf{b_i}|\widehat{\mathbf{Pr}}\right) = \mathrm{E}_{\mathbf{b_i}}\left[b_{i1} log\left(\delta(g_i)\widehat{Pr}_1\right) + \sum_{k=2}^{4} b_{ik}\left(-\frac{1}{2} log\sigma_k^2 - \frac{g_i^2}{2\sigma_k^2} + log\widehat{Pr}_k\right)\right]$$

$$= P_{i1} log\left(\delta(g_i)\widehat{Pr}_1\right) + \sum_{k=2}^{4} P_{ik}\left(-\frac{1}{2} log\sigma_k^2 - \frac{g_i^2}{2\sigma_k^2} + log\widehat{Pr}_k\right),$$

where $P_{ik} = \mathrm{E}(b_{ik}|\mathbf{y}, \widehat{Pr}_k)$, which is the posterior probability for each SNP to belong to each of the four normal distributions. The derivation of P_{ik} is explained in Additional file 1.

Next, we take the expectation over missing data $\mathbf{u}$. Only the quadratic form $\mathbf{Q} = (\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i} g_i)'(\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i} g_i)$ in the first term of Equation (6a) is related to $\mathbf{u}$. To calculate the expectation of Equation (6a) over $\mathbf{u}$, we only need to take the expectation of $\mathbf{Q}$ over $\mathbf{u}$. Applying Searle's expectation rule (Seber 2002) to $\mathrm{E}_{\hat{\mathbf{u}}}(\mathbf{Q})$, we obtain:

$$\mathrm{E}_{\hat{\mathbf{u}}}(\mathbf{Q}) = \mathrm{E}_{\hat{\mathbf{u}}}[(\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i} g_i)'(\mathbf{y}^* - \mathbf{u} - \mathbf{Z}_\mathrm{i} g_i)]$$

$$= (\mathbf{y}^* - \hat{\mathbf{u}} - \mathbf{Z}_\mathrm{i} g_i)'(\mathbf{y}^* - \hat{\mathbf{u}} - \mathbf{Z}_\mathrm{i} g_i) + tr(\mathrm{PEV}(\hat{\mathbf{u}})),$$

where $\hat{\mathbf{u}} = \sum_{\mathrm{j}\neq\mathrm{i}} \mathbf{Z}_\mathrm{j} \hat{g}_j$ and PEV is the predicted error variance.

Substituting $P_{ik} = \mathrm{E}(b_{ik}|\mathbf{y})$ and using the above $\mathrm{E}_{\hat{\mathbf{u}}}(\mathbf{Q})$, the expectation of Equation (6) over $\hat{\mathbf{u}}, \mathbf{b}$ is:

$$\mathrm{E}_{\mathbf{b_i}, \mathbf{u}|\mathbf{y}} logp\left(g_i, \mathbf{u}, \mathbf{b_i}|\mathbf{y}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}\right)$$

$$= -\frac{n}{2} log\widehat{\sigma_e^2} - \frac{(\mathbf{y}^* - \hat{\mathbf{u}} - \mathbf{Z}_\mathrm{i} g_i)'(\mathbf{y}^* - \hat{\mathbf{u}} - \mathbf{Z}_\mathrm{i} g_i) + \mathrm{tr}\big(\mathrm{PEV}(\hat{\mathbf{u}})\big)}{2\widehat{\sigma_e^2}}$$

$$+P_{i1} log\big(\delta(g_i)\widehat{Pr}_1\big) + \sum_{k=2}^{4} P_{ik}\left[log\widehat{Pr}_k - 0.5 * log\sigma_k^2 - \frac{g_i^2}{2\sigma_k^2}\right]$$

$$+ constant. \qquad (7)$$

The calculation of $\mathrm{PEV}(\hat{\mathbf{u}})$ is approximated from a GBLUP model, and is explained in Additional file 2.

The M-step of emBayesR involved estimation of the SNP effects (g_i). Differentiating Equation (7) with regard to g_i gives:

$$\frac{\partial \mathrm{E}_{\mathbf{b_i},\mathbf{u}|\mathbf{y}} log p\big(g_i, \mathbf{u}, \mathbf{b_i}\big|\mathbf{y}, \hat{\mu}, \widehat{\sigma_e^2}, \widehat{\mathbf{Pr}}\big)}{\partial g_i}$$

$$= \left[-\sum_{k=2}^{4} \frac{P_{ik}}{\sigma_k^2} - \frac{\mathbf{Z}_\mathrm{i}'\mathbf{Z}_\mathbf{i}}{\widehat{\sigma_e^2}}\right] g_i + \frac{\mathbf{Z}'(\mathbf{y} - \hat{\mathbf{u}} - \mathbf{1}_{\boldsymbol{n}}\hat{\mu})}{\widehat{\sigma_e^2}} = 0.$$

Setting this equal to 0 results in the following posterior mode estimate for each SNP effect (g_i).

$$\hat{g}_i = [\mathbf{Z}_\mathrm{i}'\mathbf{Z}_\mathbf{i} + \left(P_{i2}\frac{\widehat{\sigma_e^2}}{\sigma_2^2} + P_{i3}\frac{\widehat{\sigma_e^2}}{\sigma_3^2} + P_{i4}\frac{\widehat{\sigma_e^2}}{\sigma_4^2}\right)]^{-1}[\mathbf{Z}'\mathbf{y}^\dagger], \qquad (8a)$$

where, $\mathbf{Z}_\mathrm{i}$ is the i^{th} column of matrix $\mathbf{Z}$, and $\mathbf{y}^\dagger = \mathbf{y} - \hat{\mathbf{u}} - \mathbf{1}_{\boldsymbol{n}}\hat{\mu}$.

The mean of the posterior distribution can also be calculated as follows:

$$\mathrm{E}\big(p(g_i|\mathbf{y}, Pr_k)\big) = \frac{\int_{-\infty}^{+\infty}(\sum_{k=1}^{4} P_{ik} p(g_i|b_{ik}=1,\mathbf{y},\mathrm{Pr}) g_i \mathrm{d}g_i}{\int_{-\infty}^{+\infty}(\sum_{k=1}^{4} P_{ik} p(g_i|b_{ik}=1,\mathbf{y},\mathrm{Pr})\mathrm{d}g_i},$$

which reduces to:

$$\bar{g}_i = \sum_{k=1}^{4} P_{ik}[(\mathbf{Z}_\mathrm{i}'\mathbf{Z}_\mathbf{i} + \frac{\sigma_e^2}{\sigma_k^2})]^{-1}\,[\mathbf{Z}'\mathbf{y}^\dagger]. \qquad (8b)$$

The mode estimation of SNP effects (Equation 8a) was implemented in our EM iterations, unless otherwise stated. The posterior mean of Equation (8b) was used in some cases to evaluate the accuracy of genomic prediction using either the mode or mean estimates of SNP effects. Furthermore, to investigate the degree of shrinkage, the least square estimate of the SNP effect was also calculated for some examples:

$$g_i^{ls} = (\mathbf{Z_i}'\mathbf{Z_i})^{-1}\mathbf{Z_i}'(\mathbf{y} - \mathbf{1}_{\boldsymbol{n}}\mu).$$

Similar EM steps used for estimating $\hat{g}_i$ (but with different full models) are applied to estimate other parameters, including the proportion of SNP effects in each distribution ($\mathbf{Pr}$), the error variance (σ_e^2), and the mean (μ).

To obtain $\widehat{\mathbf{Pr}}$, we return to the full model Equation (1) with all SNP effects ($\mathbf{g}$) included. We introduce the missing variables $\mathbf{b}$, so the full likelihood is:

$p(\mathbf{Pr}, \mathbf{b}|\mathbf{y}, \mu) \propto p(\mathbf{y}|\mathbf{b})p(\mathbf{b}|\mathbf{Pr})p(\mathbf{Pr})$,

Note that $p(\mathbf{y}|\mathbf{b})$ does not involve $\mathbf{Pr}$, so when we differentiate with respect to **Pr**, this term drops out and can, therefore, be ignored, resulting in:

$p(\mathbf{b}|\mathbf{Pr}) = \prod_{i=1}^{n}\prod_{k=1}^{4}(Pr_k)^{b_{ik}}$,

$p(\mathbf{Pr}) = \prod_{k=1}^{4} Pr_k$,

$logp(\mathbf{b}|\mathbf{Pr}) = \sum_{i=1}^{n}\sum_{k=1}^{4} b_{ik} logPr_k$,

$logp(\mathbf{Pr}) = \sum_{k=1}^{4} logPr_k$, and

$\mathrm{E}_{\mathbf{b}|\mathbf{y}} logp(\mathbf{b}|\mathbf{Pr}) = \sum_{i=1}^{n}\sum_{k=1}^{4} P_{ik} logPr_k$,

where $P_{ik} = E(b_{ik}|\mathrm{y}, Pr_k)$.

Then, considering that $\sum_{k=1}^{4} Pr_k$ =1, we use Lagrange multiplier λ and

differentiate with respect to Pr_k. Given that **Pr** follows a Dirichlet distribution:

$$\frac{\partial \mathrm{E}_{\mathbf{b}|\mathbf{y}} logp(\mathbf{g}, \mathbf{Pr}, b_{ik}|\mathbf{y}, \mu) + \lambda(\sum_{k=1}^{4} Pr_k - 1)]}{\partial Pr_k}$$

$$= \frac{\sum_{i=1}^{m} P_{ik}}{Pr_k} + \frac{1}{Pr_k} + \lambda = 0.$$

Therefore, the solution is:

$$Pr_k = \frac{\sum_{i=1}^{m} P_{ik}+1}{\sum_{k=1}^{4}(\sum_{i=1}^{m} P_{ik}+1)}. \quad (9)$$

Finally, to estimate the error variance σ_e^2 and μ, we simplify Equation (5) into $\mathbf{y} = \mathbf{1_n}\mu + \mathbf{u}^* + \mathbf{e}$, $\mathbf{u}^* = \sum_{\mathrm{i}=0}^{m} \mathbf{Z}_{\mathrm{i}} \hat{g}_i$ and then the full likelihood based on this model is:

$$p(\sigma_e^2, \mu, \mathbf{u}^*|\mathbf{y}) =$$

$$\frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}} exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{u}^* - \mathbf{1_n}\mu)'(\mathbf{y} - \mathbf{u}^* - \mathbf{1_n}\mu)\right].$$

The expectation for the full log likelihood based on this model is:

$$\mathrm{E}_{\mathbf{u}^*|\mathbf{y}} logp(\sigma_e^2, \mu, \mathbf{u}^*|\mathbf{y})$$

$$= \mathrm{E}_{\mathbf{u}^*|\mathbf{y}}\left[-\frac{n}{2}\log\sigma_e^2 + \frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{u}^* - \mathbf{1_n}\mu)'(\mathbf{y} - \mathbf{u}^* - \mathbf{1_n}\mu)\right]$$

$$= -\frac{n}{2}\log\sigma_e^2 + \frac{1}{2\sigma_e^2}[(\mathbf{y} - \widehat{\mathbf{u}^*} - \mathbf{1_n}\mu)'(\mathbf{y} - \widehat{\mathbf{u}^*} - \mathbf{1_n}\mu) + tr(\mathrm{PEV}(\widehat{\mathbf{u}^*}))]. \quad (10)$$

Therefore, differentiating Equation (10) with regard to σ_e^2 and μ, we get:

$$\sigma_e^2 = \frac{1}{n}[(\mathbf{y} - \widehat{\mathbf{u}^*} - \mathbf{1_n}\mu)'(\mathbf{y} - \widehat{\mathbf{u}^*} - \mathbf{1_n}\mu) + \mathrm{tr}(\mathrm{PEV}(\widehat{\mathbf{u}^*}))], \quad (11)$$

$$\mu = \frac{1}{n}(\mathbf{1_n})'(\mathbf{y} - \widehat{\mathbf{u}^*}), \quad (12)$$

for which computation of the term $\mathrm{tr}(\mathrm{PEV}(\widehat{\mathbf{u}^*})$ is explained in Additional file 2.

In order to demonstrate the importance of the PEV correction for SNP effect estimates, the accuracy of emBayesR with and without accounting for PEV will be compared in the Results section. emBayesR without PEV has a similar EM step as emBayesR with PEV to derive the parameters P_{ik}, $\hat{g}_i$, Pr_k, σ_e^2 and μ but differs in the equations of emBayesR with PEV to calculate P_{ik} (Equation A3 in Additional file 1) and σ_e^2 (Equation 11) in that the term $tr(\mathrm{PEV}(\hat{\mathbf{u}}))$ is not included in emBayesR without PEV.

The emBayesR algorithm

The emBayesR algorithm can be described as follows:

Step 1

Initialise starting values for $\mathbf{g}$, $\mathbf{Pr}$, σ_e^2, σ_g^2, $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}_{\boldsymbol{k}}^{\mathbf{2}}$. There are two groups of parameters: fixed parameters and changing parameters. $\boldsymbol{\alpha} = (1,1,1,1)$, σ_g^2 and $\boldsymbol{\sigma}_{\boldsymbol{k}}^{\mathbf{2}}$ are fixed parameters, where $\boldsymbol{\alpha}$ is the prior parameter for $\mathbf{Pr}$, and σ_g^2 is used to set the value of $\boldsymbol{\sigma}_{\boldsymbol{k}}^{\mathbf{2}}$. The other variables ($\mathbf{g}$, $\mathbf{Pr}$, σ_e^2) are updated during EM iterations. We used $\mathbf{g} = 0.01$ and $\mathbf{Pr} = \{0.5, 0.487, 0.01, 0.003\}$, as in (Erbe *et al.* 2012). To initialise σ_e^2 and σ_g^2, we used GBLUP implemented through ASREML3.0 (Gilmour *et al.* 2002) to estimate the error variance σ_e^2 and the genetic variance σ_g^2 as inputs for the next steps. Then, as mentioned before, the value of σ_g^2 defines $\boldsymbol{\sigma}_{\boldsymbol{k}}^{\mathbf{2}}$, using $\boldsymbol{\sigma}_{\mathbf{k}}^{\mathbf{2}} = \{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2\}$ for the real data and $\boldsymbol{\sigma}_{\mathbf{k}}^{\mathbf{2}} = \{0, 0.0006 * \sigma_g^2, 0.006 * \sigma_g^2, 0.06 * \sigma_g^2\}$ for the simulated data.

Step 2

Calculate **PEV** with Equation (A7) of Additional file 2 (or it can be taken from ASREML in the step above).

Then for each SNP *i* (*i* in 1:m):

Step 3

Correct **y** for the effects of all other SNPs except the current SNP *i*, using:

$\mathbf{y}^{\dagger} = \mathbf{y} - \sum_{j \neq i} \mathbf{Z_j}\, \hat{g}_j - \mathbf{1_n}\hat{\mu}$.

Step 4

Estimate the probability that the effect of SNP *i* is from one of four normal distributions $logl_{ik}$ with Equation (A5) of Additional file 1.

Step 5

Calculate P_{ik} with Equation (A6) of Additional file 1.

Step 6

Estimate the effect of SNP *i* with Equation (8a).

Step 7

After all SNP effects have been estimated, calculate Pr_k with Equation (9), update σ_e^2 with Equation (11), and update μ with Equation (12).

Step 8

Return to Step 3 and iterate until convergence. Here, the convergence criterion evaluated at each iteration q was $(\hat{\mathbf{g}}^q - \hat{\mathbf{g}}^{q-1})'(\hat{\mathbf{g}}^q - \hat{\mathbf{g}}^{q-1})/((\hat{\mathbf{g}}^{q\prime}\hat{\mathbf{g}}^q) < \gamma$. The criterion $\gamma = 10^{-10}$ was selected after trialling the algorithm in a number of datasets and investigating changes in SNP effect estimates across iterations.

We calculated the time complexity of the algorithm (the function with parameters number of SNPs and number of animals that determines the time taken for the algorithm to run) based on the above eight steps. Time complexity is estimated in computer science applications by counting the number of innermost loops for

elementary operations, which is notated O. For example, $O(n)$ means the elementary operations in the algorithm need to be looped n times.

EmBayesR need *q* loops to be converged. For each loop, Equation (A5) of Additional file 1 (Step 4 in the EM loop of emBayesR algorithm), is located in the innermost loop for the iteration. To be mentioned, both $\mathrm{tr}\left(\mathrm{PEV}(\hat{\mathbf{u}})\right)$ and $\mathrm{tr}\left(\mathbf{Z_i}\mathbf{Z_i'}\mathrm{PEV}(\hat{\mathbf{u}})\right)$ in Equation (A5) are required, but fortunately they can be calculated outside EM iterations [See Additional file 1 for details]. Then, except for these two terms $\mathrm{tr}\left(\mathrm{PEV}(\hat{\mathbf{u}})\right)$ and $\mathrm{tr}\left(\mathbf{Z_i}\mathbf{Z_i'}\mathrm{PEV}(\hat{\mathbf{u}})\right)$, the calculation number of Equation (A5) is the number of SNPs (*m*) $\times$ the number of animals (*n*). Therefore, the time complexity of each iteration in emBayesR is $O(mn)$.

Simulated data

Simulated data were used to determine how close the genomic prediction accuracy of emBayesR was to that of BayesR. The simulated dataset described in (MacLeod *et al.* 2014a) was used. Briefly, FREGENE was used to simulate whole-genome sequence data in a population with an effective size (Ne) of 25 900 and a genome size of 50 Mb split equally over 10 chromosomes. The genome size of 50 Mb was chosen for computing efficiency. The accuracy of prediction in a c times larger genome (i.e. 50c Mb) would be approximately the same as found in our 50 Mb genome, provided the number of animals was c times larger than used here (i.e. 5000c) (Meuwissen & Goddard 2010). The mutation rate per bp was 9.38×10^{-9} and the recombination rate was 1×10^{-8} per base pair per generation (MacLeod *et al.* 2014a), based on estimates for these rates in mammals. To ensure a drift-recombination-mutation equilibrium, the population was run for 370 000 generations. A total of 10 050 markers (including 50 QTL) were randomly selected as SNPs for genomic prediction. The SNP density was equivalent to ~600 000 SNPs on a 3000 Mb genome, similar to many mammals. Fifty QTL were randomly picked from the segregating loci, which is equivalent to 3000 QTL on a human or bovine genome. To evaluate the genomic prediction

performance of emBayesR, BayesR and other algorithms, we generated two genetic architectures that differed in the distribution of true QTL effects. For this first dataset, named HD_Mix, the 50 QTL allele substitution effects were sampled from an equal mixture of three normal distributions with variances $(0, 0.0006\sigma_g^2, 0.006\sigma_g^2, 0.06\sigma_g^2)$. For the second genetic architecture (referred to as HD_One), QTL allele substitution effects were sampled from a single normal distribution. For the breeding values on simulation data, true breeding values (TBV) for individuals were obtained by summing genetic values across QTL. For each of genetic architecture, heritabilities (h^2) of either 0.45 or 0.1 were used. For each set, phenotypes of 5000 individuals were generated by means of adding a random residual value to the TBV of each individual. This residual value was sampled from a normal distribution, $N(0, \sigma^2_e)$, here $\sigma^2_e = [\sigma^2_{TBV}(1-h^2)]/h^2$, where σ^2_{TBV} is the variance of TBV in the population. Thus, we generated four datasets named HD_Mix_45 (five replicates following the mixture data model with heritability 0.45), HD_Mix_10 (five replicates following the mixture data model with heritability 0.10), HD_One_45 (five replicates following the one normal data distribution with heritability 0.45) and HD_One_10 (five replicates following the one normal distribution with heritability 0.10). Each replicate entailed sampling new SNP effects and generating new phenotypes.

To compare prediction accuracies and computing efficiencies of emBayesR, BayesR, GBLUP and fastBayesB, 5000 individuals were randomly separated into reference sets and validation sets. With an h^2 of 0.45, there were 2500 individuals in the reference set and 2500 in the validation set. With an h^2 of 0.1, there were 3750 individuals in the reference set and 1250 in the validation set. Accuracies were the correlations between GEBV and TBV.

Real data

A total of 3354 Holstein-Friesian bulls were genotyped for both the Illumina Bovine HD SNP array (632 003 SNPs following quality controls as described in (Erbe *et*

al. 2012)), and the Bovine SNP 50 array (43 025 SNPs). Bulls genotyped at the lower density were imputed to the higher density using Beagle 3.0 (Browning & Browning 2009), and applying quality controls as described in (Erbe *et al.* 2012). Phenotypes were daughter trait deviations (DTD) from two groups of traits: functional traits, including angularity, mammary conformation, stature, fertility (calving interval) and somatic cell count (SCC), and production traits, including milk yield, protein yield, protein % and fat %. For some of these traits, known QTL with moderate to large effects segregate in this population, for example a mutation in the *DGAT1* gene affects fat % (Grisart *et al.* 2002). Bulls were split into reference and validation sets by age, with the youngest bulls in the validation set. The numbers of bulls in the reference and validation sets for each trait are listed in Table 3.1. As a surrogate for prediction accuracy, the correlation of GEBV and DTD in the validation set was used. To investigate the computing time required for emBayesR relative to BayesR with different numbers of SNPs, we also ran genomic predictions in the same data but with the 50K SNP chip genotypes (38 968 SNPs) extracted from the 630K data on 3354 animals, for milk yield.

Table 3.1. Numbers of Holstein bulls in the reference and validation sets for functional traits and production traits.

	Reference set	**Validation set**
Milk	3049	262
Protein	3049	262
Fertility	2806	396
Protein%	3049	262
Fat%	3049	262
Angularity	1484	251
Mammary conformation	1484	251
Stature	1484	251
Somatic cell count	2662	410

3.5 Results

The results are presented in three sections. First, we investigated the convergence of parameters estimated by emBayesR and how close parameter estimates from emBayesR were to the true parameter values, and those estimated by BayesR, in terms of SNP effects and **Pr**, in the simulated data. We also evaluated the effect of the PEV correction on estimates of these parameters, and the accuracy of genomic prediction. Moreover, the accuracy of genomic prediction from the joint posterior mode estimation from emBayesR was compared to the accuracy when the posterior mean estimate of SNP effects was used. The mode estimation for SNP effects (Equation 8a) of emBayesR was used for the evaluation of performance of emBayesR. Thus, we also compared the accuracy of prediction with mode (8a) and mean (8b) Equations for estimates of SNP effects (Equation 8b). In the second section of results, we compared the accuracy of genomic prediction from emBayesR to that of BayesR, as well as computing speed in simulated and real datasets. Finally, the sensitivity of prediction accuracy from emBayesR to the underlying genetic architecture (multi-normal distribution, normal distribution of QTL effects, real 630K data) was investigated.

Convergence of parameter estimates with emBayesR

The algorithm is considered to have "converged" when estimated SNP effects from the previous iteration are very close to estimated SNP effects in the current iteration. The convergence criterion of emBayesR is $(\hat{\mathbf{g}}^{q} - \hat{\mathbf{g}}^{q-1})'(\hat{\mathbf{g}}^{q} - \hat{\mathbf{g}}^{q-1})/(\hat{\mathbf{g}}^{q\prime}\hat{\mathbf{g}}^{q}) < 10^{-10}$, where q is the current iteration number. Since the convergence criterion assessed only changes in SNP effect estimates, it does not guarantee that the estimates of the other parameters, i.e. **Pr** (the proportion of SNPs in each distribution) and the error variance, have converged. In the simulated dataset HD_Mix_45, convergence was reached after 2500 iterations,

and at that point, there was also very little change in the error variance and **Pr** from the previous iteration (Figure 3.1).

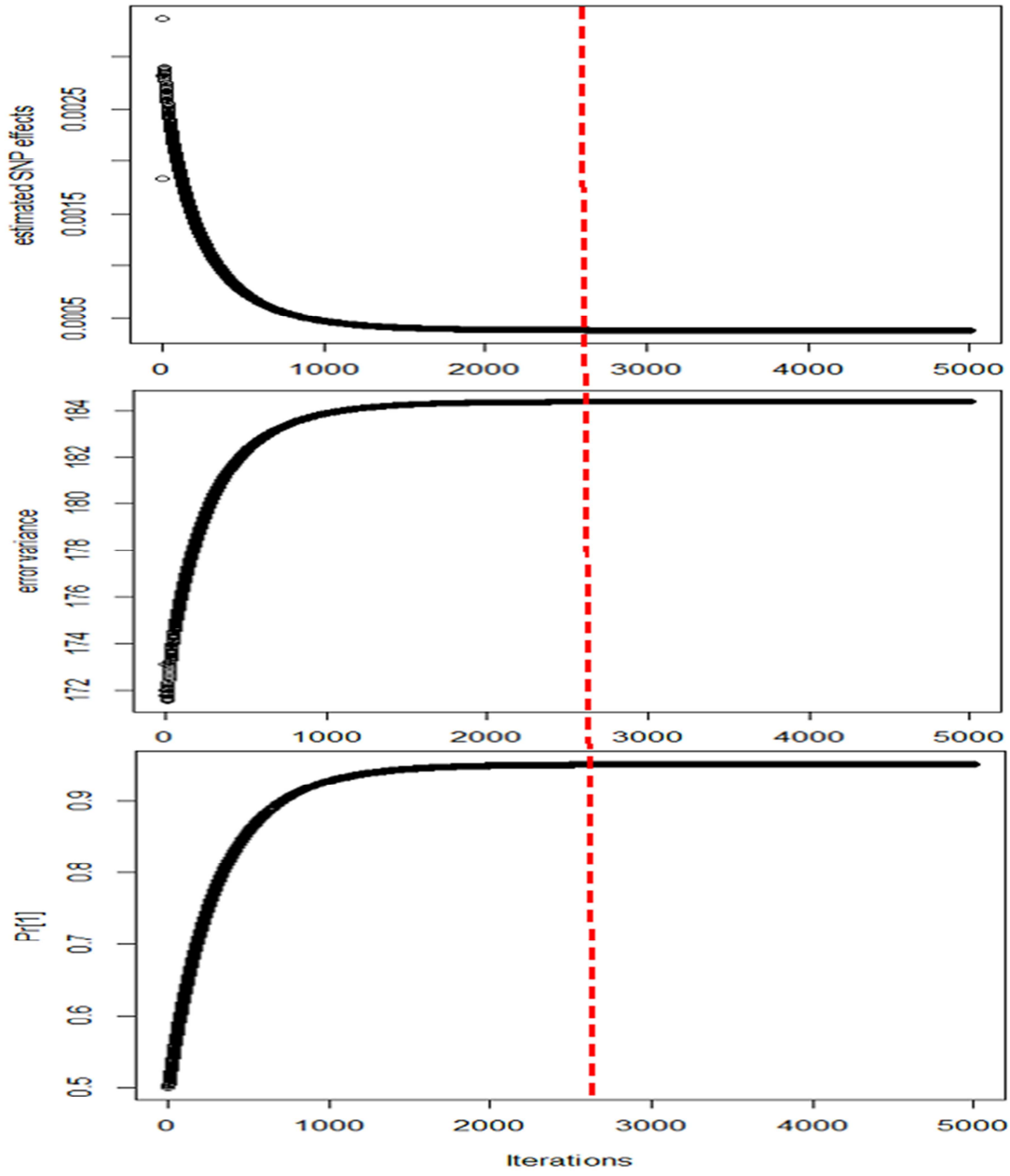


Figure 3.1. Convergence of estimated SNP effects, error variance and Pr over 5000 iterations.

The x-axis represents the number of iterations that ranged from 0 to 5000; the y-axis represents the estimated SNP effects, error variance and the first element of **Pr** (the proportion of SNPs in the distribution with zero variance).

Comparison of parameter estimates

Estimates of SNP effects and **Pr** from emBayesR can be compared to the corresponding estimates from BayesR. For the HD_Mix simulated data, estimates of large SNP effects are very similar for BayesR and emBayesR (Figure 3.2). The plot of BayesR and emBayesR estimated effects against true effects are in Figure 3.3. However, for smaller effects, emBayesR shrunk effects to a greater degree than BayesR, in some replicates.

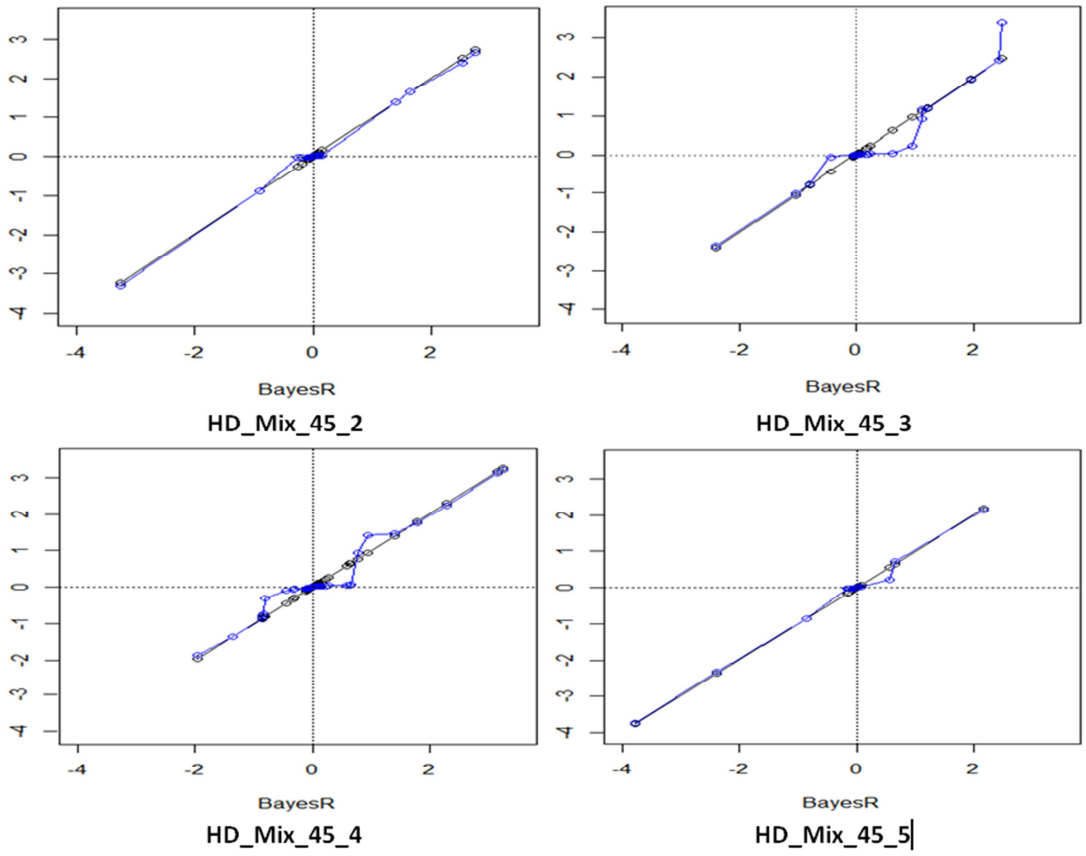


Figure 3.2. Correlation between SNP effects from BayesR and emBayesR SNP effects in four replicates of HD_Mix_45 ($\mathbf{h^2}$ = 0.45).

The x-axis represents the BayesR estimates of SNP effect; blue line plot emBayesR estimates of SNP effects on BayesR estimates of SNP effects; black line plot BayesR estimates of SNP effects on themselves for four replicates of HD_Mix with a heritability of 0.45.

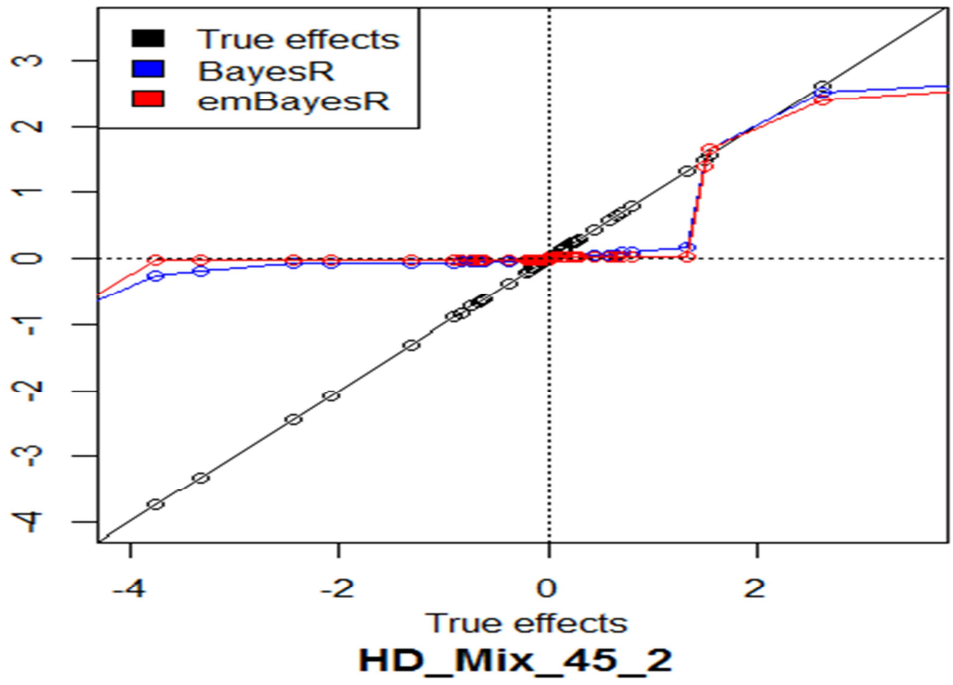


Figure 3.3. Estimates of SNP effects from BayesR and emBayesR compared with their true effects in one replicate of HD_Mix_45 (HD_Mix_45_2).

The x-axis represents true effects; blue curve plots BayesR estimates of SNP effects on true effects; red line plots emBayesR estimates of SNP effects on true effects; the black line plots true effects on themselves for one replicate of simulated data HD_Mix with a heritability of 0.45 (HD_Mix_45_2).

The degree of shrinkage from the BayesR algorithms relative to other algorithms can be demonstrated by plotting estimates of SNP effects (HD_Mix data set) from BayesR, FastBayesB, emBayesR and SNP-BLUP against their least square estimates (Figure 3.4). Both BayesR and emBayesR regressed moderate size SNP effects towards 0 more than SNP-BLUP and FastBayesB. However, BayesR and emBayesR did not shrink large SNP effects nearly as much as SNP-BLUP.

Table 3.2. Estimated mixing proportions (Pr) from BayesR and emBayesR in the 10k simulation data (HD_Mix_45).

Five replicates of 10K simulation data with h^2 = 0.45		
True value of Pr [0.9950 0.0017 0.0016 0.0017]		
	BayesR	**emBayesR**
M45_1	[0.9865 0.0110 0.0010 0.0015]	[0.9813 0.0163 0.0009 0.0015]
M45_2	[0.9861 0.0127 0.0004 0.0008]	[0.9852 0.0136 0.0003 0.0009]
M45_3	[0.9933 0.0046 0.0009 0.0012]	[0.9899 0.0083 0.0005 0.0012]
M45_4	[0.9909 0.0055 0.0022 0.0015]	[0.9864 0.0110 0.0010 0.0016]
M45_5	[0.9944 0.0043 0.0006 0.0007]	[0.9910 0.0078 0.0005 0.0007]
Five replicates of 10K simulation data with h^2 = 0.10		
True value of Pr [0.9950 0.0017 0.0016 0.0017]		
	BayesR	**emBayesR**
M10_1	[0.9759 0.0021 0.0024 0.0010]	**[0.9243 0.0741 0.0009 0.0008]**
M10_2	[0.9624 0.0343 0.0025 0.0009]	**[0.9086 0.0898 0.0010 0.0007]**
M10_3	[0.9757 0.0022 0.0018 0.0008]	**[0.9284 0.0702 0.0007 0.0007]**
M10_4	[0.9620 0.0334 0.0032 0.0014]	**[0.9146 0.0837 0.0008 0.0010]**
M10_5	[0.9664 0.0295 0.0023 0.0018]	**[0.9265 0.0715 0.0007 0.0014]**

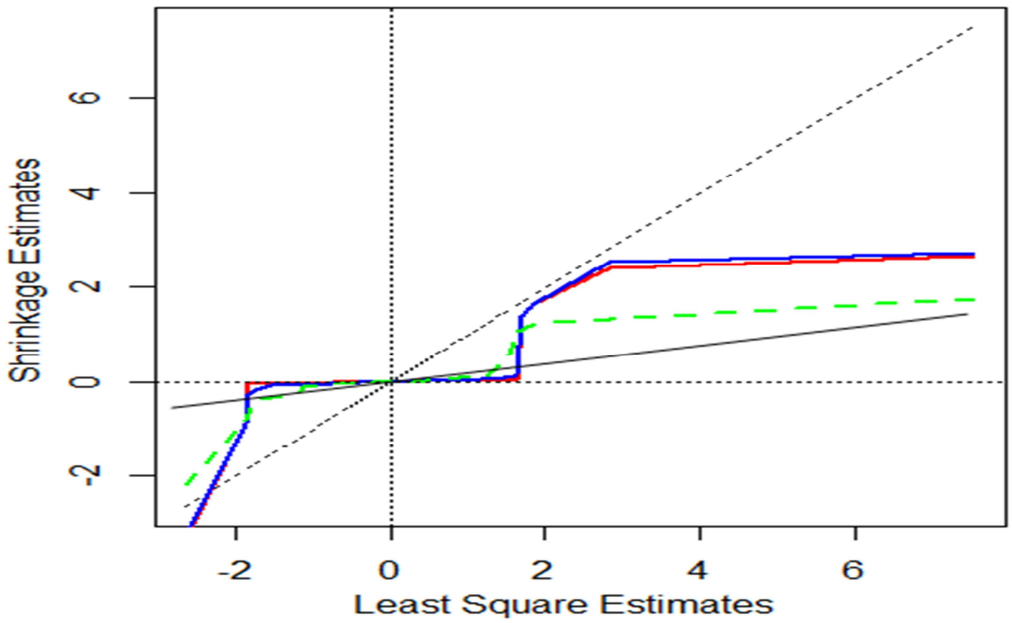


Figure 3.4. Estimates of SNP effects from SNP-BLUP, BayesR, emBayesR, FastBayesB against their least square estimates.

The x axis represents the least square estimates of SNP effects; blue line plotted BayesR estimates of SNP effects on the least square estimates; red line represents emBayesR SNP effect estimates; dotted green line represents the fastBayesB estimates of SNP effects; black line represents SNP_BLUP estimates of SNP effects for HD_Mix_45.

Estimates of **Pr** from emBayesR and BayesR are compared with the true proportion of SNP effects in each of the four normal distributions in Table 3.2. The genetic architecture of the HD_Mix data was such that 50 QTL were distributed evenly in three normal distributions with non-zero variances. The true proportion of the SNP effects (around 10 000 markers) in the four normal distributions with different variances $(0, 0.0006\sigma_g^2, 0.006\sigma_g^2, 0.06\sigma_g^2)$ was (0.995, 0.0017, 0.0016, 0.0017). As shown in Table 3.2, when h^2 = 0.45, both BayesR and emBayesR estimated the proportions of SNP effects from the four distributions to be roughly 0.99, 0.01, 0.001, and 0.001. However, when h^2 = 0.1, BayesR over-estimated the proportion of SNP effects in the smallest non-zero distribution ($\sigma_2^2 = 0.0006\sigma_g^2$)

and this tendency was even greater with emBayesR. This agreed with results in Figure 3.2, where emBayesR shrunk small effects to very small effects more than BayesR and this might have contributed to the over-estimation of the proportion of SNP effects from the distribution with the smallest non-zero variance ($0.0006\sigma_g^2$). In the 630K dairy cattle data, the posterior mean estimates of **Pr** from emBayesR were similar to those from BayesR, as shown in Table 3.3.

Table 3.3. Estimated mixing proportions (Pr) from BayesR and emBayesR for the 630k real dairy cattle data.

	BayesR	**emBayesR**
Milk	[0.99291 0.00690 0.00018 0.00001]	**[0.99511 0.00480 0.00006 0.00003]**
Protein	[0.99161 0.00831 0.00005 0.00003]	**[0.99480 0.00511 0.00007 0.00002]**
Fertility	[0.98863 0.01034 0.00092 0.00011]	**[0.99184 0.00806 0.00009 0.00001]**
Protein%	[0.99602 0.00378 0.00019 0.00001]	**[0.99902 0.00078 0.00004 0.00016]**
Fat%	[0.99480 0.00485 0.00021 0.00014]	**[0.99786 0.00204 0.00001 0.00009]**
Angularity	[0.99221 0.00739 0.00039 0.00001]	**[0.98514 0.01475 0.00009 0.00002]**
Mammary conformation	[0.99091 0.00859 0.00047 0.00003]	**[0.99276 0.00714 0.00009 0.00001]**
Stature	[0.99013 0.00927 0.00052 0.00008]	**[0.99305 0.00684 0.00006 0.00005]**
Somatic cell count	[0.98688 0.01272 0.00039 0.00001]	**[0.98761 0.01229 0.00008 0.00002]**

Sensitivity to the prior for the Dirichlet distribution

Another feature of estimates of **Pr**, may be sensitivity to its prior parameter $\boldsymbol{\alpha}$ (the pseudo-count of SNPs in each distribution in the Dirichlet distribution). To evaluate the sensitivity of emBayesR to $\boldsymbol{\alpha}$, we used different values for $\boldsymbol{\alpha}$ and investigated the effect on **Pr** with the dataset HD_Mix_45 (Table 3.4). When the prior parameter $\boldsymbol{\alpha}$ was changed from (1, 1, 1, 1) to (100, 1, 1, 1), estimates of **Pr** from emBayesR changed only slightly. Although $\boldsymbol{\alpha} =$ (100, 1, 1, 1) was closer to the true situation in the simulated datasets, estimates for **Pr** (especially *Pr[2]*, *Pr[3]*, *Pr[4]*) deviated from the true values [0.9950 0.0017 0.0016 0.0017]. When $\boldsymbol{\alpha}$ was changed to (1, 1, 1, 100) and (1, 1, 100, 1), the estimate of **Pr** was affected, with the proportion of SNP effects estimated to be in the distribution with

α[4] = 100 increasing to 0.0027 and 0.0028, respectively, instead of the simulated 0.0017. It is not surprising that a pseudo-count of 100 affected the estimate of **Pr**, since the true number of SNP effects in these distributions was equal to 17 only. Interestingly, the prediction accuracy remained at 0.97 in spite of these changes in the prior **α**.

Table 3.4. Pr estimates (proportion of SNP in each distribution) with different prior values α for the HD_Mix_45 simulated data.

	Pr_emBayesR			
α	**0**	$\mathbf{0.0006 * \sigma_g^2}$	$\mathbf{0.006 * \sigma_g^2}$	$\mathbf{0.06 * \sigma_g^2}$
(1, 1, 1, 1)	0.9861	0.0127	0.0004	0.0008
(1, 1, 1, 100)	0.9801	0.0130	0.0042	0.0027
(1, 1, 100, 1)	0.9863	0.0101	0.0028	0.0008
(100,1, 1, 1)	0.9883	0.0105	0.0003	0.0009

The prior **α** is (1, 1, 1, 1), (1, 1, 1, 100), (1, 100, 1, 1) or (100, 1, 1, 1).

Effect of PEV

We also compared estimates of parameters and accuracies of genomic prediction with and without accounting for PEV or estimates of all other SNPs in the emBayesR algorithm. When the PEV was accounted for in the emBayesR algorithm, there was a 6% improvement in the accuracy of genomic prediction in the simulated data when h^2 = 0.45, and 5% when h^2 = 0.1 (Table 3.5), compared to when PEV was not accounted for. Estimates of SNP effects from emBayesR with and without PEV were plotted against estimates of SNP effects from BayesR (Figure 3.5A). Estimates of SNP effects from emBayesR without accounting for PEV were considerably shrunken, particularly for small effects, compared with estimates of SNP effect from BayesR. Estimates of SNP effects with emBayesR when PEV were accounted for were much closer to those from BayesR, although there was still some over-shrinkage, particularly of small effects. Figure 3.5B, in which estimates of SNP effects obtained with BayesR, emBayesR, emBayesR_without_PEV were plotted, illustrates this result.

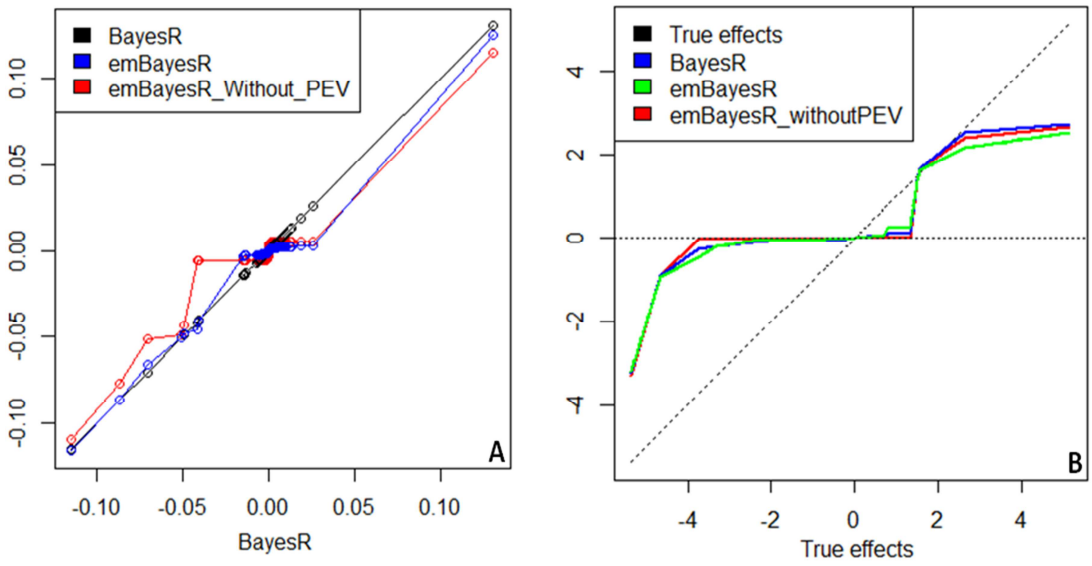


Figure 3.5. Comparison of SNP effect estimates from emBayesR with and without accounting for PEV with estimates from BayesR.

A: The x axis represents BayesR estimates of SNP effects; blue line plots emBayesR estimates of SNP effects on BayesR estimates of SNP effects; red line plots emBayesR_Without_PEV estimates of SNP effect on BayesR estimates of SNP effects; black line plots BayesR estimates of SNP effects against themselves.

B: The x axis represents true effects; blue line plots BayesR estimates of SNP effects on true effects; green line plots emBayesR estimates of SNP effects on true effect; red line plots emBayesR_without_PEV estimates of SNP effects on true effects; black line plots true effects against themselves.

Table 3.5. Accuracy of genomic prediction from emBayesR_without_PEV and emBayesR on HD_Mix dataset.

Five replicates with h^2 = 0.45 (HD_Mix_45)	**Correlation (GEBV,TBV)**				
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5
emBayesR_without_PEV	0.91	0.90	0.85	0.90	0.91
emBayesR	**0.97**	**0.96**	**0.93**	**0.97**	**0.97**
Five replicates with h^2 = 0.10 (HD_Mix_10)	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5
emBayesR_without_PEV	0.89	0.82	0.87	0.81	0.79
emBayesR	**0.91**	**0.87**	**0.93**	**0.86**	**0.87**

We also compared the accuracy of prediction based on the joint posterior mean (Equation 8b) versus the mode (Equation 8a) in the simulated data (Table 3.6). As shown in Table 3.6, using either the mean (emBayesR_Mean) or the mode (emBayesR_Mode) for estimates of SNP effect gave similar prediction accuracies.

Table 3.6. Accuracy of genomic prediction using in the algorithm posterior mode (emBayesR_Mode, Equation 8a) or posterior mean estimates of SNP effects (emBayesR_Mean, Equation 8b), in the HD_Mix dataset.

	Correlation (GEBV,TBV)				
Five replicates with h^2 = 0.45	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5
emBayesR_Mode	0.97	0.96	0.93	0.97	0.97
emBayesR_Mean	0.97	0.95	0.93	0.97	0.97
Five replicates with h^2 = 0.10	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5
emBayesR_Mode	0.91	0.87	0.93	0.86	0.87
emBayesR_Mean	0.91	0.88	0.93	0.87	0.87

Accuracy of genomic prediction with emBayesR and BayesR

In the simulation data, the accuracy of genomic prediction with emBayesR was the same as with BayesR when heritability was 0.10, but 1% lower when heritability was 0.45 (Table 3.7). However, both methods resulted in GEBV that were close to unbiased, based on the regression of TBV on GEBV being close to 1, although for HD_Mix_10, the regression was 0.89 with both BayesR and

emBayesR.

Table 3.7. Accuracy of genomic prediction and the regression coefficient of true breeding value (TBV) on genomic estimated breeding value (GEBV) for different methods for the HD_Mix simulated dataset.

	Correlation (GEBV,TBV)		**Regression coefficient (TBV on GEBV)**	
	h2 = 0.45 2500 animals	**h2 = 0.10 3750 animals**	**h2 = 0.45 2500 animals**	**h2 = 0.10 3750 animals**
BayesR	0.97±0.01	0.89±0.03	1.02±0.02	1.00±0.05
emBayesR	**0.96±0.03**	**0.89±0.02**	**0.95±0.03**	**1.00±0.04**

Accuracies of genomic prediction with BayesR, GBLUP, FastBayesB, and emBayesR on the 630K dairy data were in Table 3.8. The average accuracy of genomic prediction with emBayesR across the nine dairy cattle traits was 0.4% lower than with BayesR. The accuracy with emBayesR was on average 5% better than with FastBayesB. The average accuracy of BayesR across the nine traits was 3% higher than with GBLUP, which was due to very similar accuracies for four of the nine traits, and only protein% and fat% showing clear improvements in accuracy compared to GBLUP. For these traits, several QTL with moderate to large effects are known to exist (Grisart *et al.* 2002; Blott *et al.* 2003).

Table 3.8. Accuracy of genomic prediction from GBLUP, BayesR, fastBayesB and emBayesR for the 630K dairy cattle data for production and functional traits.

	Production traits				
	Milk	**Protein**	**Fertility**	**Protein%**	**Fat%**
GBLUP	0.57	0.63	0.40	0.63	0.77
BayesR	0.63	0.64	0.41	0.79	0.83
FastBayesB	0.57	0.60	0.35	0.70	0.80
emBayesR	**0.62**	**0.65**	**0.40**	**0.76**	**0.83**
	Functional traits				
	Angularity	**Mammary conformation**	**Stature**	**Somatic cell count**	
GBLUP	0.45	0.28	0.47	0.71	
BayesR	0.44	0.28	0.47	0.71	
FastBayesB	0.39	0.25	0.43	0.61	
emBayesR	**0.45**	**0.30**	**0.47**	**0.69**	

Computing performance of emBayesR compared with BayesR

We compared the speed of emBayesR with BayesR and fastBayesB using three criteria: the time complexity of each iteration (the function in terms of number of SNPs and individuals that determines the time taken to do one iteration), the number of iterations to convergence (or in the case of BayesR until changes in SNP estimates were sufficiently small so that the accuracy of genomic prediction did not change), and total computing time required with the 630K real data.

First, as mentioned in the method section, the time complexity for emBayesR is $O(nm)$, which is the same as with the MCMC method for BayesR and with ICE iterations for fastBayesB, and with the nonlinear A method of VanRaden (VanRaden 2008) and SNP_BLUP (Meuwissen et al. 2001).

Second, for BayesR, the accuracy of prediction for the trait milk yield exceeded 0.61 at 20 000 iterations, and did not improve with a larger number of iterations, as shown in Figure 3.6. For five traits (milk, protein, fertility, fat % and protein %) and using the 630K real data, the numbers of iterations required for convergence for emBayesR and fastBayesB are given in Table 3.9. FastBayesB required slightly more iterations to reach convergence than emBayesR for most traits.

Table 3.9. Number of iterations required for emBayesR and fastBayesB to reach convergence for five traits with the 630K dairy cattle data.

	Milk	**Protein**	**Fertility**	**Protein %**	**Fat %**
EmBayesR	**460**	**476**	**920**	**572**	**496**
FastBayesB	410	540	856	848	564

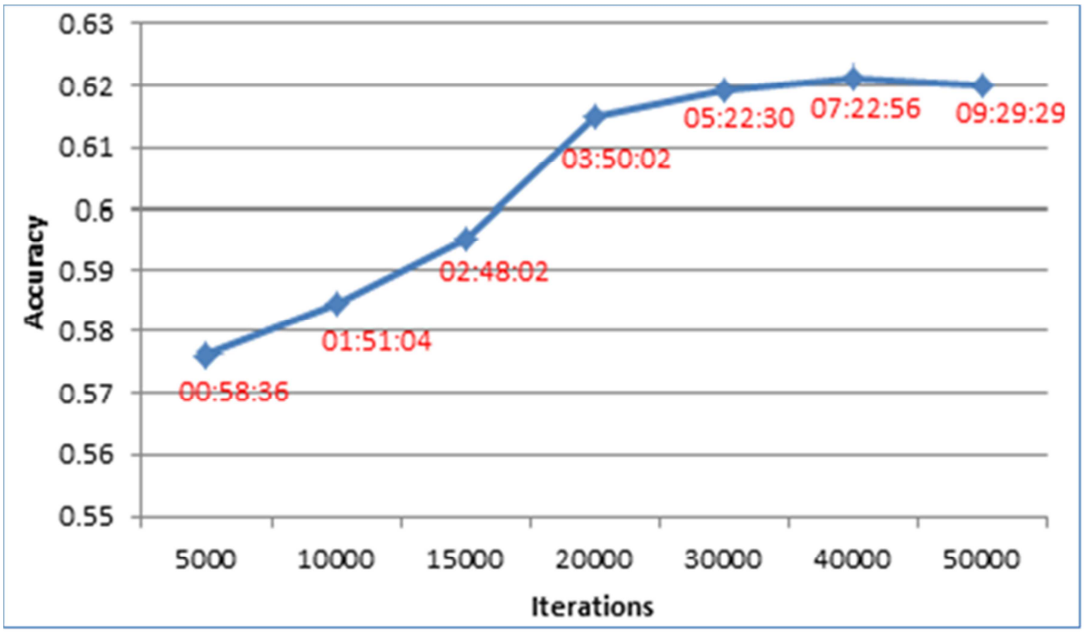


Figure 3.6. Accuracy of genomic prediction and running time for BayesR with an increasing number of iterations.

Finally, the overall computing times for emBayesR, BayesR and fastBayesB with the same implementation (each trait on one processor) were compared (Figure 3.7). The algorithms were implemented on a range of datasets with different sizes, including 10K simulated data (HD_Mix model, 2500 animals with around 10 000 SNPs), 50K data (3049 animals with 38 968 SNPs), and 630K data (3049 animals with 632 003 SNPs). As shown in Figure 3.7, the speed advantage of emBayesR compared to BayesR was greater as the number of SNPs in the dataset increases. For example, with the 630K data, BayesR needed approximately 4 days of real computing time, while emBayesR required just 4 hours (including the time to calculate PEV in GBLUP) to achieve the final solutions.

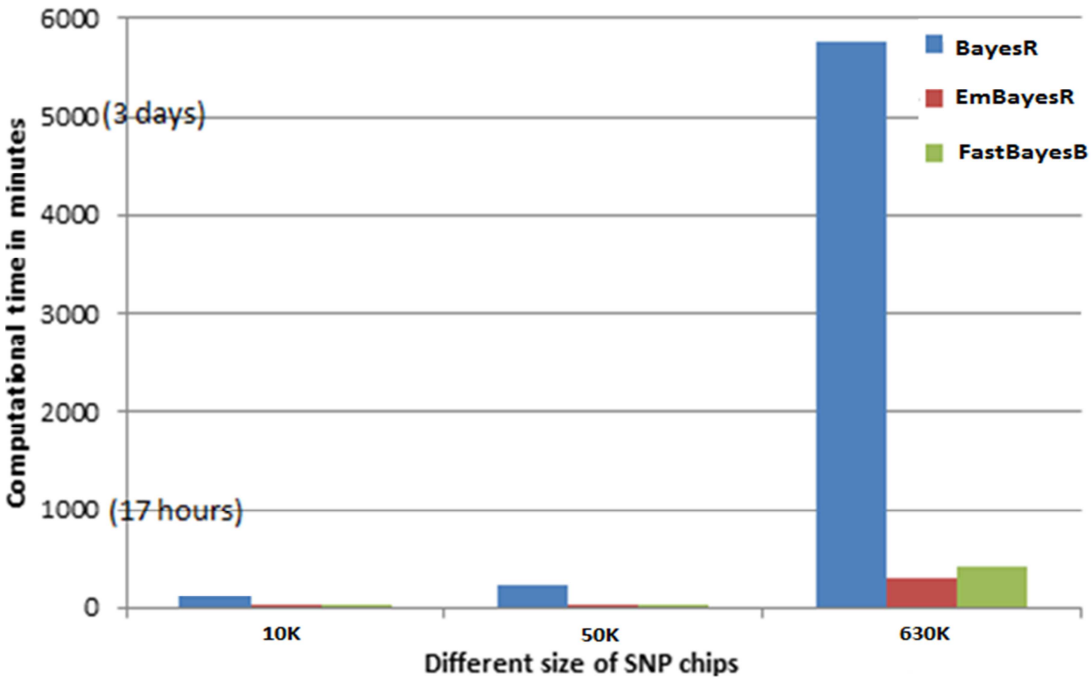


Figure 3.7. Computational time required for BayesR, emBayesR and FastBayesB on a range of SNP chips (10K, 50K and 630K).

