GENOMIC SELECTION IN GENETICALLY HETEROGENEOUS POPULATIONS

By

EL Hamidi Abdel Hay

(Under the Direction of Romdhane Rekaya)

ABSTRACT

In Livestock applications, genome wide association studies and genomic selection are regularly conducted using purebred populations. Estimation and often validation of SNP are carried out using primarily pure bred animals. This process was successful when estimated SNP effects were used to predict genomic breeding values of animals of similar breed. However, it fails at different degrees when these SNP estimates are used for genomic prediction in other breeds or crossbred animals. Current approaches for dealing with admixed and crossbred populations in genomic selection rely on using different groups of pooled animals in the training and validation sets, and hence are data dependent and often lead to reduction in accuracies for animals in the pure breed populations. In an admixture population or in presence of crossbred animals, pooled data based methods assume that SNP effects are the same across breeds or sub-populations. This assumption is inaccurate due to the fact that several parameters such as allele frequencies, strength of linkage disequilibrium, and linkage phase change across sub-populations. To remedy the problem, we proposed a multi-compartment model where the effect of an SNP could be different between breeds and parameterized as a function of its effect on one of the breeds in the pooled population through a one to one mapping function. In a simulation study, it was shown the proposed multi-compartment model is clearly superior to the pooled

breed approach as it accounts for the difference in SNP effects across divergent lines. Its superiority compared to the pooled data approach ranged from approximately from 17 to 47% and increases as the divergence between lines increases. However, the proposed multi-compartment model suffers from the high dimensionality of the unknown parameters to estimate. In fact, an extra parameter per SNP and per component in the admixed population is needed to be estimated. Although the model works well when the number of animals in each breed is reasonable, it performance degrades as the number of animals in some lines decreases, making the estimation of their corresponding SNP effects numerically instable and, in extreme cases, statistically inefficient (severely biased). To overcome this problem, we proposed not to estimate a mapping parameter for each SNP rather to build a model as a function of information already available in the genotype data via a hierarchical structural model. In this study, the genetic difference between lines was modeled as a function of the change in linkage disequilibrium and the potential change in linkage phase.

INDEX WORDS: Genomic selection, SNP, admixed, linkage disequilibrium

GENOMIC SELECION IN GENETICALY HETEROGEOUS POPULATIONS

By

El Hamidi Abdel Hay

BS, Southern Polytechnic State University, 2008

MS, University of Georgia, 2011

A Thesis Submitted to the Graduate Faculty of the University of Georgia in Partial

Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

GENOMIC SELECION IN GENETICALLY HETEROGENEOUS POPULATIONS

by

El Hamidi Abdel Hay

Major Professor: Romdhane Rekaya

Committee: Ignacy Misztal

J. Keith Bertrand

Sammy Aggrey

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

December 2014

DEDICATION

To: My Family.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to Dr Romdhane Rekaya for his guidance, support and input. Without him I could not have done it. I would like to also thank the entire animal and breeding genetics group at The University of Georgia.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the advancement of high-throughput genotyping technologies, genetic improvement of livestock species and plants has changed dramatically. Before the use of genomic information, animal breeders relied on the use of the BLUP methodology (Henderson, 1984) to rank and select genetically superior animals. Needless to say, BLUP technic worked extremely well and has led to substantial genetic gain. It is still the most used method for genetic selection in livestock and poultry species. Unfortunately, classical genetic improvement suffers from few limitations. Such limitations are the need for continuous phenotypes collection, pedigree recording and also the long generation interval in some species. The completion of the human and several livestock genomes was a major breakthrough, and it provided an unprecedented opportunity to understand and dissect traits important to the livestock industry. Currently it is a routine to genotype animals for thousands of single nucleotide polymorphisms (SNP), generating high density panels. These high density marker panels were designed to capture linkage disequilibrium between SNP markers and potential quantitative trait loci (QTL). These SNP markers could be used to compute genomic estimated breeding values (GEBV) as suggested by Meuwissen et al. (2001) leading to the so called genomic selection. Several studies have shown that high accuracies for GEBV can be obtained compared to traditional EBV and a substantial decrease in generation interval is achieved (Meuwissen et al., 2001; Shaffer et al., 2006). Currently, genomic selection is carried out using two alternative methods either multi-step procedure (Goddard and Hayes, 2007) or single step procedure based on BLUP methodology

that uses a relationship matrix merging pedigree and genomic information (Aguilar et al., 2010). Genomic selection is often conducted using purebred populations. Training and validation are mostly carried out using a select elite set of pure bred animals (i.e. proven sires). One major limitation of genomic selection is prediction equations derived from one breed couldn't be applied to other breeds or crossbred animals. This situation could be problematic to some segment of livestock industry (beef cattle, swine or poultry) where the traits of interest are measured in crossbred or admixed populations with uncertain breed composition (Kachman et al., 2013; Toosi et al., 2009). Current approaches for dealing with admixed and crossbred populations in genomic selection relies on adjusting genomic matrix to account for different breeds or pooling different breeds or lines of animals in the training set. However pooled data approach suffers from few problems such as data dependence and reduction in accuracies for animals in the pure breed populations.

In this dissertation project, an innovative method for adjusting for the heterogeneity of the data by allowing SNP marker effects to change across different subgroups was proposed and evaluated using simulated data. This new method present a clear departure for existing methods that assume constant SNP effects across breeds or lines. The objectives of this research project are the following:

1) Develop a hierarchical Bayesian model that allows for the change of SNP marker effects between different breeds.
2) To build a model for the change in SNP marker effect across lines as a function of information available in the observed marker genotype data via a hierarchical structural model.

CHAPTER 2

LITERATURE REVIEW

Use of Genomic Information

With the completion of cattle genome in 2004 and the advancement in genotyping and sequencing technologies it is now possible to efficiently genotype animals for thousands of single nucleotide polymorphisms (SNP), generating high density marker panels. These high density panels provided an opportunity to identify SNP markers in linkage disequilibrium with potential quantitative trait loci (QTL) or identify functional genes. There are many benefits of performing genome wide association studies (GWAS) in animal agriculture. By using groups of markers, the effects of genomic regions can be estimated and combined to form genomic estimated breeding values (GEBV) as suggested by Meuwissen et al. (2001). This method is a marker assisted selection which is referred to in the literature as genomic selection. These GEBVs are more accurate in estimating the true genetic potential of an animal than those obtained using the classical estimated breeding values (EBVs). Additionally, GEBVs can be calculated early on in the life of an animal, thus substantially reducing the generation interval as shown in a study by Schaeffer (2006) and others. The increased accuracy and the reduced generation interval will lead to an increase in the genetic response. Furthermore, GWAS is a useful tool in the discovery of functional variants, and /or causal mutations affecting economically important traits. Such discovery could be of great importance for the understanding of the genetic mechanisms underlying complex traits (Hirschhorn et al., 2005) and an opportunity for further improvement of selection methods. Using multiple regression or variance

component based models, substantial progress has been achieved in the last ten years in estimating genomically enhanced breeding values (König et al., 2009). Several breeding associations and breeding companies have already started using such information in their genetic selection programs. Several studies showed a substantial increase in reliability using genomic information compared to classical methods (Harris and Johnson 2010; Su et al., 2012). Unfortunately, genome wide evaluation methods suffer from few limitations such as the large size of genotypic data, population stratification and genomic pre-selection.

Genome Wide Association Studies

Genome wide association studies (GWAS) rely on estimating the association between phenotypic variation and a large number of genetic variants such as SNP markers. GWAS is a useful tool; it has led to several discoveries in human applications, livestock and plants (Hindorff et al., 2008; Visscher et al., 2012). The incentive of performing GWAS in livestock animals is the discovery of genes controlling economically important traits. Dairy cattle are ideal populations to perform GWAS due to the small effective population size and the strong artificial selection that has been applied over time. These population characteristics create strong blocks of LD, making association studies informative and more accurate. A study by De Roos et al. (2008) showed that in order to find significant associations in dairy cattle, SNP markers should be placed approximately every 10 Kb. Using dairy cattle population, Pryce et al. (2009), carried out a genome wide association study to validate and identify gene regions controlling production and fertility traits. Significant associations were observed such as a putative QTL on chromosome 18 affecting fertility. Additionally, several other mutations affecting milk production were validated. A genome wide association study by Guo et al. (2012) was carried out and revealed several candidate genes that control production traits in cattle. Daetwyler et al. (2007) performed

a genome scan in Holstein cattle and reported several associations between SNP markers and potential QTLs. Significant regions in Bos Taurus autosomes were found to be associated with milk yield and protein yield.

In human applications, genome wide association study was not as successful. Many genetic variants were reported as having an association with a certain disease or a trait, however majority of these variants have little to no established biological relevance (McClellan and King, 2010). Genome wide association studies (GWAS) suffers from a major problem which is the high dimensionality of the parameter space causing a large number of false positives. Furthermore, GWAS is still unable to identify sufficient number of variants that could explain the majority of the variability observed in traits of interest. A famous example is "lost heritability" in the case of human height (Maher, 2008). Studies showed that this lost heritability problem is largely due to the lack of power to identify variants with small effects which jointly explain large portion of the total genetic variation. Further, complex traits are often under the control of genetic and environmental factors and their potential interaction. Thus, detecting genetic variants associated with these traits is challenging especially when these variants have moderate to small effects.

Genomic Selection

Genomic selection was first proposed by Meuwissen et al. (2001), a decade later it has completely changed genetic selection and improvement of livestock species. Genomic selection consists of the prediction of the genetic merit of animals based on a high number of genetic markers that are in linkage disequilibrium with quantitative trait loci. The techniques used to implement genomic selection, the parameters that affect its accuracy, and some limitations of the method will be discussed in the next section.

Linkage Disequilibrium

Genomic selection relies on linkage disequilibrium (LD) between markers and quantitative trait loci. Linkage disequilibrium is the non-random association of alleles at different loci. When the physical distance separating two genes is short, these two genes will tend to get inherited together. LD is affected by factors such as effective population size, evolutionary forces, mutation and also admixture and migration (Hill and Robetson, 1968; Sved, 1971; Ardlie et al., 2002). Linkage disequilibrium is computed using either quantities D' or r^2. Both methods are commonly used; D' is a measure of LD from a biological perspective which measures the probability of possible haplotypes. On the other hand, r^2 is a statistical measure of the correlation between two loci; it ranges between 0 and 1, 0 being no linkage disequilibrium and 1 being perfect linkage disequilibrium.

Assuming two loci and two alleles for each locus (A, a, B, b) and the genotypic frequencies of AB, Ab, aB and ab are f_{11}, f_{12}, f_{21} and f_{22} respectively. Linkage disequilibrium, D, is calculated according to Hill and Robertson, (1968) as the following:

$$D = f_{11} - f_{22} - f_{12} f_{21}$$

Among all species, humans have the lowest extent of linkage disequilibrium due to a large effective population size. Some studies reported that LD extents to approximately 50 Kb (Pritchard and Przeworski 2001; Maniatis et al., 2002; Weiss and Clark, 2002). In Livestock species, LD is stronger than in humans as a result of stronger natural or/and artificial selection leading to smaller effective population size (Nsengimana et al., 2004). In a study by McRae et al. (2002), genetic markers were used to assess the extent of LD in two sheep populations. The study showed high levels of LD extending for tens of cM, and decreased as the physical distance

between markers increased. Further, measuring LD could be difficult as several parameters could affect the final outcome. The type of markers used to measure LD is important; LD measure in humans using SNP markers is smaller when using microsatellites (Pritchard et al., 2001). Most studies measuring LD in cattle use microsatellite markers instead of SNP markers, since microsatellites tend to have higher heterozygosity.

Genome wide association studies (GWAS) and genomic selection strongly depend on the strength of LD between SNP markers and potential QTLs affecting traits of interest. In a study by Khatkar et al. (2008) showed that in the case of association mapping in Holstein-Friesian cattle, one SNP marker is needed every 40 Kb. The need of high density SNP marker panels is justified, since they capture LD between SNPs and QTLs. However, genotyping for high density panels is still relatively expensive. To overcome this issue, several methods have been proposed to impute SNP marker genotypes.

Imputation of SNP Genotypes

The cost of genotyping is still relatively high and different sizes of SNP marker panels are available ranging in density from few thousands SNPs to hundred thousand SNPs. A plausible solution to controlling the cost of SNP genotyping is to genotype animals for cheap low density panels and then impute the missing genotypes to construct a higher marker density map. Several methods have been developed to impute missing genotypes. These methods could be classified into two categories: methods based solely on linkage disequilibrium (population based methods) and methods which exploit linkage disequilibrium and pedigree information (family based methods). The first type of imputation methods is used mainly in human applications due to the lack of pedigree information, while the second type is used in animal agriculture. For

instance, Fast-Phase algorithm (Scheet et al. 2006) utilizes population based linkage disequilibrium. It is based on the idea that haplotypes in a certain population cluster together, so a hidden markov chain was implemented (Scheet et al., 2006) to predict SNP marker genotypes. In another study, a neural network approach has been adopted to impute missing SNP genotypes and it showed decent levels of accuracy (Sun et al., 2008).

