2012

Genome-wide prediction of breeding values and mapping of quantitative trait loci in stratified and admixed populations

Ali Shaarbaf Toosi

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Genetics Commons

Recommended Citation

Shaarbaf Toosi, Ali, "Genome-wide prediction of breeding values and mapping of quantitative trait loci in stratified and admixed populations" (2012). Graduate Theses and Dissertations. 12756.

https://lib.dr.iastate.edu/etd/12756

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Genome-wide prediction of breeding values and mapping of quantitative trait loci in stratified and admixed populations

by

Ali S. Toosi

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Animal Breeding and Genetics (Quantitative Genetics)

Program of Study Committee:

Rohan L. Fernando, Major Professor

Jack Dekkers

James Reecy

Alicia Carriquiry

Dan Nettleton

Iowa State University
Ames, Iowa
2012
Copyright © Ali S. Toosi, 2012. All rights reserved.
Tribute to my late father …

To my mother …

To Nahid …

And To Saba and Yasmin
# TABLE OF CONTENTS

LIST OF TABLES ................................................................. VIII

LIST OF FIGURES .............................................................. X

ACKNOWLEDGEMENTS ....................................................... XII

ABSTRACT .............................................................................. XIV

CHAPTER 1. GENERAL INTRODUCTION ...................................................... 1

1.1 INTRODUCTION ...................................................................... 1

1.2 RESEARCH OBJECTIVES ..................................................... 8

1.3 THESIS ORGANISATION ..................................................... 8

1.4 LITERATURE CITED .......................................................... 10

CHAPTER 2. LITERATURE REVIEW ....................................................... 13

2.1 LINKAGE ANALYSIS ........................................................... 13

  2.1.1 LA of complex traits ..................................................... 14

  2.1.2 LA mapping resolution ................................................ 15

2.2 POPULATION-BASED ASSOCIATION STUDY .............................. 17

  2.2.1 Linkage disequilibrium ................................................... 18

  2.2.2 Measures of LD ............................................................ 19

  2.2.3 Linkage, LD and Hardy- Weinberg disequilibrium ............. 20

2.3 FACTORS AFFECTING LD ................................................... 21

  2.3.1 Mutation ....................................................................... 22

  2.3.2 Selection ........................................................................ 22

  2.3.3 Random genetic drift .................................................... 23

  2.3.4 Non-random mating ..................................................... 24

2.4 POPULATION STRATIFICATION .............................................. 25

  2.4.1 Measures of PS ............................................................. 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.2 Impact of PS on statistical inferences in population genetics</td>
<td>28</td>
</tr>
<tr>
<td>2.4.3 Population admixture</td>
<td>31</td>
</tr>
<tr>
<td>2.4.4 Difference between PS and population admixture</td>
<td>32</td>
</tr>
<tr>
<td>2.4.5 Confounding due to PS and admixture</td>
<td>33</td>
</tr>
<tr>
<td>2.5 TESTS OF GENETIC ASSOCIATIONS IN PBA STUDIES</td>
<td>36</td>
</tr>
<tr>
<td>2.5.1 Tests of independence or contingency tables for binomial traits</td>
<td>37</td>
</tr>
<tr>
<td>Pearson’s χ²-test</td>
<td>37</td>
</tr>
<tr>
<td>2.5.2 Alleles test</td>
<td>38</td>
</tr>
<tr>
<td>2.5.3 Fisher’s exact test</td>
<td>38</td>
</tr>
<tr>
<td>2.5.4 Cochran-Armitage trend test</td>
<td>39</td>
</tr>
<tr>
<td>2.5.5 Log-Likelihood ratio test</td>
<td>40</td>
</tr>
<tr>
<td>2.5.6 Test of hypothesis of no association in the presence of PS</td>
<td>40</td>
</tr>
<tr>
<td>2.5.7 Tests of associations for quantitative traits</td>
<td>42</td>
</tr>
<tr>
<td>2.5.8 M-sample and non-parametric tests of association for a quantitative trait</td>
<td>42</td>
</tr>
<tr>
<td>2.5.9 Generalized linear model</td>
<td>43</td>
</tr>
<tr>
<td>2.5.10 Logistic regression</td>
<td>44</td>
</tr>
<tr>
<td>2.6 ANALYTICAL CHALLENGES OF HIGH-DIMENSIONAL GWAS DATA</td>
<td>46</td>
</tr>
<tr>
<td>2.6.1 Control of false-positives in GWAS</td>
<td>46</td>
</tr>
<tr>
<td>2.6.1.1 Family-wise error rate</td>
<td>47</td>
</tr>
<tr>
<td>2.6.1.2 False discovery rate</td>
<td>47</td>
</tr>
<tr>
<td>2.6.1.3 Bonferroni correction for multiple comparisons</td>
<td>49</td>
</tr>
<tr>
<td>2.6.1.4 Effective number of tests</td>
<td>49</td>
</tr>
<tr>
<td>2.6.2 Model selection in GWAS</td>
<td>51</td>
</tr>
<tr>
<td>2.7 LITERATURE CITED</td>
<td>52</td>
</tr>
<tr>
<td>2.8 FIGURES</td>
<td>59</td>
</tr>
<tr>
<td>2.9 TABLES</td>
<td>63</td>
</tr>
<tr>
<td>CHAPTER 3. GENOMIC SELECTION IN ADMIXED AND CROSSBRED POPULATIONS</td>
<td>65</td>
</tr>
</tbody>
</table>
CHAPTER 4. GENOME-WIDE MAPPING OF QTL IN ADMIXED POPULATIONS ................. 103

