A HYBRID METHOD FOR THE IMPLEMENTATION OF GENOMIC SELECTION BASED ON FST PRIORITIZED SINGLE NUCLEOTIDE POLYMORPHISMS

by

SAJJAD TOGHIANI

(Under the Direction of Romdhane Rekaya)

ABSTRACT

Availability of high-density (HD) marker panels provides an opportunity to improve the accuracy of genomic selection (GS). Unfortunately, using HD panels resulted in no significant increase in the accuracy of GS. This lack of improvement in accuracy is more likely due to the limitations of current GS methods rather than the uselessness of HD data. Increasing variants in association models caused a reduction in statistical power. Increase in the number of genotyped animals complicated the inversion of the genomic relationship matrix. Thus, reducing the number of variants and eliminating the inversion of genomic relationship matrix are required for the full benefit from HD marker data. We proposed fixation index (F_{ST}) to prioritize SNPs for GS. To validate the usefulness of F_{ST}, a trait with heritability of 0.4 under different SNP densities was simulated. Prioritized top 2.5% markers were able to tag most significant QTL and to increase functional genomic similarity. The latter could be used as a decision-making or selection tool. In spite of being able to prioritize markers in linkage disequilibrium with relevant QTL, the latter explained only a portion of the genetic variance. This is the case because small effects QTL are often not tagged with the prioritized SNPs. These small effect QTL could be tracked, however, by a polygenic component. Thus, a hybrid model was proposed that included the prioritized SNPs and a polygenic component in the association model. The proposed approach was evaluated based on simulated data of a trait with heritability of 0.1 and 0.4 and a real data of weaning weight in beef cattle. Using only genotyped animals, the hybrid model outperformed BayesB, BayesC and GBLUP when the prioritized 2.5% SNPs were used in the association model. The hybrid model was extended to accommodate non-genotyped animals. It outperformed ssGBLUP method using simulated data under both heritability scenarios. Although the results of the evaluation are likely to depend on the data generating process including the genetic complexity of the trait, the hybrid model seemed to be competitive compared to current methods. Furthermore, its computational costs in terms of CPU time and peak memory are limited.

INDEX WORDS: Genomic selection, Marker prioritization, Hybrid method, Accuracy

A HYBRID METHOD FOR THE IMPLEMENTATION OF GENOMIC SELECTION BASED ON FST PRIORITIZED SINGLE NUCLEOTIDE POLYMORPHISMS

by

SAJJAD TOGHIANI

BS, Shiraz University, IRAN, 2005

MS, The University of Guilan, IRAN, 2007

MS, The University of Georgia, 2014

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2018

A HYBRID METHOD FOR THE IMPLEMENTATION OF GENOMIC SELECTION BASED ON FST PRIORITIZED SINGLE NUCLEOTIDE POLYMORPHISMS

by

SAJJAD TOGHIANI

Major Professor:	Romdhane Rekaya
Committee:	Ignacy Misztal
	Samual E. Aggrey
	Daniela Lino Lourenco
	El Hamidi Hay

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2018

DEDICATION

This work is dedicated to my lovely wife without her caring support it would not have been possible and to my delightful parents whose words of encouragement and push for tenacity still ring in my ears. To my sister and my brother who have never left my side and are very special.

ACKNOWLEDGEMENTS

Sincere appreciation and gratitude is extended to my advisor, Dr. Romdhane Rekaya, for his insightful guidance, immense help and support, and constructive criticism throughout the preparation and completion of this PhD dessertation. Thanks also go to other members of my advisory committee, Dr. Ignacy Misztal, Dr. Samual E. Aggrey, Dr. Daniela Lino Lourenco, and Dr. El Hamidi Hay for their support and helpful discussions during my programs. I would also like to thank all the faculty, staff and graduate students in The Department of Animal and Dairy Science UGA for providing a pleasant and stimulating environment. To my friends and faculties in the genetics and animal breeding group, thank you for listening, offering me advice, and supporting me through this entire process. To my friends scattered around the country and the world, thank you for your thoughts, well-wishes/prayers, phone calls, e-mails, texts, visits, editing advice, and being there whenever I needed as a friend. Finally, I thank the Almighty GOD for His Grace and Blessings without which this program would have been impossible.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS..v

LIST OF TABLES... vii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

2 LITERATURE REVIEW 3

3 GENOMIC DIFFERENTIATION AS A TOOL FOR SINGLE NUCLEOTIDE POLYMORPHISM PRIORITIZATION FOR GENOME WIDE ASSOCIATION AND PHENOTYPE PREDICTION IN LIVESTOCK…….……………………………………………………..…28

4 IMPLEMENTING A HYBRID MODEL FOR GENOMIC SELECTION USING PRIORITIZED SNPs AND POLYGENIC EFFECTS…………….. 62

5 EXTENSION OF THE HYBRID MODEL TO ACCOMMODATE NON-GENOTYPE ANIMALS…………………………………..…………….. 92

6 CONCLUSIONS..115

LIST OF TABLES

Page

Table 3.1: Number of selected SNPs, number of tagged QTLs, percentage of genetic variance explained, and the minimum QTL effect (% of genetic variance) captured by the selected SNPs under different marker densities, sampling distribution for the QTL effects, and cut-off point for the F_{ST} values 50

Table 3.2: Number of selected SNPs, number of tagged QTLs, percentage of genetic variance explained, and the minimum QTL effect (% of genetic variance) captured by the selected SNPs under different cut-off points of the F_{ST} distribution, heritabilities, and sample size for the Gamma and 300K marker density panel scenario. 51

Table 3.3: Distribution of genomic similarity (GS) between the 750 animals with the highest phenotypes under different marker densities, sampling distribution for the quantitative trait loci (QTL) effects, and cut-off point for the F_{ST} values…….....52

Table 3.4: Distribution of genomic similarity (GS) between the 5% animals with the highest phenotypes under different cut-off points of the F_{ST} distribution, heritabilities, and sample size for the Gamma and 300K marker density panel scenario densities 53

Table 4.1: Distribution of typed animals across the different genotyping platforms before and after quality control (QC)……………………………………………………84

Table 4.2: Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with h^2 =0.4. 85

Table 4.3: Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with h^2 =0.1. 86

Table 4.4: Estimates of variance components and heritability for weaning weight using different methods………………………………………………… 87

Table 4.5: CPU time and peak memory of the different methods used to implement genomic selection (simulated data; h^2= 0.4)…………………...……………...…88

Table 5.1: Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with h^2 =0.4. 111

Table 5.2: Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with h^2 =0.1. 112

LIST OF FIGURES

Page

Figure 3.1: Simulation parameters for the different scenarios...54

Figure 3.2: Distribution and effects of the 100 quantitative trait loci (QTLs) simulated from gamma distribution for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) marker (c) scenarios. QTL effects are expressed as percentage of genetic variance..…………….. 55

Figure 3.3: Distribution and effects of the 100 quantitative trait loci (QTLs) with predefined effects for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) marker (c) scenarios. QTL effects are expressed as percentage of genetic variance..……………..56

Figure 3.4: Distribution of estimated fixation index (F_{ST}) scores for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) markers (c) under the gamma distribution scenario..…………….. 57

Figure 3.5: Distribution of estimated fixation index (F_{ST}) scores for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) markers (c) under the predefined effect scenario..…………….. 58

Figure 3.6: Map of simulated QTLs (in Blue) and selected SNPs (in Red) across the 10 chromosomes under the 300K SNP marker density, gamma distribution for QTL effects, and 99.5 quantile as cut-off point for F_{ST} scores..……………………........59

Figure 3.7: Map of simulated QTLs (in Blue) and selected SNPs (in Red) across the 10 chromosomes under the 300K SNP marker density, gamma distribution for QTL effects, and 97.5 quantile as cut-off point for F_{ST} scores…………………..............60

Figure 3.8: Distribution of true (a) and estimated (b) breeding values of animals with high (>0.9; Green) and low (< 0.55; Red) genomic similarity with selected individual (in Blue) under Gamma distribution, 300K SNPs, and 99.5% quantile of F_{ST} distribution simulation scenario..………………...61

Figure 4.1: Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with h^2=0.4..………………......................89

Figure 4.2: Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with h^2=0.1..………………......................90

Figure 4.3: Accuracy of estimated breeding values for weaning weight in a composite beef breed using different methods...………………...91

Figure 5.1: Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with h^2=0.4..………………....................113

Figure 5.2: Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with h^2=0.1..………………....................114

CHAPTER 1

INTRODUCTION

Availability of high-density (HD) SNP marker panels and sequence data was expected to substantially improve the power of genome wide association studies (GWAS) and the accuracy of genomic selection (GS). Unfortunately, that was not the case. Together with the significant increase in the number of genotyped animals, sequence and HD panel data have created major challenges for the implementation of GWAS and GS. Lack of power due to sample size, high collinearity between markers, small effects of most quantitative trait loci (QTL), complex LD structures, and low minor allele frequencies has led to a significant reduction of statistical power of linear regression based approaches for implementation of GWAS and GS. Bayesian variable selection methods (e.g. BayesB and BayesR) rely on the magnitude of the marker effects to prioritize variants. Consequently, their efficiency decays with the increase of the number of genotyped markers as the effects of linked QTL (often small) are distributed across an increasing number of markers. Using external biological information (e.g. BayesRC) is an attractive approach to prioritize markers. Unfortunately, such external complementary biological information is limited and often is tissue or/and time specific. Furthermore, such data (e.g., gene expression) have a high noise-to-signal ratio. As a result, these methods did not increase accuracy in the presence of HD or sequence data. In spite of the increase in the number of genotyped animals, the majority of animals included in any genetic evaluation are not genotyped. Accommodating these animals in a genomic evaluation is practically impossible using linear regression methods. For mixed linear model based approaches, the increase in the number of variants does not present a major challenge at least from a computational perspective. However, the increase in the number of genotyped animals will make the direct inversion of genomic relationship matrix impossible. Approximating the inverse is a data driven process and, thus, its optimality or even adequacy is not guaranteed. A potential practical solution could be through the substantial reduction in the number of markers in the association model, eliminating the need to impute missing genotypes for non-genotyped animals, and the avoidance of the construction and inversion of the genomic relationship matrix. The purpose of this dissertation work is to tackle these major challenges with the following specific objectives 1) To evaluate the adequacy of F_{ST}, a measure of genetic differentiation, as an external and already available source of information to prioritize SNPs and to assess the genomic similarity between individuals based on the prioritized SNPs 2) to develop an alternative hybrid model to implement GS using prioritized SNPs and to evaluate its effectiveness compared to existing method using only genotyped animals 3) to extend the hybrid model to accommodate non-genotyped animals. To reach these objectives, several data sets were simulated under varying conditions and assumptions about the data generating process.

CHAPTER 2

LITERATURE REVIEW

Quantitative genetics, infinitesimal model and genomic selection

Quantitative genetics, or the genetics of complex traits, is the study of traits in which multiple influential genes and non-genetic factors jointly contribute to determine their distribution (Hill, 2010). The machinery of quantitative genetics has been successfully applied in varied research areas ranging from human genetics, to animal and plant breeding. Although the broad purpose of these disciplines is to determine how genetic and environmental factors contribute to the variance of experimental trait in a population, their specific purposes are different. Human geneticists are interested in designing association models to identify influential loci that correlate with human diseases. The main interest of evolutionary geneticists is in the dissection of the genetic architecture of traits to decipher their past and future evolutionary changes. For animal and plant breeders, the main interest is the identification of genetically superior individuals to be used as parents of the next generation with the ultimate goal of steady increase in selection response (Walsh, 2001).

The principles and theory of quantitative genetics were developed in the 1940s for plant and animal breeding (Hill, 2014). As quantitative genetics theory expanded, the genetic analysis of complex traits required sophisticated statistical and computing methods in terms of estimating quantities such as genetic variances, and heritability and breeding values (Falconer and Mackay, 1995; Lynch and Walsh, 1998). The critical requirement for quantitative genetics analyses in animal and plant breeding applications is the availability of accurate information about the relatedness among individuals (i.e., pedigree relationships). Within the framework of the classical mixed linear model, the pedigree relationships are the major tool for predicting the genetic merit or breeding value of candidates for selection. In spite of the simple assumption about the genetic mechanisms underlying quantitative traits in the classical animal model (many loci of individually small effects), spectacular selection response was achieved (Hill, 2010; Nelson et al., 2013).

With the rapid advances in genomics and the availability of high throughput data, the emphasis of quantitative genetics shifted towards mapping quantitative trait loci (QTL). Due to decreases in the cost of genotyping, mapping QTL was followed by genome-wide association studies (GWAS) aiming to identify markers associated with phenotypic variation (Hill, 2012; Gienapp et al., 2017; Visscher et al., 2017). The statistical power of GWAS to detect QTL has been a main concern for geneticists. This power is typically influenced by the effect of QTL as well as the sample size and the strength of the linkage disequilibrium (LD). Furthermore, errors in phenotyping and genotyping will also reduce the power of GWAS (Spencer et al., 2009; Wu and Zhao, 2009; Hill, 2012; Manchia et al., 2013). Thus, GWAS will greatly benefit from the increase in the sample size, the increase in the number of SNPs in the panel, and the ability to replicate significant associations in independent samples of animals (Goddard and Hayes, 2009).

Although extensive GWAS studies to identify the contributing QTL were carried out for complex traits, most of the genetic variance was unexplained by the significant SNPs (genome-wide significance). The simplest explanation is that quantitative traits are under the influence of a large number of QTL with small effects (Yang et al., 2010; Yang et al., 2011). Using over a half million SNPs (580K), Yang et al. (2010) were able to explain only 45% of the genetic variation in human height. The inability to estimate the majority of the genetic variance is likely to be due to the absence of major QTL or the insufficient LD of QTL with the genotyped SNPs. Similar results were observed using body mass index (Yang et al., 2011) and schizophrenia (Purcell et al., 2009). Even though whole genome prediction methods are promising approaches for prediction of complex traits, the improvement of prediction accuracy of unrelated individuals is largely depending on the number of closely related individuals in the studied population.

In spite of the different attempts in the field of animal breeding and genetics to use GWAS to identify QTLs and chromosomal segments influencing traits of economic interest in livestock, only limited success was achieved as identified variants explained only a minute portion of the total genetic variance (Hayes and Goddard, 2010; Sharma et al., 2015). Thus, using marker assisted selection (MAS) as a tool for genetic improvement will at best have only a limited success.

A different approach also known as genomic selection (GS) was introduced by Meuwissen and Goddard (2001) and consisted in using all the markers simultaneously to predict the breeding values. GS has two main advantageous over MAS. First, the majority of the genetic variance can be tracked by the large number of markers in the panel regardless of statistical significance. Second, there is no need for estimating individual

QTL effects, a step that is often associated with large uncertainty and bias. Therefore, GS has quickly become the method of choice for genetic improvement of complex traits by passing decades of attempts of using MAS (Hayes and Goddard, 2010; Bhat et al., 2016). To implement GS, a reference population with both genotyped and phenotyped individuals is required. Based on the assumed model, prediction equations will be developed. Estimates of the prediction equation using the reference population will be used to calculate genomic breeding value (GEBV) for selection candidates.

From the advent of GS in the field of animal breeding, two main statistical approaches formed the foundation of genomic prediction. These two models can be broadly categorized into multiple regression model and mixed linear models (Zhang et al., 2011; Garrick et al., 2014). In case of regression models for genomic prediction, two steps are required to estimate GEBV of a genotyped candidate to selection. In the first step, the SNP effects of SNPs in the panel are estimated based on the genotyped and phenotyped reference population. In the second step, the GEBV of genotyped individuals in the validation population are calculated as the sum of the product between the estimated SNP effects and their associated genotyped. Based on regression approach GS, different prediction models have been developed to estimate SNP effects. These methods include ridge regression BLUP, Bayesian variable selection (e.g., BayesA, B, C) and Bayesian LASSO (Meuwissen and Goddard, 2001; Friedman et al., 2010; Habier et al., 2010; Li and Sillanpaa, 2012). The major difference between the different regression methods for GS resides on the assumption about the distribution of the marker effects. Besides regression model, an alternative approach routinely used for genomic prediction is based on mixed linear model, which is also known as the BLUP procedure. The main advantage of the BLUP approach in genomic applications is the ability to make use all the relatives in the prediction of GEBVs through the genomic relationship matrix (GRM). The method of using GRM in BLUP procedure was introduced by Vanraden and Tooker (2007) and Habier et al. (2007) which was known as genomic-BLUP (GBLUP). Compared to regression models, GBLUP has several advantages: 1) computationally more efficient when the number of SNPs is much larger that the number of genotyped individuals reducing, thus, the dimensionality of the system of equations to be solved, 2) flexibility of adding non-genotyped animals in MME via pedigree relationships 3) predicting the GEBV of animals like traditional BLUP (VanRaden, 2008; Stranden and Garrick, 2009), 4) straightforward scaling for multivariate analyses.

Misztal et al. (2009) introduced a unified approach for GS called single-step genomic selection which eliminated the need to estimate directly the SNP marker effects. The single-step approach is an extension of the classical mixed linear by replacing the average additive relationship matrix (**A**) with the realized relationship matrix (**G**) or a blend of both in presence of non-genotyped animals as described by Legarra et al. (2009). The single-step approach has several advantages including the use of observed phenotypes, a straightforward accommodation of non-genotyped animals and multiple traits. However, it faces the challenge of inverting a very dense matrix which dimensions increase with the number of genotyped animals. Although some approximations were presented for the inversion of the genomic relationship matrix, they are data driven approaches where the optimality of their performance is not guaranteed.

Accuracy of genomic selection

Independently of the method used to implement GS, accuracy of estimated GEBVs seems to higher breeding values estimated based on genomic information. This is due to a better modelling of the Mendelian sampling, the removal of pedigree errors and the identification of unreported relationships. Additionally, the expected accuracy of genomic selection depends on the heritability of the trait and the effective size of the population. In fact, the expected prediction accuracy can be calculated using the following formula (Daetwyler et al., 2010)

$$r = \sqrt{\frac{h^2}{h^2 + \frac{M_e}{N_p}}}$$

where h^2 is the heritability of the trait, N_p is the number individuals in training population, and M_e is the number of independent chromosomal segments. The latter is a function of the effective population size and genome length (Goddard, 2009).

Accuracy of GEBVs is a crucial parameter for the successful implementation of GS and it is under the influence of several factors that could be clustered into “inflexible” and “flexible” factors. The former includes the length of genome, the effective size of the population, and the genetic architecture of trait (heritability, number of QTL). The latter includes the size and structure of training population, the density of marker panel, and prediction model used to estimate GEBV.

Even though the factors influencing the accuracy of GEBV are intrinsically interrelated, the increase in the size of the training population has more influence on accuracy of GEBVs than the other factors. The design and structure of the training population and its relatedness with validation set is crucial to maintain a high level of accuracy. Schulz-Streeck et al. (2012) shown that combining related individuals from different groups in the training population improved the predictive ability. Furthermore, studies in maize breeding population revealed including half-sibs from both parents in the training population rather than allocating large number of individuals arbitrarily increased the prediction accuracy (Riedelsheimer et al., 2013; Jacobson et al., 2014). The importance of increasing the size of the training population, especially for low-heritability traits, could be found in Hayes et al. (2009). The density of the marker panel is another factor that could affect the accuracy of GEBVs. Increase in marker density often results in an increase in LD between the markers on QTL which likely to lead to higher accuracy. Although using denser marker panels in several studies led to increase in accuracy (Calus et al., 2008; Solberg et al., 2008; Meuwissen, 2009), such accuracy will not persist across generations of selection. Basically, increasing the distance (number of generations) between the training and validation populations appears to decrease the accuracy. The key reason of this reduction of GEBV accuracy is due to change of LD structure between the markers and QTLs which could be mainly related to recombination and selection across generations. However, the threshold for marker density to reach optimum accuracy is itself variable and depends of several factors. For example, using marker density of roughly 160K for human height rapidly increased the prediction accuracy; however accuracy reached a plateau at a density of around 400K SNP markers (Makowsky et al., 2011). In a simulation study, Meuwissen (2009) reported that the prediction accuracy of unrelated individuals depends on the number of SNPs and training records. The minimum number of SNPs and training records required for unrelated individuals to obtain an accuracy of 0.88 – 0.93 are $10 \times N_e \times L$ and $2 \times N_e \times L$, respectively, where N_e and L are the effective population size and genome size in Morgan. However, for livestock populations, increasing the number of training records rather than increasing SNP density affected more on prediction accuracy, because of the relatedness structure in livestock compared to human populations.