Genotypes phasing is also of a great importance in animal agriculture. Knowing the source of alleles of the genotype is useful in QTL mapping, understanding the underlying biology of the traits of interest and in detecting genetic imprinting. Most imputation algorithms have the ability of phasing genotypes. Daetwyler et al. (2011) described an imputation and phasing algorithm (ChromoPhase) which utilizes the characteristic of finite populations to phase sections of the genome. The algorithm was applied to real Holstein data to impute missing genotypes in a 3k SNP chip panel to construct a high density 50K SNP chip panel. The algorithm performed well in imputing the missing genotypes with an accuracy of 92% for animals with a genotyped sire. Accuracy of the imputation of genotypes depends on different parameters such as the size of reference population, the origin of the reference population and the genetic relationships. Huang et al. (2012) showed that the accuracy of imputation increased with the increase of the size of the reference population. Increasing the reference population by 100 individuals increased the accuracy by 8%.

Multi-step genomic selection approach

Multi-step approach are often a regression based method (VanRaden 2008, Hayes et al. 2009) and it consists of a sequence of several steps: 1) BLUP analysis to generate pseudo phenotypes such as EBVs or de-regressed proofs or DYD, 2) estimation of SNP marker effects often using a

Bayesian based method (Meuwissen et al., 2001; Gianola et al., 2006), and 3) calculation of the GEBVs as a linear function of the estimated SNP effects.

In the multi-step genomic selection, the statistical model generally used is:

$$y_i = \mu + \sum_{j=1}^{nsnp} x_{ij} g_j + e_i$$

where **y** is a column vector of pseudo phenotypes such as de-regressed proofs or DYDs, $\mathbf{x_i}$ is the vector of SNP marker genotypes, $\mathbf{g_i}$ is the SNP marker effect, and **e** is the error.

If the number of SNP markers exceeds the number of phenotype records, the statistical model becomes non-identifiable. Implementing the model through a frequentist method is not possible and therefore a Bayesian approach is necessary. Meuwissen et al. (2001) proposed different Bayesian methods to implement the model above.

Single-Step Genomic selection

Single-step genomic selection is a unified approach eliminating the SNP markers effects estimation as in the multi-step (Mistzal et al. 2009). This approach is based on an enhanced relationship matrix, called a genomic relationship matrix which combines genomic information and pedigree information as described by Legarra et al. (2009). The model used is as follows:

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

y is a vector of observations, **b** is a vector of fixed effects and **u** is a vector of random animal effects and e is the residual. The relationship matrix can be modified to $\mathbf{H = A + A_\Delta}$ to account for genomic information and $\mathbf{A_\Delta}$ is the deviation from expected relationships. Matrix G replaces

the numerator relationship matrix for the genotyped animals (Legarra et al., 2009). Solving MME is exactly the same as in traditional mixed models. A detailed explanation on the construction of the G matrix could be found inVanRaden (2008).

$$G = \mathbf{ZZ}' / [2\Sigma p_i q_i].$$

Where **Z** is nxm genotypes matrix, n is the number of animals and m is the number of genotypes and p_i and q_i are allelic frequencies. Division by $[2\Sigma p_i q_i]$ makes **G** analogous to **A**.

Aguilar et al. (2009) implemented a single-step procedure for genomic evaluation using national evaluation framework and compared its performance to a multiple-step procedure. The single step approach performed similarly to the multi-step approach and yielded similar accuracies. It is important to note that the single step approach has many advantages compared to the multiple step approach. Multiple step procedure requires 1) classical animal model evaluation 2) generation of pseudo phenotypes such as de-regressed proofs or daughter deviations 3) estimation of a large number of parameters (VanRaden et al., 2009b; Misztal et al., 2009; Aguilar et al., 2009). Single step eliminates all these steps.

Population Stratification

Genome wide association study (GWAS) is a useful tool in identifying variants associated with a certain trait. This method has been shown to work well when the population is homogenous. However when the population consists of different subgroups of genetically distinct individuals, GWAS performs poorly. Accounting for population stratification is crucial and substantial literature has been already published (Yu et al., 2006; Kang et al., 2010; Price et al., 2010; Wang et al, 2014).

Population stratification and heterogeneity is also a problem in the animal genomic field. Genomic selection is usually conducted on purebred animals and it works well when sufficient genotypes are available. Unfortunately, these ideal conditions are often violated at different degrees when only limited number genotypes are available for a breed or a crossbred sub-population. Additionally, for some commercial level animals (often crossbreds), phenotypic information is seldom available limiting thus the potential use of genomic prediction. In genomic selection, SNP marker effects are estimated using a training population or e reference population. The latter consists of individuals with both genotypic and phenotypic records. The validation data set contains individuals without phenotypic records, SNP marker effects estimated from training data set are used to predict genomic estimated breeding values (GEBV). Several studies showed that applying SNP marker effects estimated in a certain breed do not predict accurately GEBVs in another breed (Hayes, et al., 2009; Erbe et al., 2012; Weber et al., 2012).

Genomic selection is based on assumptions which fit a single breed scenario, however sometimes the population of interest is an admixed or crossbred population as mentioned earlier. Population make-up is a major factor in determining the accuracy of genomic predictions (Goddard, 2009; Habier et al., 2007, 2010). In a study by Daetwyler et al. (2012), they decomposed the accuracy of genomic prediction into the contributions from population structure and LD between SNP and QTL in a multiple breeds of sheep.. They concluded that the accuracy of genomic predictions strongly depends on the population structure.

In a study by Wientjes et al. (2013), they investigated the effects of relationships and linkage disequilibrium between marker and QTL on the accuracy of genomic predictions. Their results showed that accuracy of genomic predictions depends on linkage disequilibrium and on the

extent of the relationship between the reference population and the validation population. Toosi et al. (2009) carried out a simulation study using different training population structure and their results showed that the accuracy of prediction is highest when the same breed is used in training and validation data sets. The accuracy decreased with the decrease in the genetic similarity between training and validation populations.

As indicated earlier, genomic selection could be implemented through two approaches: 1) variance component based approach such as the single step procedure where a a genomic relationship matrix instead of pedigree relationships is used or 2) a regression based approach where first SNP effects are estimated in a training data set and then tested on a validation set. This multi-step approach has been implemented through different procedures, especially Bayesian methods via Markov Chain Monte Carlo (MCMC) techniques.. Hayes et al. (2009) showed that using a genomic relationship matrix to carry out genomic selection is an attractive approach; however it does not perform as well as Bayesian approaches in the case of multiple breed populations. It was concluded that the accuracy of genomic prediction not only depends on LD and population structure but also on the type of method used to implement genomic selection (Hoze et al., 2014).

Population stratification is a hot area of research in both human genetics and in animal and plant breeding and genetics. Several methods have been proposed to deal with this issue.Genomic selection is based on estimating genetic merit based on a large number of SNP markers spread across the genome (Meuwissen et al., 2001). The accuracy of genomic selection depends on different factors. A major factor is the size of the reference population (VanRaden et al., 2009; Schenkel et al., 2009; Zhou et al., 2013). For instance, VanRaden et al. (2009) showed an

increase in coefficient of determination with increasing the number of bulls in the reference population.

One of the main limitations of genomic selection is prediction across different breeds or lines. SNP markers estimated in one breed are poor predictors of genomic estimated breeding values in other breeds. This is due to several genetic parameters changing across breeds, 1) linkage disequilibrium 2) allele frequencies and 3) linkage phase. The best scenario of carrying out genomic selection is training and validation on the same breed. However, in some cases the reference population consists of a mixture of different breed or/and crossbred animals with some of the components of the population have limited number of phenotypes and SNP marker genotypes. In such scenario, within breed genomic selection is limited or not possible. One potential solution that has been proposed was to to pool data from different breeds into one large multi-breed reference population. De Roos et al. (2009) assessed the accuracy of genomic prediction using a simulated multiple breed training population consisting of two divergent breeds. The results showed an increase in accuracy when the two breeds were pooled together to construct the training population. This could be beneficial when one of the breeds is too small for population specific analysis. In the case of training on one population and validating on another one, the accuracy of genomic prediction was extremely low indicating low to no predictive power. Furthermore, the study revealed that heritability and marker density has a strong impact on the accuracy of genomic predictions. In an another simulation study Toosi et al. (2010) assessed the impact of the training population structure on the accuracy of genomic predictions. Their results showed that training and validating on the same population yielded the highest accuracy, and the lowest accuracy was observed when training on one population and validating

on a different one. Pooling populations in the training population resulted in accuracies in between the two previous described scenarios.

Genomic selection is often conducted on purebred animals to improve commercial animals which are crossbred animals. In a study by Kizilkaya et al. (2010), accuracy of genomic prediction was compared between two scenarios, the first scenario is training on purebred animals and validating on muli-breed animals and the second scenario is training on multi-breed and validating on purebred animals. The results showed that the accuracy of genomic prediction was higher when training on purebred animals and validating on multi-breed animals than training on multi-bred and validating on purebred population. Their argument was that purebred animals have greater levels of linkage disequilibrium in purebred populations than multi-breed populations. Also, they argued that in order to increase genomic prediction accuracy across different breeds, a higher number of markers is required.

More recently, Kachman et al. (2013) compared the accuracy of genomic prediction using single and multiple breed training population in real beef cattle data. The study showed that prediction accuracy of genomic breeding values for breeds that were not in the training population was low. In the case of breeds included in the training population, the accuracies did not differ between using single breed training population or multiple breed training population. Similar results were reported by Weber et al. (2012). Genomic prediction accuracy when training on multi-breed populations was higher for Angus and Herford since they had large number of records in the training population. On the other hand, accuracies were lower when training on single breed due to the small number of records in the training population.

In dairy cattle, Olson et al. (2012) investigated three different methods of genomic evaluation using three dairy breeds (Holstein, Jersey and Brown Swiss). The first method used a single breed training population and validate on a different breed. The second method uses a multi-breed training population and the third method is a multi-trait model considering each breed as a different trait. The first method performed poorly, the accuracy of genomic prediction was low and even negative in some cases. The second method performed better than the first one due to the increase in the training population size especially for Brown Swiss because of its small size. Method three slightly increased the accuracy for all three breeds compared to the other two methods.

Pooling multiple breeds to construct a large training population is an attractive approach; however it suffers from several limitations due to its strong genetic assumptions.

Correcting For Population Stratification

As mentioned earlier, there are two main approaches to conduct genomic selection. Population stratification is a problem that arises using either of the two approaches. Hayes et al. (2009) showed that genomic selection using a genomic relationship matrix does not perform as well as Bayesian based methods in the case of multi-breed reference population.

Using a multiple breed training population intrinsically assumes that SNP marker effects are the same across breeds. This assumption is seldom true. Ibanez-Escriche et al. (2009) proposed a method to model breed specific SNP markers. They compared the proposed model with the classical across breed genomic selection. The results showed that using a breed specific SNP model performed better than the classical model that ignores differencesbetween breeds only when the number of markers was small. The divergence of the breeds and the size of training

population affected the results. As the breeds diverged, the breed specific SNP model performed better.

A study by Makgahlela et al. (2013) proposed a model to correct for the existence of multiple breeds when conducting genomic selection. The proposed model is a random regression like approach that accountsfor the percentage of the four breeds considered in the training population, their model was:

$$y_i = \mu + \sum_{j=1}^{4} c_{ij} b_j + \sum_{j=1}^{4} \sqrt{c_{ij} a_{ij}} + e_i$$

Where y_i is the deregressed proof of the ith bull, μ is the overall mean, b_j is the fixed regression effect of breed j (j=1, 2, 3, 4), c_{ij} is the breed proportion of bull i, a_{ij} is genomic breeding value and e_i is the residual.

Their results showed that the model accounting for breed specific SNP marker effects did not perform better than the classical genomic model ignoring breed specific marker effects which was GBLUP in their case.

Using a variance components approach, Harris and Johnson et al. (2012) proposed a method to adjust the genomic relationship matrix in the case of multiple breed populations. In their study, the genomic relationship matrix was extended to multi-breed population by taking into account the differences in allele frequencies among breeds. The regression method used to estimate the genomic matrix by VanRaden (2008) was extended to a multiple regression to adjust co-variances between relatives in multi-breed populations. This study showed that ignoring breed differences resulted in a biased genomic relationship matrix differing significantly from the expected value, the relationship matrix A.

Literature Cited

Ardlie KG, Kruglyak L, Seielstad M: Patterns of Linkage Disequilibrium in the human genome. Nat Rev genet 2002, 3(4):299-309.

Daetwyle H. D., K. E. Kemper , J. H. J. van der Werf and B. J. Hayes. Components of the accuracy of genomic prediction in a multi-breed sheep population. 2012. *J ANIM SCI vol. 90 no. 10 3375-3384*

Daetwyler D.H, Schenkel F.S, Sargolzei M., Robinson J.A.B. A Genome Scan to Detect Quantitative Trait Loci for Economically Important Traits in Holstein Cattle Using Two Methods and a Dense Single Nucleotide Polymorphism Map. J.Dairy Sci. 2007. 91:3225-3236

Daetwyler H.D, Wiggans G, Hayes B. J, Woolliams J.A, Goddard M.E: Imputation of missing genotypes from sparse to high density using long range phasing 2011. *Genetics, Vol. 189, 317-327.*

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

Erbe et al., M. Erbe, B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, Mason, M.E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels J. Dairy Sci., pp. 4114–4129

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

Goddard M.E. and Hayes B. J. Genomic Selection. J. Anim. Breed. Genet.124 (2007) 323–330

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

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

Habier D., Tetens J., Seefried F.-R., Lichtner P., Thaller G. *2010.* The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol. 42:5.*

Harris BL, Johnson DL: Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J Dairy Sci 2010, 93:1243–1252.