4.1 ABSTRACT ............................................................................................................. 103

4.2 INTRODUCTION .................................................................................................. 104

4.3 LITERATURE REVIEW ....................................................................................... 105

4.4 METHODS .......................................................................................................... 112

4.5 RESULTS ............................................................................................................ 120

4.6 DISCUSSION ....................................................................................................... 121

4.7 LITERATURE CITED ........................................................................................ 132

4.8 FIGURES ............................................................................................................. 139

4.9 TABLES .............................................................................................................. 141

CHAPTER 5. GENERAL DISCUSSION AND CONCLUSION ........................................... 145

LITERATURE CITED .................................................................................................. 148

APPENDIX 1. GENOMIC SELECTION OF PUREBREDS FOR CROSSBRED PERFORMANCE ........ 149

A1.1 ABSTRACT ......................................................................................................... 149

Background .............................................................................................................. 149

Results ....................................................................................................................... 150

Conclusion ................................................................................................................ 150

A1.2 INTRODUCTION ............................................................................................... 150

A1.3 METHODS ........................................................................................................ 152

A1.3.1 Simulation .................................................................................................... 152

A1.3.2 Statistical Models ........................................................................................ 154

A1.4 RESULTS ......................................................................................................... 155

A1.5 DISCUSSION .................................................................................................... 158

A1.6 ACKNOWLEDGEMENTS ................................................................................ 164
APPENDIX 2. APPLICATION OF WHOLE-GENOME PREDICTION METHODS FOR

GENOME-WIDE ASSOCIATION STUDIES: A BAYESIAN APPROACH

A2.1 SUMMARY .......................................................... 173
A2.3 METHODS .......................................................... 178
A2.4 DISCUSSION ......................................................... 191
A2.5 ACKNOWLEDGEMENTS ............................................. 201
A2.6 LITERATURE CITED ................................................. 201
A2.7 FIGURES .......................................................... 204
LIST OF TABLES

Table 2.1 A 2x3 contingency table for marker-trait association test in a case-control study………………………………………………………………………………………………………………..62

Table 2.2 Possible outcomes of multiple hypotheses testing of an experiment with m markers……………………………………………………………………………………63

Table 3.1 – The parameters used for the simulation program……………………………………………………………………………………………………………………………..92

Table 3.2- Average accuracy of estimated breeding values in the validation dataset (purebred B) from genomic selection with different training ……………………………….93

Table 3.3 - The impact of time since divergence of breeds on the accuracy of genomic selection when training in different datasets with 5 markers ………………………….94

Table 3.4- Average distance (in cM) between adjacent markers with r^2 greater than 0.1, 0.4 or 0.7 in different training datasets………………………………………………95

Table 4.1- QTL positions (Morgan), Mean, Standard deviation, Minimum and Maximum of h^2_{QTL} across 32 simulated datasets…………………………………………….140

