PRIORITIZATION OF RELEVANT SINGLE NUCLEOTIDE POLYMORPHISMS IN HIGH DENSITY MARKER PANELS AND ITS EFFECTS ON GENOMIC SELECTION

by

LING-YUN CHANG

(Under the Direction of Romdhane Rekaya)

ABSTRACT

The availability of large-scale genotyping platforms provided an unprecedented resource to dissect associations between traits of interest and genomic variation, and to enhance the estimation of breeding through genomic selection (GS) for animal and plant applications. The dramatic increase in the number of variants was expected to significantly increase the accuracy of GS. Unfortunately, that was not the case due to several factors including the huge increase in the number of variants in the association model, increase in co-linearity, and the reduction in statistical power. Thus, accuracy of GS using either multiple regression (RM) or mixed linear models (VC) approaches did not result in any significant increase in the accuracy of GS after a certain density of SNP markers in the genotyping panels was reached. It was clear that SNP prioritization is needed in order to harness the full potential of high density marker panels. Wright's fixation index (F_{ST}), a measure the genetic differentiation among sub-populations, was proposed to prioritize SNPs in high-density marker panels and to track relevant quantitative trait loci (QTL). Using the phenotypic distribution, three sub-populations

were created based on the 5 and 95% quantiles and were used to identify markers under selection pressure. Genomic data consisted of 200K and 400K SNP markers distributed on 10 chromosomes to mimic 770K and 1.2 million SNPs markers in the bovine genome. In the first study, the performance of F_{ST} prioritized SNPs was compared to RM based methods. Only 0.5 to 1% of SNPs were needed to achieve the same accuracy as used all markers in the panel. Furthermore, using F_{ST} method showed slight superiority compared to BayesB and BayesC. In the second study, the effects of F_{ST} prioritized SNPs on the computation of the genomic relationship matrix (**G**), genomic similarity and accuracy of GS were assessed. The results showed that a balance between genomic similarity and percentage of genetic variance explained is needed to optimize the accuracy of GS. In the third paper, F_{ST} scores were used to derive weights for computing **G**. The results clearly showed that accuracy could be improved under different weighting scenarios.

INDEX WORDS: Genomic selection, High density panel, SNP prioritization, Accuracy

PRIORITIZATION OF RELEVANT SINGLE NUCLEOTIDE POLYMORPHISMS IN HIGH DENSITY MARKER PANELS AND ITS EFFECTS ON GENOMIC SELECTION

by

LING-YUN CHANG

B.S., Taipei Medical University, Taiwan, 2004

M.S., The University of Georgia, 2007

M.S., The University of Georgia, 2010

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

PRIORITIZATION OF RELEVANT SINGLE NUCLEOTIDE POLYMORPHISMS IN HIGH DENSITY MARKER PANELS AND ITS EFFECTS ON GENOMIC SELECTION

by

LING-YUN CHANG

Major Professor:	Romdhane Rekaya
Committee:	Ignacy Misztal
	Daniela Lourenco
	Samuel Aggrey

Electronic Version Approved:

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

ACKNOWLEDGEMENTS

I would like to thank my major professor Dr. Rekaya for taking me on as his student. Thanks for his guidance, patience and support throughout my Master and PhD degrees. He always inspires me to think and provides his valuable insights into my work. Without his support, I wouldn't be where I am today. Further, I would like to express my gratitude to the committee members, Dr. Misztal, Dr. Lourenco and Dr. Aggrey for always being willing to offer their advice, both inside and outside of the classroom. It is impossible to accomplish this without their encouragements. Also, I thank the faculties, staffs and graduate students in Animal and Dairy Science Department for being such a supportive and engaging group that allowed me to grow both personally and professionally. It is grateful to be part of their life. I hope our paths will cross again in the future.

Finally, with my whole heart and soul, I would like to appreciate my family in Taiwan. Although the distance is long, their love is persistent and strong, and their support has meant so much to me. I couldn't imagine this journey without any one of them. Thanks to my mother and father for their unconditional love encourages me all the time. Thanks to my little brother and his wife for listening and giving me different perspective. Thanks to my nephew and niece for always making me laugh. Also, thanks to my grandmother and grandfather. Although they are no longer with us, I miss them all the time and wish they can share this moment with me.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

In the past decade, there has been a rapid advance in high-throughput technologies leading to the collection of a massive amount of genomic data. Both time and cost required for genotyping were dramatically reduced. Across the livestock industry thousands of animals have been already genotyped with single nucleotide polymorphisms (SNPs) panels with varying density. This massive resource provides an unprecedented opportunity to identify associations between complex traits and genetic variation. These associations can be used to identify functional or causative variants and reveal the genetic basis of complex traits. Additionally, in livestock and plant species these associations could be used to carryout genomic selection through the estimation of genomic breeding values (GEBVs)

Genomic selection has quickly become the standard tool for genetic evaluation for several livestock, plant and poultry species due its superiority to classical evaluation procedure based on pedigree information. The use of genomic information has increased the accuracy of young animals' breeding values, reduced the generation interval, and increased the rate of genetic response. Currently, genomic selection is carried out using two alternative methods: multiple regressions (RM) or variance component (VC) approaches. Genomic selection was first implemented using RM approach where SNP effects are directly estimated. Depending on the genotyped population and available data, marker effects are estimated using either the actual phenotypes or pseudo-phenotypes

such as estimated breeding values or daughter yield deviation as dependent variable in the regression model. Although simple to implement, these models suffer from an acute small *n* larger *p* problem resulting in high co-linearity, and limited statistical power. Their implementation requires the use shrinkage methods for inference. Contrarily to the RM approach, the VC approach estimates the joint effects of all SNPs (GEBV) through the use of the genomic relationships between individuals computed based on the genotyped markers. The SNP effect could be obtained through a back transformation of the GEBVs.

Although several methods have been presented for the implementation of these two approaches, results are often similar, and differences are largely due to prior assumptions about the SNP effects. Independently of the method used for implementation, accuracy of genomic selection depends largely on the density of the marker panel, LD structure between marker and QTL, training population and effective size.

For low to moderate density panels, increase in the number of SNP markers often resulted in improvement in the accuracy for GEBVs due to a higher LD between markers and QTLs for RM methods and a better estimation of the realized relationships for the VC approach. However, after a certain threshold on the number of genotyped SNPs is reached no improvement in accuracy was observed and in some cases, increase in marker density resulted in decrease of performance of genomic selection. This was the case due to the substantial increase in the number of unknown in the association model for the RM approach and the lack of any meaningful change in the realized relationship matrix for the VC method.

This lack of improvement of accuracy (and even loss of accuracy) due to increase in marker density could limit the utility of genome wide sequence (GWS) data in genomic selection. We believe that this lack of improvement in accuracy is not due to the lack of useful information in the GWS data, rather to the limitation of current methods in using this information. Because of disproportional increase in the number of markers and genotyped individuals, higher density SNP panels or GWS data will further deteriorate the already limited statistical power One solution could be to increase the size of the reference population. However, such option is prohibitively costly and time consuming which will limit is practical utility. Consequently, methodological and statistical approaches are crucially needed to address this problem. One idea that has been postulated was to reduce the dimensionality of parameter space through the prioritization of relevant variants.

In this dissertation, an innovative method to prioritize SNP markers based on the utilization of Fisher fixation index to identify genomic region under selection pressure was proposed and evaluated using simulation data. Therefore, the goals of the current studies are to: 1) estimate the accuracy of genomic selection using pre-selected SNPs based of their F_{ST} scores, 2) compare the accuracy between current SNP prioritizing methods and F_{ST} approach, 3) evaluate the impact of SNP prioritization using F_{ST} scores on the genetic similarity between individuals and on the accuracy of genomic selection, and 4) evaluate the impact of the prioritization of the realized relationship matrix (G).

CHAPTER 2

REVIEW OF LITERATURE

Genomic Information

Traditionally, genetic evaluation of livestock is implemented using phenotypic and pedigree information where breeding values for traits of economic interest are estimated. These estimated breeding values (EBVs) are then used for selection and they were a quite successful tool for the significant genetic improvement of several domestic species (Hill 2008). Unlike simple traits affected by single or few genes, most economically important traits in human, plants and animals are of complex genetic nature due to their control by many genes and the effect of the environment. The portion of genetic variance explained by each gene is usually small which complicates enormously the accurate estimation of its associate effects. In fact, only few genes are well known to be responsible for large variation in these traits (Goddard, 2009). Mapping genes associated with these complex traits can be beneficial to understand their genetic architecture.

With the availability of high-density panel of single-nucleotide polymorphisms (SNPs) and cost-effective genotyping platforms, generation and use of genomic information is becoming a routine in the genetic evaluation of livestock species. In fact, genomic selection is quickly replacing the traditional genetic evaluation based on phenotypes and pedigree information. Meuwissen et al. (2001) proposed the use genomic

selection, a form of marker-assisted selection (MAS), estimate the effects of genomic regions and generate genomic estimated breeding values (GEBVs). These dense genetic markers located across the entire genome are in linkage disequilibrium (LD) with causative gene or quantitative trait loci (QTL). They allow for the estimation of the realized additive relationships between animals and a better modeling of the Mendelian sampling. Therefore, the GEBVs are likely to more accurate compare with classical estimated breeding values. Generation interval and the cost of progeny testing also can be reduced, especially when the phenotypes of young animals are not available (Schaeffer, 2006; Konig et al., 2009). These advantages of genomic selection lead to an increase in genetic response.

In human applications, several SNPs were identified to be in association with a whole range of diseases using genome wide association studies. Currently, several millions genomic variants, including SNPs, have been identified by the 1000 genome project (The 1000 Genomes Project Consortium, 2015). The database has more than 80 million SNPs, including about 8 million common variants (allele frequency >5%). This resource has been used in genome wide association studies (Frazer et al., 2007; Weedon et al. 2008; Cantor et al., 2010; Zhang et al., 2015; Rautiainen et al., 2016; de Vries et al., 2017; Scott et al., 2017; Gorski et al., 2017). Although several relevant mutations were identified and were useful in helping understand the genetic basis of some complex diseases, only small portion of the genetic variation of major human illness has been explained by genomic information. With the continuous advances in next generation sequencing and other high throughput technology, a better understanding of the genetic basis of human diseases is likely.

In animal and plant breeding, SNP information is being used for genetic improvement. In livestock and poultry species low to medium density panels of 50 K to 60 K SNP markers developed and Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA) are being used in genomic evaluation in North America (Van Tassell et al., 2008; Matukumalli et al., 2009 Wiggans et al., 2009).

Early simulation studies showed substantial increase in accuracy compared to pedigree based genetic evaluations (Meuwissen et al., 2001; VanRaden, 2008; Habier et al., 2007). This superiority of genomic selection, although smaller, was maintained using real data (Legarra et al., 2009; Gonzalez-Recio et al., 2009; de los Campos et al., 2009; Lorenzana and Bernardo, 2009). Due to this superiority several breed associations and livestock companies started using the genomic information into their selection program.

Genome Wide Association Studies

Genome Wide Association Studies (GWAS) is an approach to identify common genetic variants that influence a phenotype and can be applied to any complex trait (Hannum et al., 2009; Hayes et al., 2010; Yang et al., 2010; Stephan et al., 2011; Cho et al., 2012; Fanous et al., 2012; Zhang et al., 2012; Li et al., 2013). The International HapMap Project provided comprehensive resource to test the statistical association between phenotypic variation and hundreds of thousands to millions of genetic common variants. In 2005, phase one was completed and all data were open to researchers all over the world.

GWAS in human are primarily focused on complex disease with a final goal of a potential use in clinical applications. Although GWAS are able to scan for potential

causative loci on a genome-wide scale, there are still several major challenges that need to be overcome including the high dimensionality of the parameter space, lack of statistical power, and the high false positive rate. Population admixture, population stratification, and the genetic relationships may confound associations at causative loci. Actually, only a small fraction of genetic variation was explained and the genetic gain is far less than expected (Vineis et al., 2010; Eichler et al., 2010; Zaitlen et al., 2012). The famous example in human height showed that the missing heritability is due to the lack of ability to determine variants with small or moderate effects (Maher, 2008). Large proportion of the total genetic variation can only be explained by a joint estimate of all SNPs. The situation is getting worse when the density of marker panel increased.

Another obstacle of human application is sample size. Some medical symptoms of diseases are difficult to confirm, and data collection is limited by patient privacy or population structure (Im et al., 2012; Stranger et al., 2011; Yoon et al., 2012). In the early stage of GWAS, costs were high. With technological advances, the price of genotyping has drop significantly and sample size increased to thousands of observations. Even with those large sample sizes, many results do not appear to replicate in subsequent studies (Lin et al., 2010; Zeggini et al., 2009). GWAS publications now involve multiple data sets to both reduce false positives and increase statistical power to find true positives (Thompson et al., 2011; Willer et al., 2010). Meta-analysis provides optimum power to find homogeneous effects and heterogeneity across cohorts (Higgins et al., 2003; Ioannidis 2007; Thompson 1994). Many of the remarkable results have come from large-scale mega-consortia and meta-analyses that combine data from dozens of studies and thousands of subjects. Therefore, Meta-analyses have become a routine part of GWAS.

In livestock, the priority was to discover genes affecting important production traits. Modern livestock production units are usually under highly controlled management and data collection is more standardized and completed. Because of wide use of artificial insemination (AI) and successful genetic selection, the effective population size (Ne) in animal population is often much smaller than in human leading to well preserved strong LD blocks. Thus, to achieve similar power of association and find significant signals, small number of SNP markers is needed (De Roos et al. 2008; Druet and Georges, 2010). In fact, the 50k to 60k SNP marker panels are still the main genotyping platform for several livestock species.

Genomic Selection

For decades, genetic evaluation of livestock species relayed on the infinitesimal model using phenotypic and pedigree information. Average relationships between individuals allowed the calculation of the expected correlation between the Breeding values of relatives. The following mixed linear model is often used:

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

where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are vectors of fixed effects and breeding values and $\boldsymbol{e}$ is a vector of residuals. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices with the appropriate dimensions. The EBVs are obtained by solving the following mixed model equations (Henderson, 1984):

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{A^{-1}\lambda} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix}$$

where $\boldsymbol{\lambda} = \sigma_e^2 / \sigma_a^2$ and $\boldsymbol{A^{-1}}$ is the inverse of the pedigree-based relationship matrix. Pedigree-based BLUP uses the expected relationships (i.e. covariances) between individuals, using the probability that for a given locus, the alleles are identical by descent (IBD) or inherited from the same common ancestor (Wright, 1951). Two full-siblings are expected to share half their alleles because each of them inherits half the genes from each parent. Which half of each parents' genome that an offspring inherits depends on recombination, the independence of chromosomes, and the random selection during gamete formation. This phenomenon, known as the Mendelian sampling, is the main reason why a pair of full sibs could share more or less than half of their alleles.

In early 1980, geneticists developed techniques that allowed the investigation of the DNA providing an opportunity to supplement the phenotypic and pedigree information in the genetic improvement of livestock. These early developments paved the way to QTL mapping and marker assisted selection (MAS). The promise of MAS was to detect and map genes underlying the traits of interest and select parents that had the desired marker profile. However, due to the complex nature of majority of traits where thousands of genes with small effects each are involved, MAS was not successful in livestock applications. Even when important genome signal was detected the target QTLs were not always observed in replicated studies or in other populations (Meuwissen et al., 2016; Andersson, 2001). Currently, there are only a few genes that contribute more than 1% of the genetic variation of important traits. Meuwissen et al. (2001) proposed a method that by passes the mapping of QTL and tries to directly estimate the breeding values. The method relies on panels of dense SNP markers in LD with the QTL. Using simulation data, they showed that accuracy can be doubled using genomic selection (GS)

compared to the classical pedigree-based evaluation. The study concluded using SNP information will increase genetic gain, reduce the generation interval, and animals could be selected early without knowing their performance or progeny testing.

Because the large number of SNPs compared to the number of genotyped animals, Meuwissen et al. (2001) proposed a two-step process. Using a training (reference) population, the effects of SNPs are simultaneously estimated. These estimated are evaluated using a validation (testing) population in a second step. The following simple regression model is ised for the estimation of the SNP effects:

$$y_i = \mu + X_{1i}\alpha_1 + X_{2i}\alpha_2 + \cdots + X_{pi}\alpha_p + e_i \quad [1]$$

where y_i is phenotype or pseudo-phenotype for animal *i*; μ is the overall mean; X_{ij} is the genotype of animal *i* for marker *j*; *p* is the number of SNPs; and e_i is the residual term. In real data, the number of animals is usually much smaller than number of markers leading to the well- known small *n* large *p* problem. Thus, the SNP effects in [1] cannot be treated as fixed. Several methods have been proposed to deal with this issue including single marker analyses and Bayesian approaches. The later assume prior distribution for the SNP effects.

Meuwissen et al., 2001 proposed two different prior distributions for the SNP effects leading the well-known BayesA and BayesB methods. BayesA assumed that the SNP effects (**α**) have identical and independent t-distributions with scale parameter S_α^2 and v_α degrees of freedom and the variance of each marker followed an inverted chi-square distribution. BayesB assumed the same prior for marker variance, but a large proportion (π) of markers had zero effect such that:

$$\alpha_i \sim \begin{cases} t(v_\alpha, S_\alpha^2) & with\ probability\ 1-\pi \\ 0 & with\ probability\ \pi \end{cases}$$

where parameter π is defined as proportion of SNPs that are assumed to have an effect of zero. In the following ten years, different authors proposed alternative ways to estimate marker effects with different prior assumptions, such as BayesC (Habier et al., 2011), Bayesian Lasso (de los Campos et al., 2009; Hastie et al., 2009), BayesR (Erbe et al., 2012), and GBLUP (Habier et al., 2007; VanRaden, 2008; Stranden and Garrick, 2009; Shen et al., 2013). Bayesian Methods all assume that SNP effects are independent. BayesC is similar to BayesB but instead of assuming marker effects have identical and independent t-distributions, the assumption is that they have identical and independent Normal distributions with a variance that is common across all markers:

$$\alpha_i \sim \begin{cases} N(0, \sigma_\alpha^2) & with\ probability\ 1-\pi \\ 0 & with\ probability\ \pi \end{cases}$$

When π = 0, this model is equivalent to the GBLUP model. An extension method of BayesC proposed by Habier et al. (2011) is BayesCπ, where π is treated as unknown parameter which follows a uniform distribution between 0 and 1

From a biological perspective, it makes more sense to use nonlinear (Bayesian) methods with different prior distributions for the SNP effects than to assume all SNPs have small effects. Bayesian methods seem to perform better than GBLUP in many simulation studies, however no genomic prediction model have shown consistent superiority across data sets (Meuwissen and Goddard, 2010; Erbe et al., 2012; Daetwyler et al., 2013). The most appropriate model for genomic prediction seems to depend on the trait heritability, genetic architecture, genetic relationships between individuals, and the number of individuals with completed genotypes and phenotypes (Daetwyler et al., 2010; Hayes et al., 2010; Habier et al., 2007; de los Campos et al., 2013; Daetwyler et al., 2013). For example, BayesA is more sensitive to the priors for genetic and phenotypic

variances because the prior information dominates the inference (Habier et al., 2011). If the prior is not informative, prediction model might become unidentifiable and force all SNPs to have the same of variance of SNP effects. BayesA could be a proper approach when we have a complex model with all SNPs having a small effect. On the contract, BayesB showed higher prediction accuracy than other methods when the trait of interest is controlled by large QTL (Meuwissen et al., 2001). Therefore, it is common to investigate the performance of different models for implementation of genomic selection.

Unlike Bayesian method, GBLUP assumed a normal prior distribution with constant variance for marker effects and it is equivalent to traditional BLUP except that the numerator relationship matrix (**A**) is substituted by a realized relationship matrix. The genomic relationship matrix (**G**) can be calculated and scaled following VanRaden (2008):

$$\mathbf{G} = \frac{\boldsymbol{ZZ'}}{k} = \frac{\boldsymbol{ZZ'}}{2\sum p_i(1-p_i)}$$

where $\boldsymbol{Z}$ is the matrix of SNP markers marker genotypes (-1 and 1 for homozygous and 0 for heterozygous), k is a scaling parameter, p_i is allele frequency for i-th SNP, and SNP effects were assumed to be independent (Gianola et al., 2009). Mixed model equations that replace $\boldsymbol{A^{-1}}$ with the inverse of the G matrix ($\boldsymbol{G^{-1}}$) are referred to as genomic BLUP (GBLUP) (VanRaden, 2008) and have been shown to improve prediction accuracy over pedigree-BLUP models (Hayes et al., 2009; Wolc et al., 2011b; Daetwyler et al., 2012). Scaling of **G** matrix is important for the precision of genomic prediction. The matrix **G** can be scaled based on different allele frequencies which are estimated from current population, base population or constant allele frequency. G matrix can also be calculated

with additional diagonal matrix with weights, which makes unequal variance for SNP markers (Leutenegger et al., 2003, Amin et al., 2007; Zhang et al., 2010).