From inflexible factors affecting GEBV accuracy, the genetic architecture of the trait including the heritability and the number of QTLs critically affect the accuracy of GS. Based on several studies high and low heritability traits (Daetwyler et al., 2008; Ornella et al., 2012; Zhao et al., 2013; Howard et al., 2014; Xu et al., 2014) there is a clear and consistent relationship between heritability and accuracy indicating an improve in accuracy with the increase heritability. Only few studies have deviated from this trend (Heffner et al., 2011; Liu et al., 2017). Although different GS models were implemented with distinct algorithms, the prediction accuracies achieved by those models are depended on the number of QTL underlying the genetic basis of the trait. According to Zhong et al. (2009) and Wang et al. (2015), the GEBV accuracies provided through different GS models were inversely related to the number of QTL. For traits controlled by a small number of QTL, Bayesian regression models (e.g., BayesB, BayesC) outperformed mixed models (e.g, GBLUP). Contrary to Bayesian regression models, the accuracy of GEBV obtained using GBLUP is stable and constant regardless of the number of QTL.

High density marker panels and sequence data

The availability of next generation sequences (NGS) data and high-density SNP panels and the substantial increase in the number of genotyped animals present great opportunities to further improve the accuracy of GS and the understanding of the genetic basis of complex traits. However, to achieve these goals, more creative implementation algorithms and modeling frameworks that reduce the noise in the estimates of the **G** matrix and make full use of NGS data in large scale GS should emerge.

First, the inclusion of all or most of sequence variants in the association model used in GRM approaches is statistically counterproductive and computationally almost non-tractable. Unfortunately, current methods used to prioritize "relevant" SNPs or variants based on statistical (BayesB, BayesR) or external biological (BayesRC) information are at best only marginally better (Meuwissen and Goddard, 2001; Erbe et al., 2012; MacLeod et al., 2016). Trying to prioritize between hundreds or even thousands of variants that are in LD with a QTL that explain a fraction of 1% of the genetic variation is not a trivial task and seldom will there be enough statistical power. Even in the best-case scenario of using the current number of sequence variants identified in the bovine genome (Daetwyler et al., 2014; Hayes et al., 2014), assuming that it is possible to prioritize variants based on statistical criteria, 300k to 1.5 Million variants will be selected at pi values of 0.99 and 0.95, respectively. In humans, the number of selected variants will range between 800k and 4 Million variants. Prioritization of variants based on biological information (i.e, BayesRC) is limited by the amount and quality of available prior information. Classifying variants based on their location in differentially expressed genes, for example, makes strong assumptions with much more consequential implications on the final results. Microarray gene expression experiments are snapshot measurements of mRNA abundance that are both time and tissue specific and are characterized by a high noise-to-signal ratio. Conditioning on the results of the association study to this prior information, when we are not even sure about which tissue to use and measure gene expression for a given trait, seems like a risky proposition.

Second, over 75% of identified polymorphisms are rare variants (MAF<1%). Any two random individuals in the population will differ at a maximum of 2% – 8% of these rare variants (Auton et al., 2015). This reality creates at least two problems: 1) Using all or the majority of these variants to compute **G** will likely lead to inaccurate estimates of the realized additive relationships because of the large overlap in variant genotypes (92 to 98%); 2) Even when a subset of variants is used to compute **G**, it is not likely to lead to a significant increase of GS accuracy, as when high density panels were used (Su et al., 2012), or a reduction of accuracy. This is true because using information provided by NGS will have small or even negative effects in the estimation of the *realized additive relationships.*

Although theoretically statistical or biological criteria are attractive to discriminate and prioritize SNPs and rare variants in GWAS, they suffer from several limitations as indicated in the previous section and highlighted by the little to no increase in accuracy of GS. Filtering variants based on their effects (BayesB, BayesR) is bounded by the limited statistical power and is unlikely to be useful in the presence of NGS data. Prior biological information could be very useful for prioritizing sequence variants, but unfortunately its abundance and quality are far from being acceptable for meaningful practical use.

The promise of population genomics

Broadly speaking, population genomics typically aims at studying simultaneous several loci to track their patterns of evolutionary processes such as mutation, genetic draft, and selection that may impact the frequency of those loci through the genome and the whole population. In the field of animal improvement, artificial selection for economic traits is expected to create a non-uniform pressure on specific regions of the genome. Furthermore, selection, as an evolutionary force, may cause different patterns of genetic variation among populations and even across genomes. The action of selection creates variation where some loci frequency (outlier loci) diverge from the rest of loci (neutral loci) on the genome. Traditionally, it was not possible to accurately identify those loci or genome segments that are under heavy selection pressure. However, the availability of high density marker panels allows for the tracking of the footprint of the outlying loci through the identification of selection signatures in regions of the genome that are under selective pressure (Luikart et al., 2003; Gholami et al., 2014; Gouveia et al., 2014).

These selection signatures are the result of either positive or balancing selection. The former occurs when favorable alleles increase in frequency and surrounding alleles tend to loss their diversity leading to the so called hitch-hiking effect or selective sweep. This phenomenon reduced the heterozygosity of the regions neighboring selected loci and introduces a deviation of frequency spectrum which results in an increase of rare variants in the regions of positive selection (Smith and Haigh, 1974; Kim and Stephan, 2002; Charlesworth, 2007). However, the balancing selection tries to maintain the polymorphism level in the region of selected loci. Contrary to positive selection, balancing selection increases the diversity of closely linked loci surrounding the selected variants, thereby deviating the frequency of the polymorphism to the intermediate level (Charlesworth, 2006; Oleksyk et al., 2010).

The footprint of selection is often blurred by several other factors acting on the genome; resulting in different patterns of selection signatures. These factors include the type and strength of selection operating on genome, the extent of the recombination rate, the history of the effective population size and the population demography and structure (Kim and Stephan, 2002; Cutter and Payseur, 2013; Gouveia et al., 2014).

Different statistical approaches have been proposed to identify and characterize signatures of selection. These approaches could categorize into three main groups based on the statistical test used for identification consisting of frequency spectrum (Tajima, 1989; Fay and Wu, 2000), linkage disequilibrium and haplotype structure (Sabeti et al., 2002; Voight et al., 2006), and index of population differentiation (Wright, 1951; Lewontin and Krakauer, 1973). The frequency spectrum method was introduced by Tajima (1989) and consists of computing the Tajima's D (Tajima, 1989) or the Fay and Wu's H-test (Fay and Wu, 2000) statistics. The second approach to detect selection signature relies on the concept of linkage disequilibrium, a measure of non-random association between alleles of two or more loci. Loci under positive selection and through the hitch-hiking effect (Smith and Haigh, 1974) tend to increase the extent of LD in the genomic region resulting in large haplotypes indicative of lower rate of decay in LD. This process in the basis of the extended haplotype homozygosity (EHH) statistic proposed by Sabeti et al. (2002) to detect signature of recent selection. The EHH statistic access the probability of two random chromosomes carrying a specific haplotype are identical by descent (IBD) based on homozygosity. To avoid the impact of different recombination rates across the genome, Voight et al. (2006) extended the EHH method to the integrated Haplotype Score (iHS) through the adjustment for the ancestral alleles within a population. The iHS approach shows that EHH areas for alleles under selection pressure are larger than for neutral alleles. Several studies used derivative of EHH statistic to investigate selection signature in different species such as cattle (Qanbari et al., 2011), poultry (Li et al., 2012; Zhang et al., 2012), swine (Ai et al., 2013) and humans (Sabeti et al., 2007). The third approach for detecting signal of selection on a genome is based on the measure of genetic differentiation between populations. Such differentiation results in variation in allele frequency of non-neutral loci between the different populations or sub-populations. One of the well-known index of genetic differentiation introduced was the F_{ST} introduced by Wright (1951) as a tool to quantify the genetic diversity due to the difference of allele frequency between populations. This statistic is the most widely used method to detect favorable loci under selection among populations (Gianola et al., 2010; Qanbari et al., 2012). Thus, genome scan for loci-specific F_{ST} scores provides evidence of selection and it can be used to identify genome regions under positive or neutral selection simply by the inspection of the distribution of the F_{ST} scores (Kullo and Ding, 2007). Various measurements were proposed to calculate the F_{ST} statistic (Akey et al., 2002; Amaral et al., 2011); however the results could be affected by several factors including genotyping errors, population stratification, and the effects of demography shaping genome-wide levels of polymorphism (Stinchcombe and Hoekstra, 2008; Narum and Hess, 2011). Regardless of the challenges facing F_{ST} estimators, geneticists developed well-accepted tools to identify loci and populations differentiation scores using the evolutionary history of population under study (Weir and Hill, 2002; Beaumont and Balding, 2004; Weir et al., 2005; Guo et al., 2009).

GS faces the challenge of accommodating a dramatic increase in the number of typed variants and a substantial increase in the number of genotyped animals. The former creates major problems for regression based methods for GS. The latter has and will further complicate the implementation of mixed linear based approaches. SNP prioritization has become a necessity. Including all the markers in a panel in the association model not only will not increase accuracy, it could lead to its reduction. Similarly, increase in the number of genotyped animals has substantially complicated the inversion of the genomic relationship matrix. Consequently, current methods using for GS did not benefit from the increase of the density of marker panels. Within this landscape, new approaches for the implementation of GS in presence of high density or sequence data and large number of genotyped animals are needed. In this study, population genomics and quantitative genetics approaches were combined to provide a powerful practical alternative. This has been accomplished through 1) Testing and validation of F_{ST}, a measure of population differentiation, as an effective tool for marker prioritization; 2) Development of a hybrid model that combined prioritized SNP and polygenic components, and 3) the extension of the hybrid model to accommodate non-genotyped animals.

References

Ai, H., L. Huang, and J. Ren. 2013. Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. PloS one 8: e56001.

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805-1814.

Amaral, A. J. et al. 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. PloS one 6: e14782.

Auton, A. et al. 2015. A global reference for human genetic variation. Nature 526: 68-74.

Beaumont, M. A., and D. J. Balding. 2004. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13.

Bhat, J. A. et al. 2016. Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. Frontiers in Genetics 7: 221.

Calus, M. P., T. H. Meuwissen, A. P. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553-561.

Charlesworth, B. 2007. A hitch-hiking guide to the genome: A commentary on'The hitch-hiking effect of a favourable gene'by John Maynard Smith and John Haigh. Genet. Res. 89: 389.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS genetics 2: e64.

Cutter, A. D., and B. A. Payseur. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Reviews Genetics 14: 262.

Daetwyler, H. D. et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46: 858-865.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021-1031.

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

Erbe, M. et al. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95: 4114-4129.

Falconer, D., and T. Mackay. 1995. Introduction to quantitative genetics. Longman Group Ltd., Harlow.

Fay, J. C., and C.-I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33: 1-22.

Garrick, D., J. Dekkers, and R. Fernando. 2014. The evolution of methodologies for genomic prediction. Livestock Science 166: 10-18.

Gholami, M. et al. 2014. Population genomic analyses based on 1 million SNPs in commercial egg layers. PloS one 9: e94509.

Gianola, D., H. Simianer, and S. Qanbari. 2010. A two-step method for detecting selection signatures using genetic markers. Genetics research 92: 141-155.

Gienapp, P. et al. 2017. Genomic Quantitative Genetics to Study Evolution in the Wild. Trends Ecol. Evol. 32: 897-908.

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

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

Gouveia, J. J. d. S., M. V. G. B. d. Silva, S. R. Paiva, and S. M. P. d. Oliveira. 2014. Identification of selection signatures in livestock species. Genetics and molecular biology 37: 330-342.

Guo, F., D. K. Dey, and K. E. Holsinger. 2009. A Bayesian hierarchical model for analysis of SNP diversity in multilocus, multipopulation samples. J Am Stat Assoc 104: 142-154.

Habier, D., R. Fernando, and J. Dekkers. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177: 2389 - 2397.

Habier, D., R. Fernando, K. Kizilkaya, and D. Garrick. 2010. Extension of the Bayesian alphabet for genomic selection. Proceedings of the 9th World Congress on Genetics applied to Livestock Production: 1-6 August 2010: 468 - 468.

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

Hayes, B., and M. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. Genome 53: 876-883.

Hayes, B. J. et al. 2014. Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, 17-22 August 2007, Vancouver,BC, Canada. p 1-6.

Heffner, E. L., J.-L. Jannink, H. Iwata, E. Souza, and M. E. Sorrells. 2011. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Science 51: 2597-2606.

Hill, W. G. 2010. Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 73-85.

Hill, W. G. 2012. Quantitative genetics in the genomics era. Curr Genomics 13: 196-206.

Hill, W. G. 2014. Applications of Population Genetics to Animal Breeding, from Wright, Fisher and Lush to Genomic Prediction. Genetics 196: 1-16.

Howard, R., A. L. Carriquiry, and W. D. Beavis. 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3: Genes, Genomes, Genetics 4: 1027-1046.

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo. 2014. General Combining Ability Model for Genomewide Selection in a Biparental Cross. Crop Science 54: 895-905.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-777.

Kullo, I. J., and K. Ding. 2007. Patterns of population differentiation of candidate genes for cardiovascular disease. BMC genetics 8: 48.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92: 4656-4663.

Lewontin, R., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.

Li, D. et al. 2012. Whole-genome scan for signatures of recent selection reveals loci associated with important traits in White Leghorn chickens. Poultry science 91: 1804-1812.

Li, Z., and M. J. Sillanpaa. 2012. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 125: 419-435.

Liu, T. et al. 2017. Assessment of the genomic prediction accuracy for feed efficiency traits in meat-type chickens. PloS one 12: e0173620.

Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet. 2003. The power and promise of population genomics: from genotyping to genome typing. Nature reviews genetics 4: 981.

Lynch, M., and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer.

MacLeod, I. M. et al. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17: 1-21.

Makowsky, R. et al. 2011. Beyond missing heritability: prediction of complex traits. PLoS Genet 7: e1002051.

Manchia, M. et al. 2013. The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. PLoS One 8: e76295.

Meuwissen, T., and M. Goddard. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 33: 605 - 634.

Meuwissen, T. H. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution 41: 35.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92: 4648-4655.

Narum, S. R., and J. E. Hess. 2011. Comparison of FST outlier tests for SNP loci under selection. Molecular ecology resources 11: 184-194.

Nelson, R. M., M. E. Pettersson, and O. Carlborg. 2013. A century after Fisher: time for a new paradigm in quantitative genetics. Trends Genet 29: 669-676.

Oleksyk, T. K., M. W. Smith, and S. J. O'Brien. 2010. Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences 365: 185-205.

Ornella, L. et al. 2012. Genomic prediction of genetic values for resistance to wheat rusts. The Plant Genome 5: 136-148.

Purcell, S. M. et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748-752.

Qanbari, S. et al. 2011. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics 12.

Qanbari, S. et al. 2012. A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. PloS one 7: e49525.

Riedelsheimer, C. et al. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194: 493-503.

Sabeti, P. C. et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832.

Sabeti, P. C. et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913.

Schulz-Streeck, T., J. O. Ogutu, Z. Karaman, C. Knaak, and H. P. Piepho. 2012. Genomic Selection using Multiple Populations. Crop Science 52: 2453-2461.

Sharma, A. et al. 2015. Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. Asian-Australasian journal of animal sciences 28: 1371-1379.

Smith, J. M., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genetics Research 23: 23-35.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and T. H. Meuwissen. 2008. Genomic selection using different marker types and densities. J Anim Sci 86: 2447-2454.

Spencer, C. C., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5: e1000477.

Stinchcombe, J. R., and H. E. Hoekstra. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100: 158.

Stranden, I., and D. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92: 2971 - 2975.

Su, G. et al. 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: 4657-4665.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Vanraden, P., and M. Tooker. 2007. Methods to explain genomic estimates of breeding value. J Dairy Sci 90: 374.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414-4423.

Visscher, P. M. et al. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101: 5-22.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4: e72.

Walsh, B. 2001. Quantitative Genetics in the Age of Genomics. Theoretical Population Biology 59: 175-184.

Wang, X., Z. Yang, and C. Xu. 2015. A comparison of genomic selection methods for breeding value prediction. Science bulletin 60: 925-935.

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Res 15: 1468-1476.

Weir, B. S., and W. G. Hill. 2002. Estimating F-statistics. Annu Rev Genet 36: 721-750.

Wright, S. 1951. The genetical structure of populations. Annals of eugenics 15: 323-354.

Wu, Z., and H. Zhao. 2009. Statistical power of model selection strategies for genome-wide association studies. PLoS Genet 5: e1000582.

Xu, S., D. Zhu, and Q. Zhang. 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proceedings of the National Academy of Sciences 111: 12456-12461.

Yang, J. et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genet 42: 565 - U131.

Yang, J. et al. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43: 519-525.

Zhang, H. et al. 2012. A genome-wide scan of selective sweeps in two broiler chicken lines divergently selected for abdominal fat content. BMC genomics 13: 704.

Zhang, Z., Q. Zhang, and X. Ding. 2011. Advances in genomic selection in domestic animals. Chinese Science Bulletin 56: 2655-2663.

Zhao, Y., J. Zeng, R. Fernando, and J. C. Reif. 2013. Genomic Prediction of Hybrid Wheat Performance. Crop Science 53: 802-810.

Zhong, S., J. C. Dekkers, R. L. Fernando, and J.-L. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182: 355-364.

CHAPTER 3

GENOMIC DIFFERENTIATION AS A TOOL FOR SINGLE NUCLEOTIDE POLYMORPHISM PRIORITIZATION FOR GENOME WIDE ASSOCIATION AND PHENOTYPE PREDICTION IN LIVESTOCK [1]

[1] Toghiani, S., L.Y. Chang, A. Ling, S.E. Aggrey, and R. Rekaya. 2017. Livestock Science. 205:24-30. Reprinted here with permission of the publisher.

Abstract

Genome-wide association studies (GWAS) have been successful in detecting associations between single nucleotide polymorphisms (SNPs) and phenotypic variation and in identifying several causative mutations. However, SNPs with significant association identified using GWAS tend to explain only small fraction of the phenotypic variations. GWAS are affected by lack of power due to small sample size, large numbers of highly correlated markers, and the moderate to small effects of most quantitative trait loci (QTLs). This situation is further complicated by the continuous increase in marker density, especially with the availability of next-generation sequencing (NGS) data. The latter generates an unprecedented number of marker variants, with a complex linkage disequilibrium (LD) structure limiting the advantage and adequacy of existing methods that internally try to prioritize (filter) SNPs (e.g. BayesB, and BayesR). Consequently, it is becoming necessary to either filter SNPs before conducting the association analysis or to enlist additional sources of information. Methods that include biological prior information (e.g. BayesRC) are limited by the amount and quality of available prior information. Knowledge of genetic diversity based on evolutionary forces is beneficial for tracking loci influenced by selection. The fixation index (F_{ST}), as a measure of allele frequency variation among sub-populations, provides a tool to reveal genomic regions under selection pressure. In order to evaluate its usefulness as an additional source of information, a simulation was carried out. A trait with heritability of 0.4 was simulated and three subpopulations were created based on the empirical phenotypic distribution (< 5% quantile; > 95% quantile; and between 5 and 95% quantiles). Marker data was simulated to mimic a bovine chip of 600K, 1 million, and 3 million SNP marker panels. Genetic complexity of the trait was modelled

by the number of QTLs, their distribution, and the magnitude of their effects. Using different empirical cut off values for F_{ST}, most QTLs were correctly detected using as few as 2.5% of SNP markers in the panels. Furthermore, the genomic similarity, calculated based on the selected SNPs, was very high (>0.80) for individuals with similar genetic and phenotypic values despite having limited to no pedigree relationship. These results indicate that filtering SNPs using F_{ST} could be beneficial for use in GWAS by focusing on genome regions under selection pressure. High functional genomic similarity based on selected markers indicates similarity in SNP signatures, regardless of relatedness, and translates into high phenotypic correlation that could be used in decision making.

Introduction

Advances in high-throughput technologies allow for genotyping with high-density single nucleotide polymorphic (SNP) marker panels. These high-density panels provide an opportunity to identify SNP markers in linkage disequilibrium (LD) with quantitative trait loci (QTLs). The SNP marker(s) effect can be estimated and used to discover functional variants and/or causal mutations. Such discovery could be of great importance for the understanding of the genetic mechanisms underlying complex traits (Hirschhorn et al., 2002). Several thousand genetic loci in association with human diseases have been already identified (Ohnishi et al., 2001; Barrett and Cardon, 2006; Eberle et al., 2007; Li et al., 2008). Unfortunately, these common variants identified through genome wide association studies (GWAS) have explained only a small portion of the observed variation in several complex traits. The "lost heritability" in the case of human height (Maher, 2008) highlights the complexity of the endeavor. Over time, it has become clear that this "lost heritability" problem is mainly due to the lack of power in identifying variants with small effects which jointly explain a large portion of the total variation (Manolio et al., 2009).