Harris, BL., Johnson DL, Spelman RJ: Genomic selection in New Zealand and the implications for national evaluation. Proceedings of the Interbull Meeting, Niagra Falls, Canada 2008.

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

Hayes, B. J, Bowman P. J, Chamberlain A. C, Verbyla K, Goddard M. E. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 2009. 41-51

Henderson, C.R., 1984. Applications of linear models in animal breeding. Univ. Guelph, Guelph, Canada.

Hill, W. G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide associations of genome wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2008, 106:9362-9367.

Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005;6:95–108.

Hoze, C., Fritz, S., Phocas, F., Boichard, D., Ducrocq, V., Croiseau, P., 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J. Dairy Sci. 97:3918-3929

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, A. J. C. M Dekkers. 2009. Genomic selection of purebreds for crossbred performance. *Genetics Selection Evolution* 2009, 41:12

Ioannidis JPA. Non-replication and inconsistency in the genome-wide association setting. Human Heredity. 2007;64:203–13.

Kang, H.M., et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354

Kizilkaya K, R. L. Fernando, Garrick D. J. Genomic prediction of simulated multi-breed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 2010. 88:544-551

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

Legarra A., I. Aguilar, I. Misztal. A relationship matrix including full pedigree and genomic information. J. Dairy Sci., 92 (2009), pp. 4656–4663

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

Makgahlela ML, Mäntysaari EA, Strandén I, Koivula M, Nielsen US, Sillanpää MJ, Juga J: Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J Anim Breed Genet 2012, doi:10.1111.

Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W,Ennis S, Ke X, Morton NE: The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci USA 2002, 99(4):2228-2233.

McClellan, J., and King M.C. Genetic Heterogeneity in human disease. 2010. Cell 141,210-217

McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J: Linkage disequilibrium in domestic sheep. Genetics 2002, 160(3): 1113-1122.

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

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

Nsengimana J, Baret P, Haley CS, Visscher PM: Linkage disequilibrium in the domesticated pig. Genetics 2004, 166(3): 1395-1404.

Olson, K. M, VanRaden P. M, Tooker M.E. Multi-breed genomic evaluations using purebred Holsteins, Jerseys and Brown Swiss. 2012 J. Dairy Sci. 95-5378-5383.

Price , L., Zaitlen, N.A., Reich, D., Patterson, N., 2010. New approaches to population stratification in genome wide association studies. Nat Rev Genet. 459–463

Pritchard JK, Przeworski M: Linkage disequilibrium in humans:models and data.Am J Hum Genet2001, 69(1):1-14.

Pryce JE, Gredler B, Bolormaa S, Bowman PJ, Egger-Danner C, Furest C, Emmerling R, Solkner J, Goddard ME, Hayes BJ: Genomic selection using a multi-breed, across country reference population. J Dairy Sci 2011. 94:2625-2630

Pryce, J. E., Bolormaa S., A. J. Chamberlain, P. J. Bowman, K. Savin, M.E Goddard, B.J. Hayes. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J Dairy Sci. 93:3331-3345

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

Scheet P and Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am. J. Hum. Genet. 2006; 78:629–644

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

Schenkel, F., Sargolzaei, M., Kistemaker, G., Jansen, G., Sullivan, P.,Van Doormaal, B.J., Vanraden, P.M., Wiggans, G.R., 2009. Reliability of genomic evaluation of Holstein cattle in Canada. Interbull. Bull. 39, 51. Theor. Popul. Biol. 125-141.

Sun YV, Kardia SL. 2008. Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. Eur J Hum Genet. 16(4):487-95.

Toosi A., R. L. Fernando and J. C. M. Dekkers. Genomic selection in admixed and crossbred populations 2009 *J ANIM SCI* 2010, 88:32-46.

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

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

Visscher PM, Brown MA, McCarthy MI, Yang J: Five years of GWAS discovery. Am J Hum Genet 2012,90:7-24.

Visscher PM, Brown MA, McCarthy MI, Yang J: Five years of GWAS discovery. Am J Hum Genet 2012,90:7-24.

Weber KL, Thallman RM, Keele JW, Snelling WM, Bennett GL, Smith TP, McDaneld TG, Allan MF, Van Eenennaam AL, Kuehn LA. Accuracy of genomic breeding values in multi-breed beef cattle populations derived from de-regressed breeding values and phenotypes. J Anim Sci. 2012 90(12):4177-90

Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. Trends Genet 2002, 18(1):19-24.

Wientjes Y. C. J., R. F. VeerKamp, P. L. Calus. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. 2013. Genetics, Vol. 193, 621-631

Yu J, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006; 38:203–208

Zhou, L., Ding, X., Zhang, Q., Wang, Y., Lund, M.S., Su, G., 2013. Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. Genet. Select. Evol. 45, 7,

CHAPTER 3

A MULTI-COMPARTMENT MODEL FOR GENOMIC SELECTION IN ADMIXTURE POPULATIONS

[1] Hay, E and R. Rekaya. 2014. Submitted to the *Journal of Livestock Science*.

Abstract

Genome wide evaluation methods are often conducted using purebred populations. Estimation and often validation are carried out using primarily select elite animals. This process is successful when estimated SNP effects are used to predict genomic breeding values of animals of similar breed. This approach fails when SNP estimates in one breed are used for genomic prediction in other breeds. In this study, we proposed a multi-compartment model where the effect of an SNP marker could differ between breeds. A simulation was carried out using an admixed population of two divergent lines (A and B) genotyped for 300 markers. Divergence between the two lines was artificially created by multiplying marker effects in one line by a variable α which was sampled from different uniform or normal distributions. The proposed method was compared to the pooled data approach based on the accuracy of predicting the true breeding values. The prediction accuracy using the pooled data approach for line A, was 0.40, 0.39 and 0.38 when **α** was generated from a uniform distribution between [-2, 2], [-4, 4] and [-8, 8] respectively. Using our proposed method, the corresponding accuracies were 0.47, 0.46 and 0.46, respectively. A similar trend was observed for line B with a clear superiority of the multi-compartment model over the pooled data approach with an increase ranging from 17 to 47% and increases as the divergence between lines increases.

Keywords: Genomic selection, Admixed population, SNP

Introduction

Recent advances in molecular genetics, especially large scale genotyping for single nucleotide polymorphisms (SNPs) have provided an unprecedented resource to study association

between traits of interest and genomic variation, to compute genomically enhanced breeding values, and to ascertain population heterogeneity and relationships between its members. Undeniable success was observed in all fronts. In fact, several studies have been successful in identifying relevant associations between complex responses and genomic variation for human diseases (Visscher et al., 2012), and livestock and plant traits of economic interests (Bennett et al., 2010; Bolormaa et al., 2010; Snelling et al., 2010). Although a detailed dissection, at the genetic level, of these complex traits is still largely elusive, continuous improvement in the quality and diversity of high-through put data as well as the development of more sophisticated statistical, bioinformatics, and computational tools are quickly moving us towards the ultimate goal. In livestock and plant applications, the benefit of using genomic information is not limited to the genetic dissection of complex traits and the potential discovery of relevant or functional variants, but also to enhance the estimation of breeding values and ultimately the increase of the genetic progress through so called genomic selection. The later uses genomic markers that are in linkage disequilibrium (LD) with Quantitative Trait Loci (QTLs) to estimate breeding values.

Genomic selection is currently implemented either through a multiple regression (RM) or variance component (VC) based models. RM approach consists in a multiple step procedure where SNP effects are first estimated in a training population and then validated in separate data set. Several procedures including single marker analyses (Habier et al., 2007), ridge regression (Xu, 2003), non and semi parametric methods (Bennewitz et al., 2009), and Bayesian approaches (Meuwissen et al., 2001) have been developed and used to implement the RM. Although these methods have different statistical and biological assumptions regarding the data generating process, they tend to yield similar results in the majority of the cases and differences are largely

due to the genetic architecture of the trait, the genetic relationships between individuals in the sample, and the chosen prior information.

Benefit of genomic selection is a more accurate pre-selection of animals that inherited genes or chromosome segments of superior merit (Meuwissen et al., 2001). In Dairy cattle, for example, accuracy of the genomically estimated breeding values (GEBVs) are 30 to 70% higher than their counterparts obtained used the classical BLUP approach (VanRaden et., 2009; Harris and Johnson 2010; Su et al., 2012). Additionally, genomic selection allows for a significant reduction of generation interval as young animals could be genomically evaluated at birth or even before; thus reducing or even eliminating the need to wait for several years (depending of the specie and the trait) until enough phenotypic information is collected and a reliable genetic evaluation is conducted. Thus, it is not surprising that genomic selection is quickly becoming the method of choice for genetic evaluation, encouraged by the continuous decrease in genotyping costs despite the substantial increase in the density of commercial single nucleotide polymorphism (SNP) marker panels. Currently, genome wide association studies and genomic selection are often conducted using purebred populations. Estimation and often validation of SNP are carried out using a select elite set of pure bred animals (i.e. proven sires). This process was successful when estimated SNP effects were used to predict genomic breeding values (or pseudo-phenotypes) on animals of the same breed. However, when these SNP estimates are used for genomic prediction in other breeds or crossbred animals, it fails at different degrees depending on the genetic similarity between breeds in the mixture (Pryce et al., 2011). Unfortunately, this situation is not rare in several segments of livestock industry (beef cattle, swine or poultry) where the traits of interest are measured in crossbred or mixed populations with uncertain breed composition (Toosi et al., 2010). The main reasons that genomic selection

is not as successful when predicting genetic merit in admixed or crossbred populations are the change in linkage disequilibrium (LD) between markers and QTLs, inconsistency of linkage phase across subpopulations, and variation in allele frequencies between breeds (De Roos et al., 2008; Kizilkaya et al., 2010).

Accuracy of genome wide evaluation methods crucially depends on the extent of LD between markers and QTLs as well as the size of the reference population (De Roos et al., 2009; Lund et al., 2011; Brondum et al., 2011). Availability of large enough reference population is not always possible especially for breeds with limited number of genotyped and phenotyped animals (VanRaden et al., 2009; Hayes et al., 2009). To deal with these limitations, one plausible solution is to pool data from different breeds; thus creating a large enough reference population. Although this approach will resolve or at least alleviate the lack of power due to limited size of the training population, it intrinsically assumes that the SNP effects are constant across all breeds in the admixed population. This assumption is seldom true due to changes in several population parameters such as minor allele frequencies, strength of LD between markers and QTLs and linkage phases across sub-populations. Several simulation and real data studies have been conducted to evaluate the adequacy of different pooling strategies for the training and validation sets (Toosi et al., 2009; Daetwyler et al., 2012; Olsen et al., 2012). Their results were mixed and even contradictory. In general, prediction accuracy increased when subpopulations are genetically close and decreases as the genetic distance between components of the admixed population increased. More recently, (Kachman et al., 2013) showed that using a multi-breed training population did not increase prediction accuracies compared to single breed analysis when reasonable number of animals are available in each breed. However, prediction accuracy increased for breeds with small number of genotyped animals.

Given the limitations of the pooled data approach, several other methods have been proposed. These methods can be clustered into two broad groups based on their mode of accommodating differences between breeds; either through SNP effects or the genomic relationship matrix. Ibanez-Escriche et al. (2009) proposed a method where marker effects were estimated based on their population of origin. Unfortunately, this method was not successful for high density SNP panels. Karoui et al. (2012) proposed using a multi-breed training population that accounts for the difference in genetic correlations between breeds. Their results showed little to no increase in accuracy compared to the classical data pooling approach. Through modifications to the genomic relationship matrix, Harris and Johnson (2010) proposed a generalization of the regression technique used to derive the relationship matrix and Makghlela et al. (2012) adopted a random regression type approach that account for breed proportions in the population which performed slightly better than models ignoring breed-specific effects. In plant breeding, Shculz-Streeck et al. (2012) proposed a model that combines marker main effects that are consistent across sub-populations and population-specific marker effects. Although in general their results showed a slight increase in accuracy using population specific marker model compared to main marker effects model, however there are some cases in which population specific model performed worse than main marker effects model.

It is clear that current approaches for dealing with admixed and crossbred populations in genomic selection are far from providing a global answer of this relevant issue. Their results are data dependent and could lead to reduction in accuracies for animals in the pure breed populations. The objective of this study is to develop a model where the effect of an SNP could be different between breeds or lines and parameterized as a function of its effect on one of the breeds in pooled population through a one to one mapping function.