The x axis represents the different sizes of the SNP chips, y axis is the computational time in minutes; blue bar is BayesR's running time; red bar is emBayesR's; green bar is FastBayesB's computing time.

Sensitivity of parameter estimates from emBayesR to the underlying genetic model

In this final Results section, we investigate the sensitivity of the accuracy of genomic prediction and estimates of **Pr** with emBayesR and BayesR to the underlying data model. Three underlying models for QTL effects were investigated: (1) an equal mixture of three non-zero normal distributions in HD_Mix; (2) all QTL effects follow a normal distribution in HD_One; and (3) an unknown model of QTL effects in the 630K real data.

Table 3.10. Estimated mixing proportions (Pr) and genomic prediction accuracy from BayesR, emBayesR and GBLUP with the HD_Mix_45 and HD_One_45 datasets.

HD_Mix_45 (h^2 = 0.45)		
	Pr	**Accuracy**
True	[0.9950 0.0017 0.0016 0.0017]	
BayesR	[0.9861 0.0127 0.0004 0.0008]	0.97
emBayesR	[0.9852 0.0136 0.0003 0.0009]	0.97
GBLUP	-	0.67
HD_One_45 (h^2 = 0.45)		
	Pr	**Accuracy**
True	[0 0 0 1]	
BayesR	[0.722 0.2621 0.0115 0.0044]	0.80
emBayesR	[0.012 0.986 0.0007 0.0013]	0.80
GBLUP	-	0.78

emBayesR and BayesR gave higher accuracies than GBLUP for the HD_Mix model data (M45_2), while for the HD_One data, the advantage of emBayesR and BayesR was smaller than that of GBLUP (Table 3.10), as might be expected given that the HD_Mix data had a proportion of QTL with larger effects. In estimating **Pr**, emBayesR generally had somewhat poorer agreement with the underlying data model than BayesR (Table 3.10), especially for the HD_One_45 data.

However, on 630K real data, emBayesR gave very similar estimates of **Pr** and accuracy of genomic prediction than BayesR and GBLUP (accuracy only for the later comparison) (Table 3.3 and Table 3.8). One conclusion from the relative performance of emBayesR to BayesR in the 10K simulated data and in the 630K real data, is that emBayesR could not distinguish SNP effects with zero variance from those with a very small variance when there is little information in small datasets, as in the HD_One simulated data. However, among the 630K SNPs there are likely more SNPs in the non-zero distributions, which should increase

the precision of estimates of **Pr.**

3.6 Discussion

Genomic prediction with non-linear Bayesian methods (under MCMC model), including BayesR, can be more accurate than GBLUP in some situations, such as when QTL with moderate to large effects segregate (VanRaden 2008; Yang *et al.* 2010), but at the cost of longer computing time. To retain the accuracy of BayesR while reducing computing time, we propose here an EM algorithm, termed emBayesR, for genomic prediction, as an alternative to the MCMC implementation of BayesR. In both 10K SNP simulated data and 630K real dairy cattle data, emBayesR gave accuracies of genomic prediction similar to BayesR, with greatly reduced computing time. As in BayesR, emBayesR estimates SNP effects, error variances and posterior probabilities of each SNP belonging to the k^{th} distribution (here, there were four distributions, one with zero variance).

Results from BayesR and emBayesR differed in three ways, albeit to a small degree. Estimates of **Pr** with emBayesR tended to have more SNP effect estimates in the smallest non-zero distribution than BayesR; emBayesR shrunk small SNP effects towards 0 somewhat more than BayesR; and the accuracy of emBayesR predictions was approximately 0.5% lower than the accuracy of BayesR. Our EM algorithm differed from the MCMC BayesR in several respects, which may explain these results. The EM algorithm estimates the SNP effect (g_i) by the mode of the posterior distribution when the mixing proportions (**Pr**) and the error variance (σ_e^2) are held at their MAP estimates, whereas the MCMC version estimates g_i by the mean of the posterior distribution while **Pr** and σ_e^2 vary over their posterior distributions. Also, when we used the mean instead of the mode of the posterior distribution of g_i as an estimate of g_i, we found that it makes no discernible difference in prediction accuracy, as shown in Table 3.6. However, varying **Pr** and σ_e^2 across their posterior distributions in BayesR, but not

emBayesR, may explain differences in results. In addition, emBayesR uses an approximation of the prediction error variance of all other SNPs when estimating g_i.

Bayesian estimates are sensitive to the prior if the data does not contain enough information to overwhelm the prior. Estimates of **Pr** with both BayesR and emBayesR were affected by the prior **α** but not to a large degree, considering that the simulated data contained only 50 causal mutations and the prior had little effect on the accuracy of genomic predictions. Results from using emBayesR with the simulated data indicate the algorithm was unable to consistently distinguish a SNP with no effect from a SNP with a very small effect. We would expect that, in data in which more causal mutations are segregating and with many more animals, estimates of **Pr** would be less sensitive to the prior.

Other EM algorithms for genomic prediction have been described using thick-tailed *t*-distributions or exponential distributions as priors for the SNP effects. These include EM-BSR (Hayashi & Iwata 2010) and FastBayesA (Sun *et al.* 2012), which aim at enhancing the computing efficiency of BayesA. emBayesR differs from most previous non-MCMC implementations of Bayesian methods for genomic prediction in two respects, i.e. it uses the BayesR model with a mixture of four normal distributions for SNP effects and it accounts for errors in all other estimated SNP effects when estimating the effect of the current SNP by including the PEV term in the model. When we implemented the EM algorithm without the PEV term, the accuracy of prediction declined by 8%. The accuracy of fastBayesB was, on average, 9% lower than that of emBayesR, suggesting that much of the loss in accuracy of fastBayesB is due to ignoring the errors in all other SNP effects when estimating a particular SNP effect. Consistent with this interpretation, both fastBayesB and our EM algorithm without accounting for the PEV shrink estimates of SNP effects more severely than emBayesR or BayesR. Most of the current fast algorithms, such as fastBayesB (Meuwissen *et al.* 2009), emBayesB

(Shepherd *et al.* 2010), em_BSR (Hayashi & Iwata 2010), and MixP (Yu & Meuwissen 2011), ignore the error produced by the estimation of other SNP effects. That is, they use an unrealistic assumption that the current solutions of all other SNPs effects are known without error when estimating the current SNP effect, which is one of the reasons why accuracies of prediction from these algorithms are typically lower than that of their counterpart MCMC methods. MCMC methods account for the error in the estimates of other SNP effects by sampling them from their posterior distributions. For the calculation of PEV, the inverse of a matrix with dimensions (number of animals $\times$ the number of animals) is required (Equation (A7) of Additional file 2). When the number of animals exceeds 50 000, this will hinder the computing efficiency of emBayesR. To reduce the computing burden of the PEV calculation, the efficient genomic recursion algorithms proposed by Misztal et al. (Misztal *et al.* 2014) could be applied but this requires further investigation.

Our results demonstrated the computing speed of emBayesR over the MCMC implementation of BayesR. The time complexity for emBayesR at each iteration is proportional to the number of markers and the number of records, as it is in the MCMC methods. However, much fewer iterations were required for the emBayesR SNP effects to converge than for BayesR to sample sufficiently from the posterior distributions of SNP effects to achieve maximum accuracy of genomic prediction. Specifically, compared with 20 000 iterations of MCMC sampling (Figure 3.6), emBayesR required only 300 to 1000 iterations with the 630K real dairy data (Table 3.9). As the size of datasets increased, this advantage could be even greater, as shown in Figure 3.7.

With high-density SNP data (630K), the prediction accuracy of emBayesR and BayesR was greater than GBLUP only for yield traits. Similar results (an advantage of a Bayesian approach over GBLUP for yield traits only) were obtained using the nonlinear iterative A method with imputed high-density data

from 15 842 reference animals and 28 traits (VanRaden *et al.* 2013). Computing time with high-density data for this nonlinear A method is also $O(nm)$, with reported times similar to emBayesR. One difference between BayesR and the nonlinear A method is that SNP effects can actually be 0 with BayesR, whereas in nonlinear A, SNPs will always have a non-zero effect, although it may be very small. This difference between the algorithms apparently does not affect accuracies of prediction with the 630K real data, although it may become more important with whole-genome sequence data, for which the number of variants is much larger. However, this is yet to be demonstrated.

It should also be noted that some reduction in computing time can be achieved by "pruning" SNPs that are in very high linkage disequilibrium from the dataset, since these SNPs carry redundant information. For example, Su et al (Su *et al.* 2012) reduced a dataset from 770K to 492K SNPs by pruning SNPs that were in very high linkage disequilibrium in a Nordic Holstein population prior to estimation of SNP effects.

Our aim is to eventually integrate emBayesR into genetic evaluations for Australian dairy cattle. Currently, the Australian National DNA reference population has more than 20 000 cattle, including 3719 Holstein bulls, 9630 Holstein cows, 1017 Jersey bulls and 4249 Jersey cows. For the evaluation of these national reference populations, GBLUP is currently used to calculate the Australia Genomic Breeding Value on 50K SNP genotypes. However, even with the current data, prediction accuracy is higher with Bayes R than with GBLUP for some traits and GBLUP is unable to take advantage of the extra information that would be contained in whole-genome sequence data. Therefore, we anticipate moving to a Bayesian method to take advantage of whole-genome sequence data and increase prediction accuracies, and we expect that an EM algorithm will be part of this methodology in order to limit computing time.

In this paper, we used only bulls in the reference and validation sets, to avoid the added complexity of weighting bull and cow trait deviations differently. However, further development of the method described in this paper is needed to include appropriate weighting of phenotypes, multi-breed effects, polygenic effects in the model (as implemented in the MCMC version (Kemper *et al.* 2015)) and to imbed the Bayesian method within a single-step genetic evaluation (Fernando *et al.* 2014b; Liu *et al.* 2014), so that it can be applied to the Australian national dairy evaluations. Also, efficient approaches for inversion of the animal by animal matrix to obtain the PEV need to be investigated to retain the efficiency advantage of emBayesR with very large numbers of animals.

3.7 Conclusions

EmBayesR uses an EM-based method to estimate the posterior mode of SNP effects, rather than the MCMC sampling used in BayesR. emBayesR can reduce computing time up to 8-fold compared to BayesR. Results with simulated data and real 630K SNP dairy cattle data show that genomic prediction accuracy of emBayesR is similar to that of BayesR (0.5% accuracy loss averaged over traits). The computing advantages of emBayesR make it attractive for implementation of genomic prediction in very large datasets.

3.8 Supporting information

All the supporting files were located in Appendix II (0) as follows:

File S1 - Calculation of $\mathrm{P_{ik}} = E(\mathrm{b_{ik}}|\mathrm{y}, \widehat{\mathrm{Pr}_k})$, which includes the details on how to derive $\mathbf{P_{ik}}$.

File S2 - PEV calculation from GBLUP, which describes the details on how to calculate PEV from GBLUP.

3.9 Acknowledgements

The authors acknowledge the support and fund from Dairy Future CRC. We would like to thank Iona Macleod (Department of Environment & Primary Industries (DEPI), 5 Ring Road, Bundoora, VIC 3083, Australia) for her work on 10K simulation.

Chapter 4 Computationally efficient schemes for multi-population genomic prediction and QTL mapping for complex traits

4.1 Chapter preface

Justification

As data sets for genomic prediction of complex traits rapidly expanded in size, the importance of computational efficiency of genomic prediction algorithms became paramount. In this paper, we described an expectation-maximization algorithm for genomic prediction (Opt_emBR) that was up to 30 times faster than MCMC implementations. The algorithm was suitable for joint analysis of data from different experiments, as it accommodated heterogeneous variances, and could accommodate effects of factors such as age, sex and additional covariates. A further advantage of the method was that QTL mapping was performed simultaneously with genomic prediction.

Publication status:

Published in the conference *the Association for the Advancement of Animal Breeding and Genetics, 2015*.

Published as

Wang TT, Chen YPP, Kemper KE, Goddard M, Hayes BJ. (2015) Opt_emBR: Computationally efficient genomic prediction and QTL mapping in multi-breed populations. Proceeding of the Association for the Advancement of Animal Breeding and Genetics, 21: 449-452.

Statement of contributions of joint authorship

Tingting Wang (Candidate): implemented the extended version of emBayesR methods with two speed-up schemes, analyzed the data for multi-breed and across-breed prediction and then drafted the manuscript.

Yi-Ping Phoebe Chen (Principle Supervisor): supervised the manuscript.

Kathryn E Kemper (Collaborator) implemented BayesR on 630 K high-density SNP panel for multi-breed and across-breed prediction.

Michael E Goddard (Collaborator): contributed the valuable idea of the speed-up schemes for the algorithm.

Ben J. Hayes (Co-Supervisor): supervised this project, suggested approaches for the implementation of the algorithm on real data, and gave a great contribution for the organizing/revising of the paper.

This chapter is an extended version of paper submitted to AAABG conference, including the more detailed methodology and more comprehensive results on the prediction accuracy and QTL mappings. The reference style, table numbers and figure numbers have been carefully formatted.

4.2 Abstract

Genomic prediction is increasing widely used in selection of livestock and crops, as well as for prediction of disease risk in species including humans. Prediction using multiple populations (and multiple data sets from different experiments) for the same species can result in higher accuracies of prediction and more precise QTL mapping, provided denser SNP and methods that allow a non-linear model of SNP effects are used. In particular, methods that allow a proportion of SNPs to be excluded from the model are necessary to take advantage of very dense SNP

genotypes. These methods are usually Bayesian and implemented through Markov Chain Monte Carlo (MCMC) sampling. However, MCMC becomes impractical with very large numbers of SNP and larger number of individuals, particularly when datasets are combined across populations. In this paper, we propose a computationally efficient scheme termed Opt_emBR (Optimized emBayesR) for multi-population genomic prediction and simultaneous QTL mapping. The method implements an expectation-maximization algorithm for a non-linear prediction model (BayesR). To increase the range of situations for which the method can be applied, we include a polygenic term to take into account genetic variance not explained by the SNP, but which can be captured by a pedigree. As well, in order to correctly model heterogeneous variances of trait observations that may come from different sources, weights are introduced into the mixture linear model. Two potential speed-up schemes for Opt_emBR are evaluated to reduce time taken for operations on large matrices of genotype data as well as to reduce the times of basic operations. The results in a large data set with individuals from two breeds (Holstein and Jersey) of dairy cattle and genotyped for 630K SNPs show the robust prediction ability of Opt_emBR for within-population, and multi-populations prediction when compared with the MCMC implementation of the same model. Moreover, the speed up schemes can make Opt_emBR up to 30 times faster than the MCMC implementation. This computational efficiency will be increasingly important as the size of genomic data sets continues to increase. Finally, we were also able to demonstrate that Opt_emBR can be used for QTL mapping and genomic prediction simultaneously, with more precision of QTL mapping than by standard GWAS approaches.

4.3 Introduction

Genomic prediction is increasingly used for prediction of disease risk in humans, and for selection programs for livestock and crops (de los Campos *et al.* 2010;

Yang *et al.* 2010). These prediction methods use genome-wide panels of SNPs to exploit linkage disequilibrium between these SNPs and the causal mutations affecting traits (Meuwissen *et al.* 2001). Genomic prediction requires deriving the prediction equation, the effect of each SNP on the trait, by estimating the SNP effects in reference population of individuals that are both genotyped for the SNP and have phenotypes for the trait of interest. For example, 45% of genetic variance for human height can be explained by common SNPs using linear genomic prediction model termed GBLUP, and accuracies of genomic prediction approach 0.5 (Yang *et al.* 2010). At the same time, genomic predictions of disease risk have been demonstrated with some accuracy for several human diseases e.g. Crohn's disease, Celiac disease and Type I(II) Diabetes (Barrett *et al.* 2008; Zhou *et al.* 2013b; Abraham *et al.* 2014; Speed & Balding 2014; Loh *et al.* 2015; Moser *et al.* 2015). In dairy cattle, many breeding programs are now based on selecting bulls for breeding on the basis of their genomic predictions (of genetic merit) for traits such as milk yield, protein yield, and fertility (Goddard & Hayes 2009; Meuwissen & Goddard 2010).

The accuracy of genomic prediction is jointly determined by the size of reference population, extent of linkage disequilibrium (LD) between SNP and causal mutations, heritability of the trait, and method of deriving the prediction equation (Hayes & Goddard 2008; Goddard 2009; Daetwyler *et al.* 2010). Across many species, a key finding is that reference populations must be very large to achieve high accuracies of genomic prediction, reflecting the limited extent of LD in these species, and large number of loci affecting most complex traits (Daetwyler *et al.* 2010; Hayes *et al.* 2010). One way to increase the size of the reference population is to combine information across populations from the same species, which can be termed multi-population prediction. There have been a wide range of previous studies attempting multi-population implementations in both livestock (e.g. dairy cattle (Habier *et al.* 2007; Erbe *et al.* 2012; Saatchi *et al.* 2013; Zhou *et*

al. 2013a; Hozé *et al.* 2014; Lund *et al.* 2014; Kemper *et al.* 2015), beef cattle (Weber *et al.* 2012; Bolormaa *et al.* 2013), sheep (Daetwyler *et al.* 2012a; Daetwyler *et al.* 2012b) and human (Wray *et al.* 2007; Visscher 2008; Wood *et al.* 2014)). In general, the finding from these studies is that small to moderate increases in genomic prediction accuracy for some traits can be achieved by combining the populations, provided SNPs are sufficiently dense that SNP-QTL associations persist across the populations (Hayes *et al.* 2010; Bolormaa *et al.* 2013; Hozé *et al.* 2014; Lund *et al.* 2014).

The potential multi-population prediction methodology includes linear and nonlinear models, which use the same Bayesian regression theory, but have different prior assumptions for SNP effects. In detail, linear models (e.g. SNP-BLUP or the mathematically equivalent GBLUP) (Meuwissen *et al.* 2001; Yang *et al.* 2010) assume the same variances for all the SNPs, with a Gaussian prior. On the contrary, nonlinear models (also termed Bayesian alphabet e.g. BayesA, B, C, D, R etc.) (Meuwissen *et al.* 2001; Habier *et al.* 2011; Erbe *et al.* 2012; Zhou *et al.* 2013b; Speed & Balding 2014) assume the SNPs are not normally distributed with different genetic variances varying across the chromosome. Compared with nonlinear models, the use of a normal prior in BLUP leads to drastic shrinkage of marker effects. As the results, the estimated effect of the causative mutation from BLUP will be “smeared” across many markers (e.g. Verbyla et al. (Verbyla *et al.* 2009), Kemper et al. (Kemper *et al.* 2015)). While these associations across many markers can hold for a number of generations within a population, they are much less likely to persist across populations (for example across breeds). Therefore, for across population prediction, SNP-BLUP or GBLUP might reduce up to 20% accuracies in contrast with nonlinear methods (non-Gaussian priors) (Hozé *et al.* 2014; Zhou *et al.* 2014; Kemper *et al.* 2015). Further, compared to BLUP methods, nonlinear models use priors, which assume a large proportion of SNP, have effects to close to zero, or actually are zero, while

a proportion of SNP have moderate to large effects. This is important not only to improve the accuracy of genomic predictions, but also to improve the precision of QTL mapping (Sun *et al.* 2011; Kemper *et al.* 2015; Moser *et al.* 2015). In contrast with other non-linear models, BayesR assumes the mixture of four normal priors for SNP effects with variances ranging from zero to three increasing variances, which leads to more straightforward computation than is possible with t distributions of effects (Erbe *et al.* 2012). As demonstrated by Moser et al. (2015), BayesR can accurately estimate genetic architecture (heritability) related to popular human diseases. Especially, for human diseases controlled by major loci (e.g. type 1 diabetes and rheumatoid arthritis), the prediction analysis and QTL mapping from BayesR outperforms BLUP method and Bayesian sparse linear mixed model (BSLMM) (Zhou *et al.* 2013b). More precise QTL mapping leads to more accurate genomic predictions across populations, as demonstrated by Kemper et al. (2015).

While the Bayesian methods are very attractive, the major difficulty with these methods is long computation time, which with very large data sets becomes intractable. The methods are typically implemented using MCMC sampling from the posterior distributions of the parameters, which leads to long run times. To speed up Bayesian methods, several heuristic convergence methods have been proposed (VanRaden 2008; Meuwissen *et al.* 2009; Hayashi & Iwata 2010; Shepherd *et al.* 2010; Yu & Meuwissen 2011; Sun *et al.* 2012; Wang *et al.* 2015) (e.g. Iterative Conditional Expectation methods and Expectation-Maximization algorithms). For example, VanRadan et al. (2008) proposed the methods termed nonlinear A and B to mimic the nonlinear shrinkage of BayesA and BayesB. Jacobi iteration was implemented on nonlinear A and B to be the approximations of both BayesA and BayesB. Meuwissen et al. (2009) and Yu et al. (2011) applied iterative conditional expectation (ICE) to the BayesLASSO model and Bayesian mixture models of two normal distributions separately. Also, an expectation-

maximization (EM) algorithm was introduced to maximize a joint posterior probability by a number of methods termed EmBayesB (Shepherd *et al.* 2010), wBSR (Hayashi & Iwata 2010), fastBayesA (Sun *et al.* 2012), and emBayesR (Wang *et al.* 2015), which are based on the prior distribution of SNP effects from BayesLASSO, BayesA, and BayesR models respectively. All of these methods were reported several orders faster than their counterparts while retaining the similar level of prediction accuracy on the simulation data or medium density of SNP panels. However, except for nonlinear A (B) and emBayesR, few fast versions of Bayesian methods are implemented on practical data of high density, not mentioned to be used for multi-populations prediction, as they generally lead to lower accuracies of prediction.

Our aim is to develop a computationally efficient algorithm (Opt_emBR for Optimized BayesR) for simultaneous multi-population genomic prediction and QTL mapping. Similar to emBayesR (Wang *et al.* 2015), Opt_emBR implements an optimized EM algorithm on the prior assumption for SNP effects and other parameters from BayesR, that is all SNP effects follow four normal distributions with the variances $0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2$ (σ_g^2 means the genetic variance). Also, Opt_emBR retains the advantage of Predicted Error Variance (PEV) correction of emBayesR (Wang *et al.* 2015) to improve the accuracy of genomic prediction. For the application on multi-population genomic prediction and QTL mapping, and application in very large data sets, Opt_emBR has four improvement compared with emBayesR (Wang *et al.* 2015): 1) weighting on phenotypes to allow for different errors in measurement across populations, or for example for combining bull phenotypes made up of with many daughter records with single cow records; 2) accommodation of fixed effects (e.g. breed, sex) into the prediction models; 3) polygenic effects to take advantage of genetic variation not captured by the SNP but captured by pedigree; 4) and two speed-up schemes to make it 30 times faster than BayesR implemented with MCMC. To evaluate the

prediction ability and computation efficiency of Opt_emBR, a data set of 630K SNP genotyped in 16,214 of three breeds of Australian dairy cattle was used.

4.4 Methods and Materials

Similar to emBayesR (Wang *et al.* 2015), Opt_emBR implements the EM algorithm on the mixture Gaussian model of BayesR to maximize the expectation of a posterior probability for SNP effects and all other parameters. In the following, we will start with the data model of Opt_emBR, including the description for various prior assumptions of the grouped parameter sets θ from the model. Then, EM derivation for the parameters set based on different priors is detailed. Next, two speed-up optimized schemes will be mainly discussed. Furthermore, the overall algorithm in terms of pseudo code is listed step by step. Afterwards, the genomic data used to test/demonstrate the algorithms will be described.

The statistical model of Opt_emBR

Compared with emBayesR, Opt_emBR extends its models by including fixed effects ($\boldsymbol{\beta}$), polygenic effects ($\mathbf{v}$), and error matrix ($\mathbf{E}$) which incorporated weights to reflect heterogeneous variances across phenotypes, while maintaining the aim of emBayesR which was to estimate each SNP effect $\mathrm{g_i}$ with the consideration of prediction error variance (PEV) produced by other SNPs. Therefore, aiming at the estimation for single SNP effect, the data model of Opt_emBR assumes that the phenotypic records of n individuals ($\mathbf{y}$), regresses fixed effects ($\boldsymbol{\beta}$), each SNP effect (g_i), random genetic effects ($\mathbf{u}$), random polygenic effects($\mathbf{v}$) and environmental errors ($\mathbf{e}$) together via

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z_i}g_i + \mathbf{u} + \mathbf{W}\mathbf{v} + \mathbf{e}, \quad (1)$$

where, $\mathbf{X}$ is the $n \times p$ design matrix, allocating phenotypes $\mathbf{y}$ to fixed effects $\boldsymbol{\beta}$ of breed and sex. p was the number of fixed effects.

$\mathbf{Z_i}$ is $n \times 1$ standardised genotype vector for each SNP i, which was each

column of $n \times m$ standardised genotype matrix $\mathbf{Z}$ with $i = 1, \ldots, m$.

$\mathbf{W}$, the $n \times q$ design matrix, allocates the $q \times 1$ vector of polygenic effects to the phenotypes $\mathbf{y}$.

The parameters set from the data model (1) includes fixed effects ($\boldsymbol{\beta}$), each SNP effect ($\mathrm{g_i}$), random genetic effects ($\mathbf{u}$), random polygenic effects ($\mathbf{v}$) and environmental errors ($\mathbf{e}$), written as $= \{g_i, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \mathbf{e}\}$. Here, for the clarification, we group the parameter set θ as the one related with SNP effects $\theta_1 = \{g_i, \ldots\}$, the one related with genomic breeding values $\mathbf{u}$ by other estimated SNPs effects which was introduced for PEV calculation , and the set with other parameters $\theta_2 = \{\boldsymbol{\beta}, \mathbf{v}, \mathbf{e}\}$.

For the parameter set $\theta_1$related to SNP effects, the prior distribution of SNP effects was assumed to be the mixture of four normal distributions with the variances, i.e. $\boldsymbol{\sigma_i^2} = \{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2\}$, where σ_g^2 is genetic variance. Therefore, the prior assumption of each SNP effect g_i can be written as $g_i \sim N(0, \boldsymbol{\sigma_i^2})$. Since there are four possible normal distributions for each SNP, three other derivative parameters including $b(i,k)$, $p(i,k)$ and Pr_k are required. In detail, $b(i,k)$, determines whether or not each SNP effectg_i belongs to one of four normal distributions k $(k = 1,2,3,4)$. The term $p(i,k)$ defines the probability of each SNP effect g_i in one of four normal distributions k. The term Pr_k is the proportion of all the SNPs following each of four normal distributions k, that is assumed as Dirichlet distribution with the parameter $\alpha = (1,1,1,1)^{\mathrm{T}}$. Therefore, the parameter set θ_1 for SNP effects also includes $b(i,k)$ and $\mathbf{Pr}$, written as $\theta_1 = \{g_i, b(i,k), \mathbf{Pr}\}$. The prior distribution of each SNP effect conditional on $b(i,k)$ is

$$\mathrm{P}\left(g_i \middle| b(i,k)\right) = \begin{cases} 0, & k = 1 \\ \frac{1}{\sqrt{2\pi\sigma_{\mathrm{i}}^2[k]}} \exp\left(-\frac{g_i^2}{2\sigma_{\mathrm{i}}^2[k]}\right), & k = 2,3,4. \end{cases}$$ Then, the joint distribution $\mathrm{p}\left(g_i, \mathbf{b_i} \middle| \widehat{\mathbf{Pr}}\right)$ conditional on $\widehat{\mathbf{Pr}}$ can be written as:

$$p(g_i, \mathbf{b_i}|\mathbf{Pr}) = \prod_{k=1}^{4} p\left(g_i \middle| b(i,k)\right) \times p(b(i,k)|Pr_k)$$

$$= \prod_{k=1}^{4} \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{g_i^2}{2\sigma_k^2}\right) Pr_k\right)^{b_{ik}}.$$

The parameters sets θ_2 from other effects includes the fixed effects $\boldsymbol{\beta}$, the polygenic effects $\mathbf{v}$, and the environmental errors $\mathbf{e}$. The $\boldsymbol{\beta}$ have uninformative priors, while the prior distributions for two other random effects are as follows: $\mathbf{v} \sim N(0, \mathbf{A}\sigma_a^2)$; $\mathbf{e} \sim N(0, \mathbf{E}\sigma_e^2)$ with σ_a^2 and σ_e^2 stand for polygenic variance and error variance, respectively. Moreover, $\mathbf{A}$ is $q \times q$ pedigree-based relationship matrix (q is the number of individuals in the pedigree). The $n \times n$ diagonal matrix $\mathbf{E}$ is especially designed to evaluate the different contributions of the phenotype records from different sex to the error variance. Each diagonal element e_{ii} of matrix $\mathbf{E}$ is equal to $1/w_i$, where w_i is used to weight the bull and cow records appropriately (Garrick *et al.* 2009). The weight of bulls and cows was calculated using the equations:

$$w_i(bulls) = \frac{(1-h^2)}{ch^2+(4-h^2)/d}, \text{ and } w_i(cows) = \frac{(1-h^2)}{ch^2+[1+(r-1)t]/r-h^2}, \qquad (2)$$

where, h^2 is the heritability of the trait; t is the repeatability of the traits; d is the number of the daughter of each bulls; r is the number of records; c is the proportion of genetic variance not accounted for by the SNP (Garrick *et al.* 2009).

The latent parameter, $\hat{\mathbf{u}}$ is derived by the equation $\mathbf{u} = \sum_{j \neq i} \mathbf{Z_j} g_j$, which is introduced to derive the prediction error variance (PEV) instead of being estimated from the posterior. Similar to emBayesR (Wang *et al.* 2015), Opt_emBR derives (PEV) from the estimation of all the SNP effects. i.e. $\mathrm{PEV} = \mathrm{var}(\hat{\mathbf{u}} - \mathbf{u})$. Under GBLUP model, $\hat{\mathbf{u}}$ is assumed to follow $N(0, \mathbf{G}\sigma_g^2)$, where $\mathbf{G}$ is the genomic relationship matrix (GRM) (VanRaden 2008; Yang *et al.* 2010). The PEV matrix can be calculated under the GBLUP model, which will be approximately implemented to the model (1) to correct estimation for each SNP effect.

$$\mathrm{PEV} = (\mathbf{E}^{-1}\sigma_e^{-2} + (\mathbf{G}\sigma_g^2 + \mathbf{WAW}'\sigma_v^2)^{-1})^{-1} \quad (3)$$

EM derivation of Opt_emBR

Similar to emBayesR, estimation of all the parameters sets θ including each SNP effect (g_i), the mixing proportions (**Pr**), fixed effects (**β**) and polygenic effects (**v**) as well as the error variance (σ_e^2) will be performed with maximum a posteriori (MAP) estimation under different prior assumptions. All the parameters were derived according to the similar expectation-maximization process with steps:

i) Define the log likelihood of the data under the model (1), which is based on the distribution $(\mathbf{y} - \hat{\mathbf{u}} - \mathbf{X\beta} - \mathbf{Wv}) \sim N(0, \mathbf{Z_i}\sigma_i^2\mathbf{Z_i'} + \mathbf{E}\sigma_e^2)$. Hence, the full log likelihood can be represented as:

$$logL = -0.5\ (\ log|\mathbf{H_k}| + ((\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}} - \hat{\mathbf{u}} - \hat{\mathbf{v}})'\mathbf{H_k}^{-1}(\mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}} - \hat{\mathbf{u}} - \hat{\mathbf{v}})), \quad (4)$$

where, $\mathbf{H_k} = \mathbf{Z_i}\mathbf{Z_i}'\boldsymbol{\sigma_i^2} + \mathbf{E}\sigma_e^2$.

ii) Derive the log posterior function of the parameters using Bayes' theorem. Following Bayes' theorem, the log posterior distribution of the parameters sets θ is based on the rule $logp(\theta|'data') \propto logf('data'|\theta) + logp(\theta)$ with $logf('data'|\theta) = logL$; $logp(\theta) =$ the prior for specific parameter. See below for log posterior functions of each parameter.

iii) Take the expectation on the posterior function with the missing data $\hat{\mathbf{u}}$ (the breeding values from other SNP effects) and $b(i,k)$ (whether or not the SNP i follows the normal distribution k of four);

iv) Differentiate the expected posterior function regarding the required parameters set θ, and the latent parameters.

According to the above four step-wise process, the posterior estimation of

parameters sets θ_1, θ_2 can be derived. In the following, we will firstly brief the estimation for the SNP effects related parameters set θ_1 which follows the same steps as emBayesR; Then, the parameters set θ_2 extended by Opt_emBR will be derived.

Then, we derive the posterior probability that each SNP is from each of the four distributions according to the equation $\mathrm{P}(i,k) = E_{\mathbf{b}} logp(b(i,k)|\mathbf{y}, \widehat{\mathrm{Pr}_{\mathrm{k}}})$ with $logp\big(b(i,k)\big|\mathbf{y}, \widehat{Pr_k}\big) = logL + logp\big(b(i,k) = 1|\widehat{\mathrm{Pr}_{\mathrm{k}}}\big)$, where $log\mathrm{p}\big(b(i,k) = 1|\widehat{\mathrm{Pr}_{\mathrm{k}}}\big) = log\mathrm{Pr}_{\mathrm{k}}$. Then, after treating $\hat{\mathbf{u}}$ as the missing data (introducing PEV term), $logL(i,k) = E_{\hat{\mathbf{u}}} logp\big(b(i,k)\big|\mathbf{y}, \widehat{\mathrm{Pr}_{\mathrm{k}}}\big) = E_{\hat{\mathbf{u}}} logL + logp\big(b(i,k) = 1|\widehat{\mathrm{Pr}_{\mathrm{k}}}\big)$ as follows:

$$logL(i,k) = -0.5\big\{log|\mathbf{H}_{\mathrm{k}}| + \mathbf{y}^{\dagger\prime}\mathbf{H_k}^{-1}\,\mathbf{y}^{\dagger} + \mathrm{tr}(\mathbf{H_k}^{-1}\mathrm{PEV})\big\} + log\mathrm{Pr}_{\mathrm{k}}, \qquad (5)$$

where, $\mathbf{H_k} = \mathbf{Z_i}\mathbf{Z_i}'\boldsymbol{\sigma}_{\boldsymbol{i}}^{\mathbf{2}} + \mathbf{E}\sigma_{\mathrm{e}}^2$, $\mathbf{y}^{\dagger} = \mathbf{y} - \mathrm{X}\hat{\boldsymbol{\beta}} - \hat{\mathbf{u}} - \hat{\mathbf{v}}$.

Then, $P(i,k) = E_{\mathbf{b}} logp\big(b(i,k)\big|\mathbf{y}, \widehat{Pr_k}\big) = E_{\mathbf{b}} logL(i,k) = \frac{exp(logL(i,k))}{\sum_{k=1}^{4} exp(logL(i,k))}$. (6)

Also, the log posterior function of proportion parameters $\mathbf{Pr}$ for all the SNPs can be simplified as $logp(\mathbf{Pr}, \mathbf{b}|\mathrm{y}) = logp(\mathbf{b}|\mathbf{Pr}) + logp(\mathbf{Pr})$. Since this function is the same as the one of emBayesR (Wang et al. 2015), the parameter $\mathbf{Pr}$ can be derived as:

$$Pr_k = \frac{\sum_{\mathrm{i}=1}^{m} P(i,k)+1}{\sum_{k=1}^{4}(\sum_{i=1}^{m} P(i,k)+1)} \qquad (7)$$

Afterwards, based on the posterior probability $\mathrm{P}(i,k)$, the SNP effects can be derived according to the log posterior distribution:

$$logp\big(g_i, \mathbf{u}, b(i,k)\big|\mathbf{y}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{v}}, \widehat{\sigma_{\mathrm{e}}^2}, \widehat{\mathbf{Pr}}\big) \propto logL + logp\big(g_i, b(i,k)\big|\widehat{\mathbf{Pr}}\big)$$

Following the Expectation and Maximization steps of EM, which has the same process as emBayesR, we obtain the estimation of SNP effects:

$$\hat{g}_i = [\mathbf{Z}_{\mathrm{i}}'\mathbf{E}^{-1}\mathbf{Z}_{\mathrm{i}} + \sum_{k=1}^{4}\left(P(i,k)\frac{\widehat{\sigma_{\mathrm{e}}^2}}{\sigma_{i\,[k]}^2}\right)]^{-1}[\mathbf{Z}'\mathbf{E}^{-1}\mathbf{y}^{\dagger}] \qquad (8)$$

Compared with the estimation for each SNP effects, the parameters including

fixed effects ($\mathbf{\beta}$), polygenic effects ($\mathbf{v}$) and the error variance (σ_e^2) work on the extended model from model (1), which can be written as $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}'' + \mathbf{Wv} + \mathbf{e}$, with the breeding values of all animals $\mathbf{u}'' = \mathbf{Zg}$. Therefore the distribution of the data based on this model can be transformed to $(\mathbf{y} - \mathbf{X\beta} - \mathbf{u}'' - \mathbf{Wv}) \sim N(0, \mathbf{E}\sigma_e^2)$. Accordingly, the updated log likelihood function $logL''$ can be written as:

$$logL'' = -\frac{n}{2} log\sigma_e^2 + \frac{1}{2\sigma_e^2}(\mathbf{y}^*)'\mathbf{E}^{-1}\mathbf{y}^* \text{ with } \mathbf{y}^* = \left(\mathbf{y} - \mathbf{u}'' - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\hat{\mathbf{v}}\right).$$

Using Bayes' theorem, the log posterior function of fixed effects ($\mathbf{\beta}$), polygenic effects ($\mathbf{v}$) and the error variance (σ_e^2) can be expressed separately. For fixed effects ($\mathbf{\beta}$) and the error variance (σ_e^2) with uninformative priors, $logp\left(\sigma_e^2, \widehat{\boldsymbol{\beta}}, \mathbf{u}''|\mathbf{y}\right) = logL''$; while polygenic effects is assumed as $\mathbf{v} \sim \mathbf{N}(\mathbf{0}, \mathbf{A}\sigma_a^2)$, with the log posterior function as $logp(\mathbf{v}, \mathbf{u}''|\mathbf{y}) = logL'' + logp(\mathbf{v})$.

Applying the Expectation for both of the above posterior function, we get:

$$E_{\mathbf{u}''} logp\left(\sigma_e^2, \widehat{\boldsymbol{\beta}}, \mathbf{u}''|\mathbf{y}\right) = E_{\mathbf{u}''} logL'' = -\frac{n}{2} log\sigma_e^2 + \frac{1}{2\sigma_e^2}[(\mathbf{y}^*)'\mathbf{E}^{-1}\mathbf{y}^* + \mathrm{tr}(\mathbf{E}^{-1}\mathrm{PEV})] \quad (9)$$

$$E_{\mathbf{u}''} logp(\mathbf{v}, \mathbf{u}''|\mathbf{y}) = E_{\mathbf{u}''} L'' + logp(\mathbf{v}) \quad (10)$$

Then, taking the maximization step for the above equations (9-10), the parameters are derived as follows:

$$\widehat{\sigma_e^2} = \frac{1}{n}\left[\left(\left(\mathbf{y} - \mathbf{u}'' - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\hat{\mathbf{v}}\right)\right)' \mathbf{E}^{-1}\left(\mathbf{y} - \mathbf{u}'' - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\hat{\mathbf{v}}\right) + \mathrm{tr}(\mathbf{E}^{-1}\mathrm{PEV})\right] \quad (11)$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{E}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}^{-1}(\mathbf{y} - \mathbf{u}'' - \mathbf{W}\hat{\mathbf{v}}) \quad (12)$$

$$\hat{\mathbf{v}} = (\mathbf{W}'\mathbf{E}^{-1}\mathbf{W}\sigma_a^2 + \sigma_e^2\mathbf{A}^{-1})^{-1}\sigma_a^2\mathbf{W}'^{\mathbf{E}^{-1}}\left(\mathbf{y} - \mathbf{u}'' - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \quad (13)$$

Modified Opt_emBR with two speed-up schemes

The other improvement of Opt_emBR is to implement two computationally efficient schemes to speed up the EM process. The first is by means of statistical transformation of the matrix calculation (to simplify the mathematical calculation of

the matrix to the vector) for the parameters $\mathrm{P(i,k)}$ (equation 6) and $\hat{\mathbf{v}}$ (equation 13). The second scheme aims at introducing the threshold criterion for updating SNP effects (to speed up the convergence speed of EM algorithm). In detail, when SNPs meet the threshold, the effects will not be updated anymore; otherwise, SNP effects will be estimated as usual.

Speed-up scheme I.

Scheme I is to transform the operations on large matrices (e.g. inverse, or determinant) to vector or scalar calculations that happen during the calculation of the parameter $\mathrm{P(i,k)}$ (Equation 6) and $\mathbf{v}$ (equation 13). In the equation (6), both inverse and determinant operations for matrix $\mathbf{H_k}$ are required, which is computationally demanding especially as the calculation of $\mathrm{P}(i,k)$ happens in the innermost loop of the algorithm (shown in the Pseudo code in Figure 4.1) of the following part). According to the Woodbury Identity theory, that is $(\mathbf{A}+\mathbf{XBX}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{X}(\mathbf{B}^{-1}+\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{A}^{-1}$, the equation $\mathbf{H_k}^{-1}$ can be simplified as:

$$\mathbf{H_k}^{-\mathbf{1}} = (\mathbf{Z_i}\mathbf{Z_i}'\sigma_i^2[k] + \mathbf{E}\boldsymbol{\sigma}_\mathbf{e}^\mathbf{2})^{-\mathbf{1}} = \boldsymbol{\sigma}_\mathbf{e}^{-\mathbf{2}}\left(\mathbf{E}^{-\mathbf{1}} - \frac{\mathbf{E}^{-\mathbf{1}}\mathbf{Z_i}\mathbf{Z_i}'\mathbf{E}^{-\mathbf{1}}\sigma_i^2[k]}{\sigma_i^2[k]\mathbf{Z_i}'\mathbf{E}^{-\mathbf{1}}\mathbf{Z_i}+\boldsymbol{\sigma}_\mathbf{e}^\mathbf{2}}\right), \qquad (14)$$

which have reduced the inverse of huge square matrix $\mathbf{H_k}$ to the product or dot product of vectors and the inverse of the diagonal matrix $\mathbf{E}$.

Pseudo-code for the overall methods of Opt_emBR

function Opt_emBR(σ_e^2, σ_v^2 , σ_g^2)

begin "get σ_e^2, σ_v^2 and σ_g^2 from GBLUP estimation; get vector y, matrix **Z**, lower triangle matrix **Ped**"

① Initialize $\mathbf{g}$, Pr, b, $\hat{v}$, σ_i^2 ; Construct $\mathbf{X}, \mathbf{A}, \mathbf{G}, \mathbf{E}, \mathbf{W}$ matrices

$n \leftarrow \mathrm{nrow}(\mathbf{Z})$, $m \leftarrow \mathrm{ncol}(\mathbf{Z})$, $p \leftarrow 3$, $q \leftarrow \mathrm{ncol}(\mathbf{Ped})$

② $\mathrm{PEV}_{\mathbf{u_1}}(\mathbf{e}) \leftarrow (\mathbf{E}^{-1}\sigma_e^{-2} + (\mathbf{G}\sigma_g^2 + \mathbf{WAW}'\sigma_a^2)^{-1})^{-1}$; $tr_{\mathrm{PEV}} \leftarrow \mathrm{tr}\left(\mathrm{E}^{-1}\mathrm{Z_i}\mathrm{Z_i'}\mathrm{E}^{-1}\mathrm{PEV}_{\mathbf{u_1}}(\mathbf{e})\right)$

while unconverged do

while $i \leftarrow 1$ to m do

③ $\mathrm{y}^{\dagger} \leftarrow \mathrm{y} - \sum_{j \neq i} \mathrm{Z_j}\hat{\mathrm{g}}_j - \mathrm{X}\hat{\mathrm{b}} - \mathrm{W}\hat{\mathrm{v}}$

④ If speed-up scheme **C1** was true then

$g_i \leftarrow 0$; $P(i,1) \leftarrow 1, P(i,2{:}4) \leftarrow 0$

Otherwise

while $k_1 \leftarrow 1$ to 4 do

⑤ calculate $logL(i,k_1)$ from equation (16)

end

update each $P(i,k_2)$ $(k_2 = (1,\ldots,4))$ with $\frac{exp(logL(i,k_2)}{\sum_{k_2=1}^{4} exp(logL(i,k_2))}$

⑥ If speed-up scheme **C2** was true then

$g_i \leftarrow 0$; $P(i,1) \leftarrow 1, P(i,2{:}4) \leftarrow 0$

otherwise

⑦ calculate $\hat{g}_i$ from equation (8)

end

end

end

⑧ update $\mathbf{Pr}$, σ_e^2, $\boldsymbol{\beta}$, $\mathbf{v}$ using equation (7),equation (11), equation (12),equation (13)

unconverged←False

if $(\hat{\mathrm{g}}^{\mathrm{q}} - \hat{\mathrm{g}}^{\mathrm{q}-1})'(\hat{\mathrm{g}}^{\mathrm{q}} - \hat{\mathrm{g}}^{\mathrm{q}-1})/((\hat{\mathrm{g}}^{\mathrm{q}\prime}\hat{\mathrm{g}}^{\mathrm{q}}) > 10^{-10}$ then

unconverged←True

endif

end

end function

Figure 4.1. The pseudo-code of Opt_emBR algorithm.

Again, applying Sylvester's theorem (detailed by Kemper et al. (2015)) to $log|\mathbf{H_k}|$, we obtain:

$$\log|\mathbf{H_k}| = (n-1)lo\,g\,\sigma_e^2 + lo\,g|\mathbf{E}| + \log(\sigma_i^2[k]\mathbf{Z_i'E^{-1}Z_i} + \sigma_e^2), \quad (15)$$

which have transformed high dimensional matrix calculations to dot products of vectors and the determinant calculation of the diagonal matrix. Substituting (14) and (15) into (6), while leaving out the terms irrelevant with σ_k^2 (which will be eliminated during the calculation of the equation (7)), the formula (6) can be re-written as:

$$logL(i,k) =$$

$$logPr_k - \tfrac{1}{2}\left\{log\,V - \left[\left(\mathbf{y}^{\dagger\prime}\mathbf{E^{-1}Z_i}\right)^2 - \mathbf{tr}\left(\mathbf{E^{-1}Z_iZ_i'E^{-1}PEV}(\hat{\mathbf{u}})\right)\right]\boldsymbol{\sigma}_\mathbf{i}^\mathbf{2}[\mathbf{k}]\sigma_e^{-2}/\mathrm{V}\right\}, \quad (16)$$

With the scalar $\mathrm{V} = \boldsymbol{\sigma}_\mathbf{k}^\mathbf{2}\mathbf{Z_i'E^{-1}Z_i} + \sigma_e^2$.

Also, for the estimation of polygenic effects $\mathbf{v}$, we apply the transformation of equation (13) to $(\mathbf{W'E^{-1}W}\sigma_a^2 + \sigma_e^2\mathbf{A^{-1}})\hat{\mathbf{v}} = \sigma_a^2\mathbf{W'E^{-1}}(\mathbf{y} - \hat{\mathbf{u}}_2 - \mathbf{X}\widehat{\boldsymbol{\beta}})$. Let $\mathbf{M} = (\mathbf{W'E^{-1}W}\sigma_a^2 + \sigma_e^2\mathbf{A^{-1}})$, and then $\mathbf{M} = \mathbf{M_{diag}} + \mathbf{M_{offdiag}}$ ($\mathbf{M_{offdiag}}$ means $\mathbf{M}$ matrix with zero diagonal elements; $\mathbf{M_{diag}}$ means to keep diagonal elements of $\mathbf{M}$ matrix but leave others as zero).

Then, $\mathbf{M_{diag}}\hat{\mathbf{v}} + \mathbf{M_{offdiag}}\hat{\mathbf{v}} = \sigma_a^2\mathbf{W'E^{-1}}(\mathbf{y} - \hat{\mathbf{u}}_2 - \mathbf{X}\widehat{\boldsymbol{\beta}})$.

The operation $\boldsymbol{\tau} = \mathbf{M_{offdiag}}\hat{\mathbf{v}} = \begin{bmatrix} \mathbf{M}_{1.}\hat{\mathrm{v}}_{-1}^* \\ \cdots \\ \mathbf{M}_{i.}\hat{\mathrm{v}}_{-i}^* \\ \cdots \\ \mathbf{M}_{q.}\hat{\mathrm{v}}_{-q}^* \end{bmatrix}$, where $\mathbf{M}_{i.}$ is the i^{th} row of $q \times q$ matrix $\mathbf{M}$ and $\hat{\mathbf{v}}_{-\boldsymbol{i}}^*$ is the current vector $\hat{\mathbf{v}}$ with $\hat{v}_i = 0$. In other words, when estimating current polygenic effect $\hat{v}_i$, the i^{th} element τ_i of the vector $\boldsymbol{\tau}$ means the sum of all the products of the other polygenic effects $\hat{v}_{j,j\neq i}$ and the elements of $\mathbf{M}_{ij,j\neq i}$. Therefore, instead of the inverse of 2-dimensional matrix $\mathbf{M}$, the polygenic effect $\hat{\mathbf{v}}$ can be calculated via

$$\hat{\mathbf{v}} = \begin{bmatrix} \hat{v}_1 \\ \cdots \\ \hat{v}_i \\ \cdots \\ \hat{v}_q \end{bmatrix} = \mathbf{M}_{diag}{}^{-\mathbf{1}} \left\{ \sigma_a^2 \mathbf{W}' \mathbf{E}^{-1} (\mathbf{y} - \hat{\mathbf{u}}_\mathbf{2} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \begin{bmatrix} \mathbf{M}_{1.} \hat{\mathrm{v}}^*_{-1} \\ \cdots \\ \mathbf{M}_{i.} \hat{\mathrm{v}}^*_{-i} \\ \cdots \\ \mathbf{M}_{q.} \hat{\mathrm{v}}^*_{-q} \end{bmatrix} \right\}, \quad (17)$$

which shows polygenic effects can be estimated by means of the other estimated polygenic effects and the data y. Since the inverse of matrix $\mathbf{M}$ can be transformed into the inverse of diagonal matrix and *q* times of dot product of vectors, the computational speed is much improved.

Speed-up scheme II.

Scheme II includes the modifications from Scheme I, but additionally applies updating threshold criterions for SNP effects to remain in the model at each iteration, otherwise they are discarded. This improves the convergence speed of the EM algorithm. The aforementioned prior assumption for SNP effects is that most SNPs have zero effects, while others follow three different normal distributions with the variances varying from $0.0001 * \sigma_g^2, 0.001 * \sigma_g^2$ to $0.01 * \sigma_g^2$. This Gaussian mixture prior assumption defines the concave log likelihood function (Dias & Wedel 2004). Then, EM algorithm will approximate the estimate of SNP effects gradually approaching to the final optimums by maximizing their log likelihood curves, which decides the trend of SNP effects is monotonic increasing or decreasing curve during the EM iterations shown in Figure 4.2.

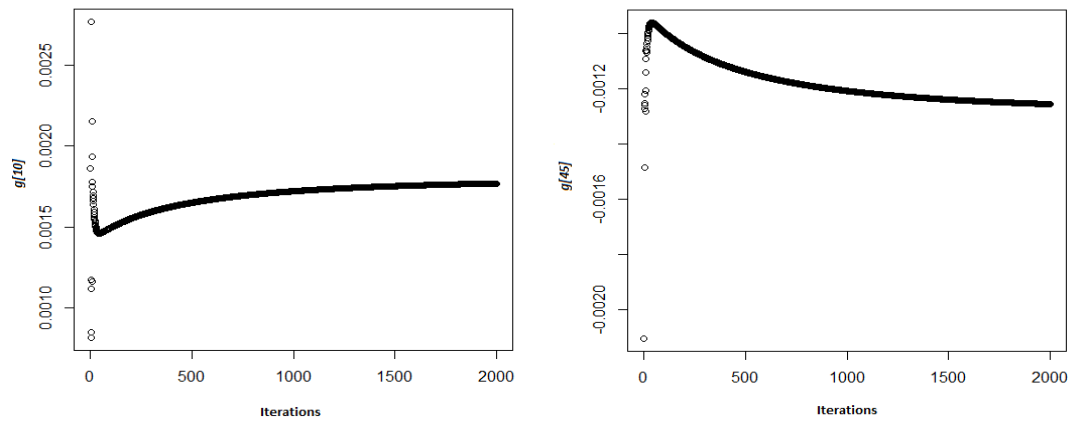


Figure 4.2. Two types of trends of SNP effects during EM iterations.

Therefore, there is a scenario: when the SNP effect $\hat{g}_i$ is zero or "very adjacent" to zero after certain numbers of iterations (e.g. 50 times), it will be considered to have zero effect until the convergence of the algorithm. In other words, once SNP effect $\hat{g}_i$ is ensured to have minor effect, there is no need to update SNP i any more during subsequent iterations. Since there are a large proportion of SNPs (Pr_1) assumed to have zero effects, such scenario can reduce the updating calculations of a large number of SNP effects, so as to speed up the algorithm. However, the difficulty is how to set up the thresholds that distinguishes SNP effects having no effects (in Pr_1) or having small effects (in Pr_2), which is very important as a substantial component of genetic variance is captured by large numbers of SNPs with very small effects for many traits (Moser *et al.* 2015). We assess one criterion for determining if SNPs had zero effects:

$$\mathbf{C1}: |g_i| \leq a;$$

$$\mathbf{C2}: p(i,0) \geq b.$$

The first condition $\mathrm{C1}$ defines the level of the adjacent extent of a SNP effect to zero according to its current value. Three values in genetic standard deviations of a are tested: $a_1 = 0.000001;\ a_2 = 0.0000001;\ a_3 = 0.00000001$. For example, when implemented with the criterion$|g_i| \leq a_2$, the effect of SNP i is assumed zero if the absolute value of its effect is less than 0.0000001, and therefore will not

be updated for the following EM iterations, and will be excluded for all future iterations. The second condition $C2$ judged whether or not SNP i have the effects based on its probability of being in the zero distribution: $p(i,0)$. A range of values are considered for this threshold: $b_1 = 0.85;\ b_2 = 0.90;\ b_3 = 0.95$. For example, the condition $p(i,0) \geq 0.90$ means SNP i have more than 90% possibility of having no effects. Once such criterion is reached, SNP i is not updated in the further iterations. We investigates the effect of thresholds of a and b for criteria $C1$ and $C2$ on the prediction accuracy in our real data set to determine the best criteria.

With the investigation on the data, the criteria C1 or C2 are applied after the EM steps were running for 50 iterations.

Pseudo-code for the algorithm

The overall procedure of Opt_emBR is described by means of the pseudo code, steps 1~9. Here we will detail these steps according to their sequence appearing in the pseudo code descriptions:

Step 1: Initialize the parameters **g**, $\mathbf{Pr}$, $\boldsymbol{\sigma}_{\mathbf{i}}^{\mathbf{2}}$ and Construct **X, A, G, E, W** matrices. Similar to emBayesR (Wang *et al.* 2015), the starting values of **g** and $\mathbf{Pr}$ were set as $\mathbf{g} = 0.01$ and $\mathrm{Pr} = \{0.5, 0.487, 0.01, 0.003\}$, while $\boldsymbol{\sigma}_{\mathbf{i}}^{\mathbf{2}} = \{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2\}$ with the genetic variance σ_g^2 obtained from GBLUP.

The $n \times 3$ matrix **X** is design matrix, allocating the phenotypes to fixed effects. In our case, matrix **X** is set up with first column being the mean, the second and third columns defining the breeds (Holstein and Jersey)and sex (bulls and cows) of the cattle. For example, if the i^{th} animal is Holstein bulls, then $x_{i,2} = 1$ with $x_{i,3} = 0$. Pedigree relationship matrix **A** is built up using the lower symmetrical matrix **Ped** detailed by Henderson (Henderson 1984); while the genomic relationship matrix $\mathbf{G}$ is constructed using the equation $\mathbf{G} = \mathbf{ZZ}'/n$. Diagonal error matrix **E** is defined with the equation (2), and the index matrix **W** refers the

individuals in the reference set to the whole pedigree. For instance, the i^{th} animal in the reference set is located in the 10^{th} of the whole pedigree, and then $w_{i,10} = 1$.

Step 2: Calculate PEV matrix under model 1, as aforementioned. Then using PEV matrix, $tr\left(\mathbf{E^{-1}Z_i Z_i' E^{-1} PEV_{u_1}(e)}\right)$ – that is the term for the equation $logL(i,k)$ is calculated in front of EM iterations, reducing computational time.

Then for each SNP i (i in 1 to m)

Step 3: Correct $\mathbf{y}$ for the effects of all other SNPs except current SNP i with equation $\mathbf{y}^{\dagger} = \mathbf{y} - \sum_{j \neq i} \mathbf{Z_j}\,\hat{g}_j - \mathbf{X\hat{b}} - \mathbf{W\hat{v}}$.

Step 4: After 50 iterations, the speed-up scheme $\mathbf{C1}$ is implemented deciding whether or not SNP i will have a non-zero effect in future iterations. If not, $\hat{g}_i = 0$; also $P(i,0)$ was set as 1.

Step 5: Estimate the probability that the effect of SNP i in from one of four normal distributions $logL(i,k)$ with the equation (23), which has been implemented by the speed-up scheme I. After this, $P(i,k)$ is calculated with the equation $exp(logL(i,k_2)/\sum_{k_2=1}^{4} exp(logL(i,k_2))$.

Step 6: After 50 iterations, the speed-up scheme $\mathbf{C2}$ is implemented according to the probability $P(i,0)$.

Step 7: When the criteria for schemes $\mathbf{C1}$ or $\mathbf{C2}$ are not met, the SNP effect $\hat{g}_i$ is updated via equation (11).