In addition to revealing SNP-trait associations, GWAS can be used to predict phenotypes and to estimate breeding values in animal and plant applications via so-called genomic selection (GS). Estimation of the breeding values requires the direct or indirect estimation of the SNP effects. Accuracy of genomic prediction depends, among others, on the density of the SNP panel and the LD between the SNPs and causative variants affecting the trait (Druet et al., 2014).

Classical GWAS suffers from the high dimensionality of the parameter space leading to high false discovery rate (Balding, 2006; Pe'er et al., 2008). Also, high LD between a given QTL and several markers, sometimes within the same gene, leads to small effects for each one of these markers and ultimately a lack of statistical power to declare any of them as being significant. Although an increase in sample size will improve the statistical power and help alleviate the problem (Cichon et al., 2009), this alternative is costly, time consuming and often not possible due to several reasons, such as the unavailability of biological samples. GWAS using common variants benefits from the strong LD between variants within a gene or genome segment where the latter is assumed to carry one or very few causal variants that could easily be tracked by the high LD with tag-SNPs. However, in the presence of rare variants, extensive allelic heterogeneity is expected within genes associated with complex traits. Furthermore, LD between rare variants, measured as the square of the correlation (r^2), within a gene is often weak due to large discrepancies in minor allele frequencies between variants.

Within this landscape of continuous increase in the density of markers maps, filtering (prioritization) of variants to be included in the association models is becoming a necessity. Traditionally, SNP filtering is conducted based on certain statistical criteria such as p-values for single marker analyses or quality of fit and model determination for Bayesian procedures such as BayesB (Meuwissen et al., 2001) and BayesR (Erbe et al., 2012). The latter showed some superiority for certain traits in the presence of low- and moderate-density marker panels compared to models that include all markers. However, they still suffer, although to a lesser degree, from high false positives, multiple testing problems, high LD and small SNP effects which have hampered at different degrees of their efficiency. Trying to prioritize between hundreds or even thousands of variants that are in LD with a QTL that explains a fraction of 1% of the genetic variation is not a trivial task and seldom will there be enough statistical power. Even in the best- case scenario, assuming that it is possible to prioritize variants based on statistical criteria, 300K to 1.5 million variants will be selected at pi values of 0.99 and 0.95, respectively, using the current number of sequence variants identified in the bovine genome (Daetwyler et al., 2014; Hayes et al., 2014). Thus, enlistment of additional sources of information seems to be an attractive alternative. BayesRC (MacLeod et al., 2016), an extension of BayesR through the inclusion of biological prior information (variant type, location in differentially expressed genes), did not lead to any meaningful increase in accuracy compared to BayesR (Erbe et al., 2012). This is not due to the inadequacy of the approach, but rather to the limitations of the prior biological information. Classifying variants based on their location in differentially expressed genes, for example, makes strong assumptions with very consequential implications on the final results. Microarray gene expression experiments are snapshot measurements of mRNA abundance that are both time- and tissue-specific and are characterized by a high noise-to-signal ratio. Conditioning the results of the association study to this prior information could be risky given the uncertainty of gene expression data. Consequently, with the continuous increase in SNP marker densities, including the availability of millions of sequence variants, and considering the limited quantity and quality of prior biological information, it is clear that statistical discriminatory criteria alone will not be enough to prioritize influential variants and that enlistment of using index of fixation (F_{ST}) as additional sources of information has become a necessity.

Modern livestock species (e.g., Dairy cattle) are highly selected for on several traits of economic interest and the history of such artificial selection is stored in the genomes of these animals. Such signatures of selection could be traced and used as external information to prioritize SNPs. As an example, F_{ST}, a measure of allele frequency variation among sub-populations resulting from genetic differentiation provides a tool to reveal selection sweeps (Lewontin and Krakauer, 1973) and can be used to identify SNPs under selection pressure due to their LD with QTLs. Small F_{ST} values indicate a similar allelic composition between populations, while high F_{ST} values are the results of lack of shared alleles between populations and represent a signature of positive directional selection. One of the main applications of F_{ST} is pinpointing genome regions which are under selection. Several studies have shown that locus-specific estimates of F_{ST} could detect SNPs showing divergent patterns of variation (Akey et al., 2002; Beaumont and Balding, 2004; Storz et al., 2004; Weir et al., 2005). Thus, it is reasonable to postulate that F_{ST} as a measure of population differentiation, can be used as screening feature to prioritize SNPs for GWAS and GS studies. In this study, a simulation was carried out under different marker densities and complexity of the genetic model (number and size of QTL effects) to: 1) evaluate the adequacy of F_{ST} as an external source of information to prioritize SNPs, and 2) to assess the genomic similarity between individuals based on the prioritized SNPs and to evaluate its adequacy as a genetic decision tool.

Materials and Methods

Simulated population structure

Simulating genomic data via QMSim software (Sargolzaei and Schenkel, 2009) consists of a two-step process. In the first step of the process, a historical population is generated. During this step, a population of 8,000 individuals was kept under random mating for 300 generations, followed by an additional 305, 310 and 320 generations with population size of 15,000, 12,000 and 17,000 individuals, respectively. The first step is carried out to initialize LD and to establish mutation-drift equilibrium in the historical generations. The mating system of historical generations was maintained based on random unions of gametes, which were randomly sampled from both the male and female gamete pools. In the second step of simulating the population structure, the founder population was generated and labelled as generation zero (G0). In our simulation scenario, the G0 population was generated from the last historical generation, based on 1500 males and 15,000 females. The mating of these individuals was random, and no selection was considered at this step. After G0, three generations were simulated and the last one (G3) was used to detect selection signatures and to evaluate the proposed approach.

From G0 to G3, animals were selected based on their estimated breeding values (EBVs) with a replacement rate of 50 and 20% for males and females, respectively. Sex ratio in the progeny was maintained at 50% and one progeny per dam was assumed throughout. One trait with either a moderate (0.4) or low (0.1) heritability was simulated where all the genetic variation was assumed explained by the simulated QTLs. Phenotypic variance was set equal to one and the residual variance was adjusted in each scenario to maintain the heritability constant at 0.4 or 0.1. The true breeding value of an individual was equal to the sum of the QTL additive effects. Phenotypes were generated by adding random errors, sampled from a normal distribution with zero mean and dispersion equal to the residual variance. To investigate the effects of the sample size on the performance of the proposed method, either all individual in the third generation (n=15,000) or a small random subset (n=5,000) were used.

Genome structure

In order to mimic high-density marker panels, a 10-chromosome genome was simulated with uniformly distributed 200K, 300K and 1 million SNP markers, resulting in a density similar to a bovine chip of 600K, 1 million, 3 million SNPs, respectively. In the three cases, 100 QTLs were simulated with effects either generated from a gamma distribution with shape parameter set equal to 0.4 or predefined as a fraction of the total genetic variance. In the later scenario, QTL effects were set to explain at least 0.5% of the genetic variance. Variation in QTL effects was used as an indicator of the complexity of the genetic model. Both SNP markers and QTLs in all simulated scenarios were assumed to be bi-allelic, and no marker loci overlapped with the QTLs. Further, it was assumed that both SNP markers and QTLs have the same allele frequency in the historical population. Complete LD was simulated between markers, between QTL and between markers and QTL in the first historical population for all simulated genome scenarios. A detailed description of the simulated genome structure of the different scenarios is presented in Figure 3.1.

Detection of signature loci

In addition to the ratio of variances for F_{ST} calculation introduced initially by Wright (1951), there are many different approaches in the literature to estimate F_{ST} (Nei, 1973; Weir and Cockerham, 1984; Hudson et al., 1992; Weir and Hill, 2002). In this study and in order to evaluate the genomic differentiation in generation G3, the population was divided into three sub-populations based on the distribution of the trait phenotype (below the 5% quantile [S1], between 5 and 95% quantiles [S0], and above the 95% quantile [S2]). Subpopulations S1 and S2 were used to estimate the differentiation values using the global F_{ST} estimator method proposed by Nei (1973). For a given locus, ***k***, the global F_{ST} value is calculated as:

$$F_{ST_k} = \frac{H_{T_k} - H_{SW_k}}{H_{T_k}}$$

with $H_{SW_k} = \frac{H_{S1_k} * n_{s1} + H_{S2_k} * n_{s2}}{n_{s1} + n_{s2}}$, $H_{T_k} = 2 * p_k * q_k$ and $H_{Si_k} = 2 * p_{Si_k} * q_{Si_k}$

where, p_{Si_k} and q_{Si_k} are the allele frequencies for locus ***k*** in subpopulation ***i*** of locus ***k***, n_{s1} and n_{s2} are the number of individuals per first and second subpopulation, H_{SW_k} is the weighted mean heterozygosity across the first and second subpopulations and H_{T_k} is the heterozygosity of the pooled subpopulations for locus ***k***.

Although theoretical approaches exist to determine loci under selection pressure based on the estimated F_{ST} values, they tend to be somewhat conservative, which could limit the predictive power of the selected set of SNPs. In this study, heuristically determined F_{ST} threshold values were used to select SNPs under selection pressure. For that purpose, the 97.5, and 99.5% quantiles of the F_{ST} distribution were used.

Functional genomic similarity

It is reasonable to expect that individuals with similar genetic values or even phenotypes (given the lack of systematic effects in our simulation model) will have high genomic similarity based on the selected SNPs. This similarity will likely be substantially higher than the expected additive relationships, and even the realized relationships, calculated using all SNPs in the panel. Conversely, individuals with different genetic values or phenotypes are likely to have much lower genomic similarity than the expected or observed additive relationships. Identity by state (IBS) analysis, which identifies the number of shared alleles between two individuals across a set of given loci, was used to calculate the genetic similarity between individuals based on the selected SNPs. In this study, similarity between individuals ***i*** and ***j*** was computed as:

$$sim(i,j) = \frac{1}{2n}\sum_{k=1}^{n} s_k\,(i,j)$$

where $s_k(i,j)$ is the number of shared alleles between individuals ***i*** and ***j*** at locus ***k***.

In order to evaluate the effectiveness of this functional similarity as a potential genetic or decision-making tool, individuals with similarity scores greater than a certain threshold (i.e. 0.90) were compared based on their true breeding values (TBVs), EBVs, and phenotypes.

Results and Discussion

Detection of loci under selection pressure

Figures 3.2 and 3.3 present the effects and distribution along the genome of the simulated 100 QTLs either from a gamma (Fig 3.2) or a uniform (Fig 3.3) distribution. In the latter scenario, the lower and upper bounds were set equal to 0.5 and 1.5% of genetic variance. Thus, each QTL will explain at least 0.5% and at maximum 1.5% of genetic variance. The gamma and uniform distributions were used to investigate the effects of the genetic complexity of the trait on the proposed method. When effects were simulated from a gamma distribution to reflect a complex genetic model, 69, 82, and 73% of QTLs explained individually less than 0.5% of genetic variance for the 200K, 300K and 1 million SNP scenarios, respectively. Obviously, all QTLs had an effect exceeding 0.5% of the genetic variance when the uniform distribution was used, reflecting, thus, a less complex genetic model. The percentage of QTLs explaining more than 1% of the genetic variance was 19, 17, and 16% when the effects were generated from a gamma distribution for 200K, 300K, and 1 million SNPs, respectively. Their corresponding percentage values were 54, 41, and 47% when QTL effects were simulated from a uniform distribution.

Figures 3.4 and 3.5 present the distribution of global F_{ST} values across the genome when the population size was equal 15,000 for the gamma distribution and the predefined QTL effect scenarios, respectively. When the QTL effects were simulated from a gamma distribution (Fig 3.2) there was a high coincidence between the distribution of QTLs along the genome (Fig 3.4) and the distribution of estimated F_{ST} across the three SNP densities. In fact, Figures 3.3 and 3.5 are almost indistinguishable, especially for the QTLs with large effects (QTL effects greater than 0.2). This is expected because SNPs linked to large effect QTLs will be under higher selection pressure and consequently a more noticeable change in their minor allele frequencies between the two extreme sub-populations S1 and S2. Although large effect QTLs were tracked with high precision, small effect QTLs were hard to detect. In fact, for the gamma distribution scenario and using the 99.5% quantile of the

F_{ST} distribution, only 14, 12 and 22% of the QTLs were tagged (LD > 0.7) by the selected SNPs for the 200K, 300K and 1 million SNP scenarios, respectively (Table 3.1). Similar percentages were observed for the predefined QTL effects scenario (Table 3.1). The distribution of these tagged QTLs along the chromosomes under the 300K SNP scenario for the 99.5% quantile cut-off point is presented in Figure 3.6. It is clear that only large QTLs were identified. However, some very small effect QTLs that are in close proximity to large QTLs were indirectly tagged. In fact, the minimum percentage of genetic variance explained by these tagged QTLs was 0.013 and 0.56% for the gamma distribution and predefined effects scenarios, respectively (Table 3.1). Collectively, these tagged QTLs explained between 48.7 to 67.9% of the total genetic variance for the gamma distribution scenario and only between 12.44 to 24.71% for the predefined QTL effects scenario across the three simulated SNP densities (Table 3.1). The striking parallel between the percentage of identified QTLs and the percentage of the total genetic variance they explain are largely due to the small range used to simulate the QTL effects in the predefined scenario. However, for the gamma distribution scenario, the top 15% of QTLs explained over 48% of the total genetic variation. Using a more relaxed cut-off point, the 97.5% quantile of the F_{ST} distribution, more SNPs were selected, as expected, leading to more QTLs being identified and a larger portion of the genetic variance explained (Table 3.1). For the predefined scenario, 80 to 83% of the QTLs were identified and between 81 to 88% of genetic variance explained across the three SNP densities. Although an even higher percentage of the genetic variance was explained for the gamma distribution scenario (82 to 94%), only 40 to 54% to the QTLs were identified (Table 3.1). Obviously, this is due to the fact that 20 to 24% simulated QTLs in this scenario have very small effect that precluded their tracking by the selected SNPs, but they had almost no effect on the explained portion of the genetic variance. Figure 3.7 presents the distribution of the simulated 100 QTLs across the different chromosomes and the selected SNPs for the 97.5% quantile scenario and 300K SNP density (similar results for the other SNP density scenarios).

Although it seems somehow contradictory, it is actually expected because, as all QTLs have effects significantly different from zero and limited range of variation for the predefined scenario, divergence in the phenotypic values of the trait between the two subpopulations could be due to selection pressure on different combinations of the QTLs or a limited pressure in all of them leading thus to only a moderate change in minor allele frequencies of linked SNP markers. These results clearly show the non-trivial relationship between genetic complexity of the trait and the ability to track markers under selection pressure and ultimately to identify QTLs. It seems that even for very complex traits, few QTLs with large relative effects (compared to the majority of other QTLs) will be easily identified, which is not always the case with less complex traits. However, these identified QTLs will likely explain only a small portion of the genetic variance. As the cut-off threshold for significant F_{ST} values is relaxed, the number of tracked QTLs will increase in both scenarios but with a much faster pace for the less complex traits. This result corroborates the well-known reports of “lost heritability” of complex traits (Maher, 2008; McCarthy and Hirschhorn, 2008; Manolio et al., 2009) when only highly-significant markers were used and that much of the genetic variance was recovered when all markers in the panel were considered. Although no formal comparison with already existing methods for marker prioritization (e.g. BayesB, BayesR) was carried out in this study, some preliminary results of our ongoing research have shown that filtering of SNPs base of F_{ST} was slightly superior to BayesB in the prediction of genomic breeding values (Chang et al., 2016).

Table 3.2 presents the results of the proposed method for the scenarios of low heritability (0.1) and small sample size (n=5,000) when QTL effects where generated from a Gamma distribution and marker density set equal at 300k SNPs. As expected, smaller sample size and low heritability reduced the number of tagged QTLs compared to the scenario when a larger population and higher heritability were used. In fact, less than half of the QTL were tagged when the heritability was reduced to 0.1 compared to the scenario when the heritability was reduced to 0.1 for both cut-off points of the F_{ST} distribution. Similar results were observed when the population size was reduced 5,000 animals. The percentage of genetic variance explained by the tagged SNPs was reduced due to lower heritability for smaller sample size. However, such reduction was smaller compared to the reduction in the number of tagged SNPs. This is in part due to the presence of some QTLs with relatively large effects under the Gamma distribution scenario. In presence of low heritability or small sample size, only QTLs with relatively large effects will be tagged under conservative F_{ST} quantile cut-off points (Table 3.2). Results in table 3.2 indicate that the proposed method performed as expected under low heritability or reduced sample size scenarios.

Functional genomic similarity

Functional genomic similarity calculated based on SNP markers identified to be under selection pressure was used to evaluate their usefulness for phenotype prediction, genetic selection and as a decision-making tool. Using all SNP markers, the functional similarity will be an estimator of the observed additive relationships (Vinkhuyzen et al., 2013; Da et al., 2014). However, in this study, the similarity is computed based only on SNP markers identified to be under selection pressure. Thus, this similarity does not reflect necessarily additive relationships, but rather the level of similarity at these selected SNPs. As these SNP markers are selected based on their level of selection pressure due to their LD with QTLs, animals with similar genetic merit, and to lesser degree similar phenotypes, are expected to have high functional genetic similarity. Across all simulation scenarios, genomic similarity increased with the increase in the F_{ST} threshold used to identify SNPs under selection pressure. In fact, when the threshold was set equal to the 99.5% quantile of the F_{ST} distribution, 7.38 to 19.92% and 15.66 to 16.66% of the genomic similarities between the 750 animals with the highest phenotypes were greater than 0.8 for the gamma distribution and predefined QTL effects scenarios, respectively (Table 3.3). When the cut-off point was relaxed (97.5% quantile), very few genomic similarities were greater than 0.8. However, in all cases the genomic similarity was at least 0.60 and greater and 0.70 in over 60% of the relationships (Table 3.3). Similar results in trend and magnitude were observed between the 750 animals with the lowest phenotypes (Table not shown). When low heritability or small sample size were used, similar trend and magnitude of genetic similarities were observed (Table 3.4).

Collectively, these results indicate that computing genomic similarity based on SNPs under selection pressure will at minimum help cluster individuals based on the magnitude of their phenotypes. Furthermore, the results from the gamma distribution scenario at the 99.5 quantile cut-off threshold seem to indicate the possibility of using this functional genomic similarity as a selection and decision-making tool. This is the case because around 20% of all genomic similarities were greater or equal to 0.80 (Table 3.3) and the tagged QTLs explained 37.84 to 52.78% of the total genetic variance (Table 3.1). To test the viability of this option, we randomly selected an individual with high true breeding value (2.51) and then we identified all animals that had high (> 0.9) and low (< 0.55) genomic similarity with such individual. The average true breeding value of the 80 animals with the high genomic similarity with the selected individual was 1.66 and a standard deviation of 0.45 (Figure 3.8a). For the 142 animals with the low genomic similarity with the selected individual, the average was -0.243 with a standard deviation of 0.43 (Figure 3.8a). The same trend was observed when estimated breeding values (EBVs) were used. In fact, the EBV of the selected individual was 1.67, whereas the average of the EBVs of the 80 and 142 animals with the high and low genomic similarity was 1.45 and -0.091, respectively (Figure 3.8b). It is clear that the genomic similarity computed based on SNPs under selection pressure could be used at least as a decision-making tool given its ability to discriminate between animals with low and high breeding values and can be used as a low-cost selection tool in some specific breeding programs.

Conclusions

The availability of high-density SNP panels and sequence variant genotypes was expected to significantly increase the accuracy of genome-wide association studies and genomic selection. Unfortunately, little to no improvement in accuracy has been observed. This lack of significant improvement of accuracy is not the result of the limited usefulness of this data, but rather due to the limitations of current methods used to implement GWAS and GS. The dramatic increase in the dimensionality of the association models has reduced the effectiveness of statistical criteria-based variant (SNP) prioritization methods and mandated the need to enlist additional sources of information. Using available biological information as a prior is an attractive idea. Unfortunately, the quantity and quality of such information have limited its usefulness. Genome segments under selection pressure could be determined based on the available genotype data. In this study, F_{ST}, as a measure of genetic differentiation, was used as an additional source of information in the analysis of high-density marker data. Prioritized markers based on F_{ST} under different scenarios were able to tag the majority of significant QTLs and were successfully used to compute genomic similarity. The latter could be used as a decision-making or selection tool. Marker prioritization using F_{ST} is currently being evaluated and compared with existing methods based on the effects on the accuracy of genomic selection. In this study F_{ST} values are not technically an external information as they were computed based on available data. This does not seem to be a major issue because only a portion of the data (extremes of the distribution) was used for their calculation, and no validation records were included.