Materials and Methods

SNP effects often change between breeds or crossbred groups due to several factors including, change of minor allele frequency, strength of LD between markers-QTLs, and linkage phase between marker and QTL alleles. From hereafter we will refer to breeds or crossbred groups (F1, F2, etc.) in an admixed population as lines. Our idea is based on the possibility of inferring change in SNP effects between lines as a function of their genetic similarity. Our hypothesis is that the genetic similarity between two lines could be either directly modeled through a one to one mapping function between SNP effects or indirectly by using information already available in the SNP genotype data. Without loss of generality, assuming an admixture population with two lines, the pooled data approach postulates that the SNP effects to be constant across lines which is seldom true and depends on the genetic similarity between the lines. Using such approach the combined data will be analyzed using the following model:

$$y_{ij} = \mu + \sum_{k=1}^{p} x_{ik} g_k + e_{ij} \qquad [1]$$

where y_{ij} is the phenotype (or pseudo-phenotype) for animal *i* in line *j* (*j*=1,2), μ is the overall mean, x_{ik} is the genotype for animal *i* at locus k $(k = 1,2,..., p)$, g_k is the k^{th} SNP effect, and e_{ij} is the residual term. A more realistic model will be to assume different SNP effects between the two lines. Assuming that animal *i* belong to breed 1 and animal *j* is a member of breed 2, and both animals were genotyped for the same set of SNP markers. Their phonotypes could be modeled as:

$$y_{i1} = \mu + \sum_{k=1}^{p} X_{ik} g_k + e_{i1} \qquad [2]$$

$$y_{j2} = \mu + \sum_{k=1}^{p} X_{ik} g_k^* + e_{i2} \qquad [3]$$

where $\mathbf{y}_1 = (y_{11}, y_{12}, ..., y_{n1})$ and $\mathbf{y}_2 = (y_{12}, y_{22}, ..., y_{n2})$ are the vectors of observations for lines 1 and 2, respectively, g_k and g_k^* are the effects of the k^{th} SNP effects in lines 1 and 2, respectively.

Furthermore, g_k^* can be written as a linear function of g_k

$$g_k^* = \alpha_k g_k \qquad [4]$$

and equation in [3] becomes

$$y_{i2} = \mu + \sum_{k=1}^{p} X_{ik} (\alpha_k g_k) + e_{i2} \qquad [5]$$

where α_k is an unknown real number indicating the similarity of the effect of SNP *k* between the two lines with:

$$\alpha_k \begin{cases} = 1 \text{ SNP has the same effect in both lines} \\ = 0 \text{ SNP has no effect in line 2} \\ > 0 \text{ change in LD strenght without change in phase} \\ < 0 \text{ change in phase and strenght} \end{cases}$$

Consequently, model in equations [2] and [5] can be rewritten in matrix notation as:

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{X}^* \mathbf{g} + \mathbf{e} \qquad [6]$$

where **y** is the vector of observations for both lines, **g** is the vector of SNP effects in line 1; $\mathbf{X}^*$ is a modified matrix of SNP genotypes where the elements in rows corresponding to individuals in line 2 are multiplied by their respective α_k. If all α_k are equal to one, the matrix $\mathbf{X}^*$ will be identical to the original matrix of SNP genotypes, **X**, as is the case in the pooled data approach.

If the vector $\boldsymbol{\alpha}$ is known, the implementation of model in [5] is straightforward using any of the existing methods for genome wide association. Unfortunately, the vector $\boldsymbol{\alpha}$ is unknown and the model in [6] is not fully identifiable. Thus, $\boldsymbol{\alpha}$ and $\mathbf{g}$ cannot be uniquely estimated. In order to deal with this non-identifiability of the model, a hierarchical Bayesian approach was adopted.

In the first stage of the hierarchy, the conditional distribution of the data given the parameters of the model was assumed to be normal

$$\mathbf{y} \mid \mu, \mathbf{X}, \boldsymbol{\alpha}, \mathbf{g}, \sigma_e^2 \sim N(\mu \mathbf{1}_n + \mathbf{X}^* \mathbf{g}, \mathbf{I}_n \sigma_e^2)$$

In the second stage, the following priors will be assumed for the model parameters

$$\mu \sim cte$$
$$g_i \sim N(0, \sigma_i^2)$$
$$\sigma_e^2 \sim \chi^{-2}(\upsilon, s_0^2)$$

These priors will lead to an implementation similar to BayesA. Although our main interest is to estimate the SNP effects in both lines, the fact that the vector $\boldsymbol{\alpha}$ was assumed to be unknown preclude us from using the classical flat or uniform prior for the later which will lead to a non-identifiable model. To overcome this problem an informative prior for $\boldsymbol{\alpha}$ is needed. The following hierarchical prior was assumed:

$$\begin{aligned} &\alpha_i \mid \lambda_i^2 \sim N(\eta_i, \lambda_i^2) \\ &\lambda_i^2 \sim \chi^{-2}(1, s_i^2) \\ &s_i^2 \sim Gamma(0.5, z_i) \\ &p(\log(z_i) \sim cte \end{aligned} \qquad [7]$$

The hierarchical prior in [7] indicates that α_i follows a mixture of normal distributions with unknown variances and that the later follows a half-Cauchy prior distribution. Other than

specifying the mean, η_i, all remaining hyper-parameters are intrinsically estimated and require no user intervention. A reasonable assumption in livestock applications is to set η_i equal to 1 and let the data modify that prior believe based on the level of genetic dissimilarity between the lines (breeds) considered in the admixed population.

Finally, the hierarchy is finalized by specifying prior for the SNP variance

$$\sigma_i^2 \sim \chi^{-2}(\upsilon_a, s_a^2)$$

The implementation of the proposed hierarchical model is straightforward as all conditional distributions are in closed form being normal for the position parameters and scaled inverted Chi square for the dispersion components. The estimates of the SNP effects in line 2 are obtained as a by-product of the sampling process of SNP effects in line 1, **g,** and the elements in the vector **α** as indicated in equation [4].

In order to assess the adequacy of the proposed method, simulated and real data sets were used. Performance of the proposed method was evaluated based on the accuracy of the estimated molecular breeding values (MBV) calculated as:

$$MBV_i = \mu + \sum_{j=1}^{p} x_{ij}\alpha_{ij}g_j \qquad [8]$$

For simulated data, accuracy was computed as the correlation between the MBV and the true BVs. Using real data, a fivefold cross validation was carried out based on the correlation between the estimated BVs and MBVs.

It is worth mentioning that when more lines are included in the admixed population, the process presented above is still valid and the only modification is to add an extra vector **α** for

each additional line. SNP effects for additional lines will be estimated as indicated in equation [4]. The proposed method was implemented and evaluated using simulated data sets.

Simulation

A real data based simulation was carried out. The population consisted of two lines with 2,799 animals (1,989 in the first line and 810 in the second line) and a pedigree file of 2,799 animals. Genomic data consisted of SNP genotypes for 300 SNPs. SNP effects, **g**, for the first line were sampled from a normal distribution with mean zero and variance σ_g^2 equal to 0.01. The true genetic merit of each animal was computed as the sum (over all SNPs) of the product between each SNP effect and its associated genotype ($\sum_{k=1}^{300} x_{ik} g_k$). Phenotypes were simulated by adding an error term to the genetic merit. The error terms were simulation for normal distribution with mean equal to zero and variance calculated based on a heritability of the trait being either 0.3 or 0.5.

To create divergence between the two lines, SNP effects of the second line were generated as the product between their counterparts in line 1 and a vector of constants, **α**. Three distributional forms were assumed to generate **α**: 1) Uniform distributions, 2) normal distributions, and 3) mixture of normal and degenerative distributions. Hyper-parameters of these distributions control the level of similarity (divergence) between the two lines.

Results and Discussion

In order to establish a base for comparison, the two lines were analyzed separately. When training and validation were conducted within line, accuracy was 0.47 and 0.29 when heritability was equal to 0.3 and 0.67 and 0.40 when heritability was equal to 0.50, for lines A and B, respectively. When validation was conducted in the line that was not used in the training, the

accuracy was -0.01 for line B and from -0.16 to -0.06 for line A (Table 3.1) when α was sampled from U[-2,2] and similar results were observed for the other two intervals (results not shown). These results are well in line with those reported in the literature (Hayes et al., 2009; Kachman et al., 2013). Correlations between TBV and estimated MBV using pooled data (A+B) for training when heritability was equal to 0.30 are presented in Table 3.2. Using uniform distributions to create divergence between lines, the correlation decreased with the increase in the variability of **α**. This is expected because the further **α** deviates from 1 the smaller the genetic similarity is between the components of the admixed population. In fact, when **α** was sampled from U[-2.2], accuracy for line A (B) was 0.40 (0.23), it decreased to 0.39 (0.21) when **α** was sampled from U[-4,4] and then a larger decrease when **α** was sampled form U[-8,8]. The divergence created using the specified uniform distributions is large especially when **α** was sampled from either a U[-4,4] or a U[-8,8]. When **α** was sampled from a normal distribution N(1,0.01) or an admixture of a normal distribution N(1,0.01) and a degenerative distribution on 1, the pooled data approach yielded results similar to those obtained using within line analysis for line A and a substantial increase for line B (Table 3.2). These results are not surprising given the small variance of the normal distribution used to simulate **α** (0.01). In fact, a close inspection of the simulated values for **α** revealed that they are very close to 1 (Figure 3.1) indicating little to no divergence between the two lines. Thus, pooling the data of both lines in this case will increase power with little to no bias on the estimation of the SNP effects given the limited divergence between lines. When a normal distribution with variance equal to 0.05 was used for simulating **α**, more divergence was created between the two lines (Figure 3.1) leading to a small decrease in accuracies (Table 3.2). However, the divergence was not large enough to affect the accuracies obtained using the pooled approach. Using the proposed procedure when **α** was sampled from normal distributions showed

similar trend to the pooled approach with an additional again of 4-6% (Table 3.2). The same trend was observed when the heritability was equal to 0.5 and $\boldsymbol{\alpha}$ was sampled from uniform distributions with different bounds (Table 3.3). As expected there has been an increase in accuracies across all simulation scenarios due to the increase in heritability. More importantly, the proposed procedure yield results similar to those obtained using the within line analysis for line A and a slight increase of accuracies for line B except the case when $\boldsymbol{\alpha}$ was sampled form U[-8,8]. Better results were observed for line B, due to its small size (810 records) thus benefiting more from the increased power using our proposed compartment model.

As the two lines diverge, LD profiles are likely to differ or even breakdown. Additionally, LD phases between some markers and QTLs may be reversed across lines (de Roos et al., 2008; Pryce et al., 2011). Under all simulation scenarios when $\boldsymbol{\alpha}$ was sampled from uniform distributions with wide ranges, data pooling approach has resulted in lower accuracies in both lines compared to the within line analyses as indicated in Table 3.5. In fact, accuracies dropped by 14 to 19% for line A and 20 to 37% for line B depending of the bounds of the uniform distribution when heritability was equal to 0.3 and 11 to 20 for line A and 25 to 35% for line B when heritability was equal to 0.5. For the same comparisons and using our proposed method, there has been little to no drop in accuracies for line A when heritability was equal 0.3 (0 to 2%) or 0.5 (0 to 13%). However for line B, our proposed method led in general to a significant increase in accuracies ranging 13 to 17% when heritability was equal to 0.3 and from -2 to 5% when heritability was equal to 0.5. As indicated before, when $\boldsymbol{\alpha}$ was sampled from normal distributions with small variances and heritability was equal to 0.3, both the pooling data approach and our proposed procedure have led to an increase in accuracies compared to the within breed analyses with a slight superiority of the multi-compartment model (Table 3.2). The

results of this simulation are of practical importance and could shed a light into the discrepancies of results reported in the literature regarding the performance of the pooled data approach. In fact, Kachman et al. (2013) and Weber et al. (2012) reported that the pooled data approach leads to a decrease in accuracies for the components of the admixed population, whereas Hayes et al. 2009, Lund et al. 2011 and Brondum et al. 2011 reported that some component of the admixed population could see their accuracy increased using the pooled data approach. Based on our results it is likely that these contradictory conclusions could be both true depending on the level of divergence between the components of the admixed populations used in these studies. For highly divergent lines, the pooled data approach will likely lead to a decrease in accuracies, especially for the components of the admixed populations with small number of records. As the divergence between lines decreases, the accuracies obtained using the pooled data method will approach those obtained using the within line analyses and even leads to increase in accuracy when the lines are genetically close.

In practice, one of the reasons for pooling data from different lines using either the classical approach or our proposed procedure is to gain power. When the components of admixture are genetically close, the pooling approach is recommended, independently of the size of each line, and it will lead to an increase of accuracies. However, when the lines are genetically dissimilar data pooling is justified only if all or some components of the admixture are of a small size that precludes a within line analysis. To investigate the performance of the proposed method compared to classical data pooling approach, different sample size for two divergent lines was simulated. When line B has only 400 observations, the proposed method was 15% superior to classical data pooling approach for line A and 60% for line B (Table 3.4). Even when both lines

have limited number of records (810 observations per line), the proposed approach performed better with an increase of accuracies of 14% for line A and 50% for line B.

The advantage of the proposed method over the pooled data based method is that our method does not assume constant SNP effects across the two lines. The assumption of constant SNP effects across sub-population is seldom true, due to change in several parameters as mentioned earlier.