Table 4.2- Accuracy, Power, Prob. Of Type I Error and Positive Predictive Value (PPV) for SMA and MLM analysis with NCHR method of finding thresholds in the ADMX population………………………………………………………………………..141
Table 4.3- Accuracy, Power, Prob. Of Type I Error and Positive Predictive Value (PPV) for SMA and MLM analysis with SLIDE method of finding thresholds in the ADMX population……………………………………………………………………..…..142

Table 4.4- Accuracy, Power, Prob. Of Type I Error (FPR) and Positive Predictive Value (PPV) for BMR analysis in the ADMX population…………………………………………………………..143

Table A1.1 Accuracy (SE) of breeding values in pure breed predicted based on two-breed cross data using ASGM or BSAM for three different………………………………………..167

Table A1.2 Accuracy (SE) of breeding values in pure breed predicted based on three-breed cross data using ASGM or BSAM for three different………………………………………..168

Table A1.3 Accuracy (SE) of breeding values in pure breed predicted based on four-breed cross data using ASGM or BSAM for three different scenarios………………………………………..169

Table A1.4 Accuracy of breeding values in pure breed predicted based on performance in the same pure breed using ASGM………………………………………………………………170

Table A1.5 Accuracy of breeding values in pure breed predicted based on crossbred data when the breeds are closely related for a simulated genome of………………………………………..171
LIST OF FIGURES

Figure 2.1 Percentage of individuals with diabetes disease in Gila Indian River population along with frequency of Gm haplotype in subpopulations with different Indian heritage ancestry................................................................. 58

Figure 2.2 An effect mediator is in the causal pathway between the genotype and the trait under study ................................................................. 60

Figure 2.3 False association at a marker locus due to population stratification in a case–control study ................................................................. 61

Figure 2.4 Stratified analysis of case-control studies, based on either genotypes or allele counts ........................................................................... 62

Figure 3.1 Schematic representation of the simulated population history (Ne = effective population size) and the different types of crossbred and admixed populations that were simulated........................................................................... 96

Figure 3.2 Average linkage disequilibrium as measured by of r^2 against distance (cM) in different training populations .................................. 97

Figure 3.3 Average distance (cM) between adjacent markers in different training populations at various levels of linkage disequilibrium (r^2) .................................................................................................................. 98

Figure 3.4 Average level of linkage disequilibrium as a function of marker density (# of markers per cM) and type of training population................................................. 99

Figure 3.5 Correlation of r between each pair of training and validation populations, as a function of marker density (in 1 cM) ........................................ 100

Figure 3.6 Plot of accuracy against between breeds variance of true breeding values ...................................................................................... 101

Figure 4.1 – Scatter plots of the first two principal components of the genome-wide
markers in the admixed (left) and the purebred (right) populations..........................138

**Figure 4.2**- Q-Q plots of the observed distribution of \(-\log_{10}(P\text{-values})\) on the null chromosomes, with different analysis approaches, vs. their expected .........................139

**Figure A1.1** Frequency of SNP alleles for purebreds A and B in generation 1050 for unrelated breeds.................................................................165

**Figure A1.2** Difference in average genotypic values of two breeds against the accuracy of breeding values predicted based on their crossbred data.................................166

**Figure A2.1** Illustration of composite genomic window \(W\) consisting of central window \(W_C\) and flanking windows \(W_L\) and \(W_R\).........................................................203

**Figure A2.2** Relationship between window posterior probability of association (WPPA) and the actual frequency of simulated QTL in analyses where .......................204

**Figure A2.3** Relationship between posterior probability of association (PPA) of individual markers and the actual frequency of simulated QTL in analyses .................205

**Figure A2.4** Relationship between window posterior probability of association (WPPA) and the actual frequency of simulated QTL........................................206

**Figure A2.5** Relationship between window posterior probability of association (WPPA) and the actual frequency of simulated QTL........................................207

**Figure A2.6** Relationship between posterior probability that variance of the central window \(W_C\) exceeds 1/1,000 of total variance and corresponding .........................208
ACKNOWLEDGEMENTS

I would like to offer my greatest thanks to Dr. Rohan Fernando who gave me the honor of doing my PhD under his supervision. I am deeply indebted to him for his encouragements, patience and support and the learning opportunities that he provided me with. Without him, I could not have this job done. Also, I would like to express my sincerest gratitude to Dr. Jack Dekkers. Without his innovative ideas and brilliant solutions this research could not have been accomplished. Dr. Jim Reecy has been always a reliable academic support for me; he has done everything that he could to make my research more productive. I have been very fortunate to have Dr. Alicia Carriquiry and Dr. Dan Nettleton in my program of study committee. I appreciate very much their insightful inputs to my research program. I greatly appreciate the time they spent to answer my questions despite their very busy schedules any time that I needed it.