After effects have been estimated for all SNP,

Step 8: Calculate σ_e^2 with equation (17), fixed effects $\boldsymbol{\beta}$ with equation (18) update Pr_k with equation (19) , and update polygenic effects with the simplified equation (22).

Step 9: Assess convergence criterion $(\hat{\mathbf{g}}^l - \hat{\mathbf{g}}^{l-1})'(\hat{\mathbf{g}}^l - \hat{\mathbf{g}}^{q-1})/((\hat{\mathbf{g}}^{l\prime}\hat{\mathbf{g}}^l) \leq 10^{-10}$ with l being the loop number of the EM iterations is applied here to test whether or not the algorithm is converged. If not converged, then return to Step 3 for the next EM iteration; Otherwise, exit the EM iterations and return the final

estimated results.

Steps from 1 to 9 describe the implementation of Opt_emBR. From these steps, the computational complexity can be calculated. Looking through the structure of the pseudo code, Step 5 of calculating $\log\mathrm{L}(\mathrm{i},\mathrm{k})$ is located in the innermost of the algorithm. According to the equation (16) of $\log\mathrm{L}(\mathrm{i},\mathrm{k})$, there are dot products of vectors with n elements, which means n loops of multiplying operations between scalars. Therefore, counting for the outer loop of each EM iteration, there are $4 \times m \times n$ times of loops for the basic operations, and hence the computational complexity can be written as $O(4mn)$. Compared with $O(4mn^3)$ for equation (6) of $\log\mathrm{L}(\mathrm{i},\mathrm{k})$, before the application speed up scheme I, the computational cost is much reduced. The other intensive calculation lies in the step 8 of polygenic effects estimation, which is iterated with q individuals in the pedigree instead of each SNP. According to the equation (17) implemented with speed-up scheme I, the time complexity is therefore $O(nq)$. The operations time is much improved in contrast with $O(nq^3)$ that is required by the equation (13), which is not yet applied with speed-up scheme I. As the operations of the equation (17) do not happen in the innermost part of the algorithm, the total time complexity is $O(4mn)$, after improved with speed-up scheme I. Moreover, because there is a large proportion of SNPs without effects, scheme II means SNPs without effects don't need to go through the inner most loop with 4 iterations shown in the step 5 of the pseudocode, and therefore approximate the time complexity to $O(mn)$.

Data Sets

Genotypes. Opt_emBR was implemented in a data set of 16,328 dairy cattle with genotypes for 632,003 SNPs from two difference genotyping arrays. One with 1,745 Holstein and Jersey cattle, and 114 Australian bulls (for validation, not included into the reference sets) were genotyped with 777K Illumina HD bovine SNP chip; while the other one including 12,049 Holstein and Jersey bulls and

cows was genotyped with 54K Illumina Bovine SNP array. After stringent quality control and SNP filtering described in (Erbe *et al.* 2012), there were 632,003 SNPs remaining from 777K SNP panels, and 43,425 SNPs from the other one. Furthermore, for animals genotyped with the 43,425 SNPs, genotypes were imputed to 632,003 SNP genotypes using beagle 3.0 (Browning & Browning 2009). The final data set was 16,328 cattle with real or imputed genotypes for 632,003 SNP.

Phenotype. Four traits under various genetic architectures (i.e. milk yield, protein yield, fat percent (fat%), and fertility) were used for evaluation. In detail, the architecture of fat percent (fat%) was known to be affected by a causal variant of large effect (e.g. *DGAT1*); while fertility was characteristic of polygenic architecture, with a large number of mutations of small effects (Lund *et al.* 2014). The phenotypes for these traits were daughter trait deviations (DTD) for bulls (the average of their daughters phenotypes, corrected for fixed effects), and trait deviations (TD) for cows. For genomic prediction, the data was separated into a reference set, where SNP effects were estimated, and validation sets, where the accuracy of genomic predictions was assessed, by year of birth. All the daughters of the bulls, which belong to the validation sets are excluded from the reference set. To evaluation the performance of Opt_emBR for within-populations, multi-populations, and across-populations prediction, the reference data included bulls and cows from two breeds of Holstein and Jersey, which could be implemented to evaluate the performance of multi-populations predictions on Holstein bulls and Jersey bulls and across-populations predictions on Australian Red bulls. The exact number of individuals in these data sets for each trait was detailed in Table 4.1.

Table 4.1. The number of individuals in the reference sets and validations sets

related to three traits including Milk yield (MilkY), Protein yield (ProtY), Fat Percent(Fat%) and Fertility.

Traits	Reference Sets				Validation Sets		
	Holstein		Jersey		Holstein Bulls	Jersey Bulls	Australian Red Bulls
	Bulls	Cows	Bulls	Cows			
MilkY/ProtY/ Fat%	3,049	8,478	770	3,917	262	105	114
Fertility	2,806	7,838	716	3,830	396	81	114

In the pseudo code of Opt_emBR, three variances parameters (σ_e^2, σ_v^2, σ_g^2) related to the reference data sets and traits were required as the input. We ran asreml4.0 (Gilmour *et al.* 2002) (which was implemented with GBLUP methods) on these data sets to estimate these variance parameters, listed in Table 4.2. Therefore, the heritability of these traits varied from 0.01 (for fertility) up to 0.65 (for fat%).

Table 4.2. Three input variance parameters related to the reference data sets.

Reference Set	Traits	$\boldsymbol{\sigma_e^2}$	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_v^2}$
Holstein and Jersey bulls & cows	Milk yield	133284.0	108619.0	34925.6
	Protein yield	132.579	68.6635	29.1662
	Fat%	0.0180012	0.0575729	0.0127094
	Fertility	3283.80	31.6187	0.000332297

The accuracy of prediction ability was calculated by means of the correlation between GEBV and DTD in the validation sets. And the bias was the coefficient of regressing DTD on GEBV – unbiased prediction would result in a regression coefficient of 1.

4.5 Results and Discussion

We first assessed the impact of the speed up schemes (Scheme I and Scheme II) on the computational time and prediction accuracy in a range of data sets of increasing size. Then the accuracy of genomic prediction from Opt_emBR was

compared to other methods within population and across populations for a number of complex traits in our dairy cattle data set. We also assessed the precision of QTL mapping with Opt_emBR and BayesR in the same data set.

The impact of speed-up schemes on the prediction accuracy and computational time

Prediction accuracy of Opt_emBR with speed-up schemes. As speed-up scheme I aimed at simplifying the matrix calculations by means of matrix transformation, the computational time could be reduced without impact on the prediction accuracy. However, the two criteria of speed-up scheme II could potentially influence the prediction accuracy. Therefore, the prediction accuracy and estimates of two other parameters (the proportion **Pr**, and error variance $\boldsymbol{\sigma_e^2}$) were assessed with different values for two criterions $\mathbf{C1}$ and $\mathbf{C2}$ in Table 4.3, using the dairy cattle data set and the milk production trait. Criteria for $\mathbf{C1}$ were $|g_i| \leq 0.000001$, $|g_i| \leq 0.0000001$, or $|g_i| \leq 0.00000001$. The results of **Pr** in the table showed, as total estimated effects of SNP s were relatively small, the first threshold was too large so as to remove too many SNPs with small effects (in the proportion *Pr[2]*). In detail, compared with Opt_emBR without **C1** (termed Original in Table 4.3), around 0.08% extra SNPs of the total (more than 500 SNPs) were shrunk to zero with $\mathbf{C1}$: $|g_i| \leq 0.000001$, leading to the accuracy reduction. The second threshold was better to distinguish the SNPs with small effects from SNPs without effects, but shrunk the SNPs with very large effects too much, leading to 9% reduction of the accuracy. As shown in Table 4.3, the algorithm with the criterion **C1**: $|g_i| \leq 0.0000001$ achieved 0.0093% SNPs (around 59 SNPs) with very large effects more than the largest proportion from Opt_emBR without C1 (Original in Table 4.3), which might reduce the prediction accuracy as well. Comparably, the third threshold $|g_i| \leq 0.00000001$ performed well for the detection of the SNPs in each proportion of Pr, and the calculation of the error variance, as well as for the

prediction accuracy. For the criterion $\mathbf{C2}$, three possible choices were tracked, but all of them had the problems to differentiate SNPs with small effects and SNPs with zero effects, resulting up to 7% loss of the prediction accuracy Table 4.3. Therefore, $|g_i| \leq 0.00000001$ of criterions $\mathbf{C1}$ was the criteria used in all implementations of Opt_emBR that followed.

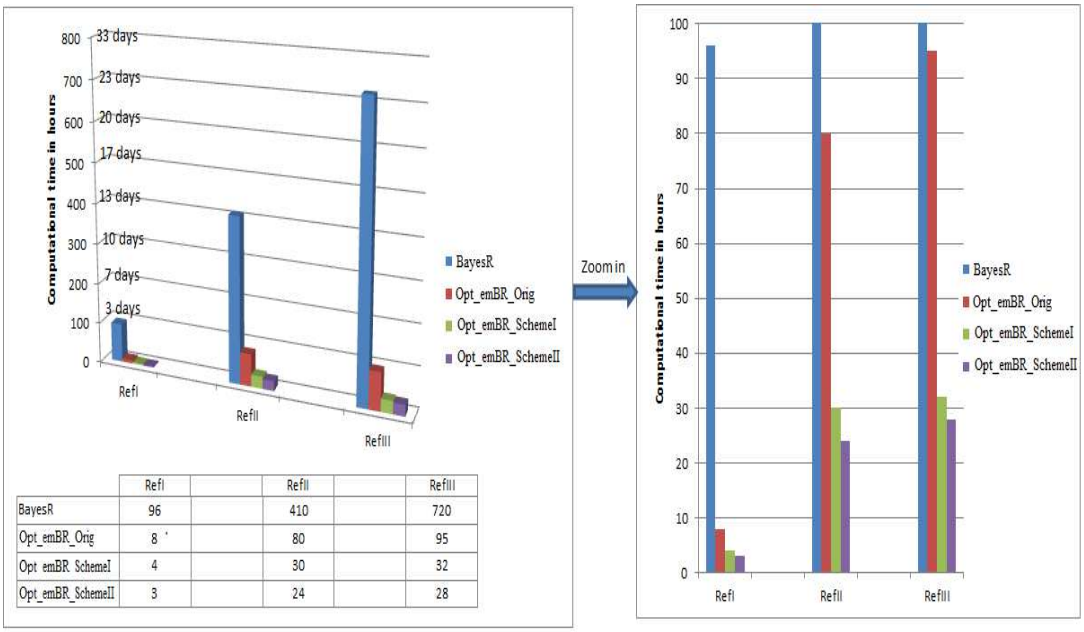
Table 4.3. The estimated results of Acc. (Accuracy), Pr(the proportion), and σ_e^2 (error variance) according to different criteria of speed-up scheme II.

Speed-up Scheme II	**Opt_emBR**		
	Acc.	**Pr**	$\boldsymbol{\sigma_e^2}$
Original[a]	0.66	[0.998371, 0.001583, 0.000007, 0.000039]	239409
Apply C1[b]: $\|g_i\| \leq 0.000001$	0.57	[0.999113, 0.000396, 0.000428, 0.000063]	306361
Apply C1[b]: $\|g_i\| \leq 0.0000001$	0.66	[0.998792, 0.001011, 0.000065, 0.000132]	221533
Apply C1[b]: $\|g_i\| \leq 0.00000001$	0.68	[0.997545, 0.002394, 0.000009, 0.000052]	247965
Apply C2[c]: $P(i,0) \geq 0.85$	0.59	[0.999938, 0.000002, 0.000002, 0.000058]	327295
Apply C2[c]: $P(i,0) \geq 0.9$	0.63	[0.999910, 0.000002, 0.000003, 0.000085]	293821
Apply C2[c]: $P(i,0) \geq 0.95$	0.61	[0.999941, 0.000002, 0.000003, 0.000054]	329221

[a] means Opt_emBR without scheme II;
[b] and [c] are two criteria C1 and C2 set up the thresholds for SNP effect g_i to define whether or not g_i was zero.

Computational time of Opt_emBR with speed-up schemes on different datasets. Since BayesR and Opt_emBR shared the same data model and prior assumption for all the parameters, they had exactly the same computation complexity $O(mn)$ where m and n were the number of markers and individuals separately. Then the difference lied in the number of iterations for Opt_emBR versus number of MCMC cycles for BayesR. BayesR required approximately 50,000 iterations to reach good estimates for SNP effects with this size of data set

(Wang *et al.* 2015). Comparatively, Opt_emBR implemented EM method to heuristically converge to the results. To access the computational performance of Opt_emBR, the computational time required for BayesR and Opt_emBR was compared in Figure 4.3. Also we compared the time demands between BayesR, Opt_emBR (Opt_emBR without speed-up schemes), Opt_emBR_Schem I (Opt_emBR implemented with first speed-up scheme) and Opt_emBR_Schem II (Opt_emBR with two speed-up schemes) to investigate the efficiency of two speed-up schemes. Three reference data sets related to milk yield were applied here, which had 632,003 SNPs with different sizes of animals ranging from 3,049 in RefI, 11,527 in RefII, to 16,214 in RefIII. The results from Figure 4.3 demonstrated an obvious advantage of Opt_emBR over BayesR: 72 hours for BayesR compared with 3 hours for Opt_emBR_Schem II on RefI, 408 hours for BayesR but 24 hours for Opt_emBR_Schem II on RefII, and 720 hours for BayesR but 28 hours for Opt_emBR_Schem II on RefIII. In addition, Figure 4.3 demonstrated that the speed-up scheme I could help to reduce up to 2/3 of the total computational time that was required for the Opt_emBR without the implementation of two optimized schemes while adding speed-up scheme II could reduce further to approximately 1/8 of the time.



	RefI		RefII		RefIII
BayesR	96		410		720
Opt_emBR_Orig	8		80		95
Opt_emBR_SchemeI	4		30		32
Opt_emBR_SchemeII	3		24		28

Figure 4.3. The computational time in hours compared between BayesR, Opt_emBR_Orig, Opt_emBR_SchemeI, Opt_emBR_SchemeII on three reference data sets (RefI, RefII, and RefIII). Opt_emBR_Orig means Opt_emBR without speed-up schemes; Opt_emBR_SchemeI means Opt_emBR with speed_up Scheme I; Opt_emBR_SchemeII is Opt_emBR with speed-up SchemeI and II. Sub-fig A is the overall comparison between BayesR with Opt_emBR families. Sub-fig B is zoomed in from Sub-fig A, with the label on y-axis scaling from 0 to 100.

Prediction accuracy and bias evaluation for four complex traits

The prediction accuracy and bias between BayesR, GBLUP and Opt_emBR (using the optimal speed up scheme described above) was compared for milk yield, protein yield, fat% and fertility, Table 4.4 and Table 4.5. The impact of the method of prediction was compared in three ways,

1. The accuracy of prediction for validation individuals from a population that was included in the reference set, where the reference set included only individuals from that population (e.g. a Holstein reference set to predict a Holstein validation set), with accuracy calculated as the

correlation of genomic estimated breeding value and DTD in the validation bulls.

2. The accuracy of prediction for validation individuals from populations that was included in the reference set, where the reference set included two populations (e.g. Holstein and Jersey reference set used to predict a Holstein validation.
3. The accuracy of prediction for validation individuals from a third population, not included in the reference set (Holstein and Jersey reference set used to predict Australian Red bulls and cows).

Table 4.4. The within-population and multi-populations prediction ability of BayesR, GBLUP, and Opt_emBR on Holstein bulls.

	Holstein reference to predict Holstein validation							
	Milk Yield		Protein Yield		Fat%		Fertility	
	Acc.[a]	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
BayesR	0.63 (-0.02)[b]	0.88 (+0.09)	0.64 (-0.01)	1.01 (+0.04)	0.81 (-0.02)	1.03 (+0.07)	0.43 (-0.02)	1.18 (+0.00)
GBLUP	0.57 (-0.01)	0.86 (+0.10)	0.63 (-0.04)	0.87 (+0.11)	0.73 (-0.02)	0.96 (+0.09)	0.43 (-0.01)	1.19 (+0.01)
Opt_emBR	0.62 (-0.00)	0.79 (+0.13)	0.65 (-0.01)	0.85 (+0.09)	0.77 (-0.03)	0.98 (+0.08)	0.42 (-0.01)	1.15 (+0.00)
	Holstein and Jersey reference to predict Holstein validation							
	Milk Yield		Protein Yield		Fat%		Fertility	
	Acc.[a]	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
BayesR	0.68 (-0.01)[b]	0.84 (+0.07)	0.68 (-0.01)	0.88 (+0.15)	0.81 (-0.02)	0.90 (+0.08)	0.44 (-0.02)	1.53 (+0.00)
GBLUP	0.63 (-0.01)	0.83 (+0.07)	0.65 (-0.08)	0.85 (+0.03)	0.74 (-0.02)	0.85 (+0.05)	0.44 (-0.02)	1.66 (+0.00)
Opt_emBR	0.68 (-0.03)	0.90 (+0.01)	0.68 (-0.02)	0.79 (+0.06)	0.77 (-0.02)	0.83 (+0.04)	0.44 (-0.00)	1.27 (+0.00)

[a] means the accuracy

[b] means the difference when not using the polygenic effects.

Table 4.5. The within-population and multi-populations prediction ability of BayesR, GBLUP, and Opt_emBR on Jersey bulls data.

	Jersey reference to predict Jersey validation							
	Milk Yield		Protein Yield		Fat%		Fertility	
	Acc. [a]	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
BayesR	0.64 (-0.01) [b]	0.94 (+0.04)	0.68 (-0.00)	0.93 (+0.08)	0.71 (-0.02)	0.87 (+0.06)	0.15 (-0.01)	1.02 (+0.02)
GBLUP	0.59 (-0.01)	0.93 (+0.12)	0.65 (-0.01)	0.91 (+0.18)	0.54 (-0.00)	0.71 (+0.06)	0.15 (-0.01)	1.05 (+0.03)
Opt_em BR	0.64 (-0.00)	0.87 (+0.11)	0.68 (-0.02)	0.92 (+0.09)	0.69 (-0.02)	0.75 (+0.04)	0.15 (-0.00)	1.09 (+0.00)
	Holstein and Jersey reference to predict Jersey validation							
	Milk Yield		Protein Yield		Fat%		Fertility	
	Acc.	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
BayesR	0.69 (-0.01)	0.85 (+0.10)	0.71 (-0.00)	0.99 (+0.10)	0.76 (-0.02)	0.88 (+0.06)	0.26 (-0.01)	1.23 (+0.01)
GBLUP	0.64 (-0.00)	0.78 (+0.12)	0.68 (+0.01)	0.85 (+0.17)	0.66 (-0.02)	0.73 (+0.07)	0.24 (-0.00)	1.12 (+0.00)
Opt_em BR	0.66 (-0.03)	0.84 (+0.02)	0.69 (-0.01)	0.71 (+0.02)	0.75 (-0.05)	0.76 (+0.06)	0.23 (-0.00)	1.13 (+0.00)

[a] means the accuracy

[b] means the difference when not using the polygenic effects.

For the Holstein validation set, and with a Holstein reference, BayesR was 1%~8% better than GBLUP on milk production traits (i.e. Milk yield, Protein yield, and Fat%), but had no advantage for fertility, Table 4.4. Compared with BayesR and GBLUP, Opt_emBR had similar accuracy to BayesR. Similar to the results in Table 4.4, BayesR had the consistent advantage in accuracy over GBLUP on milk production traits when predicting the small breed of Jersey shown in Table 4.5. Especially on the trait of Fat Percent, BayesR had up to 16% advantage over GBLUP for within population prediction (using Jersey only reference data) while increasing the accuracy 10% than GBLUP on multi populations prediction (using both Holstein and Jersey data). Compared with BayesR and GBLUP, Opt_emBR still had an obvious superiority over GBLUP, but in some cases (e.g. Fat percent), would have 3% accuracy reduction than BayesR. On the other hands, the bias

from Table 4.4 and Table 4.5 showed GBLUP underestimated the SNP effects on most of traits. On fertility, all the methods gave a higher regression coefficient than 1.

The multi-population (e.g. Holstein and Jersey in the reference set, Holstein and Jersey validation sets) prediction results from Table 4.4 and Table 4.5 showed the superiority of BayesR and Opt_emBR over GBLUP on all the traits. Moreover, the combination of two breeds together as the Ref sets did enlarge the population size from 3,049 in Holstein and 770 in Jersey up to more than 16,000 animals, which would influence the prediction accuracy as well. The accuracy difference between within population prediction and multi populations' prediction from Table 4.4 and Table 4.5 demonstrated this in Figure 4.4, regarding to the validation sets of Holstein bulls and Jersey bulls separately. As show in Figure 4.4, there were consistent accuracy improvements for the prediction of Holstein and Jersey bulls on most traits. In details, for all the prediction methods, there were up to 10% increase in accuracy for Jersey bulls and 6% improvement for Holstein bulls when combing two reference populations of Holstein and Jersey, which confirmed the results from (Kemper *et al.* 2015) and (Hozé *et al.* 2014). Especially for the Jersey bulls, since multi-breeds enlarge the reference size 20 times more than original Jersey only reference size, the accuracy increase from multi-populations prediction was much more obvious than the increase on Holstein data.

Also, the polygenic effects could positively improve the prediction accuracy on all the reference data related to all the traits shown in the parenthesis parts of Table 4.4 and Table 4.5. In detail, for the prediction with single or two populations, adding polygenic effects could improve the accuracy around 1~2%. However, on the fertility traits, the introduction of polygenic effects for all the prediction methods did not impact the accuracy at all.

Accuracy of across population prediction with Australian red bulls and cows. The prediction accuracy on Australian red bulls was not as high as for the accuracy of genomic predictions in Holstein and Jersey, which was not surprising given there were no Australian reds in the data set. Furthermore, the comparison between BayesR, GBLUP and Opt_emBR presented consistently higher accuracy of BayesR and Opt_emBR than GBLUP for milk productions traits. Again in the Fat Percent traits, BayesR had up to 10% advantage over GBLUP. When compared with BayesR, Opt_emBR had very similar accuracy as BayesR on most traits. There was the exception: for the prediction on both Australian red bulls and cows related to Fat percent, Opt_emBR had 3~4% accuracy reduction in contrast with BayesR. For the fertility, all three methods had similar prediction ability. The bias showed the underestimation of three methods for SNP effects on most of the traits except Fertility.

Altogether, the prediction results presented in Table 4.4, Table 4.5, and Table 4.6 confirmed the robust prediction ability of our algorithm Opt_emBR for within population, multi-populations and across-populations prediction. Especially for across population prediction, BayesR and Opt_emBR had an advantage over GBLUP (as was observed for BayesR by Kemper et al. 2015). However, there existed one main disadvantage for Opt_emBR: for the trait Fat Percent, Opt_emBR had systematically reduced (3%~5%) accuracy across all the reference sets in comparison with BayesR, which might be contributed by the factor that Opt_emBR did not have enough prediction power for the traits (e.g. Fat% and Protein%) with major genes.

Table 4.6. Across-populations prediction ability of BayesR, GBLUP, and Opt_emBR on Australian Red using the reference set Holstein and Jersey bulls&cows data.

	Across-populations prediction on Australian red bulls							
	Milk Yield		Protein Yield		Fat%		Fertility	
	Acc.[a]	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
BayesR	0.22	0.60	0.12	0.49	0.45	0.92	0.27	1.03
GBLUP	0.16	0.54	0.11	0.51	0.32	0.90	0.29	0.97
Opt_emBR	0.24	0.70	0.12	0.42	0.41	0.89	0.29	1.10
	Across-populations prediction on Australian red cows							
	Milk Yield		Protein Yield		Fat%		Fertility	
BayesR	0.26	0.80	0.17	0.51	0.54	0.94	0.08	0.68
GBLUP	0.15	0.71	0.08	0.13	0.50	1.19	0.08	0.79
Opt_emBR	0.24	0.79	0.17	0.53	0.51	0.89	0.08	0.74

[a] means the accuracy

Estimates of genetic architecture from Opt_emBR - the proportion of SNPs in each distribution and QTL mapping

In BayesR and Opt_emBR, the posterior probability that every SNP was in the zero, very small, small or moderate variance distribution was derived. This was ideal for QTL mapping – SNP with very high posterior probabilities of being in the largest distribution should be strongly associated with causal mutations of moderate to large effects (e.g. Moser et al. 2015, Kemper et al. 2015). We compared the performance of BayesR and Opt_emBR for QTL mapping, by investigating the SNP numbers in each proportion of normal distributions, and then the genome location of the effects with a high posterior probability of being in the largest distribution.

For BayesR and Opt_emBR, SNPs were classified into four different groups according to their effects: A) moderate variance distribution: very small part of SNPs had very large effects with the variance $0.01 * \sigma_g^2$; B) small variance distribution: small part of SNPs had small effects with the variance $0.001 * \sigma_g^2$; C)

very small variance distribution: relatively large part of SNPs having very small effects with the variance $0.0001 * \sigma_g^2$; And D) zero variance distribution: a large proportion of SNPs had no effects. Table 4.7 compared the estimation of such classification between Opt_emBR and BayesR. Compared with BayesR, Opt_emBR resulted in more SNP effects (more than 3,000 SNPs) from the proportion with very small effects (very small variance distribution) to zero for all the traits (zero variance distribution). Also, Opt_emBR generally detected more SNPs with very large effects (in high LD with QTLs or maybe QTLs) than BayesR (except for fertility).

Table 4.7. The number of SNPs in the proportion of each distribution.

Traits	**The proportion (Pr)**	**Opt_emBR**	**BayesR**
Milk Yield	A. $0.01 * \sigma_g^2$	12	8
	B. $0.001 * \sigma_g^2$	17	47
	C. $0.0001 * \sigma_g^2$	1,523	3,886
	D. 0	630,451	628,062
Protein Yield	A. $0.01 * \sigma_g^2$	10	5
	B. $0.001 * \sigma_g^2$	37	32
	C. $0.0001 * \sigma_g^2$	1,842	4,431
	D. 0	630,114	627,535
Fat%	A. $0.01 * \sigma_g^2$	19	23
	B. $0.001 * \sigma_g^2$	206	46
	C. $0.0001 * \sigma_g^2$	1,206	2882
	D. 0	630,572	629,052
Fertility	A. $0.01 * \sigma_g^2$	8	10
	B. $0.001 * \sigma_g^2$	114	147
	C. $0.0001 * \sigma_g^2$	8,572	3,949
	D. 0	623,309	627,897

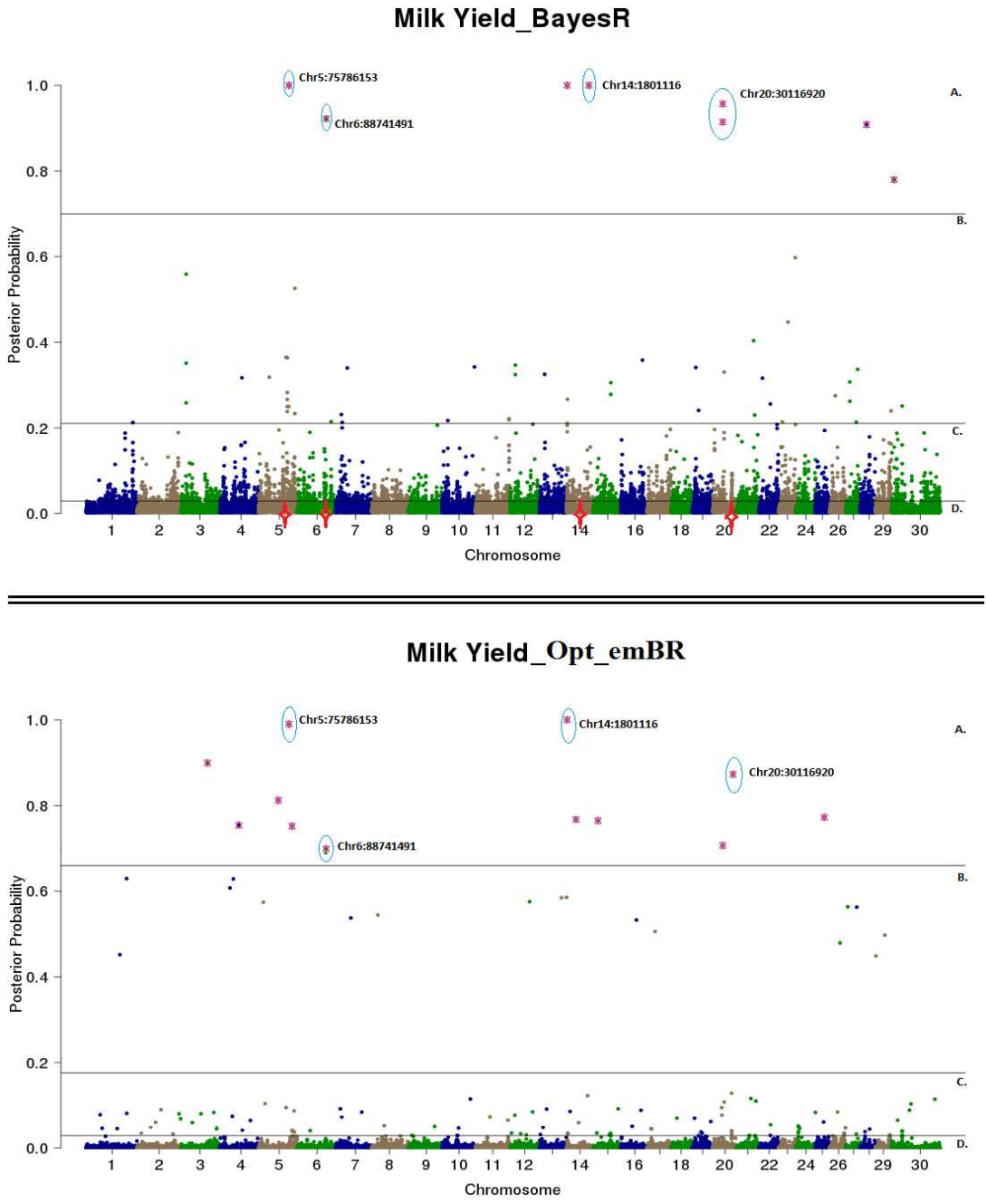


Figure 4.4. The mapping of all the SNPs estimated from BayesR and Opt_emBR on the whole chromosome related to milk yield by the posterior probability. The blue circle is the SNPs with location information mapped to known genes. In this Figure, there are four different levels (A, B, C, and D) of SNPs that are classed into different proportions according to their posterior possibility.

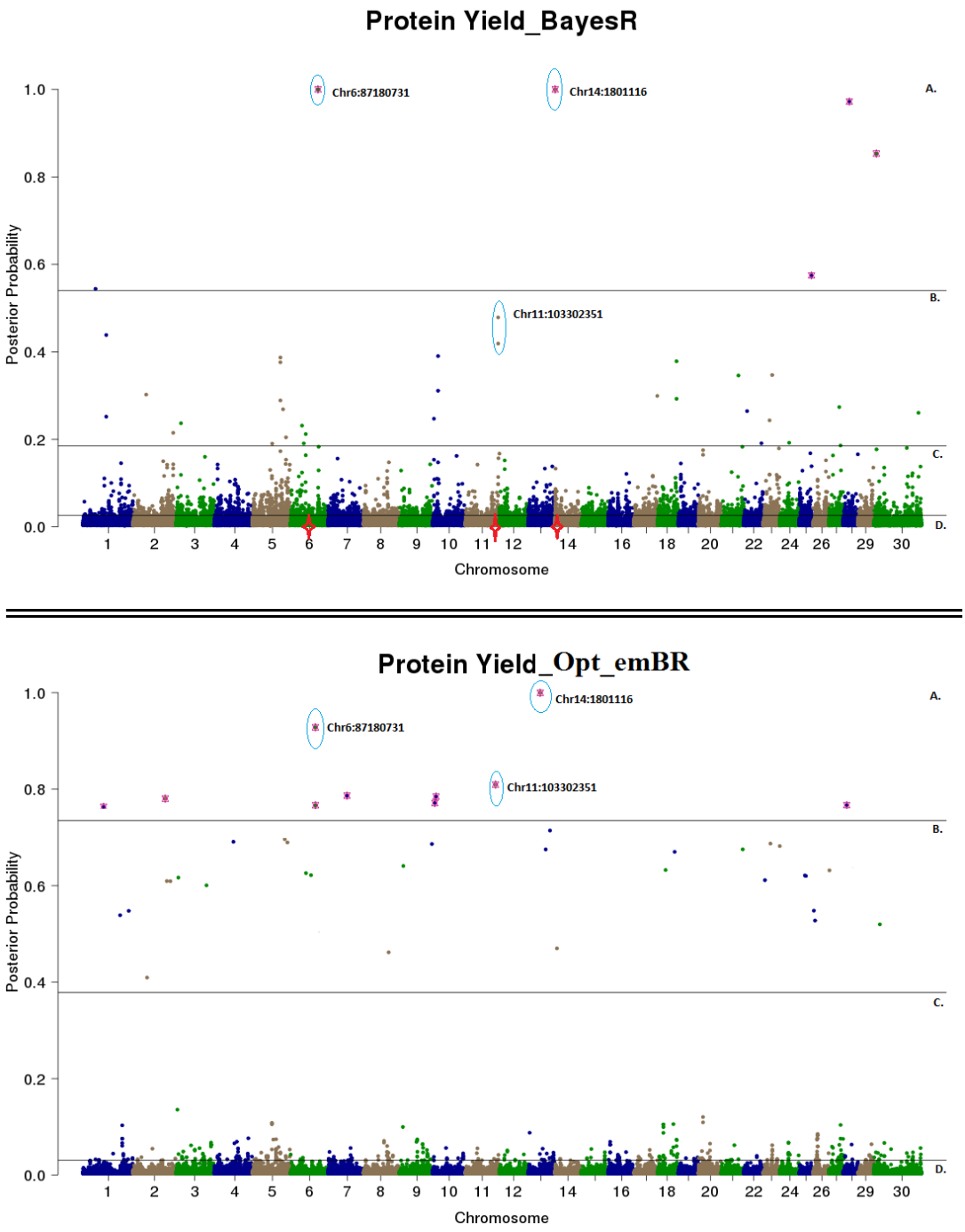


Figure 4.5. The mapping of all the SNPs estimated from BayesR and Opt_emBR on the whole chromosome related to protein yield by the posterior probability. The blue circle is the SNPs with location information mapped to known genes. In this Figure, there are four different levels (A, B, C, D) of SNPs that are classed into different proportions according to their posterior possibility.

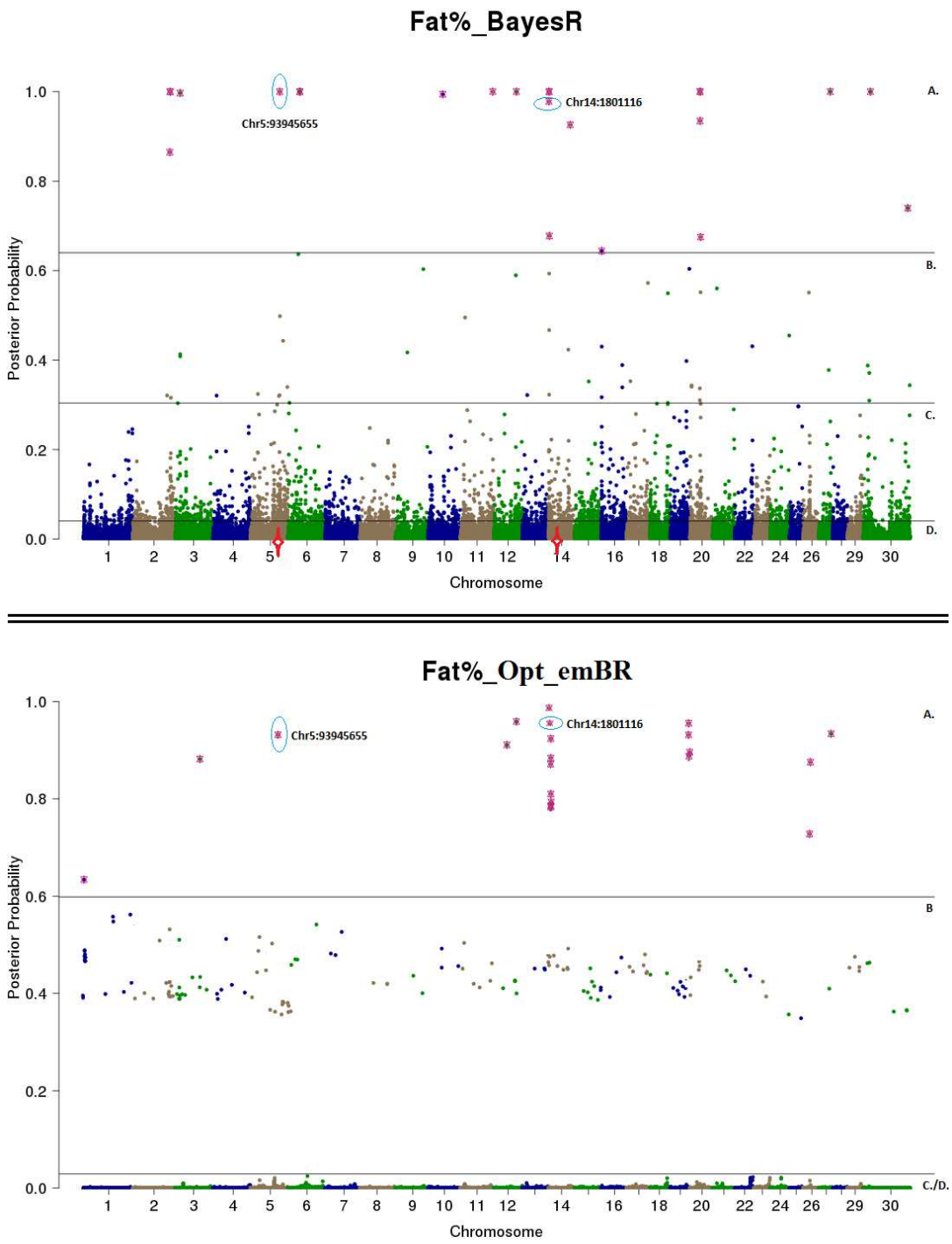


Figure 4.6. The mapping of all the SNPs estimated from BayesR and Opt_emBR on the whole chromosome related to Fat% by the posterior probability. The blue circle is the SNPs with location information mapped to known genes. In this Figure, there are four different levels (A, B, C, and D) of SNPs that are classed into different proportions according to their posterior possibility.

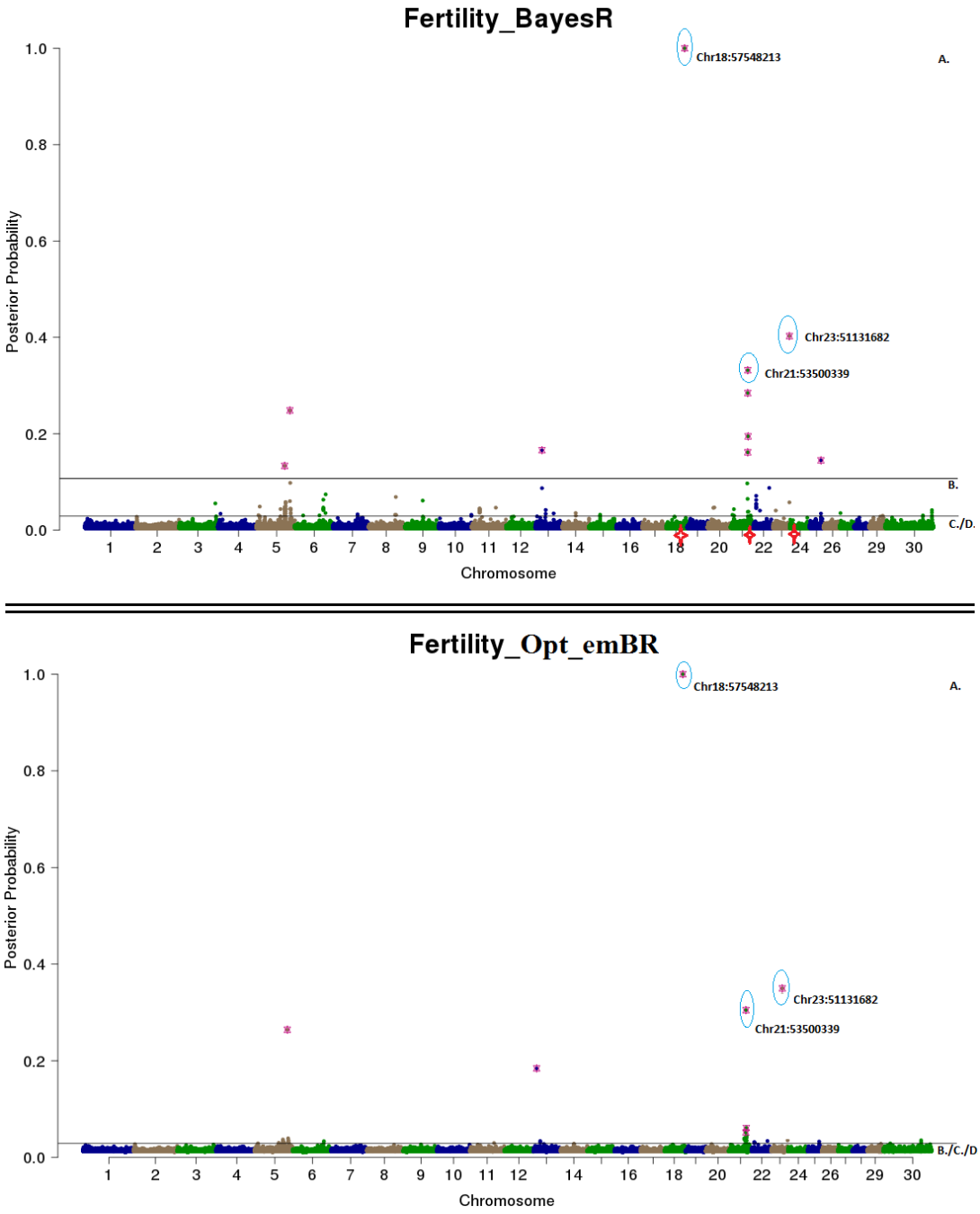


Figure 4.7. The mapping of all the SNPs estimated from BayesR and Opt_emBR on the whole chromosome related to Fertility by the posterior probability. The blue circle is the SNPs with location information mapped to known genes. In this Figure, there are four different levels (A, B, C, and D) of SNPs that are classed into different proportions according to their posterior possibility.

The genome location of SNPs with high posterior probabilities of being in the largest distribution was shown in Figure 4.4, Figure 4.5, Figure 4.6, and Figure 4.7, for Opt_emBR and BayesR, for each trait. Figure 4.4, Figure 4.5, and Figure 4.6 demonstrated the property of Opt_emBR by contrast with BayesR: 1) more SNPs were in the zero variance distribution; 2) more SNPs were in the moderate variance distribution. For fertility, even though Opt_emBR detected 2 less SNPs with large effects than BayesR, Opt_emBR still shrunk more SNP effects close to zero than BayesR in Figure 4.7. This potentially offered Opt_emBR the advantage of QTL mapping to filter large proportions of SNPs with very small effects. In the meantime, Opt_emBR could still detect SNPs in LD with known causal genes as accurately as (or even better than) BayesR. For milk yield, similar to BayesR, Opt_emBR found SNPs related to genes names *CSF2RB* (impacting milk yield) located at Chromosome 5 ranging between 75.575 and 75.775 MBP (millions of Base pairs); SNPS related to the Casein Gene *CSN1S1* on Chromosome 6 (~87MBP); And SNPs related to *CCL28/GHR* on Chromosome 20 ranging between 29.225MBP and 32.125MBP. For protein yield, the SNP close to *CSN1S1* on Chromosome 6 (~87MBP) was also detected by Opt_emBR; while *PAEP* specifically for Protein yield was also found on Chromosome 11 (~103MBP). Especially, Opt_emBR could detect *PAEP* with higher posterior probability much more clearly than BayesR. For Fat%, the gene termed *MGST1* on chromosome 5 influencing fat yield was obviously detected by both Opt_emBR and BayesR. Moreover, the well-known gene *DGAT1* (on Chromosome 14) (Grisart *et al.* 2002) was precisely mapped by BayesR and Opt_emBR. For Fertility, both BayesR and Opt_emBR detected an excellent SNP located on Chromosome 18 (~57MBP), which had been also detected by Pryce et al. (Pryce *et al.* 2010) and Cole et al. (Cole *et al.* 2011). Also the gene point on the chromosome 21 (ranging ~53K) had been described by McClure et al. (McClure *et al.* 2010), which controlled the percentage of unassisted births in first calf heifers in Angus cattle. And the SNP located on chromosome 23 (ranging ~51K) might have the linkage with the known

gene *GMDS*, which had not been proved yet for the fertility of cattle. Therefore, All above results demonstrated that Opt_emBR could perform QTL mapping with similar precision to BayesR with shrinking SNPs with smaller effects towards zero in a higher level. Additionally, these figures also showed that Opt_emBR would systematically detect more SNPs in high LD with causal mutations than BayesR except the fertility. The implementation principle of Opt_emBR and BayesR might decide this. As mentioned above, BayesR implemented MCMC and then averaged the solutions during the iterations with first certain 10,000~20,000 iterations as burn-in. For several SNPs with large effects, the effects might shrink too much due to averaging calculations. On the contrary, Opt_emBR applied EM loops to converge to a final optimal solution, which would not offset SNPs with very large effects during the iterations. For example, the detection for well-known gene termed *PAEP* influencing Protein Yield proved this in the Figure 4.5.

All the results demonstrate good performance of Opt_emBR on genomic predictions and QTLs mapping, but showing several limitations of Opt_emBR as well. Firstly, the EM algorithm of Opt_emBR can reduce up to 3% accuracy in comparison with BayesR on the traits with major gene effects (e.g. Fat percent). Theoretically, EM algorithms itself have the latent danger of falling into local optimal for the posterior distribution with multiple modes. To amend such problems, future work will develop hybrid schemes of our EM algorithms and Gibbs sampling to improve the prediction accuracy. A further improvement, particularly for QTL mapping, would be to include prior biology information into the model, which has already been attempted by Macleod et al. (MacLeod *et al.* 2016). In the meantime, as the 1000 bulls projects have been successfully conducted, the full sequence data of dairy cattle has been put into practice (MacLeod *et al.* 2016). Compared with 600K SNP panels that we use in this paper, the causal mutations actually exist in the whole genome sequence data, which can therefore improve the accuracy of gene identification.

Secondly, there are still one key part of the algorithm of Opt_emBR that consumes the running time and memory: calculates $tr(\mathbf{E^{-1}Z_i Z_i' E^{-1} PEV})$ for each SNP in front of EM loops. In detail, the calculation of $tr(\mathbf{E^{-1}Z_i Z_i' E^{-1} PEV})$ requires the time complexity of $(\frac{1}{2}mn^2)$, which accounts for almost 2/3 of the total computational time even though it happens in front of EM loops. Therefore, a future task is to implement this part on parallel processing with multi-thread to optimize the speed, which could improve the computational efficiency even more.

Our Opt_emBR has some features in common with other QTL mapping algorithms, such as BOLT-LMM, which incorporates Bayesian mixture models (fastBayesB (Meuwissen *et al.* 2009)) to improve the power of genetic associations identification with appealing outcomes. There are several common features between BOLT-LMM and our Opt_emBR methods. Theoretically, similar to Opt_emBR, BOLT-LMM borrows the variance information of the output from GBLUP and then fits into Bayesian models and associations studies as the start points. In addition, BOLT-LMM implement iterative conditional expectation (ICE) algorithm on Bayesian Lasso to transform the computational complexity from $O(mn^2)$ to $O(mn)$ (m is the number of markers and n is the number of individuals); while Opt_emBR implement another type of heuristic algorithms EM on BayesR model with the approximated computational complexity $O(mn)$. Compared with two distributions prior assumed by BOLT-LMM, the mixture of four normal distributions allows Opt_emBR to predict a wider range of gene effects so as to capture causal mutations more precisely than BOLT_LMM. Therefore, due to above theoretical analysis, the evidence provide by BOLT-LMM that Bayesian model can increase the computational efficiency and accuracy of association study in large size of human data sets gives us the confidence of implementing our Opt_emBR into Human disease prediction and causal gene identifications.

4.6 Conclusion

In summary, Opt_emBR is a computationally efficient method for simultaneous genomic prediction and QTL mapping in data sets including multiple populations, and with different accuracies of phenotypes. The heuristic expectation-maximization algorithm is implemented to make it converge to the final optimal using several orders less iteration than BayesR. The introduction of polygenic effects and weights terms into the linear model makes it applicable to multi-populations and across-populations predictions. Furthermore, the applications of two speed-up schemes make it up to 30 times faster than BayesR, while maintaining the similar accuracy. All of these results prove Opt_emBR a bright future of genomic predictions and QTL mapping on human disease data and whole genome sequence data of cattle.

4.7 Supporting information

4.8 Acknowledgements

The authors acknowledge the support and fund from Dairy Future CRC.

Chapter 5 A hybrid expectation maximization and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping

5.1 Chapter preface

Justification

In this paper, we propose a hybrid scheme for genomic prediction (termed HyB_BR) of an EM algorithm, followed by MCMC iterations without the burn-in stage. The EM algorithm effectively substitutes for the burn-in iterations usually required for the MCMC, but in a much shorter time. The result for accuracy of prediction of quantitative traits in cattle and disease trait in humans demonstrate that HyB_BR could perform equally well as Bayesian mixture models implemented with full MCMC. However, HyB_BR was up to 17 times faster than the full MCMC implementations, with the speed advantage increasing as the size of the data set increased. HyB_BR also performed as well as the full MCMC implementation of the Bayesian mixture model for QTL mapping, and for the inference of the underlying genetic architecture of human disease traits.

Publication status:

Published in the journal BMC Genomic*s, 2016*.

Submitted as

Wang T, Chen YPP, Bowman PJ, Goddard ME, Hayes BJ. (2016) A hybrid expectation maximization and MCMC sampling algorithm to implement Bayesian mixture model based genomic prediction and QTL mapping, BMC Genomics, 17:744.

Statement of contributions of joint authorship

Tingting Wang (Candidate): Developed, and implemented the hybrid algorithm on 600K SNP panels for multi-breed and across-breed prediction. Then, the author drafted the manuscript.

Yi-Ping Phoebe Chen (Principle Supervisor): gave instructions on the writing of the manuscript.

Phil J Bowman (Collaborator): provided help with C++ programming。

Michael E Goddard (Collaborator): contributed the idea about hybrid scheme.

Ben J. Hayes (Co-Supervisor): supervised this project, gave important instructions on organizing and revising the manuscript.

This chapter is an exact copy of the version submitted to BMC genomics. The reference style, table numbers and figure numbers have been carefully formatted.

5.2 Abstract

Background

Bayesian mixture models in which the effects of SNP are assumed to come from normal distributions with different variances are attractive for simultaneous genomic prediction and QTL mapping. These models are usually implemented with Monte Carlo Markov Chain (MCMC) sampling, which requires long compute times with large genomic data sets. Here, we present an efficient approach (termed HyB_BR), which is a hybrid of an Expectation-Maximization algorithm, followed by a limited number of MCMC without the requirement for burn-in.

Results

To test prediction accuracy from HyB_BR, dairy cattle and human disease trait data were used. In the dairy cattle data, there were four quantitative traits (milk volume, protein kg, fat% in milk and fertility) measured in 16,214 cattle from two breeds genotyped for 632,002 SNPs. Validation of genomic predictions was in a subset of cattle either from the reference set or in animals from a third breeds that were not in the reference set. In all cases, HyB_BR gave almost identical accuracies to Bayesian mixture models implemented with full MCMC, however computational time was reduced by up to 1/17 of that required by full MCMC. The SNPs with high posterior probability of a non-zero effect were also very similar between full MCMC and HyB_BR, with several known genes affecting milk production in this category, as well as some novel genes. HyB_BR was also applied to seven human diseases with 4,890 individuals genotyped for around 300K SNPs in a case/control design, from the Welcome Trust Case Control Consortium (WTCCC). In this data set, the results demonstrated again that HyB_BR performed as well as Bayesian mixture models with full MCMC for genomic predictions and genetic architecture inference while reducing the computational time from 45 hours with full MCMC to 3 hours with HyB_BR.

Conclusions

The results for quantitative traits in cattle and disease in humans demonstrate that HyB_BR can perform equally well as Bayesian mixture models implemented with full MCMC in terms of prediction accuracy, but with up to 17 times faster than the full MCMC implementations. The HyB_BR algorithm makes simultaneous genomic prediction, QTL mapping and inference of genetic architecture feasible in large genomic data sets.

5.3 Background

Genomic prediction of genetic merit, using SNP markers, is now routinely used in animal and plant breeding to identify superior breeding individuals and so

accelerate genetic gain (Meuwissen *et al.* 2001; Goddard & Hayes 2009; Meuwissen *et al.* 2013). Genomic prediction methodology is also increasingly used in human disease studies for the inference of genetic architecture, the identification of causal mutations (QTL mapping), and prediction of disease risk (de los Campos *et al.* 2010; Yang *et al.* 2010; Zhou *et al.* 2013b; Speed & Balding 2014; Moser *et al.* 2015).

Genomic predictions are often implemented using linear prediction models, especially best linear unbiased prediction (BLUP) or Genomic BLUP (GBLUP), which assume that all the SNPs contribute small effects to the trait and these effects are derived from a normal distribution (Meuwissen *et al.* 2001; VanRaden 2008; Yang *et al.* 2010). While GBLUP (Mäntysaari 2014), or its single-step implementation (Aguilar *et al.* 2010; Christensen & Lund 2010; Wolc *et al.* 2015), is one of the most popular genomic prediction methods implemented for official genomic evaluation in many countries, including Canada, New Zealand, Australia, Germany and Ireland, this approach does have some limitations. One limitation is that the prediction accuracy does not persist well across multiple generations, because the severe shrinkage in these models results in the effect of causative mutation being “smeared” across many markers encompassing long chromosome segments – in other words a linear combination of effects of a large number of markers is used to capture the effect of a QTL. After several generations, the association between markers and QTL might be broken down by recombination, thereby reducing prediction accuracy. The smearing of effect of causative mutations across many SNP also results in imprecise QTL mapping with BLUP methods.

To address these problems, Bayesian mixture models (nonlinear prediction e.g. Bayes A, B, C, and R) (Meuwissen *et al.* 2001; Habier *et al.* 2011; Erbe *et al.* 2012; Gianola 2013; Zhou *et al.* 2013b; Kemper *et al.* 2015) assume non-normal prior

distributions of SNP effects. One example of a flexible approach, BayesR (Erbe *et al.* 2012) defines a mixture model with SNP effects following a mixture of four normal distributions with zero, very small, small and moderate variances. In practice, the prediction accuracy of Bayesian mixture models (including BayesR) has been shown to be equal or superior to that of GBLUP for both human diseases and dairy cattle milk production traits (Ng-Kwai-Hang 1997; Grisart *et al.* 2002; Blott *et al.* 2003; Zhang *et al.* 2010; Lippert *et al.* 2011; Listgarten *et al.* 2012; Wang *et al.* 2012b; Zhou & Stephens 2012; Yang *et al.* 2014; Kemper *et al.* 2015; Loh *et al.* 2015).

In addition to the prediction of breeding values and future phenotypes using genotype data, Bayesian models (such as BayesR) can be used to map the causal polymorphisms (quantitative trait loci or QTL). For this purpose they have some advantages over traditional single SNP regression, which is commonly used to analyze genome wide association studies (GWAS) (Ng-Kwai-Hang 1997; Grisart *et al.* 2002; Blott *et al.* 2003; Zhang *et al.* 2010; Lippert *et al.* 2011; Listgarten *et al.* 2012; Wang *et al.* 2012b; Zhou & Stephens 2012; Yang *et al.* 2014). Single SNP regression fits one SNP at a time as a fixed effect and tests the significance of the association between the SNP and the trait, while ignoring all other SNPs. To protect against performing multiple tests, stringent P-values ($P<5*10-8$) are used. This method of analysis has three limitations:1) The effect of those SNPs declared significant is usually over-estimated; 2) multiple SNPs in linkage disequilibrium with the same QTL may show an association with the trait leading to imprecision in mapping the QTL; 3) many QTL are not detected at all because no SNP reaches the stringent P-value for association with the trait. By comparison, Bayesian mixture models fit all SNPs simultaneously by treating the SNP effects as random effects drawn from a prior distribution. For example, the BayesR model has been implemented for QTL detection in the dairy cattle genome and for human disease traits (Kemper *et al.* 2015). The results show that

BayesR can increase the power of identifying the known genes in contrast with GBLUP and GWAS.

Even though nonlinear models are attractive, one limitation is that nonlinear models typically require long computational times. Due to the hierarchical estimation of posterior distributions of SNP effects and their variances, nonlinear models have usually been implemented with Markov Chain Monte Carlo (MCMC). This requires a large number of iterations with time per iteration scaled linearly with the number of markers (m) and the number of individuals (n). Genomic data sets are now often very large and are rapidly becoming larger. For human, 300,000 to 9 million SNPs arrays genotyped on up to 253K individuals (The Wellcome Trust Case Control Consortium 2007; Wood *et al.* 2014) are available for association studies and disease/fitness prediction. In dairy cattle, whole genome sequence data including 39 million variants has been published by the 1,000 bull genomes project (Daetwyler *et al.* 2014). When confronted with such huge genomic data, Bayesian methods can be so computationally expensive that it is not possible to use them.