Genomic evaluation - Single step approach

Rather than estimating the SNP effects directly and then calculate the GEBVs as a linear combination of those effects, the single step approach models directly the GEBV within the frameworks of mixed linear methodology. The approach is a straightforward extension of the classical mixed model used in genetic evaluation of livestock species with the exception that the observed (realized) relationships matrix (**G**) is used instead of the its expectation (Misztal et al., 2009). In presence of non-genotyped animals, the genetic covraince between animals is obtained by blending the **A** and **G** in a single matrix often noted as **H** Legarra et al. (2009) derived a joint relationship matrix based on pedigree and genomic relationships. The ungenotyped animal can benefit from the genotyped animal through the pedigree relationships. The blended relationship matrix (**H**) can be presented as:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} (\mathbf{G} - \mathbf{A}_{22}) [\mathbf{I} \quad \mathbf{I}] \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & 0 \\ 0 & \mathbf{I} \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{G} \end{bmatrix}$$

where subscripts 1 and 2 represent ungenotyped and genotyped animals, respectively, **A** is the pedigree relationship matrix, and **G** is a genomic relationship matrix.

The main idea of single-step is based on a combined relationship matrix that integrated **G** and **A**. The inverse of **H** replaced inverse of **A** in the mixed model equation and the single-step genomic BLUP can be re-written as:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \mathbf{H^{-1}}\boldsymbol{\lambda} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are vectors of fixed effects and breeding values, and $\boldsymbol{y}$ is a vector of phenotypes. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices with the appropriate dimensions. The matrix **H** is not easy to be derived, but $\mathbf{H}^{-1}$ is very simple (Christensen and Lund, 2010; Aguilar et al., 2010). The inverse of **H** can be obtained as:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

For multiple step method, the pseudo phenotype (e.g. daughter yield deviation, de-regressed progeny-test EBV) is required if the genotyped animals are young and the phenotypes cannot be collected. Furthermore, using pseudo phenotype might cause several problems, such as low reliability for young animals, selection bias, and double counting. These problems can be solved by single-step GBLUP approach. It can be easily applied for multiple traits evaluations, reduced the number of estimated parameters and avoided double counting of phenotypic and pedigree information. Currently, ssGBLUP has been implemented in different species and applied by some companies/institution in US (Chen et al., 2011; Christensen and Lund, 2009; Forni et al., 2011; Misztal et al., 2009; Misztal et al., 2002; Vitezica et al., 2011).

.

Evaluating Performance of Genomic Prediction

The performance of genomic prediction models can be evaluated by splitting the data into training and validation sets. The training set is used to estimate the SNP effects and validation set is used to evaluate model performance and prediction accuracy (Daetwyler et al., 2013). Prediction accuracy is the standard measurement used to assess

model performance (Hayes et al., 2009; Wolc et al., 2012) because of its direct relationship to genetic gain. Prediction accuracy is commonly calculated as:

$$cor(y, \hat{u}) = \frac{cov(y, \hat{u})}{\sqrt{var(y)var(\hat{u})}}$$

where y is the phenotype and $\hat{u}$ is the estimated breeding values. This correlation can also be divided by the square root of heritability to approximate the correlation between true breeding values (u) and estimated breeding values ($\hat{u}$):

$$\frac{cor(y, \hat{u})}{\sqrt{h^2}} = \frac{cov(y, \hat{u})}{\sqrt{var(y)var(\hat{u})}} \sqrt{\frac{var(y)}{var(\hat{u})}} = cor(u, \hat{u})$$

where y is phenotype, u is the true breeding value, and $\hat{u}$ is the EBV. It is necessary to approximate this correlation because true breeding values are not directly observed. However, this approximation can result in correlations above 1 if the estimate of heritability is different from the heritability in the validation set.

The goal of evaluating the performance of a genomic prediction model is to estimate how accurate the prediction would be in another set of individuals that do not have phenotypes. The choice of training and validation data sets can have a large impact on estimates of prediction accuracy and bias. An increase in prediction accuracy can be achieved through the increase in the size of the training data set. However, given a finite data set to test the model, a small validation set might lead to unreliable accuracy estimates (Erbe, 2012). Prediction accuracy relies on these relationships between individuals in the training and validation set. Accuracy will be higher if more relatives are included in both the training and validation set (Habier et al., 2007). It may be advantageous to combine closely-related breeds into one training set, particularly when

there are few individuals in a single-breed training set (Brondum et al., 2011; Pryce et al., 2010; Kachman et al., 2013). However, it is worth to evaluate prediction accuracy separately for each breed since prediction accuracy may be different among breeds.

Two commonly used methods for evaluating the reliability of an estimate of accuracy are cross-validation and bootstrapping. A common approach to cross-validation was described by Saatchi et al. (2011) and is based on k-means clustering. Genotyped individuals are separated into k groups that reduce the relatedness between groups and maximize it within groups. The training set is specified as k-1 groups, the validation set is the remaining group. The process is repeated k times until the mean and standard deviation of an accuracy estimate is obtained for all groups (Daetwyler et al., 2013). Cross-validation is able to obtain a standard error estimates for the accuracy, however such estimates can be sensitive to outlier groups and the relationship between individuals in training and validation sets which sometimes is not reflective of the relationships in the full data. The accuracy obtained using cross-validation depends on the number of groups (k) as indicated by Erbe, 2013.

Cross-validation, such as like jackknife, utilized all r data to assess model selection and to measure the predictive accuracy. The bootstrap is another flexible and powerful statistical tool that resamples data randomly and quantifies the uncertainty in the model. The variability of the estimate can be evaluated without generating additional samples. Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement. Therefore, each of these "bootstrap data sets" is created by sampling with replacement and has the same size as the original dataset (Pryce et al.,

2014). In general, bootstrap primarily was used to obtain standard errors of estimates. Prediction accuracy is estimated with the new set and this process has to be repeated several times (5,000 to 10,000 times) to get sufficiently large sample. The mean and mean and the standard error of the prediction accuracy can be easily obtained using the moments method.

Next-generation sequencing

NGS has dramatically changed the speed, coverage and costs of sequencing whole genomes. In fact, several sequencing efforts, including the 1000 Bull Genomes Project, are underway, while several thousand human and animals have already been fully sequenced (The 1000 Genomes Project Consortium, 2015). These projects are crucial for characterizing the source of genetic variation. In fact, 84 and 31.8 million common and rare variants have been already identified in human and dairy cattle, respectively (The 1000 Genomes Project Consortium, 2015; Hayes et al. 2014). Although the majority of these variants are rare (MAF <1%), over 8 million common SNPs (MAF > 5%) have already been identified in humans. Thus, it is already a reality that GWAS and ultimately genomic selection will be implemented using several millions of directly- or indirectly-imputed sequence variant genotypes. In fact, GWAS using 17 and 19 million SNPs were carried out in human and dairy cattle (Daetwyler et al., 2014) applications, respectively. Although theoretically there are no doubts about the potential usefulness of the sequence data in GWAS and GS, major challenges are limiting the harnessing of these benefits.

When all variants are considered (i.e, BayesA), the highly informative prior will lead to excessive shrinkage that together with the high LD precludes the identification of

causative mutations or even of significant tag variants. As the effect of a QTL (often small for complex traits) is distributed in a non-trivial manner between all markers that are in LD with the causal mutation, there is little statistical power to accurately estimate its effect. Variance components based approaches in their current form are not likely to benefit from the use of information provided by NGS, and in extreme cases a decrease in performance could occur. The superiority of GBLUP compared to classical BLUP is due to the use of the observed (**G**) rather than the expected (**A**) additive relationship matrix.

An increase in SNP density, after a certain threshold, seems to not affect the quality of the estimated observed relationship matrix **G.** The accuracy obtained from using the 777K SNP panel is not any different from using the 54K SNP panel (Su et al. 2012a, b). This is because the 777K panel did not improve the quality of **G** in any significant way. Results from the human 1000 Genomes Project indicated that the majority of variants observed in a given genome are common and only 40 to 200k are rare (MAF <0.5%). Assuming similar trends will be observed in livestock species, computing the genomic similarity based on common and all or the majority of rare variants will lead to inaccurate estimates of the observed additive relationships and very likely a reduction in the performance of genomic selection.

On top of these challenges there are the added computational costs. Although the computational cost increases almost linearly with increasing number of genotyped animals in RM approaches, that is not the case with increasing number of variants. Thus, the approach will become almost impossible computationally when using sequence variant genotypes. Such costs will not be reduced even when methods for variant prioritization (BayesB, BayesR) are implemented because the detection of “relevant”

variants in each round will offset the computational benefit generated by estimating the effects of only a subset of SNPs. For VC-based approaches, the number of sequence variants will have very little computational cost. However, those computational costs increase cubically with the number of genotyped animals, making direct inversion of the matrix **G** difficult even with medium size data sets. The Algorithm for Proven and Young animals (APY) method developed by Fragomeni et al. (2015a, b) to approximate the inverse of **G** is intrinsically data-driven and could result in computational problems. As a data-driven approximation, its performance is not guaranteed with a continuous increase in the number of genotyped animals spanning several generations and more complex pedigree structures (inbreeding). Furthermore, the APY method will not benefit from the availability of high density SNP panels or NGS data.

Given these limitations, filtering (prioritization) of variants to be included in the association models has become 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 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 the efficiency of these methods. Consequently, with the current density of sequence variants, it is clear that statistical discriminatory criteria alone will not be enough to prioritize influential variants, and enlistment of additional external sources of information seems to be an attractive alternative. BayesRC

(MacLeod et al., 2016), which is 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 (Hayes et al., 2014).

Our objectives in this dissertation is to develop methodologies for implementing GS in the advent of ever increasing (1) density of SNP panels and NGS data and (2) number of genotyped animals. We aim to develop alternative methods to the current variant-prioritizing approaches through the use of population genetic parameters, such as F_{ST}.

Reference

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S. & Lawlor, T. J. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743-752.

Amin, N., van Duijn, C. M., and Aulchenko, Y. S. (2007). A genomic background based method for association analysis in related individuals. PLoS One 2:e1274.

Andersson, L. (2001). Genetic dissection of phenotypic diversity in farm animals. Nat. Rev. Genet. 2: 130- 138.

Brøndum, R. F., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse, W. F., Lund M. S. (2011). Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations J. Dairy Sci., 94, pp. 4700-4707

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am. J. Hum. Genet 86, 6–22.

Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A. and Muir, W. M. (2011). Effect of different genomic relationship matrices on accuracy and scale. Journal of Animal Science 89, 2673-2679.

Christensen, O. F., and Lund, M. S. (2009). Genomic relationship matrix when some animals are not genotyped. Page 299 in Proc. 60th Annual Meeting EAAP, Barcelona, Spain. Wageningen Press, Wageningen, the Netherlands.

Christensen, O. F. and Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42, 2.

Cho, Y. S., Chen, C. H., Hu, C., Long, J., Ong, R. T., Sim, X., Takeuchi, F., Wu, Y., Go, M. J., Yamauchi, T., Chang, Y. C., Kwak, S. H., Ma, R. C., Yamamoto, K., Adair, L. S., Aung, T., Cai, Q. et al. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 44: 67-72.

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

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H., and Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. J. Anim. Sci. 90:3375-3384.

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R. F., et al. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858-65.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182, 375–385.

De Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327-345.

De Roos, A., Hayes, B. J., Spelman, R. and Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 179:1503–1512.

De Vries, P. S., Sabater-Lleal, M., Chasman, D. I., Trompet, S., Ahluwalia, T. S., Teumer, A. et al. (2017). Comparison of HapMap and 1000 genomes reference panels in a large-scale genome-wide association study. PLoS One 12: e0167742.

Druet, T. Georges, M. (2010). A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184, 789–798.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11, 446–450.

Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., Mason, B., Goddard, M. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 95(7):4114-29.

Erbe, M. (2013) Accuracy of genomic prediction in dairy cattle. PhD Thesis, George-August University, Göttingen, Germany.

Fanous, A. H., Zhou, B., Aggen, S. H., Bergen, S. E., Amdur, R. L., Duan, J., Sanders, A. R., Shi, J., Mowry, B. J., Olincy, A. et al. (2012). Genome-Wide Association Study. 2012. Genome-wide association study of clinical dimensions of schizophrenia: polygenic effect on disorganized symptoms. Am J Psychiatry 169: 1309-1317.

Forni, S., Aguilar, I. & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43, 1.

Fragomeni, B. O., Lourenco, D. A. L., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., Lawlor, T. J., Misztal, I. (2015a). Use of genomic recursions in single-step genomic BLUP with a large number of genotypes. J. Dairy Sci. 98:4090-4094.

Fragomeni, B. O., Lourenco, D. A. L., Tsuruta, S., Masuda, Y., Aguilar, I., Misztal, I. (2015b). Use of genomic recursions and Algorithm for Proven and Young animals

for single-step genomic BLUP analyses — A simulation study. J. Anim. Breed. Genet. 132:340-345.

Frazer, K. A., Ballinger, D. G., Cox, D. R. et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. Genetics 183: 347-363.

Gorski, M., Van der Most, P. J., Teumer, A., Chu, A. Y., Li, M., Mijatovic, V., Nolte, I. M., Cocca, M., Taliun, D., Gomez, F. et al. (2017). 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. Scientific Reports 7:45040

González-Recio, O., Gianola, D., Rosa, G. J., Weigel, K. A., Kranis, A. (2009). Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Genet. Sel. Evol 41, 3.

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

Habier, D., Fernando, R. L., Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389-97.

Habier, D., Fernando, R. L., Kizilkaya, K., Garrick, D.J. (2011). Extension of the Bayesian alphabet for genomic selection. BMC Bioinf 12(1):1.

Hannum, G., Srivas, R., Guenole, A., van Attikum, H., Krogan, N, J., Karp, R. M., and Ideker, T. (2009). Genome-wide association data reveal a global map of genetic interactions among protein complexes. PLoS Genet 5: e1000782.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Ed. 2. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

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

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet 6: e1001139.

Hayes, B. J., Macleod, I., Daetwyler, H. D., Bowman, P. J., Chamberlian, A., Vander Jagt, C., Capitan, A., Pausch, H., Stothard, P., Liao, X. (2014). Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, 17-22 August 2014, Vancouver, BC, Canada, pp. 1–6.

Henderson, C. R. (1984). Applications of Linear Models in Animal Breeding. Guelph Univ. Press, Guelph, Canada.

Higgins, J.P., Thompson, S. G., Deeks, J. J., Altman, D. G. (2003). Measuring inconsistency in meta-analyses. BMJ 327:557–560.

Hill, W. G. (2008). Estimation, effectiveness and opportunities of long term genetic improvement in animals and maize. Lohmann Inf 43: 3–20.

Im, H. K., Gamazon, E. R., Nicolae, D. L., and Cox, N. J. (2012). On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet 90: 591-598.

Ioannidis, J. P. (2007). Non-replication and inconsistency in the genome-wide association setting. Hum. Heredity 64:203–213

Kachman, S., Spangler, M., Bennett, G., Hanford, K., Kuehn, L., Snelling, W., Thallman, R., Saatchi, M., Garrick, D., Schnabel, R., Taylor, J., Pollak, E. (2013). Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genet. Sel. Evol. 45:30.

König, S., Simianer, H., and Willam, A. (2009). Economic evaluation of genomic breeding programs. J. Dairy Sci 92:382–391.

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

Leutenegger, A. L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 73: 516-523.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., Liu, J., Warburton, M. L., Cheng, Y., Hao, X., Zhang, P., Zhao, J., Liu, Y., Wang, G., Li, J., and Yan, J. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45: 43-50.

Lin, D. Y., Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet. Epidemiol 34:60–66.

Lorenzana, R., and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet 120:151- 161.

MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J. et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17:144.

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

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P., Sonstegard, T. S., and Van Tassell, C. P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350.Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Meuwissen, T. and Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics185, 623–631.

Meuwissen, T. H. E., Hayes, B., and Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. Anim. Front. 6: 6-14.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T. and Lee, D. H. (2002). BLUPF90 and related programs (BGF90). In "The 7th World Congress Genetitcs Application Livestock Production", pp. 28, Montpellier, France.

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci 92: 4648-4655.

Pryce, J. E., Bolormaa, S., Chamberlain, A. J., Bowman, P. J., Savin, K., Goddard, M. E., Hayes, B. J. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J Dairy Sci. 93(7):3331–45.

Pryce, J. E., Haile-Mariam, M., Goddard, M. E., Hayes, B.J. (2014) Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet Sel Evol 46:71.

Rautiainen, M. R., Paunio, T., Repo-Tiihonen, E., Virkkunen, M., Ollila, H. M., Sulkava, S., Tiihonen, J. (2016). Genome-wide association study of antisocial personality disorder. Translational Psychiatry 6(9), e883.

Saatchi, S. S. et al. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. Proc. Natl Acad. Sci. USA 108, 9899–9904.

Schaeffer, L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet 123: 218–223.

Scott, R. A. et al. (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. Diabetes 66, 2888–2902.

Shen, X., Alam, M., Fikse, F., Ronnegard, L. (2013). A novel generalized ridge regression method for quantitative genetics. Genetics 193: 1255–1268.

Stephan, R., Sanders, A. R., Kendler, K. S., Levinson, D. F., and Sklar, P. et al. (2011). Genome-wide association study identifies five new schizophrenia loci Nat Genet No. 43. p 969-976.

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

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187: 367-383.

Su, G., Madsen, P., Nielsen, U. S., Mäntysaari, E., Aamand, G. P., Christensen, O. F. et al. (2012a). Genomic prediction for Nordic red cattle using one-step and selection index blending. J Dairy Sci 95:909-917.

Su, G., Brøndum, R. F., Ma, P., Guldbrandtsen, B., Aamand, G. P., Lund, M. S. (2012b). 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. J Dairy Sci. 95(8):4657–65.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. BMJ 309:1351–1355

Thompson, J. R., Attia, J., Minelli, C. (2011). The meta-analysis of genome-wide association studies. Brief. Bioinformatics. 12:259–269.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. J Dairy Sci 91:4414-23.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C. and Sonstegard,

T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247-252.

Vineis, P., Schatzkin, A. and Potter, J. D. (2010) Models of carcinogenesis: an overview. Carcinogenesis 31: 1703 – 1709.

Vitezica, Z. G., Aguilar, I., Misztal, I. & Legarra, A. (2011). Bias in genomic predictions for populations under selection. Genetics Research 93, 357-366.

Weedon, M. N. and Frayling, T. M. (2008). Reaching new heights: insights into the genetics of human stature. Trends Genet 24, 595-603

Willer, C. J., Li, Y., Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190–2191.

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

Wolc, A., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., Habier, D., Fernando, R., Garrick, D. J., Hill, W. G., and Dekkers, J. C. (2012). Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. Anim Genet 43 Suppl 1: 87-96.

Wright, S. (1951). The genetical structure of populations. Ann Hum Genet 15:323–54.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N, G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.

Yoon, D., Kim, Y. J., and Park, T. (2012). Phenotype prediction from genome-wide association studies: application to smoking behaviors. BMC Syst Biol 6 Suppl 2: S11.

Zaitlen, N., Lindström, S., Pasaniuc, B., Cornelis, M., Genovese, G., Pollack, S., Barton, A., Bickeböller, H., Bowden, D. W., and Eyre, S. (2012). Informed conditioning on clinical covariates increases power in case-control association studies. PLoS genetics 8: e1003032.

Zeggini, E., Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. Pharmacogenomics.10:191–201.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE 5:e12648.

Zhang, R., Yan, J. D., Valenzuela, R. K., Lu, S. M., Du, X. Y., Zhong, B., Ren, J., Zhao, S. H., Gao, C. G., Wang, L., Guo, T. W., and Ma, J. (2012). Further evidence for the association of genetic variants of ZNF804A with schizophrenia and a meta-analysis for genome-wide significance variant rs1344706. Schizophr Res 141: 40-47.

Zhang, J., Song, Q., Cregan, P. B., Nelson R. L., Wang, X., Wu, J., and Jiang, G, L. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. BMC Genomics 16:217.

1000 Genomes Project Consortium, Auton A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A.,

Abecasis, G. R. (2015). A global reference for human genetic variation. Nature 526:68–74.

CHAPTER 3

HIGH DENSITY MARKER PANELS, SNPS PRIORITIZING AND ACCURACY OF GENOMIC SELECTION [1]

[1] Chang, LY, S. Toghiani, A. Ling, S. E. Aggrey, R. Rekaya. 2018. *BMC Genetics* 19:4. Reprinted here with permission of publisher.

Abstract

The availability of high-density (HD) marker panels, genome wide variants and sequence data creates an unprecedented opportunity to dissect the genetic basis of complex traits, enhance genomic selection (GS) and identify causal variants of disease. The disproportional increase in the number of parameters in the genetic association model compared to the number of phenotypes has led to further deterioration in statistical power and an increase in co-linearity and false positive rates. At best, HD panels do not significantly improve GS accuracy and, at worst, reduce accuracy. This is true for both regression and variance component approaches. To remedy this situation, some form of single nucleotide polymorphisms (SNP) filtering or external information is needed. Current methods for prioritizing SNP markers (i.e. BayesB, BayesCπ) are sensitive to the increased co-linearity in HD panels which could limit their performance. In this study, the usefulness of F_{ST}, a measure of allele frequency variation among populations, as an external source of information in GS was evaluated. A simulation was carried out for a trait with heritability of 0.4. Data was divided into three subpopulations based on phenotype distribution (bottom 5%, middle 90%, top 5%). Marker data were simulated to mimic a 770K and 1.5 million SNP marker panel. A ten-chromosome genome with 200K and 400K SNPs was simulated. Several scenarios with varying distributions for the quantitative trait loci (QTL) effects were simulated. Using all 200K markers and no filtering, the accuracy of genomic prediction was 0.77. When marker effects were simulated from a gamma distribution, SNPs pre-selected based on the 99.5, 99.0 and 97.5% quantile of the F_{ST} score distribution resulted in an accuracy of 0.725, 0.797, and 0.853, respectively. Similar results were observed under other simulation scenarios.