Also, I would like to express my gratitude to Drs. Fields Gunsett, Fabiano Pitta and Archie Clutter, who have been my managers during my four-year of internship in Newsham Choice Genetics Co. Further, I appreciate the financial support of NCG during my PhD program.

I am totally speechless when it comes to acknowledge my family’s support. No words can explain how much they supported me during these years. I am deeply indebted to my wife for her unconditional love, support and patience. Without her sacrifice I was never able to do my PhD. I would like to express my heartfelt gratitude to my beautiful daughters, whom I could not have spent much time with them while I was a graduate student.

Further, I would like to thank all postdocs and fellow graduate students whom I have been lucky to share my thoughts with them and learn from their experience. Lastly, I would
like to offer my final gratitude to my dear friend, Dr. M.R. Nassiri whose support and friendship has been with me during the past several years.
ABSTRACT

Ideally genome-wide association studies require homogenous samples originating from randomly mating populations with minimal pedigree relationship. However, in reality such samples are very hard to collect. Non-random mating combined with artificial selection has created complex pattern of population structure and relationship in commercial crop and livestock populations. This requires proper modeling of population structure and kinship a necessary step of all genome-wide association studies. Otherwise, the risk of both false-positives (declaring a marker as significant without it be linked to a QTL) and false-negatives (markers linked to a QTL declared as non-significant) increases dramatically.

In this thesis, we first applied genomic selection (GS) approach to develop equations for prediction of breeding values of purebred candidates based on a model trained on an admixed or crossbred population. In this approach all markers effects are treated as random and are fitted simultaneously. It was hypothesized that given a high-density marker data and using the GS approach; training in a crossbred or admixed population could be as accurate as training in a purebred population that is the target of selection. In a stochastic simulation study, it was shown that both crossbreed and admixed populations could predict breeding values of a purebred population, without the need for explicitly modeling of breed composition and pedigree relationship. However, accuracy of GS was greatly reduced when genes from the target pure breed were not included in the admixed or crossbred training population. In addition, it was shown that the accuracy of GS depends on the genetic distance between the training and validation population, the closer the relationship between the two the higher was the prediction accuracy. Further, increasing of marker density improved the accuracy of prediction especially when a crossbred population has been used as the training
dataset. Considering haplotypes with weak linkage disequilibrium (LD), the crossbreds showed extensive LD, whereas the LD in the purebreds was confined to smaller segments. In contrast, examination of the length of haplotypes with strong LD indicated that these haplotypes are much shorter in crossbreds than that in purebreds. Our results showed that in crossbred populations the number of haplotypes with strong LD is less than that in the purebred populations. The findings of this research suggested that the crossbred populations are more suitable for QTL fine mapping than the purebreds.

In addition, in another simulation study we compared power, false-positive rate, accuracy and positive predictive value of QTL mapping in an admixed population with and without modeling of breed composition. The performance of ordinary least square (OLS) and mixed model methods (MLM), both fitting one-marker-at-a-time, were compared to that of a Bayesian multiple-regression (BMR) method that fitted all markers simultaneously. The OLS method showed the highest rate of false-positives due to ignoring breed composition and pedigree relationship. The MLM approach showed spurious false-positives when breed composition was not accounted for. The BMR outperformed both OLS and MLM approaches. It was shown that BMR could mitigate the confounding effects of breed composition and relationship without compromising its power. In contrast to the MLM where fitting of breed composition reduced both its power and false-positive rates, when breed composition was considered in the BMR it resulted in loss of power without a change of false-positive rate. It was concluded that the BMR is able to self-correct for the effects of population structure and relatedness.
CHAPTER 1. GENERAL INTRODUCTION