References

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805-1814.

Balding, D. J. 2006. A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781-791.

Barrett, J. C., and L. R. Cardon. 2006. Evaluating coverage of genome-wide association studies. Nat Genet 38: 659-662.

Beaumont, M. A., and D. J. Balding. 2004. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13.

Chang, L. Y., S. Toghiani, S. E. Aggrey, and R. Rekaya. 2016. High density marker panels, SNPs prioritizing and accuracy of genomic selection. Journal of Animal Science 94: 141-142.

Cichon, S. et al. 2009. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. The American journal of psychiatry 166: 540-556.

Da, Y., C. Wang, S. Wang, and G. Hu. 2014. Mixed Model Methods for Genomic Prediction and Variance Component Estimation of Additive and Dominance Effects Using SNP Markers. PLoS ONE 9: e87666.

Daetwyler, H. D. et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46: 858-865.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112: 39-47.

Eberle, M. A. et al. 2007. Power to detect risk alleles using genome-wide tag SNP panels. PLoS Genet 3: e170.

Erbe, M. et al. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95: 4114-4129.

Hayes, B. J. et al. 2014. Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, 17-22 August 2007, Vancouver,BC, Canada. p 1-6.

Hirschhorn, J. N., K. Lohmueller, E. Byrne, and K. Hirschhorn. 2002. A comprehensive review of genetic association studies. Genetics in medicine : official journal of the American College of Medical Genetics 4: 45-61.

Hudson, R. R., M. Slatkin, and W. Maddison. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583-589.

Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.

Li, M., C. Li, and W. Guan. 2008. Evaluation of coverage variation of SNP chips for genome-wide association studies. Eur J Hum Genet 16: 635-643.

MacLeod, I. M. et al. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17: 1-21.

Maher, B. 2008. Personal genomes: The case of the missing heritability. Nature 456: 18-21.

Manolio, T. et al. 2009. Finding the missing heritability of complex diseases. Nature 461: 747 - 753.

McCarthy, M. I., and J. N. Hirschhorn. 2008. Genome-wide association studies: potential next steps on a genetic journey. Human Molecular Genetics 17: R156-R165.

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

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences 70: 3321-3323.

Ohnishi, Y. et al. 2001. A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46: 471-477.

Pe'er, I., R. Yelensky, D. Altshuler, and M. J. Daly. 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic Epidemiology 32: 381-385.

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680-681.

Storz, J. F., B. A. Payseur, and M. W. Nachman. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Molecular biology and evolution 21: 1800-1811.

Vinkhuyzen, A. A., N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher. 2013. Estimation and partition of heritability in human populations using whole-genome analysis methods. Annu Rev Genet 47: 75-95.

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Res 15: 1468-1476.

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38: 1358-1370.

Weir, B. S., and W. G. Hill. 2002. Estimating F-statistics. Annu Rev Genet 36: 721-750.

Wright, S. 1951. The genetical structure of populations. Annals of eugenics 15: 323-354.

Table 3.1. Number of selected SNPs[1], number of tagged QTLs[2], percentage of genetic variance explained, and the minimum QTL effect (% of genetic variance) captured by the selected SNPs under different marker densities, sampling distribution for the QTL effects, and cut-off point for the F_{ST} values.

	97.5% quantile[3]						**99.5% quantile**					
	Gamma[4]			Predefined[5]			Gamma			Predefined		
	200K	**300K**	**1M**	**200K**	**300K**	**1M**	**200K**	**300K**	**1M**	**200K**	**300K**	**1M**
Selected SNP	4579	6257	22794	6026	10021	34035	650	464	2109	244	557	1033
Tagged QTLs[6]	40	50	54	83	81	80	14	12	22	11	13	21
% GV[7]	94.38	82.48	92.06	87.85	81.08	80.92	63.73	48.7	67.9	12.44	13.13	24.71
Min. QTL[8] effect	0.013	0.017	0.086	0.62	0.57	0.56	1.52	1.09	0.086	0.93	0.62	0.76

[1] SNPs = Single Nucleotide Polymorphisms, [2] QTLs = Quantitative Trait Loci, [3] cutoff point for the fixation index (F_{ST}) distribution, [4] QTL effects sampled from a Gamma distribution, [5] QTL effects pre-defined to explain at least 0.5% of genetic variance, [6] QTLs with linkage disequilibrium >0.70 with at least one selected SNP, [7] GV= Genetic Variance, and [8] minimum percentages of GV explained by a tagged QTL.

Table 3.2. Number of selected SNPs[1], number of tagged QTLs[2], percentage of genetic variance explained, and the minimum QTL effect (% of genetic variance) captured by the selected SNPs under different cut-off points of the F_{ST} distribution, heritabilities, and sample size for the Gamma[3] and 300K marker density panel scenario.

	97.5% quantile[4]			**99.5% quantile**		
	h^2=0.4 **n[5]=15,000**	h^2=0.4 **n=5,000**	h^2=0.1 **n=15,000**	h^2=0.4 **n=15,000**	h^2=0.4 **n=5,000**	h^2=0.1 **n=15,000**
Selected SNP	6257	6257	6257	464	464	464
Tagged QTL[6]	50	37	23	12	6	6
%GV[7]	82.48	78	72.2	48.7	40.33	43.99
Min. QTL[8] effect	0.017	0.37	0.089	1.09	3.21	4.54

[1] SNPs = Single Nucleotide Polymorphisms, [2] QTLs = Quantitative Trait Loci, [3] QTL effects sampled from a Gamma distribution, [4] cutoff point for the fixation index (F_{ST}) distribution, [5] sample size, [6] QTLs with linkage disequilibrium >0.70 with at least one selected SNP, [7] GV= Genetic Variance, and [8] minimum percentages of GV explained by a tagged QTL.

Table 3.3. Distribution of genomic similarity (GS) between the 750 animals with the highest phenotypes under different marker densities, sampling distribution for the quantitative trait loci (QTL) effects, and cut-off point for the F_{ST} values.

	97.5% quantile[1]						**99.5% quantile**					
	Gamma[2]			Predefined[3]			Gamma			Predefined		
	200K	**300K**	**1M**	**200K**	**300K**	**1M**	**200K**	**300K**	**1M**	**200K**	**300K**	**1M**
GS[4] > 0.90	0	0	0	0	0	0	0.036	2.11	0.37	1.31	1.06	0.07
0.80< GS < 0.90	0.007	0.041	0	0	0.001	0	7.38	19.92	14.29	15.66	16.66	10.55
0.70 <GS < 0.80	67.28	71.49	67.7	70.7	71.62	71.7	52.78	37.84	46.05	37.11	35.74	46.55
0.60< GS < 0.70	32.71	28.46	32.3	29.3	28.38	28.3	38.27	32.58	35.04	35.68	36.84	39.82
GS <0.60	0	0	0	0	0	0	1.527	7.543	3.97	10.23	9.71	3.00

[1] cutoff point for the fixation index (F_{ST}) distribution, [2] QTL effects sampled from a Gamma distribution, [3]QTL effects pre-defined to explain at least 0.5% of genetic variance, [4] GS= genomic similarity

Table 3.4. Distribution of genomic similarity (GS) between the 5% animals with the highest phenotypes under different cut-off points of the F_{ST} distribution, heritabilities, and sample size for the Gamma[1] and 300K marker density panel scenario densities.

	97.5% quantile[2]			99.5% quantile		
	h^2=0.4 n[3]=15,000	h^2=0.4 n=5,000	h^2=0.1 n=15,000	h^2=0.4 n=15,000	h^2=0.4 n=5,000	h^2=0.1 n=15,000
GS[4] >0.90	0	0	0	2.11	1.58	1.38
0.80 < GS< 0.90	0.041	0.0289	0.0078	19.95	20.08	17.73
0.70 < GS< 0.80	71.49	76.69	66.28	37.84	39.13	36.46
0.60 < GS< 0.70	28.46	23.28	33.71	32.58	32.94	35.48
0.60 < GS	0	0	0	7.54	6.26	8.95

[1] QTL effects sampled from a Gamma distribution, [2] cutoff point for the fixation index (F_{ST}) distribution, [3] sample size, [4] GS= genomic similarity

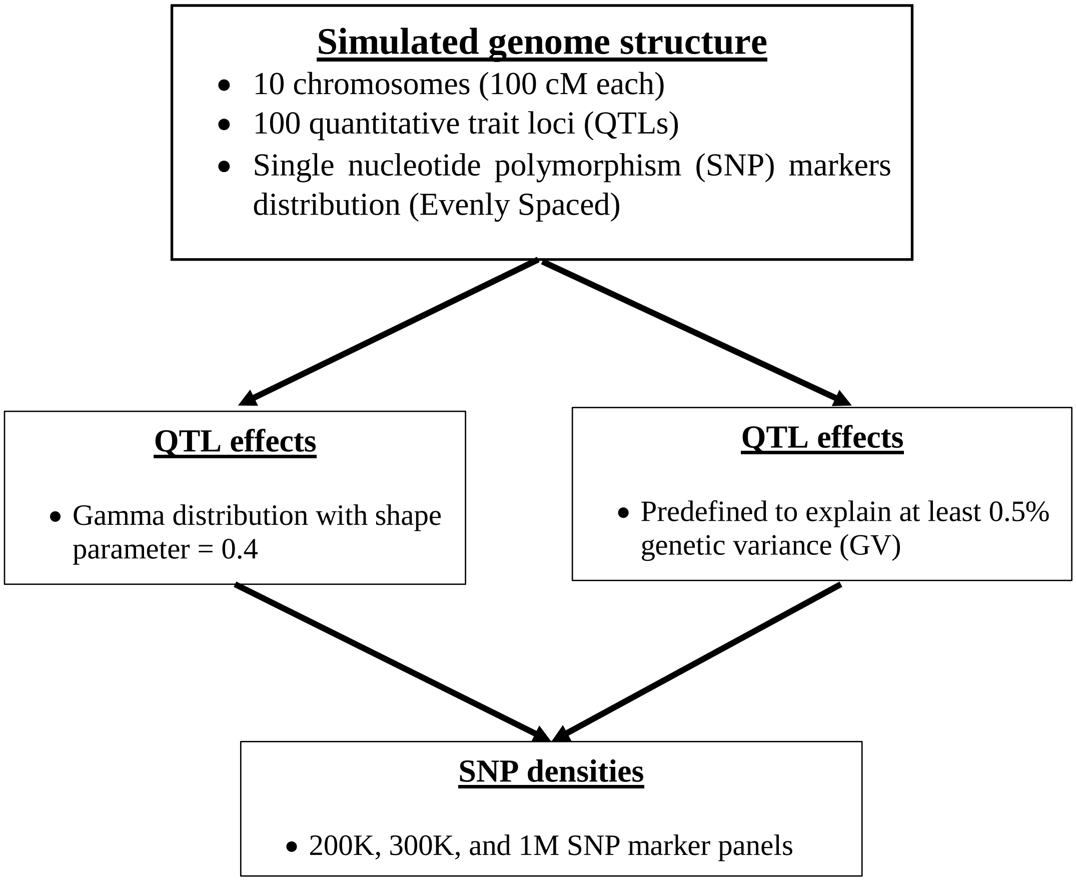
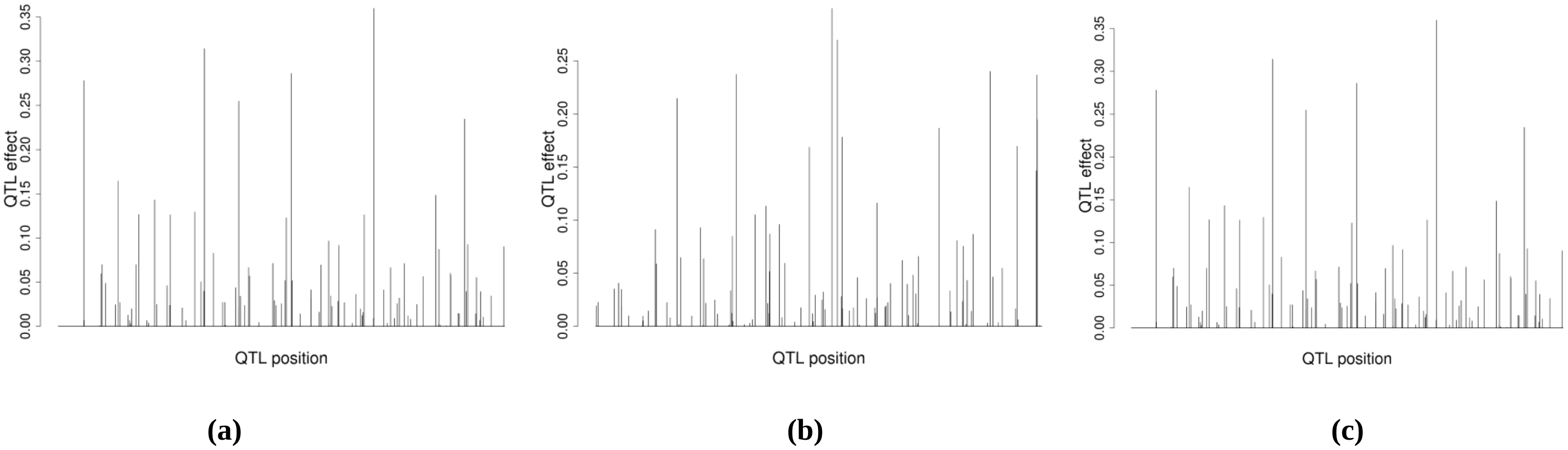

Simulated genome structure
10 chromosomes (100 cM each)
100 quantitative trait loci (QTLs)
Single nucleotide polymorphism (SNP) markers distribution (Evenly Spaced)
QTL effects
Gamma distribution with shape parameter = 0.4
QTL effects
Predefined to explain at least 0.5% genetic variance (GV)
SNP densities
200K, 300K, and 1M SNP marker panels

Figure 3.1. Simulation parameters for the different scenarios.



(a) **(b)** **(c)**

Figure 3.2. Distribution and effects of the 100 quantitative trait loci (QTLs) simulated from gamma distribution for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) marker (c) scenarios. QTL effects are expressed as percentage of genetic variance.

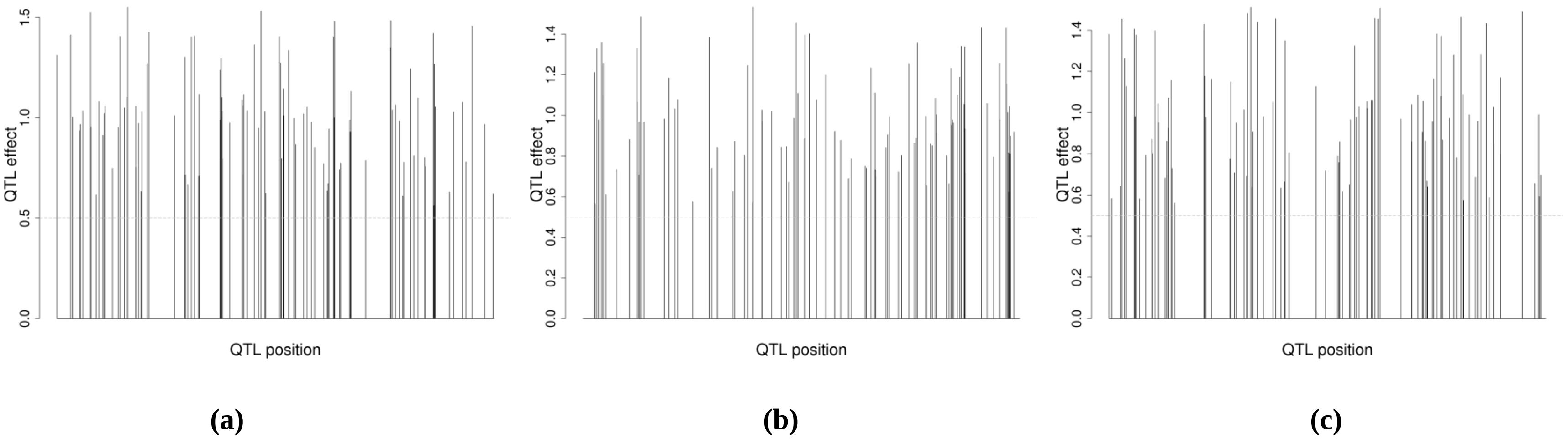


Figure 3.3. Distribution and effects of the 100 quantitative trait loci (QTLs) with predefined effects for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) marker (c) scenarios. QTL effects are expressed as percentage of genetic variance.

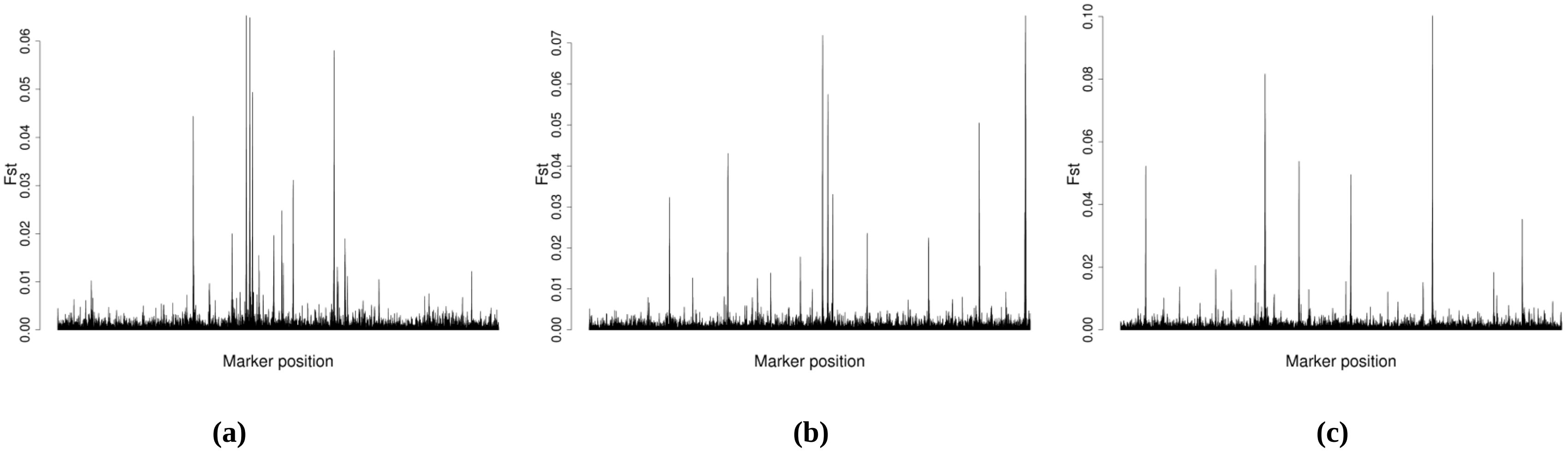


Figure 3.4. Distribution of estimated fixation index (F_{ST}) scores for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) markers (c) under the gamma distribution scenario.

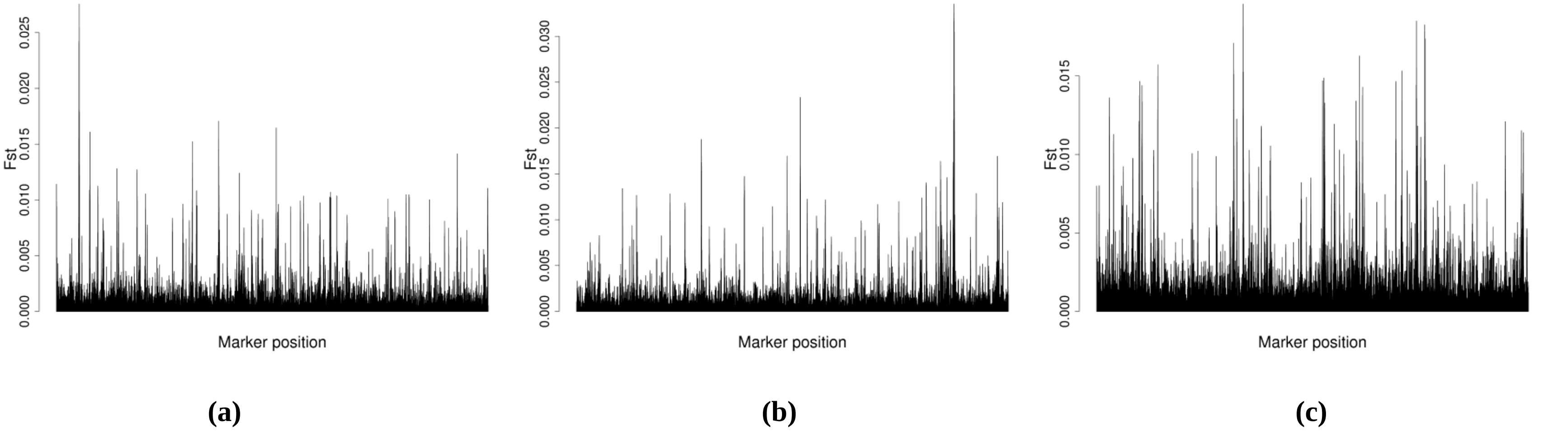


Figure 3.5. Distribution of estimated fixation index (F_{ST}) scores for the 200K (a), 300K (b), and 1 million single nucleotide polymorphism (SNP) markers (c) under the predefined effect scenario.