Clearly, the accuracy obtained using all SNPs can be easily achieved using only 0.5 to 1% of all markers. These results indicate that SNP filtering using already available external information could increase the accuracy of GS. This is especially important as next-generation sequencing technology becomes more affordable and accessible to human, animal and plant applications.

Keywords: SNP prioritizing, genomic selection, high density

Introduction

Large-scale genotyping for single-nucleotide polymorphisms (SNPs) has provided an unprecedented resource to study associations between traits and genomic variation and to compute genomic enhanced breeding values (GEBVs). Although a detailed dissection at the genetic level of these complex traits is still largely elusive, continuous improvements in the quality and diversity of high-throughput data, as well as the development of more sophisticated statistical and computational tools, are quickly moving us towards a better understanding of the genetic basis of these traits. Genomic selection (GS) is rapidly becoming the tool of choice for genetic evaluation of several livestock species due to an increase in accuracy and substantial reduction of the generation interval (VanRaden et al., 2009; Su et al., 2010; Su et al., 2012; Schefers and Weigel, 2012; Zeng et al., 2013). Genomic selection is currently being implemented either through a multiple regression (RM) or variance component (VC) based models. The RM approach consists of a multiple step procedure where SNP effects are first estimated in a training population and then validated in a separate data set. Several methods have been developed and used to implement this approach (Meuwissen et al., 2001; Xu, 2003; Fernando et al., 2007; Habier et al. 2007; Muir, 2007; Xu, 2007; VanRaden, 2008; Bennewitz et al. 2009; Gianola et al., 2009; Habier et al., 2011). 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; differences are largely due to the genetic architecture of the trait, the genetic relationships between individuals in the data, and the chosen prior information.

Next generation sequencing (NGS) has dramatically changed the speed, coverage and costs of sequencing whole genomes. Several sequencing efforts, including the 1000 Bull Genomes Project (Daetwyler et al., 2014), are underway, while several thousand humans and animals have already been fully sequenced (The 1000 Genomes Project Consortium, 2015). These projects are crucial for characterizing the source of genetic variation. In fact, 84 and 31.8 million common and rare variants have been already identified in human and dairy cattle, respectively (The 1000 Genomes Project Consortium, 2015; Hayes et al. 2014). Although a majority of these variants are rare (MAF <1%), over 8 million common SNPs (MAF > 5%) have already been identified in humans. Thus, it is already a reality that genome-wide association study (GWAS) and ultimately GS will be implemented using several millions of directly- or indirectly-imputed sequence variant genotypes. GWAS using 17 and 19 million SNPs were carried out in human (The 1000 Genomes Project Consortium, 2015) and dairy cattle (Daetwyler et al., 2014) applications, respectively. Although theoretically there are no doubts about the potential usefulness of the sequence data in GWAS and GS, major challenges are limiting the harnessing of these benefits.

The major problem of the analysis of high dimensional SNP data and sequence variant genotypes stems mainly from the high dimensionality of the parameter space. When all variants are considered (i.e., BayesA), the highly informative prior will lead to excessive shrinkage that together with the high linkage disequilibrium (LD) precludes the identification of causative mutations or even of significant tag variants. As the effect of a QTL (often small for complex traits) is distributed in a non-trivial manner between all markers that are in LD with the causal mutation, there is little statistical power to

accurately estimate its effect. Given these limitations, filtering (prioritization) of variants to be included in the association models has become a necessity. Traditionally, SNP filtering is conducted based on certain statistical criteria such as p-values for single-marker analyses (Balding, 2006; Zheng et al. 2009) 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 including 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 the efficiency of these methods. Consequently, with the current density of sequence variants, it is clear that statistical discriminatory criteria alone will not be enough to prioritize influential variants, and enlistment of additional external sources of information seems to be an attractive alternative. BayesRC (MacLeod et al., 2016) which is an extension of BayesR through the inclusion of biological prior information (variant type, location in differentially expressed genes), has only led to slight increase in accuracy compared to BayesR (Hayes et al., 2014).

The limited success so far of these SNP/variant-prioritizing methods is due to several reasons: 1) the "artificial" reduction in the number of parameters in the model. Although marker prioritization methods based on statistical criteria (BayesB, and BayesR) reduce the number of parameters (variants) fitted in the association model in every round of the iterative process, the total number of unknowns to be inferred in each round is at least equal to the number of parameters in a fully parameterized model (i.e., BayesA). This is due to the need to identify those markers with zero effects which is

often accomplished either through a Metropolis-Hastings step or through the estimation of indicator variables in a data augmentation approach; 2) currently available biological information is often limited (tissue specific, time specific, etc.) and with a high noise-to-signal ratio (gene expression, methylation profiles, etc.); and 3) small QTL effects in LD with a large number of variants.

Consequently, other sources of prior information need to be investigated. Livestock species are under heavy artificial selection. The signature of such selection pressure can be traced through changes in allele frequencies of markers in LD with QTL. F_{ST}, a measure of allele frequency variation among sub-populations, provides a tool to reveal selection sweeps (Lewontin & Krakauer 1973) and can be used to identify SNPs under selection pressure. In this study, a simulation was carried out under different marker densities and complexity of the genetic model to evaluate the usefulness of F_{ST} scores as an external source of information to prioritize SNP markers in the association models and to compare its performance with currently used approaches.

Methods

Simulated SNP genotypes and phenotypes:

Simulation was carried out using QMSim software (Sargolzaei & Schenkel 2009). A historical population was generated based on random mating of 8,000 animals for 300 generations followed by an additional 15 generations of random mating with population size ranging between 12,000 and 17,000 animals. This random mating was carried out to initialize LD and to establish mutation-drift equilibrium in the historical population. The founder population or generation zero (G0) was created 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, four generations were simulated. The third generation (G3) was used to detect selection signatures and the last one (G4) was used to evaluate the proposed approach.

In the last four generations (G1 to G4), animals were selected base on their estimated breeding values (EBVs). Replacement rate for males and females were set to 50 and 20%, respectively. In all generations, one progeny per dam and a sex ratio of 50% were assumed. Only animals in generation three and four were assumed to be genotyped. In order to mimic high-density marker panels, a 10-chromosome genome was simulated with uniformly-distributed 200K and 400K SNP markers, resulting in a density similar to a bovine chip of 600K and 1 million SNPs, respectively. The additive effects of one hundred QTL were sampled either from a Gamma distribution with shape parameter equal to 0.4 or predefined as a fraction of the total genetic variance. In the predefined scenario, QTL effects were set to explain at least 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. A detailed description of the simulated genome structure of the different scenarios is presented in Table 3.1.

The phenotype consisted of one trait with 40% heritability. Phenotypic variance was set equal to one and the residual variance was adjusted in each scenario to maintain the heritability constant at 0.4. 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.

Measure of selection pressure as source of external information:

Wright's F statistics (Wright, 1951) are fixation indexes that measure the rate of fixation through the increase in homozygosity. In particular, F_{ST}, a measure of population structure, is one of the most frequently used scores in the field of genetics. It measures the rate of genetic differentiation between subpopulations through the assessment of the changes in allele frequencies. The larger the F_{ST} values, the higher the genetic differentiation (Wright 1978; Nagylaki 1998; Hedrick 1999; Balloux et al. 2000). Among its multiple uses, F_{ST} can be used to assess signatures of natural and artificial selection.

Although there are several methods to estimate F_{ST} (Nei 1973; Weir & Cockerham 1984; Hudson *et al.* 1992; Weir & Hill 2002), the global estimator proposed by Nei (1973) was used in this study. Animals in generation G3 were divided into three sub-populations based on their simulated phenotype (below the 5 quantile [**S1**], between 5 and 95 quantiles [**S0**], and above the 95 quantile [**S2**]) and F_{ST} score for a given locus was calculated as:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

with $H_T = 2 * p * q$, $H_S = \frac{H_{S1} * n_{s1} + H_{S2} * n_{s2}}{n_{s1} + n_{s2}}$, and $H_{Si} = 2 * p_{Si} * q_{Si}$

where, p_{Si} and q_{Si} are the allele frequencies in subpopulation i, n_{s1} and n_{s2} are the number of individuals of the subpopulations, H_S is the average of sub-population heterozygosities and H_T is the heterozygosity based on the total population.

Animals in subpopulations S1 (below 5 quantile) and S2 (above 95 quantile) of the third generation of simulation data (G3) were used to calculate the F_{ST} scores. A total

of 1,500 genotyped animals equally divided between both groups were used. Three heuristically defined threshold values of F_{ST} scores (Table 3.2) were used to select SNPs that are potentially under genetic differentiation. For 200K SNP panels, the number of selected SNPs was 935, 1,956, and 4,932 for the three threshold values in the gamma distribution scenario, respectively. The number of selected SNPs was 1,076, 2,171, and 5,620 in the predefined distribution scenario, respectively.

Data Analysis:

Each simulated data set was analyzed using BayesB, BayesC, and the proposed method where SNPs selected based on their F_{ST} scores are used as explanatory variables in a regression model similar to BayesA. Implementation of BayesB and BayesC was carried out using GenSel software (Fernando and Garrick, 2009) with (1-π) values set equal to 0.9, 0.95, 0.98, or 0.99. Scaled inverted Chi square prior distributions were assumed for the genetic and residual variances with scaling factors equal to the true values used in the simulation and degrees of freedom of 1 and 4, respectively.

The general statistical model used for analysis in BayesB, BayesC and the proposed method can be presented as:

$$y_i = \mu + \sum_{j=1}^{p} X_{ij}\,\beta_j\gamma_j + e_i$$

where y_i is the phenotype for individual i; μ is an overall mean; X_{ij} is the genotype of individual i for SNP j taking the value of 0, 1, or 2; β_j is the effect of the SNP j; and γ_j is an indicator factor that takes the value of 1 if SNP j is included in the model and 0 otherwise. For the proposed method, $\gamma_j$was equal to 1 for all preselected SNPs. e_i is the

error term and p was equal to the preselected SNPs for the proposed method or the total number of SNPs times (1-π) for BayesB and BayesC.

Point estimates of the SNPs effects were used to compute the estimated genomic breeding values as:

$$GEBV_i = \sum_{j=1}^{p} z_{ij}\hat{a}_j$$

where $\hat{a}_j$ is the estimated effect of SNP j

Genomic and phenotype accuracies were calculated based on the correlation between the true breeding values and the GEBVs and between the GEBVs and the observed phenotypes adjusted for the systematic effects.

For each simulated data set, randomly 10,000 genotyped animals in the third generation (G3) were assigned to the training population and randomly 5,000 genotyped animals in the last generation (G4) were used for validation. Each simulation scenario was replicated 5 times. For BayesB and BayesC, four $(1-\pi)$ values (0.99, 0.98, 0.95, and 0.9) were evaluated.

Results and discussion

Distribution of QTL and estimated F_{ST} scores:

Figure 3.1 presents the distribution and effects of the 100 QTL simulated from a gamma distribution with a shape parameter of 0.4 (Fig 3.1a) and the F_{ST} scores for the 200 K SNPs (Fig. 3.1b). The largest QTL explained about 13.2% of the total genetic variance (GV). The top 15 QTL explained over 70% of the GV while the bottom 50% of QTL explained less than 0.05% of the GV each. The distribution of estimated F_{ST} scores

(Fig. 3.1b) showed a striking similarity to the true QTL distribution (Fig. 3.1a), especially for large effect QTL. For QTL with effect greater than 0.2 (Fig. 3.1a) there were three distinguished peaks that are easily captured by the F_{ST} scores under the three threshold cutoff values (Fig. 3.1b). This result was not unexpected given the large simulated effects for the top QTL. Due to selection, SNPs in LD with these QTL will experience quick and substantial change in their minor allele frequencies that will be easily captured by the F_{ST} scores. When the QTL effects were pre-defined (each QTL explains at least 0.5% of GV), the QTL with the largest effect explained 1.5% of GV and the bottom 50% of QTL explained between 0.5 and 1% of the GV each (Fig. 3.1c and 3.1d). Even under this complex genetic model and absence of large effect QTL, SNPs selected based on F_{ST} scores were able to track the majority of QTL with as little as 3% of all SNP in the panel (Fig. 3.1d). Similar results were observed when a 400 K SNP panel was considered (Fig. 3.S1).

Accuracy of genomic selection: Population genetics approach:

Table 3.3 presents the accuracy of prediction of both of the true breeding values and the simulated phenotypes in the case where the SNPs were preselected based on F_{ST} scores. Accuracy was calculated based on the correlation between the true parameters (breeding values or phenotypes) and their associated prediction on the validation data set (G4). All results are based on the average of 5 replicates for each simulated data set. Using all SNPs in the 200K and 400K panels resulted in genomic accuracy of 0.777 and 0.775, respectively when the QTL effects were generated from a gamma distribution. When the QTL effects were predefined, the corresponding accuracies were 0.741 and

0.735. This drop in accuracy is due in part to the increased complexity of the genetic model in the case of predefined QTL effects which resulted in a reduction in the portion of GV explained compared to the scenario when QTL were simulated from a gamma distribution. When SNPs were preselected based on their F_{ST} scores under the 200K marker panel and gamma distribution for the simulation of the QTL effects scenario, genomic accuracy increased from 0.725 to 0.853 when the preselected SNPs were based either on the 99.5 (1076 SNPs) or 97.5 quantile (4932 SNPs) of the distribution of the F_{ST} scores. Similarly, the number of tagged QTL ($r^2 > 0.7$ with at least one selected SNP) and the portion of GV explained increased from 13 to 33 and 64.08 to 83.70%, respectively. When the QTL effects were pre-defined to explain at least 0.5% of the GV, the same trend was observed as when the QTL were simulated from a gamma distribution, except that the accuracies and portion of genetic variance explained were smaller and the number of tagged QTL was larger for the same quantile. At the 97.5 quantile, 69% of the QTL were tagged for the predefined scenario versus 33% in the gamma distribution scenario. However, the predefined scenario explained only 71.27% of the GV compared to 83.70% in the gamma distribution scenario. This is obviously due to the change in the complexity of the genetic model. Using the 400K marker panel, accuracies, number of tagged QTL and portion of GV explained increased compared to the 200K SNP scenario (Table 3.3). This is likely due to an increase in LD between preselected SNPs and QTL. However, the difference between the two marker density scenarios is small for the 97.5 quantile case. This indicates that in this case, around 5,000 SNPs are needed to track the majority of the QTL and any additional markers will increase accuracy marginally. Across all simulation scenarios, phenotype prediction accuracy has the same trend as the

accuracy of genomic enhanced breeding values (GEBV) although with a much lower magnitude, as expected (Table 3.3). It is worth mentioning that although the optimum number of preselected SNPs was not determined in this study, a continuous increase in the number of markers in the association model will at some point lead to a decrease in accuracy. This is well supported by the lower accuracy when all SNPs were included in the model (Table 3.3).

Figure 3.2 presents the distribution of simulated QTL across the 10 chromosomes and the preselected SNPs based on 99.5 (Fig. 3.2a) and 97.5 quantile (Fig. 3.2b) of the F_{ST} score distribution for the 200K marker panel and predefined QTL effect scenario (Fig. 3.S2.1-S2.3 present the results for the remaining scenarios). It is clear that when only SNPs with a large F_{ST} score were preselected (Fig. 3.2a), only large QTL were tagged. As more SNPs are preselected (Fig. 3.2b), most of the QTL (70%) were tagged and a large proportion of the GV was explained. When the QTL were simulated from a gamma distribution, although only the most influential QTL were effectively tagged, the majority of the GV was explained even when SNPs were preselected based on their F_{ST} score exceeding the 99.5 quantile of the distribution (Table 3.3).

In order to further evaluate the performance of the SNP prioritization approach based on F_{ST} scores, a comparison with well-established and extensively used methods was carried out. The same simulated data sets were analyzed using BayesB, and BayesC implemented by GenSel software (Fernando and Garrick, 2009). For BayesB and BayesC, four π values (0.01, 0.02, 0.05 and 0.10) were evaluated. Table 3.4 presents the accuracies using BayesB. For both marker densities, the accuracies increased with the decrease of π with the maximum at π=0.01. For the 200K SNP scenario, accuracy of

predicted GEBV ranged from 0.797 to 0.845 and from 0.770 to 0.833 when QTL effects were simulated from a gamma distribution or predefined, respectively. A similar trend was observed for the 400K SNP scenario, although the magnitude of the accuracies was slightly smaller. Using BayesC, the results were very similar to those obtained using BayesB, although they tended to be slightly higher for the latter (Table 3.5). When compared with the proposed method, BayesB and BayesC have slightly lower accuracies of genomic prediction in all simulation scenarios, except the 200K SNP marker density and predefined QTL effect scenario using BayesB (Table 3.6). In fact, the superiority in the remaining scenarios ranged from 0.74 to 3.60% and 1.08 to 4.19% compared to BayesB and BayesC, respectively. Similar trend was observed for the phenotype prediction accuracy. Phenotypic accuracy was lower using the proposed method only for the 200K SNP marker panel and predefined QTL effect scenario (Table 3.6).

In the 200K SNP marker panel simulation scenario we tried to mimic the LD observed in the Bovine 770K chip. Thus, simulations were carried out with LD between adjacent SNPs ranging between 0.65-0.70. In order to test the performance of the proposed method when LD is lower, the 200K SNP marker density and gamma QTL effect scenario was re-simulated with LD between adjacent markers of around 0.3. The results showed that across all three methods (BayesB, BayesC and our proposed method), accuracy decreased by 18 to 20% compared with the scenario with higher LD. Furthermore, the three methods have similar results with a slight superiority (0.53%) for BayesB.

Bayesian methods for prioritizing SNPs rely on sound statistical foundation. However, their performance is expected to decay with the increase in the density of the

marker panel at least for two reasons: 1) an increase in the number of unknowns in the association model leading to an increase in the statistical cost of finding the relevant SNPs (non-zero effect SNPs) and 2) an increase in the number of markers in the panel increasing the number of SNPs that are in high LD with the QTL. Consequently, the effect of each QTL will be partitioned across an increasing number of markers leading to smaller effects of associated SNPs. Because these methods rely on the magnitude of the estimated marker effects to determine the relevant SNPs, their performance will undoubtedly decay due to a lack of statistical power. However, the proposed method pre-selects markers based on the change of their minor allele frequencies rather than the magnitude of their effects. Thus, it does not suffer from the problem indicated before, but it is prone to some redundancy in the selected SNPs because markers with very high LD will have similar F_{ST} score.

In this study, a homogeneous population was assumed. The proposed method could be modified in presence of admixed populations. Specifically, in the presence of an admixed population the change in the minor allele frequency (MAF) of SNPs and consequently of F_{ST} scores could be the result of selection pressure on linked QTLs or simply due to difference in MAF between components (breeds) of the population. The latter will not be useful to prioritize SNPs. However, in presence of admixed population we suggest performing within breed SNP prioritization which will take care largely of the difference in MAF. SNPs prioritized in more than one breed (at least those with the largest F_{ST} scores) should be tested for LD phase consistency. This could be a manageable task given the limited number of prioritized SNPs. Furthermore, selected SNPs will have effect only in the subpopulations (breeds) where they were prioritized

increasing potentially the power of the association model. However, within subpopulation SNP prioritization could be problematic for breeds with small number of genotyped individuals. In such case grouping for genetically closer breeds could be used for SNP prioritization.

Across all simulation scenarios, we tried to mimic high density SNP panels used in livestock applications where causal variants were assumed not to be genotyped. However, with the recent availability of sequence data large portion of causative variants will be genotyped. Furthermore, these variants could have rare frequencies (MAF <1%). These two issues could have impact on the performance of our method as well as other approaches. However, it is intuitive to think that the F_{ST} method will perform even better because causative variants or those in very high or complete LD with them will, in general, see their minor allele frequencies change more significantly than other variants resulting in higher F_{ST} score and easy prioritization. This might not be the case for competing methods (BayesB and BayesC) were prioritization is based on the effect of variants.

Conclusion

A continuous increase in the density of SNP marker panels and the availability of whole genome sequence data provide an unprecedented opportunity to dissect the genetic basis of complex traits and to enhance the estimation of genetic merit in animal and plant applications. Unfortunately, this dramatic increase in the available genomic data has created some implementation problems and most importantly did not lead to any significant increase of accuracy of genomic selection using single- and multiple-step

approaches. For the latter, the massive increase in the number of explanatory variables has led to an over-parametrization of the association model which resulted in increased co-linearity and loss of statistical power. Together these factors led to no increase in accuracy of genomic selection. Limitations of current models stem from the lack of information on the genotyped individual to prioritize SNPs marker to be considered in the association model. Furthermore, methods based on statistical criteria to filter SNPs will see their performance decay as the marker density increases due to the reduction in the effects of SNPs associated with QTL. Using external information (i.e. gene expression data) is attractive and could compensate for the limited information in the data. Unfortunately, such external information is not always available, often is tissue or time specific, and could have high noise-to-signal ratio. In this study, we proposed using F_{ST} score as an alternative to existing method to prioritize SNPs in high-density marker panels. Although this information is internal to the data, the results of this study suggest that it could provide a reliable tool for prioritization of SNPs.

Reference

Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7(10):781–791. doi: 10.1038/nrg1916.

Balloux F, Brunner H, Lugon-Moulin N, Hausser J, Goudet J. Microsatellites can be misleading: an empirical and simulation study. Evolution. 2000;54 (4): 1414-1422.

Bennewitz J, Solberg T, Meuwissen T. Genomic breeding value estimation using nonparametric additive regression models. Genet Sel Evol. 2009;41: 20-10.1186/1297-9686-41-20.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858-65.

Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, Mason B, Goddard M. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114-29.

Fernando RL, Garrick D: GenSel-User Manual for a portfolio of Genomic Selection related Analyses. Animal Breeding and Genetics. Iowa State University, Ames, IA. 2009. [http://taurus.ansci.iastate.edu] Accessed 10 March 2017.

Fernando RL, Habier D, Stricker C, Dekkers JCM, Totir LR. Genomic selection. Acta Agriculturae Scandinavica. 2007;57(4):192-5.

Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando RL. Additive genetic variability and the Bayesian alphabet. Genetics. 2009;183(1):347-63.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinf. 2011;12(1):1.

Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177(4):2389-97.

Hayes, B. J., I.M. MacLeod, H.D. Daetwyler, P.J. Bowman, A.J. Chamberlian, C.J. Vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten , E.

Mullaart, R. Fries, B. Guldbrandtsen, M.S. Lund, D.A. Boichard, R.F. Veerkamp, C.P. VanTassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D.J. deKoning, E. Santus, and M.E. Goddard. Genomic Prediction from Whole Genome Sequence in Livestock: the 1000 Bull Genomes Project. Proceedings of the 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada. 17-22 August 2014.

Hedrick PW. Highly variable loci and their interpretation in evolution and conservation. Evolution. 1999;53,313-318.

Hudson GS, Evans JR, von Caemmerer S, Arvidsson YBC, Andrewset TJ. Reduction of ribulose-1,5-bisphosphate carboxylase/oxygenase content by antisense RNA reduces photosynthesis in transgenic tobacco plants. Plant Physiol. 1992;98:294-302.

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

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17:144.

Meuwissen THE, Goddard ME. Prediction of identity by descent probabilities from marker haplotypes. Genet Sel Evol. 2001;33:605-634. 10.1186/1297-9686-33-6-605.

Muir WM. Genomic selection, accuracy and comparisons with traditional BLUP under alternative marker density and generations of training. Journal of Animal Breeding and Genetics. 2007;6:342-355.

Nagylaki T. Fixation indices in subdivided populations. Genetics. 1998;148,1325-1332.

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

Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25(5):680-1.

Schefers JM, Weigel KA. Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Animal Frontiers. 2012;2:4-9.

Su G, Guldbrandtsen B, Gregersen VR, Lund MS. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J Dairy Sci. 2010;93(3):1175-83.

Su G, Madsen P, Nielsen US, Mäntysaari E a, Aamand GP, Christensen OF, et al. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J Dairy Sci. 2012;95:909-17.

VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414-23.

VanRaden P, Van Tassell C, Wiggans G, Sonstegard T, Schnabel R, Taylor J, Schenkel F. Invited review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92(1):16–24.

Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38(6):1358-1370.

Weir BS, Hill WG. Estimating F-statistics. Annu Rev Genet. 2002;36:721–50.

Wright S. The genetical structure of populations. Ann Hum Genet. 1951;15:323–54.

Wright S. Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations. University of Chicago Press, Chicago. 1978.

Xu S. Estimating polygenic effects using markers of the entire genome. Genetics. 2003;163–789:801.

Xu S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics. 2007;63:513–21.

Zeng J, Toosi A, Fernando RL, Dekkers JC, Garrick DJ. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol. 2013;45(1):1.

Zheng G, Xu J, Yuan A, Gastwirth JL. Single Marker Association Analysis for Unrelated Samples. Methods in molecular biology (Clifton, NJ). 2012;850:10.1007/978-1-61779-555-8_18. doi:10.1007/978-1-61779-555-8_18.

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.

Table 3.1. Descriptive statistics of simulation schemes

Historical Population (HP)	
Number of generation	315
Mutation rate for markers	$1.0*10^{-4}$
Mutation rate for QTL	$1.0*10^{-4}$
Founder Population (G0)	
Number of generation	3
Number of male	1,500
Number of female	15,000
Selection Population (G3)	
Number of chromosomes	10
Length per chromosome (cM)	100
Number of markers per generation	200,000 / 400,000
Marker distribution	Evenly spaced
Number of QTL per generation	100
QTL distribution	Randomly distributed
QTL effect	Sampled from gamma with shape 0.4
Heritability	0.4
Genetic variance	0.4
Residual variance	0.6

Table 3.2. Preselected SNPs based on different cutoff values for the FST scores and different simulation scenarios

Panel density	QTL effects[1]	Quantile	F_{ST} Score[3]	Selected SNPs[4]
200K	Gamma[5]	99.5	0.02	935
		99.0	0.01	1,956
		97.5	0.004	4,932
	Predefined[6]	99.5	0.009	1,076
		99.0	0.007	2,171
		97.5	0.005	5,620
400K	Gamma	99.5	0.015	2,078
		99.0	0.009	3,586
		97.5	0.004	10,178
	Predefined	99.5	0.009	2,036
		99.0	0.007	4,646
		97.5	0.004	10,651

[1] Distribution used for the simulation of the QTL effects; [2] quantiles of the F_{ST} score distribution; [3] cutoff point for the fixation index (F_{ST}); [4] number of selected SNPs based on the F_{ST} score cutoff; [5] Gamma distribution with shape parameter equal 0.4; [6] QTL effects pre-defined to explain at least 0.5% of genetic variance each.

Table 3.3. Number of selected SNPs, number of tagged QTLs, percentage of genetic variance explained, and accuracies of genomic and phenotype prediction under different quantile of the distribution of FST scores, sampling distribution for the QTL effects and density of the marker panel using the proposed method. Standard errors of accuracies are listed between parentheses

	200K SNP marker panel							
	All SNPs		97.5 quantile[1]		99.0 quantile		99.5 quantile	
	Gamma[2]	Predefined[3]	Gamma	Predefined	Gamma	Predefined	Gamma	Predefined
Selected SNP	200K	200K	4932	5620	1956	2171	935	1076
Tagged QTL[4]	95	97	33	69	18	47	13	31
% GV[5]	91.29	98.60	83.70	71.27	73.57	49.69	64.08	35.10
Acc_P[6]	0.462	0.445	0.503	0.490	0.472	0.415	0.434	0.359
	(0.018)	(0.012)	(0.017)	(0.014)	(0.015)	(0.018)	(0.028)	(0.032)
Acc_G[7]	0.777	0.741	0.853	0.830	0.797	0.704	0.725	0.617
	(0.017)	(0.012)	(0.019)	(0.023)	(0.017)	(0.031)	(0.037)	(0.026)
	400K SNP marker panel							
Selected SNP	400K	400K	10173	10651	3586	4646	2078	2037
Tagged QTL	95	99	38	74	20	53	13	34
% GV	96.73	99.01	84.03	75.09	73.83	56.66	66.12	43.79
Acc_P	0.456	0.438	0.506	0.485	0.473	0.433	0.448	0.350
	(0.015)	(0.017)	(0.014)	(0.017)	(0.029)	(0.021)	(0.039)	(0.028)
Acc_G	0.775	0.735	0.860	0.813	0.807	0.722	0.765	0.685
	(0.020)	(0.012)	(0.015)	(0.012)	(0.041)	(0.025)	(0.059)	(0.052)

[1] quantile of the distribution of the F_{ST} scores; [2] QTL effects sampled from a Gamma distribution; [3] QTL effects pre-defined to explain at least 0.5% of genetic variance (GV); [4] QTL with $r^2 > 0.7$ with at least one selected SNP; [5] GV = Genetic Variance; [6] accuracy of phenotype prediction; [7] accuracy of genomic prediction

Table 3.4. Number of selected SNPs, number of tagged QTL, percentage of genetic variance explained, and accuracies of genomic and phenotype prediction under different π values, sampling distribution for the QTL effects and density of the marker panel using BayesB method. Standard errors of accuracies are listed between parentheses

	200K marker density							
	(1-π) =0.90		(1-π) =0.95		(1-π) =0.98		(1-π) =0.99	
	Gamma[1]	Predefined[2]	Gamma	Predefined	Gamma	Predefined	Gamma	Predefined
# SNP	20K	20K	10K	10K	4K	4K	2K	2K
Tagged QTL[3]	78	98	63	97	54	94	48	91
% GV[4]	89.31	98.16	86.43	97.88	84.30	95.76	83.88	93.20
Acc_P[5]	0.473	0.463	0.478	0.471	0.489	0.487	0.499	0.500
	(0.018)	(0.009)	(0.018)	(0.009)	(0.018)	(0.008)	(0.018)	(0.007)
Acc_G[6]	0.797	0.770	0.807	0.785	0.827	0.810	0.845	0.833
	(0.017)	(0.008)	(0.017)	(0.007)	(0.018)	(0.007)	(0.018)	(0.005)
	400K marker density							
# SNP	40K	40K	20K	20K	8K	8K	4K	4K
Tagged QTL	86	99	75	98	59	97	53	96
% GV	92.36	98.46	91.88	98.16	91.20	97.78	91.03	96.69
Acc_P	0.465	0.450	0.470	0.457	0.478	0.469	0.488	0.481
	(0.015)	(0.018)	(0.015)	(0.018)	(0.014)	(0.018)	(0.013)	(0.019)
Acc_G	0.790	0.756	0.799	0.767	0.813	0.787	0.829	0.807
	(0.019)	(0.013)	(0.017)	(0.013)	(0.016)	(0.014)	(0.015)	(0.014)

[1] QTL effects sampled from a Gamma distribution; [2] QTL effects pre-defined to explain at least 0.5% of genetic variance (GV); [3] QTL with $r^2 > 0.7$ with at least one selected SNP; [4] GV = Genetic Variance; [5] accuracy of phenotype prediction; [6] accuracy of genomic prediction

Table 3.5. Number of selected SNPs, number of tagged QTL, percentage of genetic variance explained, and accuracies of genomic and phenotype prediction under different π values, sampling distribution for the QTL effects and density of the marker panel using BayesC method. Standard errors of accuracies are listed between parentheses

	200K marker density							
	(1-π) =0.90		(1-π) =0.95		(1-π) =0.98		(1-π) =0.99	
	Gamma	Predefined	Gamma	Predefined	Gamma	Predefined	Gamma	Predefined
# SNP	20K	20K	10K	10K	4K	4K	2K	2K
Tagged QTL[3]	76	97	61	96	53	94	46	91
% GV[4]	88.84	97.66	86.56	97.53	86.30	95.74	85.76	93.32
Acc_P[5]	0.453	0.451	0.467	0.459	0.484	0.477	0.496	0.493
	(0.019)	(0.009)	(0.019)	(0.009)	(0.018)	(0.008)	(0.018)	(0.008)
Acc_G[6]	0.769	0.751	0.791	0.766	0.821	0.794	0.842	0.821
	(0.017)	(0.009)	(0.018)	(0.008)	(0.018)	(0.009)	(0.018)	(0.006)
	400K marker density							
# SNP	40K	40K	20K	20K	8K	8K	4K	4K
Tagged QTL	85	99	68	98	53	97	48	95
% GV	92.05	98.97	91.59	98.37	90.98	96.95	90.16	95.81
Acc_P	0.444	0.441	0.456	0.447	0.472	0.459	0.485	0.472
	(0.013)	(0.017)	(0.013)	(0.017)	(0.014)	(0.017)	(0.014)	(0.018)
Acc_G	0.754	0.740	0.773	0.749	0.802	0.769	0.824	0.791
	(0.017)	(0.011)	(0.017)	(0.011)	(0.017)	(0.012)	(0.016)	(0.012)

[1] QTL effects sampled from a Gamma distribution; [2] QTL effects pre-defined to explain at least 0.5% of genetic variance (GV); [3] QTL with $r^2 > 0.7$ with at least one selected SNP; [4] GV = Genetic Variance; [5] accuracy of phenotype prediction; [6] accuracy of genomic prediction

Table 3.6. Comparison of best accuracies between BayesB, BayesC, and the proposed method under different sampling distribution for the QTL effects and density of the marker panel.

	200K marker panel		400K marker panel	
	Gamma[1]	Predefined[2]	Gamma	Predefined
Diff_acc_G[3]				
BayesB	-0.94	0.36	-3.60	-0.74
BayesC	-1.29	-1.08	-4.19	-2.71
Diff_acc_P[4]				
BayesB	-0.80	2.04	-3.56	-0.82
BayesC	-1.39	0.61	-4.15	-2.68

[1] QTL effects sampled from a Gamma distribution; [2] QTL effects pre-defined to explain at least 0.5% of genetic variance; [3] percentage difference in genomic accuracy compared to the proposed method; [4] percentage difference in phenotype prediction genomic accuracy compared to the proposed method

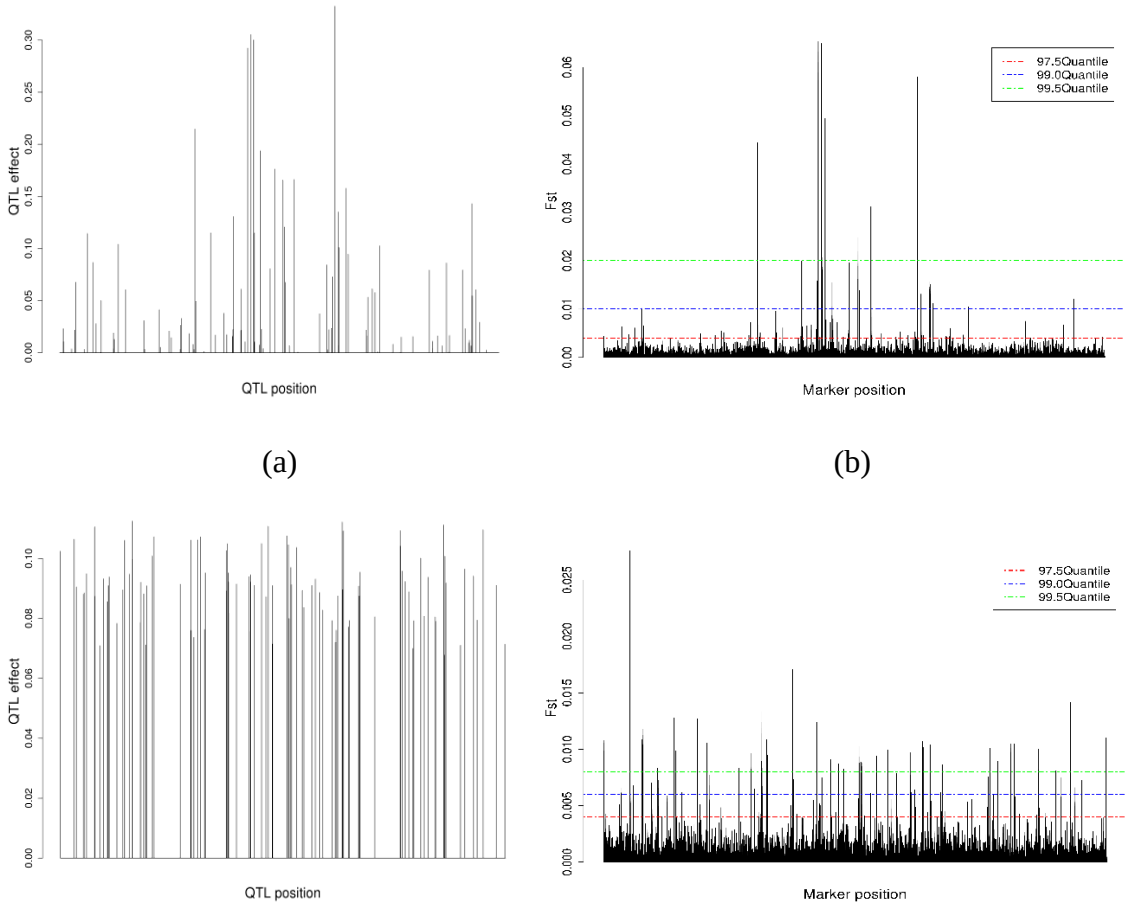
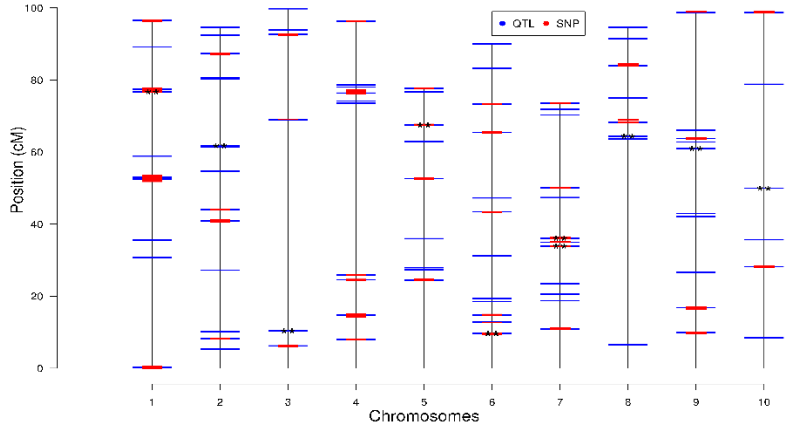
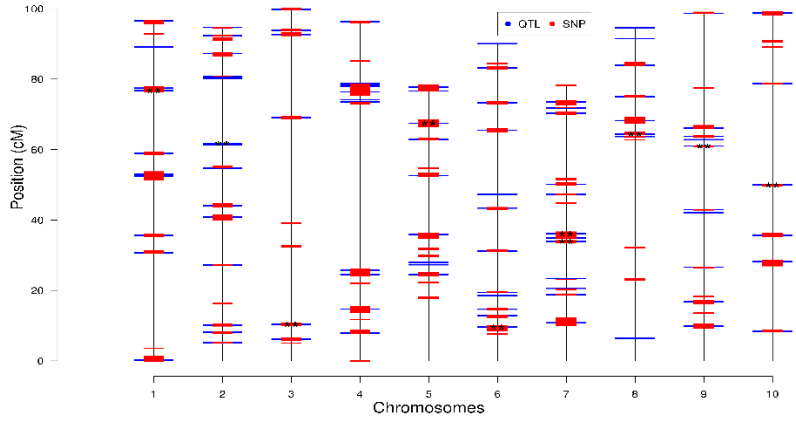


Fig. 3.1. Distribution of the simulated quantitative trait loci (QTL) along the ten chromosomes when their effects were simulated from a gamma distribution (a) or predefined (c) and their associated F_{ST} scores distribution (b) and (d) for the 200 K marker panel scenario. Horizontal dashed lines indicate the 99.5 (red), 99.0 (blue), and 97.5 (green) quantiles of the F_{ST} distribution



(a)



(b)

Fig. 3.2. Distribution of the simulated QTL (in Blue) and the preselected SNPs (in Red) across the 10 chromosomes using the 99.5 (a) and 97.5 (b) quantiles of the F_{ST} scores under the predefined QTL effect and the 200K marker panel simulation scenario. (* indicates the top 10% QTL)

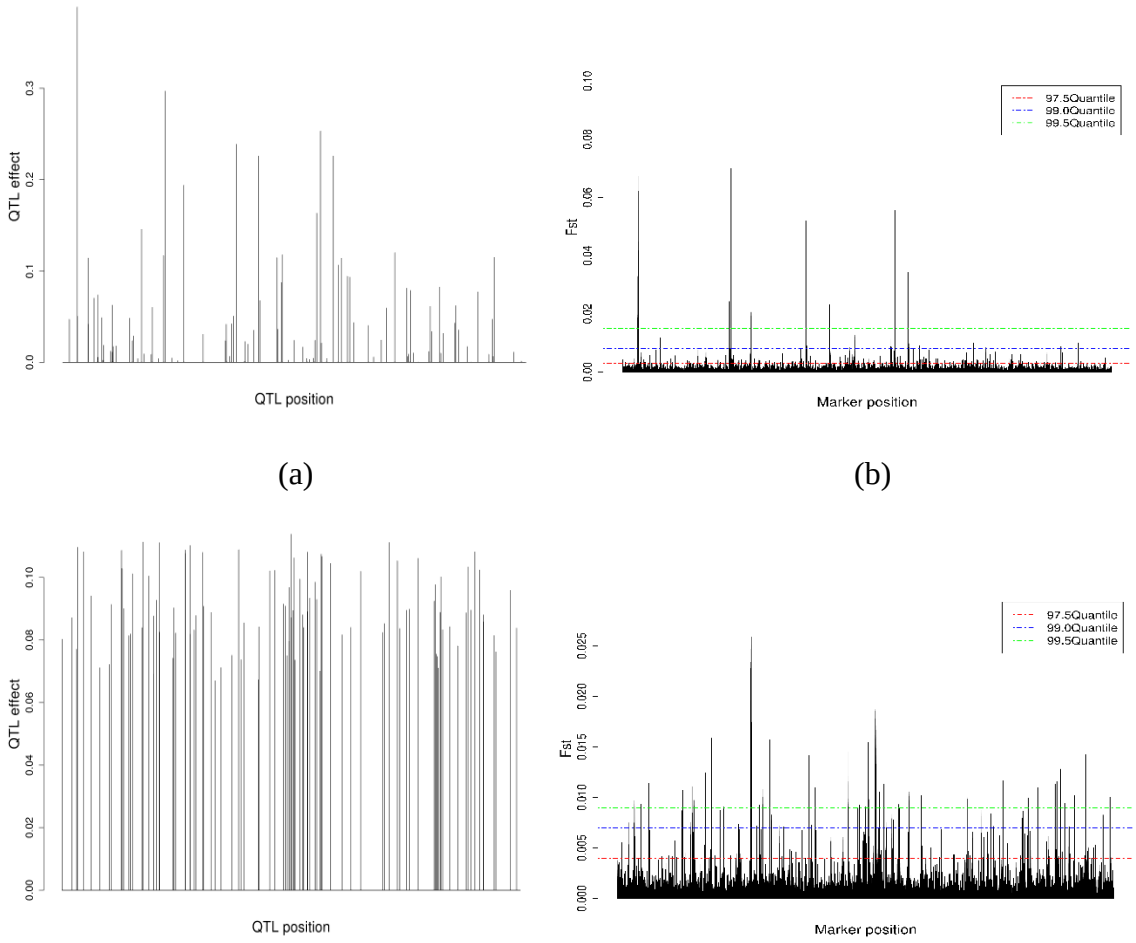
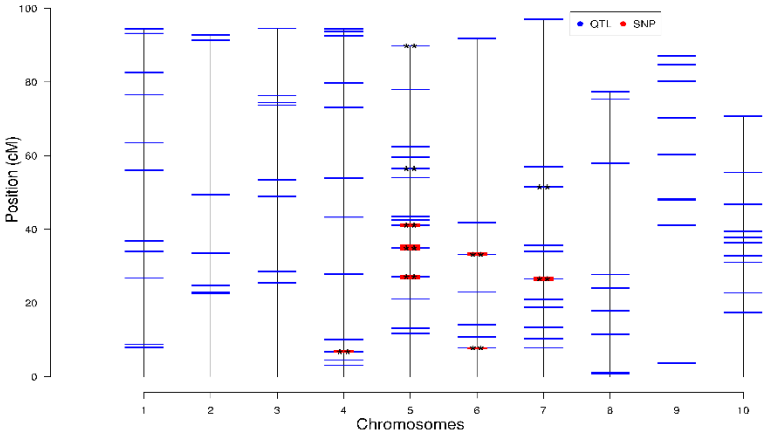
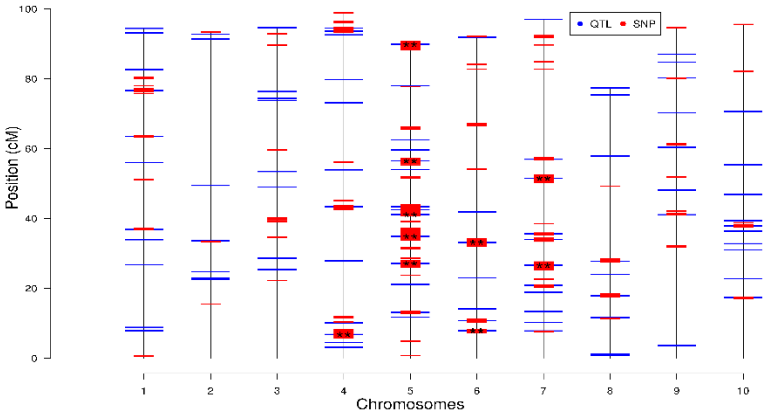