Conclusions

Pooling data from lines or breeds in the training set when conducting genome wide evaluation studies seems an attractive approach since it benefits from the increase in power. Its performance is variable and depends largely on the genetic similarity between the sub-populations in the mixture. When the sub-populations are very close genetically, the pooled data approach even in its basic form will result in an increase of accuracies, especially for the lines with limited recording. As the genetic similarity between lines decreases, the classical pooled data approach becomes inefficient with substantial decrease in accuracies for all components of the admixed populations. Thus, in such scenario it is not recommended. However, the proposed multi-compartment model and based on the simulation results is clearly superior as it allows systematically for the accounting of the difference in SNP effects across divergent lines. Its superiority compared to the pooled data approach ranged from approximately 17 to 47% and increases as the divergence between lines increases. Independently of the genetic similarity between lines, the pooled data approach is justified only when not enough data is available for each of the components of the admixed population to conduct within line analyses. The current simulation parameters do not reflect the actual SNPs density in commercially used panels. Thus,

it is needed that the performance of the proposed model be evaluated when large numbers of SNPs are genotyped. Additionally, the proposed method should be evaluated in presence of more than two lines and/or crossbreed animals.

References

Bennett, B. J., Farber, C. R., Orozco, L., Kang, H. M., Ghazalpour, A., Siemers, N., Neubauer, M., Neuhaus, I.Yordanova, R., Guan, B., Truong, A., Yang, W. P., He, A., Kayne, P., Gargalovic, P., Kirchgessner, T., Pan, C., Castellani, L. W., Kostem, E., Furlotte, N., Drake, T. A., Eskin, E. & Lusis, A. J. 2010. A high resolution association mapping panel for the dissection of complex traits in mice. Genome Research. 20, 281–290.

Bennewitz, J., Solberg, T., Meuwissen, T., 2009. Genomic breeding value estimation using nonparametric additive regression models. *Genet Sel Evol*, 41:20

Bolormaa, S., Pryce, J. E., Hayes, B. J., Goddard, M. E., 2010. Multivariate analysis of a genome-wide association study in dairy cattle. J.Dairy Sci 93, 3818–3833.

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, 94:4700–4707.

Daetwyler, H. D., Villanueva, B., Woolliams, J.A., 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3(10): e3395.

Daetwyler, H. D., Kemper, K. E., Van der Werf, J.H.J., Hayes, B.J. 2012. Components of the Accuracy of Genomic Prediction in a Multi-Breed Sheep population. J. Anim Sci. 90, pp. 3375–3384

De Roos, A. P. W., Hayes B. J., Goddard M.E. 2009. Reliability of genomic predictions across multiple populations. Genetics 183: 1545-1553

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

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

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

Harris, B.L, Johnson, D.L., 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J Dairy Sci, 93:1243–1252.

Harris, B.L., Johnson D.L, Spelman R.J., 2008. Genomic selection in New Zealand and the implications for national evaluation. Proceedings of the Interbull Meeting, Niagra Falls, Canada.

Hayes, B., Bowman, P. J., Chamberlain, A.C., Verbyla, K., Goddard, M.E., 2009. Accuracy of genomic breeding values in multiple dairy cattle populations. Genet Sel Evol 41:51.

Ibanez-Escriche, N.,Fernando., R.L., Toosi, A., Dekkers., J.C.M., 2009. Genomic selection of purebreds for crossbred performance. Genet Sel Evol, 41:12

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

Karoui, S., Carabano, M.J., Diaz, C., Legarra. A., 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol, 44-39.

Kizilkaya, K., Fernando, R.L., Garrick, D.J., 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88:544-551.

Lund, M.S., de Roos, A.P.W., de Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., Su, G., 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol, 43:43.

Makgahlela, M.L., Mäntysaari, E.A., Strandén, I., Koivula, M., Nielsen, U.S., Sillanpää, M.J., Juga, J., 2012. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J Anim Breed Genet, doi:10.1111.

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

Olson, K.M., VanRaden, P.M., Tooker, M.E., 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 95:5378-5383

Pryce, J.E., Gredler, B., Bolormaa, S., Bowman, P.J., Egger-Danner, C., Furest, C., Emmerling, R., Solkner, J., Goddard, M.E., Hayes, B.J., 2011. Genomic selection using a multi-breed, across country reference population. J Dairy Sci. 94:2625-2630

Schulz-Streeck, T., Ogutu, J.O. and Karaman, Z., Knaak, C., Piepho, H.P., 2012 Genomic Selection using Multiple Populations. Crop Science, 52 (6). pp. 2453-2461.

Snelling, W. M., Allan, M.F., Keele, J.W., Kuehn, L.A., McDaneld, T., Smith, P.L., Sonstegard, T.S., Thallman, R.M., Bennett,G.L., 2010. Genome-wide association study of growth in crossbred beef cattle. J. Anim. Sci. 88:837–848.

Su. G., Christensen, O.F., Ostersen, T., Henryon, M., Lund, M.S., 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLOS one. DOI: 10.1371.

Toosi, A., Fernando, R.L., Dekkers, J.C.M., 2009. Genomic selection in admixed and crossbred populations. J. Anim Sci. 2009. 88:32-46.

VanRaden, P. M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., 2009. Invited review: Reliability of genomic predictions for North American Hollstein bulls. J. Dairy Sci. 92:16-24.

Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J., 2012. Five years of GWAS discovery. Am. J Hum Genet, 90:7-24.

Weber, K.L., Thallman, R.M., Keele, J.W., Snelling, W.M., Bennett, G.L., Smith, T.P.L., McDaneld, T.G., Allan, Van Eenennaam, A.L., Kuehn, L.A., 2012. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes.

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

Xu, S., 2007. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics. 63:513–521

Table 3.1: Correlations between true and molecular breeding values using different training and validation datasets using heritability of 0.3 and 0.5 for lines A and B.

Training Data Set	Validation	h^2=0.3	h^2=0.5
A	A	0.47	0.67
B	B	0.29	0.40
A	B	-0.01	-0.01
B	A	-0.06	-0.16

Table 3.2: Correlations between true and molecular breeding values using pooled data method and multi-compartment model for heritability 0.3.

	Pooled data Model		Multi-compartment model	
Distribution of α	A	B	A	B
α ~ [-2,2]	0.40	0.23	0.47	0.34
α ~ [-4,4]	0.39	0.21	0.46	0.33
α ~ [-8,8]	0.38	0.18	0.46	0.33
α ~ N(1,0.01)	0.48	0.47	0.50	0.49
50% α ~ N(1,0.01) 50% α=1	0.48	0.46	0.48	0.46
α ~ N(1,0.05)	0.46	0.44	0.48	0.46

Table 3.3: Correlations between true and molecular breeding values using pooled data method and multi-compartment model for heritability 0.5.

	Pooled data Model		Multi-compartment model	
Distribution of α	A	B	A	B
α ~ [-2,2]	0.59	0.30	0.67	0.42
α ~ [-4,4]	0.56	0.28	0.60	0.42
α ~ [-8,8]	0.53	0.26	0.58	0.39

Table 3.4: Correlations between true and molecular breeding values using pooled data method and multi-compartment model for different population size in the case of α~ U [-2, 2] and a heritability of 0.3.

	Population Size: A=1989, B=400		Population Size: A=810, B=810	
	Pooled data Model	Multi-compartment model	Pooled data model	Multi-compartment model
A	0.40	0.46	0.34	0.39
B	0.10	0.16	0.21	0.32

Table 3.5: Gain (loss) in prediction accuracy using pooled data model and multi-compartment model

	h^2=0.3		h^2=0.5	
Lines	Pooled data model	Multi-compartment model	Pooled data model	Multi-compartment model
A	(-19%, -14%)	(-2%, -0%)	(-20%,-11%)	(-13%, -0%)
B	(-20%, -37%)	(+13, +17%)	(-35%, -25%)	(-2%, +5%)

Figure 3.1

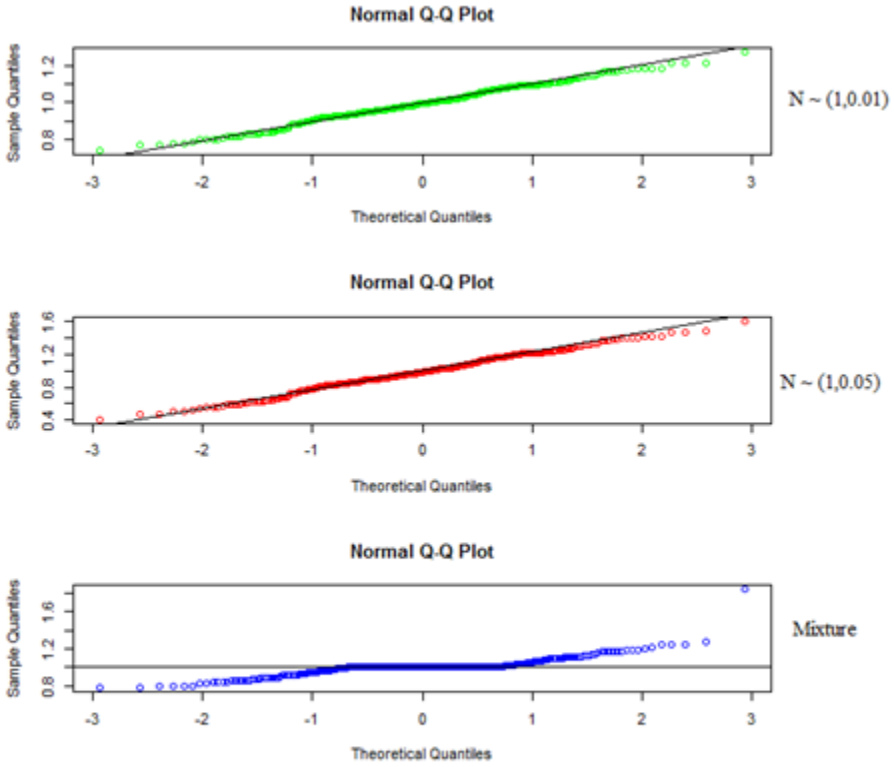

Figure 1: Q-Q plot of simulated α values from three distributions, N~(1,0.01), N~(1,0.05) and a mixture distribution (50% from N~(1,0.01), and 50% fixed to 1)

CHAPTER 4

A STRUCTURAL MODEL FOR GENETIC SIMILARITY IN GENOMIC SELECTION OF ADMIXED POPULATIONS

[1] Hay, E. and R. Rekaya. 2014. To be submitted.

Abstract

Current approaches for dealing with admixed and crossbred populations in genomic selection rely on using different groups of animals in training sets. These approaches benefit from increased power as a result of increasing the size of the training set. However, the performance largely depends on the genetic similarity between the sub-populations of the admixed population. Our proposed multi-compartment model where the effect of an SNP could be different between breeds and parameterized as a function of its effect on one of the breeds in admixed population through a one to one mapping function, was able to remediate some problems of the pooled data approaches but still suffers from the high dimensionality of the unknown parameters to estimate. To overcome this problem, we propose not to estimate the mapping parameter α for each SNP but rather to build a model for α as a function of information already available in the genotype data via a hierarchical structural model. In this study, α was modeled as a function of the change in linkage disequilibrium. An admixed population (A and B) and crossbred populations (AB, BA, BxAB, BxBAB) were simulated. Individuals were genotyped for 300 SNPs and measured for a quantitative trait with 0.5 heritability. Three analyses were conducted: 1) classical pooled data (**M1**); 2) pooled data using the multi-compartment model and α for each SNP (**M2**); and 3) pooled data using multi-compartment model and our structural model for α (**M3**). The accuracy of (**M1**) tended to be much lower than using models **M2** or **M3**. The prediction accuracies for line A using model **M1** was 0.39 compared to 0.56 and 0.43 using **M2** and **M3,** respectively. The accuracies using the structural model (**M3**) resulted in intermediate to those obtained using **M1** and **M2.** The relatively good performance obtained using M2 indicates that it is possible to model α as a function of the information already available in the genotype and to substantially reduce the number of parameters to be estimated.

Key Words**:** genomic selection, admixed population

Introduction

Genomic selection benefit is a more accurate pre-selection of animals that inherited genes or chromosome segments of superior merit (Meuwissen et al., 2001). In livestock, genomic selection is becoming a routine technique, mainly due to the decreasing cost of genotyping for large number of single nucleotide polymorphism (SNP) markers. Genomic selection uses markers that are in linkage disequilibrium with QTLs to estimate breeding values. Currently, genomic selection is conducted on purebreds, and training and validating on such data is successful. However, when training on purebreds and validating on admixed or crossbred animals, this method fails at different degrees depending on the genetic similarity between breeds in the mixture (Habier et al., 2010; Wientjes et al. 2013). The main reason genomic selection is not as successful when predicting genetic merit of admixed or crossbred animals is the change in linkage disequilibrium (LD), linkage phase, and allele frequencies between breeds (De Roos et al., 2008).