Two approaches have been used to develop more computationally efficient algorithms for implementing Bayesian mixture models. One is to modify MCMC with speed-up schemes. For example, Moser et al. (Moser *et al.* 2015) introduced a “500SNPs” scheme to pick 500 SNPs with non-zero effects to be updated instead of all the SNPs. Such modification schemes could reduce the running time by 3~6 fold. Calus et al. (Calus 2014) proposed a right-hand-side updating algorithm to cluster multiple SNPs (similar to a haplotype) to be updated as one during MCMC iterations. The results on 50K SNP panels demonstrated up to 90% reduction in computational time without reducing prediction accuracy. The other approach is to introduce heuristic methods (e.g. iterated conditional expectation, ICE; expectation maximization, EM) as an alternative to MCMC. There are a wide

range of fast versions of Bayesian approaches to genomic prediction using these methods (including fastBayesB, emBayesB, emBayesR) (Meuwissen *et al.* 2009; Hayashi & Iwata 2010; Shepherd *et al.* 2010; Yu & Meuwissen 2011; Sun *et al.* 2012; Wang *et al.* 2015), which are several orders faster than MCMC implementations. However, none of these algorithms gives consistently as high prediction accuracy as their MCMC counterparts. The EM method of Wang et al. (Wang *et al.* 2015), emBayesR, gave higher accuracy than ICE based methods but still had a reduction in accuracy of 5%~7% for traits with mutations of moderate to large effect. In other words, the heuristic approximations works best when there are no mutations of moderate to large effect, otherwise accuracy can be compromised. This is undesirable, especially when the largest advantage of the non-linear Bayesian methods over BLUP is observed when there are mutations of moderate to large effect (where moderate effect can mean a QTL explaining 1% of the variance if the data set is large)!

Motivated by the deficiency of both MCMC (long computing terms) and fast versions of nonlinear models (lower prediction accuracy with some genetic architectures), we hypothesize that a hybrid scheme, beginning with EM iterations and finishing with MCMC sampling iterations, would give similar prediction accuracy to a full MCMC implementation, while having a significant speed advantage. Here we propose a hybrid algorithm (termed HyB_BR) of Expectation-Maximization (EM) (emBayesR) and MCMC under the BayesR model. The algorithm also incorporates a speed-up scheme where only a proportion of SNP continue to be sampled in MCMC iterations. In comparison with emBayesR (Wang *et al.* 2015), the main improvement is that HyB_BR introduces a limited number of MCMC iterations after EM to improve the solutions from emBayesR.

To evaluate the predictive ability and computational efficiency of HyB_BR,

prediction accuracy was compared with BayesR and GBLUP in two data sets. The first data set was 800K SNP genotypes in 16,214 Holstein and Jersey bulls and cows. The prediction accuracy within these breeds and in a third breed (Aussie Red) was evaluated. The results showed that HyB_BR achieved very similar prediction accuracy to BayesR, while reducing the running time by up to 17 fold, and overcoming the limitations of slightly reduced accuracy of emBayesR. As a result of running the algorithm, the posterior probability of each SNP being in the model was derived, and this was used for QTL mapping. The resulting QTL regions were compared between the approaches and with previous literature reports. The results demonstrated that HyB_BR has enough power to detect the major known genes affecting milk production traits in dairy cattle as well as some novel regions. HyB_BR was also evaluated in a second data set - the Welcome Trust Case Control Consortium (WTCCC) human disease data set (The Wellcome Trust Case Control Consortium 2007). The results demonstrated that HyB_BR is a promising method for risk prediction and genetic architecture inference for human disease traits as well.

5.4 Methods and Materials

The mixture data model

The overall model at the level of the data for HyB_BR (independent of MCMC and EM implementation) including all the relevant parameters and priors is first described. The model assumes that $\mathbf{y}$, the phenotypic records of n individuals, was a linear model of fixed effects ($\boldsymbol{\beta}$), SNP effect ($\mathbf{g}$), random polygenic effects ($\mathbf{v}$) and environmental errors ($\mathbf{e}$):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{W}\mathbf{v} + \mathbf{e}, \quad (1)$$

where,

$\boldsymbol{\beta} =$ vector of p fixed effects, following uninformative priors.

$\mathbf{g} =$ vector of m SNP effects. For each SNP, $g_i \sim b(i,1) \times N\left(0,0 * \sigma_g^2\right) + b(i,2) \times N\left(0,0.0001 * \sigma_g^2\right) + b(i,3) \times N\left(0,0.001 * \sigma_g^2\right) + b(i,4) \times N\left(0,0.01 * \sigma_g^2\right)$, in which σ_g^2 is the genetic variance of the trait and $b(i,k)$ is a scalar with two possible values $\{0,1\}$, determining whether or not the effect of the *i*th SNP is derived from the *k*th distribution.

$\mathbf{Pr} =$ vector of probabilities where $Pr_k =$scalar with the value range between 0 and 1. The parameter $\mathbf{Pr}$ defines the proportion of all the SNPs following each of four normal distributions k. Pr_k is assumed to follow a Dirichlet distribution with the parameter $\alpha = (1,1,1,1)^{\mathrm{T}}$.

$\mathbf{v} =$ vector of q polygenic effects (breeding values, for the proportion of variance not explained by the SNP), with $\mathbf{v} \sim N(0, \mathbf{A}\sigma_a^2)$, $\mathbf{A}$ is the $q \times q$ pedigree-based relationship matrix, σ_a^2 is the polygenic variance, q is the number of individuals in the pedigree.

$\mathbf{e} =$ vector of n residual errors. For cattle data, $\mathbf{e} \sim N(0, \mathbf{E}\sigma_e^2)$, where **E** is a $n \times n$ diagonal matrix so that the error variance associated with different records could vary. For example, for bulls, the phenotype will be daughter yield deviations, which will have a lower error variance than the trait deviations (TD) of cows (Garrick *et al.* 2009). For human data where all phenotypes have the same magnitude of error, $\mathbf{E}$ matrix could be replaced by the identity matrix $\mathbf{I}$.

$\mathbf{X} =$ $n \times p$ design matrix, allocating phenotypes $\mathbf{y}$ to fixed effects $\boldsymbol{\beta}$.

$\mathbf{Z} =$ the $n \times m$ genotype matrix with elements $\mathbf{z}_{\mathrm{ij}} = (\mathbf{s}_{\mathrm{ij}} - 2p_i)/\sqrt[2]{2p_i(1-p_i)}$, in which $\mathbf{s}_{\mathrm{ij}}$ is the genotypes of the *j*th individual for the *i*th SNP (0, 1 or 2 copies of the second allele), and p_i is the allele frequency of each SNP *i*.

$\mathbf{W} =$ $n \times q$ design matrix, allocating the $q \times 1$ vector of polygenic effects to $\mathbf{y}$.

Note that model (1) extends the model used by Wang et al. (Wang *et al.* 2015) to include fixed effects, polygenic effects and weights.

The prior distribution of each SNP effect g_i conditional on $b(i,k)$ is

$$\mathrm{p}\big(g_i\big|b(i,k)\big)=\begin{cases}\delta(g_i), & b(i,1)=1\\ \frac{1}{\sqrt{2\pi\sigma_{\mathrm{i}}^2[k]}}\exp\left(-\frac{g_i^2}{2\sigma_{\mathrm{i}}^2[k]}\right), & b(i,k)=1(k=2,3,4)\end{cases}.$$

Where, $\delta(\mathrm{g_i})$ denotes the Dirac delta function with all probability mass at $\mathrm{g_i}=0$.

The joint distribution $\mathrm{p}(g_i,b(i,k)|Pr_k)$ (i.e. conditional on Pr_k) can be written as:

$$p(g_i,b(i,k)|Pr_k)=\prod_{k=1}^{4}p\big(g_i\big|b(i,k)\big)\times p(b(i,k)|Pr_k)$$

$$=(\delta(g_i)Pr_1)^{b(i,1)}\prod_{k=2}^{4}(\frac{1}{\sqrt{2\pi\sigma_{\mathrm{i}}^2[k]}}\exp\left(-\frac{g_i^2}{2\sigma_{\mathrm{i}}^2[k]}\right)Pr_k)^{b(i,k)} \quad (2)$$

The implementation of HyB_BR with the mixture model defined above consists of two components: 1) The Expectation-Maximization module. HyB_BR first implements the EM iterations under the mixture Gaussian model (equation 2) to give approximations for the parameter set including SNP effects $\mathbf{g}$, proportion of SNP in each distribution $\mathbf{Pr}$, error variance σ_e^2, and polygenic variance σ_a^2. For the estimation of each SNP effect, the PEV (predicted error variance) correction is introduced to account for the errors which are generated from the estimations of all other SNP effects (detailed in File S1-Chapter 10). 2) MCMC module. Once the EM steps are converged, the output values of the parameters are used in the modified MCMC iterations as the start values. For the final step, a MCMC scheme is implemented with a limited number of iterations.

EM module

In the following EM modules, the parameter set $=\{\mathbf{g},\mathbf{Pr},\boldsymbol{\beta},\mathbf{v},\ \sigma_e^2\}$ will be estimated by their maximum a posteriori (MAP) value. Similar to emBayesR

(Wang *et al.* 2015), all the parameters θ were estimated according to the expectation- maximization process with steps:

i) Define the log likelihood $f(\mathbf{y}|\theta)$ of the data under the data model (1).

ii) Derive the log posterior function of the parameters using Bayes' theorem. Following Bayes' theorem, the log posterior distribution of the parameter sets θ was based on the rule $logp(\theta|\mathbf{y}) \propto logf(\mathbf{y}|\theta) + logp(\theta)$, with $p(\theta)$ the prior for the parameter.

iii) Take the expectation on the posterior function over the missing data.

iv) Differentiate the expected posterior function and solve for this equal to zero to obtain MAP (Maximum A Posterior) of the parameter set θ.

In the Expectation-maximization implementation, the posterior distributions for each parameter $p(\theta|\mathbf{y})$ are obtained while "integrating out" the other parameters. For example, for the estimation of each SNP effect, we maximize the posterior distribution of each SNP effect $p(g_i|\mathbf{y}, b(i,k), \mathrm{Pr_k}, \boldsymbol{\beta}, \mathbf{v}, \sigma_e^2)$ by integrating out the other SNP effects $g_{j\neq i}$, the parameters $b(i,k), \boldsymbol{\beta}, \mathbf{v}$, but we fix the proportion parameter $\mathrm{Pr_k}$ and the error variance σ_e^2 at their maximum posterior estimates. In the following, we will detail the inference process for several key parameters including SNP effect ($\mathbf{g}$), the mixing parameters ($\mathbf{Pr}_k$), fixed effects ($\boldsymbol{\beta}$), polygenic effects ($\mathbf{v}$) and the error variance (σ_e^2) separately:

1) Estimation of SNP effect $\mathbf{g}$.

As in our EM version of BayesR (Wang *et al.* 2015), in HyB_BR we will update the estimated effect of SNPs one at a time. Therefore, we rewrite $\mathbf{Zg}$ in equation (1) as the sum of the effect of the current SNP $\mathbf{Z_i}g_i$ and the combined effect of all other SNP effects $\mathbf{u_i}$ ($\mathbf{u_i} = \sum_{j\neq i} \mathbf{Z_j}g_j$). We rewrite the model (1) as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_i}g_i + \mathbf{u} + \mathbf{Wv} + \mathbf{e} \qquad (1a)$$

where, $g_i(\text{the effect of SNP } i)$ is the i^{th} element of the vector $\mathbf{g}$, and $\mathbf{u_i} = \sum_{j \neq i} \mathbf{Z_j} g_j$.

We estimate of $\hat{g}_i$ by the value of g_i that maximises the posterior probability $\mathrm{P}(g_i|\mathbf{y}, \widehat{\mathbf{Pr}}, \widehat{\sigma_e^2})$ where $\widehat{\mathbf{Pr}}$ and $\widehat{\sigma_e^2}$ are the MAP estimates of $\mathbf{Pr}$ and σ_e^2 conditional on $\mathbf{y}$.

To perform this, we first introduce "missing data" $(b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u})$, and then "integrate them out" via the Expectation-maximization steps. In detail, the marginal posterior distribution of each SNP effect g_i can be written as:

$$p\left(g_i, b(i,k)\middle|\mathbf{y}, \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}, \widehat{\sigma_e^2}, \widehat{Pr_k}\right) \propto p(\mathbf{y}|g_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}, \widehat{\sigma_e^2}, \widehat{Pr_k}) p\left(g_i, b(i,k)\middle|\widehat{Pr_k}\right).$$

Under the model (1a), the first term $p(\mathbf{y}|g_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}, \widehat{\sigma_e^2}, \widehat{Pr_k})$ is obtained according to the following normal distribution:

$$\mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathbf{Z_i} g_i - \mathbf{W}\mathbf{v} - \mathbf{u} \sim N(0, \mathbf{E}\sigma_e^2),$$

which can be transformed as:

$$\mathbf{e}^* - \mathbf{Z_i} g_i \sim N(0, \mathbf{E}\sigma_e^2),$$

Where, $\mathbf{e}^* = \mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{v} - \mathbf{u}$.

Therefore, the term $p(\mathbf{y}|g_i, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}, \widehat{\sigma_e^2}, \widehat{Pr_k})$ can be written as: $p\left(\mathbf{y}\middle|g_i, \mathbf{u}, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}, \widehat{\sigma_e^2}, \widehat{Pr_k}\right) = \frac{1}{(2\pi\widehat{\sigma_e^2})^{\frac{n}{2}}} \frac{1}{|\mathbf{E}|} exp\left[-\frac{1}{2\widehat{\sigma_e^2}} (\mathbf{e}^* - \mathbf{Z_i} g_i)' \mathbf{E^{-1}} (\mathbf{e}^* - \mathbf{Z_i} g_i)\right]$

Then the log likelihood function is:

$$logp\left(\mathbf{y}\middle|g_i, \mathbf{u}, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \widehat{\sigma_e^2}, \widehat{Pr_k}\right) = -\frac{\mathrm{n}}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| - \frac{1}{2\widehat{\sigma_e^2}} (\mathbf{e}^* - \mathbf{Z_i} g_i)' \mathbf{E^{-1}} (\mathbf{e}^* - \mathbf{Z_i} g_i) \quad (3)$$

The second term $p\left(g_i, b(i,k)\middle|\widehat{Pr_k}\right)$ is defined in the equation (2). Then the log of $p\left(g_i, b(i,k)\middle|\widehat{Pr_k}\right)$ is:

$$logp\left(g_i, b(i,k)\middle|\widehat{Pr_k}\right) = b(i,1)\log\left(\delta(g_i)\widehat{\mathrm{Pr}}_1\right) + \sum_{\mathrm{k}=2}^{4} b(i,k)\left(-\frac{1}{2}\log\sigma_i^2[k] - \frac{g_i^2}{2\sigma_i^2[k]} + \log\widehat{\mathrm{Pr}}_{\mathrm{k}}\right) \quad (4)$$

Treating $(\mathbf{e}^*, b(i,k))$ as missing data and omitting the terms without g_i, the expectation of the log marginal posterior of each SNP effect is:

$$E_{\mathbf{e}^*,b(i,k)} logp\left(g_i, b(i,k)\middle|\mathbf{y}, \boldsymbol{\beta}, \mathbf{v}, \mathbf{u}, \widehat{\sigma_e^2}, \widehat{Pr_k}\right)$$

$$= E_{\mathbf{e}^*,b(i,k)} logp\left(\mathbf{y}\middle|g_i, \mathbf{u}, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \widehat{\sigma_e^2}, \widehat{Pr_k}\right) + E_{\mathbf{e}^*,b(i,k)} logp\left(g_i, b(i,k)\middle|\widehat{Pr_k}\right)$$

According to equation (3), the expectation of the first term is:

$$E_{\mathbf{e}^*,b(i,k)} logp\left(\mathbf{y}\middle|g_i, \mathbf{u}, b(i,k), \boldsymbol{\beta}, \mathbf{v}, \widehat{\sigma_e^2}, \widehat{Pr_k}\right)$$

$$\propto -\frac{1}{2\widehat{\sigma_e^2}}\{(\mathbf{e}^* - \mathbf{Z}_i \mathrm{g}_i)'\mathbf{E}^{-1}(\mathbf{e}^* - \mathbf{Z}_i \mathrm{g}_i) + tr(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e}^*))\} \quad (5)$$

According to the equation (4), the expectation of the second term is:

$$E_{\mathbf{e}^*,b(i,k)} logp\left(g_i, b(i,k)\middle|\widehat{Pr_k}\right)$$

$$\propto P(i,1)\log\left(\delta(g_i)\widehat{\mathrm{Pr}}_1\right) + \sum_{k=2}^{4} P(i,k)\left(-\frac{1}{2}\log\sigma_i^2[k] - \frac{g_i^2}{2\sigma_i^2[k]} + \log\widehat{\mathrm{Pr}}_k\right) \quad (6)$$

Where, $P(i,k) = E(b(i,k)|\mathbf{y}, \widehat{Pr_k})$. The term $P(i,k)$ can be derived as in the File S2 (Chapter 10).

Hence, in the Maximization steps of EM, we differentiate equations (5) and (6) respect to $\hat{g}_i$, and then obtain an estimate for the SNP effect as:

$$\frac{\partial E_{\mathbf{e}^*,b(i,k)} logp\left(g_i, \mathbf{u}, b(i,k)\middle|\mathbf{y}, \widehat{\boldsymbol{\beta}}, \hat{\mathbf{v}}, \widehat{\sigma_e^2}, \widehat{Pr_k}\right)}{\partial \mathrm{g}_i} = \left[-\sum_{k=2}^{4}\frac{P(i,k)}{\sigma_i^2[k]} - \frac{\mathbf{Z}_i'\mathbf{E}^{-1}\mathbf{Z}_i}{\widehat{\sigma_e^2}}\right] g_i + \frac{\mathbf{Z}'\mathbf{E}^{-1}\mathbf{e}^*}{\widehat{\sigma_e^2}}$$

Setting $\frac{\partial E_{\mathbf{e}^*,b(i,k)} logp\left(g_i,\mathbf{u},b(i,k)\middle|\mathbf{y},\widehat{\boldsymbol{\beta}},\hat{\mathbf{v}},\widehat{\sigma_e^2},\widehat{Pr_k}\right)}{\partial \mathrm{g}_i} = 0$, then we can derive the effect $\hat{g}_i$ as:

$$\hat{g}_i = [\mathbf{Z}_i'\mathbf{E}^{-1}\mathbf{Z}_i + \sum_{k=1}^{4}\left(P(i,k)\frac{\widehat{\sigma_e^2}}{\sigma_i^2[k]}\right)]^{-1}[\mathbf{Z}'\mathbf{E}^{-1}\mathbf{e}^*] \quad (7)$$

2) Estimation of parameter $\mathbf{Pr}$.

This follows a common method for an EM algorithm to analyze a mixture of distributions. We introduce the 'missing data' $b(i,k)$, which is the indicator variable that indicates which of the k=4 distributions SNP effect i is drawn from. The posterior distribution of proportion parameter $\mathbf{Pr}$ is:

$$p(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b})p(\mathbf{b}|\mathbf{Pr})p(\mathbf{Pr})$$

Where,

The term $p(\mathbf{y}|\mathbf{b})$ does not involve $\mathbf{Pr}$. So when we differentiate with respect to **Pr**, this term will drop out and therefore can be ignored.

$p(\mathbf{b}|\mathbf{Pr}) = \prod_{i=1}^{n}\prod_{k=1}^{4}(Pr_k)^{b(i,k)}$

$p(\mathbf{Pr}) = \prod_{k=1}^{4} Pr_k$

Therefore, the log posterior expression of $\mathbf{Pr}$ can be written as:

$logp(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) \propto logp(\mathbf{b}|\mathbf{Pr}) + logp(\mathbf{Pr})$

$= \sum_{i=1}^{n}\sum_{k=1}^{4} b(i,k) logPr_k + \sum_{k=1}^{4} logPr_k.$

Treating $\mathbf{b}$ as the missing data and defining $P(i,k) = E(b(i,k)|\mathbf{y}, Pr_k)$, the expectation of the posterior can be written as:

$$E_{\mathbf{b}|\mathbf{y}} logp(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) = \sum_{i=1}^{n}\sum_{k=1}^{4} P(i,k) logPr_k + \sum_{k=1}^{4} logPr_k. \qquad (8)$$

Introducing Lagrange multiplier λ to account for the constraint that $\sum_{k=1}^{4} Pr_k$=1 and differentiate with respect to Pr_k, the parameter $\mathbf{Pr}$ can be estimated by:

$$\frac{\partial \mathrm{E}_{\mathbf{b}|\mathbf{y}}[logp(\mathbf{Pr}, \mathbf{b}|\mathbf{y}) + \lambda(\sum_{\mathrm{k}=1}^{4} Pr_k - 1)]}{\partial Pr_k} = \frac{\sum_{i=1}^{m} P(i,k)}{Pr_k} + \frac{1}{Pr_k} + \lambda = 0$$

$$Pr_k = \frac{\sum_{\mathrm{i}=1}^{m} P(i,k)+1}{\sum_{k=1}^{4}(\sum_{i=1}^{m} P(i,k)+1)} \qquad (9)$$

The computation of $P(i,k)$ is given in the File S2 (Chapter 10).

3) Estimation of fixed effects (β) and the error variance (σ_e^2).

Since fixed effects (**β**) and the error variance have uninformative priors, their posterior distribution is:

$$p(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y}) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}}\frac{1}{|\mathbf{E}|} exp\left[-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\hat{\mathbf{v}})'\mathbf{E}^{-1}(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\hat{\mathbf{v}})\right]$$

Due to the equation $\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\hat{\mathbf{v}} = \mathbf{e}$, the full log likelihood based on this model is:

$$logp(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y}) = -\frac{\mathrm{n}}{2}\log\sigma_{\mathrm{e}}^2 + \frac{1}{2\sigma_{\mathrm{e}}^2}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e} \quad (10)$$

Treating $\mathbf{e}$ as the missing data, the expectation of the equation (10) can be expressed as

$$\mathrm{E}_{\mathbf{e}|\mathbf{y}} logp(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y}) = \mathrm{E}_{\mathbf{e}|\mathbf{y}}\left[-\frac{\mathrm{n}}{2}\log\sigma_{\mathrm{e}}^2 + \frac{1}{2\sigma_{\mathrm{e}}^2}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e}\right]$$

$$= -\frac{\mathrm{n}}{2}\log\sigma_{\mathrm{e}}^2 + \frac{1}{2\sigma_{\mathrm{e}}^2}[\mathbf{e}'\mathbf{E}^{-1}\mathbf{e} + tr(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e}))]$$

In theory, $\mathrm{PEV}(\mathbf{e}) \neq \mathrm{PEV}(\mathbf{e}^*)$ due to $\mathbf{e} = \mathbf{e}^* + \mathbf{Z}_{\mathrm{i}}g_i$. However, since each SNP effect is shrunk severely towards zero by GBLUP (Yang *et al.* 2010), we approximate $\mathrm{PEV}(\mathbf{e}) \cong \mathrm{PEV}(\mathbf{e}^*)$. The calculation of the term $\mathrm{PEV}(\mathbf{e}^*)$ is detailed in the File S1 (Chapter 10).

Therefore, differentiating the equation $\mathrm{E}_{\mathbf{e}|\mathbf{y}} logp(\sigma_e^2, \boldsymbol{\beta}, \hat{\mathbf{g}}|\mathbf{y})$ with regard to σ_{e}^2 and $\mathbf{b}$, we achieve:

$$\widehat{\sigma_{\mathrm{e}}^2} = \frac{1}{\mathrm{n}}[(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\hat{\mathbf{v}})'\mathbf{E}^{-1}(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\hat{\mathbf{v}}) + tr(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e}^*))] \quad (11)$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{E}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}^{-1}(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{W}\hat{\mathbf{v}}) \quad (12)$$

4) Estimation of polygenic effects (v)

Under the model (1), the conditional posterior density function of polygenic effects $\mathbf{v}$ is:

$$p(\mathbf{v}|\mathbf{y}) = p(\mathbf{y}|\mathbf{v}, \hat{\mathbf{g}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma_e^2})p(\mathbf{v})$$

Where,

$$p(\mathbf{y}|\mathbf{v}, \hat{\mathbf{g}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma_e^2}) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}}\frac{1}{|\mathbf{E}|}exp\left[-\frac{1}{2\sigma_e^2}\left(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\mathbf{v}\right)'\mathbf{E}^{-1}\left(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\mathbf{v}\right)\right] \quad (13)$$

$$p(\mathbf{v}) = \frac{1}{(2\pi\sigma_{\mathrm{a}}^2)^{\frac{q}{2}}}\frac{1}{|\mathbf{A}|}exp\left[-\frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \quad (14)$$

Therefore, the log posterior based on equation (13) and (14) is:

$$logp(\mathbf{v}|\mathbf{y}) = logf\left(\mathbf{y}\middle|\mathbf{v}, \hat{\mathbf{g}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma_e^2}\right) + logp(\mathbf{v})$$

$$= \left[-\frac{n}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| + \frac{1}{2\widehat{\sigma_e^2}}\left(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\mathbf{v}\right)'\mathbf{E}^{-1}\left(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\mathbf{v}\right)\right]$$

$$+ \left[-\frac{q}{2}\log\sigma_a^2 - \log|\mathbf{A}| + \frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \quad (15)$$

According to the equation $\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\hat{\mathbf{v}} = \mathbf{e}$, the equation (15) can be written as:

$$logp(\mathbf{v}|\mathbf{y}) = logf\left(\mathbf{y}\middle|\mathbf{v}, \hat{\mathbf{g}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma_e^2}\right) + logp(\mathbf{v})$$

$$= \left[-\frac{n}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| + \frac{1}{2\widehat{\sigma_e^2}}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e}\right]$$

$$+ \left[-\frac{q}{2}\log\sigma_a^2 - \log|\mathbf{A}| + \frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \quad (16)$$

Taking expectation over the missing data $\mathbf{e}$, we get:

$$\mathrm{E}_{\mathbf{e}|\mathbf{y}} logp(\mathbf{v}|\mathbf{y}) = \left[-\frac{n}{2}\log\widehat{\sigma_e^2} - \log|\mathbf{E}| + \frac{1}{2\widehat{\sigma_e^2}}\mathbf{e}'\mathbf{E}^{-1}\mathbf{e} + tr\left(\mathbf{E}^{-1}\mathrm{PEV}(\mathbf{e})\right)\right]$$

$$+ \left[-\frac{q}{2}\log\sigma_a^2 - \log|\mathbf{A}| + \frac{1}{2\sigma_a^2}\mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\right] \quad (17)$$

Differentiating the equation (17) with regards to $\mathbf{v}$, we get:

$$\hat{\mathbf{v}} = (\mathbf{W}'\mathbf{E}^{-1}\mathbf{W}\sigma_a^2 + \sigma_e^2\mathbf{A}^{-1})^{-1}\sigma_a^2\mathbf{W}'^{\mathbf{E}^{-1}}\left(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \quad (18)$$

where, for simplicity, the variance σ_a^2 will be fixed as the specified value from GBLUP estimation.

All the parameters and their equation derived from EM steps are listed in Table 5.1.

Table 5.1. The list of all the estimated parameters. The parameter list includes the possibility for each SNP ($\mathbf{P(i,k)}$), the proportion parameter ($\mathbf{Pr}$), each SNP effect ($\mathbf{g_i}$), error variance ($\boldsymbol{\sigma_e^2}$), fixed effect ($\boldsymbol{\beta}$), and polygenic effects $\mathbf{v}$ and the according equation derived from EM steps.

Parameters	The data model	According equations derived from EM
$E_{e^*}logP(i,k)$	The expected likelihood parameters for $P(i,k)$	Equation (S3)
$P(i,k)$	SNP effects related parameters under the extended model (1a)	Equation (S4)
Pr		Equation (9)
g_i		Equation (7)
σ_e^2	The overall model (1)	Equation (11)
β		Equation (12)
v		Equation (18)

Steps for EM module. The overall procedure of EM is described by means of the pseudo code, steps ①~⑦. Here we will detail these steps according to their sequence appearing in the pseudo code descriptions:

Step $EM\_①$: Initialise the parameters **g**, $\mathbf{Pr}$, $\boldsymbol{\sigma_i^2}$ and Construct **X, A, G, E, W** matrices. Similar to emBayesR (Wang *et al.* 2015), the starting values of **g** and $\mathbf{Pr}$ are set as $\mathbf{g} = 0.01$ and $\mathrm{Pr} = \{0.5, 0.487, 0.01, 0.003\}$, while $\boldsymbol{\sigma_i^2} = \{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2\}$. The genetic variance σ_g^2 and polygenic variance σ_a^2 are obtained from GBLUP. Both variances will not be updated during EM iterations.

The $n \times 3$ matrix **X** is design matrix, allocating the phenotypes to fixed effects. In our case, matrix **X** is set up with first column being the mean, the second and third columns defining the breeds (Holstein or Jersey) and sex (bulls or cows) of the cattle. For example, if the i^{th} animal is a Holstein bull, then $x_{i,2} = 1$ with $x_{i,3} = 0$.

The Pedigree relationship matrix **A** is built using Henderson's rules (Henderson 1984); while the genomic relationship matrix $\mathbf{G}$ is constructed using the equation $\mathbf{G} = \mathbf{ZZ}'/n$. Diagonal error matrix **E** is calculated following Garrick et al. (Garrick *et al.* 2009), and the index matrix **W** maps individuals in the pedigree

to phenotypes if they had one .

Step $EM\_②$: Calculate the PEV matrix under model 1, as previously described. Then using PEV matrix, calculate $tr\left(\mathbf{E^{-1}Z_i Z_i' E^{-1} PEV_{u_1}(e)}\right)$ which is used in the equation for $E_{e^*}logP(i,k)$ (File S2 (Chapter 10)) and is calculated before EM iterations, to save computational time.

Then for each SNP i (i in 1 to m)

Step $EM\_③$: Correct **y** for the effects of all other SNPs except current SNP i with equation $\mathbf{y}^{\dagger} = \mathbf{y} - \sum_{j \neq i} \mathbf{Z_j}\,\hat{g}_j - \mathbf{X\hat{b}} - \mathbf{W\hat{v}}$.

Step $EM\_④$: Estimate the probability that the effect of SNP i is from one of four normal distributions $E_{e^*}logP(i,k)$ with the equation (S3). After this, $P(i,k)$ is calculated with the equation $exp(E_{e^*}logP_{ik}/\sum_{k=1}^{4} exp(E_{e^*}logP_{ik}))$ (S4).

Step $EM\_⑤$: the SNP effect $\hat{g}_i$ is updated via equation (7).

After effects have been estimated for all SNP,

Step $EM\_⑥$: Estimate σ_e^2 with equation (11), fixed effects **β** with equation (12), update Pr_k with equation (9) , and update polygenic effects **v** with the equation (18).

Step $EM\_⑦$: Assess convergence criterion $(\hat{\mathbf{g}}^{l} - \hat{\mathbf{g}}^{l-1})'(\hat{\mathbf{g}}^{l} - \hat{\mathbf{g}}^{q-1})/(\hat{\mathbf{g}}^{l'}\hat{\mathbf{g}}^{l}) \leq 10^{-10}$ with l being the EM iterations number. If not converged, then return to Step ③ for the next EM iteration; otherwise, exit the EM iterations and return the estimates of parameters from the final iterations.

Modified MCMC module with speed-up scheme

The outputs of the EM including SNP solutions, polygenic effects, error variance and genetic variance are used as starting values of parameters for the MCMC module, which allows MCMC to begin with no burn in.

The MCMC module of HyB_BR implements the same Gibbs sampling processes

as BayesR (Kemper *et al.* 2015) but modified with one speed-up scheme as follows. Over the first 500 iterations, the average probability that the SNP effect was zero ($p(i,1)$) is calculated. If $p(i,1) \geq \mathrm{a}$, then the SNP effect is set to zero and was not updated in future iterations.

The test for selecting a reasonable value of the parameter a was conducted as follows. The impact of value of *a* from 0.85 to 0.95 on prediction accuracy was investigated for the milk production traits and fertility, Figure 5.1. The results show that criterion $p(\mathrm{i},1) \geq 1$, is the lowest threshold which gives an accuracy very close to the maximum. The criterion means SNP i has more than 90% probability of having no effect.

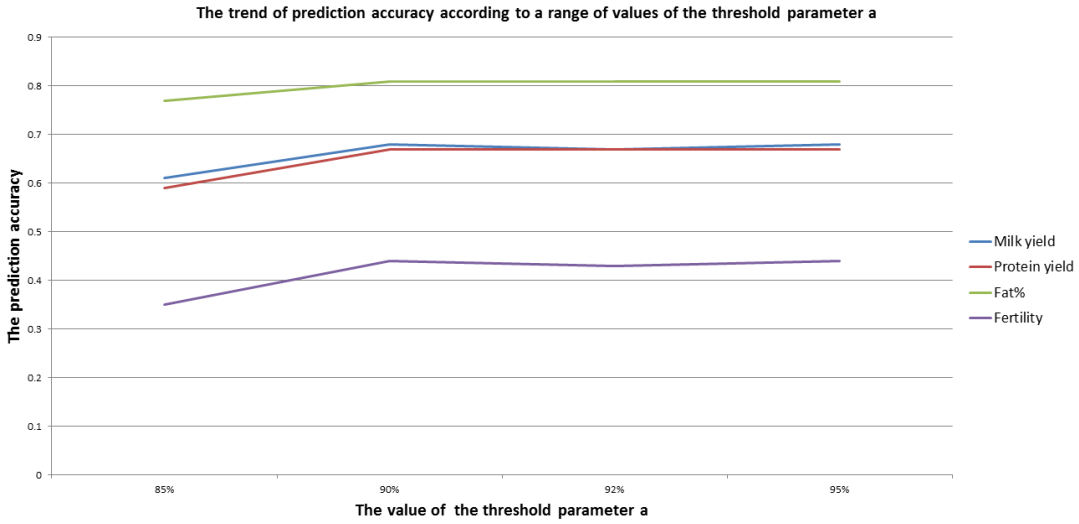


Figure 5.1. The trend of prediction accuracy according to a range of values of the threshold parameter $\mathbf{a}$.

The modified MCMC steps can then be described as follows:

Step $MCMC\_①$: sample the error variance $\hat{\sigma}_{\mathrm{e}}^2$ according to the distribution $\hat{\sigma}_{\mathrm{e}}^2 \sim Inv-\chi^2(n-2, \frac{\mathbf{y}^{*\prime}\mathbf{E}^{-1}\mathbf{y}^{*}}{n-2})$, with $\mathbf{y}^{*} = \left(\mathbf{y} - \mathbf{Zg} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{W}\hat{\mathbf{v}}\right)$.

Step $MCMC\_②$: sample the fixed effects $\boldsymbol{\beta}$ from the distribution $N(\boldsymbol{\beta}_{\mu}, (\mathbf{X}'\mathbf{E}^{-1}\mathbf{X})^{-1}\hat{\sigma}_{\mathrm{e}}^2)$, with $\boldsymbol{\beta}_{\mu} = (\mathbf{X}'\mathbf{E}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}^{-1}(\mathbf{y} - \mathbf{Z}\hat{\mathbf{g}} - \mathbf{W}\hat{\mathbf{v}})$.

Step $MCMC\_③$: Polygenic variance is sampled $\hat{\sigma}_{\mathrm{a}}^2 \sim Inv-\chi^2(n-2, \frac{\hat{\mathbf{v}}'\mathbf{A}^{-1}\hat{\mathbf{v}}}{n-2})$.

Step $MCMC\_④$: The polygenic effects are sampled from normal distribution $N(\mu, s)$, with the mean $\mu = \hat{\mathbf{v}}$ from equation (20) and the variance $s = (\mathbf{W}'\mathbf{E}^{-1}\mathbf{W} + \mathbf{A}^{-1}\sigma_e^2/\sigma_a^2)^{-1}$.

Then for each SNP i (i in 1 to m),

Step $MCMC\_⑤$: Implement the speed-up scheme **:** if (iterations > 500) and ($P(i,1) > 0.9$), then stop updating this SNP i.

Else,

Step $MCMC\_⑥$: Estimate the log likelihood that the effect of SNP i is from one of four normal distributions $L(g_i|\sigma_i^2[k])$. Following the derivation steps of Kemper et al. (Kemper *et al.* 2015), the optimized equation of the log likelihood function is

$$L(g_i|\sigma_i^2[k]) = -\frac{1}{2}\{\log(\sigma_i^2[k]\mathbf{Z_i'Z_i} + \sigma_e^2) + ((e^*)'\mathbf{E^{-1}Z_i})^2\sigma_i^2[k]\sigma_e^{-2}/(\sigma_i^2[k]\mathbf{Z_i'E^{-1}Z_i} + \sigma_e^2)\} + log\Pr_k,$$

with $e^* = \mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathbf{u} - \mathbf{Wv}$.

After this, $P(i,k)$ is calculated with the equation:

$$exp(L(g_i|\sigma_i^2[k])/\sum_{k=1}^{4} exp(L(g_i|\sigma_i^2[k])).$$

Step $MCMC\_⑦$: Sample $\hat{\mathrm{g}}_i$ with $\mathrm{N}(\mu, s)$, $\mu = [\mathbf{Z_i'E^{-1}Z_i} + \frac{\widehat{\sigma_e^2}}{\sigma_i^2[k]}]^{-1}[\mathbf{Z'E^{-1}}\, e^*]$, and $s = [\mathbf{Z_i'E^{-1}Z_i} + \frac{\widehat{\sigma_e^2}}{\sigma_i^2[k]}]^{-1}$.

Step $MCMC\_⑧$:Update $\Pr \sim \mathrm{Dirichlet}(\beta_1 + 1, \beta_2 + 1, \beta_3 + 1, \beta_4 + 1)$,where $\beta_1, \beta_2, \ldots, \beta_4$ are the number of SNPs in one of four normal distributions.

Return to MCMC step 1.

HyB_BR is written in the C++ programming language.

Data

Cattle. 1,745 Holstein and Jersey cattle and 114 Australian Red bulls were genotyped with the 777K Illumina HD bovine SNP chip. 15,049 Holstein and Jersey bulls and cows, 249 Australian red bulls and cows were genotyped with the 54K Illumina Bovine SNP array. After stringent quality control and SNP filtering described in (Erbe *et al.* 2012), there were 632,003 SNPs remaining for animals genotyped with the 777K SNP panel, and 43,425 SNPs remaining for animals genotyped with the 54K SNP array. Animals genotyped with the 43,425 SNPs were imputed to 632,003 SNP genotypes using Beagle 3.0 (Browning & Browning 2009). Therefore, the total data set was 17,157 cattle of three breeds with real or imputed genotypes for 632,003 SNP.

The phenotypes include milk yield, protein yield, fat percent(fat%), and cow fertility. The heritability of these traits varies from 0.33 (for milk yield, protein yield and fat%), to 0.03 (for cow fertility). The fertility (reproductive performance of dairy cows) is usually measured according to calving interval (CI, the number of days between successive calvings), days from calving to first service (CFS), pregnancy diagnosis, lactation length (LL), and survival to second lactation on Australian Holstein and Jersey cows (Haile-Mariam *et al.* 2013; Haile-Mariam *et al.* 2015). Here, the fertility phenotype was calving interval (CI). The phenotypes for all the traits were daughter trait deviations (DTD) for bulls (the average of their daughters phenotypes, corrected for fixed effects), and trait deviations (TD) for cows (as described by Kemper et al. 2015 (Kemper *et al.* 2015)). For genomic prediction, the data was separated into a reference set, where SNP effects were estimated, and validation sets, where the prediction accuracy was assessed, and the division of animals into reference and validation sets was by year of birth (youngest animals in validation set). The reference data includes bulls and cows from two breeds (Holstein and Jersey), and the predictions were evaluated in the other

animals of the same breeds or in a breed (Aussie red) not included in the reference set. The exact number of individuals in these data sets for each trait is given in

Table 5.2.

Table 5.2. The number of individuals in the reference sets and validations sets related to three traits including Milk yield (MilkY), Protein yield (ProtY), Fat Percent (Fat%) and Fertility.

Traits	**Reference Sets**				**Validation Sets**		
	Holstein		**Jersey**		**Holstein Bulls**	**Jersey Bulls**	**Australian Red Bulls**
	Bulls	**Cows**	**Bulls**	**Cows**			
MilkY/ProtY/ Fat%	3,049	8,478	770	3,917	262	105	114
Fertility	2,806	7,838	716	3,830	396	81	114

To compare the computational time required by the different genomic prediction methods, we also used three reference sets with increasing different numbers of animals; Ref 1_ CATTLE had 3,049 Holstein bulls; Ref 2_CATTLE had 11,527 Holstein bulls and cows, while Ref 3_CATTLE was the complete reference data set with 16,214 animals.

For the EM module, estimates of three variance components (σ_e^2, σ_v^2 , σ_g^2) were required as the input. We ran Asreml4.0 (Gilmour *et al.* 2002) (which was implemented with GBLUP methods) on these data sets to estimate these variance parameters and the results were listed in Table 5.3.

Table 5.3. Three input variance parameters related to the reference data sets.

Reference Set	**Traits**	$\boldsymbol{\sigma_e^2}$	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_v^2}$
Holstein and Jersey bulls & cows	Milk yield	133284.0	108619.0	34925.6
	Protein yield	132.579	68.6635	29.1662
	Fat%	0.0180012	0.0575729	0.0127094
	Fertility	3283.80	31.6187	0.000332297

The variances including error variance ($\boldsymbol{\sigma_e^2}$), genetic variance ($\boldsymbol{\sigma_g^2}$), and polygenic variance ($\boldsymbol{\sigma_v^2}$) were estimated by ASReml 4.

The correlation between GEBV and DTD in the validation sets was used as a proxy for accuracy of prediction. The regression of DTD on GEBV in the validation sets was used to investigate if any of the methods resulted in biased predictions.

Case/Control human disease trait data. For predicting human disease risk, seven disease traits of the Welcome Trust Case Control Consortium (WTCCC) genomic data (The Wellcome Trust Case Control Consortium 2007) including bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), Hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D) were used. Following the steps of strict QC on SNP data (Lee *et al.* ; Speed & Balding 2014; Moser *et al.* 2015) with Plink (Purcell *et al.* 2007), we had seven combined case/control data sets (one for each trait) with different number of markers and records listed in Table 5.4. The input parameters of seven datasets for HyB_BR including error variance and genetic variance were calculated by GCTA (Yang *et al.* 2011), given in Table 5.4. To assess accuracy of genomic predictions, for each disease, we randomly generated 20 splits of the data with 80% of individuals for the reference set and 20% for the validation set. To assess the prediction ability, the area under the ROC curve (AUC) (Wray *et al.* 2010) was calculated.

Table 5.4. The size and genetic architecture of seven combined control/case data sets.

Diseases	Number of Records	Number of Markers	σ_e^2	σ_g^2	h^2
BD	4,722	292,496	0.070509	0.17156	0.71
CAD	4,864	296,610	0.149782	0.09189	0.38
CD	4,577	301,579	0.073900	0.16056	0.69
HT	4,890	294,404	0.113621	0.12816	0.53
RA	4,704	295,890	0.070900	0.07120	0.50
T1D	4,824	296,228	0.064739	0.12567	0.66
T2D	4,722	294,641	0.099866	0.14497	0.59

The error variance (σ_e^2) and genetic variance (σ_g^2) were estimated by GCTA; the heritability (h^2) was estimated by the equation $h^2 = \sigma_g^2 / (\sigma_e^2 + \sigma_g^2)$.

5.5 Results

Compute time comparison of HyB_BR and BayesR

To compare computational efficiency, HyB_BR without the speed-up scheme (labelled as Hyb_BR_Orig), HyB_BR with the speed-up scheme and pure MCMC BayesR were implemented on three data sets with 632,003 markers but different numbers of records, varying from 3,049 in Ref 1_CATTLE, 11,527 in Ref 2_CATTLE, to 16,214 in Ref 3_CATTLE. As used by Kemper et al. (Kemper *et al.* 2015), pure MCMC BayesR required 40,000 iterations of complexity $O(mn)$ with parameters estimated from samples from the posterior distributions (m is the number of markers and n is the number of individuals). The first 20,000 iterations were removed as burn in. The MCMC module of HyB_BR used only 4,000 iterations and burn-in stage was replaced by the EM (400 iterations to convergence). 4,000 cycles for the MCMC module were used after comparing results with increasing number of iterations. The results showed that 4,000 were necessary to achieve maximum accuracy (Figure 5.2).

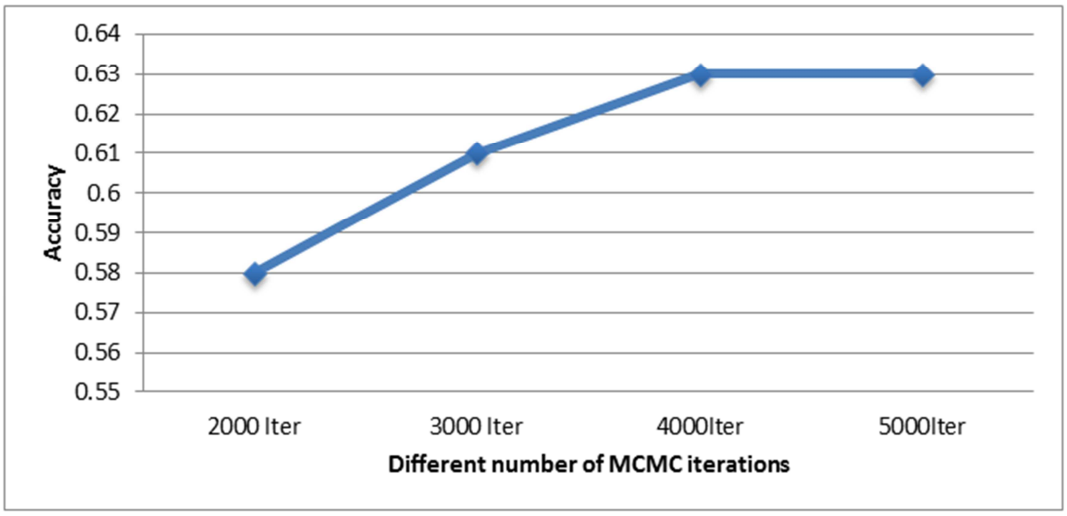


Figure 5.2. Accuracy of genomic prediction with an increasing number of MCMC iterations for BayesR.

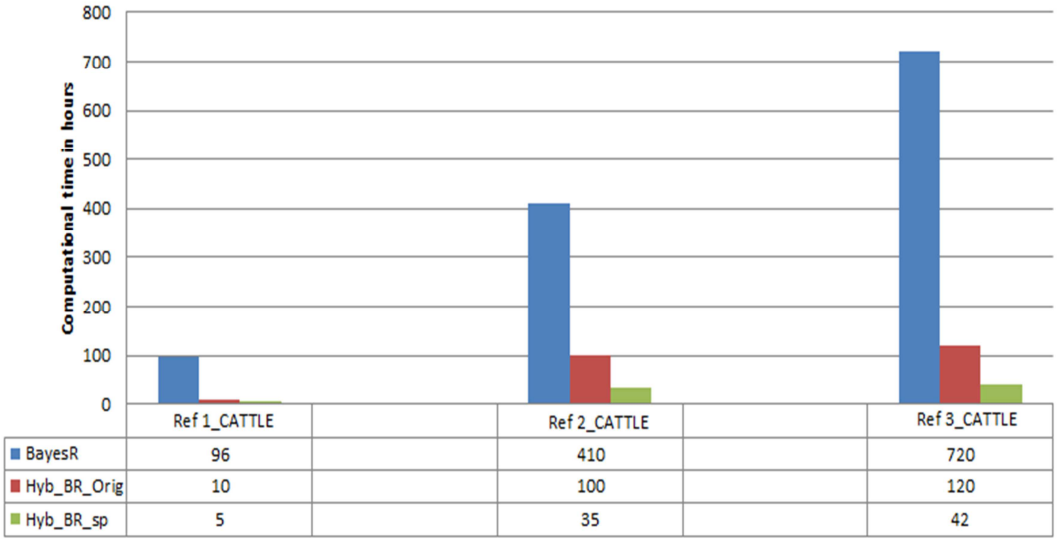


Figure 5.3. Computational time in hours required for BayesR, HyB_BR_Orig, and HyB_BR_sp on three reference sets (Ref 1_CATTLE, Ref 2_CATTLE, Ref 3_CATTLE).

The prediction accuracy was evaluated for milk yield with a reference set containing the Holstein and Jersey bulls&cows data. With the smallest data set (Ref 1_CATTLE), 5 hours compute time were required for HyB_BR compared with

96 hours for BayesR MCMC (Figure 5.3); 35 hours required by HyB_BR instead of 410 hours of BayesR for Ref 2_CATTLE; And in Ref 3_CATTLE, 42 hours for HyB_BR_sp but 720 hours for BayesR. These results suggested HyB_BR was at least 10 times faster than BayesR MCMC, with the speed advantage increasing as data sets became larger (17 times faster with the largest data set). The HyB_BR speed-up scheme reduced compute time by approximately 50%, compared to HyB_BR_Orig without the speed up scheme (Figure 5.3), with no reduction in the accuracy of genomic prediction (Table 5.5, Table 5.6, and Table 5.7).

These timings were recorded on a server with Intel E5-2680 2.7GHz processors and 384GB of 1333MHz RAM.

Table 5.5. The accuracy and bias of with-population prediction of GBLUP, BayesR(BR), emBayesR (EM), and HyB_BR (HB). +Poly means polygenic effects were included in the predictions; while -Poly means the predictions do not include polygenic effects into the model.

		Holstein reference to predict Holstein validation							
		Milk Yield		**Protein Yield**		**Fat%**		**Fertility**	
		Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	**+Poly**[a]	0.57	0.96	0.63	0.98	0.73	0.96	0.43	1.26
	-Poly[b]	0.56	0.86	0.59	0.87	0.71	1.15	0.42	1.27
BR	**+Poly**[a]	0.63	0.91	0.64	1.01	0.79	1.06	0.43	1.19
	-Poly[b]	0.61	1.00	0.63	1.06	0.77	1.13	0.41	1.19
EM	**+Poly**[a]	0.62	0.79	0.63	0.85	0.77	0.98	0.42	1.15
	-Poly[b]	0.62	0.92	0.62	0.94	0.74	1.06	0.41	1.15
HB	**+Poly**[a]	0.63	0.93	0.63	0.97	0.79	1.09	0.43	1.19
	-Poly[b]	0.63	1.03	0.62	1.06	0.76	1.17	0.42	1.19
		Jersey reference to predict Jersey validation							
		Milk Yield		**Protein Yield**		**Fat%**		**Fertility**	
		Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	**+Poly**[a]	0.59	0.93	0.65	0.91	0.54	0.71	0.15	1.05
	-Poly[b]	0.58	1.05	0.64	1.09	0.54	0.77	0.14	1.08
BR	**+Poly**[a]	0.64	0.94	0.68	0.93	0.71	0.87	0.15	1.02
	-Poly[b]	0.63	0.98	0.68	1.01	0.69	0.93	0.14	1.04
EM	**+Poly**[a]	0.64	0.87	0.68	0.92	0.69	0.75	0.15	1.09
	-Poly[b]	0.64	0.98	0.66	1.01	0.67	0.79	0.15	1.09
HB	**+Poly**[a]	0.64	0.97	0.68	0.90	0.71	0.89	0.15	1.02
	-Poly[b]	0.64	1.06	0.66	0.96	0.69	0.87	0.15	1.02

Table 5.6. The accuracy and bias of multi-population prediction of GBLUP, BayesR(BR), emBayesR (EM), and HyB_BR (HB). +Poly means polygenic effects were included in the predictions; while -Poly means the predictions do not include polygenic effects into the model.

		Holstein and Jersey reference to predict Holstein validation							
		Milk Yield		**Protein Yield**		**Fat%**		**Fertility**	
		Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	**+Poly**[a]	0.63	0.83	0.65	0.85	0.74	0.85	0.44	1.66
	-Poly[b]	0.62	0.90	0.57	0.88	0.72	0.90	0.42	1.66
BR	**+Poly**[a]	0.68	0.84	0.68	0.88	0.81	0.90	0.44	1.53
	-Poly[b]	0.67	0.91	0.67	1.03	0.79	0.98	0.42	1.53
EM	**+Poly**[a]	0.68	0.90	0.68	0.79	0.77	0.83	0.44	1.27
	-Poly[b]	0.65	0.91	0.66	0.85	0.75	0.87	0.44	1.27
HB	**+Poly**[a]	0.68	0.82	0.67	0.88	0.81	0.94	0.44	1.33
	-Poly[b]	0.67	0.89	0.67	0.95	0.80	1.08	0.44	1.33
		Holstein and Jersey reference to predict Jersey validation							
		Milk Yield		**Protein Yield**		**Fat%**		**Fertility**	
		Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	**+Poly**[a]	0.64	0.78	0.68	0.85	0.66	0.73	0.24	1.12
	-Poly[b]	0.64	0.90	0.69	1.02	0.64	0.80	0.24	1.12
BR	**+Poly**[a]	0.69	0.85	0.71	0.99	0.76	0.88	0.26	1.23
	-Poly[b]	0.68	0.95	0.71	1.09	0.74	0.94	0.25	1.24
EM	**+Poly**[a]	0.66	0.84	0.69	0.71	0.75	0.76	0.23	1.13
	-Poly[b]	0.63	0.86	0.68	0.73	0.70	0.82	0.23	1.13
HB	**+Poly**[a]	0.71	0.89	0.74	0.94	0.77	0.89	0.26	1.02
	-Poly[b]	0.69	0.98	0.73	1.02	0.73	0.97	0.26	1.02

Table 5.7. The accuracy and bias of across-breeds prediction of BayesR, GBLUP, emBayesR and HyB_BR.

	Across breeds prediction on Australian red bulls							
	Milk Yield		Protein Yield		Fat%		Fertility	
	Acc.	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
GBLUP	0.16	0.54	0.11	0.51	0.32	0.90	0.29	0.97
BR	0.22	0.60	0.12	0.49	0.45	0.92	0.27	1.03
EM	0.24	0.70	0.12	0.42	0.41	0.89	0.29	1.10
HB	0.23	0.74	0.17	0.49	0.50	0.90	0.30	0.98
	Across breeds prediction on Australian red cows							
	Milk Yield		**Protein Yield**		**Fat%**		**Fertility**	
GBLUP	0.15	0.71	0.08	0.13	0.50	1.19	0.08	0.79
BR	0.26	0.80	0.17	0.51	0.54	0.94	0.08	0.68
EM	0.24	0.79	0.16	0.53	0.51	0.89	0.08	0.74
HB	0.26	0.81	0.16	0.57	0.55	0.91	0.08	0.70

The accuracy and bias of within-breeds, multi-breeds and across-breeds prediction for four complex dairy traits

Genomic prediction with a single breed reference. For the within-breed prediction (that is, when a Holstein reference was used to estimate SNP effects used for calculating GEBV in a Holstein validation set, and likewise for Jersey) in Table 5.5, HyB_BR performed as well as BayesR for all traits, including fat%. Both BayesR and HyB_BR had a 1%~6% superiority of accuracy over GBLUP for Milk yield, Protein yield and Fat%, but had no advantage for fertility. Similarly, for the prediction of Jersey validation with Jersey reference, BayesR and HyB_BR had a consistent advantage over GBLUP for milk production traits, but not for fertility. Especially, for the trait Fat%, BayesR and HyB_BR gave very similar results, with a 17% increase in accuracy (0.79 vs 0.73 in Holstein and 0.71 vs 0.54 in Jersey) of genomic prediction over GBLUP, as well as an 5% increase in accuracy over emBayesR. HyB_BR and BayesR also gave regression coefficients closer to one than GBLUP for most traits.

Genomic prediction with a multi-breed reference. When predicting the Holstein or Jersey validation with the combined Holstein and Jersey reference, HyB_BR had the same accuracy as BayesR, Table 5.5. Compared with GBLUP, BayesR and HyB_BR gave consistently higher accuracy increase for the milk production traits, though this was not observed for fertility. And for the prediction of Jersey validation set, BayesR and HyB_BR improved accuracy for the milk production traits by 11% compared with GBLUP. The results showed that there were small but consistent accuracy improvements as a result of using the multi-breed reference (compare Table 5.5 and Table 5.6), consistent with the results of Kemper et al. (Kemper et al. 2015) and Hoze et al. (Hozé et al. 2014).

In addition, including polygenic effects (estimated using the pedigree) in the model could improve the prediction accuracy by 1~2%, at least for milk production traits, Table 5.5 and Table 5.6. However, for fertility the introduction of polygenic effects for all the prediction methods did not affect the accuracy at all.

Compared with GBLUP and emBayesR, BayesR and HyB_BR gave less biased predictions for milk production traits. However for fertility the regression values far from one indicate bias, from all methods – the low accuracy of fertility phenotypes, including in the validation set, likely contributes to this.

Genomic prediction across breeds. For predicting Australian Red validation bulls (an additional breed to those in the reference set), BayesR and HyB_BR gave higher accuracy than GBLUP for all traits (Table 5.7).

Across all the prediction results shown in Table 5.5, Table 5.6, and Table 5.7, emBayesR had a 2%~5% reduction in accuracy compared with BayesR and HyB_BR for fat%, while BayesR and HyB_BR gave almost identical accuracies in

all cases.

Inferred genetic architecture and QTL mapping for dairy production and fertility traits.

Bayes R described the genetic architecture of a trait by the posterior proportion of SNPs in each of the 4 different distributions. Table 5.8 reported the estimated proportion in each of four distributions from BayesR, emBayesR, and HyB_BR. The number of SNPs falling into the distribution with the largest variance was similar for all three methods. Compared with BayesR, HyB_BR tended to estimate more SNPs (up to 5%) in the distribution with variance $0.001 * \sigma_g^2$, and $0.0001 * \sigma_g^2$. In contrast to HyB_BR, emBayesR tended to estimate that a higher proportion of SNPs had no effect than did BayesR. This might explain the lower accuracy it achieved.