Figure 3.S1. Distribution of the simulated quantitative trait loci (QTL) along the ten chromosomes when their effects were simulated from a gamma distribution (a) or predefined (c) and their associated F_{ST} scores distribution (b) and (d) for the 400 K marker panel scenario. Horizontal dashed lines indicate the 99.5 (red), 99.0 (blue), and 97.5 (green) quantiles of the F_{ST} distribution

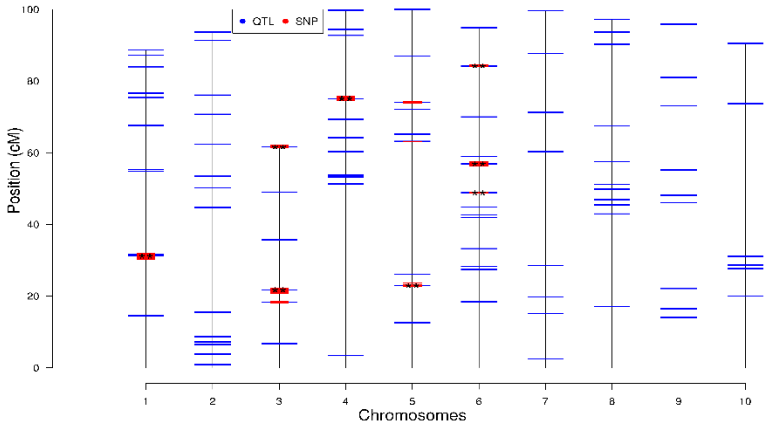


(a)

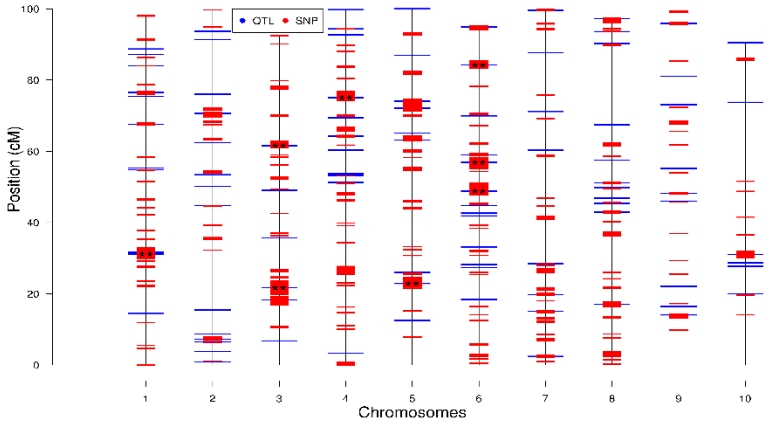


(b)

Figure 3.S2.1. Distribution of the simulated QTL (in Blue) and the preselected SNPs (in Red) across the 10 chromosomes using the 99.5 (a) and 97.5 (b) quantiles of the F_{ST} scores under the QTL effect sampled a Gamma distribution with shape parameter equal to 0.4. and the 200K marker panel simulation scenario. (* indicates the top 10% QTL)

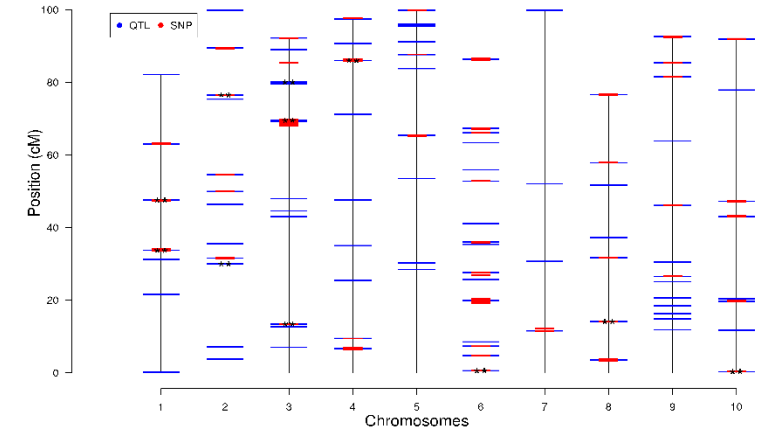


(a)

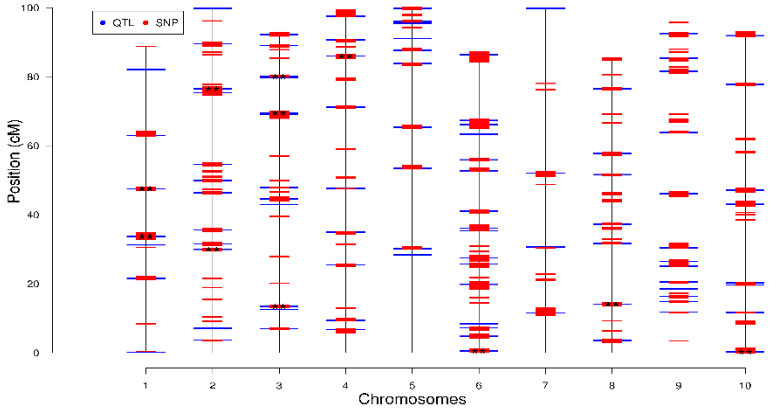


(b)

Figure 3.S2.2. Distribution of the simulated QTL (in Blue) and the preselected SNPs (in Red) across the 10 chromosomes using the 99.5 (a) and 97.5 (b) quantiles of the F_{ST} scores under the QTL effect sampled a Gamma distribution with shape parameter equal to 0.4. and the 400K marker panel simulation scenario. (* indicates the top 10% QTL)



(a)



(b)

Figure 3.S2.3. Distribution of the simulated QTL (in Blue) and the preselected SNPs (in Red) across the 10 chromosomes using the 99.5 (a) and 97.5 (b) quantiles of the F_{ST} scores under the predefined QTL effect and the 400K marker panel simulation scenario. (* indicates the top 10% QTL)

CHAPTER 4

INCREASING ACCURACY OF GENOMIC SELECTION IN PRESENCE OF HIGH DENSITY MARKER PANELS THROUGH THE PRIORITIZATION OF RELEVANT POLYMORPHISMS [1]

[1] Chang, LY, S. Toghiani, S. E. Aggrey, R. Rekaya. 2018. Submitted to the *BMC Genetics* (Under Review)

Abstract

Because of the limitations of current methods used for implementation, increase in the density of marker panels did not result in any significant increase in the accuracy of genomic selection (GS) using either regression (RM) or variance component (VC) approaches. Increasing the number of variants in the association models led to an over-parameterization, increased co-linearity and lack of statistical power. For VC based approach, increase in marker density, after a certain threshold, did not improve the genomic relationships. To remedy these problems, the fixation index (F_{ST}), a measure of genetic differentiation, was used as an additional source of information to prioritize SNPs in high-density marker panels and track the majority of significant QTL. A trait with heritability of 0.4 was simulated in two populations (P_1 and P_2) with average LD between adjacent markers of 0.3 and 0.7. Genomic data consisted of 400K SNP markers distributed on 10 chromosomes to mimic 1.2 million SNPs markers in the bovine genome. The matrix **G** was calculated for each set of selected SNPs based on their F_{ST} score and similar numbers of SNPs were selected randomly for comparison. Using all 400 K SNPs, 46% of the off-diagonal elements (OD) were between -0.01 and 0.01. The same portion was 31, 23 and 16% when 80 K, 40 K and 20 K SNP s were selected based on F_{ST} scores. For randomly selected 20K SNP subsets, around 33% of the OD fell within the same range. Genomic similarity computed using SNPs selected based on F_{ST} scores was always higher than using the same number of SNPs selected randomly. Maximum accuracies of 0.741 and 0.828 were achieved when 20 and 10 K SNPs were selected based on F_{ST} scores in P_1 and P_2, respectively. Genomic similarity could be maximized by the decrease in the number of selected SNPs, but it also leads to a decrease in the

percentage of genetic variation explained by the selected markers. Finding the balance between these two parameters could optimize the accuracy of GS in presence of high density marker panels.

Keywords: Genomic selection, High density panel, SNP prioritization

Introduction

Current high-throughput genotyping and sequencing techniques allow for the generation of dense marker maps on a large number of individuals. Polymorphic variants including single-nucleotide polymorphisms (SNPs), rare variants and indels are being massively identified and genotyped at an ever-decreasing cost. This abundant genomic information provides an unprecedented resource to identify the association between complex traits and genetic variation. Thus, this genomic information, especially SNP markers have been used to detect functional or causative variants and reveal the genetic mechanisms of complex trait in human, plants and livestock species (Goddard et al., 2009; Karlsson et al., 2007; Bennett et al., 2010; Bolormaa et al., 2010; Pryce et al., 2010).

In livestock and plants, genomic information has been extensively and successfully used for selection via the estimation of gnomically enhanced breeding values (GEBV) which are computed as linear function of the SNP effects and their associated genotypes. Due to its superiority, mainly due to shorter generation interval and higher accuracies, compared to classical pedigree-based evaluation, genomic selection (GS) is quickly becoming the standard tool for genetic evaluation in several livestock species. In spite of its undeniable success, GS is facing several challenges due to the continuous increase in the density of marker panels, the complexity of accommodating low (rare) frequency variants, and the increase in the number of genotyped individuals. Although high-throughput genotyping and sequencing techniques have substantially changed the speed and cost and discovered huge number of rare variants (MAF < 1%), some challenges are limiting the performance of analyzing high density data. The major

problem is the high dimensionality of the parameter space. Regression (RM) based approaches model directly the association between the phenotype and all or a subset of the genotyped variants. Thus, their problems stem mainly from the high dimensionality of the model. When all variants are considered (i.e., BayesA), the highly informative prior will lead to excessive shrinkage that together with the high linkage disequilibrium (LD) precludes the identification of causative mutations or even of significant tag variants. As the effect of a QTL (often small for complex traits) is distributed in a non-trivial manner between all markers that are in LD with the causal mutation, there is little statistical power to accurately estimate its effect. Given these limitations, filtering (prioritization) of variants to be included in the association models has become 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 the efficiency of these methods. Consequently, with the current density of sequence variants, it is clear that statistical discriminatory criteria alone will not be enough to prioritize influential variants, and enlistment of additional external 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 (Hayes et al., 2014). More recently, Chang et al. (2018) proposed using F_{ST}, a measure of genetic differentiation, to prioritize markers. The results showed some superiority compared to BayesB at different values for the portion of SNPs included in the association model.

Variance component (VC) based approaches, such as GBLUP (Goddard et al., 2011) or ssBLUP (Aguilar et al., 2010) are not directly affected by the increase in the number of variants in the association model. However, in their current form they are unlikely to benefit from the use of information provided by high density marker panels and NGS, and in extreme cases a decrease in performance could occur. The superiority of GBLUP compared to classical BLUP is due to the use of the observed (**G**) rather than the expected (**A**) additive relationship matrix. In real applications, the genomic relationship matrix allows for: 1) correction of potentially erroneous pedigree information, 2) detection of and accounting for unreported relationships, and 3) a better modeling of Mendelian sampling (Cole and VanRaden, 2011; Hill and Weir, 2011, Clark et al., 2013).

An increase in SNP density, after a certain threshold, seems to not affect the quality of the estimated observed relationship matrix **G**. The accuracy obtained from using the 777 K SNP panel is not any different from using the 54 K SNP panel (Su et al. 2012a, b). This is because the 777K panel did not improve the quality of **G** in any significant way. Results from the human 1000 Genomes Project indicated that the majority of variants observed in a given genome are common and only 40 to 200 K are rare (MAF <0.5%). Thus, out of the 64 million rare variants only 1 to 4% are polymorphic in a given genome and two random individuals will differ at a maximum of 80 to 400 K rare variants. If similar trends were to be observed in livestock species,

computing the genomic similarity based on common and all or the majority of rare variants will lead to inaccurate estimates of the observed additive relationships and very likely a reduction in the performance of GS.

On top of these challenges there are the added computational costs. Although the computational cost increases almost linearly with increasing number of genotyped animals in RM approaches, that is not the case with increasing number of variants. Thus, the approach will become almost impossible computationally when using sequence variant genotypes. Such costs will not be reduced even when methods for variant prioritization (BayesB, BayesR) are implemented because the detection of "relevant" variants in each round will offset the computational benefit generated by estimating the effects of only a subset of SNPs. For VC-based approaches, the number of sequence variants will have very little computational cost. However, those computational costs increase cubically with the number of genotyped animals, making direct inversion of the matrix **G** difficult even with medium size data sets. The Algorithm for proven and young animals (APY) method developed by Fragomeni et al. (2015a, b) to approximate the inverse of **G** is intrinsically data-driven and could result in computational problems trying to construct the inverse of the blended matrix **H**. As a data-driven approximation, its performance is not guaranteed with a continuous increase in the number of genotyped animals spanning several generations and more complex pedigree structures (inbreeding). Furthermore, the APY method will not benefit from the availability of high density SNP panels or NGS data.

Although prioritization methods based on statistical (i.e., BayesB), external prior information (i.e., BayesRC), and population genetics criteria (i.e., F_{ST}) have been

frequently applied in RM, little has been done to evaluate the impact of marker prioritization on the estimation of the genomic relationship matrix (**G**) and the potential impact in GS using VC approaches. The later will benefit from SNP marker prioritization for two reasons: 1) only relevant markers will be used to compute **G** removing thus the contribution of non-influential SNPs that could increase or decrease the realized genetic similarity between two individuals, especially for low MAF markers, 2) some prioritization methods (i.e., based on F_{ST}) could provide a simple and systematic approach for weighting the contribution of different markers to the estimation of **G**. For example, this could be accomplished by using the individual marker F_{ST} score as a weight factor. In this study, marker prioritization method presented by Chang et al. (2018) will be assessed for its impact on the estimation of the genetic similarity between individuals and on the accuracy of GS. For that purpose, SNP markers in high density panels will be prioritized using F_{ST} score as suggested by Chang et al. (2018).

Material and method

Simulation: population structure:

Data was simulation to mimic high-density marker panels using the QMSim simulation software (Sargolzaei and Schenkel, 2009). In the first step of the process, a randomly mated historical population was generated to initialize LD and to establish mutation-drift equilibrium and was used as a base to create two populations (P_1 and P_2) with average LD between adjacent markers of 0.3 and 0.7, respectively. Gametes were randomly sampled from both male and female gamete pools. To produce a realistic level of LD in population P_1, 300 historical generations were generated based on random

mating of an initial 8,000 animals, increasing to 15,000 animals at generation 305, decreasing to 12,000 animals at generation 1000, and then increasing to 17,000 animals at the last generation. For population P_2, the initial 8,000 animals were also simulated for 300 generations but followed by additional 5 generations with 15,000, 5 generations with 12,000, and 5 generations with 17,000 animals. In the second step, the founder population (G_0) for P_1 and P_2 was founded by 1,000 males and 15,000 females randomly selected from the historical population. A trait with heritability equal to 0.30 was generated and all genetic variation was assumed to be due to the simulated QTL. The mating system was at random throughout up to generation G_0. An additional 7 selection generations (G_1-G_7) of 15,000 animals each were simulated. Parents were chosen based on their estimated breeding values (EBVs). The replacement rate for males and females was 50 and 20%, respectively. Throughout, one progeny per mating was assumed and the sex ratio of progeny was set to 50%. The average effective population size ranged between 323 and 350 for P_1 and P_2, respectively. The sixth generation (G_6) was considered as the training population and the last generation (G_7) was used to evaluate (validation population) our proposed method.

In both populations, only animals in the training and validation populations were genotyped. Genotypes were simulated for 400,000 biallelic SNP markers uniformly-distributed along 10 chromosomes of 100 cM in length each to roughly mimic 1.2 million SNP markers in the bovine genome. Two hundred biallelic QTL were sampled from a Gamma distribution with shape parameter equal to 0. 4. No overlap between SNP markers and QTL was allowed. Additionally, QTL were assumed not to be genotyped. In general, the genotype structure for P_1 and P_2 were similar except that P_2 had higher LD

between adjacent markers. The residual variance was scaled accordantly in each scenario of selected SNPs such that the heritability and phenotypic variance were constant at the values of 0.3 and 1, respectively. Trait phenotypes were generated as the sum of an overall mean, the random additive effects of QTL and their associated genotypes, and the residual terms. The later were sampled from a normal distribution with zero mean and variance-covariance matrices $I\sigma_e^2$ where σ_e^2 is the residual variance.

Method of prioritizing SNPs: F_{ST} approach:

Wright's fixation indexes, F_{ST} in particular, have been used to measure the level of variation among subpopulation with respect to the variation in the total population. F_{ST} is a measure of genetic differentiation among groups and depends on the allele frequencies at a locus. The larger the F_{ST} values are, the greater is the genetic differentiation. In this study, F_{ST} scores were calculated following the estimators presented by Nei (1973) and Chang et al. (2018). Based on the distribution of the trait phenotypes, generation 6 (G_6) was divided into three sub-populations (below the 5% quantile [S1], between 5 and 95% quantiles [S0], and above the 95% quantile [S2]). Genotypes of individuals (1,500) in sub-populations S1 and S2 were used to calculate the F_{ST} scores. For each locus, the global F_{ST} estimator was defined as:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

with $H_T = 2 * p * q$, $H_S = \frac{H_{S1} * n_{s1} + H_{S2} * n_{s2}}{n_{s1} + n_{s2}}$, and $H_{Si} = 2 * p_{Si} * q_{Si}$

where, p_{Si} and q_{Si} are the allele frequencies in subpopulation i, n_{s1} and n_{s2} are the number of individuals of the subpopulations, H_S is the average of sub-population heterozygosities and H_T is the heterozygosity based on the total population.