Accuracy of genome wide evaluation methods crucially depends on the extent of LD between markers and QTLs (De Roos et al., 2009) as well as the size of the reference population (Goddard, 2009). Availability of large reference population is not always guaranteed especially for breeds with limited number of genotyped and phenotyped animals (VanRaden et al., 2009; Hayes et al., 2009). A plausible solution is the pooled data approach where multiple breed data is pooled to create a large reference population (Lund et al., 2011; Brøndum et al., 2011; Kizilkaya et al. 2010). Intrinsically, the pooled data method assumes constant SNP effects across breeds. This assumption is seldom true due to changes in several parameters such as minor allele frequency, strength of LD between markers and QTLs and linkage phases across sub-populations. Several simulation and real data studies have been conducted to evaluate the accuracy of using different training population pooling strategies (Toosi et al., 2009; Heringstad

et al., 2011; Daetwyler et al., 2012; Olsen et al., 2012; Zhou et el., 2014a). Their results showed an increase in prediction accuracy when subpopulations are genetically close. Kachman et al. (2013) showed that using a multi-breed training population did not increase prediction accuracies than single breed training population when the breed had a reasonable number of animals; however the prediction accuracy increased for breeds with small number of genotyped animals. Similar results have been reported in dairy cattle (Pryce et al., 2011; Hoze et al., 2014). Furthermore, accuracies tend to drop when breeds other than the one used in the validation were included in the training set (Toosi et al., 2009; Hayes et al., 2009; Ibanez-Escriche et al., 2009). More recently, Hay and Rekaya (2014) presented a multi-compartment model to analyze genomic information with pooled multi-breed data. Although their approach yielded better accuracies in general, it suffers from the high dimentionality of the model and numerical instabilities when the number of SNPs is large. This is due to fact that an additional unknown parameter is estimated for each SNP in the panel. Numerical instabilities occur when the effect of an SNP in one line is very small (close to zero) which in turn leads to an estimate of the mapping parameter that tends towards infinity.

The objective of the present study is to expand the multi-compartment approach for admixed populations presented by Hay and Rekaya (2014) through the reduction of the dimensionality of the model and the elimination of the numerical instabilities. This will be achieved by modeling the mapping parameter as a function of some characteristics of the observed SNP genotypes.

Materials and Methods

Statistical method

Multi-compartmental model for genomic selection in admixed populations:

The model proposed by Hay and Rekaya (2014) postulates that changes in SNP effects between breeds or crossbred groups can be accommodates through a one to one mapping function. Such function allows for the expression of SNP effects for all components of the mixture population as a function of the estimates of those SNPs in only one of the breeds (lines). Although such model showed a clear superiority compared to the pooled data approach using relatively small number of SNPs, it suffers for high dimentionality and numerical instabilities. This is true especially when high density SNP panels are considered. Similarly to other proposed methods for dealing with admixed populations, the approach proposed by Hay and Rekaya (2014) does not allow for the estimation of genomic breeding values for non-phenotyped sub-populations. To deal with these issues, we hypothesize that the changes in SNP effects between the components of an admixed population could be inferred based on criteria already available in the observed SNP marker genotypes.

The multi-compartment model presented by Hay and Rekaya (2014), assuming an admixture population of two lines, postulates that any genetically heterogeneous pooled data set is generated by different data generating processes governed primarily by the SNP genotypes and the effects of associated QTLs. Thus, a simple model to reflect such reality will consists of as many compartments as number of components in mixture. Following Hay and Rekaya (2014), the model could be presented as:

$$y_{i1} = \mu + \sum_{k=1}^{p} X_{ik} g_k + e_{i1} \qquad [1]$$

$$y_{j2} = \mu + \sum_{k=1}^{p} X_{ik} g_k^* + e_{i2} \qquad [2]$$

where $\mathbf{y}_1 = (y_{11}, y_{12}, ..., y_{n1})$ and $\mathbf{y}_2 = (y_{12}, y_{22}, ..., y_{n2})$ are the vectors of observations for lines 1 and 2, respectively. μ is the overall mean, x_{ik} is the genotype for animal *i* at locus k $(k = 1,2,..., p)$, g_k and g_k^* are the effects of the k^{th} SNP effects in lines 1 and 2, respectively, and e_{ij} is the residual term.

Furthermore, g_k^* can be written as a linear function of g_k

$$g_k^* = \alpha_k g_k \qquad [3]$$

and equation in [2] becomes

$$y_{i2} = \mu + \sum_{k=1}^{p} X_{ik} (\alpha_k g_k) + e_{i2} \qquad [4]$$

where α_k is an unknown real number indicating the similarity of the effect of SNP *k* between the two lines.

From the presentation in equations [2] and [4], it is clear that a parameter alpha must be estimated for each SNP. Furthermore, such parameter indirectly captures the change of LD between the marker and the potential associated QTL(s). Thus, it is reasonable to postulate that such change of LD between the SNP markers and QTLs (ΔLD_{M-Q}) could be predicted or at least approximated using change in LD structures across markers (ΔLD_{M-M}) between different components of an admixed population. Although the relationship between ΔLD_{M-Q} and ΔLD_{M-M} could be complex (non-linear), some heuristically defined models could be developed. In this study, the following model was used to model the relation between of between ΔLD_{M-Q} and ΔLD_{M-M}.

$$(\Delta LD\_M\_Q)_k = a_0 + a_1 m_k + a_2 s_k \qquad [5]$$

where $(\Delta LD\_M\_Q)_k$ α_k is the change in LD between SNP marker *k* and the linked QTL(s) across the two lines, m_k and s_k are the mean and standard deviation of the difference of LD between marker-marker in the two lines respectively.

Using small simulated data sets where marker and QTL genotypes were known for two divergent lines (see next chapter), the model presented in equation [5] was able to predict $(\Delta LD\_M\_Q)_k$ with sufficient accuracy. In fact, the R^2 of the model ranged from 0.54 to 0.67. Although the prediction was not perfect, it is high enough to warrant its consideration to model the change in alpha. Additionally, modeling alpha using equation [5] will reduce the dimentionality of the model considerably. In fact, rather than estimating one alpha for each SNP, only three parameters (a_0, a_1, and a_2) will be needed using this new parametrization. Furthermore, modeling alpha using equation [5] will provide an easy and straightforward way to predict genomic breeding values for non-phenotyped lines or crossbred groups and it suffice to calculate m_k and s_k.

As indicated earlier, the parameter alpha in equation [4] tries to directly model the change of LD between markers and QTLs across breed or lines. Thus, it seems reasonable to assume equation [5] could be used to model alpha, leading to:

$$\alpha_k = a_0 + a_1 m_k + a_2 s_k \qquad [6]$$

And equation in [4] can be rewritten as:

$$y_{i2} = \mu + \sum_{k=1}^{p} x_{ik} (a_0 + a_1 m + a_2 s) g_k + e_{i2} \qquad [7]$$

In order to evaluate the adequacy of this new reparametrization, three models were implemented and compared using simulated data: 1) classical pooled data (**M1**), multi-compartment model as presented by Hay and Rekaya (2014) where alpha was directly modeled (**M2**); and 3) the new model as presented in equation [5] where alpha is indirectly modeled (**M3**). All three models were compared based on the accuracy of the predicted genomic breeding values defined as the correlation between the latter and the true BVs. Additionally, the new method (**M3**) was compared with the classical pooled data approach (**M1**) in predicted genomic breeding values for non-phenotyped crossbred populations. The GEBVs were computed as follow:

$$GEBV_i = \sum_{k=1}^{m} x_{ki} \; g_k \qquad [8]$$

where x_{ki} is the genotype of animal i, g_k is the effect of genotype k

Simulation

QMSim software (Sargolzaei and Schenkel. 2009) was used for data simulation. A randomly mated historical population was generated and used as a base population in order to create two pure divergent lines (A and B) with 16,790 and 16,776 animals, respectively. A genome of 100cM in length and harboring 300 evenly spaced SNP markers and 3 QTLs was simulated. Minor allele frequency of simulated markers was greater or equal to 0.05. QTL additive effects were sampled from a gamma distribution with shape and scale parameter equal to 0.4. Phenotypes were simulated based on a heritability of 0.5. A descriptive summary of the simulated data and genotypes is presented in Table 4.1.

Several crosses were generated using selected sets of males and females from the two lines. Reciprocal F1 crosses (AB and BA) were generated using 200 males from the first line and 4000 females from the second. Similarly, two backcrosses (BxAB; and BxBAB) were generated

using the same approach used for F1 crosses, except that only 2,000 and 1,000 females were used, respectively.

Linkage disequilibrium between different populations was calculated and then used in the implementation of the structural model in **M3**. Linkage disequilibrium between pair of SNP markers was computed using the r^2 coefficient which is a statistical measure of the correlation between a pair of loci.

Results and Discussion

To provide a basis for comparison and to evaluate the performance of the proposed model (**M3**), correlations between estimated and true breeding values when training and validating on the same population were calculated and are shown in Table 4.2. Accuracies were reasonably high for all populations as expected in the case of training and validating on the same population. Accuracies were slightly higher for pure lines A and B (0.85 and 0.88) then crossbred populations due to a larger number of individuals in the pure lines. In order to test performance of the multi-compartment model (**M2**) and the structural model (**M3**) compared to the pooled data approach (**M1**) using a multi-breed population, a reference population comprised of the two divergent lines A and B was used. As shown in Table 4.2, when pooling both lines without accounting for differences between sub-populations (**M1**) accuracies tended to be much lower than using models **M2** or **M3**. The prediction accuracies for line A using model **M1** was 0.39 compared to 0.56 and 0.43 using **M2** and **M3,** respectively. The accuracies using the structural model (**M3**) resulted in intermediate performance between **M1** and **M2.** This is very likely due to the fact that the model used to explain the change in LD between marker and QTLs has an R^2 smaller than one. However, a substantial increase in accuracy was still seen compared to **M1**. In

fact, using **M3** resulted in an increase in prediction accuracy compared to **M1** by 10% and 18% for lines A and B, respectively. The largest increase in prediction accuracy, as expected, was for the multi-compartment model (**M2**). In fact, for line A the accuracy increased by 43% (from 0.39 to 0.56) and 62% (from 0.32 to 0.52) for line B compared to **M1**. Comparing the two models which account for differences between the two sub-populations (**M2** and **M3)**, the multi-compartment model (**M2**) has results in 30% and 36% superiority for lines A and B, respectively.

These results indicate that when sufficient number of observations are available in the training set, as it is the case in this study, and sufficient genetic dissimilarity exists between the sub-population, pooled data approaches will result in a decrease of accuracies for some components of the population as it was observed in previous studies (Hayes et al., 2009; Erbe et al., 2012; Kachman et al., 2013). However, the rate of loss of accuracies depends on the ability of the approach to accommodate the genetic differences between components of the population. It is clear that **M2** has a better handle of the genetic dissimilarity. However, when the size of one or all sub-populations is limited, pooled data approached will often result in an increase of accuracy for at least some components of the population. Such increase is a function of the genetic similarity. As clearly shown in Hay and Rekaya (2014), **M2** will result in a better performance even in the presence of limited data and extensive genetic dissimilarity

One of the limitations of **M1** and **M2** is their inability of predict accurate GEBV for non-phenotyped populations. To test the ability of the proposed method (**M3**) to deal with this issue, different genotyped but non-phenotyped crossbred populations were used as validation sets to mimic real applications, in the sense that phenotypes are not always available for crossbred or commercial animals. As shown in Table 4.4 prediction accuracies are substantially higher for

M3 compared to **M1** across all validation populations. For example, validating on crossbred population (AB) using **M1** resulted in an accuracy of 0.11 compared to 0.28 when using model **M3**. The same behavior was seen across all crossbred populations, with a substantial superiority of model **M3** ranging from 71% to 154%. As expected, as the percentage of line B in the crossbred populations increased so did the prediction accuracy using either models **M1** or **M3**. This is due to fact that the pooled data of lines A and B were used in the training. Furthermore, for (BxBAB) cross (87.5% B), the accuracy (0.36) was very similar to the one obtained for line B (0.38) when using **M3** but was substantially smaller (0.21) compared to 0.32 (Table 3) using **M1**.

In general, although the proposed method resulted in a substantial increase in accuracies for the non-phenotyped sub-populations compared to **M1**, its performance is roughly 50% lower than the results obtained when training and validation were conducted within the same crossbred population. However, this could still be of substantial practical and commercial importance especially in situation where no other alternatives are available.

This study shows that using models which adjust for different sub-populations have a positive impact on the prediction accuracy of GEBV. Although the structural model (**M3**) did not perform in the same level as the multi-compartment model (**M2**), it did however increase prediction accuracy compared to the pooled data approach. This increase in prediction accuracy was most apparent when crossbred populations were considered as shown in Table 4.4. Pooling data approach could be beneficial when the sub-populations are genetically similar. In fact, Olsen et al. (2012) reported an increase in the accuracy of genomic prediction using a multiple breed reference population, especially for breeds with limited records. Further, Lund et al. (2011) and Brøndum et al. (2011) showed a notable increase in accuracy when pooling multiple breeds

in the reference population. Unfortunately, few other studies reported no benefit in genomic prediction accuracy when pooling breeds in the reference population (Kachman et al., 2013; Weber et al., 2012). However, as discussed earlier, the pooled data approach makes a strong assumption of constant SNP marker effects across sub-population (Hayes et al. 2009). The proposed model **M3** and model **M2** tend to relax such assumption. In all simulation scenarios, **M2** and **M3** resulted in higher prediction accuracies compared to **M1**. As presented in Table 4.2, an increase of 10% and 18% in accuracy was observed in line A and line B respectively using M3. Genomic selection is primarily conducted in purebred animals. It is of great importance to investigate the performance of genomic selection in non-phenotyped commercial animals, such as crossbreds and animals with unknown genetic composition. Ignoring differences in the genetic parameters (i.e. linkage disequilibrium) between crossbred populations could lead to low genomic prediction accuracy or spurious associations. It has been reported that linkage disequilibrium is stronger and extends over longer intervals in pure populations than outbred populations (Shifman et al., 2003; Lindblad-Toh et al., 2005). To investigate this, genomic prediction accuracy was computed validating on four crossbred populations. As Table 4.4 displays, using the pooled data approach (**M1**), resulted in low prediction accuracies for all four crossbred animals. Accounting for population structure by using model **M3**, which estimates the change of SNP marker effects using the difference in linkage disequilibrium between the different crossbred populations increased prediction accuracy substantially.