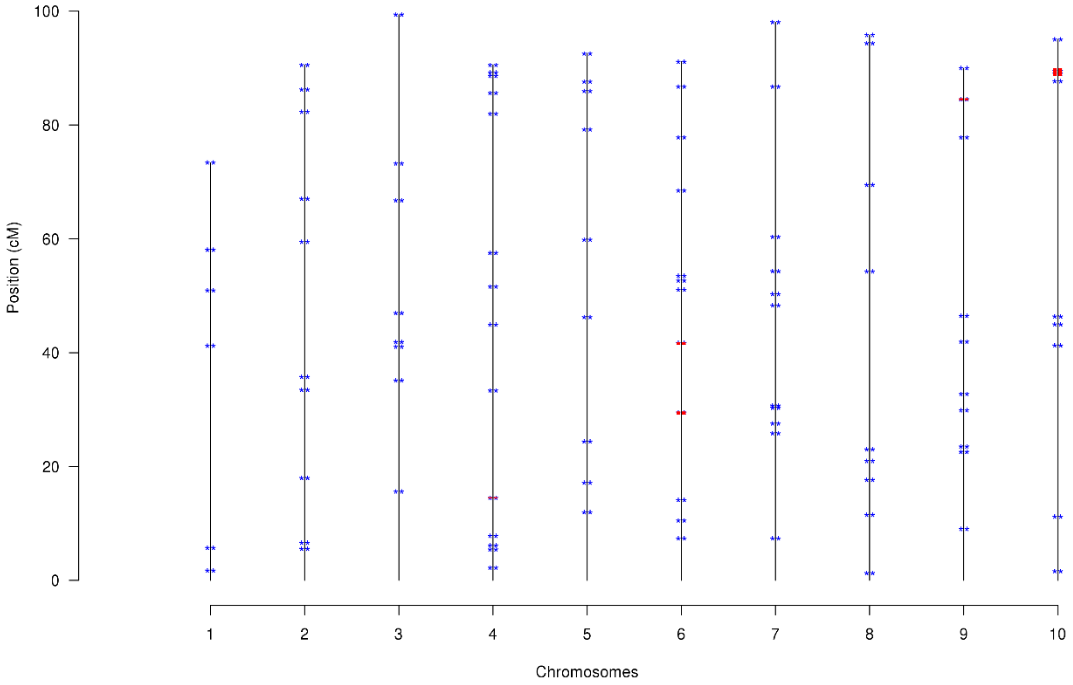


Figure 3.6. Map of simulated QTLs (in Blue) and selected SNPs (in Red) across the 10 chromosomes under the 300K SNP marker density, gamma distribution for QTL effects, and 99.5 quantile as cut-off point for F_{ST} scores.

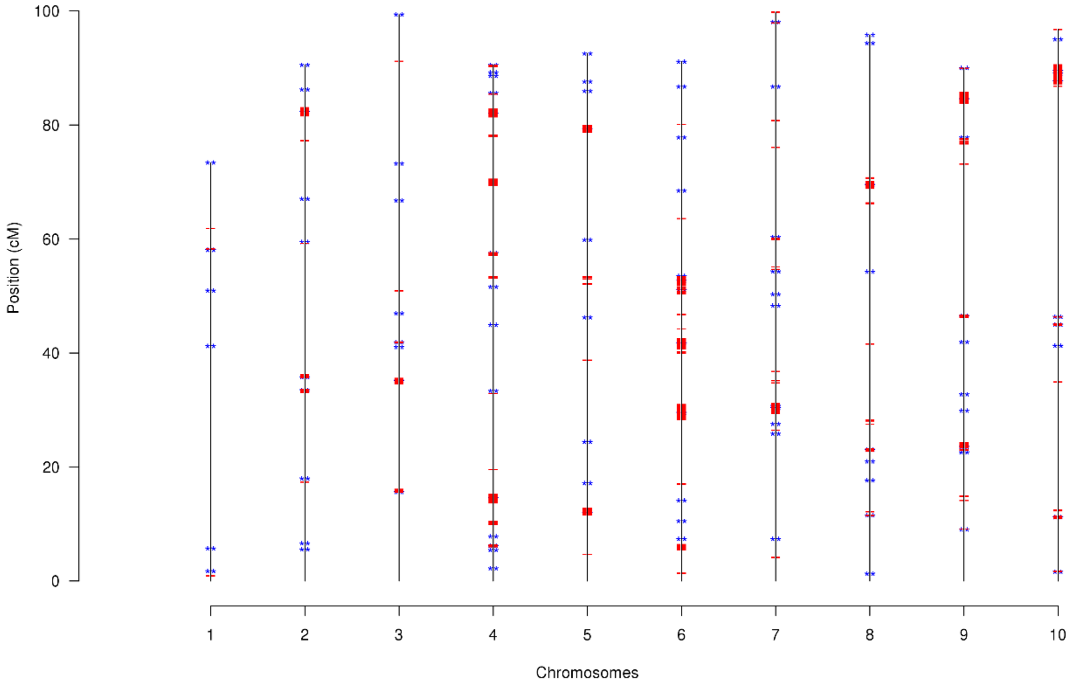


Figure 3.7. Map of simulated QTLs (in Blue) and selected SNPs (in Red) across the 10 chromosomes under the 300K SNP marker density, gamma distribution for QTL effects, and 97.5 quantile as cut-off point for F_{ST} scores.

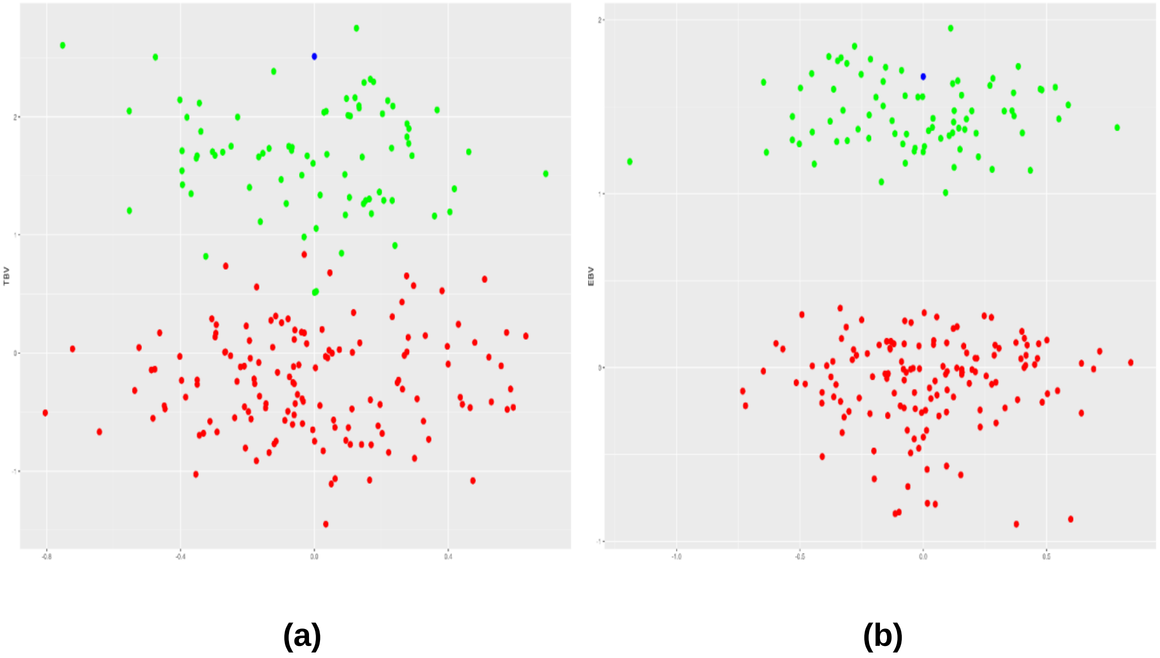


Figure 3.8. Distribution of true (a) and estimated (b) breeding values of animals with high (>0.9; Green) and low (< 0.55; Red) genomic similarity with selected individual (in Blue) under Gamma distribution, 300K SNPs, and 99.5% quantile of F_{ST} distribution simulation scenario**.**

CHAPTER 4

IMPLEMENTING A HYBRID MODEL FOR GENOMIC SELECTION USING PRIORITIZED SNPs AND POLYGENIC EFFECTS[2]

[2] Toghiani, S., L.Y. Chang, S.E. Aggrey, and R. Rekaya. To be submitted to Livestock Science.

Abstract

The vast majority of SNP markers in a given panel are not in high linkage disequilibrium (LD) with quantitative trait loci (QTL). Continuous increase in the density of marker panels has further reduced the statistical power in association analyses, resulting in limited to no improvement in the accuracy of genomic selection (GS). Similarly, increase in the number of genotyped animals has made the direct inversion of the genomic relationship matrix (**G**) impossible in some applications and before too long it will be the case for the majority of livestock and poultry populations. Although some data driven approximations of the inverse of **G** have been proposed, their optimality is not guaranteed. Furthermore, constructing the matrix **G** using all the available markers, on top of being computationally costly, will not improve accuracy and could even lead to lower performance. To overcome these challenges, a hybrid approach that uses only a limited number of prioritized variants and a polygenic component in the association model was proposed. The Fixation index (F_{ST}) scores were used to prioritize relevant markers that are potentially under selection pressure. Because the prioritized markers are unlikely to account for all the genetic variance, a polygenic component was added to the model. The effectiveness of the hybrid model was assessed by comparing its performance to BayesB, BayesC and GBLUP using simulated and real data sets. A trait with heritability equal to 0.1 or 0.4 was simulated. Two hundred QTL sampled from predefined uniform distributions were generated. The real dataset consisted of weaning weight in a composite beef cattle population. In both simulated and real datasets, 1 and 2.5% of total SNPs were prioritized based on the quantile distribution of the F_{ST} scores. When the heritability was equal to 0.4 in simulated situation, the proposed hybrid model

increased accuracy by 10.1 to 11.5%, 9.2 to 10.1%, and 26.1 to 29% compared to BayesB, BayesC and GBLUP models, respectively. When the heritability was equal to 0.1 in simulated situation and only 1% of the markers were prioritized, GBLUP was superior to BayesB, BayesC, and the hybrid method. However, when 2.5% of markers were prioritized, the hybrid model outperformed all the other methods with a superiority ranging from 3.3 to 7.7%.

Introduction

The paper by Meuwissen et al. (2001) presented the general idea for the potential use of high marker maps to estimate breeding values through so-called genomic selection (GS). Such idea became a reality when the BovineSNP50k chip panel (Illumina Inc, San Diego, USA) became available in 2007. Today, genomic information is being systematically used to estimate genomically enhance breeding values (GEBV) for several livestock and poultry species. In fact, GS is becoming the standard tool for genetic evaluation due to the increase in accuracy and the substantial reduction in generation interval (VanRaden et al., 2009; Su et al., 2010; Schefers and Weigel, 2012; Su et al., 2012). Several methods based on multiple regression or mixed linear models have been developed to implement GS. Although these methods have different statistical and biological assumptions regarding the data generating process, they tend to yield similar results in most cases, at least when low to moderate density panels are used, and differences are largely due to the genetic architecture of the trait, the genetic relationships between individuals in the data, and the chosen prior information.

Continuous improvement in high throughput technologies and the dramatic decrease in genotyping and sequencing costs have substantially increased the number of genotyped animals for several livestock species, especially dairy cattle. In fact, more than a million dairy cattle animals were genotyped by 2016 (García-Ruiz et al., 2016) and it is anticipated that over 3 million Holsteins will be genotyped by 2021 (Decker, 2015). This increase in the number of genotyped animals will create major computational challenges for the different methods used to implement GS. More importantly, it will further complicate the already formidable task of inverting large genomic relationship matrices

(Faux et al., 2012; Aguilar et al., 2014; Misztal, 2016). A more immediate and pressing issue in GS is the explosion in the number of genotyped variants (rare and common) due to the major advances in next generation sequencing. Currently, an increasing number of animals are being genotyped with high density panels (> 250K SNP markers) or have their genome fully sequenced, resulting in tens of millions of genotyped variants. This dramatic increase in the number of genotyped markers represents a major challenge, especially for multiple regression based approaches for GS implementation. This is the case due to loss of statistical power, the high shrinkage of variant effect estimates, and the increase in computational costs. Collectively, these challenges have limited the benefits of high density and sequence data on the accuracy of genomic selection. In fact, little to no improvement was achieved using HD compared to low or moderate panels (Harris and Johnson, 2010; VanRaden et al., 2013). Additionally, most of the animals in genetic evaluations are still non-genotyped. Accommodating non-genotyped animals could be achieved as proposed by Fernando et al. (2014) where the genotypes and non-typed animals could be imputed based on the observed genotypes, and the average additive relationships matrix. Thus, the computational cost of imputing the missing genotypes will dramatically increase with the number of markers in the panels.

Consequently, including all or the majority of high density or sequence variants in the association model used in multiple regression approaches is statistically counterproductive and computationally almost non-tractable. Unfortunately, current methods used to prioritize “relevant” SNPs or variants based on statistical (BayesB, BayesR) or external biological (BayesRC) information are at best only marginally better. Filtering variants based on their effects (BayesB, BayesR) is bounded by the limited statistical power and is unlikely to be useful in the presence of sequence data. Prior biological information could be very useful for prioritizing variants, but unfortunately its abundance and quality are far from being acceptable for meaningful practical use.

Toghiani et al. (2017) showed that prioritizing SNPs using F_{ST} score, a measure of genetic differentiation, had the ability to track the majority of influential quantitative trait loci (QTL). Chang et al. (2018) reported that using SNPs prioritized based on their F_{ST} scores resulted in higher accuracy than when all SNP were used. Compared to BayesB and BayesC, the proposed prioritization was superior. Furthermore, Chang et al. (2018; unpublished results) showed that F_{ST} prioritized SNPs could increase accuracy when used (weighted or unweighted) in the computation of the genomic relationship matrix (**G**) with a mixed linear framework. Independently of the prioritization methods, the subset of selected SNPs will not explain the totality of the genetic variance (GV). In fact, the fraction of GV explained by the prioritized SNPs will depend on the heritability, number of QTL, and the genetic complexity of the trait. Additionally, accommodating non-genotyped animals in a multi-step procedure is non-trivial and could limit the utility of these prioritization methods. Accounting for the portion of GV that is not explained by the prioritized SNPs could be achieved by accommodating a "polygenic" component using either pedigree or genomic information. In other words, the breeding value will be decomposed into two components: 1) the part explained by the prioritized SNPs and 2) the "polygenic" components.

In this study, a hybrid (multiple regression and variance component) approach using F_{ST} prioritized SNPs was implemented and compared to GBLUP and Bayesian models. Simulated and real data were used for the assessment.

Material and Methods

Simulated Data

Population structure: Simulation was carried out using QMsim software (Sargolzaei and Schenkel, 2009). The simulation process consisted of two steps. In the first step, a historical population was generated. This population was initiated with 10,000 individuals and steadily decreased to 5,000 individuals after 1,000 generations. Then, the population size gradually increased for 250 generations to 17,000 individuals. The first step is carried out to initialize LD and to establish mutation-drift equilibrium in the historical generations. The mating was at random in the historical generations. In the second step of simulating the population structure, the founder population was generated and labelled as generation zero (G0). In this study, the G0 population was generated from the last historical generation based on 1,500 males and 15,000 females. The mating of these individuals was random, and no selection was considered at this step. After G0, three generations were simulated and the last one (G3) was used to evaluate the proposed approach. From G0 to G3, animals were selected based on their estimated breeding values (EBVs). Sex ratio in the progeny was maintained at 50% and one progeny per dam was assumed throughout. Two quantitative traits, one with low (0.1) and the other with moderate (0.4) heritability, were simulated. The true breeding value (TBV) of an individual was equal to the sum of the QTL additive effects. Because the inability of the QMsim to simulate systematic effects, two fixed effects with 100 and 4 levels were simulate separately. Phenotypes were generated by adding fixed effects to the TBVs and the random residual terms. The simulation process was replicated five times.

Genome structure: A 30-chromosome genome, each with 100 centimorgans (cM) in length, was simulated with uniformly distributed 50K SNP markers to mimic a medium density marker panel for bovine. Two hundred QTL were simulated with their effects generated from uniform distributions to explain a predefined fraction of the total genetic variance. Specifically, 40 QTL were assumed to explain 1% to 1.5% of the genetic variance each $U \sim [1, 1.5]$ and the remaining 160 QTL were simulated from $U \sim [0.2, 0.5]$ so that each of them will explain between 0.2 to 0.5% of the genetic variance. Both SNP markers and QTL in all simulated scenarios were assumed to be bi-allelic, and no marker loci overlapped with the QTLs. Further, it was assumed that both SNP markers and QTLs have the same allele frequency in the historical population. The desired level of LD between markers was created based on the simulated historical population.

Real Data

The real data used in this study consisted of weaning weight (WW) records of 3,012 animals from a Composite Gene Combination breed (CGC; 50% Red Angus, 25% Charolais, 25% Tarentaise) born between 2002 and 2011 at USDA-ARS, Fort Keogh Livestock and Range Research Laboratory, Miles City, MT (Newman et al. (1993a, 1993b) The pedigree file consisted of 5,374 animals including 128 sires and 1,723 dams. The range of WW records was between 110.22 and 303.91 kg. Moreover, the mean and standard deviation of WW records were 209.58 and 30.73 kg, respectively. The systematic effects associated with this data consisted of sex (2 classes), feeding treatment (2 classes), year of birth (10 classes) and three covariates: age of dam, age at weaning weight, and birth weight.

A total of 4,457 CGC animals born between 2001 and 2015 were genotyped with a mixture of different density SNP arrays (Table 4.1). Across the different arrays, SNPs with call rate smaller than 0.90, minor allele frequency (MAF) less than 0.05, and heterozygous deviation greater than 15% from Hardy-Weinberg Equilibrium (HWE) were removed. In addition, animals with call rate less than 0.90 were also discarded. Number of animals and SNPs remaining after quality control (QC) edits are presented in Table 4.1. Animals genotyped with low-density panels were imputed to the 50K SNP array using FImpute software (Sargolzaei et al., 2011) where population and pedigree information were used in the imputation process. FImpute was implemented using default parameters in all imputation analyses. Each group of animals genotyped with a specific array was imputed separately. In all cases, animals genotyped with the 50K SNP panel were used as reference. Furthermore, SNP markers present in low density arrays but not in the 50K SNP array were removed. After imputation, the same QC process indicated before was reapplied, resulting in a dataset of 3,902 animals (1,387 males and 2,516 females) with genotype information on 41,694 SNPs. The total number of animals with marker genotypes and WW records was 2,193.

SNP prioritization via Fst scores

F_{ST} scores (Wright, 1951), a measure of population differentiation, were used to prioritize SNPs following Toghiani et al. (2017) and Chang et al. (2018). Briefly, the genotyped population was divided into three sub-populations based on the distribution of the trait phenotype (below the 10% quantile [S1], between 10 and 90% quantiles [S0], and above the 90% quantile [S2]). Subpopulations S1 and S2 were used to estimate the differentiation values using the global F_{ST} estimator method proposed by Nei (1973). For a given locus, *k*, the global F_{ST} value is calculated as:

$$F_{ST_k} = \frac{H_{T_k} - H_{SW_k}}{H_{T_k}}$$

with $H_{SW_k} = \frac{H_{S1_k} * n_{s1} + H_{S2_k} * n_{s2}}{n_{s1} + n_{s2}}$, $H_{T_k} = 2 * p_k * q_k$ and $H_{Si_k} = 2 * p_{Si_k} * q_{Si_k}$

where, p_{Si_k} and q_{Si_k} are the allele frequencies for locus k in subpopulation i of locus k, n_{s1} and n_{s2} are the number of individuals per first and second subpopulation, H_{SW_k} is the weighted mean heterozygosity across the first and second sub-populations and H_{T_k} is the heterozygosity of the pooled subpopulations for locus k. In this study, F_{ST} threshold values were heuristically determined to select SNPs under selection pressure. For that purpose, only top 97.5% and 99% quantiles of the F_{ST} distribution were used in the association model for the hybrid method in both simulated and real data sets.