Genetic similarity:

Historically, genetic similarity between individual is measured by their average expected additive relationships derived from pedigrees. With the availability of genetic markers, SNP panels with reasonable density provide an alternative tool to estimate genetic similarity based on realized relationships or other measurements. Currently, genomic relationships are calculated based on identity by state (IBS) between alleles of SNP markers (VanRaden, 2008). It basically measures the similarity of marker genotypes between two individuals at a large number of loci independently of their mode of inheritance. Although estimated realized relationships using IBS are in general better than their counterparts obtained using pedigree information, they still suffer from several problems including the non-zero estimates of realized relationship between two individuals that are not related by ancestry as it was shown by (Li et al., 1993; Blouin et al., 1996; Csilléry et al., 2006) and the inevitable noisy associated with these estimates. More importantly though is that as the SNP marker density increases, after a certain threshold, it seems not to affect the quality of the estimated observed relationships. The accuracy obtained from using the 777 K SNP panel is not any different from using the 54 K SNP panel (Su et al. 2012b). This is because the 777 K panel did not improve the quality of realized genomic relationships in any significant way. Thus, in presence of high density marker data, using all SNPs to estimate genetic similarity will not improve

the genomic relationships. To the contrary, it could lead to less accurate estimates of genetic similarity. This clearly indicates that true genetic relationships could be accurately estimated by a reasonably small number of well distributed SNP markers. From genomic association and selection perspectives, the lack of improvement using high density panels is not due to the lack of useful information in the additional marker genotypes rather due to the limitations of current methods used. One way to better use this high-density data is to maximize the functional genomic similarity between individuals rather than simply to estimate the average or observed additive relationships. This functional similarity will likely be higher than the additive relationships if it is calculated based on a selected subset of SNPs prioritized based on their ability to increase genetic or phenotypic similarity between individuals. As the marker density increases, especially in presence of SNPs with low minor allele frequency (closer to 0.05 cut-off threshold value), prioritization of SNP markers to be included in the calculation of the genomic similarity becomes more relevant. This is the case because as the number of SNPs increases, the genomic relationships inch closer to the expected relationships. It is reasonable to expect that individuals with similar genetic values or even phenotypes 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 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:

$$sim\ (i,j) = \frac{1}{2n}\sum_{k=1}^{n} S_k(i,j) \qquad [1]$$

where $S_k(i,j)$ is the number of shared alleles between individuals *i* and *j* at locus *k*. Genetic similarity was computed based on all SNPs in the panel and subsets of 2.5, 5, 10, 20, 40, 80 and 160 K markers selected either based on FST scores or at random.

Statistical model and data analysis:

For both simulated populations (P_1 and P_2), 10,000 and 5,000 animals were randomly selected from G6 and G7, respectively. For each population, several data sets with different number of SNPs (from 10 to 400 K) selected either using F_{ST} scores or at random were generated. Data was analyzed using the following mixed linear model:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e}$$

Where $\boldsymbol{y}$ is the vector of phenotypes, $\boldsymbol{b}$ is the vector of fixed effects, $\boldsymbol{u}$ is the vector of genomic breeding values, and $\boldsymbol{e}$ is the vector of random residuals. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known incidence matrices with the appropriate dimensions. Additionally, it was assumed that $\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$ with **G** is the genomic relationship matrix and σ_u^2 is the genetic variance.

AIREMLF90 program, a modification of restricted maximum likelihood (REML) approach with the Average-Information algorithm (Misztal et al., 2002), was used to estimate variance components and genomic breeding values under the different scenarios. Accuracy of genomic evaluation was defined as the correlation between true breeding value and the genomic estimated breeding value in validation population. In this study, each simulation scenario was replicated 5 times.

Result and discussion

Distribution of QTL and estimated F_{ST} values:

The efficiency of a marker prioritization method depends on its ability to track all the QTL controlling the trait and at least it should track the most influential ones. Figure 4.1 presents the distribution and effects of the simulated 200 QTL (Figure 4.1a) and the F_{ST} scores of the 400 K SNPs (Fig. 4.1b) for the scenario when the LD between adjacent markers was equal to 0.3 (Figure 4.S1 presents the results for populations P_2). As in Toghiani et al. (2016) and Chang et al. (2018), there is a striking similarity between the distribution of QTL effects and F_{ST} scores. In fact, there is an almost perfect overlap between the peaks in Figure 4.1a (QTL with large effects) and Figure 4.1b (SNPs with large F_{ST} scores). Such overlap persists even for QTL with moderate to small effects indicating the ability of F_{ST} scores to track the distribution and effects of the majority of simulated QTL. Obviously, the ability to track QTL using F_{ST} scores depends primarily on the heritability of trait, the genetic variance explained by the QTL, the population structure, LD between markers and QTL and among markers. For population P_1, 55%, 30% and 15% of QTL explained less than 0.1%, between 1 and 0.1 and greater than 1% of genetic variance each. Similar percentages were observed for population P_2. Although this distribution of QTL effects in unlikely in human populations, it is not unexpected in highly selected plant and animal populations. Figure 4.2 presents the distribution of simulated QTL across the 10 chromosomes and the top 10 K (Fig. 2a) and 5 K (Fig. 2b) SNPs selected based on their F_{ST} score for population P_2 (LD=0.7). It is clear that the majority of QTL are tracked by more than one SNP and only QTL with very small effects

(< 0.01% of genetic variance) were not effectively tracked (i.e., first QTL in the lower end of chromosome 4). In fact, over 85% and 78% of the genetic variance was tracked by the 10 K and 5 K preselected SNPs, respectively.

Dissection of Genomic relationship matrix and genetic similarity:

Table 4.1 presents estimates of functional genomic similarity based on different number of selected SNPs. Under the random scenario, genomic similarity was the same across the different SNP densities and when all 400 K markers were used. This is in line with the limited improvement in the estimation of the genomic relationships with the increase of marker density (VanRaden et al., 2011; Erbe et al., 2012; Su et al., 2012b). However, when SNPs were prioritized based on F_{ST} scores, functional similarity increased with the decrease in the number of selected markers and was higher than its counterpart in the random selection scenario. In fact, prioritization based on F_{ST} scores resulted in 0.5 to 1.5% increase in genetic similarity across the different marker densities (Table 4.1). Off diagonal elements of the genomic relationships matrix also capture the magnitude of the genetic similarity between individuals. When constructed from a sufficiently large number of SNP markers, the **G** matrix is likely to better reflect the real additive relationships between individuals compared to a pedigree-based kinship matrix (**A**). Although the contribution of a SNP marker to the estimation of **G** is intrinsically weighted by its MAF, unfortunately it is not weighed by the magnitude of its effect. Thus, after a certain threshold on the number of SNP markers is reached, little to no improvement is expected in **G** and ultimately in the performance of the association model with additional markers. The limited change in **G** with additional markers could be an

indicator of the sufficiency of available SNPs in estimating the realized relationships. However, such sufficiency is not a guaranty of the optimality of such matrix for the implementation of association analyses.

In fact, as the number of randomly selected SNPs increased from 40 to 400 K, the matrix **G** gets closer and closer to the expected additivity relationship matrix (**A**) as indicated in Table 4.2. The matrix **G** computed based on a selected subset of 20 K markers is markedly different from **A**, especially on the tails of the distribution of off-diagonal elements indicating higher genetic similarity between individuals (Figure 4.3). More importantly, larger genomic similarities between training and validation individuals were observed when subsets of SNP markers were selected based on F_{ST} scores (Table 4.2). In fact, the portion of genomic relationships between training and validation individuals exceeding 0.05 ranged between 0.50 and 3.83% when all 400 K or random subsets (80 K, 40 K and 20 K) of SNPs were used. The same portion was 4.98, 14.55 and 30.75% when 80 K, 40 K and 20 K SNP s were selected based on F_{ST} scores (Table 4.2). Figure 4.4 presents a heat map of genomic similarity between training and validation individuals. It is clear that those genomic relationships are higher using the same number of SNPs selected based on F_{ST} scores than at random and they are markedly larger when 40 K or less SNPs are prioritized. It is worth mentioning that in order to increase the number of selected SNPs, the cutoff point for the F_{ST} scores has to be decreased. Consequently, SNPs with non-noticeable selection pressure get selected leading a decrease in performance of the prioritization method as it is the case when 80 K SNPs were preselected (Table 4.1 and Figure 4.4).

Variance components and accuracy of estimated breeding value:

Table 4.3 presents the estimates of the variance components and their associated standard deviations as a function of the number and mode of preselection of SNP markers used to compute the genomic relationship matrix. As expected, the percentage of the genetic variance recovered increased with the increase in the number of SNPs used to compute **G** for both populations P_1 and P_2. When the LD between adjacent markers was equal to 0.3 (population P_1), less than half of the genetic variance was recovered when **G** was estimated based on 2.5 K selected either randomly or using F_{ST} scores. Such percentage increased steadily to reach a maximum when all 400 K SNP markers were used at which point over 83% of the genetic variance was recovered. The inability to recover all the genetic variance in this case is due to the large number of QTL with very small effects. In fact, 55% of QTL have a true effect smaller than one tenth of one percent and an additional 20% of QTL have an effect smaller 0.5% of the total genetic variance. These small QTL are hard track effectively when the LD is moderate to low. Although the general trend was similar when LD was set equal to 0.7 (population P_2), the percentage of genetic variance explained at a given number of SNPs was in general higher than in P_1 (Table 4.3). This is especially the case for the random selection scenario and when the number of SNPs used to estimate **G** was small for the F_{ST} scores based selection approach. Estimates of the residual variance were almost identical to the true value (0.7) when all 400 K SNPs were used to compute **G**. For the random selection scenario, there was an over-estimation of the residual variance, except for the case when 160 K SNP were used (Table 4.3). This is large due to under estimation of the genetic variance. When SNPs were prioritized based on their F_{ST} scores, the residual variance is

over estimated when the number of marker used to calculate **G** was small (< 5 K) and under-estimated when the number of markers exceed 40 K. In between these two numbers of selected SNPs, the residual variance is precisely estimated (Table 4.3).

Using all SNPs in the 400K panel resulted in genomic accuracy of 0.716 and 0.760 for P_1 and P_2, respectively (Table 4.4). Genomic selection relies on the assumption that QTL are in LD at least with one of the SNPs in the panel. Thus, the higher accuracy in P_2 is due to the increase in LD between adjacent SNP markers and ultimately between markers and QTL. Across all random subsets (2.5 to 160 K SNPs), accuracy increased with the increase of the number of selected SNPs under both 0.3 and 0.7 LD scenarios. Further, accuracy was always smaller than when all 400 K SNPs were used (Table 4.4). When SNPs were prioritized based on their F_{ST} scores, accuracy ranged between 0.723 to 0.741 and 0.784 to 0.828 for P_1 and P_2, respectively (Table 4.4). However, accuracy did not increase continuously with the increase in the number of SNPs as it was the case in the random scenario. Accuracy reached a maximum of 0.741 and 0.828 at around 20 and 10 K selected SNPs for P_1 and P_2, respectively. This intermediate optimum behavior of the accuracy seems to be the result of a balancing act between the percentage of the genetic variance explained by the selected SNPs and the resulting genetic similarity between individual based on those markers. Increase in the number of prioritized SNPs will increase the percentage of the captured genetic variance (Table 4.3) and will ultimately results in a higher accuracy. However, such increase in the number of selected SNPs will reduce the genetic similarity between individuals in the training and validation sets (Table 4.1) which will lead to reduction of accuracy. At some point, the benefits resulting for the increase in the percentage of captured genetic variance will not offset the

cost (loss of accuracy) due to the reduction in genetic similarity. This behavior does not occur in the random selection scenario due to the little change in the genetic similarity will the increase of the number of SNPs (Table 4.1). Thus, accuracy is largely under the control of the percentage of captured genetic variance.

Conclusions

High-density SNP panels and whole genome sequence data were expected to increase the accuracy of genomic selection in livestock. However, because of the limitations of current methods used for implementation, increase in genomic data did not result in any significant improvement of accuracy. The dramatic increase in the dimensionality of the association models led to an over-parameterization problem, such as increased co-linearity and lack of statistical power. F_{ST}, a measure of genetic differentiation, was used as an additional source of information to prioritize SNPs in high-density marker panels. Prioritized markers based on F_{ST} under different scenarios were able to track the majority of significant QTL and to increase the functional genetic similarity between individuals. The later could be maximized by the decrease in the number of selected SNPs. Unfortunately, that will lead to a reduction in the percentage of genetic variation explained by the selected markers. Thus, a balance between these two parameters is needed in order to maximize the accuracy of GS in presence of high density marker panels. This balance is likely to depend on the heritability of the trait and its genetic complexity. However, given the simplicity and flexibility of marker prioritization using F_{ST} the balance could be easily identified empirically. Finally, high density

genomic data is relevant for GWAS and GS and should not be dismissed simple because it did improve accuracies based on using current methods.

Reference

Aguilar I, Misztal I, Johnson D, Legarra A, Tsuruta S, Lawlor T. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010; 93(2):743–52

Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, Neubauer M, Neuhaus I, Yordanova R, Guan B, et al. A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res. 2010; 20(2):281–90.

Blouin MS, Parsons M, Lacaille V, Lotz S: Use of microsatellite loci to classify individuals by relatedness. Mol Ecol 1996; 5(3):393–401.

Bolormaa S, Pryce JE, Hayes BJ, Goddard ME. Multivariate analysis of a genome-wide association study in dairy cattle. J Dairy Sci. 2010; 93:3818–33

Chang LY, Toghiani S, Aggrey SE, Rekaya R. High density marker panels, SNPs prioritizing and accuracy of genomic selection. BMC Genetics 2018; 19:4.

Clark SA, Kinghorn BP, Hickey JM, van der Werf JHJ. The effect of genomic information on optimal contribution selection in livestock breeding programs. Genet Sel Evol. 2013; 45:44.

Cole JB, and VanRaden PM. Use of haplotypes to estimate Mendelian sampling effects and selection limits. J. Anim. Breed. Genet. 2011; 128: 446–455.

Csilléry K, Johnson T, Beraldi D, Clutton-Brock T, Coltman D, Hansson B, Spong G, Pemberton JM: Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. Genetics 2006, 173(4):2091–2101.

Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, Mason B, Goddard M. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012; 95(7):4114–29.

Fragomeni BO, Lourenco DAL, Tsuruta S, Masuda Y, Aguilar I, Legarra A, Lawlor TJ, Misztal I. Use of genomic recursions in single-step genomic BLUP with a large number of genotypes. J. Dairy Sci. 2015a; 98:4090-4094.

Fragomeni BO, Lourenco DAL, Tsuruta S, Masuda Y, Aguilar I, Misztal I. Use of genomic recursions and Algorithm for Proven and Young animals for single-step genomic BLUP analyses — A simulation study. J. Anim. Breed. Genet. 2015b; 132:340-345.

Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011; 128(6):409–21.

Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009; 10:381–391.

Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009; 10:381–391.

Hayes BJ, Macleod I, Daetwyler HD, Bowman PJ, Chamberlian A, Vander Jagt C, Capitan A, Pausch H, Stothard P, Liao X. Genomic prediction from whole

genome sequence in livestock: the 1000 Bull Genomes Project. Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, 17-22 August 2014, Vancouver, BC, Canada, pp. 1–6.

Hill W, and Weir B. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 2011; 93: 47–64.

Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet. 2007; 39:1321–8.

Li CC, Weeks DE and Chakravarti A. Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered. 1993; 43:45–52.

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016; 17:144.

Meuwissen THE, Hayes BJ, and Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001; 157: 1819-1829.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee D. "BLUPF90 and related programs (BGF90)". Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, 19-23 August 2002, Montpellier, France, 28–27.

Nei M. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences of the USA. 1973; 70:3321–3.

Pryce JE, Bolormaa S, Chamberlain AJ, Bowman PJ, Savin K, Goddard ME, Hayes BJ. A validated genome-wide association study in 2 dairy cattle breeds for milk

production and fertility traits using variable length haplotypes. J Dairy Sci. 2010; 93(7):3331–45.

Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009; 25(5):680–1.

Su G, Madsen P, Nielsen US, Mäntysaari E a, Aamand GP, Christensen OF, et al. Genomic prediction for Nordic red cattle using one-step and selection index blending. J Dairy Sci 2012a; 95:909-917.

Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. 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. J Dairy Sci. 2012b; 95(8):4657–65.

Toghiani S, Aggrey SE, Rekaya R. Multi-generational imputation of single nucleotide polymorphism marker genotypes and accuracy of genomic selection. Animal. 2016; 10:1077–85.

VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008; 91:4414–23.

VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA. Genomic evaluations with many more genotypes. Genet Sel Evol. 2011; 43:10.

Table 4.1. Functional genomic similarity under different subsets of F_{ST} based and randomly selected SNPs for the scenario when LD[1] between adjacent markers was equal to 0.7

	Genomic similarity	
SNPs	F_{ST} based	Random
2.5K	0.7013	0.6695
5K	0.6862	0.6687
10K	0.6752	0.6682
20K	0.6718	0.6678
40K	0.6712	0.6675
80K	0.6708	0.6673
160K	0.6705	0.6672
400K	0.6671	0.6671

[1] LD= linkage disequilibrium

Table 4.2. Distribution of off-diagonal elements (OD) of the genomic relationships matrix corresponding to the training and validation individuals under different selection criteria of SNP markers (in %)

	20 K SNPs		40 K SNPs		80 K SNPs		400K SNPs	Pedigree
	S[1]	R[2]	S	R	S	R	-	-
OD < -0.05	15.47	1.79	7.30	1.64	2.42	0.66	0.11	0
-0.05<OD<-0.03	11.71	8.80	11.97	8.56	9.54	6.35	3.30	0
-0.03<OD<-0.01	14.96	23.79	19.60	23.93	23.16	24.72	24.43	0
-0.01<OD<0.01	16.19	32.57	22.98	33.1	30.78	37.96	45.91	60.09
0.01<OD<0.03	14.85	22.75	19.98	22.85	22.46	23.39	22.72	32.55
0.03<OD<0.05	11.54	8.26	11.62	8.02	9.09	5.96	3.15	5.25
ODE>0.05	15.28	2.04	7.25	1.90	2.56	0.95	0.39	2.11

[1] SNPs selected based on F_{ST} scores; [2] SNPs randomly selected

Table 4.3. Variance component estimates (standard deviation) under different subsets of F_{ST} based and randomly selected SNPs for populations[1] P_1 and P_2 (average over 5 replicates)

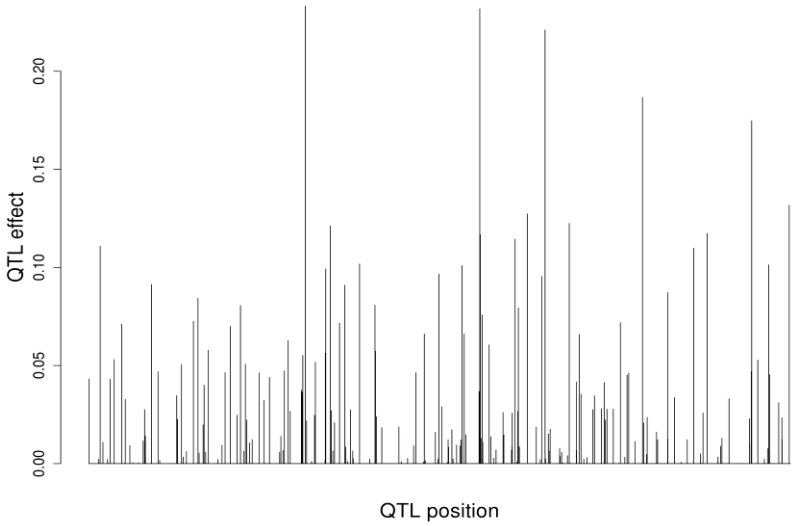
	P_1 (LD =0.3)		P_2 (LD = 0.7)	
	GV[2]	RV[3]	GV	RV
F_{ST} based				
2.5K	0.126 (0.017)	0.728 (0.027)	0.198 (0.029)	0.736 (0.006)
5K	0.149 (0.016)	0.706 (0.030)	0.204 (0.005)	0.711 (0.001)
10K	0.175 (0.023)	0.684 (0.037)	0.195 (0.009)	0.697 (0.004)
20K	0.203 (0.031)	0.663 (0.044)	0.195 (0.007)	0.686 (0.007)
40K	0.226 (0.041)	0.649 (0.052)	0.203 (0.009)	0.677 (0.007)
80K	0.247 (0.048)	0.641 (0.055)	0.217 (0.008)	0.671 (0.008)
160K	0.264 (0.045)	0.642 (0.047)	0.235 (0.008)	0.670 (0.008)
Random				
2.5K	0.104 (0.013)	0.834 (0.012)	0.155 (0.012)	0.788 (0.006)
5K	0.139 (0.016)	0.796 (0.013)	0.185 (0.013)	0.757 (0.005)
10K	0.173 (0.019)	0.762 (0.006)	0.215 (0.011)	0.730 (0.012)
20K	0.203 (0.023)	0.733 (0.013)	0.234 (0.010)	0.712 (0.008)
40K	0.227 (0.026)	0.710 (0.015)	0.242 (0.007)	0.703 (0.005)
80K	0.238 (0.027)	0.770 (0.015)	0.246 (0.008)	0.699 (0.007)
160K	0.242 (0.027)	0.696 (0.016)	0.250 (0.008)	0.696 (0.006)
Full panel				
400K	0.247 (0.027)	0.692 (0.016)	0.251 (0.007)	0.695 (0.006)

[1] P_1: 200 QTLs and linkage disequilibrium (LD) between adjacent markers equal to 0.3 and P_2: 200 QTLs and LD between adjacent markers equal to 0.7; [2] genetic variance; [3] residual variance

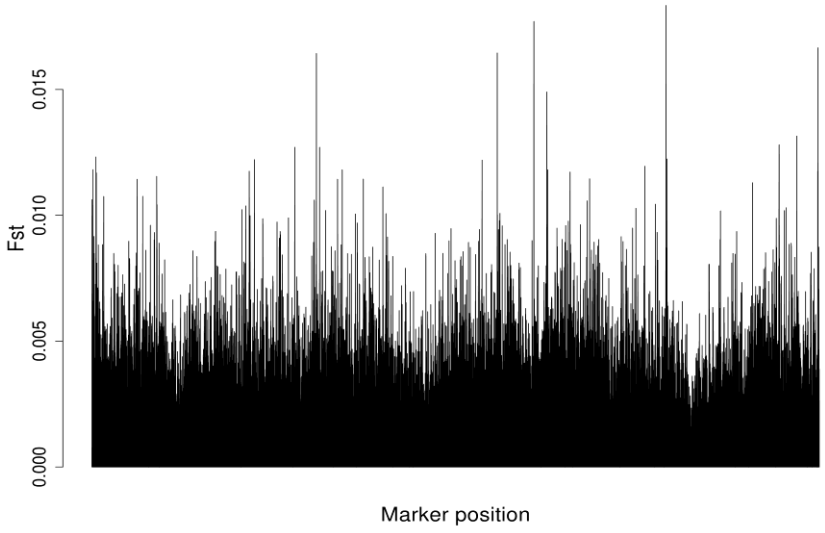
Table 4.4. Accuracy of genomic prediction (standard deviation) under different subsets of F_{ST} based and randomly selected SNPs for populations[1] P_1 and P_2 (average over 5 replicates)