Conclusions

Using a model which allows for SNP marker effects to change across sub-populations did improve the prediction accuracy of GEBVs. Pooling data approaches are attractive methods since they benefit from the increase in power due to the increase of the size of the training set.

However, their performance is variable and depends largely on size and the genetic similarity between the sub-populations in the mixture. One of the limitations of current methods for genomic prediction in is their inability of predict accurate GEBV for non-phenotyped populations. Our proposed method resulted in a substantial increase in accuracies for the non-phenotyped sub-populations and it could be of substantial practical and commercial importance especially in situation where no other alternatives are available. Further, the presented structural model could be improved through the inclusion of other explanatory variables that could be calculated based on already available data.

References

Bolormaa, S., Pryce, J. E., Hayes, B. J., Goddard, M. E., 2010. Multivariate analysis of a genome-wide association study in dairy cattle. J.Dairy Sci 93, 3818–3833.

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, 94:4700–4707.

Daetwyler, H. D., Kemper, K. E., Van der Werf, J.H.J., Hayes, B.J. 2012. Components of the Accuracy of Genomic Prediction in a Multi-Breed Sheep population. J. Anim Sci. 90, pp. 3375–3384

De Roos, A. P. W., Hayes B. J., Goddard M.E. 2009. Reliability of genomic predictions across multiple populations. Genetics 183: 1545-1553

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

Erbe et al., M. Erbe, B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, Mason, M.E. Goddard. 2012. Improving accuracy of genomic predictions within and

between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels J. Dairy Sci., pp. 4114–4129

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

Habier, D., J. Tetens, F. Seefried, P. Lichtner, G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Select. Evol., 42, p. 5

Hayes, B., Bowman, P. J., Chamberlain, A.C., Verbyla, K., Goddard, M.E., 2009. Accuracy of genomic breeding values in multiple dairy cattle populations. Genet Sel Evol 41:51

Heringstad, B., G. Su, T.R. Solberg, B. Guldbrandtsen, M. Svendsen, M.S. Lund. 2011. Genomic predictions based on a joint reference population for Scandinavian red breeds. In Book of Abstracts of the 62nd Annual Meeting of the European Federation of Animal Science Waageningen Academic Publishers, Stavanger, Norway, p. 29.

Hoze, C., Fritz, S., Phocas, F., Boichard, D., Ducrocq, V., Croiseau, P., 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J. Dairy Sci. 97:3918-3929

Ibanez-Escriche, N.,Fernando., R.L., Toosi, A., Dekkers., J.C.M., 2009. Genomic selection of purebreds for crossbred performance. Genet Sel Evol, 41:12

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

Kizilkaya, K., Fernando, R.L., Garrick, D.J., 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88:544-551.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., et al., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Lund, M.S., de Roos, A.P.W., de Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., Su, G., 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol, 43:43.

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

Olson, K.M., VanRaden, P.M., Tooker, M.E., 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 95:5378-5383

Pryce, J.E., Gredler, B., Bolormaa, S., Bowman, P.J., Egger-Danner, C., Furest, C., Emmerling, R., Solkner, J., Goddard, M.E., Hayes, B.J., 2011. Genomic selection using a multi-breed, across country reference population. J Dairy Sci. 94:2625-2630

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

Shifman, S., J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi. 2003. Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. 12:771–776.

Toosi, A., Fernando, R.L., Dekkers, J.C.M., 2009. Genomic selection in admixed and crossbred populations. J. Anim Sci. 2009. 88:32-46.

VanRaden, P. M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., 2009. Invited review: Reliability of genomic predictions for North American Hollstein bulls. J. Dairy Sci. 92:16-24.

Weber, K.L., Thallman, R.M., Keele, J.W., Snelling, W.M., Bennett, G.L., Smith, T.P.L., McDaneld, T.G., Allan, Van Eenennaam, A.L., Kuehn, L.A., 2012. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes.

Wientjes, Y. C. J., Veerkamp, R.F., Calus M.P.L., 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics,, pp. 621–631

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. Meuwissen, M. Svendsen, H. Grove, U.S. Nielsen, M.S. Lund. 2014. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J. Dairy Sci. 10.3168.

Table 4.1. Simulation parameters

Parameter	
Population structure	Two divergent Lines (Line A, Line B)
Heritability	0.50
Phenotypic variance	1.0
QTL heritability	0.5
Genome size	100 cM
Number of Chromosomes	1
Number of markers	300
Minor allele frequency	0.05

Table 4.2: Accuracy[1] of genomic prediction using the pooled data (M1), the multi-compartment (M2) and the structural (M3) models

Lines	M1	M2	M3
A	0.39	0.56	0.43
B	0.32	0.52	0.38

[1]The accuracy was calculated as the correlation between the GEBV and EBV, averaged over five replications

Table 4.3: Genomic prediction accuracy[2] when training and validating are conducted on the same sub-population

Sub-population	Accuracy
A	0.85
B	0.88
(AB)	0.74
(BA)	0.75
(BxAB)	0.78
(BxBAB)	0.78

[1]The accuracy was calculated as the correlation between the GEBV and EBV, averaged over five replications

Table 4.4. Accuracy[1] of genomic prediction in different crossbred sub-populations (Line A and Line B were used for training)

Validation data set	Model estimating α	Pooled data method
(AB)	0.28	0.11
(BA)	0.32	0.16
(BxAB)	0.35	0.18
(BxBAB)	0.36	0.21

[1]The accuracy was calculated as the correlation between the GEBV and EBV, averaged over five replications

CHAPTER 5

USE OF OBSERVED GENOMIC INFORMATION TO INFER LINKAGE DISEQUILIBRIUM BETWEEN MARKERS AND QTLS

[1] Hay, E. and R. Rekaya. To be submitted

Abstract

Conducting genomic selection in admixed populations is challenging and its accuracy in this case largely depends on the persistence of linkage disequilibrium between markers and QTLs. Inferring linkage disequilibrium between markers and QTLs could be important in understanding the change of SNP marker effects across different breeds. Predicting the change in linkage disequilibrium between markers and QTLs across two divergent breeds was explored using information from the genotype data. Two different models (M1, M2) that differ in the definition of the explanatory variables were used to infer the level of LD between SNP markers and QTLs using all markers in the panel or windows of fixed number of markers. Three simulation scenarios were conducted using different number of SNPs and QTLs. In the first scenario, the resulting coefficient of determination (R^2) was 0.65 and 0.52 using M1 and M2, respectively. In the second scenario, average R^2 equaled 0.12 using all markers in the panel and 0.25 using 100 marker windows. Across the three simulation scenarios, it was clear that a significant portion of the variation in the change in LD between SNP markers and QTLs could be explained by information already available in the observed SNP marker data.

Key words: genomic selection, linkage disequilibrium, SNP.

Introduction

Genomic selection is a type of marker assisted selection which involves the estimation of genomic breeding values (GEBV) based on a large number of markers across the genome (Meuwissen et al., 2001). Genomic selection relies on the assumption that all relevant quantitative loci (QTL) are in linkage disequilibrium (LD) with genotyped SNP markers. Thus,

linkage disequilibrium or the non-random association of alleles at different loci (Hill and Robertson, 1968) across genotyped markers and between the later and QTLs will fundamentally condition the efficiency of the association analysis and it is of great importance in QTL mapping, genomic selection and genome wide association studies. Although the strength of LD between genotyped SNP markers is easy to calculate, inferring the level of LD between SNP markers and QTLs is a complex problem due to the unavailability of QTL genotypes in the majority of genomic association studies. Although the knowledge of the QTL(s) genotypes or their LD with SNP markers in the panel is not needed in association studies, such information could be of great interest in some applications such as multi-breed and crossbred genomic selection.

Genomic selection has been successful in prediction of genomic breeding values. However this success did not extend to admixed breeds or crossbreds. Several studies showed that the structure of the reference population strongly impacts the accuracy of genomic predictions (VanRaden et al., 2009; Hayes et al., 2009; Thomasen et al., 2012; Erbe et al., 2012). Moreover, SNP marker estimates derived from one breed have little to no predictive power of GEBVs of animals in a different breed (Pryce et al., 2011; Hayes et al., 2009). A potential solution would be to use a pooled multi-breed reference population to predict GEBV of animals in other breeds or crossbred animals (Heringstad et al., 2011; Daetwyler et al., 2012; Olsen et al., 2012; Zhou et al., 2013; Hoze et al., 2014). This method showed promising results in improving prediction accuracy in the case when a breed has a limited number of records. However, the performance of this approach, as expected, depends largely on the genetic similarity between components of the admixed population.

Although simple in its concept, the multi-breed reference population approach makes strong genetic and population structure assumptions. In its most basic formulation, it assumes a

genetically homogenous population where SNP marker effects are constant across sub-populations or breeds. Furthermore, it assumes that linkage disequilibrium (LD) between SNPs and QTLs is the same across the reference and validation populations. Although that is the case for within breed genomic selection, such assumption is often violated when breeds with different genetic structure and background are being considered. This genetic difference between breeds is manifested by varying allele frequencies for markers and QTLs, change in LD strength and structure, and linkage phase (Goddard et al., 2009; De Roos et al., 2008; Kizilkaya et al., 2010; Wientjes et al., 2013). Furthermore, several studies have evaluated LD blocks in various population structures and reported differences in the extent of LD. For example, Thomasen et al. (2012) reported differences in LD between the Danish Jersey population and the North American Jersey population. In addition, Shifman and Darvasi (2001) showed that LD was several folds higher in isolated population than outbred populations very likely due to higher inbreeding. Similarly, Lindbladtoh et al. (2005) reported, as expected, larger LD blocks within breeds than across breeds. Hay and Rekaya (2014a and 2014b) showed that accommodating the potential change in SNP effects between the different components of an admixed population, increased accuracies of genomic prediction. When change in SNP effects was directly modeled, substantial increase in accuracies was observed compared to the classical pooled data approach. Unfortunately, such model suffers from high dimensionality and numeral instability especially in presence for large number of SNPs. Their indirect approach to account for change in SNP effects was based on heuristically developed structural model using available information on marker genotypes. Although it remedies the problems associated with the direct approach and yields better results than the classical pooled data model, its performance are significantly lower than the direct approach. These results indicate that change in the distribution of SNP marker

genotypes between sub-populations is likely to carry relevant information about change of LD structure and strength between markers and QTLs across components of the admixed population that could be garnished to account for change in SNP effects. Since genomic selection largely depends on LD structure, it is of great importance to be able to evaluate and infer the magnitude of change in LD between SNP markers and QTLs in different populations. This information might shed some light on the change of SNP effects across different breeds or lines and how to adjust for this change. The objective of this study is to evaluate and infer the change of LD between markers and QTLs across two breeds using simulated data sets.

Materials and Methods

As indicated in the introduction section, genetic heterogeneity between sub-populations leads to change in estimates of SNP effects due to change in LD between observed markers and putative QTLs. The foundation of genome wide associations is that QTL effects can be inferred indirectly through their correlation (LD) with genotyped markers. Across sub-population, LD structure between markers as well between markers and QTLs changes. Consequently, it is reasonable to postulate that change in LD between SNP markers across two sub-populations (ΔLD_{M-M}) could explain, at least partially, the change in LD between markers and QTLs (ΔLD_{M-Q}).

In order to evaluate this hypothesis, several small scale simulations were carried out. In these simulations, the genotypes of the QTL(s) and associated SNPs markers were all assumed known. Thus, LD between SNP markers and QTL(s) was available. In all cases our goal was to test the ability of ΔLD_{M-M} to predict ΔLD_{M-Q}.

Simulation scenarios: Three simulation scenarios with varying number of SNP markers and QTLs were carried out to test the postulated hypothesis. In all cases, two divergent sub-populations for a trait with heritability equal to 0.5 were generated. A full description of the simulation parameters are presented in the next section. Two models (M1, M2) were evaluated in their ability to predict the change in ΔLD_{M-Q}:

$$\Delta LD_{M_k-Q} = a_0 + a_1 \mathrm{M_k} + a_2 \mathrm{S_k} + \mathrm{e_k} \qquad \text{(M1)}$$

$$\Delta LD_{M_k-Q} = b_0 + b_1 MR_{\mathrm{k}} + b_2 \mathrm{SR_k} + \mathrm{e_k} \qquad \text{(M2)}$$

where ΔLD_{M_k-Q} is the difference of LD between marker *k* and the QTL across the two sub-populations, $\mathrm{M_k}$and $\mathrm{S_k}$are the mean and standard deviation of the difference of LD between marker *k* and the remaining SNPs or a specified numbers of markers with a fixed genomic window (for example 100 adjacent SNP markers), respectively. $\mathrm{MR_k}$and $\mathrm{SR_k}$are the same as $\mathrm{M_k}$and $\mathrm{S_k}$, except they represent the relative mean and standard deviation of the difference in LD. $\mathrm{a_j}$and $\mathrm{b_j}(j = 0, 1, 2)$are unknown regression coefficients.