1.1 INTRODUCTION

The field of commercial plant and livestock improvement has experienced considerable advances over the past 60 years. This has been made possible through the application of quantitative genetic theory and artificial selection based on phenotypic measurements (Dekkers and Hospital 2002). Despite the tremendous achievements made, selection based on phenotypes and pedigree has several shortcomings: (1) the phenotype of interest might be an imperfect predictor of an individual’s true breeding value (BV); this is most relevant for lowly heritable traits, (2) some phenotypes are difficult and/or expensive to measure or are expressed late in an individual’s life, or even not expressed in one gender, i.e. the so-called sex-limited traits, (3) sometimes the real genetic potential of an individual is masked by the epistatic interaction among different genes or by the unfavorable association between genes contributing to the trait (Dekkers and Hospital 2002). In contrast, scores based on non-functional DNA polymorphisms, known as DNA markers, such as the single nucleotide polymorphism (SNP): (1) are stable and more reliable than phenotypic scores, because they are not influenced by environmental factors and are perfectly heritable, i.e., their heritability is one. Thus the error margins of a marker score tends to be much narrower than that of a phenotypic assay (Peleman et al. 2005), (2) can be assessed at any age, and hence they are very advantageous for evaluating traits that are expressed subsequent to reproductive age or at a point when selection decision is being made (Dekkers and Hospital 2002; Peleman et
The first step of marker assisted selection (MAS), that is selection of genetically elite individual using molecular markers rather than phenotypic scores, is identification of markers that are linked to quantitative trait loci (QTL) (MALOSETTI et al. 2007). A successful genetic association study leads to the identification of two sets of genetic loci that can be used towards a MAS program: functional polymorphisms or causal mutations (so-called *direct association*) and flanking non-functional genetic markers that are highly correlated with QTL (*indirect association*) (CLAYTON 2008; DEKKERS and HOSPITAL 2002). The resource population for such studies could be a specialized one, such as an F2, a back-cross, a recombinant-inbred population or a natural population (PELEMAN et al. 2005).

The success of a genome-wide association (GWA) study that is based on the indirect association depends on the magnitude of the covariance or linkage disequilibrium (LD) that exist between the marker being tested and the unobserved QTL affecting the trait of interest. The magnitude of the LD is determined by biological factors like population history and distance between the two loci. While we are especially interested in the LD that is due to close linkage, we want to exclude confounding factors such as selection, kinship and population stratification (PS) from contributing to the association signal (SILLANPAA and BHATTACHARJEE 2005). A GWA study with samples from a randomly mating population with minimal relatedness typically has the greatest statistical power (VISSCHER et al. 2008; YU et al. 2006). A statistical association between the marker being tested and the trait of interest in such a population might imply a physical linkage because LD between unlinked loci dissipates very rapidly with time (PRITCHARD and ROSENBERG 1999). However, in the
presence of PS any marker that has different allele frequencies across population strata will be in LD with other loci across the genome and thus might show an association with the phenotype of interest (PRITCHARD and ROSENBERG 1999). A spurious association, i.e., an association without linkage, due to PS occurs if both of these conditions are met: First, allele frequencies of the marker being tested must differ among subpopulations. Second, the mean trait value of interest must vary across subpopulations (CHAKRABORTY and WEISS 1988; DENG 2001; OCHIENG et al. 2007; PRITCHARD and ROSENBERG 1999). The spurious association occurs simply because many markers throughout the genome are likely to be slightly informative of an individual’s subpopulation of origin; therefore, they could be predictive of any phenotype that varies across subpopulations (ASTLE and BALDING 2009).