Data analysis

The following mixed linear model that includes the prioritized SNPs and the polygenic components was used:

$$\boldsymbol{y_i} = \boldsymbol{w_i}\boldsymbol{\alpha} + \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{u_i} + \boldsymbol{e_i} \qquad [1]$$

where $\boldsymbol{y_i}$ is the phenotype for animals i, $\boldsymbol{\alpha}$ is the vector of systematic effects, $\boldsymbol{\beta}$ is the vector of the effects of the prioritized SNPs, $\boldsymbol{u_i}$ and $\boldsymbol{e_i}$ is the polygenic and random residual effects for animal i, respectively. $\boldsymbol{x}_i$ is the vector of genotypes of the prioritized SNPs for animal i and $\boldsymbol{w_i}$ is a known incidence vector relating the phenotype to the systematic effects.

Hierarchical Bayesian implementation

Let $v_i = \boldsymbol{x}_i\boldsymbol{\beta}$, the model in equation [1] can be implemented using a two-stage hierarchical Bayesian model. In the first stage, the conditional distribution of data ($\boldsymbol{y}$) adjusted for the genomic contributions, $\widehat{\boldsymbol{v}} = (\hat{v}_1, \hat{v}_2, \dots, \hat{v}_n)$, is given by:

$$(\boldsymbol{y} - \widehat{\boldsymbol{v}})|\boldsymbol{\alpha}, \boldsymbol{u}, \sigma_e^2 \sim N(\boldsymbol{W\alpha} + \boldsymbol{Zu}, \boldsymbol{I}\sigma_e^2) \qquad [2]$$

where $(\boldsymbol{y} - \widehat{\boldsymbol{v}})$ is the vector of adjusted phenotypes, $\boldsymbol{\alpha}$ is the vector of systematic effects, $\boldsymbol{u}$ is the vector to polygenic effects distributed as $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_u^2)$. σ_u^2 and σ_e^2 are the is the polygenic and residual variances, respectively. W and Z are known incidence matrices with the appropriate dimensions.

In the second stage, the conditional distribution of data adjusted for the systematic and polygenic effects follows a multivariate normal given by:

$$\boldsymbol{y}^*|\boldsymbol{X}, \boldsymbol{\beta}, \sigma_e^2 \sim N(\boldsymbol{Hv}, \boldsymbol{I}\sigma_e^2) \qquad [3]$$

where $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{\alpha}} - \boldsymbol{Z}\widehat{\boldsymbol{u}})$, $\boldsymbol{v} = (v_1, v_2, \dots v_n)'$, $\mathbf{X}$ is the matrix of genotypes of the prioritized SNPs, and $\mathbf{H}$ is a known incidence matrix.

The estimated breeding value base on the hybrid model is given by:

$$EBV_i = \boldsymbol{x_i}\widehat{\boldsymbol{\beta}} + \hat{u}_i$$

where $\hat{\beta}$ is the vector of estimated effects for the prioritized SNPs and x_i is the associated vector of genotypes for animal *i*, and $\hat{u}_i$ is the polygenic effect for animal *i*.

The Bayesian implementation of the model presented in equations [1-3] is straightforward, as all conditional distributions were in closed forms. To evaluate the performance of the hybrid model, the real and simulated data sets were analyzed BayesB and BayesC (with π set equal to 0.99 and 0.975) using Gensel software (Fernando and Garrick, 2009) and GBLUP using BLUPf90 program (Misztal et al., 2016) were compared. For the simulated data, 10K and 5K animals were randomly assigned to the training and validation sets, respectively. For the real data, a five-fold cross validation was implemented where each time 80% of the data was used for training and the remaining 20% of the data was used for validation. Accuracy was calculated as the correlation between true and estimated breeding values for the simulated data. For the real data, accuracy was calculated as the correlation between the adjusted phenotypes (for the systematic effects) and the estimated breeding values.

Results and Discussions

Tables 4.2 and 4.3 present the estimates of the variance components and heritabilities using the different approaches. When the true heritability in the base population was equal to 0.4, the highest estimates of genetic variance and heritability were obtained using the GBLUP method where all 50K SNPs were included in the calculation of the genomic relationship matrix. For the hybrid method, BayesB, and BayesC, the portion of genetic variance explained by genomic contribution increased with the increase in the number of selected SNPs (Table 4.2). The estimates of the genetic variance explained by the selected SNPs ranged from 0.21 to 0.23 when only 1% of the SNPs were included in the association model and 0.27 to 0.28 when 2.5% of the SNPs were prioritized. However, for the hybrid method the estimates of the polygenetic variance were 0.10 and 0.03 when 1 and 5% of SNPs were prioritized, respectively. Thus, the estimated genetic variance using the hybrid method ranged between 0.30 and 0.31 which is very similar to the estimate obtained using GBLUP. When π was equal to 0.99, the hybrid method estimate of genetic variance was substantially higher than estimates obtained using BayesB (0.23) and BayesC (0.22). In all cases, there was an overestimation of the residual variance due to the underestimation of the genetic variance. This overestimation ranged between 0.07 for GBLUP to 0.16 for BayesB (π = 0.99). For the hybrid method, the estimate of residual variance was equal to 0.69 (Table 4.2). Due to the underestimation of the genetic variance and the overestimation of the residual variance, the heritability was severely underestimated, especially when only 1% of the SNPs were prioritized.

For the scenario where the heritability was equal to 0.1, a similar trend was observed. The genetic variance and heritability were underestimated, and the residual variance tended to be overestimated, except when the hybrid model was used (Table 4.3). Across two different heritability scenarios, the proposed hybrid model with 1 and 2.5% of prioritized SNPs captured a larger portion of the genetic variance compared to their Bayesian model counterparts (Tables 4.2 and 4.3). Several studies using livestock genomic data (Tsuruta et al., 2011; Jensen et al., 2012; Haile-Mariam et al., 2013) have noted that the fraction of the genetic variance explained by the markers in the panel ranged from 35 to 96%. Obviously, the variation depends on the complexity of the trait, the heritability, and the number of SNP markers in the panel. In human, the percentage of the genetic variance explained by the SNP markers tends to be smaller than in livestock and plant applications. In fact, Yang et al. (2010) showed that at best only 45% of additive genetic variance in human height was explained by approximately 300K SNP markers. These differences are mainly due to the smaller effective population size in selected livestock populations compared to humans. Additionally, selected livestock populations, especially in dairy cattle, are well structured and stratified. In this study, 80% of the QTL were assumed to explain between 0.2 and 0.5% of the genetic variance each. It is very likely that a large portion of these QTL were not effectively tagged by the selected SNPs which explain the underestimation of the genetic variance.

Accuracy, defined as the correlation between true and predicted breeding values, when the heritability was equal to 0.4 and 0.1 is presented in Figures 4.1 and 4.2, respectively. When the heritability was equal to 0.4, the proposed hybrid model increased accuracy by 11.1 to 11.3%, 9.7 to 10.4%, and 25.6 to 30.4% compared to BayesB, BayesC and GBLUP models, respectively (Figure 4.1). When the heritability was equal to 0.1, GBLUP was superior to BayesB, BayesC, and the hybrid method for π equal to 0.99 (Figure 4.2). However, when π was equal 0.975, the prediction hybrid model outperformed all the other methods with a superiority ranging from 7 (compared GBLUP) to 11.2% (compared to GBLUP).

Several factors were reported to affect the accuracy of GS (Hayes et al., 2009; Zhong et al., 2009; Daetwyler et al., 2012). These factors include the genetic architecture of the trait, the relatedness between training and validation populations, the marker density, the size of training population, the LD between SNPs and QTL, and the heritability of the trait. Some of these factors correlate directly with the fraction of the genetic variance to be explained by the SNP markers. Thus, the superiority of the hybrid method, at least compared to BayesB and BayesC, could be due to the fact that a larger number of QTL were tracked using the F_{ST} scores. This argument does not seem to be valid when comparing the results between the hybrid and GBLUP methods. However, Chang et al. (2018) showed that accuracy depends on a balance between the percentage of genetic variance explained by the prioritized SNPs and the genetic similarity. Thus, although GBLUP explained a slightly larger portion of the genetic variance, it is likely that the hybrid method resulted in a much higher genetic similarity. Additionally, the GBLUP method assumes that all the genetic variance is captured by the SNPs. A violation of such assumption will result in lower prediction accuracy (Kemper and Goddard, 2012). BayesB and BayesC are variable selection approaches that work well for traits with a genetic architecture that includes large or moderate effect SNPs (Wimmer et al., 2013). However, in presence of small QTL effect, identifying relevant linked SNPs (based on magnitude of their effects) becomes more challenging. On the other hand, the hybrid method does not select markers based on their effects but rather based on the change of their allele frequencies due to selection pressure. This difference in the prioritization is extremely important in presence of small QTL effect.

Table 4.4 presents the estimates of the variance components and heritability using the real data. Estimates of WW heritability ranged between 0.25 and 0.27 using Bayesian methods. As with the simulated data, the estimated genetic variance and heritability increased with the increase in the number of prioritized SNPs. Similar pattern was observed using the hybrid method, although estimates of the total genetic variance and heritability were higher compared to the Bayesian methods. Using GBLUP, the estimate of heritability was higher compared to the Bayesian methods but slightly smaller compared to the hybrid approach (Table 4.4). Across the different methods, estimates of heritability were within the range of estimates reported in the literature (de Mattos et al., 2000; Vargas et al., 2000; Pico et al., 2004; Gutiérrez et al., 2007; Dezfuli and Mashayekhi, 2009; Vergara et al., 2009; Chud et al., 2014). Accuracy, defined as the correlation between adjusted phenotypes and predicted breeding values based on a five-fold cross validation, is presented in Figure 4.3. The proposed hybrid approach outperformed all competing methods with an accuracy of 0.36. Bayesian methods and GBLUP had an accuracy of around 0.27.

Table 4.5 presents the CPU time and peak memory required by the different methods. The hybrid model has significantly lower computational costs. In fact, when 2.5% of the SNPs were prioritized, it required only 120 minutes of CPU time compared to 745, 1011, 954 minutes for GBLUP, BayesB and BayesC, respectively. Comparisons were based on a single chain of 100K rounds for the Bayesian and hybrid methods. Compared to the Bayesian methods, the hybrid approach reduced the CPU time by 6 to 15 folds and this superiority could further increase in presence of higher density panels or/and larger number of genotyped animals. Compared to GBLUP, the hybrid model reduced the computational cost by 6 to 11 folds. It is expected that the computational advantage of the hybrid method is increased with the increase of the number of genotyped animals due the significant increase in the costs of inverting (or approximating the inverse) of the **G** matrix for the GBLUP method. Regarding the peak memory required to implement, the hybrid method required an insignificant amount of memory compared to the Bayesian and GBLUP methods. In fact, it only required 5-10% and 0.5-1% of the peak memory needed by the Bayesian methods and GBLUP, respectively (Table 4.5).

Conclusions

The substantial increase in the number of genomic variants or/and genotyped individuals is creating serious challenges for the implementation of genomic selections. The lack of improvement in prediction accuracy using high density or sequence data clearly highlights the need to prioritize markers to be included in the association model or to compute the genomic relationship matrix. Continuous increase in the genotyped animals is further complicating the inversion of **G**. Variable selection models are computationally intensive. Furthermore, they prioritize SNPs based on their relative effects, which is inefficient in presence of high density marker panels or/and when large and moderate effect QTL explain only a small fraction of the total genetic variance. SNP prioritization based on changes in allele frequencies due to selection is an attractive alternative. However, in presence of large number of small effect QTL, a hybrid method that makes use of the prioritized SNPs and a polygenic component could be advantageous. Using simulated and real data, the hybrid model was superior to the competing methods in terms of accuracy of prediction. Furthermore, the proposed hybrid model has low computational costs compares to BayesB, BayesC and GBLUP. It requires no inversion of any matrix other than the average relationship matrix.

References

Aguilar, I., I. Misztal, S. Tsuruta, A. Legarra, and H. Wang. 2014. PREGSF90–POSTGSF90: computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped individuals in BLUPF90 programs. Proc. 10th World Congr. Genet. Appl. Livest. Prod.

Chang, L.-Y., S. Toghiani, A. Ling, S. E. Aggrey, and R. Rekaya. 2018. High density marker panels, SNPs prioritizing and accuracy of genomic selection. BMC genetics 19: 4.

Chud, T. C. et al. 2014. Genetic analysis for gestation length, birth weight, weaning weight, and accumulated productivity in Nellore beef cattle. Livestock Science 170: 16-21.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2012. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193: 347-365.

de Mattos, D., I. Misztal, and J. K. Bertrand. 2000. Variance and covariance components for weaning weight for Herefords in three countries. Journal of Animal Science 78: 33-37.

Decker, J. E. 2015. Agricultural Genomics: Commercial Applications Bring Increased Basic Research Power. PLoS Genet 11.

Dezfuli, B. T., and M. R. Mashayekhi. 2009. Genetic study of birth weight and weaning weight in Najdi calves. Journal of Animal and Veterinary Advances 8: 276-280.

Faux, P., N. Gengler, and I. Misztal. 2012. A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. J Dairy Sci 95: 6093 - 6102.

Fernando, R., J. Dekkers, and D. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution 46: 50.

Fernando, R., and D. Garrick. 2009. GenSel–user manual for a portfolio of genomic selection related analyses, create 9.1.

García-Ruiz, A. et al. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proceedings of the National Academy of Sciences 113: E3995-E4004.

Gutiérrez, J. P., F. Goyache, I. Fernández, I. Alvarez, and L. J. Royo. 2007. Genetic relationships among calving ease, calving interval, birth weight, and weaning weight in the Asturiana de los Valles beef cattle breed1. Journal of Animal Science 85: 69-75.

Haile-Mariam, M., G. J. Nieuwhof, K. T. Beard, K. V. Konstatinov, and B. J. Hayes. 2013. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J Anim Breed Genet 130: 20-31.

Harris, B., and D. Johnson. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. Interbull Bulletin: 40.

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

Jensen, J., G. Su, and P. Madsen. 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC genetics 13: 44.

Kemper, K. E., and M. E. Goddard. 2012. Understanding and predicting complex traits: knowledge from cattle. Hum Mol Genet 21: R45-51.

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

Misztal, I. 2016. Inexpensive Computation of the Inverse of the Genomic Relationship Matrix in Populations with Small Effective Population Size. Genetics 202: 401-409.

Misztal, I. et al. 2016. Manual for BLUPF90 family of programs. Retrieved from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all5.pdf.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences 70: 3321-3323.

Newman, S., M. MacNeil, W. Reynolds, B. Knapp, and J. Urick. 1993a. Fixed effects in the formation of a composite line of beef cattle: I. Experimental design and reproductive performance. J. Anim. Sci. 71: 2026-2032.

Newman, S., M. MacNeil, W. Reynolds, B. Knapp, and J. Urick. 1993b. Fixed effects in the formation of a composite line of beef cattle: II. Pre-and postweaning growth and carcass composition. J. Anim. Sci. 71: 2033-2039.

Pico, B., F. Neser, and J. Van Wyk. 2004. Genetic parameters for growth traits in South African Brahman cattle. S. Afr. J. Anim. Sci. 34: 44-46.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2011. FImpute - An efficient imputation algorithm for dairy cattle populations. J. Dairy Sci. 94(E-Suppl. 1): 421.

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680-681.

Schefers, J. M., and K. A. Weigel. 2012. Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Animal Frontiers 2: 4-9.

Su, G. et al. 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: 4657-4665.

Su, G., B. Guldbrandtsen, V. Gregersen, and M. Lund. 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. Journal of Dairy Science 93: 1175-1183.

Toghiani, S., L.-Y. Chang, A. Ling, S. E. Aggrey, and R. Rekaya. 2017. Genomic differentiation as a tool for single nucleotide polymorphism prioritization for Genome wide association and phenotype prediction in livestock. Livestock Science 205: 24-30.

Tsuruta, S., I. Misztal, I. Aguilar, and T. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J Dairy Sci 94.

VanRaden, P. et al. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92: 16 - 24.

VanRaden, P. M. et al. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science 96: 668-678.

Vargas, C. A., M. A. Elzo, J. C. C. Chase, and T. A. Olson. 2000. Genetic parameters and relationships between hip height and weight in Brahman cattle. Journal of Animal Science 78: 3045-3052.

Vergara, O. D., M. F. Ceron-Muñoz, E. M. Arboleda, Y. Orozco, and G. A. Ossa. 2009. Direct genetic, maternal genetic, and heterozygosity effects on weaning weight in a Colombian multibreed beef cattle population12. Journal of Animal Science 87: 516-521.

Wimmer, V. et al. 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195: 573-587.

Wright, S. 1951. The genetical structure of populations. Annals of eugenics 15: 323-354.

Yang, J. et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genet 42: 565 - U131.

Zhong, S., J. Dekkers, R. Fernando, and J. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182: 355 - 364.

Table 4.1. Distribution of typed animals across the different genotyping platforms before and after quality control (QC)

	Raw Data[1]		QC Data[2]	
	# Animals	# SNP	# Animals	# SNP
50K SNP array	64	54,166	88	42,264
	24	54,209		
27K SNP array	380	25,856	790	8,126
	14	25,948		
	96	25,890		
	326	25,887		
20K SNP array	396	19,642	379	7,945
9K SNP array	391	8,727	909	6,754
	185	8,781		
	344	8,777		
3K SNP array	1,944	2,866	1,739	2,727
	197	2,877		
	96	2,882		

[1] 4,457 genotyped animals before the QC

[2] 3,902 genotyped animals remained after the QC

Table 4.2. Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with $h^2 = 0.4$

	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_{poly}^2}$	$\boldsymbol{\sigma_e^2}$	h^2
hybrid.fst(0.99)	0.21	0.10	0.69	0.31
hybrid.fst(0.975)	0.27	0.03	0.69	0.30
BayesB(0.99)	0.23	--	0.76	0.23
BayesB(0.975)	0.28	--	0.70	0.29
BayesC(0.99)	0.22	--	0.76	0.22
BayesC(0.975)	0.28	--	0.70	0.29
GBLUP	0.31	--	0.67	0.32

$\boldsymbol{\sigma_g^2}$= genomic variance;$\boldsymbol{\sigma_{poly}^2}$= polygenic variance; $\boldsymbol{\sigma_e^2}$= residual variance; h^2= heritability.

Table 4.3. Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with $h^2 = 0.1$

	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_{poly}^2}$	$\boldsymbol{\sigma_e^2}$	**h^2**
hybrid.fst(0.99)	0.077	0.025	0.865	0.105
hybrid.fst(0.975)	0.080	0.009	0.851	0.095
BayesB(0.99)	0.074	--	0.926	0.074
BayesB(0.975)	0.087	--	0.912	0.087
BayesC(0.99)	0.065	--	0.931	0.065
BayesC(0.975)	0.076	--	0.920	0.076
GBLUP	0.077	--	0.912	0.078

$\boldsymbol{\sigma_g^2}$= genomic variance;$\boldsymbol{\sigma_{poly}^2}$= polygenic variance; $\boldsymbol{\sigma_e^2}$= residual variance; h^2= heritability.

Table 4.4. Estimates of variance components and heritability for weaning weight using different methods

	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_{poly}^2}$	$\boldsymbol{\sigma_e^2}$	**h^2**
hybrid.fst(0.99)	76.35	71.35	306.12	0.325
hybrid.fst(0.975)	80.13	66.56	299.87	0.328
BayesB(0.99)	105.80	--	326.27	0.245
BayesB(0.975)	112.44	--	319.24	0.260
BayesC(0.99)	115.01	--	325.23	0.261
BayesC(0.975)	119.17	--	320.99	0.271
GBLUP	128.08	--	306.69	0.295

$\boldsymbol{\sigma_g^2}$= genomic variance;$\boldsymbol{\sigma_{poly}^2}$= polygenic variance; $\boldsymbol{\sigma_e^2}$= residual variance; h^2= heritability.