Table 5.8. The number of SNPs in each of four distributions.

Traits	The proportion (Pr)	BayesR	emBayesR	HyB_BR
Milk Yield	A. $0.01 * \sigma_g^2$	8	6	8
	B. $0.001 * \sigma_g^2$	47	17	327
	C. $0.0001 * \sigma_g^2$	3,886	1,523	4039
	D. 0	628,062	630,457	627,629
Protein Yield	A. $0.01 * \sigma_g^2$	5	4	6
	B. $0.001 * \sigma_g^2$	32	37	297
	C. $0.0001 * \sigma_g^2$	4,431	1,842	6,604
	D. 0	627,535	630,120	625,096
Fat%	A. $0.01 * \sigma_g^2$	23	19	20
	B. $0.001 * \sigma_g^2$	46	206	119
	C. $0.0001 * \sigma_g^2$	2882	1,206	1,852
	D. 0	629,052	630,572	630,012
Fertility	A. $0.01 * \sigma_g^2$	10	8	12
	B. $0.001 * \sigma_g^2$	147	114	202
	C. $0.0001 * \sigma_g^2$	3,949	8,572	7,597
	D. 0	627,897	623,309	624,192

QTL mapping for dairy production and fertility traits.

Both BayesR and HyB_BR estimated the posterior probability that every SNP had a non-zero effect on the trait. This was useful for QTL mapping – SNP with very high posterior probabilities of having a non-zero effect should be strongly associated with causal mutations (e.g. Moser et al. (Moser *et al.* 2015), Kemper et al. (Kemper *et al.* 2015)). Then, QTL mapping from BayesR and HyB_BR could be conducted by plotting the posterior probability of each SNPs having a non-zero

effect on the trait by genome position, and then comparing the genome location of the effects with a high posterior probability of being in the largest distribution for each method.

The estimated posterior possibilities of all the SNPs (y axis) related to four different traits were plotted according to the positions (base pairs) of SNPs on the whole genome (x axis) in Figure 5.4, Figure 5.5, Figure 5.6, and Figure 5.7. The top SNPs with high posterior possibilities were picked up according to the number of SNPs in the variance $0.01 * \sigma_g^2$ (the total number of markers * Pr[4]). Table 5.9 listed all the top SNPs in the variance related to the previously reported genes with a effect on milk production including CSF2RB (Chamberlain *et al.* 2015), GC (Sanders *et al.* 2006), GHR/CCL28 (Blott *et al.* 2003), PAEP (Ng-Kwai-Hang), MGST1 (Raven *et al.* 2015), and DGAT1 (Grisart *et al.* 2002). Both BayesR and HyB_BR identified all of these regions, which demonstrated that HyB_BR could perform QTL mapping with similar precision to BayesR. For example, HyB_BR could detect the DGAT1 as well as BayesR shown in Figure 5.6 (Fat%).

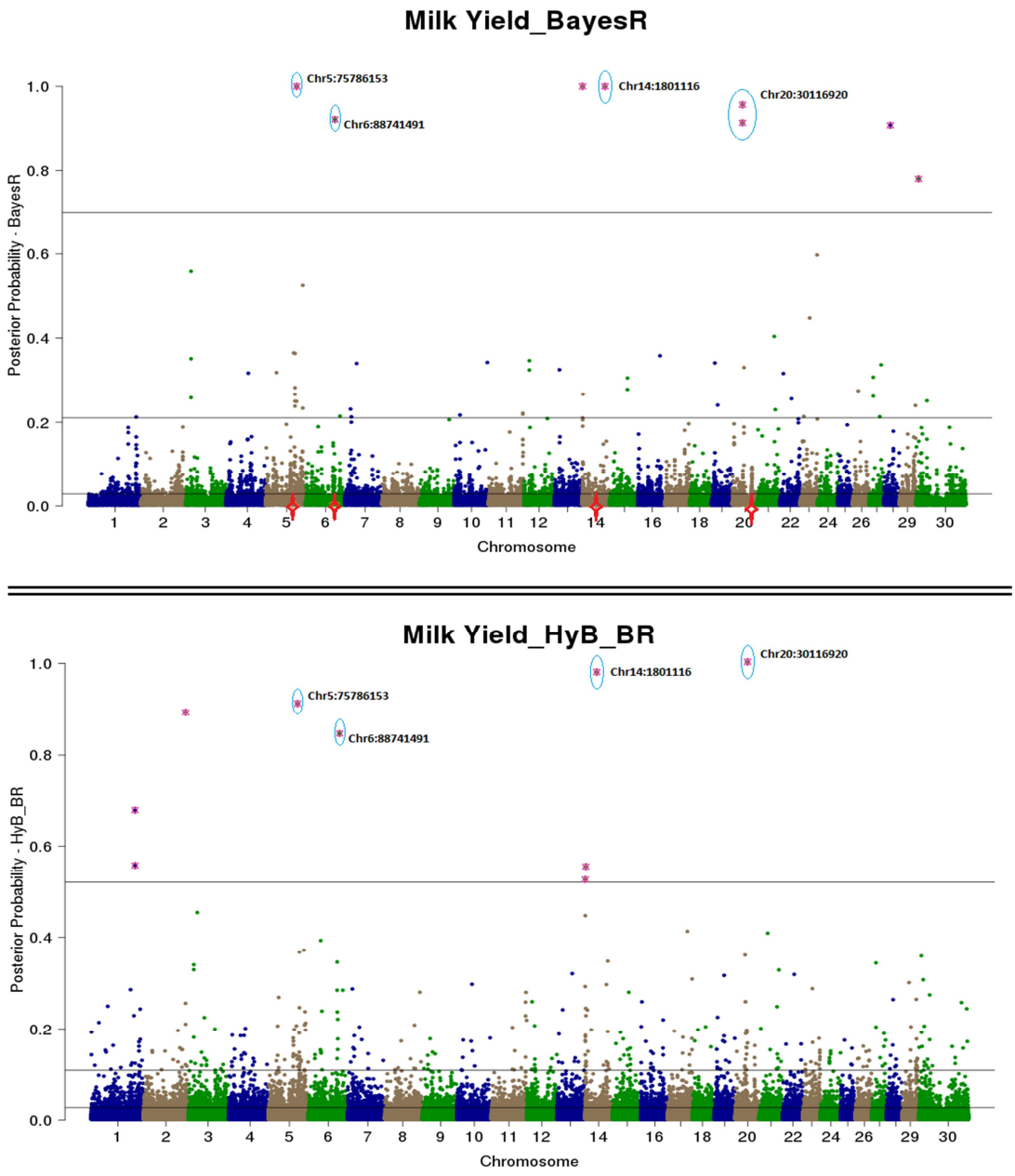


Figure 5.4. Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to milk yield. The posterior possibility is calculated based on the sum of the posterior possibilities $P(i,k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i,k)$. The blue circle is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes.

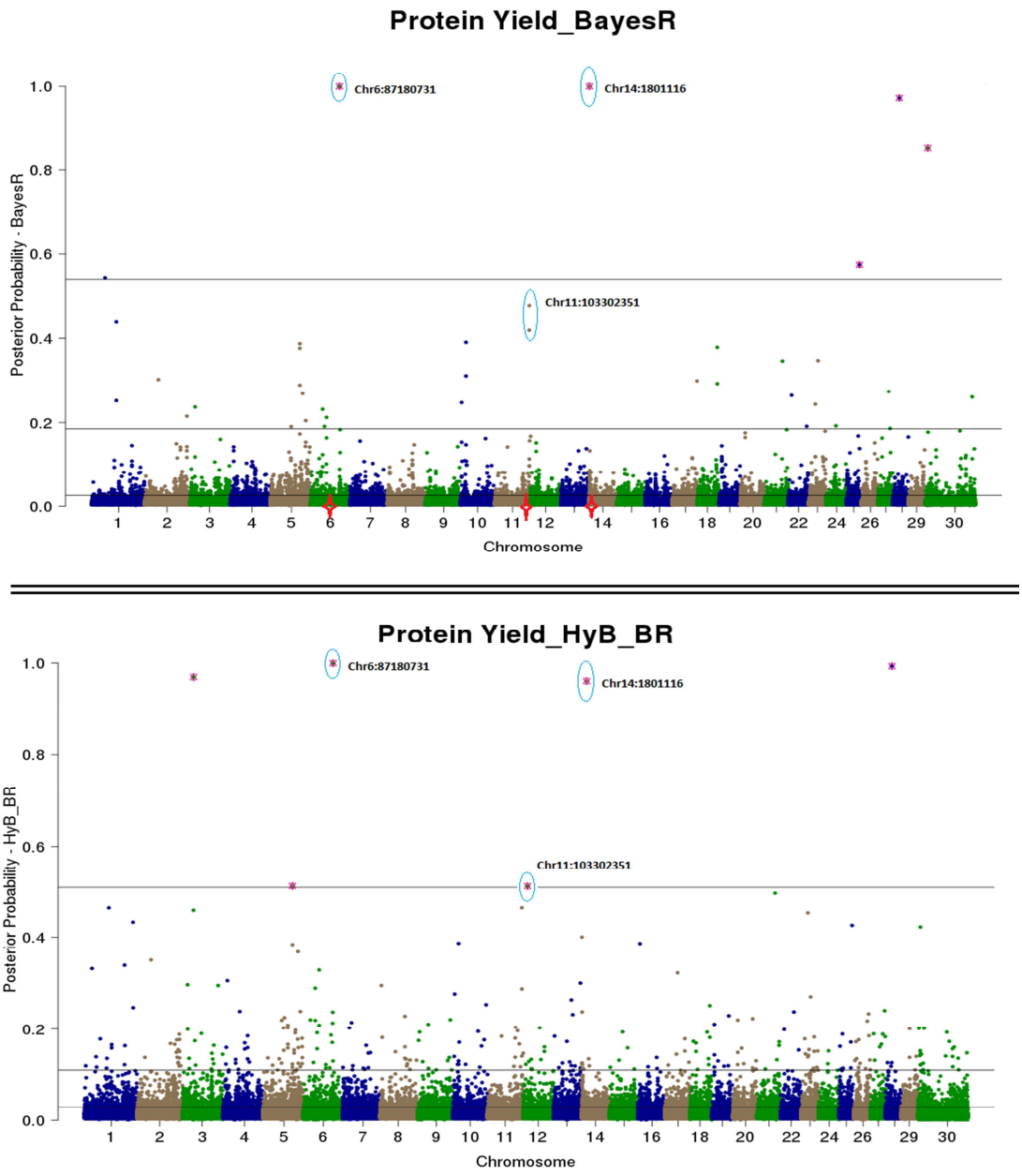


Figure 5.5. Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to protein yield. The posterior possibility is calculated based on the sum of the posterior possibilities $P(i,k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i,k)$. The blue circle is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes.

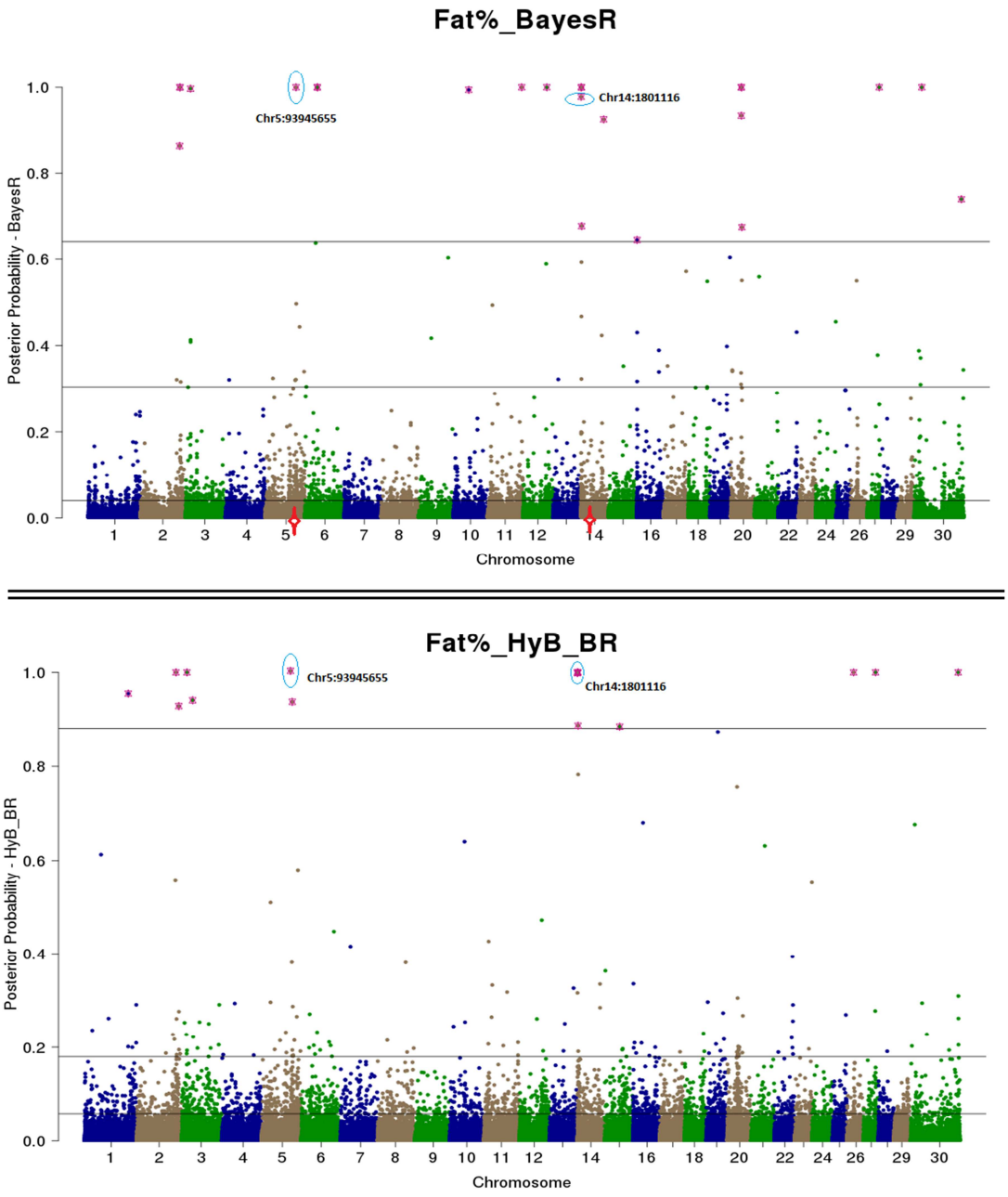


Figure 5.6. Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to Fat percent (Fat%). The posterior possibility is calculated based on the sum of the posterior possibilities $P(i,k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i,k)$. The blue circle is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes.

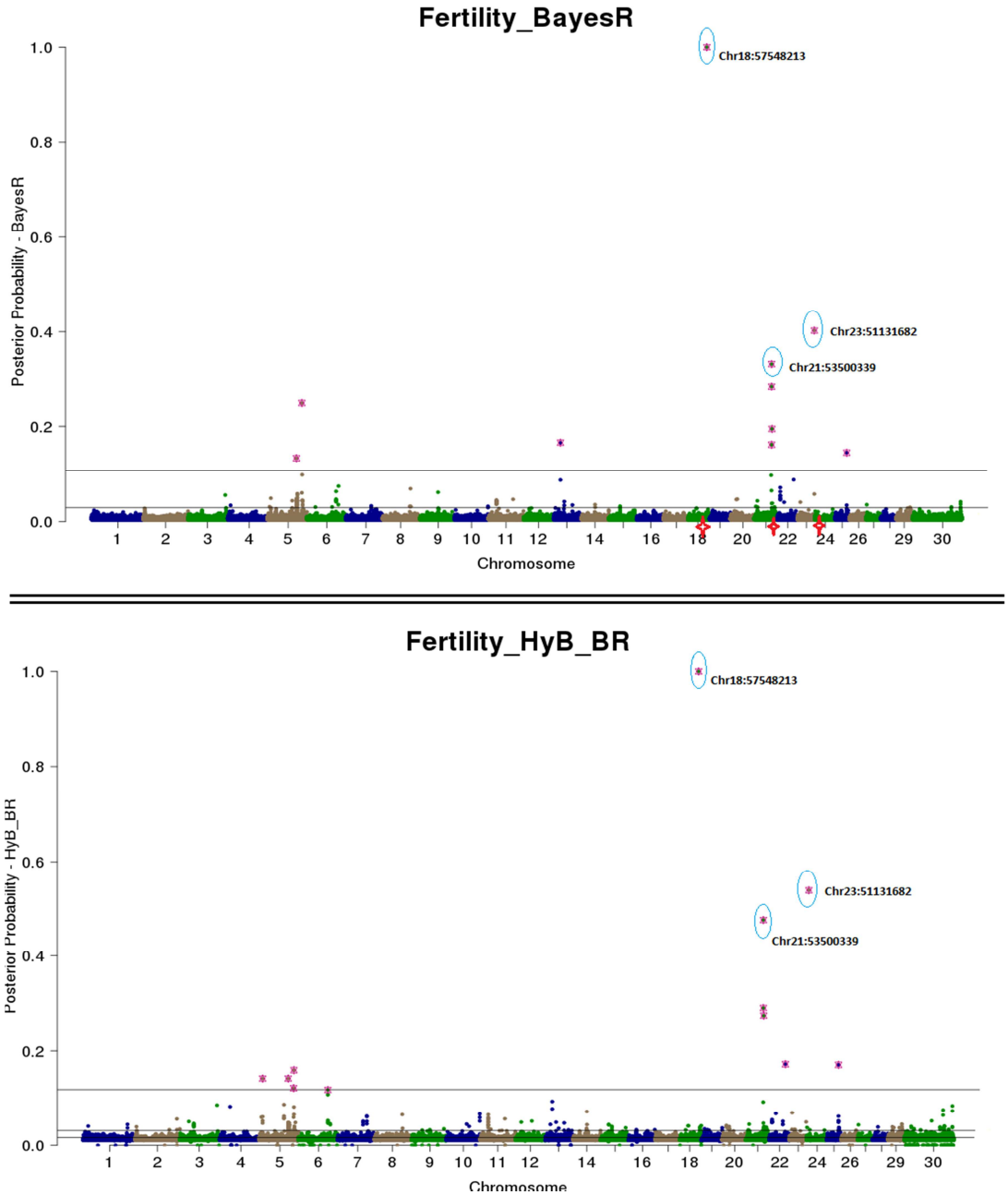


Figure 5.7. Mapping all the SNPs' posterior possibilities estimated from BayesR and HyB_BR across the whole chromosome related to fertility. The posterior possibility is calculated based on the sum of the posterior possibilities $P(i,k)$ of each SNP with non-zero variances written as $\sum_{k=2}^{4} P(i,k)$. The blue circle is the SNPs (picked up based on the high posterior possibility following in the distribution with largest variances) with location information mapped to known genes.

Table 5.9. The list of identified causal mutations by both BayesR and HyB_BR.

Traits	**loci**	**Information (known genes)**	**Range (bp) [Startpoints~End points]**
Milk yield	Chr5:75786153	CSF2RB impacting milk yield (Chamberlain *et al.* 2015).	[75724620~75745819]
	Chr6:88741491	GC, encoding the vitamin D binding protein, positively impacting the milk yield (Sanders *et al.* 2006).	[88695940~88749180]
	Chr20:30116920	In association with CCL28/GHR impacting milk production (Blott *et al.* 2003).	[31890736~32199996]
Protein yield	Chr6:87180731	CSN1S1 positively impacting protein yield (Sanders *et al.* 2006).	[87141556~ 87159096]
	Chr11:103302351	PAEP impacting protein yield (Wang *et al.* 2012b).	[103301664~103306381]
Fat%	Chr5:93945655	MGST1 for Fat percent (Raven *et al.* 2015).	[93926791~3950162]
Fertility	Chr18:57548213	-In association with the gene CEACAM18, Detected by (Pryce *et al.* 2010), (Cole *et al.* 2011).	~57MBP
	Chr21:53500339	- Control the percentage of unassisted births in first calf heifers (McClure *et al.* 2010).	~53MBP
	Chr23:51131682	In the linkage with the known gene GMDS (Wickramasinghe *et al.* 2011).	~51MBP
All the traits	Chr14:1801116	DGAT1 impacting Fat percent (Grisart *et al.* 2002).	[1795351~1804562]

The application of HyB_BR to predict the risk of Human disease traits and infer genetic architecture for these traits

In the human data, cross validation was used to estimate the accuracy of HyB_BR. As there were 20 replicates of 20/80 split (validation/reference), we evaluated the mean of the AUC for each disease shown in Table 5.10. Analysis methods compared were GBLUP implemented in GCTA (Yang *et al.* 2011), BayesR from Moser et al. (Moser *et al.* 2015), and HyB_BR. The standard deviations of the accuracy (across the 20 replicates) were also listed in the parenthesis of Table 5.10. HyB_BR and BayesR performed equally well across all seven traits, with the same prediction accuracy for each trait. For the diseases of CD, RA, and T1D, BayesR and HyB_BR had significantly higher accuracy than GBLUP. Especially for T1D, BayesR and HyB_BR could have up to 12% accuracy increase compared with GBLUP. However, for other traits including BD, CAD, HT, and T2D, BayesR and HyB_BR did not show any superiority over GBLUP. The underlying architecture of these traits might explain this, as suggested by Moser et al. (Moser *et al.* 2015). In detail, for CD, RA and T1D, there were known mutations of moderate to large effect, and the mixture assumptions of BayesR and HyB_BR could take advantage of this. However, for four other diseases including BD, CAD, HT, and T2D, there were no known mutations of moderate to large effect, and this was reflected in the genetic architecture for these diseases inferred by HyB_BR.

Table 5.10. The prediction performance evaluated by the Area under curve (AUC) of GBLUP, BayesR and HyB_BR on seven diseases.

Diseases	GBLUP		BayesR		HyB_BR	
	AUC	h^2	AUC	h^2	AUC	h^2
BD	0.63(0.0135)	0.71	0.63(0.0131)	0.63	0.64(0.0174)	0.63
CAD	0.58(0.0116)	0.38	0.59(0.0118)	0.38	0.58(0.0131)	0.38
CD	0.60(0.0134)	0.69	0.65(0.0159)	0.61	0.65(0.0158)	0.61
HT	0.58(0.0125)	0.53	0.58(0.0131)	0.52	0.58(0.0140)	0.51
RA	0.58(0.0109)	0.50	0.70(0.0104)	0.45	0.70(0.0107)	0.45
T1D	0.64(0.0133)	0.66	0.86(0.0099)	0.63	0.86(0.0102)	0.63
T2D	0.59(0.0139)	0.59	0.60(0.0117)	0.52	0.60(0.0122)	0.52

The heritability (h^2) was estimated by the equation $h^2 = {\sigma_g^2}/{(\sigma_e^2 + \sigma_g^2)}$; σ_e^2 was derived separately by three methods; fixed genetic variance of σ_g^2 for BayesR and HyB_BR was obtained from GCTA.

The genetic architecture of human disease traits. The inferred genetic architecture was different for each of the seven diseases (Table 5.11). For example, the genetic architecture of BD was controlled by many SNPs (9,077 for HyB_BR; 9,611 for BayesR) with small effects (the variance $0.0001\sigma_g^2$), but just 3 SNPs with large effects (the variance $0.01\sigma_g^2$). These numbers demonstrated the polygenic architecture of BD. On the contrary, for T1D, there was relatively smaller number of SNPs (3,544 for HyB_BR; 2,750 for BayesR) with small effects but many more SNPs (almost 200) with large effects. The proportion numbers from Figure 5.8 also demonstrated this (in accordance with the results from Moser et al. (Moser *et al.* 2015)). Large proportion of SNPs with small effects (the variance $0.0001\sigma_g^2$) controlled the polygenic architecture of the diseases BD (98.76% for HyB_BR; 99.55% for BayesR), CAD (97.31% for HyB_BR; 96.8% for BayesR), HT (96.96% for HyB_BR; 98.09% for BayesR), and T2D (95.14% for HyB_BR; 97.79% for BayesR). For these diseases, the mixture model of BayesR and HyB_BR did not have much advantage. However, relatively larger proportions of SNPs with moderate effects (the variance $0.001\sigma_g^2$) existed for the traits RA

(0.77% for HyB_BR; 0.93% for BayesR) and T1D (5.02% for HyB_BR; 5.54% for BayesR). For these two traits controlled by major genes, BayesR and HyB_BR gave substantially greater accuracy than GBLUP, which explained the results for accuracy of prediction (Table 5.10).

Table 5.11. The number of SNPs in each proportion of four distributions estimated by BayesR, and HyB_BR on seven human diseases.

Diseases	**BayesR**				**HyB_BR**			
	Pr[1]	**Pr[2]**	**Pr[3]**	**Pr[4]**	**Pr[1]**	**Pr[2]**	**Pr[3]**	**Pr[4]**
BD	282,843	9,611	39	3	283,306	9,077	110	3
CAD	289,491	6,892	214	13	289,203	7,211	183	13
CD	294,423	6,878	269	9	294,463	6,576	331	9
HT	286,152	8,094	150	8	286,160	7,993	243	8
RA	291,401	4,172	275	42	290,420	5,025	403	42
T1D	293,366	2,607	54	200	292,523	3,396	104	207
T2D	286,489	7,972	173	7	288,365	5,971	298	7

Compared with BayesR, HyB_BR detected the same number of SNPs with moderate variance (the variance $0.01 * \sigma_g^2$) but appeared to systematically detect more SNPs in the proportion of small variance (the variance $0.0001 * \sigma_g^2$), similar to the results observed for the comparison of BayesR and HyB_BR in in dairy cattle data (Table 5.8).

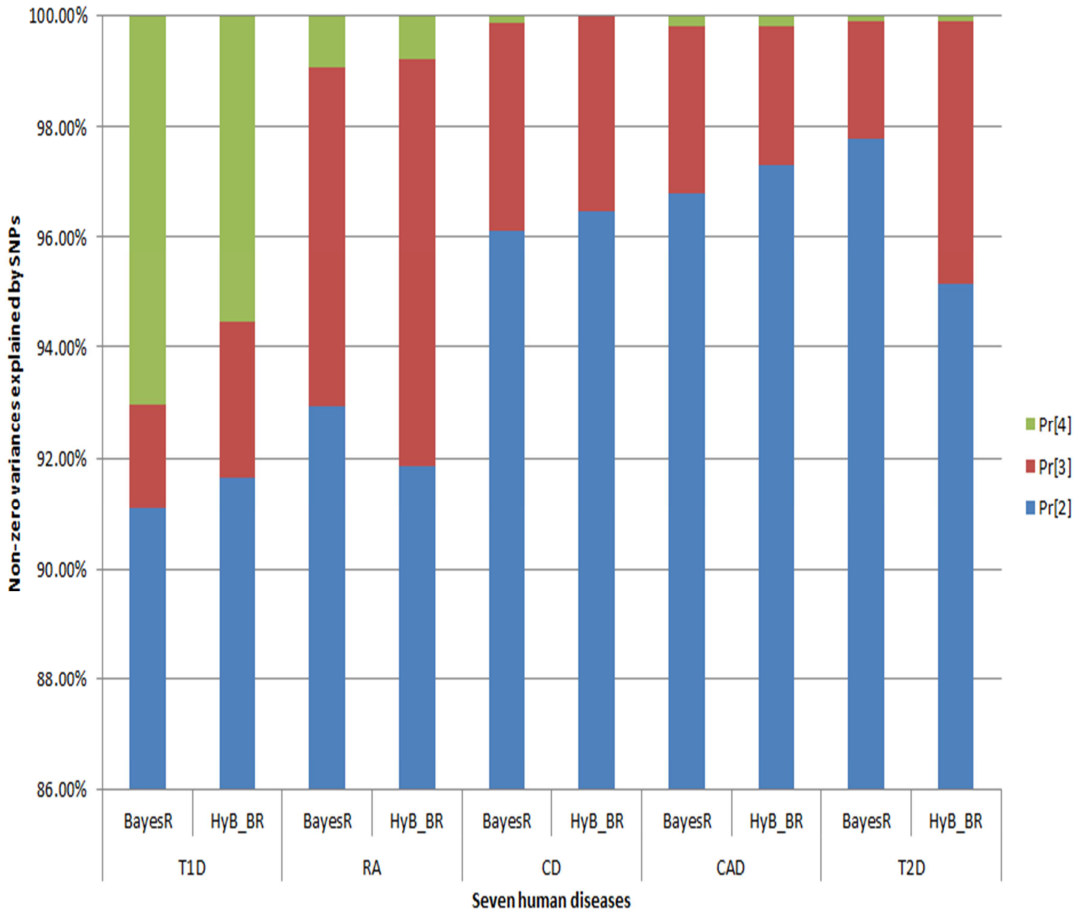


Figure 5.8. The inferred genetic architecture of seven human diseases from BayesR and HyB_BR.

The blue bar is the proportion of SNPs in Pr[2] (with the variance $0.0001 * \sigma_g^2$), which is estimated by the number of SNP in Pr[2] divided by the total number of SNPs with nonzero variance. The red bar is the proportion of SNPs with the variance $0.001 * \sigma_g^2$, estimated by the number of SNP in Pr[3] divided by the total number of SNPs with nonzero variance. The green bar is the proportion of SNPs with the variances $0.01 * \sigma_g^2$, estimated by the number of SNPs in Pr[2] divided by the total number of SNPs with non-zero variance.

5.6 Discussion

We have presented a novel and computationally efficient algorithm termed HyB_BR for simultaneous genomic prediction and QTL mapping. A pure EM

algorithm was less accurate for some traits, while pure MCMC requires very long computation times. Therefore, HyB_BR implements the EM algorithm followed by a limited number of MCMC iterations. In this way, the algorithm takes advantage of the features of an EM algorithm (rapid convergence) and the higher accuracy from MCMC implementations in a hybrid scheme. Our accuracies of genomic prediction for complex traits in human and cattle from HyB_BR are almost identical to those from the full MCMC implementation of the Bayesian mixture model, with a 10 fold or greater reduction in computing time required.

For the pure MCMC algorithm, the burn-in stage can account for up to 50% of the total running time. One of the key advantages of HyB_BR is that the EM module effectively replaces the burn-in cycles that are usually required for MCMC. Based on the starting point from EM (with very limited number of iterations; less than 500 iterations), the running time of HyB_BR can be much reduced.

The pure EM algorithm, EmBayesR (Wang *et al.* 2015) has been demonstrated to be much faster than BayesR, but had lower accuracy for some traits, particularly those with mutations of moderate to large effect. For example, when implemented on the trait fat% in dairy cattle, emBayesR had a decreased accuracy of 5%~7% compared to BayesR. One possible explanation is that emBayesR shrinks SNP effects too much (shown in Table 5.8). This could be because the PEV that is used to account for the error of the effects of all the other SNPs while estimating the effect of the current SNP is only an approximation. The introduction of PEV correction is based on one observation: previous fast algorithm studies (especially Iterative conditional expectation algorithms) assumed the effect of the other SNP were estimated perfectly while estimating the effect of the current SNP, leading to poor performance (Wang *et al.* 2015). Therefore, EmBayesR and the EM part of HyB_BR allow for the errors in the effect of other SNPs and other location parameters by using the PEV. The calculation of the PEV from GBLUP is carried

out before the iterations to estimate the effects of each SNP. And since the normal priors from GBLUP model do not allow for SNPs of moderate to large effects, such PEV calculation is an approximation and this may be one reason for loss of accuracy in the EM. To deal with this, HyB_BR further implements a small number of MCMC iterations to improve the outcome of pure EM steps.

HyB_BR has three advantages. First, as the size of genomic data increases, the computational efficiency of HyB_BR without burn-in stage (a small number of $O(mn)$ iterations), is greater than BayesR by full MCMC. And when implemented with the speed-up scheme described in the methods, computational time can be reduced even further, by sampling a reduced set of SNPs in the MCMC module, apparently with no loss of accuracy (but critically the information from the SNPs that are not sampled remains in the posterior proportions of SNPs in each distribution). Second, the prediction accuracy of HyB_BR is comparable to BayesR in all cases including dairy cattle and human disease prediction shown in Table 5.5, Table 5.6, Table 5.7 and Table 5.10. Third, HyB_BR, like BayesR, is flexible with respect to the genetic architecture of complex traits. As shown in Table 5.5, Table 5.6, and Table 5.7, HyB_BR performs well on four different complex traits, with architecture ranging from highly polygenic architecture to genetic architecture controlled by major genes. In addition to the prediction on the continuous quantitative traits of dairy cattle, the investigation on the risk prediction of seven case/control human diseases with binary 0/1 phenotypes shows HyB_BR and BayesR perform on this type of data, Table 5.10. Finally, the posterior probabilities of SNP having a nonzero effect from HyB_BR can be used for QTL mapping, Figure 5.6.

Implementing genomic prediction methods with whole genome sequence data may improve the prediction accuracy and accelerate the discovery of causal variants. However, for this to occur, more computationally efficient genomic

prediction algorithms are required. Compared with BayesR, the predicted time of HyB_BR on different number of markers with the same reference phenotypes is listed in Table 5.12. The time is estimated linearly on the number of markers and individuals. When the number of markers reaches 30 million (the number of variants discovered in the 1000 bull genomes project, Daetwyler et al. (Daetwyler *et al.* 2014)), the running time of BayesR is around 34,170 hours, which is impractical. On the contrary, on the same data with 30 million of variants, HyB_BR is predicted to require 2,010 hours. It may be possible to reduce this further by optimising the code even more. Therefore, as the size of genomic data increases, HyB_BR will remain feasible well beyond the point where the use of BayesR is impractical.

Table 5.12. The predicted computational time (in hours) of HyB_BR and BayesR on high-density data with different number of variants and the same number of individuals (16,214).

	Different number of markers			
	800K SNP panel	**1 million**	**2 million**	**30 millions**
BayesR	720h	1,139h	2,278h	34,170h
HyB_BR	42h	67h	134h	2,010h

While HyB_BR performs well with computational efficiency and robust prediction accuracy, there are at least still two strategies that could be used to further improve efficiency. There is one key part of EM module that consumes running time and memory: the calculation of $tr(\mathbf{E^{-1}Z_i Z_i' E^{-1}PEV})$ for each SNP in front of EM iterations. In detail, the calculation of $tr(\mathbf{E^{-1}Z_i Z_i' E^{-1}PEV})$ requires the time complexity of $(\frac{1}{2}mn^2)$, which accounts for almost 2/3 of the total computational time even though it happens in front of EM iterations. Therefore, a future task is to implement a multi-threaded version to improve speed. The threshold of limiting the number of SNPs to be updated requires further study. Currently we define the threshold as $\mathbf{T}$: if$(P(i,1) > 0.9)$, which is applicable for the current data. However,

it's uncertain whether or not such a threshold is suitable for other types of data.

HyB_BR has some features in common with other mixture methods such as BSLMM (Zhou *et al.* 2013b), and BOLT-LMM (Loh *et al.* 2015). All of these methods declared the merit of computational efficiency with time complexity $O(mn)$ but under different mixture models. In detail, BSLMM assumed a large proportion of SNPs with small effects (under BLUP models), while others had large effects (under Bayesian sparse regression models; the mixture of two normal priors). Due to limited number of SNPs implemented for MCMC sampling (large proportion of SNPs were under GBLUP models), BSLMM could be computationally efficient. However, compared with the mixture of four normal distributions by BayesR, which provided great flexibility with respect genetic architecture, the flexibility of BSLMM with respect to different genetic architectures required further investigation. Another algorithm was BOLT-LMM, which had been developed mainly for the association studies. BOLT-LMM incorporated Bayesian mixture models to improve the power of GWAS with appealing outcomes. Instead of MCMC sampling, BOLT-LMM implemented iterative conditional expectation (ICE) algorithm on a mixture of two normal distributions to improve the computational speed with the approximated computational complexity $O(mn)$. There could be three limitations with this method: 1) ICE algorithms did not account for the PEVs from all other SNP effects during the estimation of current SNP effect. On practical data sets, ICE could lead to the loss of prediction accuracy. BOLT-LMM introduced LD score regression technique to calibrate the prediction errors. However, since the calibrating factor was constant across all the SNPs (the prediction error variance regarding each SNP differed according to our equation $tr(\mathbf{E^{-1}Z_i Z_i' E^{-1} PEV})$), such calibration scheme seem not to be effective to solve the problem. 2) The leave-one-chromosome-out scheme implemented in BOLT-LMM might perform well for GWAS but not be suitable for simultaneous genomic prediction. 3) BOLT-LMM treated each SNP effect as a fixed effect for

the association statistics. This, combined with the stringent significance threshold for multiple testing, leaded to the over-estimation for SNP effects. Another efficient method for genomic prediction termed MultiBLUP (Speed & Balding 2014) introduced SNPs clusters into BLUP models according to its adaptive algorithm. For each SNP class, the linear combination models (using genomic relationship matrix) similar to GBLUP were implemented. MultiBLUP has been demonstrated to be computationally efficient with robust prediction accuracy in the human data sets. However, when moved to dairy cattle genomic data sets, there is long Linkage disequilibrium (LD) between markers, which might be easily broken up by multiBLUP models.

5.7 Conclusion

In summary, HyB_BR is a computationally efficient method for simultaneous genomic prediction, QTL mapping and inference of genetic architecture. The hybrid scheme of MCMC and EM decreases computational time by a factor of at least 10 fold with no reduction in prediction accuracy. The HyB_BR algorithm makes simultaneous genomic prediction, QTL mapping and inference of genetic architecture feasible in extremely large genomic data sets including whole genome sequence data.

5.8 Supporting information

All the supporting files were located in Appendix III (Chapter 10) as follows:

File S1 - PEV calculation from GBLUP.

File S2 - Calculation of $\mathrm{P}(\mathrm{i},\mathrm{k})$.

5.9 Acknowledgements

The authors acknowledge the support from Dairy Futures CRC project.

Chapter 6 Application of Hybrid to Sequence data for genomic prediction and QTL mapping

6.1 Chapter preface

Justification

While using whole genome sequence data was attractive for genomic prediction of complex traits, the computational burden that would be imposed, particularly for non-linear Bayesian methods, make this currently infeasible. In this paper we implemented the HyB_BR methods on a large subset of whole genome sequence data from dairy cattle to assess feasibility of genomic predictions from sequence. The performance of HyB_BR was evaluated in terms of computational efficiency and genomic prediction accuracy. The computation advantage of HyB_BR over GBLUP and BayesR showed that HyB_BR was 10 fold faster than BayesR and 5 fold faster than GBLUP. In addition, the accuracy of HyB_BR on a range of traits with different genetic architectures was investigated. The results showed similar accuracy of HyB_BR and BayesR. A further advantage of the method was that a similar precision of QTL mapping to BayesR was demonstrated.

Publication status:

In preparation for submission

Statement of contributions of joint authorship

Tingting Wang (Candidate): implemented the hybrid algorithm on the whole genome sequence data for multi-breed and across-breed prediction. Then, the author drafted the manuscript.

Yi-Ping Phoebe Chen (Principle Supervisor): supervised on the paper.

Iona Macleod (Collaborator): provided help to implement BayesR on the whole genome sequence data.

Michael E Goddard (Collaborator): contributed the idea about the speed-up scheme.

Ben J. Hayes (Co-Supervisor): supervised this project, gave important instructions on organizing and revising the manuscript.

6.2 Abstract

Using whole genome sequence data for genomic predictions can improve the prediction accuracy, compared with what is possible from high-density SNP arrays, and can lead to identification of variants affecting complex traits. The most accurate genomic predictions for some traits are achieved with non-linear Bayesian methods. However, as the number of variants and the size of the reference population increases, the computational time required to implement these Bayesian methods (typically with Monte Carlo Markov Chain sampling) becomes unfeasibly long. Here, we apply a new method, HyB_BR (for Hybrid BayesR), which implements a mixture of normal model and hybridizes an Expectation-Maximization (EM) algorithm followed by Markov Chain Monte Carlo (MCMC) sampling, to genomic prediction in a large subset of whole genome sequence data in a dairy cattle reference population. The imputed whole genome sequence data includes 994,019 variant genotypes in 16,214 bulls and cows from the Holstein and Jersey breeds. Traits include Fat yield, Milk volume, Protein kg, Fat% and Protein% in milk, fertility and heat tolerance. HyB_BR achieves

genomic prediction accuracies as high as a full MCMC implementation of BayesR, both for predicting a validation set of Holstein and Jersey bulls (multi-breed prediction) and a validation set Australian red bulls (across-breed prediction).. The computation time on the data sets shows that HyB_BR (48 hours) can reduce compute time by tem fold compared with the MCMC implementation of BayesR (594 hours). We also demonstrate that in some cases HyB_BR can identify similar mutations to BayesR in the sequence data, including mutations in or close to the genes DGAT1, FASN, CSN2, CSN3, and CEACAM18, affecting milk production or fertility traits. For heat tolerance, both HyB_BR and BayesR find nine potential causative mutations not detected by previous studies. The results demonstrate that HyB_BR is a feasible method for simultaneous genomic predictions and QTL mapping with whole genome sequence in large reference populations.

6.3 Introduction

Whole genome sequence data is available for an increasing number of species, and in some cases, enough individuals have been sequenced to serve as a reference panel for imputation, for the thousands of individuals that have been genotyped with SNP arrays, to whole genome sequence variant genotypes. A good example of such a reference set is the 1000 bull genomes project which includes 234 bulls with whole-genome sequencing data and 28.3 million genotyped sequence variants (MacLeod *et al.* 2016). Compared with dense SNP arrays, the advantage of using whole genome sequence data might include more accurate genomic predictions, better persistence of accuracy of genomic predictions across generations, more accurate genomics predictions across breeds (Clark *et al.* 2011; Druet *et al.* 2014; MacLeod *et al.* 2014c; MacLeod *et al.* 2016), and more precise QTL mapping (MacLeod *et al.* 2016), all as a result of including the causal mutation genotypes in the data set.

As the resulting data sets will be extremely large (large numbers of individuals

with millions of imputed genotypes), the algorithms used to derive genomic predictions must be extremely computationally efficient. Ideally, they should also implement a non-linear model at the level of the SNP effects, including the possibility of excluding some SNP from the model, as such models have been demonstrated to give higher accuracies of genomic predictions for some traits with high-density genotype data (Kemper *et al.* 2015; MacLeod *et al.* 2016). Although computationally efficient, GBLUP and BLUP do not satisfy the second criteria (they implement a liner model and all SNPs are in the model). BayesR (Erbe *et al.* 2012) is one type of flexible non-linear model, which assumes that SNP effects follow a mixture of four normal distributions (with zero variance, very small variance, small variance, and moderate variance). Compared with GBLUP, BayesR results in superior accuracy of genomic prediction for some traits (VanRaden *et al.* 2011; Bolormaa *et al.* 2013; VanRaden *et al.* 2013; MacLeod *et al.* 2014a; Kemper *et al.* 2015; Moser *et al.* 2015). However, as the Bayesian models are typically implemented with MCMC (Markov Chain Monte Carlo) sampling, application of BayesR with sequence data is currently not feasible.

Another advantage of non-linear models such as BayesR is the application for QTL mapping (Speed & Balding 2014; Kemper *et al.* 2015; Loh *et al.* 2015; Moser *et al.* 2015; MacLeod *et al.* 2016). In detail, Loh et al. (Loh *et al.* 2015) pointed out that Bayesian mixed-model with speed-up schemes (termed fastBayesB (Meuwissen *et al.* 2009)) could improve the power of detecting genes in association with human diseases. Speed and Balding (Speed & Balding 2014) developed an efficient approach termed multiBLUP (the mixture model of SNP effects, similar to nonlinear models) on the Welcome Trust Case Control Consortium (WTCCC) human disease data to perform simultaneous disease risk prediction and causal variants analysis with complex traits. The results showed that multiBLUP could efficiently detect the genome regions associated with seven diseases. Later, Kemper et al. (Kemper *et al.* 2015) implemented nonlinear model

BayesR for QTL mapping in dairy cattle. Kemper et al. (2015) showed that BayesR could be successfully used to detect causal mutations (e.g. DGAT1) affecting milk production traits with local GEBV windows. Then, Moser et al (Moser *et al.* 2015) applied modified BayesR (updating the genetic variance in the MCMC chain instead of fixing it as in the original BayesR) to WTCCC human disease data. The posterior probability for each SNP being in the model (e.g. Not having a zero effect) was used to simultaneously map variants in association with seven diseases. The precision of BayesR was demonstrated to be higher than a mixed linear model and GWAS (Moser et al. 2015). Furthermore, Macleod et al (MacLeod *et al.* 2016) proposed the algorithm termed (BayesRC), which modified BayesR to incorporate biological prior information. Compared with GWAS, BayesRC could improve the power and precision of discovering known causal mutations (e.g. in or close to DGAT1 and PAEP, and many other genes),as well as several novel mutations, in dairy cattle data. All these previous studies demonstrated that nonlinear models, which might exclude SNPs from the models with the assumptions of Bayesian mixture priors for SNP effects, could actually help to improve the precision of QTL mapping or association studies in human or dairy cattle.

To take advantage of the accuracy superiority of MCMC nonlinear models but improve their time-efficiency, a hybrid scheme (termed HyB_BR) was proposed by Wang et al. (Wang *et al.* 2016). This scheme had three steps: 1) Implement the mixture model of BayesR, which had been demonstrated to be quite flexible for genomic prediction. 2) Converge the parameters to the optimum with Expectation-Maximization steps to speed up the program; 3) Using as starting points the solutions from the EM, run a limited number of MCMC iterations to improve the parameter estimates. The results of Hybrid algorithm on 600K dairy cattle data and 300K human disease data from Welcome Trust Case Control Consortium (WTCCC) demonstrated that Hybrid algorithm performed as well as

BayesR while requiring half of the running time demanded by MCMC iterations (Wang *et al.* 2016).

With the aim of investigating whether HyB_BR gives comparable accuracies of genomic prediction and precision of QTL mapping with whole genome sequence data to BayesR, we implemented HyB_BR on a large subset of imputed whole-genome sequence data with 994,019 variants in 16,214 cattle and. This genotype data came from the imputed sequence variants in or close to gene coding regions and 600K Bovine HD SNP genotypes. The HyB_BR algorithm was evaluated on this data set with three criteria: 1) computational performance (speed) compared to a full MCMC implementation, 2) prediction accuracy for a range of traits of different genetic architectures. The traits included Fat yield, Milk yield, Protein yield, Fat percent, Protein percent, fertility and heat tolerance (defined as the decrease in milk yield, fat yield or protein yield with increasing heat stress). 3) the precision of HyB_BR for QTL mapping for the milk production, fertility and heat tolerance traits.

6.4 Materials and Methods

High density and Sequence genotypes

Two types of genomic data, 600K Bovine HD SNP array, and imputed sequence data were used in this study. As described by Kemper et al. (Kemper *et al.* 2015), 10,311 Holstein, 4,738 Jersey and 249 Australian red bulls and cows were genotyped with the Bovine SNP50 Array (Illumina, San Diego, CA). In addition, 1,620 Holstein bulls and cows, 125 Jersey bulls, and 114 Australian Red bulls were genotyped with the 777K bovine HD SNP panel. After quality control steps described in Erbe et al. (Erbe *et al.* 2012), all genotypes were imputed to 632,003 SNP using Beagle 3.0 (Browning & Browning 2009).

For the Sequence data (termed SEQ), the sequences of 136 Holstein and 27 Jersey bulls from 1000 Bulls Genome Project (Daetwyler *et al.* 2014) were used as a reference for imputation. Using this reference set, all the animals described above with real or imputed 600K SNP genotypes were imputed into the whole genome sequence data using FImpute software (Sargolzaei *et al.* 2014). In total, there were 2.785 million sequence variants imputed, including both SNPs and indels (MacLeod *et al.* 2016). After quality control including minor allele frequency check and LD pruning by PLINK (Purcell *et al.* 2007), there were 994,019 variants remaining including 370,259 markers from the 600K SNP panel, and 623,760 sequence variants in gene coding region or 5000bp up- and down-stream of the gene start stop positions, as detailed by (MacLeod *et al.* 2016).

Genomic predictions from 600K HD SNP panel and sequence data were compared in the following investigation.

Phenotypes

The phenotype data included bulls (daughter trait deviations; DTD) and cows (trait deviations on their lactation records; TD) from Holstein and Jersey cattle, for milk production traits and fertility, detailed in Table 6.1. For milk production traits including fat yield, milk yield, protein yields, fat percent and protein percent, there were 16,214 bulls and cows from Holstein and Jersey breeds. For fertility, the number of bulls and cows in the reference set was 15,190. Then, for the validation sets, Holstein bulls and Jersey bulls (closely related to the reference set) were used to assess the accuracy of within-breed prediction. These bulls were the youngest cohort in the data set. In addition, Australian Red bulls (a third breed) were included for the validation set to evaluate the performance of across-breed prediction. We implemented the calculation of Garrick et al. (Garrick *et al.* 2009) to appropriately weight phenotypes of bulls and cows as follows:

$$w_i(bulls) = \frac{(1-h^2)}{ch^2+(4-h^2)/d} \text{ , and } w_i(cows) = \frac{(1-h^2)}{ch^2+[1+(r-1)t]/r-h^2}$$

where, h^2 is the heritability of the trait; t is the repeatability of the traits; d is the number of the daughter of each bulls; r is the number of records; c is the proportion of genetic variance not accounted for by the SNP (Garrick *et al.* 2009). To compare the prediction accuracy of GBLUP, BayesR and HyB_BR for multi-breeds and across-breed, the weight calculation is included into all of three models.

For heat tolerance, the traits were the rate of the decline of the production traits (e.g. fat, milk and protein yield) with increasing heat stress. The rate of decline for each trait was estimated for each cow in the data set with a linear random regression of yield on daily temperature-humidity index (THI), when the THI was above a threshold of 60 units (Hayes *et al.* 2003; Haile-Mariam *et al.* 2008; Nguyen *et al.* 2016). The total number of animals recorded for heat tolerance was 5,657 from Holstein and Jersey, including cows and bulls. The validation set for heat tolerance was a set of Holstein bulls and a set of Jersey bulls, Table 6.1.

Table 6.1. The number of animals in the reference sets and validation sets.

Traits	**Reference Sets**				**Validation Sets**		
	Holstein		**Jersey**		**Holstein Bulls**	**Jersey Bulls**	**Australian Red Bulls**
	Bulls	**Cows**	**Bulls**	**Cows**			
Milk production traits (FatY/MilkY/ProtY/Fat%/Protein%)	3,049	8,478	770	3,917	262	105	114
Fertility	2,806	7,838	716	3,830	396	81	114
Heat Tolerance traits (FatY decline/MilkY decline/ProtY decline)	2,028	2,037	476	1,116	252	101	-

The input parameters for HyB_BR included genetic variance, error variance, and

polygenetic variance, which were estimated from the data with ASReml4 (Table 6.2). Compared with milk production traits, low heritabilities for heat tolerance traits and fertility were estimated.

Table 6.2. The genetic architecture of milk production traits, Fertility, and Heat tolerance traits estimated by ASReml.

	Genetic variance (σ_g^2)	**Polygenetic variance (σ_a^2)**	**Error variance (σ_e^2)**	**Heritability (h^2)**
FatY	118.594	48.6891	234.326	0.42
MilkY	114827.0	38532.3	135598.0	0.53
ProtY	72.4877	36.0716	140.417	0.44
Fat%	0.0555	0.0082	0.0183827	0.78
Protein%	0.01155	0.0028	0.0032	0.82
Fertility	42.9901	0.0003	3402.87	0.01
HT_Fat[a]	0.041	0.58e-07	0.571	0.07
HT_MK[b]	0.004	0.35e-06	0.035	0.09
HT_Prot[c]	0.035	0.56e-07	0.561	0.06

[a] means Heat tolerance traits with Fat decline (HT_Fat); [b] means Heat tolerance traits with milk decline (HT_MK); [c] means Heat tolerance with protein decline (HT_Prot).

Genomic prediction methods

GBLUP. GBLUP assumes all marker effects follow the normal distribution with the same genetic variance. The overall model of GBLUP is:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Su} + \mathbf{Wv} + \mathbf{e} \quad (1)$$

Where,

$\mathbf{y}$ = vector of *n* phenotypes.

$\boldsymbol{\beta}$ = vector of b fixed effects, following uninformative priors.

$\mathbf{u} =$ vector of q random genetic values (q=number of animals) captured by the SNP, with $N(0, \mathbf{G}\sigma_g^2)$. $\mathbf{G}$ is the q x q genomic similarity matrix between pairs of individuals; σ_g^2 is the additive genetic variance.

$\mathbf{v} =$ vector of q polygenic effects (q=number of animals), with $\mathbf{v} \sim N(0, \mathbf{A}\sigma_a^2)$, $\mathbf{A}$ is the $q \times q$ pedigree-based relationship matrix, σ_a^2 is the polygenic variance.

$\mathbf{e} =$ vector of n residual errors. For cattle data, $\mathbf{e} \sim N(0, \mathbf{E}\sigma_e^2)$, the $n \times n$ diagonal matrix $\mathbf{E}$ is especially designed to evaluate the different contributions of the phenotype records from different sex to the error variance, de-regressing estimated breeding values and weighting information for genomic regression analyses (Garrick *et al.* 2009).

$\mathbf{X} = n \times b$ design matrix, allocating phenotypes $\mathbf{y}$ to fixed effects $\boldsymbol{\beta}$. b is the number of fixed effects

$\mathbf{W} = n \times q$ design matrix, which aims at allocating the $q \times 1$ vector of polygenic effects to $\mathbf{y}$.

$\mathbf{S} = n \times q$ design matrix, allocating the $q \times 1$ vector of genetic values to $\mathbf{y}$.

ASReml 4 (Gilmour *et al.* 2002) iss used to estimate variance components and genomic breeding values, and $\mathbf{G}$ iss constructed as described by (Yang *et al.* 2010).

BayesR. Compared with the common prior distributions of GBLUP, BayesR (Erbe *et al.* 2012) assumes SNP effects are drawn from the mixture of four normal distributions. BayesR aims at estimating each SNP effects instead of estimating breeding values directly for each animal. Therefore, the genetic values $\mathbf{u}$ in the model (1) is substituted into $\mathbf{Zg}$ in the BayesR model. Briefly, the data model of BayesR can be written as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zg} + \mathbf{Wv} + \mathbf{e} \quad (2)$$

Where,

$\mathbf{g} =$ m vector of SNP effects, $\mathbf{g} \sim N\left(0, \mathrm{I}\boldsymbol{\sigma}_\mathbf{i}^\mathbf{2}\right)$, $\boldsymbol{\sigma}_\mathbf{i}^\mathbf{2} = \{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 *$

$\sigma_{\mathrm{g}}^2\}$. Therefore, each SNP has four possible normal distributions: $N(0,0 * \sigma_g^2)$, $N(0,0.0001 * \sigma_g^2)$, $N(0,0.001 * \sigma_g^2)$, and $N(0,0.01 * \sigma_g^2)$. Related to such mixture priors, there are two other parameters including $b(i,k)$ and $\mathbf{Pr}$.

$b(i,k) = \{0,1\}$, which defines where or nor SNP i follows normal distribution k ($k = 1,2,3,4$). Therefore, the prior distribution of each SNP i conditional on $b(i,k)$ could be written as:

$$\mathrm{p}\big(g_i \big| b(i,k)\big) = \begin{cases} b(i,1) = 1 \\ \frac{1}{\sqrt{2\pi\sigma_{\mathrm{i}}^2[k]}} \exp\left(-\frac{g_i^2}{2\sigma_{\mathrm{i}}^2[k]}\right), \; b(i,k) = 1 (k = 2,3,4) \end{cases}.$$

$\mathbf{Pr} =$ the vector of proportions parameter, which defines the proportion SNPs in each of four normal distributions. The prior of $\mathbf{Pr}$ is drawn from Dirichlet distribution $\mathbf{Pr} \sim \mathrm{Dirichlet}(\alpha)$, with $\boldsymbol{\alpha} = [\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1}]$. The conditional distribution of SNP effect on the proportion parameter $\mathbf{Pr}$ is: $p(g_i|\mathbf{Pr}) = Pr_1 \times N(0,0 * \sigma_g^2) + Pr_2 \times N(0,0.0001 * \sigma_g^2) + Pr_3 \times N(0,0.001 * \sigma_g^2) + Pr_4 \times N(0,0.01 * \sigma_g^2)$.

$\mathbf{Z}$ is the standardized (for mean and variance) $n \times m$ genotype matrix.

To implement the BayesR model, and arrive at posterior estimates of parameters, Gibbs sampling has been used as detailed by Kemper et al (Kemper *et al.* 2015). On the sequence data, we use five independent replicate chains of the Gibbs sampling, and for each independent chain, there are 40,000 iterations, with the first 20,000 iterations discarded as burn in, as described by Kemper et al (2015) (for 630K SNP data).

HyB_BR. Motivated by improving the speed of BayesR, the HyB_BR model (Wang *et al.* 2016) incorporates the same assumption for SNP effects as BayesR but serially hybridizes the expectation-maximization (EM) and MCMC to reduce

large number of iterations required by MCMC. That is, HyB_BR first implements an EM algorithm to perform the Maximum A Posterior (MAP) estimation until converged. Then, to improve accuracy, a limited number of MCMC iterations are performed to improve parameter estimates (Wang *et al.* 2016).