Admixture is the presence of several genetically distinct subgroups within a population (WANG et al. 2005). A well-know example of population admixture is a sample consisting of a mixture of breeds (GODDARD and HAYES 2009). A more subtle example of admixture is relatedness of individuals within a sample (GODDARD and HAYES 2009). Admixture not only creates new LD between loci but also alters the extent of it for loci that were in LD in the parental populations (CHAKRABORTY and WEISS 1988; DU et al. 2007). In addition, it may cause highly significant LD between polymorphisms that are fairly apart from each other (e.g., KAPLAN et al. 1998) or even are located on different chromosomes (FLINT-GARCIA et al. 2003; GODDARD and MEUWISSEN 2005; HIRSCHHORN and DALY 2005; PFAFF et al. 2001; RABINOWITZ and LAIRD 2000). Therefore, admixture and PS can seriously elevate the false-positive rates of GWA studies, the extent of which depends on the degrees of population differentiation and admixture (DENG 2001). Unequal relatedness within a sample can result
in increased false-positive rates in two ways: first, regions where QTL are residing may be co-inherited with regions devoid of QTL (Goddard and Meuwissen 2005; Payseur and Place 2007) and second, genotype correlations within larger families can have a larger impact on the association results compared to the smaller ones (Goddard and Meuwissen 2005; Pryce et al. 2010). In essence, PS and relationship are two different aspects of the same factor, i.e., the large unobserved pedigree (Astle and Balding 2009). Relatedness is concerned with having a “common ancestor” in the recent past and PS represents having a “common ancestor” in the distant past (Clayton 2008). Therefore, association analysis approaches that model these two factors in a unified way are expected to be better in controlling of false-positive rates compared to approaches that treat them separately (Kang et al. 2010; Zhang et al. 2010).

The effect sizes of QTL contributing to complex traits are relatively small (Visscher et al. 2008). The magnitude of signals from these QTL may be comparable to confounding signals from PS and thus, the risk of false-positives for such traits might be higher than that for monogenic traits (Teo et al. 2009). A plethora of approaches have been developed over the past decades to overcome the confounding effect of PS and unequal relatedness. These methods have been mainly built on the single marker association analyses (SMA), i.e., fitting one-marker-at-a-time (Ingvarsson and Street 2011). When we employ SMA to conduct a GWA study on a complex trait, the problem of PS might be better thought as incorrectly modeling of the trait (Atwell et al. 2010). The SMA model simply ignores the multifactorial background of the phenotypic variance and implicitly assumes that a single QTL is causing all the variation (Atwell et al. 2010).
One of the main concerns with SMA is that it ignores the information contained in the joint distribution of all markers (Balding 2006; Zhang et al. 2011). This would not be an issue if markers were widely spaced such that they could be considered literally independent or a very dense marker array was available so that every QTL could have been on the chip (Balding 2006). However, with the current genotyping densities and a polygenic trait it is very unlikely that all QTL are included on the marker chip. For complex traits there are possibly multiple genes across the genome that each have a small effect picked up by markers adjacent to them; therefore, a multi-marker association (MMA) model better explains the true underlying genetic architecture of the trait than a SMA model (Chapman and Whittaker 2008; Fridley 2009; He and Lin 2011).

SMA has several major drawbacks: (1) For most complex polygenic traits SMA only detects a very small proportion of genetic variation and might lack enough power to detect weaker associations, which are being penalized through adjustments for multiple comparisons (Cho et al. 2010; Gu et al. 2009; Han and Pan 2010; Hoggart et al. 2008; Shriver and Vaughan 2011; Zhang et al. 2011). (2) Performance of SMA largely depends on the magnitude of LD between the marker being tested and the potential QTL, hence this method could be underpowered if the LD is low (Pan 2009). (3) SMA tends to underestimate marker effects, because the effects of marker alleles are marginalized over all genetic and environmental effects (Shriver and Vaughan 2011). (4) SMA not only fails to characterize complex network of gene-by-gene interactions (Pan 2009) but also it lacks power and precision to identify GxE interactions (Li et al. 2010). (5) It cannot distinguish between the set of markers in LD with each other (Punyani et al. 2010) and tends to miss causal signals that are marginally uncorrelated with the phenotype (He and Lin 2011).
Considering these limitations, most statisticians would prefer to run GWA studies in a multiple linear regression (MLR) framework in order to see predictors in concert (Wu et al. 2009). However, MLR might not be powerful enough for large-scale GWA studies with hundreds of thousands of markers being tested due to its large degrees of freedom cost and the collinearity that might exist between marker genotypes (Shriner and Vaughan 2011; Wang and Abbott 2008; Zhang et al. 2011). The main difficulty when the number of predictors is much larger than the number of observations is to decide which set of predictors should be kept in the joint prediction model and which ones should be dropped (He and Lin 2011). In the second chapter of this thesis, we will briefly discuss this issue in the context of the model selection.