Table 4.5. CPU time and peak memory of the different methods used to implement genomic selection (simulated data; $h^2 = 0.4$)

	CPU (min)[1]	**Peak memory**
BayesB(0.99)	964.03	2.39
BayesB(0.975)	1011.48	2.39
BayesC(0.99)	756.22	2.39
BayesC(0.975)	954.47	2.39
GBLUP	745.52	26.65
hybrid.fst(0.99)	65.02	0.125
hybrid.fst(0.975)	119.37	0.24

[1] Bayesian and hybrid method were implemented based on a single chain of 100,000 rounds

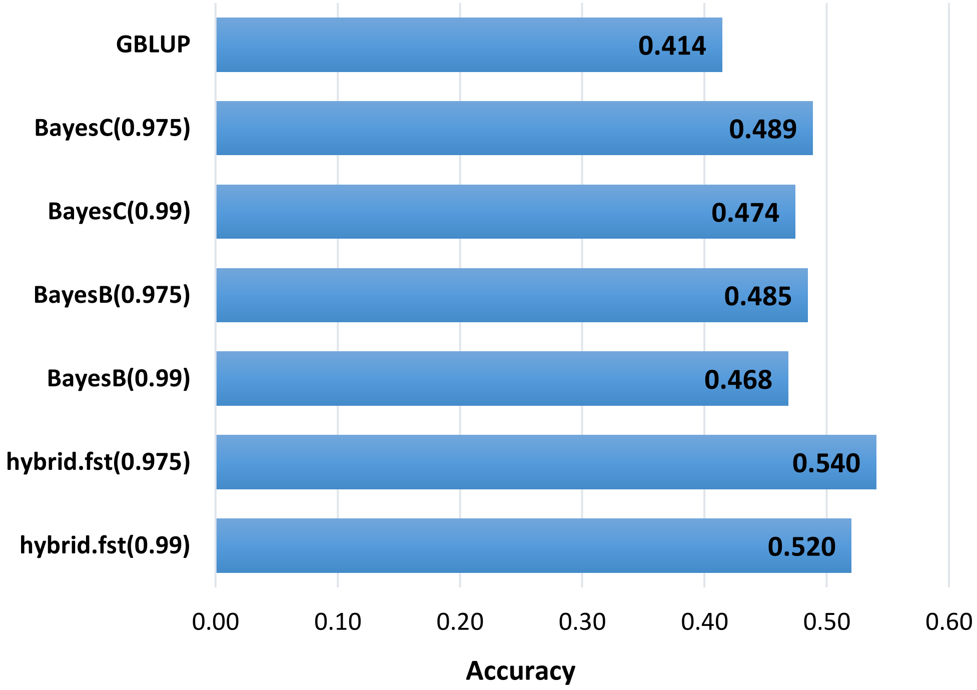


Figure 4.1. Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with $h^2 = 0.4$

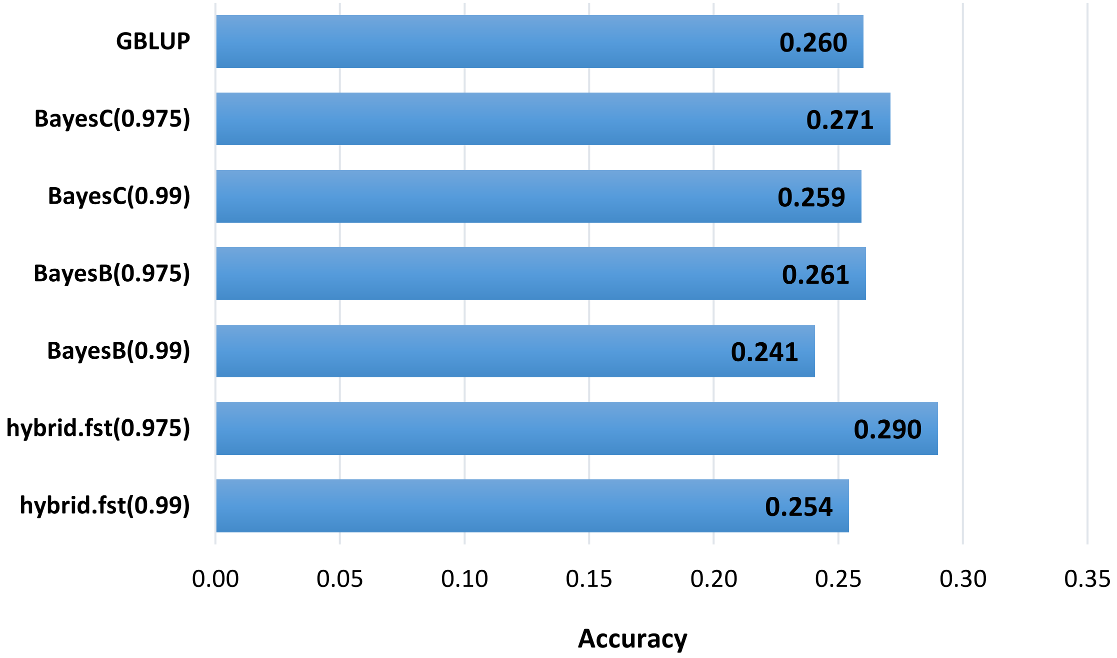


Figure 4.2. Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with $h^2 = 0.1$

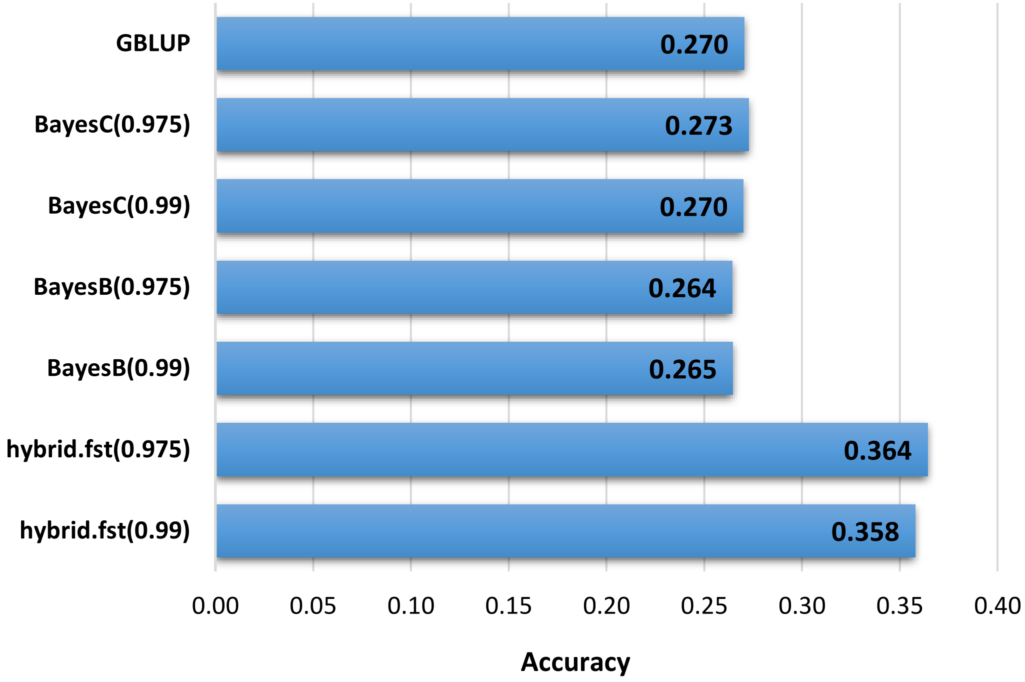


Figure 4.3. Accuracy of estimated breeding values for weaning weight in a composite beef breed using different methods

CHAPTER 5

EXTENSION OF THE HYBRID MODEL TO ACCOMMODATE NON-GENOTYPED ANIMALS [3]

[3] Toghiani, S., L.Y. Chang, S.E. Aggrey, and R. Rekaya. To be submitted to Livestock Science.

Abstract

The dramatic advance in genotyping and sequencing technology has greatly reduced the complexity and cost of genotyping. This has led to a significant increase in the number of genotyped variants and typed individuals, resulting in major complications for the implementation of genomic selection (GS). Furthermore, continuous increase in the number of variants did not improve the accuracy of GS. Increase in the number of genotyped animals severely impacted the cost of inverting the genomic relationship matrix (**G**). In spite of the substantial increase in the number of genotyped animals, the majority of animals included in any genetic evaluation are not genotyped. Including these animals in a genomic evaluation requires the imputation of their missing genotype using linear regression methods. To overcome these issues, the hybrid approach was extended to accommodate non-genotyped animals. Only markers prioritized using the fixation index (F_{ST}) scores were included in the association model. Because the prioritize markers are unlikely to account for all the genetic variance, a polygenic component was added. An asymmetric prior was used to account for the genomic contribution of the prioritized marker for non-genotyped animals with the need to impute their missing genotypes. The effectiveness of the hybrid model with the extension to non-genotyped animals was assessed by comparing its performance to ssGBLUP using simulated data sets. A trait with heritability equal to 0.1 or 0.4 was simulated. Two hundred QTL sampled from predefined uniform distributions were generated. Either 1 or 2.5% of total SNPs were prioritized based on the quantile distribution of the F_{ST} scores. When the heritability was equal to 0.4, the proposed hybrid model resulted in an accuracy of 0.46-0.48 compared to 0.29 for the ssGBLUP model. When the heritability was equal to 0.1, the hybrid model

outperformed ssGBLUP with a superiority ranging from 15 to 27%. The hybrid method required only a small fraction of the computation needed to implement ssGBLUP.

Introduction

The theoretical basis of genomic selection (GS) was represented by Meuwissen et al. (2001). However, practical implementation of GS was not possible until 2008 when the BovineSNP50k chip panel became available for dairy cattle (Matukumalli et al., 2009). The use of SNP information allows for the prediction of genomically enhanced breeding values (GEBV) through the use of marker information, pedigree and phenotypic records. The rapid decrease in genotyping and whole genome sequencing costs led to a spectacular increase in the number of genotyped variants and animals. In fact, García-Ruiz et al. (2016) analyzed a dairy data set consisting of more than one million genotyped animals. Decker (2015) anticipated that around 3 million genotyped Holstein by 2021. These simultaneous increases in the number of variants and genotyped individuals created a major computational challenge for GS implementation. This challenge affects both the mixed linear (ML) and the linear regression (LR) implementation approaches. When the number of genotyped animals is in the order of the millions, the direct inversion of the genomic relationship matrix, needed for the implementation of the ML approach, becomes impossible. Although some approximations for the inverse have been presented (Misztal et al., 2014), their optimality is data driven. Increase in the number of genotyped variants, especially those with low minor allele frequency, could negatively impact the quality of the genomic relationship matrix and ultimately the GS implementation using ML. In fact, using high-density panels showed no improvement in prediction accuracy compared to medium-density chips (Harris and Johnson, 2010; VanRaden et al., 2013). These problems are not limited to the ML approach. In fact, they are even more extreme for LR methods, especially as a result of the increase in the number of genotyped variants. Including all variants of high-density panels or sequence data in the association model using LR approaches is statistically impractical and computationally cumbersome. Variable selection models (e.g., BayesB and BayesC), which use the magnitude of the marker effect to prioritize relevant SNPs, did not reduce the computational cost or improve accuracy significantly. In fact, in presence of high density marker data, these methods suffer from severe lack of statistical power. Using external information to prioritize SNP markers is an attractive alternative as it will improve the statistical power of LR methods. Unfortunately, the availability and quality of such data is at best limited. Furthermore, results by (MacLeod et al., 2016; Fang et al., 2017) showed an insignificant increase of accuracy using gene expression data as external information to prioritize SNPs.

Despite the significant increase in the number of genotyped animals, the vast majority of animals included in any genetic evaluation are non-genotyped animals. Although the ML approach can accommodate non-typed animals in a straightforward manner, that it is not the case for LR methods. Fernando et al. (2014) proposed a way to accommodate non-genotyped animals for GS procedure using LR approaches. Their study proposed predicting the GEBV of non-genotyped animals through the imputation of their missing genotypes using the information available of the typed animals. Although the idea is sound, it is computationally very demanding, making its implementation in real applications almost impossible.

Toghiani et al. (2017) showed that using F_{ST}, a measure of population differentiation, the majority of significant QTL could be tracked using the prioritized SNPs. Chang et al. (2018) reported higher prediction accuracy using only the F_{ST}

prioritized in the association model instead of including all SNPs. Furthermore, Toghiani et al. (2018; unpublished results) reported that using the hybrid model resulted in higher prediction accuracy compared to LM and LR approached using only genotyped animals. Due to the excess of non-genotyped over genotyped animals in commercial farms, extending the hybrid model to accommodate non-genotyped animals becomes more applicable for genetic evaluation.

The objective of this study is to extend the hybrid model to accommodate non-genotyped animals for implementation of GS using a two-stage hierarchical model. Furthermore, the effectiveness of this proposed method will be evaluated using simulation data and compared to single-step GBLUP.

Material and Methods

Simulated Data

Population structure: Simulation was carried out using QMsim software (Sargolzaei and Schenkel, 2009). The simulation process consisted of two steps. In the first step, a historical population was generated. This population was initiated with 10,000 individuals and steadily decreased to 5,000 individuals after 1,000 generations. Then, the population size gradually increased for 250 generations to reach 17,000 individuals. The first step is carried out to initialize LD and to establish mutation-drift equilibrium in the historical generations. The mating was at random in the historical generations. In the second step of the simulation of the population structure, the founder population was generated and labelled as generation zero (G0). In this study, the G0 population was generated from the last historical generation based on 1,500 males and

15,000 females. The mating of these individuals was at random and no selection was considered at this step. After G0, three generations were simulated and the last one (G3) was used to evaluate the proposed approach. From G0 to G3, animals were selected based on their estimated breeding values (EBVs). Sex ratio in the progeny was maintained at 50% and one progeny per dam was assumed throughout. Two quantitative traits, one with low (0.1) and the other with moderate (0.4) heritability, were simulated. The true breeding value (TBV) of an individual was set equal to the sum of the QTL additive effects. Because the inability of the QMsim to simulate systematic effects, two fixed effects with 100 and 4 levels were simulate separately. Phenotypes were generated by adding fixed effects to the TBVs and the random residual terms. The simulation process was replicated five times.

Genome structure: A 30-chromosome genome, each with 100 centimorgans (cM) in length, was simulated with uniformly distributed 50K SNP markers to mimic a medium density marker panel for bovine. Two hundred QTL were simulated with their effects generated from uniform distributions to explain a predefined fraction of the total genetic variance. Specifically, 40 QTL were assumed to explain 1% to 1.5% of the genetic variance, each $U{\sim}[1, 1.5]$, and the remaining 160 QTL were simulated from $U{\sim}[0.2, 0.5]$ so that each of them will explain between 0.2 to 0.5% of the genetic variance. Both SNP markers and QTL in all simulated scenarios were assumed to be bi-allelic, and no marker loci overlapped with the QTL. Further, it was assumed that both SNP markers and QTL have the same allele frequency in the historical population. The desired level of LD between markers was created based on the simulated historical population.

SNP prioritization via Fst scores

F_{ST} scores (Wright, 1951), a measure of population differentiation, were used to prioritize SNPs following Toghiani et al. (2017) and Chang et al. (2018). Briefly, the genotyped population was divided into three sub-populations based on the distribution of the trait phenotype (below the 10% quantile [S1], between 10 and 90% quantiles [S0], and above the 90% quantile [S2]). Subpopulations S1 and S2 were used to estimate the differentiation values using the global F_{ST} estimator method proposed by Nei (1973). For a given locus, *k*, the global F_{ST} value is calculated as:

$$F_{ST_k} = \frac{H_{T_k} - H_{SW_k}}{H_{T_k}}$$

with $H_{SW_k} = \frac{H_{S1_k} * n_{s1} + H_{S2_k} * n_{s2}}{n_{s1} + n_{s2}}$, $H_{T_k} = 2 * p_k * q_k$ and $H_{Si_k} = 2 * p_{Si_k} * q_{Si_k}$

where, p_{Si_k} and q_{Si_k} are the allele frequencies for locus *k* in subpopulation *i* of locus *k*, n_{s1} and n_{s2} are the number of individuals per first and second subpopulation, H_{SW_k} is the weighted mean heterozygosity across the first and second sub-populations and H_{T_k} is the heterozygosity of the pooled subpopulations for locus *k*. In this study, F_{ST} threshold values were heuristically determined to select SNPs under selection pressure. For that purpose, only top 97.5% and 99% quantiles of the F_{ST} distribution were used in the association model for the hybrid method in simulated data sets.

Data analysis

The following mixed linear model that includes the prioritized SNPs and the polygenic components was used to accommodate non-genotyped animals:

$$\boldsymbol{y_i} = \boldsymbol{w_i}\boldsymbol{\alpha} + \boldsymbol{v_i} + \boldsymbol{u_i} + \boldsymbol{e_i} \qquad \textbf{[1]}$$

where $\boldsymbol{y_i}$ is the phenotype for animal *i*, $\boldsymbol{\alpha}$ is the vector of systematic effects, $\boldsymbol{w_i}$ is a known incidence vector relating the phenotype to systematic effects of genotyped and non-genotyped animals, $\boldsymbol{u_i}$ and $\boldsymbol{e_i}$ is the polygenic and random residual term, respectively. The term $\boldsymbol{v_i} = \boldsymbol{x_i}\boldsymbol{\beta}$ is the genomic contribution of the prioritized SNPs, where $\boldsymbol{x_i}$ is the vector of genotypes of the prioritized SNPs for animal *i* and $\boldsymbol{\beta}$ is the vector of prioritized SNP effects.

The mixed linear model presented in equation [1] was extensively used in the field of animal breeding and it presents no major implementational challenges when all incidence vectors and matrices are completely known. In our case, the genotypes of the prioritized SNPs for the non-typed animals are unknown. Thus, the genomic contribution for non-genotyped animals cannot be computed.

Hierarchical Bayesian implementation

The model in equation [1] can be implemented using a two-stage hierarchical Bayesian model. In the first stage, the conditional distribution of data ($\boldsymbol{y}$) adjusted for the genomic contributions, $\hat{\boldsymbol{v}} = (\hat{v}_1, \hat{v}_2, \dots, \hat{v}_n)$, is given by:

$$(\boldsymbol{y} - \hat{\boldsymbol{v}})|\boldsymbol{\alpha}, \boldsymbol{u}, \sigma_e^2 \sim N(\boldsymbol{W\alpha} + \boldsymbol{Zu}, \boldsymbol{I}\sigma_e^2) \quad \textbf{[2]}$$

where $(\boldsymbol{y} - \hat{\boldsymbol{v}})$ is the vector of adjusted phenotypes, $\boldsymbol{\alpha}$ is the vector of systematic effects, $\boldsymbol{u}$ is the vector to polygenic effects distributed as $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_u^2)$. σ_u^2 and σ_e^2 are the is the polygenic and residual variances, respectively. $\boldsymbol{W}$ and $\boldsymbol{Z}$ are known incidence matrices with the appropriate dimensions.

In the second stage, the conditional distribution of data adjusted for the systematic and polygenic effects follows a multivariate normal given by:

$$\boldsymbol{y}^*|\boldsymbol{X},\boldsymbol{\beta},\sigma_e^2 \sim N(\boldsymbol{Hv},\boldsymbol{I}\sigma_e^2) \qquad \textbf{[3]}$$

where $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{\alpha}} - \boldsymbol{Z}\widehat{\boldsymbol{u}})$, $\boldsymbol{v} = (v_1, v_2, \dots v_n)'$, $\boldsymbol{X}$ is the matrix of genotypes of the prioritized SNPs, and $\boldsymbol{H}$ is a known incidence matrix.

Let $\boldsymbol{y}^* = (\boldsymbol{y}_1^*, \boldsymbol{y}_2^*)$ be the vector of adjusted (for the systematic effects) phenotypes for the genotyped and non-genotyped animals and $\boldsymbol{v} = (\boldsymbol{v_1}, \boldsymbol{v_2})$ be the corresponding vectors of the genomic contribution of genotyped and non-genotyped animals respectively, where $\boldsymbol{v_1} = (v_1, v_2, \dots v_{n1})$ and $\boldsymbol{v_2} = (v_{n1+1}, v_{n1+2}, \dots v_n)$. The adjusted phenotypes for the genotyped and non-genotyped animals could be written as:

$$\boldsymbol{y_1^*} = \boldsymbol{H_1 v_1} + \boldsymbol{e_1}$$

$$\boldsymbol{y_2^*} = \boldsymbol{H_2 v_2} + \boldsymbol{e_2}$$

The model for $\boldsymbol{y_1^*}$ is straightforward**.** However, the model for the non-genotyped animals ($\boldsymbol{y_2^*}$) is not identifiable.