As described in Wang et al. (Wang *et al.* 2016), the HyB_BR model for a SNP effect is :

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}_i g_i + \mathbf{u} + \mathbf{Wv} + \mathbf{e} \quad (3)$$

Assumptions in the model are 1) each SNP effect g_i follows the same prior assumption as BayesR with $\mathbf{Z}_i$ being the standardized genotype for SNP i. 2) to correct the prediction errors generated by all other SNPs, HyB_BR introduces the genetic values $\mathbf{u}$, whereby a correction based on the prediction error variance (PEV) is introduced to account for the effects of all the other SNP with a GBLUP model as detailed by Wang et al. (Wang *et al.* 2016). Then under the model (3), the posterior distribution for all related parameter sets including $\{g_i, \mathbf{Pr}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \sigma_e^2\}$ are derived according to the theory: $p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)p(\theta)$ where $f(\mathbf{y}|\theta)$ is the likelihood function based of model (3) and $p(\theta)$ is the prior density function for the parameter sets θ. Based on the derived marginal posterior distribution $p(\theta|\mathbf{y})$, the expectation- maximization steps are implemented to estimate each parameter while “integrating out” the other parameters detailed by Wang et al. (2016). Therefore, the process of EM module can be presented according to the pseudo code (Figure 6.1).

Pseudo-code for the EM module

function EM(σ_e^2, σ_v^2 , σ_g^2)

begin "get σ_e^2, σ_v^2 and σ_g^2 from GBLUP estimation; get vector y, matrix **Z**, lower triangle matrix **Ped**"

① Initialize **g**, Pr, b, $\hat{v}$, σ_i^2 ; Construct **X**, **A**, **G**, **E**, **W** matrices

$n \leftarrow \text{nrow}(\mathbf{Z}), m \leftarrow \text{ncol}(\mathbf{Z}), p \leftarrow 3, q \leftarrow \text{ncol}(\mathbf{Ped})$

② $\text{PEV}_u(\mathbf{e}) \leftarrow (\mathbf{E}^{-1}\sigma_e^{-2} + (\mathbf{G}\sigma_g^2 + \mathbf{WAW}'\sigma_a^2)^{-1})^{-1}$; $tr_{\text{PEV}} \leftarrow \text{tr}\left(\text{E}^{-1}\text{Z}_i\text{Z}_i'\text{E}^{-1}\text{PEV}_u(\mathbf{e})\right)$

while unconverged do

while $i \leftarrow 1$ to m do

③ $y^{\dagger} \leftarrow y - \sum_{j \neq i} Z_j \hat{g}_j - X\ddot{b} - W\hat{v}$

④ for $k_1 = 1,..,4$ calculate $logL(i,k_1)$ with the equations:

$$\boldsymbol{V} = (\boldsymbol{\sigma}_k^2 \mathbf{Z}_i'\mathbf{E}^{-1}\mathbf{Z}_i + \sigma_e^2)$$

$$logL(i,k) \leftarrow logPr_k - \frac{1}{2}\left\{log\, \boldsymbol{V} - \left[\left(\mathbf{y}^{\dagger'}\mathbf{E}^{-1}\mathbf{Z}_i\right)^2 - tr_{\text{PEV}}\right]\boldsymbol{\sigma}_i^2[\mathbf{k}]\sigma_e^{-2}/\boldsymbol{V}\right\}$$

update each $P(i,k_2)$ $(k_2 = (1,\ldots,4))$ with $\frac{exp(logL(i,k_2))}{\sum_{k_2=1}^{4} exp(logL(i,k_2))}$

⑤ calculate $\hat{g}_i$ from equation

$$\hat{g}_i \leftarrow [\mathbf{Z}_i'\mathbf{E}^{-1}\mathbf{Z}_i + \sum_{k=1}^{4}\left(P(i,k)\frac{\widetilde{\sigma_e^2}}{\sigma_i^2[k]}\right)]^{-1}[\mathbf{Z}'\mathbf{E}^{-1}\mathbf{y}^{\dagger}]$$

end

⑥ update **Pr**, σ_e^2, **β**, **v** using the equations:

$$Pr_k \leftarrow \frac{\sum_{i=1}^{m} P(i,k)+1}{\sum_{k=1}^{4}(\sum_{i=1}^{m} P(i,k)+1)}$$

$$\widetilde{\sigma_e^2} \leftarrow \frac{1}{n}\left[\left(\left(\mathbf{y} - \mathbf{Zg} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\hat{\mathbf{v}}\right)\right)' \mathbf{E}^{-1}\left(\mathbf{y} - \mathbf{Zg} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{W}\hat{\mathbf{v}}\right) + \text{tr}\left(\mathbf{E}^{-1}\text{PEV}_u(\mathbf{e})\right)\right]$$

$$\widehat{\boldsymbol{\beta}} \leftarrow (\mathbf{X}'\mathbf{E}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}^{-1}(\mathbf{y} - \mathbf{Zg} - \mathbf{W}\hat{\mathbf{v}})$$

$$\hat{\mathbf{v}} \leftarrow (\mathbf{W}'\mathbf{E}^{-1}\mathbf{W}\sigma_a^2 + \sigma_e^2\mathbf{A}^{-1})^{-1}\sigma_a^2\mathbf{W}'\mathbf{E}^{-1}(\mathbf{y} - \mathbf{Zg} - \mathbf{W}\hat{\mathbf{v}} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

unconverged←False

⑦ if $(\hat{g}^q - \hat{g}^{q-1})'(\hat{g}^q - \hat{g}^{q-1})/((\hat{g}^{q'}\hat{g}^q) > 10^{-10}$ then

unconverged←True

endif

end

end function

Figure 6.1. The pseudo-code of the EM module.

As shown in Figure 6.1, the EM module begins by initializing all the input

parameters including SNP effects (**g**), Proportion parameter (**Pr**), the variance for each SNP ($\boldsymbol{\sigma_i^2}$), the fixed matrix (**X**), the pedigree based relationship matrix (**A**), the genomic relationship matrix (**G**), the error matrix (**E**), and index matrix for polygenic effects (**W**). Similar to emBayesR (Wang *et al.* 2015), the starting values of $\boldsymbol{g}$ and **Pr** were set as $\boldsymbol{g} = 0.01$ and $\mathrm{Pr} = \{0.5, 0.487, 0.01, 0.003\}$, while $\boldsymbol{\sigma_i^2} = \{0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2\}$. The genetic variance σ_g^2, error variance σ_e^2, and polygenic variance σ_a^2 are obtained from ASReml, with the value of genetic variance and polygenic variance then fixed. The $n \times 3$ matrix **X** is design matrix, allocating the phenotypes to fixed effects. In our case, matrix **X** is set up with first column being the mean, the second and third columns defining the breeds (Holstein and Jersey) and sex (bulls and cows) of the cattle. The pedigree relationship matrix **A** is built up using the lower symmetrical matrix **Ped** detailed by Henderson (Henderson 1984); while the genomic relationship matrix **G** is constructed using the equation $\mathbf{G} = \mathbf{Z}^s\mathbf{Z}^{s\prime}/n$, $\mathbf{Z}^s$ is the standardized **Z** matrix with $\mathbf{Z}^s{}_{ij} = {(\mathbf{Z}_{ij} - 2p_i)}\Big/{\sqrt{2p_i(1-p_i)}}$. Diagonal error matrix **E** is constructed according to the equation defined by Garrick et al. (Garrick *et al.* 2009).

Then, after initializing step, the pseudo code of Figure 6.1 describes the process of the EM module (which was detailed in Wang et al. (2016)).

The EM steps require the time complexity $O(mn)$. For the calculation of $tr\left(\mathbf{E^{-1}Z_i}\mathbf{Z_i'E^{-1}}\mathrm{PEV_u}(\mathbf{e})\right)$ which is calculated prior to the EM steps, the required time is $O(m^2n)$. This calculation accounts for 40% of the total computational time. Since the calculation is independent SNP by SNP, we parallelize the operations by chromosomes, which can reduce around 30% of the total running time.

Once the EM has converged using the criterion ($(\hat{\mathrm{g}}^{\mathrm{q}} - \hat{\mathrm{g}}^{\mathrm{q}-1})'(\hat{\mathrm{g}}^{\mathrm{q}} - \hat{\mathrm{g}}^{\mathrm{q}-1})/$

$((\hat{g}^{q\prime}\hat{g}^{q}) > 10^{-10}$ with q be the iteration number, the parameter estimates from the EM are used as starting points of parameter values in the MCMC iterations. The steps of MCMC iterations were detailed by Kemper et al. (Kemper *et al.* 2015). Furthermore, Wang et al. (Wang *et al.* 2016) suggested a speed-up scheme to improve computational efficiency. The scheme is as follows. After 500 MCMC iterations, the SNPs with high probability in the distribution with zero variance will be excluded from the model. In other words, when $P(i,1) \geq 0.90$, the SNP effects will be set as zero. Also, as investigated by Wang et al. 2016, HyB_BR requires 4,000 MCMC iterations for both 600K SNP panel and imputed sequence data to maximize accuracy of genomic prediction (Wang *et al.* 2016).

To compare the computational cost between BayesR and HyB_BR and how this changes with an increasing number of individuals in the reference set, we subseted the data of Table 6.1 into three different reference sets (Ref1, Ref2, and Ref3) (with the number of sequence variants hold constant). Ref1 had Holstein bulls only with 3,049 bulls; Ref2 included Holstein bull and cow data with 12,527 animals; Ref3 had all data with 16,214 animals.

On all three reference sets, the speed advantage of HyB_BR compared with BayesR was investigated. Then the accuracy of genomic prediction from BayesR, HyB_BR and GBLUP was compared in the full data (including the sequence variants). In addition, the precision of mapping QTL from the three methods was compared.

6.5 Results

Computational time comparison between GBLUP, BayesR and HyB_BR

On both 600K and SEQ data sets, HyB_BR was more than 10 times faster than BayesR, Figure 6.2. As the size of the data set increased (from Ref 1 to Ref3 or from 600K to SEQ data), the computational time required for HyB_BR reduced by a greater and greater margin relative to BayesR. On 600K data, HyB_BR had a similar compute time to GBLUP. For the SEQ data, HyB_BR was up to four fold faster than GBLUP.

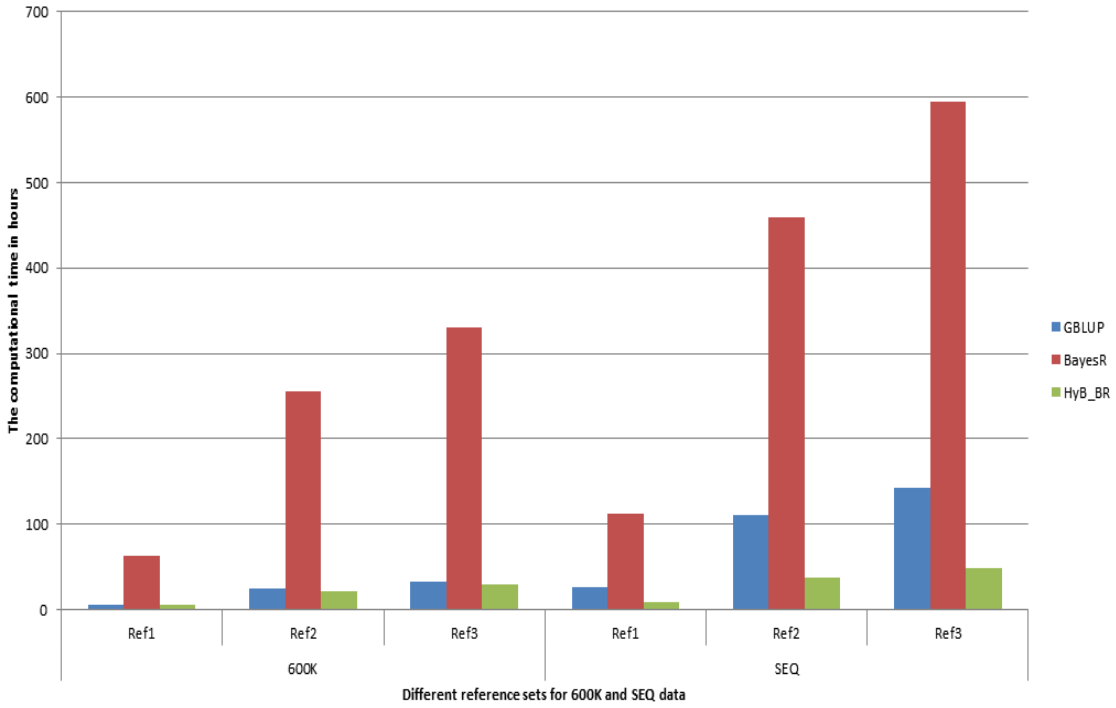


Figure 6.2. The computational time comparison between GBLUP, BayesR and HyB_BR on 600K and SEQ data. Three reference sets (Ref1, Ref2 and Ref3) with the same number of variants (600K or SEQ) are used here. Ref1 has Holstein bulls data only with 3,049 animals; Ref2 has Holstein bull and cow data with 12,527 animals; Ref3 has Holstein and Jersey bulls and cows with 16,214 animals.

Accuracy of genomic prediction for GBLUP, BayesR, and Hybrid with sequence data

Prediction accuracy for milk production traits and fertility. For the milk production and fertility traits, the combined Holstein and Jersey reference sets were used to predict three validation sets including Holstein bulls (Table 6.3), Jersey bulls (Table 6.3), and Australian red bulls & cows (Table 6.4).

When predicting the Holstein validation bulls data, BayesR and HyB_BR performed equally well. Compared with GBLUP, BayesR and HyB_BR had consistent accuracy improvement for the milk production traits except Protein percent. For Fat% trait, BayesR and HyB_BR gave 5% higher accuracy than GBLUP. However, on the traits Protein% and Fertility, there was no difference between these methods. The results were similar when the Jersey validation set was used. However there were several exceptions for the prediction of Jersey bulls as the validation set: 1) on the Jersey validation set, the accuracy superiority of HyB_BR and BayesR over GBLUP became more obvious. For example, for Fat percent, BayesR and HyB_BR gave a 10% higher accuracy than GBLUP; 2) HyB_BR even had 1% accuracy increase than BayesR on most cases. HyB_BR and BayesR also gave regression coefficients (DTD on GEBV) closer to one than GBLUP for most traits.

In addition, when incorporating polygenic effects into the prediction model, a small but consistent accuracy improvement was observed, Table 6.3. However, for fertility, including polygenic effects did not affect the prediction accuracy at all.

When predicting Australian red bulls and cows using the combined Holstein and Jersey reference set (across breed prediction), both HyB_BR and BayesR had a considerable accuracy advantage (up to 12% increase) over GBLUP for all the traits (Table 6.4). Compared with BayesR, HyB_BR performed equal or better in terms of accuracy for all traits except fat yield.

Accuracy of genomic prediction for heat tolerance. The accuracy of genomic prediction for the heat tolerance traits was similar for GBLUP, BayesR, and HyB_BR, Table 6.5. There were two exceptions when predicting the validation set of Jersey bulls: 1) On the heat tolerance for fat yield, there was 6% accuracy reduction of BayesR and HyB_BR in comparison with GBLUP; 2) For the milk yield, 9% increase in accuracy from BayesR and HyB_BR over that from GBLUP was observed. HyB_BR and BayesR also gave regression coefficients closer to one than GBLUP for all the traits.

Table 6.3. The multi-breed prediction accuracy (Holstein and Jersey validation sets) and bias of GBLUP, BayesR, and HyB_BR on SEQ data related to Fat Yield, Milk Yield, Protein Yield, Fat%, Protein% and Fertility. +Poly is the prediction accuracy when adding the polygenic term in the model; While -Poly is the prediction accuracy when leaving out the polygenic term from the model.

		Holstein and Jersey reference to predict Holstein validation											
		Fat Yield		**Milk Yield**		**Protein Yield**		**Fat%**		**Protein%**		**Fertility**	
		Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	+Poly [a]	0.64	1.07	0.66	0.92	0.63	0.95	0.76	0.95	0.83	0.98	0.42	1.70
	-Poly [b]	0.62	1.32	0.60	0.83	0.58	1.15	0.75	1.01	0.81	1.09	0.42	1.70
BayesR	+Poly [a]	0.65	1.27	0.69	0.91	0.68	1.04	0.81	1.01	0.83	0.99	0.42	1.32
	-Poly [b]	0.63	1.17	0.67	0.85	0.65	0.91	0.80	1.01	0.82	0.96	0.42	1.32
HyB_BR	+Poly [a]	0.66	1.04	0.69	0.89	0.68	0.96	0.81	0.99	0.83	0.96	0.42	1.32
	-Poly [b]	0.63	0.96	0.69	0.89	0.66	0.88	0.81	0.99	0.81	0.94	0.42	1.32
		Holstein and Jersey reference to predict Jersey validation											
		Fat Yield		**Milk Yield**		**Protein Yield**		**Fat%**		**Protein%**		**Fertility**	
		Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**

GBLUP	+Poly [a]	0.54	0.76	0.65	0.88	0.69	0.94	0.67	0.86	0.77	0.94	0.23	1.13
	-Poly [b]	0.52	0.93	0.65	1.03	0.68	1.24	0.66	0.93	0.75	1.02	0.23	1.13
BayesR	+Poly [a]	0.57	0.88	0.70	0.96	0.72	1.22	0.77	0.97	0.77	0.89	0.23	1.03
	-Poly [b]	0.52	0.73	0.68	0.87	0.67	1.02	0.76	0.95	0.77	0.87	0.23	1.02
HyB_BR	+Poly [a]	0.58	0.87	0.69	0.95	0.73	0.91	0.77	0.93	0.79	0.87	0.23	0.97
	-Poly [b]	0.57	0.74	0.69	0.85	0.73	0.91	0.76	0.93	0.78	0.85	0.23	0.97

The bulls and cows from two breeds of Holstein and Jersey are used as reference set to predict Holstein bulls and Jersey bulls separately.

Table 6.4. The across breed prediction accuracy (validation data set Australian red bulls and Australian red cows) of GBLUP, BayesR, and HyB_BR on SEQ data related to Fat Yield, Milk Yield, Protein Yield, Fat%, Protein% and Fertility.

	Across breeds prediction on Australian red bulls											
	Fat Yield		**Milk Yield**		**Protein Yield**		**Fat%**		**Protein%**		**Fertility**	
	Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	0.13	0.58	0.21	0.59	0.15	0.71	0.39	0.61	0.50	1.32	0.22	0.96
BayesR	0.35	1.31	0.22	0.77	0.24	0.92	0.40	0.61	0.53	0.86	0.27	0.97
HyB_BR	0.28	0.74	0.36	0.70	0.26	0.74	0.47	0.66	0.53	0.88	0.27	0.95
	Across breeds prediction on Australian red cows											
	Fat Yield		**Milk Yield**		**Protein Yield**		**Fat%**		**Protein%**		**Fertility**	
	Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	0.15	0.77	0.11	0.37	0.12	0.57	0.31	0.92	0.34	1.09	0.07	0.61
BayesR	0.28	1.02	0.22	0.55	0.16	0.60	0.37	0.94	0.34	0.93	0.07	0.52
HyB_BR	0.25	0.88	0.23	0.54	0.16	0.59	0.37	0.91	0.34	0.91	0.07	0.57

The bulls and cows from two breeds of Holstein and Jersey are used as reference set to predict Australian red bulls and cows.

Table 6.5. The multi-breed prediction accuracy (Holstein and Jersey validation sets) and bias of GBLUP, BayesR, and HyB_BR with SEQ data and heat tolerance traits.

	Holstein and Jersey reference Prediction on Holstein bulls					
	Fat		**Milk**		**Protein**	
	Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	0.35	1.47	0.24	0.84	0.32	1.24
BayesR	0.35	1.05	0.29	0.88	0.33	0.92
HyB_BR	0.35	1.05	0.28	0.86	0.33	1.01
	Holstein and Jersey reference Prediction on Jersey bulls					
	Fat		**Milk**		**Protein**	
	Acc.	**Bias**	**Acc.**	**Bias**	**Acc.**	**Bias**
GBLUP	0.33	1.25	0.37	1.11	0.35	0.72
BayesR	0.27	0.89	0.46	0.89	0.35	0.76
HyB_BR	0.27	0.88	0.46	0.89	0.35	0.77

The bulls and cows from two breeds of Holstein and Jersey are used as reference set to predict Holstein bulls and Jersey bulls separately.

The impact of sequence data on the prediction accuracy of GBLUP, BayesR, and HyB_BR.

Compared with 600K SNP panels, the impact of sequence data (SEQ) on the prediction accuracy of GBLUP, BayesR, and HyB_BR depended on trait and validation population (Figure 6.3). For the prediction of the validation sets of Holstein or Jersey bulls (which were closely related to the reference set), only small accuracy gain (1%~2%) was observed from using sequence data compared to using the 600K panel. However, for the validation set Australian Red bulls and cows, there was more advantage of using the sequence data, provided BayesR or HyB_BR were used. For example, the accuracy using BayesR and HyB_BR on sequence data was up to 13% higher than when the 600K SNP panel was used. In the same situation, GBLUP gave only a very limited increase (or even reduction for Fat Yield trait) when using the sequence data.

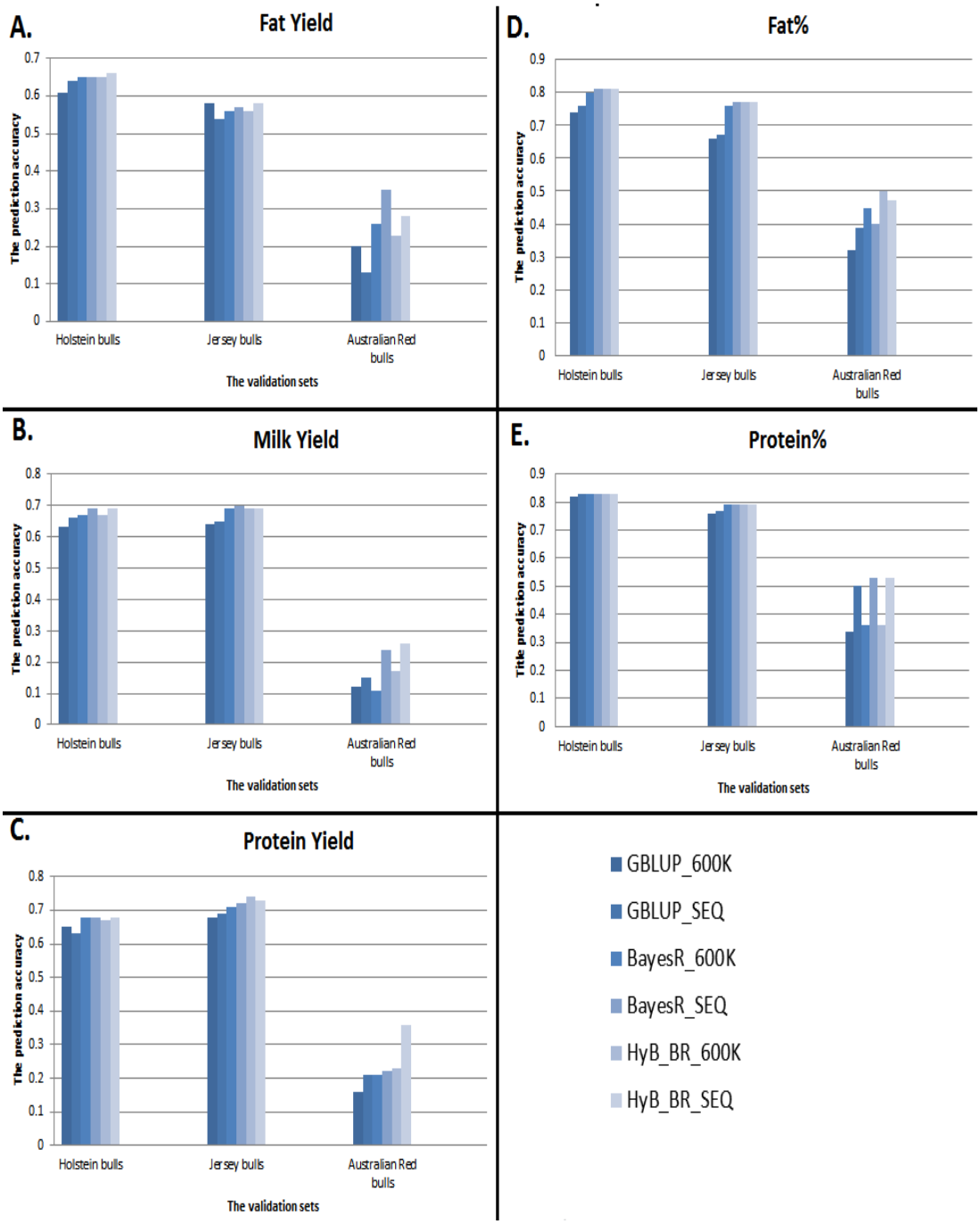


Figure 6.3. The prediction accuracy of GBLUP, BayesR, and HyB_BR on 600K and SEQ data related to three milk production traits including Fat Yield (A.), Milk Yield (B.), Protein Yield (C.), Fat Percent (D.), and Protein Percent (E.).

Inference of genetic architecture

To infer the underlying genetic architecture, the proportion of SNPs contributing the total additive genetic variances was expressed as the posterior estimate of the proportion of SNPs in each of the three non-zero distributions (with the variance $0.0001 * \sigma_g^2$, $0.001 * \sigma_g^2$, or $0.01 * \sigma_g^2$) relative to the total number of SNPs fitted in the model (Figure 6.4). Across all the traits, BayesR and HyB_BR gave a similar proportion of SNP in each distribution. For example, both BayesR and HyB_BR estimated a relatively larger proportion of SNPs in the largest distribution ($0.01 * \sigma_g^2$; the red bars) for Fat percent and Protein percent than for the other traits. In detail, as shown in the Figure 6.4, the red bars for trait Fat percent is5.5% more than Fat yield, Milk yield, and protein yield; while on trait Fat percent, there is 5.8% more. Such inference agreed with the discovery of several mutations of moderate or large effects for the % traits, including DGAT1 (Grisart *et al.* 2002) on Chromosome 14 and GHR (Blott *et al.* 2003) on Chromosome 20.

For fertility, HyB_BR and BayesR both gave estimates of a very large number of SNP in the smallest non-zero distribution, consistent with a highly polygenic architecture for fertility investigated by previous researches (Pryce *et al.* 2010; Sahana *et al.* 2011; Schulman *et al.* 2011). In addition, from Figure 6.4, we could detect some differences between HyB_BR and BayesR. For milk yield and protein yield, HyB_BR derived more SNPs in the distribution with variance $0.001 * \sigma_g^2$ but a smaller number of SNPs in the distribution with very small variance $0.0001 * \sigma_g^2$, than BayesR.

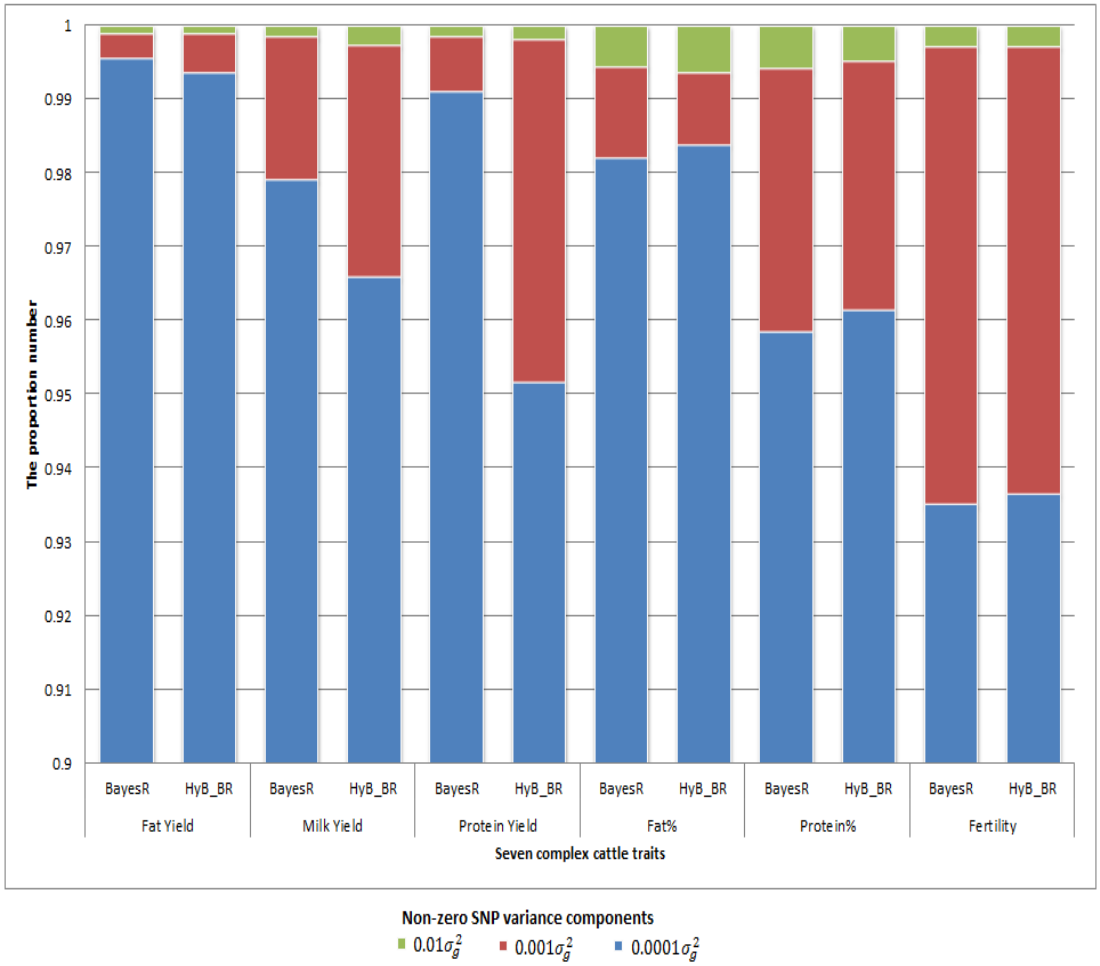


Figure 6.4. The genetic architecture of six traits inferred by BayesR and HyB_BR. The proportion parameter $\Pr$ is the proportion of SNPs in each of four possible normal distributions with variances $(0, 0.0001 * \sigma_g^2, 0.001 * \sigma_g^2, 0.01 * \sigma_g^2)$. Three bars with different colors label the proportion of SNPs contributing the total additive genetic variances, which is standardized from the proportion of SNPs per non-zero distribution relative to the total number of SNPs fitted in the model.

QTL mapping

For each trait, the top variants with highest posterior probability of being in the distribution with the largest variance $(0.01 * \sigma_g^2)$, and largest effects, from BayesR and HyB_BR were investigated.

QTL mapping for milk production traits and fertility. Table 6.6 listed all the top variants influencing milk production and fertility detected by BayesR and HyB_BR, which have also been demonstrated by previous studies (novel variants were described below). The top variants detected by both BayesR and HyB_BR (Table 6.6) were in or close to many previously described causal mutations involved in milk productions. For example, in Table 6.6, there were some well-known mutations impacting milk fatty acid synthesis include DGAT1 (Grisart *et al.* 2002; Schennink *et al.* 2007; Schennink *et al.* 2008), FASN (Roy *et al.* 2006), SCD (Mele *et al.* 2007), PAEP (Ng-Kwai-Hang), AGPAT6 (Wang *et al.* 2012b; Littlejohn *et al.* 2014), and CNS2/3 (MacLeod *et al.* 2016). In addition, HyB_BR could detect some novel causal mutations including GC (encoding the vitamin D binding protein, affecting milk yield), SMEK1 (regulating the Insulin/IGF pathway, indirectly impacting milk production and fertility) and MYH9 (myosin, heavy chain 9, non-muscle; impacting protein yield (Chamberlain *et al.* 2015; Raven *et al.* 2015; MacLeod *et al.* 2016). In fertility, the causal mutations located on Chromosome 18 including (CTU1 and CEACAM18) detected by BayesR and HyB_BR had been demonstrated to significantly affect direct calving traits (Mao *et al.* 2015; Purfield *et al.* 2015).

Table 6.6. Known genes (impacting milk production traits and fertility) identified by HyB_BR using the variants with the largest variances $0.01 * \sigma_g^2$.

Gene	BTA	Position (bp)	Fat	Milk	Protein	Fat %	Fertility	Description
ROBO1	1	26212317			✓			Roundabout, axon guidance receptor; Positively impacted the protein yield related to milk productions (Chamberlain *et al.* 2015; Raven *et al.* 2015).
SLC37A1	1	144437682	✓	✓				Glucose transport, which negatively impacted the milk and Fat yield, but with higher Fat% and Protein% (Fritz *et al.* 2013).
MYH9	5	75157624			✓			Myosin, heavy chain 9, non-muscle; Positively impacting the protein yield (Chamberlain *et al.* 2015; Raven *et al.* 2015; MacLeod *et al.* 2016).
CSF2RB	5	75736516	✓	✓				The JAK-STAT signal pathway, which strongly contributed the Milk and Fat yield (Chamberlain *et al.* 2015).
LDHB	5	88975951					✓	lactate dehydrogenase B, the highest expression level on the lactation (Mishra *et al.* 2013).
GYS2	5	89080460					✓	Involvement in glycogen biosynthesis, showing significant under-expression in mammary tissue (Hwang *et al.* 2012; Chamberlain *et al.* 2015).

MGST1	5	93950493~ 93954943	✓	✓		✓		Microsomal glutathione S-transferase, which was reported to negatively impact fat yield and fat% (Wang *et al.* 2012b; Chamberlain *et al.* 2015; Raven *et al.* 2015), but positively contributed milk yield.
GRID2	6	32205789			✓			Encoding an ionotropic glutamate receptor, which impacted protein yield (Chamberlain *et al.* 2015; Raven *et al.* 2015; MacLeod *et al.* 2016).
GC	6	88741762		✓			✓	Group-specific Component, encoding the vitamin D binding protein, which had been investigated to positively impact the milk yield (Raven *et al.* 2015).
CSN2	6	87180731			✓			Well-known casein gene cluster, strongly impacting the protein content of bovine milk (MacLeod *et al.* 2016).
CSN3	6	87390576			✓			
HSD17B3	8	84379597					✓	Hydroxysteroid–dehydrogenases, known to affect reproductive processes (e.g. steroidogenesis) (Cochran *et al.* 2013).
PAEP	11	103303475		✓	✓	✓		The alias beta-lactoglobulin gene, encoding the primary whey protein of bovine milk. PAEP had been reported to had a large effect on protein yield and smaller effects on MY and Fat%(Ng-Kwai-Hang 1997).
SLC39A4	14	1716713	✓		✓	✓		The member of the Zinc/Iron-regulated transporter-like family, encoding a zinc-specific transporter (Schmitt *et al.* 2009). In bovine, SLC39A4 affecting fat/fat percent and protein content of milk

								productions (D'Alessandro *et al.* 2011)
CPSF1	14	1726659	✓					The gene near to DGAT1, impacting milk fat composition.
DGAT1	14	1801116~ 1802266	✓	✓	✓	✓		The diacylglycerol O-acyltransferase 1, well-known gene which had a large influence on the milk fat composition (Grisart *et al.* 2002; Schennink *et al.* 2007; Schennink *et al.* 2008).
CTU1	18	57521276~ 57527946					✓	The missense variant, affecting direct carving difficulty (Purfield *et al.* 2015).
CEACAM 18	18	57548213					✓	The member of the carcinoembryonic antigen (CEA) gene family, which had been reported to significantly affect direct calving traits (Mao *et al.* 2015).
KRT19	19	42366926	✓					The member of a family of cytokeratins responsible for the structural integrity of epithelial cells. KRT19 was reported to indirectly affect fat content of milk yield (Chamberlain *et al.* 2015).
FASN	19	51381233				✓		The multifunctional protein that carried out the synthesis of fatty acids highly affecting the fat milk content (Roy *et al.* 2006).
GHR	20	31699535		✓		✓		The growth hormone receptor, positively impacting the milk productions (Blott *et al.* 2003).
SMEK1	21	56798101		✓	✓		✓	The gene, regulating the Insulin/IGF pathway. SMEK1 had been investigate to indirectly impact the milk and protein content of the milk productions (MacLeod *et al.* 2016). Also, SMEK1 had been

								demonstrated to regulate the differentiation of embryonic stem cells for fertility (Lyu *et al.* 2011).
GMDS	23	51280200		✓			✓	The enzyme GDP-mannose-4, 6-dehydratase, which was reported to indirectly impacting the milk production (Wickramasinghe *et al.* 2011).
SCD	26	21139935	✓					The stearoyl-CoA desaturase, which was in milk fat synthesis pathways and highly impacted the fat content of milk productions (Mele *et al.* 2007).
GINS4	27	36155097				✓		Near to AGPAT6 gene, which had been reported to impact the Fat milk (Littlejohn *et al.* 2014; Raven *et al.* 2015).
AGPAT6	27	36211252				✓		A family of 1-acylglycerol-3-phosphate acyltransferases (AGPATs), which had been reported to be strongly associated with high milk fat percentage (Wang *et al.* 2012b; Littlejohn *et al.* 2014).

The blue bar highlights the genes that cannot be detected by BayesR in the proportion with the largest variances.

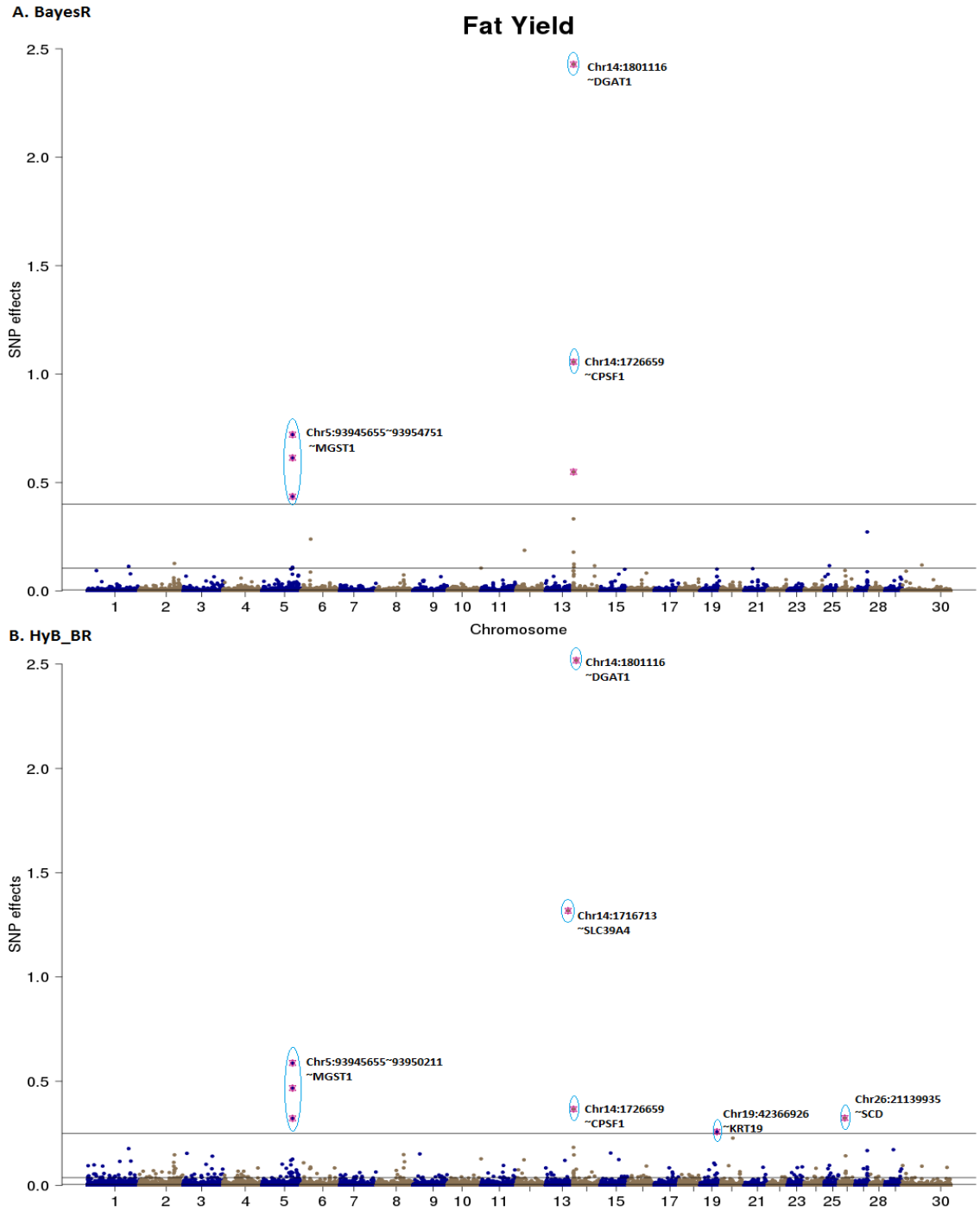


Figure 6.5. Effects of all the variants on fat yield estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole genome. The top SNPs with moderate to large effects are labelled with blue circle.

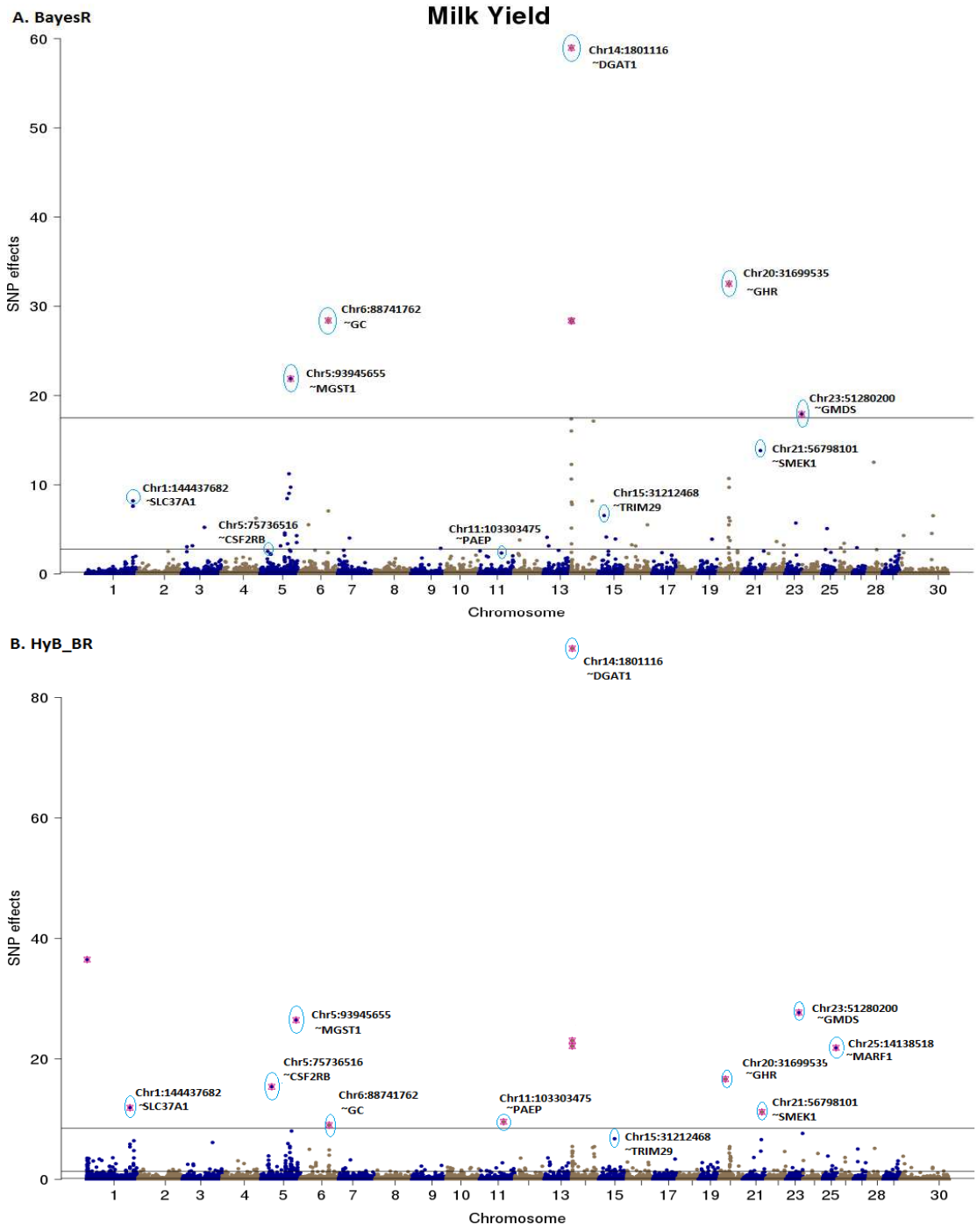


Figure 6.6. Effects of all the variants for milk yield estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole genome. The top SNPs with moderate to large effects are labelled with blue circle.

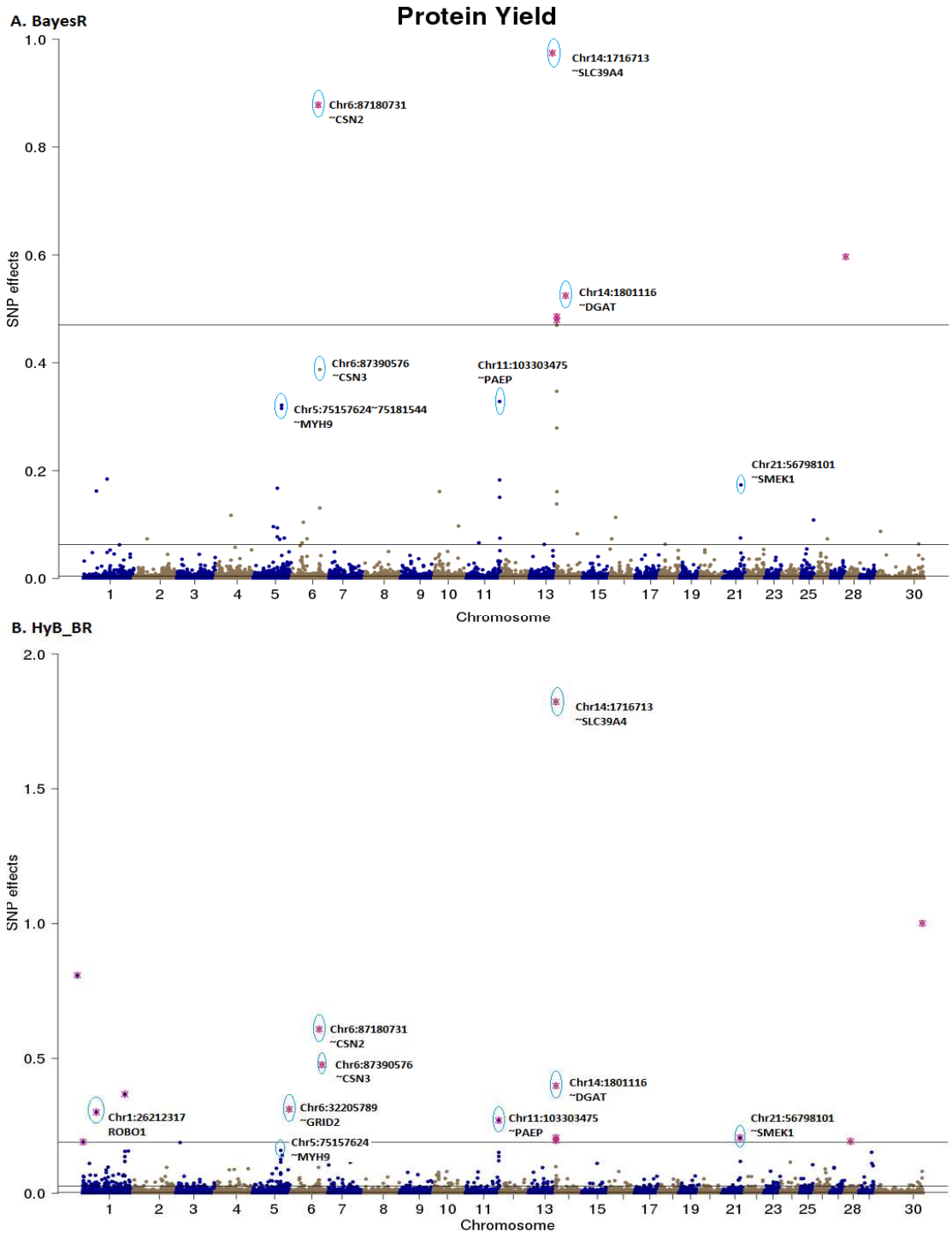


Figure 6.7. Effects of all the variants for protein yield estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole chromosome genome. The top SNPs with moderate to large effects are labelled with blue circle.

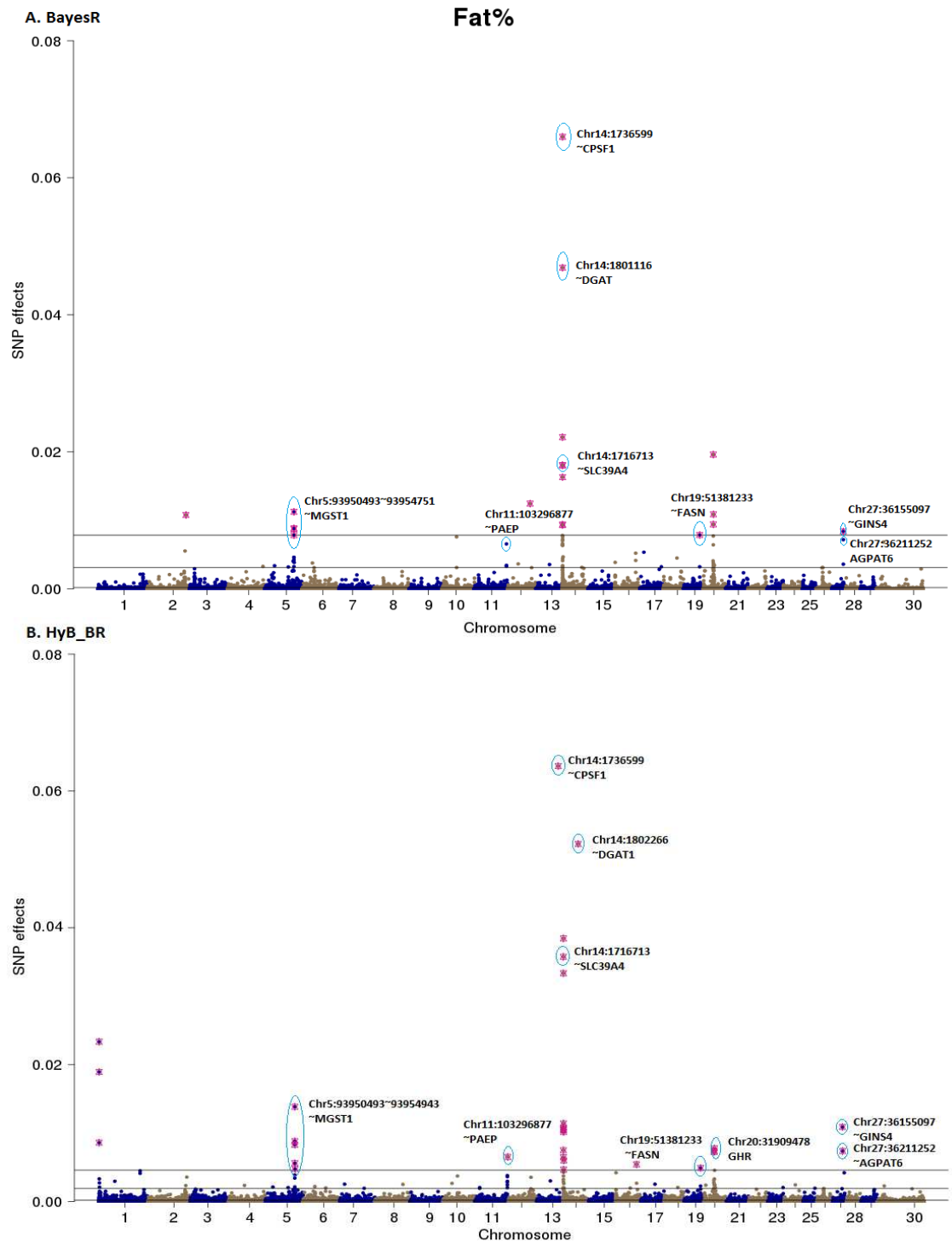


Figure 6.8. Effects of all the variants for fat percent estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole genome. The top SNPs with moderate to large effects are labelled with blue circle.

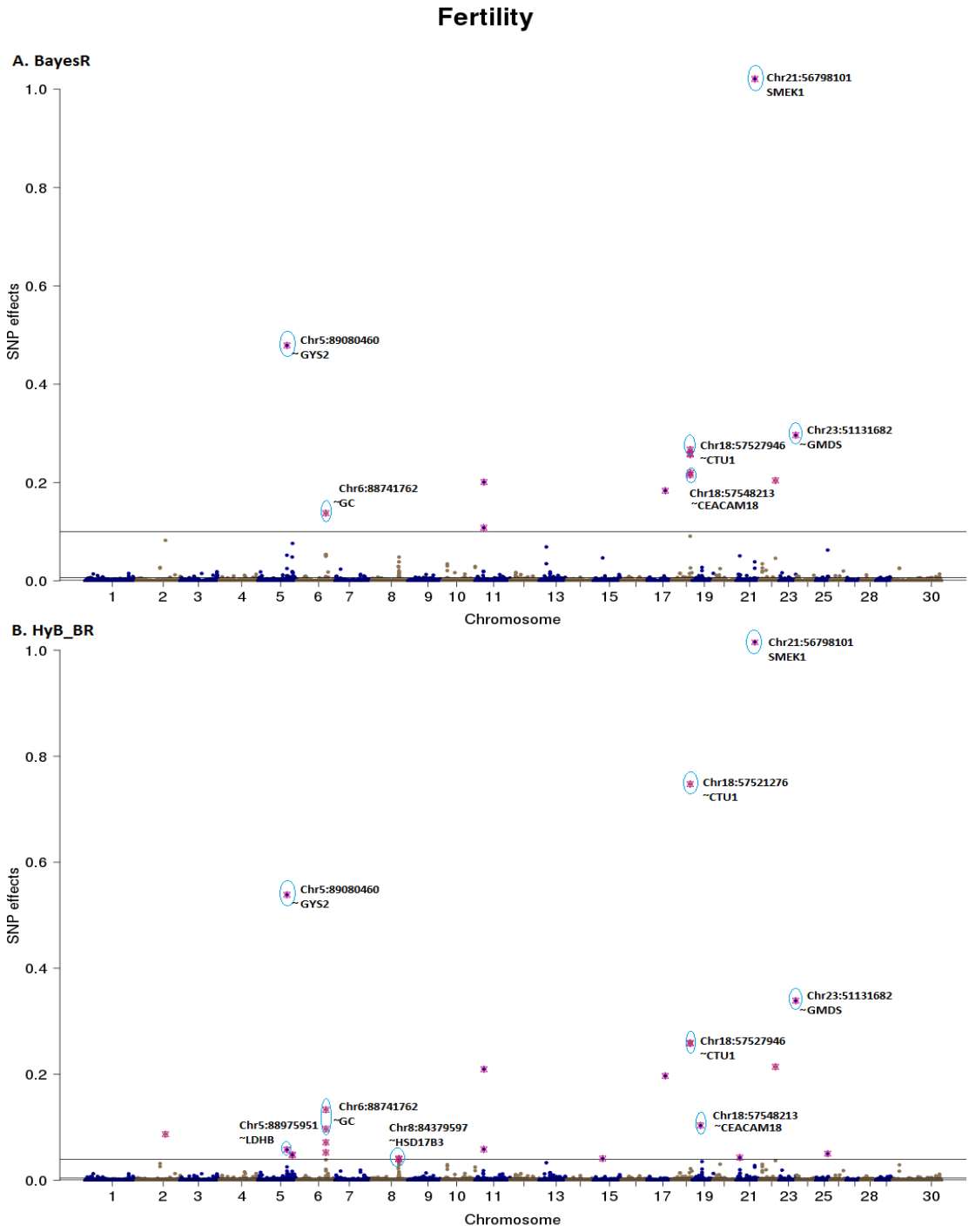


Figure 6.9. Effects of all the variants on fertility estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole genome. The top SNPs with moderate to large effects are labelled with blue circle.

QTL mapping for Heat tolerance traits. There were relatively little previous literatures reporting QTL for heat tolerance in cattle. A QTL (close to FGF4) (Hayes *et al.* 2009), later suggested to be SHANK2 by (Dikmen *et al.* 2015) was located at Chromosome 29 with the position 48329079 bp. In our study, neither BayesR nor HyB_BR detected this gene with a high posterior probability of non-zero effect. In Figure 6.11, the gene SHANK2 was detected but not in the list of top causal mutations.

To avoid the impact of several major causal mutations (e.g. DGAT1) affecting milk production traits, we first fitted these well-known variants including DGAT1, ROBOT1, PAEP, and MGST1 as fixed effects for our further investigation. Then, aiming at detecting all the top variants, the posterior possibilities of all the variants estimated by HyB_BR and BayesR were plotted across the whole genome in Figure 6.10, Figure 6.11, and Figure 6.12. In total, we found fourteen novel variants (Table 6.7), which were in response to heat tolerance in human or other species. There were several typical instances. YBEY (Rasouly *et al.* 2009; Grinwald & Ron 2013), located at BTA1 with the position 147710807 bp, has been reported to be important in Escherichia coli of human or other animals under heat-shock response. Two unknown genes locating at BTA2: 112901035 (near the region harbored by SERPINE2) and BTA22: 47737890 (near the region harbored by CACNA1D) have been reported to impact the sweating rate and respiration rate of dairy cattle (Dikmen *et al.* 2015). DYRK3 (The dual specificity tyrosine-phosphorylation-regulated kinase 3), has been proved to affect respiration rate (breaths per minute) in dairy cattle (Dikmen *et al.* 2015). HSF1, heat shock factor protein 1, coordinated stress-induced transcription in Human (Rabindran *et al.* 1991). One single nucleotide polymorphism (SNP) in the 3'-untranslated region (g.4693G>T) of HSF1 has been reported to be in association with thermo tolerance in Chinese Holstein cattle (Li *et al.* 2015).