	Accuracy[2]	
	P_1 (LD = 0.3)	P_2 (LD = 0.7)
F_{ST} based		
2.5K	0.724 (0.021)	0.805 (0.014)
5K	0.736 (0.022)	0.823 (0.012)
10K	0.740 (0.023)	0.828 (0.013)
20K	0.741 (0.027)	0.824 (0.013)
40K	0.735 (0.027)	0.815 (0.014)
80K	0.728 (0.028)	0.802 (0.012)
160K	0.723 (0.031)	0.784 (0.013)
Random		
2.5K	0.600 (0.054)	0.669 (0.019)
5K	0.640 (0.047)	0.709 (0.015)
10K	0.676 (0.036)	0.736 (0.019)
20K	0.695 (0.370)	0.746 (0.014)
40K	0.707 (0.034)	0.754 (0.010)
80K	0.712 (0.033)	0.757 (0.013)
160K	0.715 (0.031)	0.759 (0.011)
Full panel		
400K	0.716 (0.032)	0.760 (0.011)

[1] P_1: 200 QTLs and linkage disequilibrium (LD) between adjacent markers equal to 0.3 and P_2: 200 QTLs and LD between adjacent markers equal to 0.7; [2] correlation between true and predicted breeding values

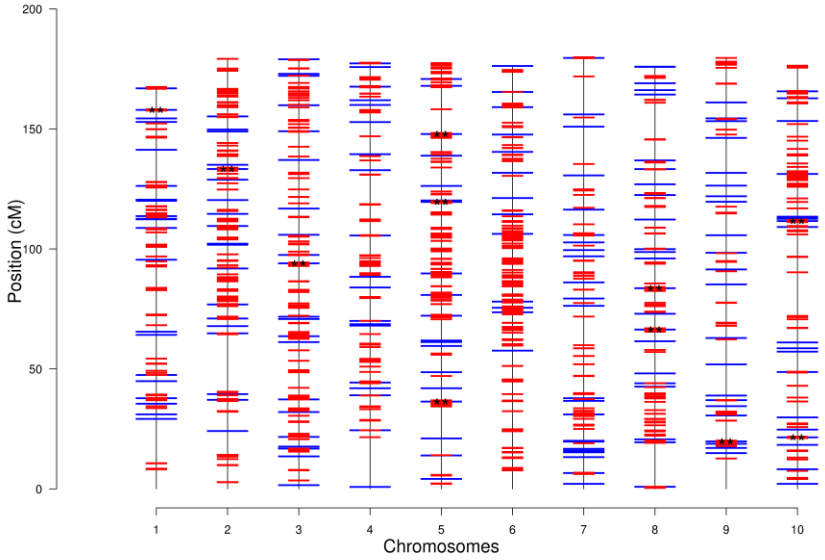


(a)

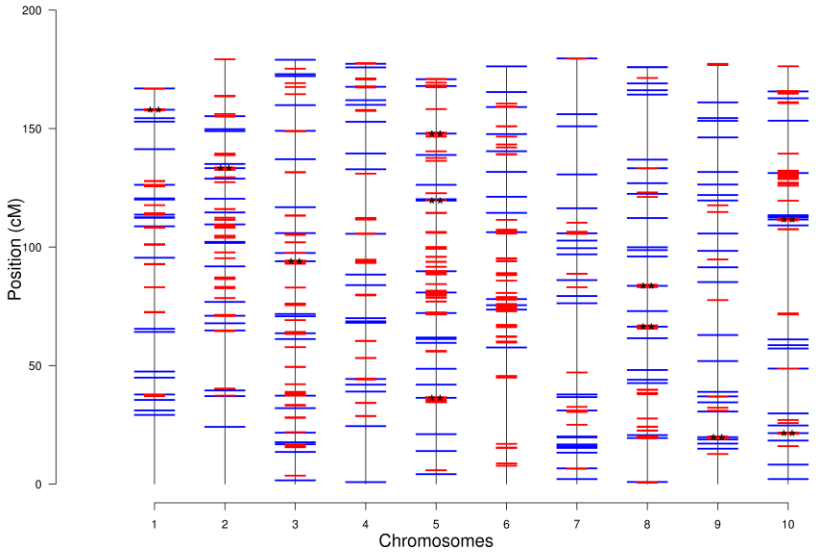


(b)

Figure 4.1. Effects and distribution of the 200 simulated quantitative trait loci (QTL) along the ten chromosomes (a) and their associated F_{ST} scores distribution (b) when the LD between adjacent markers was equal to 0.3.



(a)



(b)

Figure 4.2. Distribution of the 200 simulated QTL (in blue) and 10 K (a) and 5K (b) preselected SNPs based on F_{ST} scores (in red) across the 10 chromosomes when LD between adjacent markers was equal to 0.7 (* indicates the top 10% QTL)

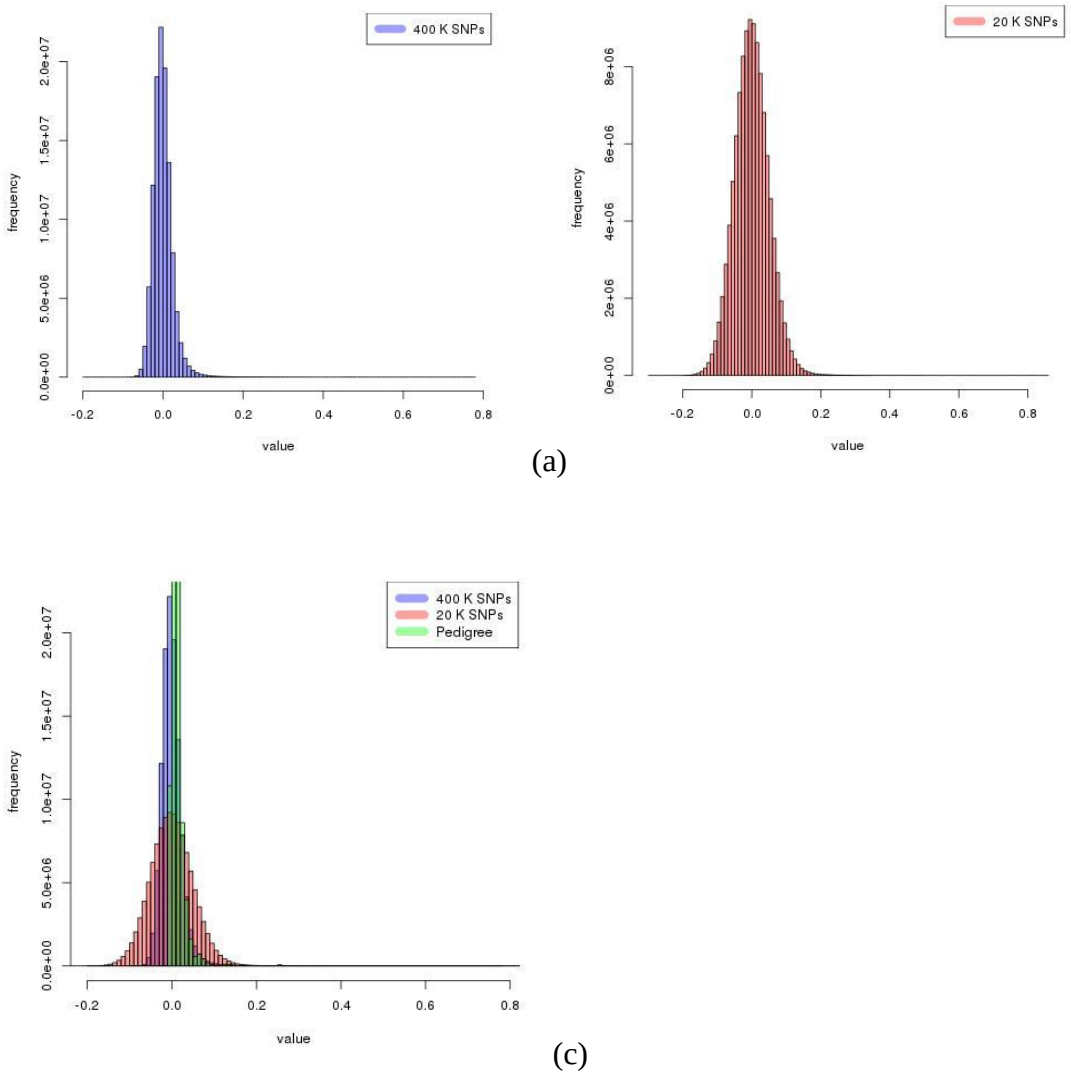


Figure 4.3. Distribution of off-diagonal elements of the additive relationship matrix using a) all 400K SNP markers (in blue), b) 20 K SNPs prioritized based on their F_{ST} scores (in red), and c) pedigree information (in green)

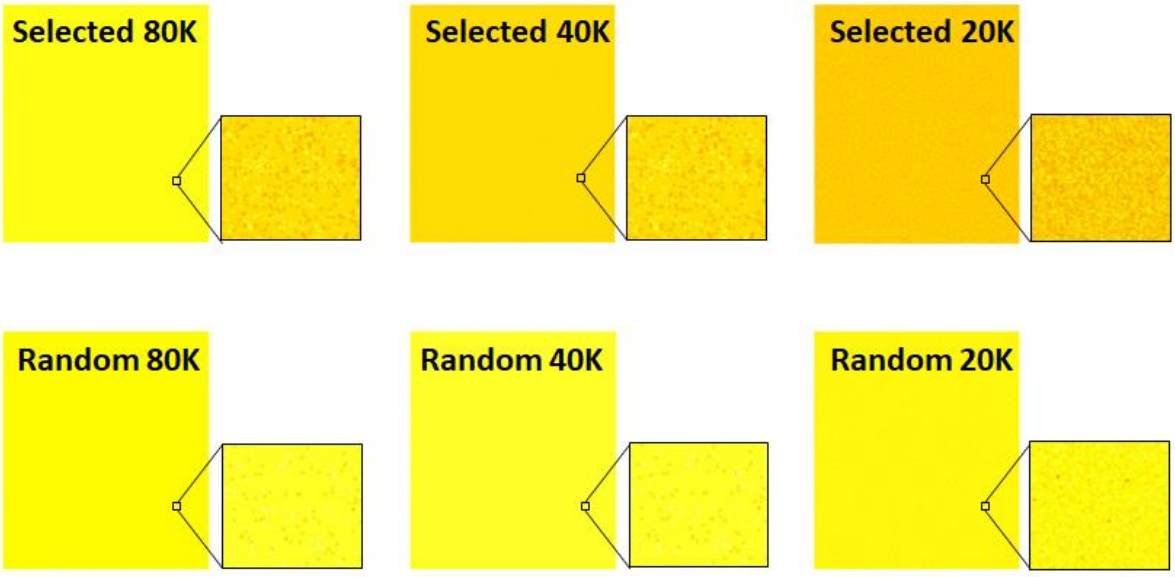
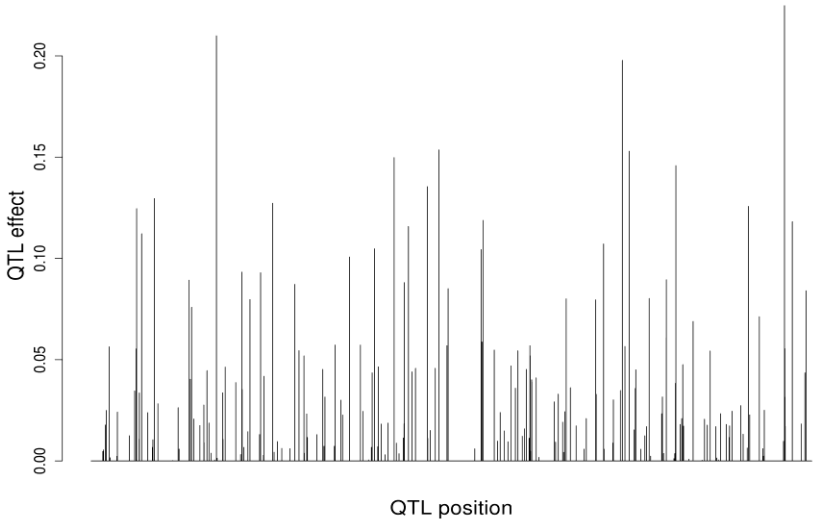
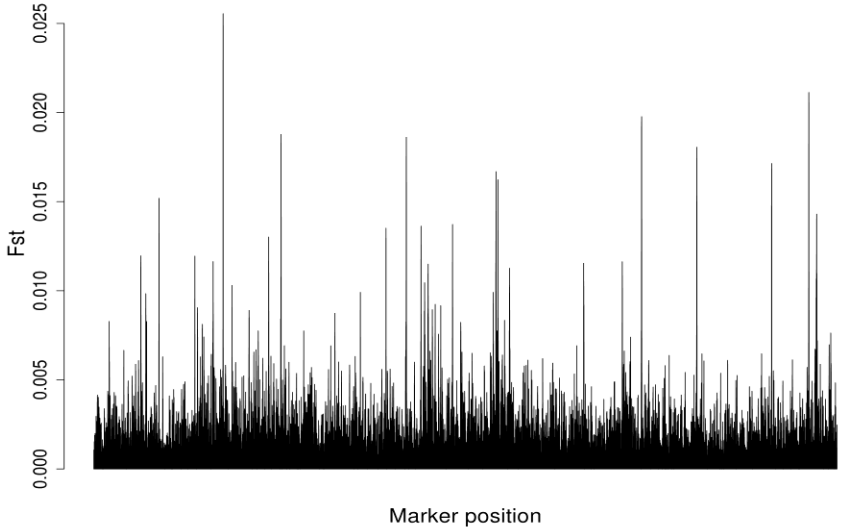


Figure 4.4. Heat map representing of the genomic similarity between training and validation individuals based on different subsets of SNPs selected based on FST score or randomly when LD between adjacent markers was equal to 0.3 (the darker the color the higher is the similarity)



(a)



(b)

Figure 4.S1. Effects and distribution of the 200 simulated quantitative trait loci (QTL) along the ten chromosomes (a) and their associated F_{ST} scores distribution (b) when the LD between adjacent markers was equal to 0.7.

CHAPTER 5

A WEIGHTED GENOMIC RELATIONSHIP MATRIX BASED ON FST PRIORITIZED SNPS FOR GENOMIC SELECTION [1]

[1] Chang, LY, S. Toghiani, S. E. Aggrey, R. Rekaya. 2018. To be submitted to the *Journal of Animal Science*

Abstract

An increasing number of animals are being sequenced or genotyped using high density marker panels with more than a quarter million SNP. The dramatic increase in the number of variants was expected to significantly increase the accuracy of genome wide association studies (GWAS) and genomic selection. Unfortunately, little to no improvement was observed. It is unlikely that this lack of improvement is the result of the limited usefulness of sequence data, but rather because of the limitations of current methods used in GWAS and genomic selection. Including all the variants in the association model will dramatically increase the dimensionality of the problem and it will undoubtedly reduce the statistical power resulting in a worst model. Using all the SNPs to compute the genomic relationship matrix will not increase accuracy as the additive relationships between individuals can be accurately estimated using a much smaller number of SNP markers. Due to these limitations, variant prioritization has become a necessity to improve accuracy. F_{ST} as a measure population differential has been successfully used to identify genome segments and variants under selection pressure. Using these variants has increased accuracy. Additionally, F_{ST} scores provide an easy way to weight the relative contribution of prioritized SNPs in the calculation of the genomic relationship matrix. In this study, F_{ST} scores based relative weights for SNPs included in the computation of **G** were developed and assessed on their impact on the estimation of variance components and the accuracy of genomic selection. The results of this study clearly show that prioritizing SNP markers based on their F_{ST} score and using the latter to compute relative weights has increased the genetic similarity between

training and validations animals. Furthermore, it resulted in more than 5% improvement in accuracy.

Keywords: High density, sequence data, genomic selection, accuracy

Introduction

Recent advances in High-throughput genotyping and sequencing techniques led to the generation of dense marker panels and facilitated the genotyping large numbers of individuals. Because of the availability of these cost-effective genotyping technologies and the increase in sequencing speed, large-scale genotyping for single-nucleotide polymorphisms (SNP) becomes more affordable and accessible. This genomic data provides an unprecedented opportunity to dissect the genetic basis of complex traits and to identify relevant functional associations.

From animal breeding perspective, the use of genomic information allows for a substantial reduction in generation interval and in the increase of the accuracy of predicted breeding values; leading undoubtedly to an improvement in the genetic response (VanRaden et al., 2009; Su et al., 2010; Su et al., 2012a; Schefers and Weigel, 2012; Zeng et al., 2013). Genomic selection is often carried out using multiple regression or mixed linear models. For both methods, the density of the SNP marker panel and the LD structure between markers and QTL have a great impact on accuracy. Regression based approaches directly model the association between the phenotypes and all or a subset of the genotyped variants. Thus, their problems stem mainly from the high dimensionality of the parameter space. When all variants are considered (e.g., BayesA), the highly informative prior will lead to excessive shrinkage that together with the high LD precludes the identification of causative mutations or even of significant tag variants. As the effect of a QTL (often small for complex traits) is distributed in a non-trivial manner between all markers that are in LD with the causal mutation, there is little statistical power to accurately estimate its effect. 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 (Meussoissan 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 the efficiency of these methods. Increase in SNP density, after a certain threshold, seems to not affect the quality of the estimated observed relationship matrix **G** and thus the performance of mixed linear model based approaches. In fact, the accuracy obtained from using the 777K SNP panel is not any different from using the 54K SNP panel (Su et al. 2012b). This is because the 777K panel did not improve the quality of **G** in any significant way.

Due to these limitations, prioritization of variants to be included in the association model or to compute the genomic relationship matrix has become a necessity. Commercial livestock species are under heavy artificial selection. The effects of such selection on the genome can be traced through the changes in allele frequencies. Chang et al. (2018) proposed utilizing F_{ST} which measures the allele frequency among sub-populations to identify genomic region under selection pressure. Using F_{ST} scores to prioritize SNP markers in high density panels have led to increase in genomic similarity and in an improvement of the accuracy genomic selection compared to BayesB (Meuwissen et al. 2001) and BayesC (Habier et al. 2011) approaches. Furthermore, they showed that genomic relation matrix and accuracy could be improved using prioritized SNPs based on F_{ST} scores. Wang et al. (2012) evaluated a weighted single-step GBLUP

(WssGBLUP) approach using simulation data. They showed that weighting SNP could be effective in improving the accuracy of GEBV prediction and the estimate of marker effects. However, the challenge was on how to derive the optimum set of weights to compute the genomic relationship matrix. In this study, the Fst score based prioritization method developed by Toghiani et al. (2017) and Chang et al. (2018) will be expended to derive the needed weights to compute **G.**

The specific objective of this study was to derive F_{ST} score based relative weights for SNPs included in the computation of **G** and to assess the impact of different strategies on the estimation of variance components and the accuracy of genomic selection.

Materials and methods

Simulation of the univariate data:

Marker genotype data was simulation using the QMSim simulation software (Sargolzaei and Schenkel, 2009). A randomly mated historical population was generated to initialize LD and to establish mutation-drift equilibrium and was used as a base to create a population with average LD between adjacent markers of 0.3. Three hundred historical generations were generated based on random mating of an initial 8,000 animals, increasing to 15,000 animals at generation 305, decreasing to 12,000 animals at generation 1000, and then increasing to 17,000 animals at the last generation. The base population (G_0) was founded by 1,000 males and 15,000 females randomly selected from the historical population. A trait with heritability equal to 0.30 was simulated and all genetic variation was assumed to be due to the simulated QTL. The mating system was at random throughout up to generation G_0. An additional 7 selection generations (G_1-G_7) of

15,000 animals each were simulated. Parents were chosen based on their estimated breeding values (EBVs). The replacement rate for males and females was 50 and 20%, respectively. Throughout, one progeny per mating was assumed and the sex ratio of progeny was set to 50%. The average of the effective population size was equal to 323. The sixth generation (G_6) was used as a training population and the last generation (G_7) was used to evaluate (validation population) our proposed method.

All animals in the training and validation populations were genotyped. Genotypes were simulated for 400,000 biallelic SNP markers uniformly-distributed along 10 chromosomes of 100 cM in length each to roughly mimic 1.2 million SNP markers in the bovine genome. Two hundred biallelic QTL were sampled from a Gamma distribution with shape parameter equal to 0.4. No overlap between SNP markers and QTL was allowed. Additionally, QTL were assumed not to be genotyped. The residual variance was scaled accordantly in each scenario of selected SNPs such that the heritability and phenotypic variance was constant at the values of 0.3 and 1, respectively. Trait phenotypes were generated as the sum of an overall mean, the random additive effects of QTL and their associated genotypes, and the residual terms. The later were sampled from a normal distribution with zero mean and variance-covariance matrices $I\sigma_e^2$ where σ_e^2 is the residual variance.