Let $\boldsymbol{v}_1 = \boldsymbol{X_1\beta}$ be the vector of genomic contributions of the prioritized SNPs for the genotyped animals where $\boldsymbol{X_1}$ is the matrix of genotypes**.** To make the model for y_2^* identifiable, the following multivariate normal prior was assumed for $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$

$$p(\boldsymbol{v_1}, \boldsymbol{v_2}|\boldsymbol{A}) \sim \boldsymbol{N}(\boldsymbol{0}, (\boldsymbol{A}.\boldsymbol{D})\boldsymbol{\sigma_a^2}) \ \ \textbf{[4]}$$

where $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} \end{bmatrix}$ is the average additive relationship matrix, $\boldsymbol{D} = \begin{bmatrix} \gamma\boldsymbol{I_{n1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I_{n2}} \end{bmatrix}$ is a diagonal matrix, and $\boldsymbol{\sigma_a^2}$ is the part of the genetic variance explained by the prioritized SNPs. $\boldsymbol{I_{n1}}$ and $\boldsymbol{I_{n2}}$ are the identity matrix with dimension n_1 (number of genotyped animals) and n_2 (number of non-genotyped animals**).** Using the joint prior in equation [4], the resulting system of equations at the second stage of the model is given by:

$$[\boldsymbol{H}'\boldsymbol{R}^{-1}\boldsymbol{H} + \boldsymbol{D}^{-1}\boldsymbol{A}^{-1}\sigma_a^{-2}]\begin{bmatrix}\boldsymbol{v}_1\\ \boldsymbol{v}_2\end{bmatrix} = [\boldsymbol{H}'\boldsymbol{R}^{-1}\boldsymbol{y}^*] \qquad \textbf{[5]}$$

Replacing $\boldsymbol{D}^{-1}$ in equation [5] and multiplying it with the inverse of the matrix $\boldsymbol{A}$ leads to:

$$(\boldsymbol{H}'\boldsymbol{R}^{-1}\boldsymbol{H} + \begin{bmatrix}{}^{\boldsymbol{1}}\!/_{\gamma}\,\boldsymbol{A}^{\boldsymbol{11}} & {}^{\boldsymbol{1}}\!/_{\gamma}\,\boldsymbol{A}^{\boldsymbol{12}}\\ \boldsymbol{A}^{\boldsymbol{21}} & \boldsymbol{A}^{\boldsymbol{22}}\end{bmatrix}\sigma_a^{-2})\begin{bmatrix}\boldsymbol{v}_1\\ \boldsymbol{v}_2\end{bmatrix} = [\boldsymbol{H}'\boldsymbol{R}^{-1}\boldsymbol{y}^*] \qquad \textbf{[6]}$$

In general, to make the system of equations in [6] identifiable, it suffices by assigning a value to γ. In our specific case, it is preferable that the effects of the prioritized SNPs **($\boldsymbol{\beta}$)** be estimated using only the genotyped animals. That could be achieved by letting γ tend toward infinity ($\gamma \to \infty$). The genomic contribution of the prioritized SNPs for the non-genotyped animal could be obtained as:

$$\hat{\boldsymbol{v}}_2 = (\boldsymbol{H}_2{}'\boldsymbol{R}_2^{-1}\boldsymbol{H}_2 + \boldsymbol{A}^{\boldsymbol{22}}\sigma_a^{-2})^{-1}[\boldsymbol{H}_2{}'\boldsymbol{R}_2^{-1}\boldsymbol{y}_2^* - \boldsymbol{A}^{\boldsymbol{21}}\hat{\boldsymbol{v}}_1\sigma_a^{-2}] \qquad \textbf{[7]}$$

where $\boldsymbol{R}_2 = \boldsymbol{I}_{n2}\boldsymbol{\sigma}_e^2$

It is worth mentioning that the proposed hybrid approach requires only the inverse of the average additive relationship matrix. Furthermore, the approach could be expending by allowing a more general model for the genomic contributions such that $\boldsymbol{v}_1 = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon}$ a vector of error term following a specified distribution.

The Bayesian implementation of the model presented in equations [1-7] is straightforward, as all conditional distributions were in closed forms. The simulated data was used to evaluate the extended hybrid model for non-genotyped animals and to compare its performance with ssGBLUP using BLUPf90 program (Misztal et al., 2016). The simulated data for the third generation was partitioned into 3 groups: 1) 5K animals with phenotypes and genotypes (training), 2) 8K animals with phenotypes only and 3) 2K animals with genotypes only (validation). Accuracy was calculated as the correlation between true and estimated breeding values.

Results and Discussion

Tables 5.1 and 5.2 present the estimates of the variance components and heritabilities using the different approaches. When the true genetic variance in the base population was equal to 0.4, the highest estimates of genetic variance due to genomic contribution was obtained using ssGBLUP, where all 50K SNPs were included in the calculation of **G**. For the hybrid method, the portion of genetic variance explained by prioritized SNP increased, as expected, with the increase in the number of selected SNPs (Table 5.1). For the hybrid method, the estimates of the polygenetic variance decreased with the number of selected SNPs due to the increase in the fraction of the total genetic variance explained by the prioritized SNPs. The misalignment between the decrease of the polygenic and the increase of the genomic variance with the increase of the number of prioritized SNPs could indicate that the selected SNPs reflect more the phenotypic similarity between individuals rather than the additive relationships. The estimated genetic variance using the hybrid method ranged between 0.36 and 0.38, which is slightly larger than the estimate obtained using ssGBLUP. Across the different models, there was a slight tendency of overestimation of the residual variance, likely due to the underestimation of the genetic variance. This overestimation ranged between 0.01-0.02 for hybrid models to 0.04 for ssGBLUP (Table 5.1). Similar trend in the estimates of variance components and heritabilities was observed for the scenario when the simulated heritability was equal to 0.1 (Table 5.2). The estimated genetic variance using the hybrid method ranged between 0.091 and 0.095, which is similar to the estimate obtained using ssGBLUP (0.091). However, the portion of genetic variance explained by the genomic contribution decreased slightly with the increase in the number of prioritized SNPs. Across two different heritability scenarios, the proposed hybrid model with 1 and 2.5% of prioritized SNPs captured a larger portion of the genetic variance compared to the ssGBLUP model (Tables 5.1 and 5.2). Several genomic studies explained that the proportions of genetic variance detained by SNP markers varied from 35-80% depending on the trait and study population (Tsuruta et al., 2011; Jensen et al., 2012; Haile-Mariam et al., 2013). The number of SNP markers on the panels more likely explain the amount of genetic variance. For instance, Jensen et al. (2012) indicated that using SNP panels less than 5K will explain an expected portion of <85% additive genetic variance. However, increasing the density of SNPs to 44K rapidly increased the proportion of expected genetic variance explained to 96%. In contrast, Yang et al. (2010) demonstrated that around 45% of additive genetic variance in human height explained using 300K SNP panels. The reason denser SNP marker panels in human captured smaller percentage of genetic variance compared to livestock is due to larger effective population size in human compared to livestock population.

Prediction accuracy, defined as correlation between true and estimated breeding values, when heritability was equal to 0.4 and 0.1, is presented in Figures 5.1 and 5.2, respectively. When the heritability was equal to 0.4, accuracy using ssGBLUP was 0.29 compared to 0.46 and 0.48 using the proposed hybrid model with 1 and 2.5% of prioritized SNPs, respectively (Figure 5.1). Similarly, when the heritability was equal to 0.1, accuracy using the proposed method was 15 to 27% higher compared to the ssGBLUP model (Figure 5.2). Based on numerous studies, it has been shown that several factors affect the accuracy of GS. These factors include the genetic architecture of the trait, relatedness between training and validation populations, marker density, size of training population, LD between SNPs and QTL and heritability of the trait (Hayes et al., 2009; Zhong et al., 2009; Daetwyler et al., 2012). In this case, the clear superiority of the hybrid method compared to ssGBLUP could be in part due to the relatively large effect of the simulated QTL that are easily tagged using F_{ST} scores. Toghiani et al. (2017) indicated that the functional genomic similarity based on SNP markers identified by F_{ST} scores reflects the similarity at the selected SNPs. As these SNP markers are prioritized based on the intensity of selection pressure they receive, animals with similar genetic merit are expected to have higher functional genomic similarity. Thus, although ssGBLUP explained similar portion of the genetic variance compared to the hybrid model, it is likely that the latter resulted in a much higher genetic similarity. In addition, ssGBLUP assumes that the total genetic variance can be expressed by the SNP markers. If that is not the case, the accuracy of prediction will be smaller than the expected (Kemper and Goddard, 2012). In this study, around half of the genetic variance was explained by 80% of the QTL with effects ranging between 0.2 and 0.5% of the genetic variance. Therefore, in the presence of small QTL effect, identifying relevant linked SNPs based on the magnitude of their effects becomes more challenging. Prioritized SNP markers using the hybrid method are based on change in allele frequency, due to selection pressure, but not directly of the effect of the SNP. This difference in the prioritization for the hybrid model is extremely important in the presence of small QTL effect and could explain, in part, the higher accuracy against ssGBLUP (Fig 5.1 and 5.2).

Conclusions

The substantial increase in the density of marker panels and the number of genotyped animals is creating significant challenges implementing genomic selection. Accommodating non-genotyped animals is essential in order to eliminate potential selection bias. All these factors create challenges of different magnitude for the different approaches used to implement genomic selection. High-density marker panels or sequence data will greatly increase the computational costs of LR models and limit the utility of Bayesian variable selection methods such as BayesB and BayesC due to lack of statistical power. Accommodating non-genotyped animals will become practically impossible. For linear mixed model based approach, the increase in the number of variants does not present a major challenge, at least from a computational perspective. However, the increase in the number of genotyped animals will make the direct inversion of genomic relationship matrix impossible. Approximating the inverse is a data driven process and, thus, its optimality or even adequacy is not guaranteed. A potential practical solution could be through the substantial reduction in the number of markers in the association model, eliminating the need to impute missing genotypes for non-typed animals, and the avoidance of the construction and inversion of the genomic relationship matrix. The hybrid method presented in this study seems to have successfully tackled all these challenges. At a fraction of the computational costs, the hybrid method resulted in higher accuracy compared to ssGBLUP. The results of this study are based on simulated data with predefined distributions for the QTL effects and need to be validated in more diverse simulation scenarios. However, the results of these studies and those of Chang et al. (2018) seem to indicate the competitiveness of the hybrid method.

References

Chang, L.-Y., S. Toghiani, A. Ling, S. E. Aggrey, and R. Rekaya. 2018. High density marker panels, SNPs prioritizing and accuracy of genomic selection. BMC genetics 19: 4.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2012. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193: 347-365.

Decker, J. E. 2015. Agricultural Genomics: Commercial Applications Bring Increased Basic Research Power. PLOS Genetics 11: e1005621.

Fang, L. et al. 2017. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. Genet Sel Evol 49.

Fernando, R., J. Dekkers, and D. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution 46: 50.

García-Ruiz, A. et al. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proceedings of the National Academy of Sciences 113: E3995-E4004.

Haile-Mariam, M., G. J. Nieuwhof, K. T. Beard, K. V. Konstatinov, and B. J. Hayes. 2013. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J Anim Breed Genet 130: 20-31.

Harris, B., and D. Johnson. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. Interbull Bulletin: 40.

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

Jensen, J., G. Su, and P. Madsen. 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC genetics 13: 44.

Kemper, K. E., and M. E. Goddard. 2012. Understanding and predicting complex traits: knowledge from cattle. Hum Mol Genet 21: R45-51.

MacLeod, I. M. et al. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17: 1-21.

Matukumalli, L. et al. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4: e5350.

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

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. Journal of Dairy Science 97: 3943-3952.

Misztal, I. et al. 2016. Manual for BLUPF90 family of programs. Retrieved from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all5.pdf.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences 70: 3321-3323.

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680-681.

Toghiani, S., L.-Y. Chang, A. Ling, S. E. Aggrey, and R. Rekaya. 2017. Genomic differentiation as a tool for single nucleotide polymorphism prioritization for Genome wide association and phenotype prediction in livestock. Livestock Science 205: 24-30.

Tsuruta, S., I. Misztal, I. Aguilar, and T. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J Dairy Sci 94.

VanRaden, P. M. et al. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science 96: 668-678.

Wright, S. 1951. The genetical structure of populations. Annals of eugenics 15: 323-354.

Yang, J. et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genet 42: 565 - U131.

Zhong, S., J. Dekkers, R. Fernando, and J. Jannink. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182: 355 - 364.

Table 5.1. Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with $h^2 = 0.4$

	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_{poly}^2}$	$\boldsymbol{\sigma_e^2}$	h^2
hybrid.fst(0.99)	0.22	0.16	0.62	0.38
hybrid.fst(0.975)	0.26	0.10	0.61	0.37
ssGBLUP	0.34	--	0.64	0.34

$\boldsymbol{\sigma_g^2}$= genomic variance;$\boldsymbol{\sigma_{poly}^2}$= polygenic variance; $\boldsymbol{\sigma_e^2}$= residual variance; h^2= heritability.

Table 5.2. Estimates of variance components and heritability (averages over 5 replicates) using different methods for a simulated trait with $h^2 = 0.1$

	$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma_{poly}^2}$	$\boldsymbol{\sigma_e^2}$	h²
hybrid.fst(0.99)	0.062	0.033	0.882	0.097
hybrid.fst(0.975)	0.060	0.031	0.876	0.094
ssGBLUP	0.091	--	0.897	0.092

$\boldsymbol{\sigma_g^2}$= genomic variance;$\boldsymbol{\sigma_{poly}^2}$= polygenic variance; $\boldsymbol{\sigma_e^2}$= residual variance; h^2= heritability.

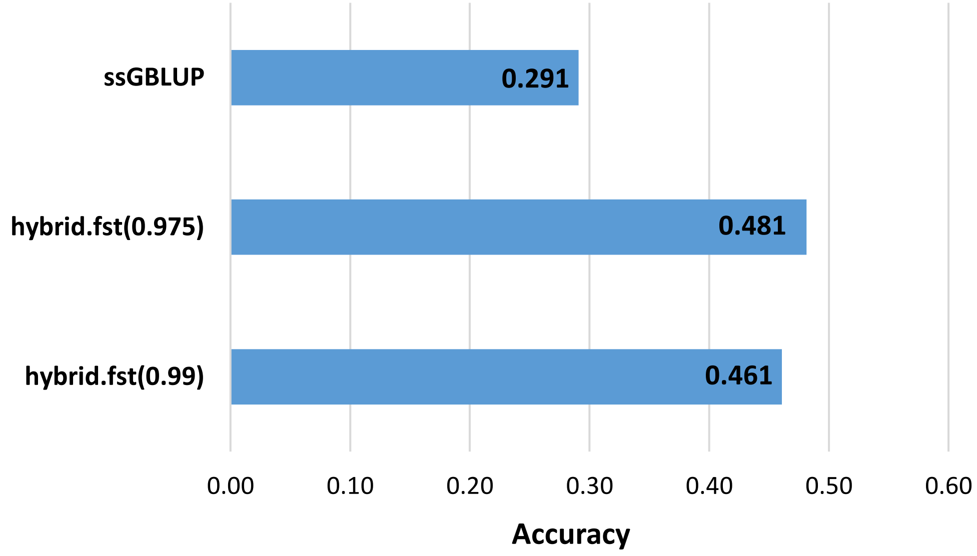


Figure 5.1. Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with $h^2 = 0.4$

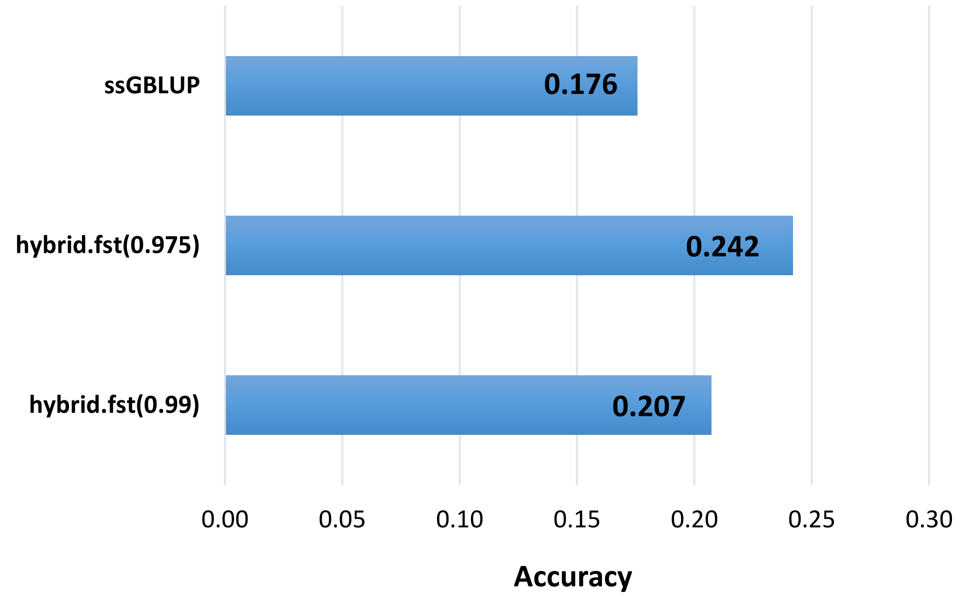


Figure 5.2. Accuracy of estimated breeding values (average over 5 replicates) using different methods for a simulated trait with $h^2 = 0.1$

CHAPTER 6

CONCLUSIONS

Using either high-density (HD) marker panels or next generation sequence (NGS) data resulted in no significant improvement in the accuracy of genomic selection (GS). This lack of improvement of the accuracy of GS is not due to the uselessness of HD and NGS data rather than the limitation of currently used methods for implementation of GS. Including all variants of HD and NGS simultaneously in the association model will lead to a dramatic increase in the number of unknown parameters and a substantial reduction in statistical power. Increase in the number of genotyped animals has substantially complicated the inversion of the genomic relationship matrix (**G**). Thus, reducing the number of variants to include in the association model and the elimination of the need to invert the **G** are needed to harness the full benefits of HD and NGS data.

SNP prioritization using fixation index (F_{ST}) is an attractive tool to identify marker under selection pressure based on the change of their allele frequencies. In this study, using the 97.5% quantile of the F_{ST} distribution to prioritize SNPs captured between 40 to 80% of the significant QTL under different simulation scenarios. Furthermore, the genomic similarity calculated based on prioritized SNP markers proved to be a useful tool for decision-making, phenotype prediction, and genetic selection.

Increase in the number of genotyped animals has made the direct inversion of the genomic relationship matrix (**G**) impossible. Although some data driven approximations of the inverse of **G** have been proposed, their optimality is not guaranteed. Furthermore, constructing the matrix **G** using all the available markers is on top of being computationally costly it will not improve accuracy and it could even lead to lower performance. To overcome these limitations, a hybrid approach that uses only a limited number of prioritized variants and a polygenic component in the association model was proposed. Because the prioritized markers will unlikely account for all the genetic variance, a polygenic component was added to model. The effectiveness of the hybrid model was assessed by comparing its performance to BayesB, BayesC and GBLUP using simulated and real data sets for a trait with heritability equal to 0.1 or 0.4. In both simulated and real datasets, 1 and 2.5% of total SNPs were prioritized based on the quantile distribution of the F_{ST} scores. When the heritability was equal to 0.4, the proposed hybrid model increased accuracy by 10.1 to 11.5%, 9.2 to 10.1%, and 26.1 to 29% compared to BayesB, BayesC and GBLUP models, respectively. When the heritability was equal to 0.1 and only 1% of the markers were prioritized, GBLUP was superior to BayesB, BayesC, and the hybrid method. However, when 2.5 of markers were prioritized, the hybrid model outperformed all the other methods with a superiority ranging from 3.3 to 7.7%. The hybrid model was extended to accommodate non-genotyped animals and its performance was assessed compared to ssGBLUP. When the heritability was equal to 0.4, the proposed hybrid model resulted in an accuracy of 0.46-0.48 compared to 0.29 for ssGBLUP model. When the heritability was equal to 0.1, the hybrid model was outperformed ssGBLUP with a superiority ranging from 15 to 27%. The hybrid method required only a small fraction of the computational needed to implement ssGBLUP.

Based on these results, the hybrid method seems to have successfully tackling some of the challenges facing genomic selection. The results of this study are based on simulated data with predefined distributions for the QTL effects and need to be validated in more diverse simulation scenarios. However, the results of these studies seem to support the competitiveness of the hybrid method.