STIP1, stress inducible protein 1, has been reported to be homologous to hsc70/hsp90 in human (Schmid *et al.* 2012). In mouse, STIP1 could play a key role on in the ability of germ cells to survive in stress conditions including high temperatures (Mizrak *et al.* 2006). Further investigation is required for all of these genes.

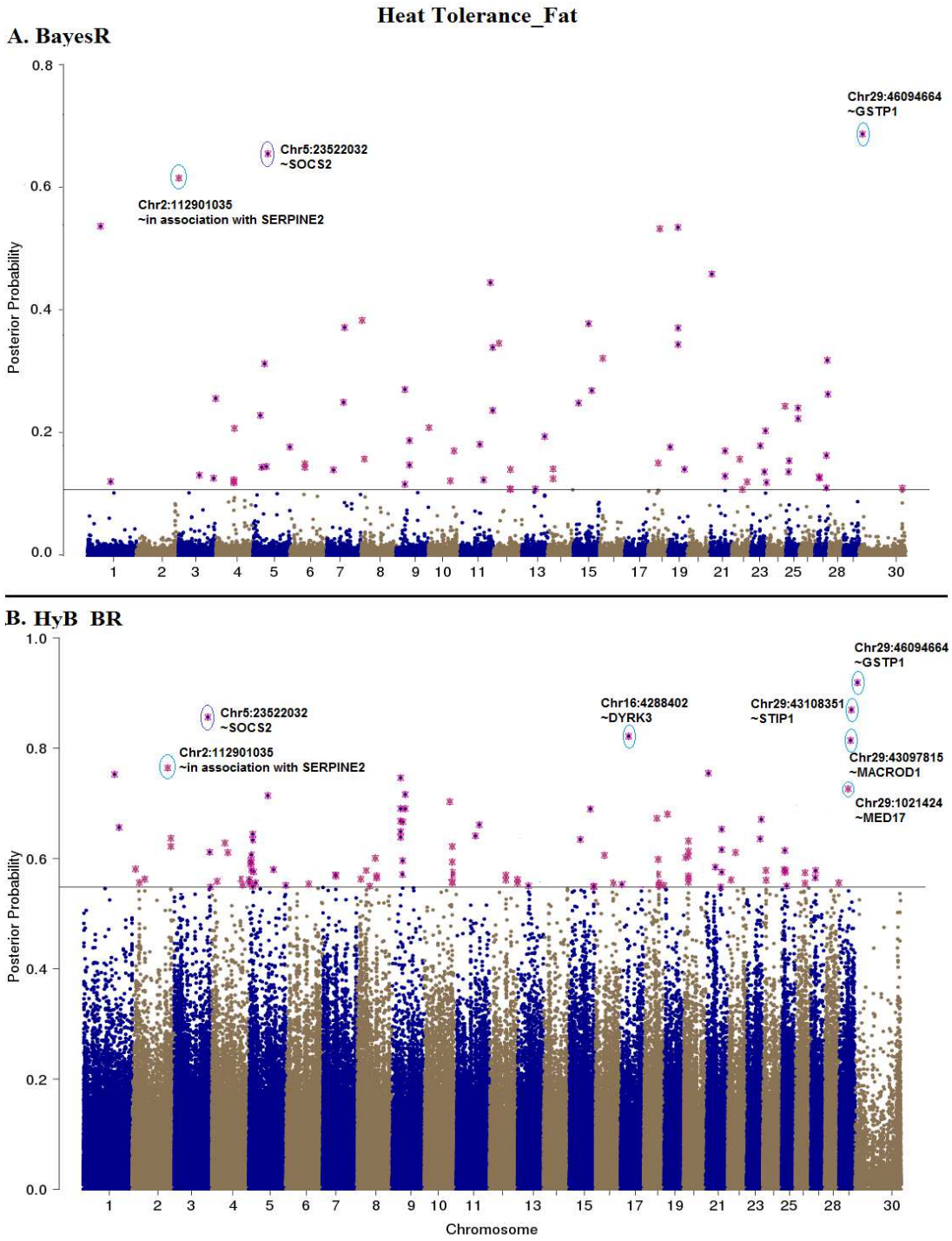


Figure 6.10. Mapping posterior probabilities of all the variants estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole chromosome related to Fat yield affected by heat tolerance.

The top SNPs with highest posterior possibilities are labelled with blue circle.

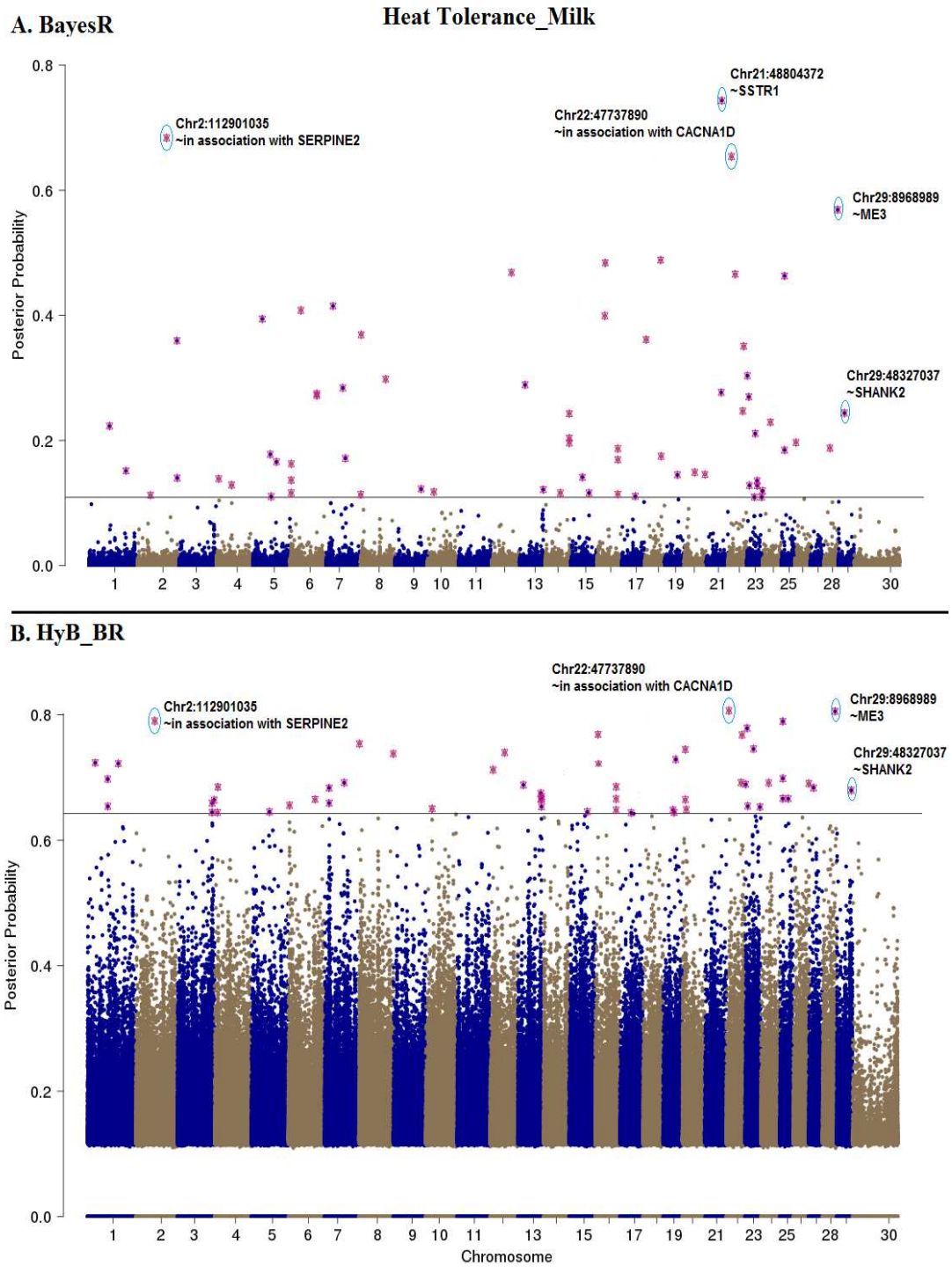


Figure 6.11. Mapping the posterior probabilities of all the variants estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole chromosome related to Milk yield affected by heat tolerance.

The top SNPs with highest posterior possibilities are labelled with blue circle.

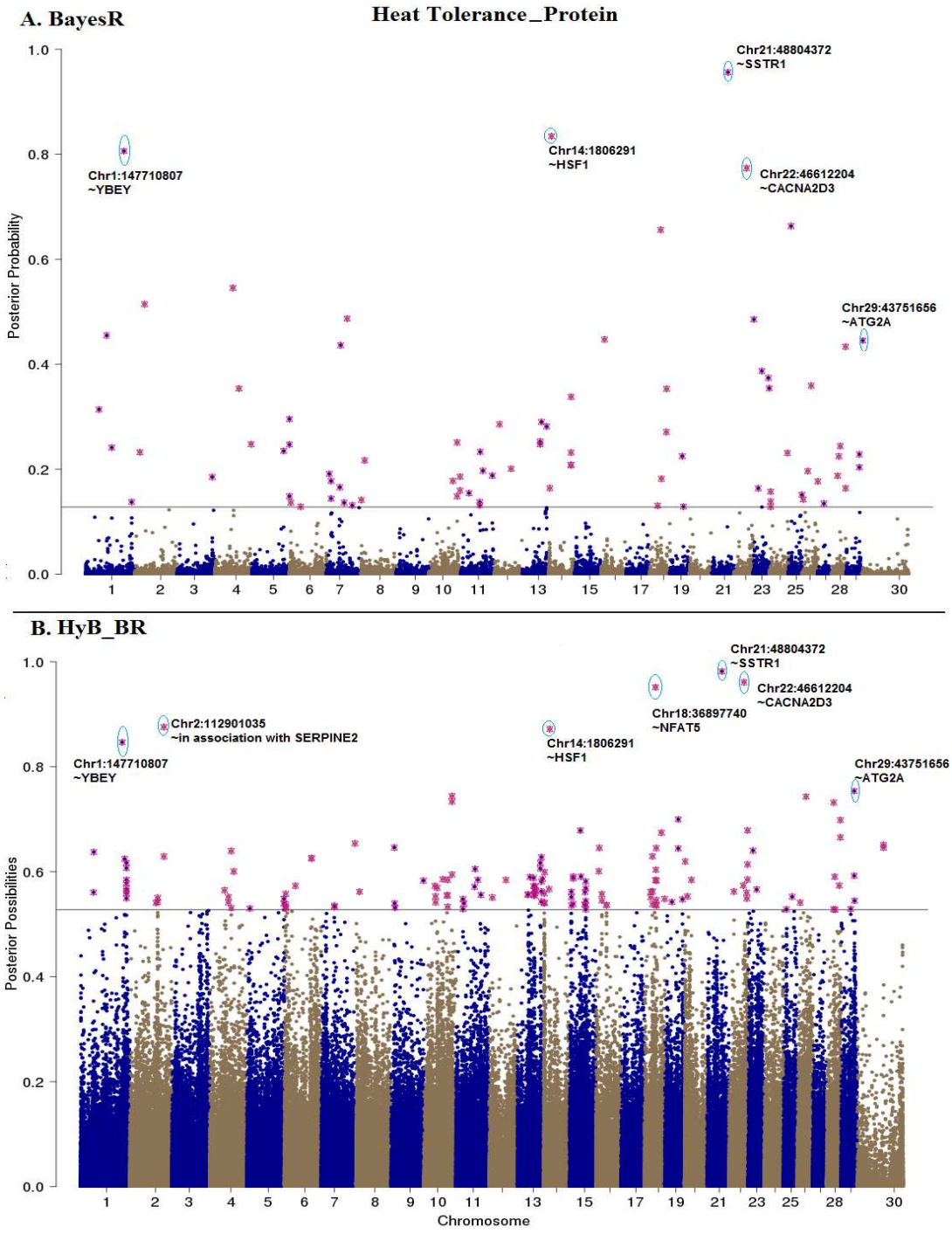


Figure 6.12. Mapping the posterior probabilities of all the variants estimated from BayesR (A.) and HyB_BR (B.) according to their positions (base pairs) across the whole chromosome related to protein yield affected by heat tolerance.

The top SNPs with highest posterior possibilities are labelled with blue circle

Table 6.7. Known genes interacting with heat stress.

Gene	BTA	Position	Traits			Description
			Fat	Milk	Protein	
YBEY	1	147710807			✓	The translation-associated heat shock genes, playing key roles in the heat-shock response of *E. coli* under heat shock stress (Rasouly *et al.* 2009; Grinwald & Ron 2013).
Unknown	2	112901035	✓	✓	✓	In association with the gene SERPINE2, which had been proved to impact the sweating rate of dairy cattle (Dikmen *et al.* 2015)
SOCS2	5	23522032	✓			Suppressor of cytokine signaling 2, might regulate with heat tolerance abatement during the dry period of dairy cattle (do Amaral *et al.* 2011).
HSF1	14	1806291			✓	Genes involved in the bovine heat stress response (Collier *et al.* 2008; Li *et al.* 2015).
DYRK3	16	4288402	✓			The dual specificity tyrosine-phosphorylation-regulated kinase 3, impacting Respiration rate (breaths per minute) in dairy cattle (Dikmen *et al.* 2015)
NFAT5	18	36897740			✓	Nuclear factor of activated T cells, simulating transcription of Heat shock protein 70 (Woo *et al.* 2002).
SSTR1	21	48804372			✓	Somatostatin receptor 1, played a role in heat stress sensing or communicating stress status between cells (Raychaudhuri *et al.* 2014).
CACNA2D3	22	46612204			✓	Methylation of the Calcium Channel-Related Gene, showing impaired behavioral heat pain sensitivity in mice and human studies (Neely *et al.* 2010).
MED17	29	1021424	✓			The mediator mutant yeast, which was temperature-sensitive (Paul *et al.* 2015).

ME3	29	8968989		✓		Malic Enzyme 3, conferring heat-stable resistance to root-knot nematodes in plants (Djian-Caporalino *et al.* 2001).
MACROD 1	29	43097815	✓			Heat shock protein 90kDa alpha (cytosolic), class A member 1, which might be association with PAR (had been proved to function heat shock response) (Petesch & Lis 2012; Di Giammartino *et al.* 2013).
STIP1	29	43108351	✓		✓	Stress inducible protein 1, was homologous to the human heat shock cognate protein 70 (hsc70)/heat shock protein 90 (hsp90) (Mizrak *et al.* 2006).
GSTP1	29	46094664	✓			Glutathione S-transferase Pi, which was reported to plays positive role under heat stress in controlling cellular toxicants and to alleviate the destructive effect on cattle (Rao *et al.* 2013).
ATG2A	29	43751656			✓	Autophagy Related 2 Homolog A, which had been referred as the Heat Stress-repressed target genes by (Niskanen *et al.* 2015).

All the listed genes are identified by HyB_BR using the variants with the largest variances $0.01 * \sigma_g^2$.

6.6 Discussion

In this paper, we demonstrated that HyB_BR (Wang *et al.* 2016) could be efficiently implemented for simultaneous prediction of genomic estimated breeding values, inference of genetic architecture, and causal mutation discovery using whole-genome sequence data. As mentioned by Wang et al (2016), HyB_BR was developed to overcome two challenges:

1) The heavy computation burden has been the main limitation of traditional MCMC Bayesian models to be applied to the whole genome sequence data with very huge data size. Therefore, Expectation-Maximization converge scheme was introduced to largely reduce the iteration times of MCMC.

2) Fast schemes (mainly including Iterative Conditional Expectation, and Expectation-Maximization algorithms) implemented for Bayesian models has been criticized due to the accuracy limitation for practical application.

HyB_BR implemented EM algorithm to quickly convergence for estimates of SNP effects and other parameters, followed by a limited number of MCMC iterations to optimize the posterior estimation for SNP effects. When implemented on the whole genome sequence data, our results indicated HyB_BR had similar accuracy of genomic prediction and precision of QTL mapping to BayesR implemented with full MCMC, but with 10 fold less computational time required. Furthermore, compared with the prediction accuracy on 600K SNP panels, we demonstrated that using the sequence data improved the accuracy of genomic prediction for some traits.

The key improvement for computational efficiency was that HyB_BR reduced the

iteration times. BayesR required a huge number of MCMC iterations, which was dependent on the size of the data. For example, on the whole genome sequence data with 16,214 animals and almost 1 million of variants, 40,000 iterations with first 20,000 as burn-in were required. For each MCMC iteration, the basis operation times were $O(mn^2)$. In comparison with BayesR, HyB_BR kept the same number of basic operations. But after the EM converged (with very small number of iterations as demonstrated by Wang et al. (Wang *et al.* 2015)), HyB_BR implemented its MCMC iterations with speed-up schemes, which could reduce the iteration number to 4,000 iterations. Therefore, the main advantage of HyB_BR was to reduce huge amount of random-walking time of BayesR to very limited number. The results from Figure 6.2 provided the evidence that HyB_BR showed up to 10 orders faster than BayesR.

In addition to the computational time, the prediction accuracy of HyB_BR for multi-breed prediction and across-breed prediction was very similar to BayesR for a range of traits with various genetic architectures, shown in Table 6.3, Table 6.4, and Table 6.5. The accuracy advantage of HyB_BR and BayesR over GBLUP for across-breeds prediction demonstrated the benefit of the non-linear Bayesian models. Also, increase in accuracy using whole genome sequence data for across-breed prediction confirmed the results from (MacLeod *et al.* 2016). That is, nonlinear model from Bayesian methods could take full advantage of the high-density data to predict the validation set, particularly when the predicted set was less related to the reference set.

There was one difference between BayesR and HyB_BR for the inference of genetic architecture, Figure 6.4. In comparison with BayesR, HyB_BR systematically detected larger proportion of SNPs in the distribution with the variance $0.001 * \sigma_g^2$; while smaller proportion SNPs in the variance $0.0001 * \sigma_g^2$.

Such results suggested less from HyB_BR for a proportion of the SNP effects. Such difference mainly happened on the traits including Fat yield, Milk yield, and protein yield. One explanation was that the speed-up scheme from HyB_BR contributed this. In detail, during MCMC sampling, the speed-up scheme helped HyB_BR to quickly decide a large proportion of SNPs to be excluded out of the model. Then, for the remaining SNPs, MCMC iterations could more accurately estimate SNP effects instead of shrinking them too hardly. The evidence could be detected in the QTLs discovery. That is, less shrinkage feature of HyB_BR for variant effects might help to increase the power of causal mutations detection shown in Figure 6.5, Figure 6.6, Figure 6.7, Figure 6.8, and Figure 6.9.

Also, QTL mapping performance of BayesR and HyB_BR was evaluated on traits with low heritability (heat tolerance and fertility). As been discussed by previous researches, heat stress (e.g. temperature, humidity) leaded to the reduction of milk production in dairy cattle (Hayes *et al.* 2003; Haile-Mariam *et al.* 2008; Nguyen *et al.* 2016). Therefore, the identification of mutations improving heat tolerance became a valuable field for investigation. Only two genetic variants (located at Chromosome 29), affecting milk production traits under heat stress, were previously detected (Hayes *et al.* 2009). Using sequence data, BayesR and HyB_BR did not find these two genes with strong signals. However, the two methods did pick up mutations in or close to twelve genes (e.g. YEBY, HSF1, DYRK3, MED17, ME3, STIP1, etc.), which have been investigated by previous studies to be in response with the heat shock stress in human, mice, or other species. In addition, HyB_BR also detected two other unknown variants. The position information from these two variants conveyed that they were very close to the regions harbored by two known genes (SERPINE2 and CACNA1D), which were investigated to impact the sweating rate and respiration rate in dairy cattle (Dikmen *et al.* 2015) . All these nine variants required the further investigation in

regards to their functions of regulation between milk productions and heat tolerance.

In the future research, the speed-up scheme of HyB_BR requires more investigation for optimization. For the current stage, the threshold of the speed-up scheme ($P(i,1) \geq 0.90$) might not be perfect for different traits and genomic data, which will hinder its flexibility for different applications. Therefore, a reasonable rule to define the threshold is required.

6.7 Conclusion

A hybrid scheme of Expectation-Maximization algorithm and MCMC sampling was implemented on the whole-genome sequence data for simultaneous genomic prediction, inference of genetic architecture inference and causal mutation identification. The accuracy of HyB_BR for multi-breed and across breed prediction for all traits was very similar to the results from BayesR (implemented with full MCMC) while requiring only 1/10 of the total running time of BayesR. In addition, HyB_BR could identify some well-known mutations (e.g. DGAT1) with the highest posterior probability, which proved its power of QTL mapping in complex traits. In the near future, HyB_BR could be implemented on very huge size of genomic data for genomic prediction and QTL mapping.

6.8 Acknowledgements

The authors acknowledge the support from Dairy Futures CRC project.

Chapter 7 General Discussion

7.1 Introduction

The aim of this thesis was to develop a computationally efficient algorithm for genomic prediction with robust prediction ability for complex traits with varying genetic architectures. The study started with the development of the emBayesR method, which introduced a Expectation-Maximization (EM) algorithm to implement the BayesR mixture of normal models for genomic prediction (In Chapter 3). The emBayesR algorithm was extended to incorporate a polygenic term and the ability to appropriately weight bull and cow phenotypes (Chapter 4). With the extension, emBayesR was implemented on the combined reference sets from Holstein and Jersey, which was used to predict the breed within these breeds and a third breed. The results from both Chapter 3 and Chapter 4 showed that emBayesR performed well on most traits but not for the traits where there were mutations of moderate or large effects affecting the underlying genetic architectures. With the aim of improving prediction accuracy across the full range of genetic architectures, in Chapter 5 and Chapter 6, a scheme was developed which hybridized the EM algorithm with limited number of MCMC iterations, called HyB_BR. In Chapter 6, it was demonstrated that this scheme was fast enough to apply to whole genome sequence data, while maintaining the accuracy of a full MCMC scheme.

7.2 The key findings

Computation time

The advantage of computational time of both emBayesR and HyB_BR could be easily detected through Chapter 3 – Chapter 6. All the methods of BayesR, emBayesR and HyB_BR required the same number of basic operations that was

$O(mn)$ for each loop. In other words, their time complexity was proportional to the number of markers (m) times the number of individuals (n). However, the difference happened with how many loops were required by each method. In detail, BayesR required 40,000 iterations with the first 20,000 iterations as burn in. For emBayesR, a very limited number (300~1,000 depending on the trait) of EM loops for the convergence was required (Chapter 3). Compared with emBayesR, HyB_BR needed the extra limited number of MCMC loops (around 4,000 iterations) (Chapter 4), which made HyB_BR require longer time than emBayesR. Compared with BayesR (the full MCMC implementation), the main contribution of HyB_BR was that it reduced large number of burn-in iterations of MCMC sampling. In practice, when moved to genomic data with very huge size, BayesR, emBayesR, HyB_BR demanded quite different running time. The curves from Figure 7.1 were the predicted time for BayesR, emBayesR, and HyB_BR according to different number of markers with the same number of individuals. For example, on the 800K SNP panels, the running time was 24 hours for emBayesR and 29 hours for HyB_BR in comparison with 330 hours required by BayesR. And when moved to huge genomic data with 10 million of markers, it took BayesR up to 5,940 hours, whereas emBayesR required 370 hours with 480 hours for HyB_BR. Therefore, for such data, BayesR required extremely huge computational time, which was intractable for the practical application. Compared with BayesR, both emBayesR and HyB_BR improved the computational speed by up to 17 fold. The time difference between emBayesR and HyB_BR could be detected as well. However, when accounting for the trade-off between running time and prediction accuracy, we could easily accept minor increase of the running time by HyB_BR.

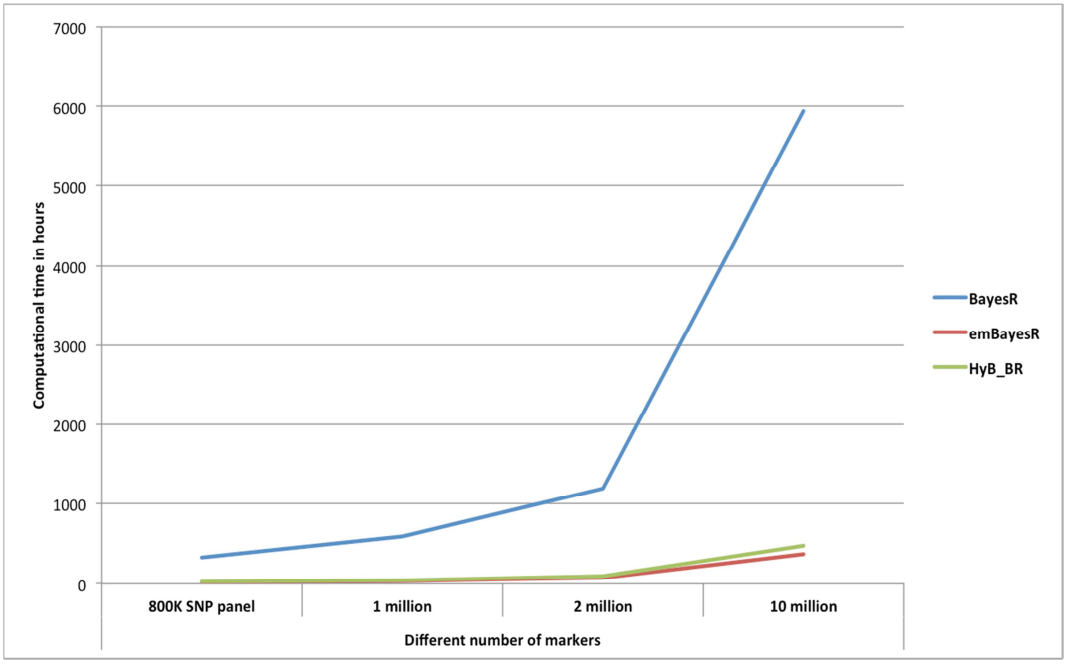


Figure 7.1. The predicted computational time required by BayesR, emBayesR and HyB_BR according to increasing number of markers with the same number of individuals (16,214) related to milk yield trait.

Predicted Error Variance correction

One key advance with emBayesR and HyB_BR was the introduction of prediction error variance (PEV) correction. Other fast methods (Meuwissen *et al.* 2009; Hayashi & Iwata 2010; Shepherd *et al.* 2010; Yu & Meuwissen 2011; Sun *et al.* 2012) hypothesized that the estimations of other SNP effects were 100% correct during the estimation of current SNP effect, which was unrealistic. Therefore, one of motivation of emBayesR was to introduce the prediction error variance to correct the calculation for current SNP effect. In theory, PEV correction should happen during each EM loop. That was, before the estimation of each SNP effect, the PEV matrix needed to be generated according to the estimated effects of all other SNPs. Such strategy could be very time consuming. To avoid the huge computational burden from the above theoretical PEV correction, there was another type of strategy, which approximated the PEV calculation (from GBLUP

models) in front of EM loops. Such approximate steps of PEV correction had been detailed in Chapter 3 and Chapter 5. The results from Chapter 3 had demonstrated the improvement of the PEV correction as shown in Table 3.5. The prediction accuracy could be improved up to 6% by the PEV calculation. The shrinkage feature illustrated from Figure 3.5 could explain this. In detail, compared with estimates of SNP effect from BayesR, emBayesR without accounting for PEV considerably shrunk SNP effects (particularly for small effects). However, with PEV correction, estimates of SNP effects with emBayesR were much closer to those from BayesR, although there was still some over-shrinkage, particularly for SNPs with small effects.

The other finding from the results of Chapter 3 and Chapter 4 was that the optimization steps of PEV correction were not perfect enough. On the traits controlled by major mutations (e.g. Fat%), emBayesR (also including the extension models in Chapter 4) had up to 5% accuracy reduction in comparison with BayesR (Table 4.4-Table 4.5). As demonstrated in Chapter 4, the extended emBayesR with PEV correction still over-shrunk SNPs with small effects too heavily especially on the bigger size of data incorporating multi-breeds of animals. The finding related to the deficiency of PEV correction was the reason that we introduced Hybrid schemes to improve the prediction performance of emBayesR (Chapter 5 and Chapter 6).

The prediction accuracy of HyB_BR in comparison with emBayesR, BayesR and GBLUP.

The performance of HyB_BR was evaluated by comparing the prediction accuracy between HyB_BR, emBayesR, BayesR, and GBLUP on two situations (Chapter 4, Chapter 5 and Chapter 6). On the one hand, the combined Holstein

and Jersey reference sets (separately genotyped with 800K SNP panel and whole genome sequence data) were used to predict the Holstein or Jersey bulls, termed as multi-breed prediction. On the other hand, the combined reference of Holstein and Jersey was also used to predict the other breed (Australian Red bulls), which was termed as across-breed prediction. The results demonstrated the following conclusions:

1) For multi-breeds prediction, GBLUP gave a consistent accuracy reduction (up to 11%) when compared with BayesR, emBayesR, and HyB_BR for the milk production traits. Especially for the traits such as Fat%, GBLUP showed much lower prediction accuracy than BayesR, emBayesR, and HyB_BR. This result conformed to the conclusions from other studies (Kemper *et al.* 2015; MacLeod *et al.* 2016). When performed for across-breed prediction, BayesR and HyB_BR gave higher accuracy (up to 18%) than GBLUP for all traits. Such results confirmed the hypothesis that the linear combination of SNPs from GBLUP could be easily broken down when the prediction happens between the breeds, which were distantly far from each other.

2) EmBayesR had 2%~7% reduction in accuracy compared with BayesR and HyB_BR for fat%. As mentioned before, the heavy shrinkage feature of emBayesR contributed this.

3) The comparison of the results between BayesR and HyB_BR showed that the prediction accuracy of HyB_BR was comparable to BayesR on all the cases. In detail, Both BayesR and HyB_BR were implemented on a range of traits with different heritabilities including milk production, fertility and heat tolerance traits in cattle. That was, there were higher heritabilities (ranging from 33% or 45%) in milk production traits, compared with the fertility (only having 3% heritability) and heat

tolerance (around 5%~9% heritabilities). Furthermore, the underlying genetic architectures of Fat% and all other traits varied. In detail, for most of the traits (e.g. milk yield, protein yields), the genetic architectures were decided by a number of variants with moderate or small effects. In comparison with the above traits, Fat% was controlled by one well-known major gene (DGAT1) with very large effects. Rather than the accuracy reduction of emBayesR on the trait Fat%, the results from HyB_BR demonstrated that HyB_BR could perform as well as BayesR on all these traits including Fat%, which demonstrated that HyB_BR was flexible for various applications.

4) Increasing the number of animals (multi-breeds reference sets) could had small but consistent accuracy improvements, in accord with the results from previous researches (Hozé *et al.* 2014; Kemper *et al.* 2015).

Furthermore, in addition to the prediction on the continuous quantitative traits of dairy cattle, the investigation on the risk prediction of seven case/control human diseases with binary 0/1 phenotypes showed HyB_BR and BayesR performed equally well. Moreover, for the traits (e.g. CD, RA, and T1D), HyB_BR and BayesR outperformed GBLUP with higher accuracy.

The impact of the sequence data in comparison with 800K SNP panels

The results from Chapter 6 demonstrated the impact of the sequence data. For the prediction of the validation sets Holstein or Jersey bulls using combined Holstein and Jersey reference sets, there was minimal accuracy gain (1%~2%) when implementing both BayesR and HyB_BR on the sequence data. However, for the validation set Australian Red distantly far from the reference set, both BayesR and HyB_BR could take full advantage of sequence data to improve the

prediction accuracy. In detail, the accuracy using BayesR and HyB_BR on sequence data was up to 13% higher than on 800K SNP panels. On the same situation, GBLUP could only had very minimal increase (or even reduction for Fat Yield trait) when using the sequence data. One explanation was that the linear combination of BLUP model shrunk all the SNPs with small effects, which therefore did not allow true causal mutations to had large effects, while some others had zero effects. On the high-density SNP panels (e.g. 800K) or whole genome sequence data, it was unreasonable for BLUP model to define all the SNPs with small effects. Therefore, a number of such little errors from BLUP models added up to the reduction in prediction accuracy.

The causal variants discovery across 800K SNP panel and whole genome sequence data

Both BayesR and HyB_BR could calculate posterior possibilities for each SNP effect following in the distribution with largest variance. Therefore, they could easily be used to pick up top SNPs with large effects. HyB_BR and BayesR were implemented on both 800 SNP panels (in Chapter 5) and whole genome sequence data (in Chapter 6) for QTL mapping. The top SNPs with highest non-zero posterior possibilities were picked up.

In regards with causal mutation identification, there were several key findings:

1) In comparison with BayesR, HyB_BR showed the similar precision of detecting the causal mutations in or near known genes affecting milk production, fertility, and heat tolerance. In addition, when compared with BLUP models, both BayesR and HyB_BR were more powerful to detect some causal mutations, which were in the region of some famous genes. For example, both BayesR and HyB_BR found the variant located at 103302351 of Chromosome 11, which was

in the region of well-known gene PAEP impacting milk yield and protein yield (Chapter 6).

2) The QTL mappings of BayesR and HyB_BR on 800K SNP panels and imputed sequence data (Chapter 5 and 6) demonstrated that BayesR and HyB_BR could detect many causal mutations on the sequence data, which did not appear on the 800K SNP panels. For example, for several causal mutations in the region of famous genes including CSN2/CSN3 (MacLeod *et al.* 2016), KRT19 (Chamberlain *et al.* 2015), FASN (Roy *et al.* 2006), and SCD (Mele *et al.* 2007), which had been proved to impact the milk production traits by previous studies, HyB_BR and BayesR could detect them from whole genome sequence data but not from 800K SNP panels.

3) Another important discovery was that HyB_BR and BayesR could detect nine novel variants affecting milk productions but also being in response with heat tolerances. One of them (HSF1 at Chromosome 14 with the position 1806291) had been reported to be associated with heat tolerance in Chinese dairy cattle (Li *et al.* 2015). DYRK3 (Dikmen et al. 2015) was reported to affect Respiration rate (breaths per minute) in dairy cattle. Two variants with unknown names might be associated with two genes (SERPINE2 and CACNA1D), which had been reported to impact the sweating rate and respiration rate of dairy cattle (Dikmen *et al.* 2015). Six other variants including YEBY (Rasouly *et al.* 2009; Grinwald & Ron 2013), MED17 (Paul *et al.* 2015), ME3 (Djian-Caporalino *et al.* 2001), MACROD1 (Petesch & Lis 2012; Di Giammartino *et al.* 2013), STIP1 (Mizrak *et al.* 2006), and ATG2A (Niskanen *et al.* 2015) and had been detected to be in response with heat shock stress in human, mice or other species. Nevertheless, none of these genes was previously reported to impact milk production traits in dairy cattle. Therefore, the results might be valuable in the further studies for detecting the function of

these variants in the interaction between milk production traits and heat shock stress of dairy cattle.

7.3 Future investigation

In the near future, optimally using whole genome sequence data with millions of variants and increasing number of individuals for genomic prediction would benefit from the following research regarding to three aspects:

1) Incorporate the biological priors into the model. To date, Macleod et al., (2016) developed a new method (termed BayesRC) to group markers into two or more categories according to their biology properties (i.e. variant annotation, candidate gene lists and known causal variants). Then, BayesRC implemented a similar approach to BayesR with a uniform prior, but processing each category independently across the MCMC iterations. The results demonstrated the accuracy improvement of BayesRC in comparison with BayesR, especially when the reference sets for genomic prediction were not closely related to the validation sets. And, it had been detected that the combination of biological information could help to improve the precision of causal mutations identification. Therefore, we had confidence that incorporating biological information into HyB_BR models could help to improve the prediction performance of HyB_BR.

2) Apply HyB_BR for multi-traits genomic prediction. A number of previous researches (Calus *et al.* 2008; Hayashi & Iwata 2010; Aguilar *et al.* 2011; Jiao *et al.* 2012; Schulthess *et al.* 2016) on multi-traits genomic predictions aimed to improve the prediction accuracy of the traits, which were difficult or expensive to record. In cattle, some economically important traits (e.g. Feed efficiency) were quite hard to be recorded due to the properties, which were sex-linked or expressed late in life. To improve the prediction accuracy of these traits, multiple traits (in genetically relationship) prediction could be used. In practice, the results

from previous researches provided the evidence that the multiple traits based genomic prediction could improve the prediction accuracy of rarely recorded traits (with missing phenotypes). In addition, HyB_BR implemented for multi-traits prediction could be even more usefully when including gene expression (RNASeq data) or protein expression as additional traits.

3) Speed up HyB_BR even further. As genomic data increases to very huge size (especially the dramatically growing number of individuals), HyB_BR might also be threatened with heavy computational problems in terms of computational time and memory cost. The key part was due to the calculation of $tr(\mathbf{E^{-1}Z_i Z_i' E^{-1} PEV})$) (linearly scaled on the number of markers and quarterly on the number of individuals), which account for more than half of the total running time of EM module. Therefore, one further investigation was to implement HyB_BR with multi-threaded programming technology. Actually, some investigations at testing stage had already been conducted in our research team.

7.4 Conclusions

The main outcome of this thesis was the design and development of a computationally efficient and robust method (HyB_BR) for the prediction on unknown phenotypes, genetic architecture dissection, and genes identifications. The computational time demonstrated HyB_BR could be more than 17 times faster than BayesR. The prediction accuracy of HyB_BR across a wide range of traits covering milk productions, fertility, and heat tolerance proved that HyB_BR could perform as well as BayesR. In addition, the gene identifications of HyB_BR on all the above traits also demonstrated the precision of HyB_BR for QTL mapping. In summary, the efficient speed, flexible prediction performance, and precise QTL mapping all shed light on the application of HyB_BR to the whole-genome sequence data with very large number of variants and animals.

Chapter 8 Appendix I

(For Chapter 2)

8.1 File S1 - Non Bayesian Penalized regression and orthogonal linear regression models for genomic prediction

Penalized regression, of which Least Angle Regression (LAR) was the most widely used (Usai *et al.* 2009) shrinks the estimation towards zero or other fixed points when minimizing the residual sum of squares. For example, the least absolute shrinkage and selection operator (LASSO), a variant of LAR, implements the LARS algorithm to estimate SNP effects, subsequently followed with a cross-validation step to select the best subset of SNPs (Usai *et al.* 2009). Unlike BLUP, LASSO sets a subset of SNPs to had zero effects while others have effects shrunk towards zero. For traits where there are some mutations of large effect, LASSO could result in a higher accuracy of genomic prediction. However, when the number of markers dramatically exceeds the number of records, LASSO loses its superiority over BLUP (Usai *et al.* 2009). The LARS algorithm is also computationally very intensive, making it impractical for large data sets.

Similar to BLUP models, orthogonal linear regression models (Jannink 2010; Macciotta *et al.* 2010; Colombani *et al.* 2012) are also linear models, in that the predicted breeding values or future phenotypes are linear combinations of effects. However, rather than considering individual markers as in the BLUP models, orthogonal linear models regress the phenotypes on linear combination of “components”, constructed from the markers. In detail, there are two key steps for orthogonal linear regression methods: 1) build orthogonal linear combinations (the

components) of the markers according to the correlation among the markers, or among the markers and the phenotypes; 2) regress the phenotypes on to a small number of the above linear combinations. Principal component (PC) (Solberg *et al.* 2009) and Partial Least Squares (PLS) (Colombani *et al.* 2012) are two typical approaches of orthogonal linear regression models, and both have been used for genomic prediction. The main difference between them is the way in which the orthogonal linear combinations are constructed. Specially, PLS builds the combinations using the markers in a maximum correlation with the phenotypes. PC derives the components by calculating the eigenvalues or the regression sum of square (SS) contribution from the genotype data. The orthogonal linear regression approaches reduce the dimensionality of genomic data, and therefore improve the computational efficiency of genomic prediction. However, the dimensionality reduction could cause some loss of genomic prediction accuracy. As the density of genomic data increases, the potential advantage of selecting individual SNP in higher linkage disequilibrium to use in the genomic prediction equation might be lost in the PLS or PC approaches (Solberg *et al.* 2009; Colombani *et al.* 2012). Further investigation is required before applying these approaches to higher density SNP data or whole genome sequence data.

8.2 File S2 - The description of the model and prior density function

In genomic prediction, all the SNPs were fitted simultaneously:

$$\mathbf{y} = \mu + \mathbf{Zu} + \mathbf{e}$$

Where, $\mathbf{y}$ was the $n \times 1$ vector of phenotypes, n was the number of individuals;

μ was the mean;

$\mathbf{u}$ was the $m \times 1$ vector of SNP effects, m was the number of SNPs;

$\mathbf{e}$ was $n \times 1$ vector of random residual error terms, following the distribution $\boldsymbol{e} \sim N(0, \sigma_e^2)$, where σ_e^2 was the error variance.

$\mathbf{Z}$ was the $n \times m$ matrix of standardised genotypes, where the genotype for the i^{th} individual at the j^{th} SNP, x_{ij} was coded as 0=*aa*, 1=*Aa*,2=*AA*, where *a* and *A* were the two alleles at the SNP, and genotypes were standardized as $z_{ij=}\frac{x_{ij}-2p_i}{\sqrt{p_i(1-p_i}}$, where p_i was the allele frequency of the i^{th} SNP.

For the prior density function $p(\mathbf{u})$, the prior density function $p(\mathrm{u_i}|\omega,\cdots)$ for each SNP *i* conditional on the hyper parameters could be represented as:

$$p(u_i|\omega) = \int N\left(u_i\middle|0,\sigma_{u_i}^2\right)p\left(\sigma_{u_i}^2\middle|\omega\right)p(\omega)d\sigma_{u_i}^2$$

where, $N\left(\mathrm{u_i}\middle|0,\sigma_{\mathrm{u_i}}^2\right)$ was the prior normal distribution of SNP effects conditional on its variance $\sigma_{\mathrm{u_i}}^2$;

$p\left(\sigma_{\mathrm{u_i}}^2\middle|\omega\right)$ was the prior distribution for the variance of SNPs conditional on its hyper parameters ω;

$p(\omega)$ was the density function of the hyper parameters set ω.

The density function $p(\mathrm{u_i}|\omega,\cdots)$ encompasses three stages of prior assumption detailed in Table 2.1:

1) The first stage defines the normal prior for each SNP effect (where each SNP was assumed to come from a normal distribution with it's own variance) conditional on the genetic variance $\sigma_{\mathrm{u_i}}^2$, which could be written as $N\left(\mathrm{u_i}\middle|0,\sigma_{\mathrm{u_i}}^2\right)$. The difference between BLUP and Bayesian alphabet families happens at this stage. In detail, BLUP assumed the common genetic variances $\sigma_{\mathrm{u_i}}^2$ across all the SNPs written as $\sigma_{\mathrm{u_i}}^2 = \sigma_{\mathrm{u_j}}^2 = \sigma_{\mathrm{u}}^2$. On the contrary, Bayesian alphabet families assign the variance $\sigma_{\mathrm{u_i}}^2$ specific to each SNP i or each group of SNPs.

2) The second stage (applicable for Bayesian regression model) relates to the

density function $p(\sigma^2_{u_i}|\omega)$ which defines the variance $\sigma^2_{u_i}$ conditional on the hyper parameters ω. Different assumption for the distribution of the variances $p(\sigma^2_{u_i}|\omega)$ generate a range of Bayesian regression models. For example, BayesA assumed that the variance $\sigma^2_{u_i}$ for each SNP was independently draw from the inverted chi-square distribution, resulting in a Students *t* distribution at the level of the SNP; BayesB assumed the small proportion (π) of the variance $\sigma^2_{u_i}$ with inverted chi-square distribution while allowing others to be zero (1-π).

3) For the third stage (only used in a small number of Bayesian genomic prediction models to date), the density function $p(\omega)$ defines the prior distributions of hyper parameters for the variance. For example, BayesD, which could be treated as a modified BayesB method, assumed the hyper parameter S follows gamma distribution.

With the above three-stage prior assumption, the conditional density function $p(u_i|\omega,\cdots)$ conveys two types of information: the proportion of SNPs effects around zero area and the kurtosis feature of the density (the thickness of the tail). Such information from the priors classifies the Bayesian regression models into four groups shown in Figure 2.2: normal priors (black dotted curve; e.g, BLUP), thick tail such as *t* distribution (red curve; e.g. BayesA), Spike-around-zero & slabs such as the mixture of two normal priors with small variances (small but nonzero) and relatively large variances (green curve; e.g. BayesSSVS), and Spike-at-zero &Slabs such as the mixture distributions with zero variances and non-zero variances (purple curve; e.g. BayesB/BayesR).

8.3 File S3 - An example of deriving the conditional prior density function

BayesA (Meuwissen *et al.* 2001) assumed two stages prior assumption for SNP effects shown in Table 2.1:

$$u_i \sim \mathrm{N}(0, \sigma^2_{\mathrm{u_i}}),\ \sigma^2_{\mathrm{u_i}} \sim \chi^{-2}(v, \mathrm{S}^2) \text{ with fixed hyper-parameters } \mathrm{v}, S.$$

Then, the prior density function $p(\mathrm{u_i}|\mathrm{v}, S)$ for BayesA could derived according to the following steps:

$$\begin{aligned} p(u_i|v,S) &= \int_0^{+\infty} N\left(u_i|0,\sigma^2_{u_i}\right) p\left(\sigma^2_{u_i}|v,\mathrm{S}^2\right) d\sigma^2_{u_i} \\ &\propto \int_0^{+\infty} (\sigma^2_{u_i})^{-\frac{1}{2}} \times \exp(-\frac{{u_i}^2}{2\sigma^2_{u_i}}) \times (\sigma^2_{u_i})^{-1-\frac{v}{2}} \times \exp(-\frac{v\mathrm{S}^2}{2\sigma^2_{u_i}})\ ) d\sigma^2_{u_i} \\ &\propto \int_0^{+\infty} (\sigma^2_{u_i})^{-1-\frac{v+1}{2}} \exp(-\frac{{u_i}^2+v\mathrm{S}^2}{2\sigma^2_{u_i}}) \\ &\propto\ (1+\frac{\sigma^2_{u_i}}{v\mathrm{S}^2})^{-\frac{v+1}{2}} \text{ (the } t \text{ distribution density)} \end{aligned}$$

Based on the above deriving process, the conditional prior density of BayesA on the hyper-parameters follows a *t* distribution $t(u_i|0, v, S)$.

8.4 File S4 - The detailed fast version of Bayesian algorithms

VanRaden et al. (VanRaden 2008) proposed the methods termed nonlinear A and B to mimic the nonlinear shrinkage of BayesA and BayesB. Jacobi iteration was implemented on nonlinear A and B to be approximations of BayesA and BayesB.

Meuwissen (Meuwissen *et al.* 2009) described a method termed fastBayesB by using ICE in the BayesLASSO model. FastBayesB iteratively calculated each SNP's posterior mean, conditioning on current estimates of all the other SNPs as

if they were true effects. With the same Bayesian model, a method dubbed EmBayesB applied an expectation- maximization (EM) algorithm by maximizing a joint posterior probability based on the prior distribution of SNP effects.

Hayashi and Iwata (Hayashi & Iwata 2010) proposed the method termed em_BSR, which introduced SNP weights to define the association between SNP and the traits. Then, treating such association as missing data, em_BSR performed partial maximization so as to implement EM for both BayesA and BayesB models.

Yu and Meuwissen (Yu & Meuwissen 2011) described a method termed MixP, which first used the pareto principle method (80:20 rules) to define the mixture of two normal distributions with big and small variance (similar to BayesSSVS models). Then under such mixture prior, MixP implemented the ICE algorithm to approximate the mean of the SNP effects.

Sun et al. (Sun *et al.* 2012) developed fastBayesA, which implemented an EM algorithm instead of MCMC on the BayesA model. Unlike other EM methods, fastBayesA maximized the posterior estimation for all the SNPs simultaneously using BLUP model.

Chapter 9 Appendix II

(For Chapter 3)

9.1 File S1– Calculation of $\mathbf{P_{ik}} = E(\mathbf{b_{ik}}|\mathbf{y}, \widehat{\mathbf{Pr}_k})$

In the expectation step of the EM algorithm we require the $E_{b|y}$ of equation 6b. This requires the $E(\mathbf{b_{ik}}|\mathbf{y}, \widehat{\mathbf{Pr}_k})$ which was derived in this appendix.

The model was $\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Z}_i g_i + \mathbf{u} + \mathbf{e}$,

Then,

$$\begin{aligned} \boldsymbol{E}_{\mathbf{u}}(b_{ik}|\mathbf{y}, \widehat{\Pr}_k) &= p\left(b_{ik} = 1 \middle| \mathbf{y}, \widehat{\Pr}_k\right) \\ &= \frac{p(\mathbf{y}|b_{ik}=1) \times p\left(b_{ik}=1|\widehat{\Pr}_k\right)}{p(\mathbf{y})} \\ &\propto p(\mathbf{y}|b_{ik} = 1) \times p\left(b_{ik} = 1|\widehat{\mathbf{Pr}_k}\right) \qquad \text{(A1)} \end{aligned}$$

where,

$p\left(b_{ik} = 1|\widehat{\mathbf{Pr}_k}\right) = \widehat{\mathbf{Pr}_k}$ **, and**

$p(\mathbf{y}|b_{ik} = 1) = \frac{1}{\sqrt{|\mathbf{W_k}|}}\exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{1}_n\mu - \mathbf{u})'\mathbf{W_k}^{-1}(\mathbf{y} - \mathbf{1}_n\mu - \mathbf{u})\right)$,

so $logp(\mathbf{y}|b_{ik} = 1) = -0.5\,(\,log|\mathbf{W_k}| + ((\mathbf{y} - \mathbf{1}_n\mu - \mathbf{u})'\mathbf{W_k}^{-1}(\mathbf{y} - \mathbf{1}_n\mu - \mathbf{u}))$

based on $(y - \mathbf{1}_n\mu - \mathbf{u})|(b_{ik} = 1) \sim N(0, \mathbf{W_k})$, and $\mathbf{W_k} = \mathbf{Z_i}\mathbf{Z_i}'\sigma_k^2 + \mathbf{I}\sigma_e^2$.

Therefore,

$$logl_{ik} = log\, p\left(b_{ik} = 1 \middle| \mathbf{y}, \widehat{\Pr}_k\right) = logp(\mathbf{y}|b_{ik} = 1) + logp\left(b_{ik} = 1|\widehat{\mathbf{Pr}_k}\right) + constant$$

The $constant$ appears on both denominator term and numerator term of equation (A7), and therefore could be ignored.

The expression above for $logp(\mathbf{y}, |b_{ik} = 1)$ involves the unknown $\mathbf{u}$. Therefore, we take the expectation over $\mathbf{u}|\mathbf{y}$. That was,

$$logp(\mathbf{y}|b_{ik} = 1) = -0.5\boldsymbol{E}_{\mathbf{u}|\mathbf{y}}\left\{lo\, g|\mathbf{W_k}| + (\mathbf{y} - \mathbf{1}_n\mu - \mathbf{u})'\mathbf{W_k}^{-1}(\mathbf{y} - \mathbf{1}_n\mu - \mathbf{u})\right\}$$

Only the quadratic form $Q = (\mathbf{y} - \mathbf{1}_{\boldsymbol{n}}\mu - \mathbf{u})'\mathbf{W}_{\mathrm{k}}{}^{-1}(\mathbf{y} - \mathbf{1}_{\boldsymbol{n}}\mu - \mathbf{u})$ of $logp(\mathbf{y}|\mathrm{b_{ik}} = 1)$ involves $\mathbf{u}$. Therefore, apply Searle's expectation rule for Q as follows:

$$\mathrm{E}_{\widehat{\mathbf{u}}}Q = (\mathbf{y} - \mathbf{1}_{\boldsymbol{n}}\mu - \widehat{\mathbf{u}})'\mathbf{W}_{\mathrm{k}}{}^{-1}(\mathbf{y} - \mathbf{1}_{\boldsymbol{n}}\mu - \widehat{\mathbf{u}}) + \mathrm{tr}(\mathbf{W}_{\mathrm{k}}{}^{-1}PEV((\widehat{\mathbf{u}}))$$

Hence, $logp(\mathbf{y}, |\mathrm{b_{ik}} = 1) = -0.5\{lo\ g|\mathbf{W}_{\mathrm{k}}| + \mathrm{E}_{\widehat{\mathbf{u}}}Q\}$

$$= -0.5\left\{lo\ g|\mathbf{W}_{\mathrm{k}}| + \mathbf{y}^{\dagger'}\mathbf{W}_{\mathrm{k}}{}^{-1}\,\mathbf{y}^{\dagger} + \mathrm{tr}(\mathbf{W}_{\mathrm{k}}{}^{-1}PEV((\widehat{\mathbf{u}}))\right\}$$

where, $\mathbf{y}^{\dagger} = (\mathbf{y} - \mathbf{1}_{\boldsymbol{n}}\mu - \widehat{\mathbf{u}})$.

Although $\mathbf{W}_{\mathrm{k}}$ was a $n \times n$ matrix. the calculation of $log|\mathbf{W}_{\mathrm{k}}|$ and $\mathbf{W}_{\mathrm{k}}{}^{-1}$ could be simplified by using the Woodbury identity so that

$$\mathbf{W}_{\mathrm{k}}{}^{-1} = \left(\mathbf{Z_i}\mathbf{Z_i}'\sigma_{\mathrm{k}}^2 + \mathbf{I}\sigma_{\mathrm{e}}^2\right)^{-1} = \sigma_{\mathrm{e}}^{-2}\left(I - \frac{\mathbf{Z_i}\mathbf{Z_i}'\sigma_{\mathrm{k}}^2}{\sigma_{\mathrm{k}}^2\mathbf{Z_i}'\mathbf{Z_i} + \sigma_{\mathrm{e}}^2}\right) \quad (A2)$$

$$|\mathbf{W}_{\mathrm{k}}| = \sigma_{\mathrm{e}}^{(2\mathrm{n}-2)}(\sigma_{\mathrm{k}}^2\mathbf{Z_i}'\mathbf{Z_i} + \sigma_{\mathrm{e}}^2),$$

so,

$$log|\mathbf{W}_{\mathrm{k}}| = (2\mathrm{n} - 2)\mathrm{lo\,g}\,\sigma_{\mathrm{e}}^2 + log\left(\sigma_{\mathrm{k}}^2\mathbf{Z_i}'\mathbf{Z_i} + \sigma_{\mathrm{e}}^2\right) \quad (A3)$$

Such transformation could transfer the inverse calculation of a large matrix $\mathbf{W}_{\mathrm{k}}$ to the multiplication of the vectors, which could reduce the cost for matrix calculation.

Therefore, substitute (A3) and (A4) into $logp(\mathbf{y}|\mathrm{b_{ik}}, \widehat{\mathbf{u}})$ as follow:

$$logp(\mathbf{y}|\mathrm{b_{ik}} = 1) = -0.5\left\{(\mathrm{n} - 1)\mathrm{lo\,g}\,\sigma_{\mathrm{e}}^2 + \mathrm{log}\left(\sigma_{\mathrm{k}}^2\mathbf{Z_i}'\mathbf{Z_i} + \sigma_{\mathrm{e}}^2\right)\right\}$$
$$-0.5\left\{\left(\mathbf{y}^{*'}\mathbf{y}^{*}\right)\sigma_{\mathrm{e}}^{-2} - \left(\mathbf{y}^{*'}\mathbf{Z_i}\right)^2\sigma_{\mathrm{k}}^2\sigma_{\mathrm{e}}^{-2}/\left(\sigma_{\mathrm{k}}^2\mathbf{Z_i}'\mathbf{Z_i} + \sigma_{\mathrm{e}}^2\right)\right\}$$
$$-0.5\left\{\mathrm{tr}\left(\mathrm{PEV}(\widehat{\mathbf{u}})\right)\sigma_{\mathrm{e}}^{-2} - \mathrm{tr}\left(\mathbf{Z_i}\mathbf{Z_i}'\mathrm{PEV}(\widehat{\mathbf{u}})\right)\sigma_{\mathrm{k}}^2\sigma_{\mathrm{e}}^{-2}/\left(\sigma_{\mathrm{k}}^2\mathbf{Z_i}'\mathbf{Z_i} + \sigma_{\mathrm{e}}^2\right)\right\} \quad (A4)$$

Then,

$$logl_{ik} = logp(\mathbf{y}|\mathrm{b_{ik}} = 1) + logp\left(\mathrm{b_{ik}} = 1|\widehat{\mathbf{Pr}_{\boldsymbol{k}}}\right)$$

$$= log\mathrm{Pr_k} - 0.5\{2(\mathrm{n}-1)\log\sigma_e^2 + \log \mathrm{V}\}$$

$$-0.5\left\{\left(\mathbf{y}^{\dagger\prime}\mathbf{y}^{\dagger}\right)\sigma_e^{-2} - \left(\mathbf{y}^{\dagger\prime}\mathbf{Z_i}\right)^2\sigma_k^2\sigma_e^{-2}/V\right\}$$

$$-0.5\left\{\mathrm{tr}\left(\mathrm{PEV}(\widehat{\mathbf{u}})\right)\sigma_e^{-2} - \mathrm{tr}\left(\mathbf{Z_i}\mathbf{Z_i'}\mathrm{PEV}(\widehat{\mathbf{u}})\right)\sigma_k^2\sigma_e^{-2}/V\right\} \quad (A5)$$

where, $\mathbf{y}^{\dagger} = \mathbf{y} - \mathbf{1_n}\mu - \widehat{\mathbf{u}}$, $V = \sigma_k^2\mathbf{Z_i'Z_i} + \sigma_e^2$ and n was the number of animals. $\mathrm{PEV}(\widehat{\mathbf{u}})$ ($n \times n$ symmetric matrix) could be approximated by $\mathrm{PEV}(\widehat{u^*})$ as derived in File S2 (Chapter 10) and could be calculated based on GBLUP, outside the iterations of EM algorithm. The term $\mathrm{tr}\left(\mathbf{Z_i}\mathbf{Z_i'}\mathrm{PEV}(\widehat{\mathbf{u}})\right)$ means to add up the diagonal elements of symmetric matrix. In other words, we just need to calculate and then add up the diagonal elements of the multiplication of $\mathbf{Z_i}\mathbf{Z_i'}$ (also a $n \times n$ symmetric matrix) and $\mathrm{PEV}(\widehat{\mathbf{u}})$. Because $\mathrm{tr}\left(\mathbf{Z_i}\mathbf{Z_i'}\mathrm{PEV}(\widehat{\mathbf{u}})\right)$ and $\mathrm{tr}\left(\mathrm{PEV}(\widehat{\mathbf{u}})\right)$ did not change each iterations, they could be calculated once and stored in front of the EM steps.