SNPs prioritization based on F_{ST} scores:

Divergence between populations and subpopulations is often due to differential selection pressure. This differentiation between populations can be accessed using Wright's fixation indexes. In particular, F_{ST} has been used to measure the level of

variation among subpopulations. This measure of genetic differentiation among groups depends on the allele frequencies at the different loci. Furthermore, the size of the genetic differentiation is positively correlated with the magnitude of the F_{ST} scores. Although several estimators of genetic differentiation were presented in literature, in this study, F_{ST} scores were calculated following Nei (1973) and Chang et al. (2018). Conversely, the trait phenotypes for animals in generation 6 (G_6) were divided into three sub-populations based on the 5 and 95% quantiles (below the 5% quantile [S1], between 5 and 95% quantiles [S0], and above the 95% quantile [S2]). Genotypes of individuals (1,500) in sub-populations S1 and S2 were used to calculate the F_{ST} scores. For each locus, the global F_{ST} estimator was defined as:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

with $H_T = 2 * p * q$, $H_S = \frac{H_{S1} * n_{s1} + H_{S2} * n_{s2}}{n_{s1} + n_{s2}}$, and $H_{Si} = 2 * p_{Si} * q_{Si}$

where, p_{Si} and q_{Si} are the allele frequencies in subpopulation i, n_{s1} and n_{s2} are the number of individuals of the subpopulations, H_S is the average of sub-population heterozygosities and H_T is the heterozygosity based on the total population.

Prioritized SNPs and genomic relationships:

Several methods have been proposed to calculate the genomic relationships (Amin et al. 2007; Gengler et al. 2007; VanRaden 2008; Legarra and Misztal 2009; Yang et al. 2010). In the field of animal breeding and genetics, the genomic relationship matrix is generally calculated based on the method proposed by VanRaden (2008). It basically measures the similarity of marker genotypes between two individuals at a large number

of loci independently of their mode of inheritance. Although estimated observed additive relationships using identity by state are in general better than their counterparts obtained using pedigree information, they still suffer from several problems including the non-zero estimates of realized relationship between two individuals that are not related by ancestry as it was shown by (Li et al., 1993; Blouin et al., 1996; Csilléry et al., 2006), negative of diagonal elements, and the inevitable noisy associated with these estimates. Furthermore, several studies (VanRaden, 2011; Su et al. 2012a, b) have shown that little to no improvement in genomic relationship matrix were observed with the increase in the number of SNPs used for it calculation. In fact, Sue et al. (2012b) showed that accuracies obtained using a genomic matrix calculated based on 777K SNP panel were not any different from those obtained using the 54K SNP panel. These results clearly indicate that additive relationships between individuals could be accurately estimated with a reasonably small number of well distributed SNP markers. However, that does not mean that accuracy of genomic selection cannot be improved using high density marker panels or even sequence data. To achieve that goal, the genomic matrix has to evolve from a measure of additive relationship matrix to an optimum measure of genetic similarity between individuals. The current method used to calculate the genomic relationship matrix gives the same weight to all the markers and thus could not guaranty the optimality of genetic similarity. For that purpose, contributions of SNPs used to compute the genomic relationship matrix have to be weighted according to their importance on the phenotype (strength of association with the phenotype). To maximize the functional genomic similarity between individuals, SNPs have to be prioritized based on their ability to increase genetic or phenotypic similarity between individuals. Conversely, individuals

with different genetic values or phenotypes are likely to have much lower genomic similarity than the expected or observed additive relationships.

The challenge in maximizing the genomic similarities is finding the relative weights to the SNPs used in the calculation of the genomic relationship matrix. In this study, F_{ST} scores were used to prioritize and to assign relative weights to the SNP markers. The top 20K SNPs based on their F_{ST} scores were used either alone or with the remaining 380K SNPs to compute the genomic relationship matrix with or without weighting. When only the top 20K SNPs were used to compute the genomic relationship matrix, two scenarios were considered: 1) equal weights for all SNPs or 2) weights proportional to the SNP F_{ST} scores. When all 400K SNP markers were used, the different weighting scenarios evaluated are presented in Table 5.1.

The relative weights were calculated using the following equation:

$$w_i = \frac{Fst_i}{\sum_{j=1}^{N} Fst_j} * N$$

where w_i is the relative weight for SNP i, Fst_i is the F_{ST} score for SNP i and N is the total number of SNPs (400 K or 20K)

Statistical model and data analysis:

For all scenarios, 10,000 and 5,000 animals are randomly selected from G6 and G7, respectively. For each scenario, the genomic relationship matrix was computed with the appropriate number of markers and the weighting factors and the analysis was carried out using following mixed linear model:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the vector of phenotypes, $\boldsymbol{b}$ is the vector of fixed effects, $\boldsymbol{u}$ is the vector of genomic breeding values, and $\boldsymbol{e}$ is the vector of random residuals. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known incidence matrices with the appropriate dimensions. Additionally, it was assumed that $\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$ with **G** is the genomic relationship matrix and σ_u^2 is the genetic variance.

AIREMLF90 program, a modification of restricted maximum likelihood (REML) approach with the Average-Information algorithm (Misztal et al., 2002), was used to estimate variance components and genomic breeding values under the different scenarios. Accuracy of genomic evaluation was defined as the correlation between true breeding value and the genomic estimated breeding value in validation population. In this study, each simulation scenario was replicated 5 times.

Results and discussions

Table 5.2 presents the estimates of the variance components and heritability and their associated standard deviations for the different scenarios when all 400K SNP markers were used to compute the genomic relationship matrix. In general, the percentage of the genetic variance recovered increased with the decrease of the percentage weight assigned to the prioritized top 20K SNPs to reach a maximum when the top SNPs (based on FST scores) accounted for 25% or less of the weights used to compute **G**. In all cases, the genetic variance was underestimated when no weights were used (scenario 7 in Table 5.2). Similarly, approximately only two thirds of the genetic variance were recovered when zero weights were assigned to the 380K non-prioritized SNPs. The inability to recover all the genetic variance is due to the large number of QTL with very small effects. In fact, 55% of QTL have a true effect smaller than one tenth of

one percent and an additional 20% of QTL have an effect smaller 0.5% of the total genetic variance. These small QTL are hard track effectively when the LD is moderate to low. Across the different scenarios, there is an underestimation trend of the residual variance, although it does not seem to be any systematic bias. Heritability was clearly underestimated when the vast majority of weight (≥ 90%) was allocated to the prioritized top 20K SNPs. In fact, for those scenarios, estimates of the heritability are likely to be biased. For the remaining scenarios, although there is a general trend of an underestimation of the heritability, estimates are not likely not to be biased. When only the unweighted top 20K prioritized SNPs were used to compute **G**, the genetic and residual variances were very similar to the estimates obtained using scenario 1 in Table 5.2.

Intrinsically, the contribution of a SNP marker to the estimation of **G** is weighted by its minor allele frequency (MAF), favoring thus markers with low MAF. However, it is not weighed by the size of the marker effect. Consequently, after a certain number of SNP markers are included in the computation of **G**, little to no improvement is expected. Chang et al. (2018) showed that the limited change in **G** with additional markers could be an indicator of the sufficiency of available SNPs in estimating the realized relationships. However, such sufficiency is not a guaranty of the optimality of such matrix for the implementation of association and genome selection analyses. In fact, as the number of randomly selected SNPs increased from 40 K to 400 K, the matrix **G** inched closer to the expected additivity relationship matrix (**A**). Furthermore, they showed that a genomic relationship matrix computed based on a selected subset on 20 K markers was markedly different from **A**. In this study we further prove that within those selected 20K SNPs

additional improvement could be achieved through appropriate weighting of the contribution of these SNPs in the calculated of **G**. Table 5.3 presents the distribution of off-diagonal elements of **G** under different weighting scenarios. In fact, the portion of genomic relationships between training and validation individuals exceeding 0.03 was 5.24% when all 400K SNPs were used with equal weight. The same portion was 5.59, 7.22, 11.13, 14.38, and 16.78% when the relative weight assigned to the top 20K prioritized SNPs in the calculation of **G** was 25, 50, 75, 90 and 100%, respectively. When only the top 20K prioritized SNPs were used to compute **G**, weighting the contribution of each marker by its F_{ST} score resulted in an increase in the off-diagonal elements exceeding 0.03 (Table 5.4). The increase in the percentage of off-diagonal elements exceeding 0.03 is an indicator of increased similarity between the training and validation data sets and could lead to increase in accuracy.

When the same weight ($w_i = 1$) was used for all 400K SNP markers to compute **G**, the accuracy of genomic prediction (correlation between true and predicted BVs) was 0.690 (Figure 5.1) and it increased to 0.718 when all SNPs in the panel were weighted by their relative F_{ST} score. When the relative weight of the top 20K prioritized SNPs in the calculation of G increased, higher accuracy was achieved. In fact, accuracy increased by 4.3, 5.2, 5.4, 5.3, and 5.2% compared the scenario where all markers had the same weight ($w_i = 1$) when the relative weight assigned to the top 20K prioritized SNPs in the calculation of **G** was 25, 50, 75, 90 and 100%, respectively (Figure 5.1). Weighting all markers with their relative F_{ST} scores resulted in a 4.3% increase in accuracy compared to the same weight scenario. Using only the prioritized 20K SNPs with or without weights resulted in a 5.2 and 3.5% increase in accuracy compared to the same weight scenario.

These results clearly indicate that as the density of the marker panel increases, using all SNPs to compute **G** in the not the best option. Furthermore, weighting the prioritized markers could further improve the genetic similarity and ultimately the accuracy of genomic prediction. F_{ST} scores seem to be an efficient prioritization and weighting tool.

Conclusions

Using low to moderate density SNP marker panels, a substantial increase in accuracy was achieved. The dramatic increase in the number of identified common and rare variants due to advances in NGS was expected to significantly increase the accuracy of GWAS and GS. Unfortunately, little to no improvement in accuracy was observed using NGS or high-density marker data. In spite of the repeated argument that all needed information is already captured by the available marker panels, the results of this study clearly show that the lack of improvement in accuracy is due to the limitations of the methods used rather than the limited additional information in the high density/sequence data. Prioritizing SNP markers based on their F_{ST} score and using the latter to compute relative weights has increased the genetic similarity between training and validations animals. Furthermore, it resulted in more than 5% improvement in accuracy. These results clearly indicate that additive relationships between individuals could be accurately estimated with by a reasonably small number of well distributed SNP markers. However, that does not mean that accuracy of genomic selection cannot be improved using high density marker panels. The genomic matrix has to evolve from a measure of additive relationship matrix to an optimum measure of genetic similarity between individuals. The

current method used to calculate the genomic relationship matrix gives the same weight to all the markers and thus could not guaranty the optimality of genetic similarity.

Reference

Amin N, van Duijn CM, Aulchenko YS: A genomic background based method for association analysis in related individuals. PLoS ONE 2007, 2(12):e1274.

Balloux F, Brunner H, Lugon-Moulin N, Hausser J, Goudet J. Microsatellites can be misleading: an empirical and simulation study. Evolution. 2000;54 (4): 1414-1422.

Blouin MS, Parsons M, Lacaille V, Lotz S: Use of microsatellite loci to classify individuals by relatedness. Mol Ecol 1996; 5(3):393–401.

Chang LY, Toghiani S, Aggrey SE, Rekaya R. High density marker panels, SNPs prioritizing and accuracy of genomic selection. BMC Genetics 2018; 19:4.

Csilléry K, Johnson T, Beraldi D, Clutton-Brock T, Coltman D, Hansson B, Spong G, Pemberton JM: Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. Genetics 2006, 173(4):2091–2101.

Erbe M, Hayes B, Matukumalli L, Goswami S, Bowman P, Reich C, Mason B, Goddard M. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95(7):4114-29.

Gengler, N., Mayeres, P., and Szydlowski, M. A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007; 1: 21–28

Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinf. 2011;12(1):1.

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

Li CC, Weeks DE and Chakravarti A. Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered. 1993; 43:45–52.

Meuwissen THE, Goddard ME. Prediction of identity by descent probabilities from marker haplotypes. Genet Sel Evol. 2001;33:605-634. 10.1186/1297-9686-33-6-605.

Meuwissen THE, Hayes BJ, and Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001; 157: 1819-1829.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee D. "BLUPF90 and related programs (BGF90)". Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, 19-23 August 2002, Montpellier, France, 28–27.

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

Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25(5):680-1.

Schefers JM, Weigel KA. Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Animal Frontiers. 2012;2:4-9.

Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. 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. J Dairy Sci. 2012b; 95(8):4657–65.

Su G, Guldbrandtsen B, Gregersen VR, Lund MS. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J Dairy Sci. 2010;93(3):1175-83.

Su G, Madsen P, Nielsen US, Mäntysaari E a, Aamand GP, Christensen OF, et al. Genomic prediction for Nordic red cattle using one-step and selection index blending. J Dairy Sci 2012a; 95:909-917.

Toghiani S, Chang LY, Aggrey SE, Rekaya R. Genomic differentiation as a tool for single nucleotide polymorphism prioritization for Genome wide association and phenotype prediction in livestock. Livestock Science 2017; 205:24–30.

VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414-23.

VanRaden P, Van Tassell C, Wiggans G, Sonstegard T, Schnabel R, Taylor J, Schenkel F. Invited review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92(1):16–24.

VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA. Genomic evaluations with many more genotypes. Genet Sel Evol. 2011; 43:10.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb.) 94:73–83.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Health AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: Common

SNPs explain a large proportion of the heritability for human height. Nature Genet 2010, 42:565-569.

Zeng J, Toosi A, Fernando RL, Dekkers JC, Garrick DJ. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol. 2013;45(1):1.

Table 5.1. Percentage of the weight allocated to the prioritized 20K and the remaining 380K SNPs when the full panel (400K SNPs) was used to compute the genomic relationship matrix

	400K	
Scenario	20K[1]	380K
1	100	0
2	90	10
3	75	25
4	50	50
5	25	75
6	PS[2]	PS
7	Equal weights	Equal weights

[1] Top 20K SNPs based on F_{ST} scores; [2] Contribution proportional to the SNP F_{ST} score

Table 5.2. Variance component and heritability (SE) using all 400K SNPs and different weighting scenario to compute the genomic relationship matrix (average over 5 replicates)

Scenario[1]	Genetic variance	Residual Variance	Heritability
1= [100, 0]	0.196 (0.026)	0.671 (0.042)	0.228 (0.033)
2= [90, 10]	0.213 (0.018)	0.648 (0.032)	0.247 (0.023)
3= [75,25]	0.232 (0.015)	0.633 (0.025)	0.268 (0.018)
4= [50,50]	0.257 (0.016)	0.618 (0.021)	0.294 (0.018)
5= [25,75]	0.279 (0.021)	0.619 (0.021)	0.311 (0.023)
6= [PS^2,PS]	0.251 (0.032)	0.629 (0.037)	0.285 (0.037)
7= Equal weights	0.247 (0.027)	0.692 (0.016)	0.263 (0.025)

[1] [x,y] are the percentages of the weights allocated to the prioritized top 20K and the remaining 380K SNPs, respectively; [2] Contribution proportional to the SNP F_{ST} score

Table 5.3. Distribution of off-diagonal elements (OD) of the genomic relationships matrix corresponding to the training and validation individuals using all 400 SNPs and under different weighting scenarios for the prioritized[1] (20K) and non-prioritized (380K) SNPs (in %)

	Weights ($w_i \neq w_j$)	No weight ($w_i = 1$)	[100, 0]	[90, 10]	[75, 25]	[50, 50]	[25, 75]
OD<-0.05	0.92	0.00	2.32	1.66	0.92	0.25	0.03
-0.05<OD<-0.03	4.60	0.77	9.85	8.72	6.81	3.67	1.42
-0.03<OD<-0.01	30.26	29.77	28.18	29.16	30.45	31.72	31.22
-0.01<OD<0.01	43.93	52.43	33.49	35.54	38.86	44.57	49.81
0.01<OD<0.03	13.52	11.52	17.21	16.74	15.74	13.64	11.89
0.03<OD<0.05	4.04	3.3	5.52	5.01	4.36	3.68	3.36
ODE>0.05	2.73	2.21	3.46	3.19	2.86	2.49	2.27

[1] SNPs selected based on F_{ST} scores; [2] [x,y] are the percentages of the weights allocated to the prioritized top 20K and the remaining 380K SNPs, respectively

Table 5.4. Distribution of off-diagonal elements (OD) of the genomic relationships matrix corresponding to the training and validation individuals using the prioritized[1] 20K SNPs and under different weighting scenarios (in %)

	Weights ($w_i \neq w_j$)	No weight ($w_i = 1$)
OD < -0.05	2.32	1.61
-0.05<OD<-0.03	9.85	8.39
-0.03<OD<-0.01	28.18	29.35
-0.01<OD<0.01	33.48	36.14
0.01<OD<0.03	17.21	16.52
0.03<OD<0.05	5.52	4.86
ODE>0.05	3.46	4.86

[1] SNPs selected based on F_{ST} scores

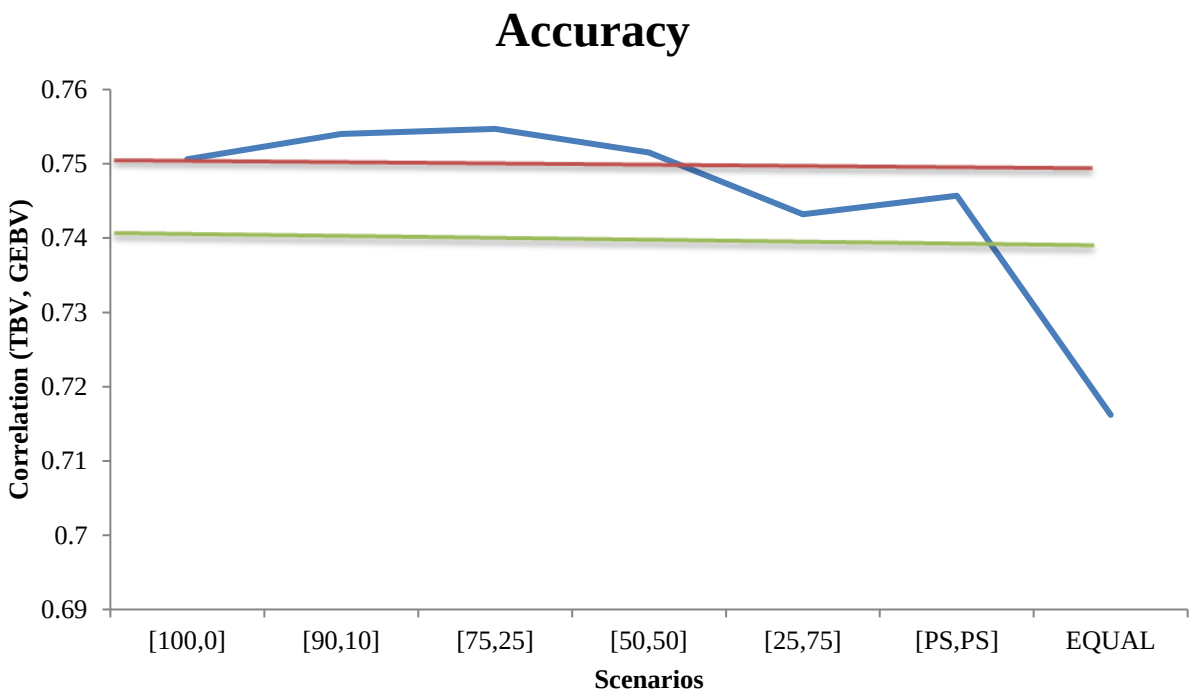


Figure 5.1. Accuracy of genomic prediction under different weighting scenarios for the contribution of the 20K prioritized SNPs and the remaining 380K markers [x,y]. Horizontal lines indicate the accuracy using only the top 20K SNPs with (red) or without (green) weights SNPs.

CHAPTER 6

CONCLUSION

The availability of high density marker panels and whole genome sequence data provide an unprecedented opportunity to dissect the associations between traits and genomic variation and was expected to increase the accuracy of genomic selection (GS) in livestock. Using low to moderate density SNP marker panels, a substantial increase in accuracy was observed. Unfortunately, the dramatic increase in genomic data did not lead to the expected increase in accuracy of GS. This lack of improvement is likely not the result of the limited additional information in high density panels, rather it is the result of the limitations of current methods used to implement GS. Independently of the method used, increase in marker density resulted in reduction of the statistical power for multiple regression based methods and to no improvement in the estimation of genomic relationship matrix (**G**) for mixed linear model approaches. Thus, further improvement of accuracy requires some form of prioritization of SNPs to be included in the association model or in the computation of **G**. Effect based prioritization (e.g., BayesB, BayesC) or using prior biological information (BayesRC) had limited success dealing with the problem and did not result in any meaningful increase in accuracy.

In this study, F_{ST}, a competitive measurement of genetic variation in different populations, was proposed as an alternative to prioritize SNPs. Prioritized markers based on different F_{ST} thresholds were able to track the majority of significant QTL, to increase genetic similarity between individuals, and to improve the accuracy of GS. A balance

between the increase in genetic similarity and the percentage of genetic variance explained is needed to reach optimum accuracy. Furthermore, accuracy could be optimized by assigning the proper weights to preselected SNP markers based on their F_{ST} scores. In fact, accuracy was increased by more than 5% under different weighting scenarios for the SNPs used to compute the genomic relationship matrix. When using prioritized SNPs to compute **G**, it is recommended that some non-prioritized markers be included. This additional SNPs although they do not contribute to the genetic similarity they help with the accurate estimation of the additive relationships.