Linkage disequilibrium across SNP markers and between SNP markers and QTLs in both lines was calculated using the r^2coefficient as proposed by (Hill and Robertson 1968) using the following general equation.

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

where D is calculated as *D= f(AB) – f(A)f(B)* and f(AB), f(A), f(a), f(B) and f(b) are observed frequencies of haplotype AB and of alleles A, a, B, and b, respectively.

For all cases and for both models, unknown coefficients were estimated using the proc glm of SAS software (SAS Institute, Cary NC).

Data simulation: QMSim software (Sargolzaei and Schenkel. 2009) was used for data simulation. A historical population of unrelated individuals was simulated and used as a base population for two pure breeds (A and B). Breeds A and B consisted of 1677 and 1668 individuals respectively. The simulated genome consisted of 1 chromosome, with varying number of QTLs and varying number of SNP markers with equal spacing of an average 50Kb. Minor allele frequency was set to 0.05. QTL additive effects were sampled from a gamma distribution with shape and scale parameter equal to 0.4. Phenotypes were simulated based on a heritability of 0.5. Three simulation scenarios were carried out. In the first scenario, 10 SNP markers and 1 QTL were considered. The QTL was positioned in close proximity to SNP marker 5. In the second scenario the number of markers was increased to 300 SNP markers and also increased the number of QTLs to 3. Finally, in the last simulation scenario, the number of SNP markers was increased to 3000 SNPs and the number of QTLs increased to 30. These QTLs were randomly positioned across the genome.

Results and Discussion

Linkage disequilibrium between the SNP markers and the QTL for lines A and B as well as ΔLD_{M-Q} for the first simulation scenario are presented in Table 5.1. Since the QTL was placed in the center of the simulated segment, the LD_M_QTL was, as expected, higher for markers 4, 5 and 6. Figure 5.1 shows the trend of LD between the SNP markers and QTL for the two lines. Similarly, the LD between markers (LD_M_M) for the two lines as well as the difference in LD (ΔLD_M_M) were calculated. In order to infer ΔLD_{M-Q} between the two breeds, the mean and standard deviation of ΔLD_M_M were calculated and later used as explanatory variables in the regression model (Table 5.2). Fitting model M1 resulted in an R^2 of 0.65; indicated that the mean and standard deviation of ΔLD_M_M explained around two thirds

of the variation in ΔLD_{M-Q} between breeds A and B. On the other hand, fitting model M2 resulted in 25% decrease in R^2 (0.52). Although M2 resulted in a decrease in R^2, the model still was able to explain a significant portion of the variation in ΔLD_{M-Q} across the two breeds. When the number of SNP markers and QTLs were increased to 30 and 3, respectively (second simulation scenario), the coefficients of determination tended to decrease using either all the SNP markers (300) or fixed size widows of 100 SNPs to calculate the parameters of the regression model. Table 5.3 shows the resulting coefficients of determination (R^2) for models M1 and M2 using all markers and using fixed windows of 100 SNPs. Using M1 resulted in R^2 equal to 0.14, 0.12 and 0.12 for QTLs 1, 2 and 3 respectively using all 300 markers. In the case of using 100 marker windows, R^2 increased to 0.26 for QTL 1, 0.24 for QTL 2, and 0.27 for QTL 3. This increase in R^2 is due for at least two reasons: 1) a QTL was positioned in each 100 SNP marker window, and 2) including all 300 SNP markers where a large portion of them has no LD with the QTL, resulted in a less informative mean and standard deviation of ΔLD_M_M to explain variation in ΔLD_{M-Q}. The highest increase in R^2 was for QTL 3, from 0.12 to 0.27. Using M2, a substantial decrease in R^2 was observed across all QTLs using either 100 marker windows or all markers. Table 5.4 shows the average R^2 across all 3 markers, it is clear that M1 performed better than M2 in this simulation scenario.

In the third simulation scenario, a larger SNP panel (3000 SNPs), and a higher number of QTLs (30) were simulated. Table 5.4 shows the average R^2 obtained using M1and M2. Clearly, M1 performed notably better than M2 using either all markers or 100 marker windows. For example, fitting M1 using all markers resulted in an average R^2 of 0.27 compared to 0.01 for M2. It should be mentioned that M2 did not explain any variation in the change of LD_M_QTL across breed A and B.

Across the three simulation scenarios, it is clear that a significant portion of the variation in variation in ΔLD_{M-Q} could be explained by information already available in the observed SNP marker data. Furthermore, the statistical model as well as the extent of the window of SNPs considered in the calculation of the parameters of the regression line plays a crucial role in estimating change in LD between markers and QTLs in both breeds. Based on the results of this simulation study and the structure of LD generated, it seems that small windows are preferable. This is true because including large number of SNPs with little to no LD with the QTL(s) will render the mean and standard deviation non-informative about the variation in ΔLD_{M-Q}. Using real data, the situation will be more complex due to a larger number of SNP markers and QTLs where the latter have a random and unknown distribution. In such case, information about LD blocks should be used in determining the length of SNP windows to be used. Additionally, the relationship between ΔLD_{M-Q} and the observed information in the SNP genotypes could be non-linear and cannot be approximated well with simple regression models.

Conclusions

In this simulation study, inferring change of linkage disequilibrium between marker and QTL between two pure breeds proved to be possible. This might help in inferring the change of SNP marker effects when having different breeds or lines in the population. Whether or not this could be used to in genomic selection in the case of admixed populations, further testing and research is required.

References

Daetwyler, H. D., Kemper, K. E., Van der Werf, J.H.J., Hayes, B.J. 2012. Components of the Accuracy of Genomic Prediction in a Multi-Breed Sheep population. J. Anim Sci. 90, pp. 3375–3384

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

Erbe et al., M. Erbe, B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, Mason, M.E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels J. Dairy Sci., pp. 4114–4129

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

Hayes, B., Bowman, P. J., Chamberlain, A.C., Verbyla, K., Goddard, M.E., 2009. Accuracy of genomic breeding values in multiple dairy cattle populations. Genet Sel Evol 41:51

Heringstad, B., G. Su, T.R. Solberg, B. Guldbrandtsen, M. Svendsen, M.S. Lund. 2011. Genomic predictions based on a joint reference population for Scandinavian red breeds. In Book of Abstracts of the 62nd Annual Meeting of the European Federation of Animal Science Waageningen Academic Publishers, Stavanger, Norway, p. 29.

Hoze, C., Fritz, S., Phocas, F., Boichard, D., Ducrocq, V., Croiseau, P., 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J. Dairy Sci. 97:3918-3929

Kizilkaya, K., Fernando, R.L., Garrick, D.J., 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88:544-551.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., et al., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803‑819.

Lund, M.S., de Roos, A.P.W., de Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., Su, G., 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol, 43:43.

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

Olson, K.M., VanRaden, P.M., Tooker, M.E., 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 95:5378-5383

Pryce, J.E., Gredler, B., Bolormaa, S., Bowman, P.J., Egger-Danner, C., Furest, C., Emmerling, R., Solkner, J., Goddard, M.E., Hayes, B.J., 2011. Genomic selection using a multi-breed, across country reference population. J Dairy Sci. 94:2625-2630

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

SAS Institute Inc. 2011. Base SAS® 9.3 Procedures Guide. Cary, NC: SAS Institute Inc.

Shifman, S., J. Kuypers, M. Kokoris, B. Yakir, and A. Darvasi. 2003. Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. 12:771‑776.

Toosi, A., Fernando, R.L., Dekkers, J.C.M., 2009. Genomic selection in admixed and crossbred populations. J. Anim Sci. 2009. 88:32-46.

VanRaden, P. M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., 2009. Invited review: Reliability of genomic predictions for North American Hollstein bulls. J. Dairy Sci. 92:16-24.

Wientjes, Y. C. J., Veerkamp, R.F., Calus M.P.L., 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics,, pp. 621–631

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. Meuwissen, M. Svendsen, H. Grove, U.S. Nielsen, M.S. Lund. 2014. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. J. Dairy Sci. 10.3168.

Table 5.1: Linkage disequilibrium between markers and QTL for breed A and B

LD_M_QTL A[1]	LD_M_QTL B[2]	ΔLD_M_QTL[3]
0.131	0.101	0.029
0.107	0.008	0.026
0.107	0.008	0.024
0.758	0.333	0.419
0.999	0.649	0.350
0.622	0.363	0.259
0.296	0.222	0.074
0.132	0.195	-0.063
0.128	0.172	-0.043
0.005	0.106	-0.051

[1]LD between marker and QTL for breed A; [2]LD between marker and QTL for breed B, [3]Difference in marker and QTL LD between breed A and B.

Table 5.2: Mean and standard deviation of change of LD between markers[1]

ΔLD_M_M[1]	
mean	SD
0.010	0.187
0.018	0.166
0.019	0.167
-0.044	0.198
0.014	0.259
-0.091	0.225
-0.134	0.197
-0.082	0.095
-0.087	0.088
-0.072	0.120

[1]Difference in LD of marker and marker between breeds A and B.

Table 5.3: Coefficient of determination for models M1 and M2 in the second simulation scenario

	M1		M2	
ΔLD_M_QTL	All markers	100 marker window	All markers	100 marker window
QTl_1	0.14	0.26	0.07	0.03
QTL_2	0.12	0.24	0.02	0.02
QTL_3	0.12	0.27	0.01	0.01

Table 5.4: Average coefficient of determination over all QTLs for models M1 and M2 in the second and third simulation scenarios

	M1		M2	
Genome	All markers	100 marker window	All markers	100 marker window
300 SNP,3 QTLs	0.12	0.25	0.05	0.03
3000 SNP, 30 QTLs	0.27	0.10	0.03	0.01

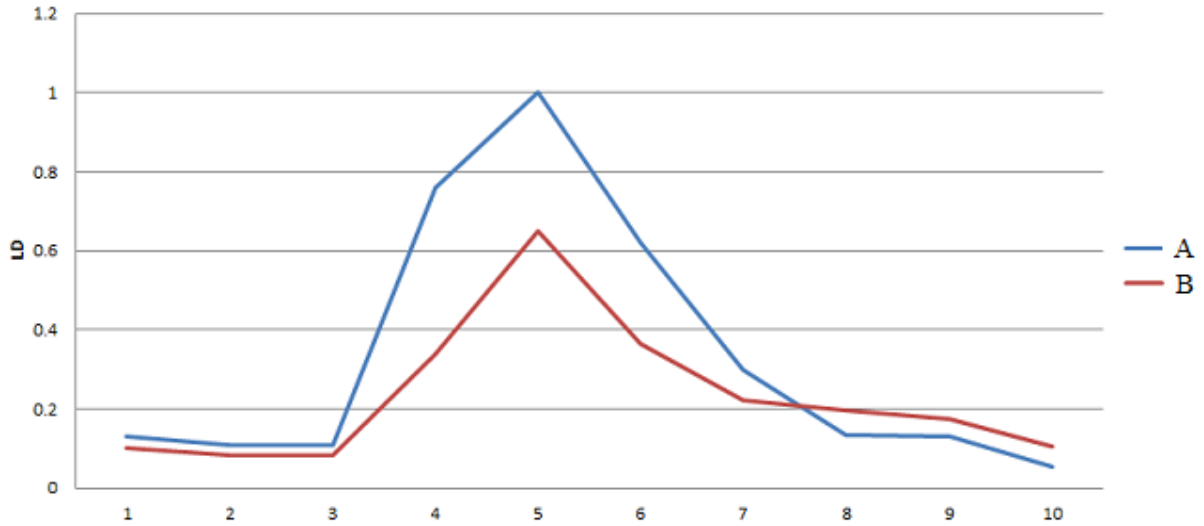

Figure 1: Linkage disequilibrium between markers and QTL for breeds A and B

CHAPTER 6

CONCLUSIONS

Current genomic selection methods for dealing with admixed populations assume homogeneity of the population and ignore the change in genetic parameters. Few approaches have been proposed to account for sub-population differences; however most of these studies only slightly improved the accuracy of genomic prediction and similar in some cases to approaches assuming homogeneity.

The pooled data approach is an attractive method since it increases the size of training data, therefore increasing power. However, this approach only works when the sub-populations are genetically similar. In this study, both proposed models do not assume homogeneity of the population and allow SNP marker effects to differ across sub-populations. Given the results of this study, it is evident that both proposed models performed notably better than classical pooled data approach. As presented in chapter 3 the proposed multi-compartment model is clearly superior then pooled data approach as it accounts for the difference in SNP effects across sub-populations. Its superiority compared to the pooled data approach ranged from approximately 17 to 47% and increases as the divergence between lines increases. In chapter 4, the multi-compartment model resulted in 43%-62% increase of prediction accuracy. The structural model also performed better than the pooled data method increasing the prediction accuracy by 10%-18%. The results from the structural model suggest the possibility to model the change of SNP marker effects as a function of the information already available in the genotypes data such as change in linkage disequilibrium. Furthermore, the major benefit from the proposed structural

model is its ability to predict genomic breeding values for non-phenotyped individuals as is the case for some commercial animals. Although promising, the results from this study are largely based on small simulated data sets and thus their testing and validation on real data are needed. Furthermore, the proposed models to explain change in LD between lines were rather simplistic and could be easily. Improved of these models could be achieved through the refinement of the sets on explanatory variables included or by the assumption of non and semi parametric approaches including machine learning based methods.