With the expression for $\mathrm{logl_{ik}} = logp\left(\mathrm{b_{ik}} = 1\middle|\mathbf{y}, \widehat{\mathbf{Pr}_k}\right)$, we could now calculate the probability that each SNP was in one of four normal distributions:

$$P_{ik} = \boldsymbol{E}_{\mathbf{u}}\left(\mathrm{b_{ik}}\middle|\mathbf{y}, \widehat{\mathrm{Pr}_k}\right) = \frac{\exp(logl_{ik})}{\sum_{k=1}^{4}\exp(logl_{ik})} \quad (A6)$$

9.2 File S2– PEV calculation from GBLUP

In the EM algorithm, we would need the prediction error variance of $\hat{\mathbf{u}}$ (PEV). $\hat{\mathbf{u}}$ is the sum of the estimated effects of all the SNP multiplied by the genotypes for each animal but we approximate its PEV by assuming it is normally distributed and therefore could be calculated by the GBLUP model. That was,

For *n* animals, the phenotype could be modelled as a simplified model:

$$\mathbf{y} = \mathbf{1}_{\boldsymbol{n}}\mu + \mathbf{u}^{*} + \mathbf{e}$$

Where, $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_{\mathrm{e}}^{2})$

$\mathbf{u}^{*}$ is the breeding value for all of the animals($\mathbf{u}^{*} = \mathbf{Zg}$), $\mathbf{u}^{*} \sim N(\mathbf{0}, \mathbf{G}\sigma_{\mathrm{g}}^{2})$**.** Here **G** is the genomic relationship matrix (Yang *et al.* 2010), written as $\mathbf{G} = \frac{\boldsymbol{WW'}}{\boldsymbol{m}}$, with $\boldsymbol{w_{ij}} = \frac{\boldsymbol{z_{ij}} - p_i}{\sqrt[2]{2\sum p_i(1-p_i)}}$, m is the number of SNPs, and p_i is the frequency of the second allele "1" for SNP i**.**

Then, the prediction error variance of $\widehat{\mathbf{u}^{*}}$ is:

$$\mathrm{PEV}(\widehat{\mathbf{u}^{*}}) = \mathrm{Var}(\mathbf{u}^{*} - \widehat{\mathbf{u}^{*}}) = (\mathbf{G}^{-1}\sigma_{\mathrm{g}}^{-2} + \mathbf{I}\sigma_{\mathrm{e}}^{-2})^{-1} \qquad (A7)$$

In emBayesR, we also use the model

$$\mathbf{y} = \mathbf{1}_{\boldsymbol{n}}\mu + \mathbf{u} + \mathrm{Z_i g_i} + \mathbf{e}$$

Where $\mathbf{u} = \mathbf{u}^{*} - \mathrm{Z_i g_i}$. But we assume $\mathrm{PEV}(\hat{\mathrm{u}}) = \mathrm{PEV}(\widehat{\mathrm{u}^{*}})$. In fact, u^{*} differs from u. That is, u^{*} includes the effect of the current SNP ($\mathrm{Z_i g_i}$) whereas u does not. Consequently, $\mathrm{PEV}(\hat{\mathrm{u}}) < \mathrm{PEV}(\widehat{\mathrm{u}^{*}})$. However, the difference should be small because the effect of each SNP is small and the estimated effect is even smaller because it is shrunk especially in the GBLUP model.

Chapter 10 Appendix III

(For Chapter 5)

File S1 - PEV calculation from GBLUP

The prediction error variance $\mathrm{PEV}(\mathbf{e}^*)$ was derived using GBLUP under the data model as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}^* + \mathbf{Wv} + \mathbf{e},$$

here, $\mathbf{u}^* = \mathbf{Zg}, \mathbf{u}^* \sim N(0, \mathbf{G}\sigma_g^2)$. In theory, $\mathbf{e}^* = \mathbf{e} + \mathbf{Z}_{\mathrm{i}}\mathrm{g}_{\mathrm{i}}$, therefore, $\mathrm{PEV}(\mathbf{e}^*) \neq \mathrm{PEV}(\mathbf{e})$. However, the difference should be small since the estimated effect from GBLUP model was shrunk to very small number. Therefore, $\mathrm{PEV}(\mathbf{e})$ could be treated as the approximated calculation of $\mathrm{PEV}(\mathbf{e}^*)$.

The calculation of $\mathrm{PEV}(\mathbf{e})$ was described as follows:

$$\mathrm{PEV(e)} = \mathrm{Var}(\mathbf{e} - \hat{\mathbf{e}}) = \mathrm{Var}(\mathbf{e} - \mathbf{R}'\mathbf{V}^{-1}\mathbf{y})$$

where, $\mathbf{V}$ was the variance of phenotype $\mathbf{y}$; $\mathbf{V} = \mathbf{G}' + \mathbf{Q} + \mathbf{R}$ ($\mathbf{G}'$ was the variance and co-variance matrix of $\mathbf{u}$ by SNPs, $\mathbf{Q}$ was the variance and co-variance matrix by polygenes, **R** was the variance and co-variance error matrix).

Therefore,

$$\begin{aligned}
\mathrm{PEV}(\mathbf{e}) &= \mathrm{Var}(\mathbf{e} - \mathbf{RV}^{-1}\mathbf{y}) \\
&= \mathrm{Var}(\mathbf{e}) + \mathrm{Var}(\mathbf{RV}^{-1}\mathbf{y}) - 2\mathrm{cov}(\mathbf{u}, \mathbf{RV}^{-1}\mathbf{y}) \\
&= \mathbf{R} + \mathbf{RV}^{-1}\mathrm{Var}(\mathbf{y})\mathbf{V}^{-1}\mathbf{R} - 2\mathbf{RV}^{-1}\mathbf{cov}(\mathbf{u}, \mathbf{y}) \\
&= \mathbf{R} + \mathbf{RV}^{-1}\mathbf{R} - \mathbf{2RV}^{-1}\mathbf{R} \\
&= \mathbf{R} - \mathbf{RV}^{-1}\mathbf{R} \\
&= [(\mathbf{Q} + \mathbf{G}')^{-1} + \mathbf{R}^{-1}]^{-1}
\end{aligned}$$

Substitute it with $\mathbf{G}' = \mathbf{G}\sigma_g^2$, $\mathbf{R} = \mathbf{E}\sigma_e^2, \mathbf{Q} = \mathbf{A}\sigma_a^2$, we get:

$$\mathrm{PEV} = (\mathbf{E}^{-1}\sigma_e^{-2} + (\mathbf{G}\sigma_g^2 + \mathbf{A}\sigma_a^2)^{-1})^{-1} \quad \text{(S1)}$$

Afterwards, such PEV matrix would be used in File S2 (Chapter 10).

File S2 - Calculation of $\mathbf{P(i,k)}$

The parameter $P(i,k)$ defines the probability that each SNP i follows in the k^{th} normal distribution ($k = 1,2,3,4$) conditional on the data. An important part of the EM algorithm for the mixture BayesR model was estimating:

$$P(i,k) = p(b(i,k) = 1|\mathbf{y}, Pr_k, \sigma_e^2, \boldsymbol{\beta}, \mathbf{v}).$$

Suppressing the parameters not involving $b(i,k)$, we get:

$$\begin{aligned} P(i,k) &= p(b(i,k) = 1|\mathbf{y}, Pr_k, \sigma_e^2, \boldsymbol{\beta}, \mathbf{v}) \\ &\propto p(b(i,k) = 1|\mathbf{y}, Pr_k) \\ &\propto p(\mathbf{y}|b(i,k) = 1)p(b(i,k) = 1|Pr_k) \end{aligned}$$

Under the model (1a), we introduce the "missing data" $e^* = \mathbf{y} - \mathrm{X}\boldsymbol{\beta} - \mathbf{u} - \mathbf{W}\mathbf{v} = \mathbf{Z_i}g_i + e$. Therefore, $e^* \sim N(0, \mathbf{H_k})$, with $\mathbf{H_k} = \mathbf{Z_i}\mathbf{Z_i}'\sigma^2[k] + \mathbf{E}\sigma_e^2$. The posterior expression of P_{ik} could be rewritten as:

$$P(i,k) \propto p(e^*|b(i,k) = 1)p(b(i,k) = 1|Pr_k)$$

$$\begin{aligned} logP(i,k) &= logp(e^*|b(i,k) = 1) + logp(b(i,k) = 1|Pr_k) \\ &= -\tfrac{1}{2}(log|\mathbf{H_k}| + (e^*)'\mathbf{H_k}^{-1}e^* + logPr_k) \end{aligned}$$

Take expectation of $logP_{ik}$ regarding the missing data e^* as follows:

$$E_{e^*}logP(i,k) = -\tfrac{1}{2}(log|\mathbf{H_k}| + (e^*)'\mathbf{H_k}^{-1}e^* + \mathrm{tr}(\mathbf{H_k}^{-1}\mathrm{PEV}(e^*)) + logPr_k). \quad \text{(S2)}$$

Here, $\mathbf{H_k} = \mathbf{Z_i}\mathbf{Z_i}'\sigma^2[k] + \mathbf{E}\sigma_e^2$, $\mathrm{PEV}(e^*)$ was estimated in the File S1(Chapter 10). According to the Woodbury Identity theory, the calculation of the equation $\mathbf{H_k}^{-1}$ and $log|\mathbf{H_k}|$ could be simplified as

$$\mathbf{H_k}^{-1} = (\mathbf{Z_i}\mathbf{Z_i'}\sigma_i^2[k] + \mathbf{E}\boldsymbol{\sigma_e^2})^{-1} = \boldsymbol{\sigma_e^{-2}}\left(\mathbf{E^{-1}} - \frac{\mathbf{E^{-1}}\mathbf{Z_i}\mathbf{Z_i'}\mathbf{E^{-1}}\sigma_i^2[k]}{\sigma_i^2[k]\mathbf{Z_i'}\mathbf{E^{-1}}\mathbf{Z_i} + \boldsymbol{\sigma_e^2}}\right)$$

$$log|\mathbf{H_k}| = (\mathrm{n} - 1)\log\sigma_e^2 + \log|\mathbf{E}| + log(\sigma_i^2[k]\mathbf{Z_i'}\mathbf{E^{-1}}\mathbf{Z_i} + \sigma_e^2)$$

Therefore, the equation (S2) could be simplified as:

$$E_{e^*} logP(i,k) = log\Pr_{\mathrm{k}} - \tfrac{1}{2}\{(\mathrm{n}-1)\log \sigma_{\mathrm{e}}^2 + \log|\mathbf{E}| + \log(\sigma_i^2[k]\mathbf{Z_i'Z_i} + \sigma_{\mathrm{e}}^2)\} -$$

$$\tfrac{1}{2}\left\{\left((e^*)'^{\mathbf{E^{-1}}e^*}\right)\sigma_{\mathrm{e}}^{-2} - \left((e^*)'^{\mathbf{E^{-1}Z_i}}\right)^2 \sigma_i^2[k]\sigma_{\mathrm{e}}^{-2}/(\sigma_i^2[k]\mathbf{Z_i'E^{-1}Z_i} + \sigma_{\mathrm{e}}^2)\right\}$$

$$-\tfrac{1}{2}\{\mathrm{tr}(\mathbf{E^{-1}}\mathrm{PEV}(e^*))\sigma_{\mathrm{e}}^{-2} - \mathrm{tr}(\mathbf{E^{-1}Z_iZ_i'E^{-1}}\mathrm{PEV}(e^*))\sigma_i^2[k]\sigma_{\mathrm{e}}^{-2}/(\sigma_i^2[k]\mathbf{Z_i'E^{-1}Z_i} + \sigma_{\mathrm{e}}^2)\} \quad \text{(S3)}$$

Then, $P(i,k) = {exp(E_{e^*} logP(i,k))}\Big/{\sum_{k=1}^{4} exp(E_{e^*} logP(i,k))}$. (S4)

Bibliography

Abraham G., Tye-Din J.A., Bhalala O.G., Kowalczyk A., Zobel J. & Inouye M. (2014) Accurate and Robust Genomic Prediction of Celiac Disease Using Statistical Learning. *PLoS Genet* **10**, e1004137.

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. & Lawlor T.J. (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. *Journal of Dairy Science* **93**, 743-52.

Aguilar I., Misztal I., Legarra A. & Tsuruta S. (2011) Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal Breeding and Genetics* **128**, 422-8.

Badke Y.M., Bates R.O., Ernst C.W., Schwab C. & Steibel J.P. (2012) Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics* **13**, 1-10.

Barrett J.C., Hansoul S., Nicolae D.L., Cho J.H., Duerr R.H., Rioux J.D., Brant S.R., Silverberg M.S., Taylor K.D., Barmada M.M., Bitton A., Dassopoulos T., Datta L.W., Green T., Griffiths A.M., Kistner E.O., Murtha M.T., Regueiro M.D., Rotter J.I., Schumm L.P., Steinhart A.H., Targan S.R., Xavier R.J., Libioulle C., Sandor C., Lathrop M., Belaiche J., Dewit O., Gut I., Heath S., Laukens D., Mni M., Rutgeerts P., Van Gossum A., Zelenika D., Franchimont D., Hugot J.-P., de Vos M., Vermeire S., Louis E., Cardon L.R., Anderson C.A., Drummond H., Nimmo E., Ahmad T., Prescott N.J., Onnie C.M., Fisher S.A., Marchini J., Ghori J., Bumpstead S., Gwilliam R., Tremelling M., Deloukas P., Mansfield J., Jewell D., Satsangi J., Mathew C.G., Parkes M., Georges M. & Daly M.J. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62.

Beavis W. (1998) *QTL analyses: Power, precision, and accuracy*. CRC Press, New York.

Blott S., Kim J.-J., Moisio S., Schmidt-Küntzel A., Cornet A., Berzi P., Cambisano N., Ford C., Grisart B., Johnson D., Karim L., Simon P., Snell R., Spelman R., Wong J., Vilkki J., Georges M., Farnir F. & Coppieters W. (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **163**, 253-66.

Bolormaa S., Pryce J.E., Kemper K., Savin K., Hayes B.J., Barendse W., Zhang Y., Reich C.M., Mason B.A., Bunch R.J., Harrison B.E., Reverter A., Herd R.M., Tier B., Graser H.U. & Goddard M.E. (2013) Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle1.

Journal of Animal Science **91**, 3088-104.

Brøndum R.F., Su G., Janss L., Sahana G., Guldbrandtsen B., Boichard D. & Lund M.S. (2015) Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science* **98**, 4107-16.

Browning B.L. & Browning S.R. (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics* **84**, 210-23.

Calus M.P.L. (2014) Right-hand-side updating for fast computing of genomic breeding values. *Genetics Selection Evolution* **46**, 24.

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**, 553-61.

Chamberlain A.J., Vander Jagt C.J., Hayes B.J., Khansefid M., Marett L.C., Millen C.A., Nguyen T.T.T. & Goddard M.E. (2015) Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* **16**, 1-20.

Chen Q., Chen Y.-P.P. & Zhang C. (2007) Detecting inconsistency in biological molecular databases using ontologies. *Data Mining and Knowledge Discovery* **15**, 275-96.

Chen Y.-P.P. & Chen F. (2008) Using Bioinformatics Techniques for Gene Identification in Drug Discovery and Development. *Current Drug Metabolism* **9**, 567-73.

Christensen O.F. & Lund M.S. (2010) Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* **42**, 1-8.

Christensen O.F., Madsen P., Nielsen B., Ostersen T. & Su G. (2012) Single-step methods for genomic evaluation in pigs. *animal* **6**, 1565-71.

Clark S.A., Hickey J.M. & van der Werf J.H.J. (2011) Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution* **43**, 1-9.

Cleveland M.A. & Hickey J.M. (2013) Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation1. *Journal of Animal Science* **91**, 3583-92.

Cleveland M.A., Hickey J.M. & Forni S. (2012) A common dataset for genomic analysis of livestock populations. *G3 (Bethesda)* **2**, 429-35.

Cochran S.D., Cole J.B., Null D.J. & Hansen P.J. (2013) Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. *BMC Genetics* **14**, 1-23.

Cole J.B., Wiggans G.R., Ma L., Sonstegard T.S., Lawlor T.J., Crooker B.A., Van Tassell C.P., Yang J., Wang S., Matukumalli L.K. & Da Y. (2011) Genome-wide association analysis of thirty one production, health,

reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics* **12**, 1-17.

Collier R.J., Collier J.L., Rhoads R.P. & Baumgard L.H. (2008) Invited review: genes involved in the bovine heat stress response. *J Dairy Sci* **91**, 445-54.

Colombani C., Croiseau P., Fritz S., Guillaume F., Legarra A., Ducrocq V. & Robert-Granié C. (2012) A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. *Journal of Dairy Science* **95**, 2120-31.

D'Alessandro A., Zolla L. & Scaloni A. (2011) The bovine milk proteome: cherishing, nourishing and fostering molecular complexity. An interactomics and functional overview. *Molecular BioSystems* **7**, 579-97.

Daetwyler H.D., Calus M.P.L., Pong-Wong R., de los Campos G. & Hickey J.M. (2013) Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics* **193**, 347-65.

Daetwyler H.D., Capitan A., Pausch H., Stothard P., van Binsbergen R., Brondum R.F., Liao X., Djari A., Rodriguez S.C., Grohs C., Esquerre D., Bouchez O., Rossignol M.-N., Klopp C., Rocha D., Fritz S., Eggen A., Bowman P.J., Coote D., Chamberlain A.J., Anderson C., VanTassell C.P., Hulsegge I., Goddard M.E., Guldbrandtsen B., Lund M.S., Veerkamp R.F., Boichard D.A., Fries R. & Hayes B.J. (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* **46**, 858-65.

Daetwyler H.D., Kemper K.E., van der Werf J.H. & Hayes B.J. (2012a) Components of the accuracy of genomic prediction in a multi-breed sheep population. *J Anim Sci* **90**, 3375-84.

Daetwyler H.D., Pong-Wong R., Villanueva B. & Woolliams J.A. (2010) The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics* **185**, 1021-31.

Daetwyler H.D., Swan A.A., van der Werf J.H.J. & Hayes B.J. (2012b) Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genetics Selection Evolution* **44**, 1-11.

de los Campos G., Gianola D. & Allison D.B. (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat Rev Genet* **11**, 880-6.

de los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D. & Calus M.P.L. (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* **193**, 327-45.

de los Campos G., Klimentidis Y.C., Vazquez A.I. & Allison D.B. (2012) Prediction of Expected Years of Life Using Whole-Genome Markers. *PLoS ONE* **7**, e40964.

de los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E.,

Weigel K. & Cotes J.M. (2009) Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics* **182**, 375-85.

Di Giammartino Dafne C., Shi Y. & Manley James L. (2013) PARP1 Represses PAP and Inhibits Polyadenylation during Heat Shock. *Molecular cell* **49**, 7-17.

Dias J.G. & Wedel M. (2004) An empirical comparison of EM, SEM and MCMC performance for problematic Gaussian mixture likelihoods. *Statistics and Computing* **14**, 323-32.

Dikmen S., Wang X.z., Ortega M.S., Cole J.B., Null D.J. & Hansen P.J. (2015) Single nucleotide polymorphisms associated with thermoregulation in lactating dairy cows exposed to heat stress. *Journal of Animal Breeding and Genetics* **132**, 409-19.

Djian-Caporalino C., Pijarowski L., Fazari A., Samson M., Gaveau L., O'Byrne C., Lefebvre V., Caranta C., Palloix A. & Abad P. (2001) High-resolution genetic mapping of the pepper (Capsicum annuum L.) resistance loci Me3 and Me4 conferring heat-stable resistance to root-knot nematodes (Meloidogyne spp.). *Theoretical and Applied Genetics* **103**, 592-600.

do Amaral B.C., Connor E.E., Tao S., Hayen M.J., Bubolz J.W. & Dahl G.E. (2011) Heat stress abatement during the dry period influences metabolic gene expression and improves immune status in the transition period of dairy cows. *Journal of Dairy Science* **94**, 86-96.

Druet T., Macleod I.M. & Hayes B.J. (2014) Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* **112**, 39-47.

Duchemin S.I., Colombani C., Legarra A., Baloche G., Larroque H., Astruc J.M., Barillet F., Robert-Granié C. & Manfredi E. (2012) Genomic selection in the French Lacaune dairy sheep breed. *Journal of Dairy Science* **95**, 2723-33.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. & Goddard M.E. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science* **95**, 4114-29.

Fernando R.L., Dekkers J.C.M. & Garrick D.J. (2014a) Bayesian Methods for Genomic Prediction and Genome-Wide Association Studies combining Information on Genotyped and Non-Genotyped Individuals. *Animal Industry Report:* **AS 660, ASL R2865**.

Fernando R.L., Dekkers J.C.M. & Garrick D.J. (2014b) A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics Selection Evolution* **46**, 1-13.

Fritz S., Capitan A., Djari A., Rodriguez S.C., Barbat A., Baur A., Grohs C., Weiss

B., Boussaha M., EsquerrÈ D., Klopp C., Rocha D. & Boichard D. (2013) Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. *PLoS ONE* **8**, e65550.

Gao H., Lund M.S., Zhang Y. & Su G. (2013) Accuracy of genomic prediction using different models and response variables in the Nordic Red cattle population. *Journal of Animal Breeding and Genetics* **130**, 333-40.

Garrick D., Taylor J. & Fernando R. (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution* **41**, 55.

Gianola D. (2013) Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. *Genetics* **194**, 573-96.

Gianola D., de los Campos G., Hill W.G., Manfredi E. & Fernando R. (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* **183**, 347-63.

Gilmour A., Cullis B., Welham S. & Thompson R. (2002) ASReml Reference Manual 2nd edition. *NSW Agriculture Biometrical Bulletin 3*.

Goddard M. (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-57.

Goddard M.E. & Hayes B.J. (2007) Genomic selection. *Journal of Animal Breeding and Genetics* **124**, 323-30.

Goddard M.E. & Hayes B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* **10**, 381-91.

Grinwald M. & Ron E.Z. (2013) The Escherichia coli Translation-Associated Heat Shock Protein YbeY Is Involved in rRNA Transcription Antitermination. *PLoS ONE* **8**, e62297.

Grisart B., Coppieters W., Farnir F., Karim L., C F., Berzi P., Cambisano N., Mni M., Reid S., Simon P., Spelman R., Georges M. & Snell R. (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**, 222-31.

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

Habier D., Fernando R.L., Kizilkaya K. & Garrick D.J. (2011) Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**, 1-12.

Haile-Mariam M., Bowman P.J. & Pryce J.E. (2013) Genetic analyses of fertility and predictor traits in Holstein herds with low and high mean calving intervals and in Jersey herds. *Journal of Dairy Science* **96**, 655-67.

Haile-Mariam M., Carrick M.J. & Goddard M.E. (2008) Genotype by Environment Interaction for Fertility, Survival, and Milk Production Traits in Australian Dairy Cattle. *Journal of Dairy Science* **91**, 4840-53.

Haile-Mariam M., Pryce J.E., Schrooten C. & Hayes B.J. (2015) Including overseas performance information in genomic evaluations of Australian dairy cattle. *Journal of Dairy Science* **98**, 3443-59.

Haile‐Mariam M., Nieuwhof G.J., Beard K.T., Konstatinov K.V. & Hayes B.J. (2013) Comparison of heritabilities of dairy traits in Australian Holstein‐Friesian cattle from genomic and pedigree data and implications for genomic evaluations. *Journal of Animal Breeding and Genetics* **130**, 20-31.

Harris B. & Johnson D. (2010) The impact of high density SNP chips on genomic evaluation in dairy cattle. *Interbull Bull* **42**, 40-3.

Hayashi T. & Iwata H. (2010) EM algorithm for Bayesian estimation of genomic breeding values. *BMC Genetics* **11**, 1-9.

Hayes B.J., Bowman P.J., Chamberlain A.J., Savin K., van Tassell C.P., Sonstegard T.S. & Goddard M.E. (2009) A Validated Genome Wide Association Study to Breed Cattle Adapted to an Environment Altered by Climate Change. *PLoS ONE* **4**, e6676.

Hayes B.J., Carrick M., Bowman P. & Goddard M.E. (2003) Genotype × Environment Interaction for Milk Production of Daughters of Australian Dairy Sires from Test-Day Records. *Journal of Dairy Science* **86**, 3736-44.

Hayes B.J. & Goddard M.E. (2008) Technical note: Prediction of breeding values using marker-derived relationship matrices. *Journal of Animal Science* **86**.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J. & Goddard M.E. (2010) Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. *PLoS Genet* **6**, e1001139.

Heidaritabar M., Calus M.P.L., Megens H.J., Vereijken A., Groenen M.A.M. & Bastiaansen J.W.M. (2016) Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *Journal of Animal Breeding and Genetics* **133**, 167-79.

Henderson C. (1984) *Application of linear models in animal breeding*. University of Guelph, Canada.

Hozé C., Fritz S., Phocas F., Boichard D., Ducrocq V. & Croiseau P. (2014) Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. *Journal of Dairy Science* **97**, 3918-29.

Huang X., Wei X., Sang T., Zhao Q., Feng Q., Zhao Y., Li C., Zhu C., Lu T., Zhang Z., Li M., Fan D., Guo Y., Wang A., Wang L., Deng L., Li W., Lu Y., Weng Q., Liu K., Huang T., Zhou T., Jing Y., Li W., Lin Z., Buckler E.S., Qian Q., Zhang Q.-F., Li J. & Han B. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42**, 961-7.

Hwang J.Y., Lee E.J., Jin Go M., Sung Y.A., Lee H.J., Heon Kwak S., Jang H.C., Soo Park K., Lee H.J., Byul Jang H., Song J., Park K.H., Kim H.L., Cho M.C. & Lee J.Y. (2012) Genome-wide association study identifies GYS2 as

a novel genetic factor for polycystic ovary syndrome through obesity-related condition. *J Hum Genet* **57**, 660-4.

Jannink J.-L. (2010) Dynamics of long-term genomic selection. *Genetics Selection Evolution* **42**, 1-11.

Jannink J.-L., Lorenz A.J. & Iwata H. (2010) Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* **9**, 166-77.

Jiao Y., Zhao H., Ren L., Song W., Zeng B., Guo J., Wang B., Liu Z., Chen J., Li W., Zhang M., Xie S. & Lai J. (2012) Genome-wide genetic changes during modern breeding of maize. *Nat Genet* **44**, 812-5.

Kemper K.E., Reich C.M., Bowman P.J., vander Jagt C.J., Chamberlain A.J., Mason B.A., Hayes B.J. & Goddard M.E. (2015) Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genetics Selection Evolution* **47**, 29.

Kranis A., Gheyas A.A., Boschiero C., Turner F., Yu L., Smith S., Talbot R., Pirani A., Brew F., Kaiser P., Hocking P.M., Fife M., Salmon N., Fulton J., Strom T.M., Haberer G., Weigend S., Preisinger R., Gholami M., Qanbari S., Simianer H., Watson K.A., Woolliams J.A. & Burt D.W. (2013) Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* **14**, 1-13.

Lee Sang H., Wray Naomi R., Goddard Michael E. & Visscher Peter M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics* **88**, 294-305.

Legarra A. & Ducrocq V. (2012) Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *Journal of Dairy Science* **95**, 4629-45.

Li Q.L., Zhang Z.F., Xia P., Wang Y.J., Wu Z.Y., Jia Y.H., Chang S.M. & Chu M.X. (2015) A SNP in the 3'-UTR of HSF1 in dairy cattle affects binding of target bta-miR-484. *Genet Mol Res* **14**, 12746-55.

Li Y., Huang Y., Bergelson J., Nordborg M. & Borevitz J.O. (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **107**, 21199-204.

Lin Z., Hayes, B. J., Daetwyler, H. D. (2014) Genomic selection in crops, trees and forages: a review *Crop & Pasture Science* **65**, 1177-91.

Lippert C., Listgarten J., Liu Y., Kadie C.M., Davidson R.I. & Heckerman D. (2011) FaST linear mixed models for genome-wide association studies. *Nat Meth* **8**, 833-5.

Listgarten J., Lippert C., Kadie C.M., Davidson R.I., Eskin E. & Heckerman D. (2012) Improved linear mixed models for genome-wide association studies. *Nat Meth* **9**, 525-6.

Littlejohn M.D., Tiplady K., Lopdell T., Law T.A., Scott A., Harland C., Sherlock R.,

Henty K., Obolonkin V., Lehnert K., MacGibbon A., Spelman R.J., Davis S.R. & Snell R.G. (2014) Expression Variants of the Lipogenic AGPAT6 Gene Affect Diverse Milk Composition Phenotypes in Bos taurus. *PLoS ONE* **9**, e85757.

Liu Z., Goddard M.E., Reinhardt F. & Reents R. (2014) A single-step genomic model with direct estimation of marker effects. *Journal of Dairy Science* **97**, 5833-50.

Loh P.-R., Tucker G., Bulik-Sullivan B.K., Vilhjalmsson B.J., Finucane H.K., Salem R.M., Chasman D.I., Ridker P.M., Neale B.M., Berger B., Patterson N. & Price A.L. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90.

Lund M.S., Su G., Janss L., Guldbrandtsen B. & Brøndum R.F. (2014) Genomic evaluation of cattle in a multi-breed context. *Livestock Science* **166**, 101-10.

Lyu J., Jho E.-h. & Lu W. (2011) Smek promotes histone deacetylation to suppress transcription of Wnt target gene brachyury in pluripotent embryonic stem cells. *Cell Res* **21**, 911-21.

Macciotta N.P.P., Gaspa G., Steri R., Nicolazzi E.L., Dimauro C., Pieramati C. & Cappio-Borlino A. (2010) Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *Journal of Dairy Science* **93**, 2765-74.

MacLeod I.M., Bowman P.J., Vander Jagt C.J., Haile-Mariam M., Kemper K.E., Chamberlain A.J., Schrooten C., Hayes B.J. & Goddard M.E. (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* **17**, 144.

MacLeod I.M., Hayes B.J. & Goddard M.E. (2014a) The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. *Genetics* **198**, 1671-84.

MacLeod I.M., Hayes B.J. & Goddard M.E. (2014b) Will sequence SNP data improve the accuracy of genomic prediction in the presence of long term selection? In: *roceedings of the Twentieth Conference of the Association for the Advancement of Animal Breeding and Genetics,Translating Science into Action,* , pp. 215-9, Napier, New Zealand.

MacLeod I.M., Hayes B.J., VanderJagt C.J., Kemper K.E., Haile-Mariam M., Bowman P.J., Schrooten C. & Goddard M.E. (2014c) A Bayesian analysis to exploit imputed sequence variants for QTL discovery. In: *Proceedings on 10th World Congress of Genetics Applied to Livestock Production*, p. 193, Vancouver, BC, Canada.

Mäntysaari E. (2014) Challenges in industry application of genomic prediction experiences from dairy cattle. In: *Proceedings on 10th World Congress of Genetics Applied to Livestock Production*, Vancouver, BC, Canada.

Mao X., Kadri N.K., Thomasen J.R., De Koning D.J., Sahana G. & Guldbrandtsen

B. (2015) Fine mapping of a calving QTL on Bos taurus autosome 18 in Holstein cattle. *Journal of Animal Breeding and Genetics*, n/a-n/a.

McClure M.C., Morsci N.S., Schnabel R.D., Kim J.W., Yao P., Rolf M.M., McKay S.D., Gregg S.J., Chapple R.H., Northcutt S.L. & Taylor J.F. (2010) A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. *Animal Genetics* **41**, 597-607.

Mele M., Conte G., Castiglioni B., Chessa S., Macciotta N.P.P., Serra A., Buccioni A., Pagnacco G. & Secchiari P. (2007) Stearoyl-Coenzyme A Desaturase Gene Polymorphism and Milk Fatty Acid Composition in Italian Holsteins. *Journal of Dairy Science* **90**, 4458-65.

Meuwissen T. & Goddard M. (2010) Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics* **185**, 623-31.

Meuwissen T., Hayes B. & Goddard M. (2013) Accelerating Improvement of Livestock with Genomic Selection. *Annual Review of Animal Biosciences* **1**, 221-37.

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **157**, 1819-29.

Meuwissen T.H.E., Odegard J., Andersen-Ranberg I. & Grindflek E. (2014) On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genetics Selection Evolution* **46**, 1-8.

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

Mishra C., Palai T.K., Sarangi L.N., Prusty B.R. & Maharana B.R. (2013) Candidate gene markers for sperm quality and fertility in bulls. *Veterinary World* **6**, 905-10.

Misztal I., Legarra A. & Aguilar I. (2014) Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science* **97**, 3943-52.

Mizrak S.C., Bogerd J., Lopez-Casas P.P., Párraga M., del Mazo J. & de Rooij D.G. (2006) Expression of stress inducible protein 1 (Stip1) in the mouse testis. *Molecular Reproduction and Development* **73**, 1361-6.

Morota G., Abdollahi-Arpanahi R., Kranis A. & Gianola D. (2014) Genome-enabled prediction of quantitative traits in chickens using genomic annotation. *BMC Genomics* **15**, 1-10.

Moser G., Lee S.H., Hayes B.J., Goddard M.E., Wray N.R. & Visscher P.M. (2015) Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLoS Genet* **11**, e1004969.

Nahar J., Chen Y.-P.P. & Ali S. (2007) KERNEL-BASED NAIVE BAYES CLASSIFIER FOR BREAST CANCER PREDICTION. *Journal of Biological Systems* **15**, 17-25.

Neely G.G., Hess A., Costigan M., Keene A.C., Goulas S., Langeslag M., Griffin

R.S., Belfer I., Dai F., Smith S.B., Diatchenko L., Gupta V., Xia C.-p., Amann S., Kreitz S., Heindl-Erdmann C., Wolz S., Ly C.V., Arora S., Sarangi R., Dan D., Novatchkova M., Rosenzweig M., Gibson D.G., Truong D., Schramek D., Zoranovic T., Cronin S.J.F., Angjeli B., Brune K., Dietzl G., Maixner W., Meixner A., Thomas W., Pospisilik J.A., Alenius M., Kress M., Subramaniam S., Garrity P.A., Bellen H.J., Woolf C.J. & Penninger J.M. (2010) A Genome-wide Drosophila Screen for Heat Nociception Identifies α2δ3 as an Evolutionarily Conserved Pain Gene. *Cell* **143**, 628-38.

Ng-Kwai-Hang K. (1997) A Review of the Relationship between Milk Protein Polymorphism and Milk Composition/Milk Production. In: *Proceedings of the International Dairy Federation Seminar*, pp. 22-37, Palmerston North, New Zealand.

Nguyen T.T.T., Bowman P.J., Haile-Mariam M., Pryce J.E. & Hayes B.J. (2016) Genomic selection for tolerance to heat stress in Australian dairy cattle. *Journal of Dairy Science* **99**, 2849-62.

Niskanen E.A., Malinen M., Sutinen P., Toropainen S., Paakinaho V., Vihervaara A., Joutsen J., Kaikkonen M.U., Sistonen L. & Palvimo J.J. (2015) Global SUMOylation on active chromatin is an acute heat stress response restricting transcription. *Genome Biology* **16**, 1-19.

Ober U., Ayroles J.F., Stone E.A., Richards S., Zhu D., Gibbs R.A., Stricker C., Gianola D., Schlather M., Mackay T.F.C. & Simianer H. (2012) Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster. *PLoS Genet* **8**, e1002685.

Olson K.M., VanRaden P.M. & Tooker M.E. (2012) Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *Journal of Dairy Science* **95**, 5378-83.

Park T. & Casella G. (2008) The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681-6.

Paul E., Zhu Z.I., Landsman D. & Morse R.H. (2015) Genome-wide association of mediator and RNA polymerase II in wild-type and mediator mutant yeast. *Mol Cell Biol* **35**, 331-42.

Petesch S.J. & Lis J.T. (2012) Activator-Induced Spread of Poly(ADP-Ribose) Polymerase Promotes Nucleosome Loss at Hsp70. *Molecular cell* **45**, 64-74.

Pryce J.E., Arias J., Bowman P.J., Davis S.R., Macdonald K.A., Waghorn G.C., Wales W.J., Williams Y.J., Spelman R.J. & Hayes B.J. (2012) Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. *Journal of Dairy Science* **95**, 2108-19.

Pryce J.E., Bolormaa S., Chamberlain A.J., Bowman P.J., Savin K., Goddard M.E. & Hayes B.J. (2010) A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length

haplotypes. *Journal of Dairy Science* **93**, 3331-45.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J. & Sham P.C. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81**, 559-75.

Purfield D.C., Bradley D.G., Evans R.D., Kearney F.J. & Berry D.P. (2015) Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. *Genetics Selection Evolution* **47**, 1-13.

Rabindran S.K., Giorgi G., Clos J. & Wu C. (1991) Molecular cloning and expression of a human heat shock factor, HSF1. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 6906-10.

Rao T.V.L.N., Ramesha K.P., Barani A., Chauhan S.S. & Basavaraju M. (2013) Association of GSTP1 gene polymorphisms with performance traits in Deoni cattle. *African Journal of Biotechnology* **12**, 3768-73.

Rasouly A., Schonbrun M., Shenhar Y. & Ron E.Z. (2009) YbeY, a heat shock protein involved in translation in Escherichia coli. *J Bacteriol* **191**, 2649-55.

Raven L.-A., Cocks B.G. & Hayes B.J. (2014) Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics* **15**, 1-14.

Raven L.-A., Cocks B.G., Kemper K.E., Chamberlain A.J., Jagt C.J., Goddard M.E. & Hayes B.J. (2015) Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. *Mammalian Genome* **27**, 81-97.

Raychaudhuri S., Loew C., Körner R., Pinkert S., Theis M., Hayer-Hartl M., Buchholz F. & Hartl F.U. (2014) Interplay of Acetyltransferase EP300 and the Proteasome System in Regulating Heat Shock Transcription Factor 1. *Cell* **156**, 975-85.

Riedelsheimer C., Technow F. & Melchinger A.E. (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics* **13**, 1-9.

Roy R., Ordovas L., Zaragoza P., Romero A., Moreno C., Altarriba J. & Rodellar C. (2006) Association of polymorphisms in the bovine FASN gene with milk-fat content. *Animal Genetics* **37**, 215-8.

Saatchi M., Ward J. & Garrick D.J. (2013) Accuracies of direct genomic breeding values in Hereford beef cattle using national or international training populations1. *Journal of Animal Science* **91**.

Sahana G., Guldbrandtsen B. & Lund M.S. (2011) Genome-wide association study for calving traits in Danish and Swedish Holstein cattle. *Journal of*

Dairy Science **94**, 479-86.

Sanders K., Bennewitz J., Reinsch N., Thaller G., Prinzenberg E.M., Kühn C. & Kalm E. (2006) Characterization of the DGAT1 Mutations and the CSN1S1 Promoter in the German Angeln Dairy Cattle Population. *Journal of Dairy Science* **89**, 3164-74.

Sargolzaei M., Chesnais J.P. & Schenkel F.S. (2014) A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 1-12.

Schennink A., Heck J.M.L., Bovenhuis H., Visker M.H.P.W., van Valenberg H.J.F. & van Arendonk J.A.M. (2008) Milk Fatty Acid Unsaturation: Genetic Parameters and Effects of Stearoyl-CoA Desaturase (SCD1) and Acyl CoA: Diacylglycerol Acyltransferase 1 (DGAT1). *Journal of Dairy Science* **91**, 2135-43.

Schennink A., Stoop W.M., Visker M.H.P.W., Heck J.M.L., Bovenhuis H., Van Der Poel J.J., Van Valenberg H.J.F. & Van Arendonk J.A.M. (2007) DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. *Animal Genetics* **38**, 467-73.

Schmid A.B., Lagleder S., Gräwert M.A., Röhl A., Hagn F., Wandinger S.K., Cox M.B., Demmer O., Richter K., Groll M., Kessler H. & Buchner J. (2012) The architecture of functional modules in the Hsp90 co‐chaperone Sti1/Hop. *The EMBO Journal* **31**, 1506-17.

Schmitt S., Küry S., Giraud M., Dréno B., Kharfi M. & Bézieau S. (2009) An update on mutations of the SLC39A4 gene in acrodermatitis enteropathica. *Human Mutation* **30**, 926-33.

Schulman N.F., Sahana G., Iso-Touru T., McKay S.D., Schnabel R.D., Lund M.S., Taylor J.F., Virta J. & Vilkki J.H. (2011) Mapping of fertility traits in Finnish Ayrshire by genome-wide association analysis. *Animal Genetics* **42**, 263-9.

Schulthess A.W., Wang Y., Miedaner T., Wilde P., Reif J.C. & Zhao Y. (2016) Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. *Theoretical and Applied Genetics* **129**, 273-87.

Seber G.L., AJ (2002) Linear Regression Analysis. John Wiley and Sons, Hoboken.

Shepherd R.K., Meuwissen T.H.E. & Woolliams J.A. (2010) Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. *BMC Bioinformatics* **11**, 1-12.

Solberg T.R., Heringstad B., Svendsen M., Grove H. & Meuwissen T. (2011) Genomic Predictions for production- and functional traits in Norwegian red from BLUP analysis of imputed 54K and 777K SNP data. *Interbull* **44**, 240-3.

Solberg T.R., Sonesson A.K., Woolliams J.A. & Meuwissen T.H.E. (2009) Reducing dimensionality for prediction of genome-wide breeding values.

Genetics Selection Evolution **41**, 1-8.

Speed D. & Balding D.J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research* **24**, 1550-7.

Strandén I. & Garrick D.J. (2009) Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science* **92**, 2971-5.

Su G., Brøndum R.F., Ma P., Guldbrandtsen B., Aamand G.P. & Lund M.S. (2012) Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *Journal of Dairy Science* **95**, 4657-65.

Sun X., Habier D., Fernando R.L., Garrick D.J. & Dekkers J.C.M. (2011) Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian Methods. *BMC Proceedings* **5**, 1-8.

Sun X., Qu L., Garrick D.J., Dekkers J.C.M. & Fernando R.L. (2012) A Fast EM Algorithm for BayesA-Like Prediction of Genomic Breeding Values. *PLoS ONE* **7**, e49157.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78.

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **Series B (Methodological) 58**, 267–88.

Usai M.G., Goddard M.E. & Hayes B.J. (2009) LASSO with cross-validation for genomic selection. *Genetics Research* **91**, 427-36.

van Binsbergen R., Calus M.P.L., Bink M.C.A.M., van Eeuwijk F.A., Schrooten C. & Veerkamp R.F. (2015) Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* **47**, 71.

Vanraden P., Olson K., Null D., Sargolzaei M., Winters M. & Van Kaam J. (2012) Reliability increases from combining 50,000- and 777,000- marker genotypes from four countries. *Interbull. Bull* **46**, 75-9.

VanRaden P.M. (2007) Genomic measures of relationship and inbreeding. *Interbull Bull* **37**, 33-6.

VanRaden P.M. (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91**, 4414-23.

VanRaden P.M., Null D.J., Sargolzaei M., Wiggans G.R., Tooker M.E., Cole J.B., Sonstegard T.S., Connor E.E., Winters M., van Kaam J.B.C.H.M., Valentini A., Van Doormaal B.J., Faust M.A. & Doak G.A. (2013) Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science* **96**, 668-78.

VanRaden P.M., O'Connell J.R., Wiggans G.R. & Weigel K.A. (2011) Genomic evaluations with many more genotypes. *Genetics Selection Evolution* **43**, 1-11.

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

Verbyla K.L., Bowman P.J., Hayes B.J. & Goddard M.E. (2010) Sensitivity of genomic selection to using different prior distributions. *BMC Proceedings* **4**, 1-4.

Verbyla K.L., Hayes B.J., Bowman P.J. & Goddard M.E. (2009) Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetics Research* **91**, 307-11.

Visscher P.M. (2008) Sizing up human height variation. *Nat Genet* **40**, 489-90.

Wang H., Misztal I., Aguilar I., Legarra A. & Muir W.M. (2012a) Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research* **94**, 73-83.

Wang T., Chen Y.-P.P., Goddard M.E., Meuwissen T.H.E., Kemper K.E. & Hayes B.J. (2015) A computationally efficient algorithm for genomic prediction using a Bayesian model. *Genetics Selection Evolution* **47**, 34.

Wang T., Chen Y.P., Bowman P.J., Goddard M.E. & Hayes B.J. (2016) A hybrid expectation maximisation and MCMC sampling algorithm for Bayesian mixture model based genomic prediction and QTL mapping. *BMC Genomic* **17**, 744.

Wang X., Wurmser C., Pausch H., Jung S., Reinhardt F., Tetens J., Thaller G. & Fries R. (2012b) Identification and Dissection of Four Major QTL Affecting Milk Fat Content in the German Holstein-Friesian Population. *PLoS ONE* **7**, e40711.

Weber K.L., Thallman R.M., Keele J.W., Snelling W.M., Bennett G.L., Smith T.P.L., McDaneld T.G., Allan M.F., Van Eenennaam A.L. & Kuehn L.A. (2012) Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes1,2. *Journal of Animal Science* **90**.

Whittaker JC T.R., Denham MC. (2000) Marker-assisted selection using ridge regression. *Genetic Research* **75**, 249-52.

Wickramasinghe S., Hua S., Rincon G., Islas-Trejo A., German J.B., Lebrilla C.B. & Medrano J.F. (2011) Transcriptome Profiling of Bovine Milk Oligosaccharide Metabolism Genes Using RNA-Sequencing. *PLoS ONE* **6**, e18895.

Wimmer V., Lehermeier C., Albrecht T., Auinger H.-J., Wang Y. & Schön C.-C. (2013) Genome-Wide Prediction of Traits with Different Genetic Architecture Through Efficient Variable Selection. *Genetics* **195**, 573-87.

Wolc A., Stricker C., Arango J., Settar P., Fulton J.E., O'Sullivan N.P., Preisinger R., Habier D., Fernando R., Garrick D.J., Lamont S.J. & Dekkers J.C.M. (2011) Breeding value prediction for production traits in layer chickens

using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution* **43**, 1-9.

Wolc A., Zhao H.H., Arango J., Settar P., Fulton J.E., O'Sullivan N.P., Preisinger R., Stricker C., Habier D., Fernando R.L., Garrick D.J., Lamont S.J. & Dekkers J.C.M. (2015) Response and inbreeding from a genomic selection experiment in layer chickens. *Genetics Selection Evolution* **47**, 59.

Woo S.K., Lee S.D., Na K.Y., Park W.K. & Kwon H.M. (2002) TonEBP/NFAT5 Stimulates Transcription of HSP70 in Response to Hypertonicity. *Molecular and Cellular Biology* **22**, 5753-60.

Wood A.R., Esko T., Yang J., Vedantam S., Pers T.H., Gustafsson S., Chu A.Y., Estrada K., Luan J.a., Kutalik Z., Amin N., Buchkovich M.L., Croteau-Chonka D.C., Day F.R., Duan Y., Fall T., Fehrmann R., Ferreira T., Jackson A.U., Karjalainen J., Lo K.S., Locke A.E., Magi R., Mihailov E., Porcu E., Randall J.C., Scherag A., Vinkhuyzen A.A.E., Westra H.-J., Winkler T.W., Workalemahu T., Zhao J.H., Absher D., Albrecht E., Anderson D., Baron J., Beekman M., Demirkan A., Ehret G.B., Feenstra B., Feitosa M.F., Fischer K., Fraser R.M., Goel A., Gong J., Justice A.E., Kanoni S., Kleber M.E., Kristiansson K., Lim U., Lotay V., Lui J.C., Mangino M., Leach I.M., Medina-Gomez C., Nalls M.A., Nyholt D.R., Palmer C.D., Pasko D., Pechlivanis S., Prokopenko I., Ried J.S., Ripke S., Shungin D., Stancakova A., Strawbridge R.J., Sung Y.J., Tanaka T., Teumer A., Trompet S., van der Laan S.W., van Setten J., Van Vliet-Ostaptchouk J.V., Wang Z., Yengo L., Zhang W., Afzal U., Arnlov J., Arscott G.M., Bandinelli S., Barrett A., Bellis C., Bennett A.J., Berne C., Bluher M., Bolton J.L., Bottcher Y., Boyd H.A., Bruinenberg M., Buckley B.M., Buyske S., Caspersen I.H., Chines P.S., Clarke R., Claudi-Boehm S., Cooper M., Daw E.W., De Jong P.A., Deelen J., Delgado G., Denny J.C., Dhonukshe-Rutten R., Dimitriou M., Doney A.S.F., Dorr M., Eklund N., Eury E., Folkersen L., Garcia M.E., Geller F., Giedraitis V., Go A.S., Grallert H., Grammer T.B., Graszler J., Gronberg H., de Groot L.C.P.G.M., Groves C.J., Haessler J., Hall P., Haller T., Hallmans G., Hannemann A., Hartman C.A., Hassinen M., Hayward C., Heard-Costa N.L., Helmer Q., Hemani G., Henders A.K., Hillege H.L., Hlatky M.A., Hoffmann W., Hoffmann P., Holmen O., Houwing-Duistermaat J.J., Illig T., Isaacs A., James A.L., Jeff J., Johansen B., Johansson A., Jolley J., Juliusdottir T., Junttila J., Kho A.N., Kinnunen L., Klopp N., Kocher T., Kratzer W., Lichtner P., Lind L., Lindstrom J., Lobbens S., Lorentzon M., Lu Y., Lyssenko V., Magnusson P.K.E., Mahajan A., Maillard M., McArdle W.L., McKenzie C.A., McLachlan S., McLaren P.J., Menni C., Merger S., Milani L., Moayyeri A., Monda K.L., Morken M.A., Muller G., Muller-Nurasyid M., Musk A.W., Narisu N., Nauck M., Nolte I.M., Nothen M.M., Oozageer L., Pilz S., Rayner N.W., Renstrom F., Robertson N.R., Rose L.M., Roussel R., Sanna S., Scharnagl H., Scholtens S.,

Schumacher F.R., Schunkert H., Scott R.A., Sehmi J., Seufferlein T., Shi J., Silventoinen K., Smit J.H., Smith A.V., Smolonska J., Stanton A.V., Stirrups K., Stott D.J., Stringham H.M., Sundstrom J., Swertz M.A., Syvanen A.-C., Tayo B.O., Thorleifsson G., Tyrer J.P., van Dijk S., van Schoor N.M., van der Velde N., van Heemst D., van Oort F.V.A., Vermeulen S.H., Verweij N., Vonk J.M., Waite L.L., Waldenberger M., Wennauer R., Wilkens L.R., Willenborg C., Wilsgaard T., Wojczynski M.K., Wong A., Wright A.F., Zhang Q., Arveiler D., Bakker S.J.L., Beilby J., Bergman R.N., Bergmann S., Biffar R., Blangero J., Boomsma D.I., Bornstein S.R., Bovet P., Brambilla P., Brown M.J., Campbell H., Caulfield M.J., Chakravarti A., Collins R., Collins F.S., Crawford D.C., Cupples L.A., Danesh J., de Faire U., den Ruijter H.M., Erbel R., Erdmann J., Eriksson J.G., Farrall M., Ferrannini E., Ferrieres J., Ford I., Forouhi N.G., Forrester T., Gansevoort R.T., Gejman P.V., Gieger C., Golay A., Gottesman O., Gudnason V., Gyllensten U., Haas D.W., Hall A.S., Harris T.B., Hattersley A.T., Heath A.C., Hengstenberg C., Hicks A.A., Hindorff L.A., Hingorani A.D., Hofman A., Hovingh G.K., Humphries S.E., Hunt S.C., Hypponen E., Jacobs K.B., Jarvelin M.-R., Jousilahti P., Jula A.M., Kaprio J., Kastelein J.J.P., Kayser M., Kee F., Keinanen-Kiukaanniemi S.M., Kiemeney L.A., Kooner J.S., Kooperberg C., Koskinen S., Kovacs P., Kraja A.T., Kumari M., Kuusisto J., Lakka T.A., Langenberg C., Le Marchand L., Lehtimaki T., Lupoli S., Madden P.A.F., Mannisto S., Manunta P., Marette A., Matise T.C., McKnight B., Meitinger T., Moll F.L., Montgomery G.W., Morris A.D., Morris A.P., Murray J.C., Nelis M., Ohlsson C., Oldehinkel A.J., Ong K.K., Ouwehand W.H., Pasterkamp G., Peters A., Pramstaller P.P., Price J.F., Qi L., Raitakari O.T., Rankinen T., Rao D.C., Rice T.K., Ritchie M., Rudan I., Salomaa V., Samani N.J., Saramies J., Sarzynski M.A., Schwarz P.E.H., Sebert S., Sever P., Shuldiner A.R., Sinisalo J., Steinthorsdottir V., Stolk R.P., Tardif J.-C., Tonjes A., Tremblay A., Tremoli E., Virtamo J., Vohl M.-C., The Electronic Medical R., Genomics C., The M.C., The P.C., The LifeLines Cohort S., Amouyel P., Asselbergs F.W., Assimes T.L., Bochud M., Boehm B.O., Boerwinkle E., Bottinger E.P., Bouchard C., Cauchi S., Chambers J.C., Chanock S.J., Cooper R.S., de Bakker P.I.W., Dedoussis G., Ferrucci L., Franks P.W., Froguel P., Groop L.C., Haiman C.A., Hamsten A., Hayes M.G., Hui J., Hunter D.J., Hveem K., Jukema J.W., Kaplan R.C., Kivimaki M., Kuh D., Laakso M., Liu Y., Martin N.G., Marz W., Melbye M., Moebus S., Munroe P.B., Njolstad I., Oostra B.A., Palmer C.N.A., Pedersen N.L., Perola M., Perusse L., Peters U., Powell J.E., Power C., Quertermous T., Rauramaa R., Reinmaa E., Ridker P.M., Rivadeneira F., Rotter J.I., Saaristo T.E., Saleheen D., Schlessinger D., Slagboom P.E., Snieder H., Spector T.D., Strauch K., Stumvoll M., Tuomilehto J., Uusitupa M., van der Harst P., Volzke H., Walker M., Wareham N.J., Watkins H., Wichmann H.E.,

Wilson J.F., Zanen P., Deloukas P., Heid I.M., Lindgren C.M., Mohlke K.L., Speliotes E.K., Thorsteinsdottir U., Barroso I., Fox C.S., North K.E., Strachan D.P., Beckmann J.S., Berndt S.I., Boehnke M., Borecki I.B., McCarthy M.I., Metspalu A., Stefansson K., Uitterlinden A.G., van Duijn C.M., Franke L., Willer C.J., Price A.L., Lettre G., Loos R.J.F., Weedon M.N., Ingelsson E., O'Connell J.R., Abecasis G.R., Chasman D.I., Goddard M.E., Visscher P.M., Hirschhorn J.N. & Frayling T.M. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86.

Wray N.R., Goddard M.E. & Visscher P.M. (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17**, 1520-8.

Wray N.R., Yang J., Goddard M.E. & Visscher P.M. (2010) The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. *PLoS Genet* **6**, e1000864.

Xu S. (2003) Theoretical Basis of the Beavis Effect. *Genetics* **165**, 2259-68.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E. & Visscher P.M. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9.

Yang J., Lee S.H., Goddard M.E. & Visscher P.M. (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88**, 76-82.

Yang J., Zaitlen N.A., Goddard M.E., Visscher P.M. & Price A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100-6.

Yu X. & Meuwissen T.H.E. (2011) Using the Pareto principle in genome-wide breeding value estimation. *Genetics Selection Evolution* **43**, 35.

Zhang Z., Ersoz E., Lai C.-Q., Todhunter R.J., Tiwari H.K., Gore M.A., Bradbury P.J., Yu J., Arnett D.K., Ordovas J.M. & Buckler E.S. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355-60.

Zhou L., Ding X., Zhang Q., Wang Y., Lund M.S. & Su G. (2013a) Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. *Genetics Selection Evolution* **45**, 1-7.

Zhou L., Heringstad B., Su G., Guldbrandtsen B., Meuwissen T.H.E., Svendsen M., Grove H., Nielsen U.S. & Lund M.S. (2014) Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. *Journal of Dairy Science* **97**, 4485-96.

Zhou X., Carbonetto P. & Stephens M. (2013b) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet* **9**, e1003264.

Zhou X. & Stephens M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-4.