Compared to homogenous purebred populations, heterogeneous multi-breed populations offer some advantages. For the initial QTL mapping steps, without any loss of power, they require a lot less marker density relative to the purebred populations due to their long-range LD (Gabriel et al. 2002). The improved power and accuracy of QTL mapping using multi-population datasets have been shown in several studies (Guo et al. 2008; Kim et al. 2005). Fine mapping in a multi-population sample will yield better result if the structure of LD varies significantly across the sub-populations (Teo et al. 2009). A multi-population dataset, e.g., a multi-breed sample has potentially more informative recombination events and shorter haplotype length due to narrower LD distances across breeds. Thus, QTL mapping in such populations might be more accurate (Goddard and Hayes 2009; Parker et al. 2007; Toosi et al. 2010).

Genomic selection (GS) (Meuwissen et al. 2001) is a form of marker-assisted selection that uses marker genotypes and phenotypes in a training population to simultaneously
estimate effects of a large number of markers across the genome for the purpose of predicting breeding values (BV) of selection candidates based on their marker genotypes. In animal breeding, crossbred and multi-breed populations are usually the target of selection for genetic improvement of purebreds (Dekkers 2007). PS and selection both are inherent in livestock populations and generate LD between unlinked markers (Atwell et al. 2010). A plausible concern about GS and QTL mapping in such populations has been the extent and magnitude of LD in such populations and its impact on the accuracy of predictions and the precision and power of QTL mapping. Interestingly, studies of human populations have shown that the strengths of short-range LD in admixed African-Americans is quite similar to that in Africans (Gabriel et al. 2002). The admixed and multi-breed populations are vulnerable to the spurious associations due to their variation in ancestry. Therefore, the subject of this dissertation is to investigate the extent and the magnitude of LD in different crossbred and admixed populations and to examine the effects of PS and admixture on the accuracy of prediction and false-positive rates of GWA studies.
1.2 RESEARCH OBJECTIVES

The work presented in this thesis investigates the possibility of using admixed and multi-breed populations for QTL mapping and predicting breeding values of selection candidates using the genomic selection approach and compares the distribution of LD in different crossbred populations with that in a purebred population. The objective of applying genomic selection in admixed and structured populations is described in the third chapter. The goal of association mapping in multi-breed populations is explained in chapter four. The overall objective of this thesis is to evaluate the feasibility of using the genomic selection approach in admixed and different types of crossbred populations for prediction and association mapping purposes.

1.3 THESIS ORGANISATION

The aim of the second chapter is to provide a general background on some topics of population-based association studies that are relevant to the subject of this thesis. This includes principles and concepts of genome-wide association studies and the impact of population stratification on their results. A review of literature on the most common approaches for controlling false-positives due to population stratification is included in chapter 4, thus is not discussed in Chapter 2.

Chapter 3 consists of the paper “Genomic selection in admixed and crossbred populations”. This paper was published in the Journal of Animal Science 88:32-46 (2010) and was conducted by Ali S. Toosi under the direction of Drs. Rohan L. Fernando and Jack C.M. Dekkers.
Chapter 4 consists of the paper “Genome-wide QTL mapping of quantitative traits in admixed populations”. This paper will be submitted to *Genetics* and was conducted by Ali S. Toosi under direction of Drs. Rohan L. Fernando and Jack C.M. Dekkers.

Chapter 5 provides general discussion and conclusions based on the findings of projects described in this thesis.

Appendix 1 consists of the paper “Genomic selection of purebreds for crossbred performance”. This paper was published in *Genetics Selection Evolution* 41:12 (2009) and the authors are Noelia Ibánêz-Escriche, Rohan L. Fernando, Ali S. Toosi and Jack C.M. Dekkers. This paper compared prediction accuracy of a model with breed-specific SNP effects to that of the usual model that assumes SNP effects are the same across breeds. Ali S. Toosi was involved in the simulation of various scenarios of genomic selection designed for this study.

Appendix 2 consists of the paper “Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach”. This paper has been submitted to *Animal Genetics*. This paper compares various approaches of controlling false-positive rates in GWA studies. The authors are Rohan L. Fernando, Ali S. Toosi, Dorian Garrick and Jack C.M. Dekkers. Under direction of Drs. Rohan L. Fernando and Jack C.M. Dekkers, Ali S. Toosi conducted extensive simulations to study the properties of inference based on the posterior probabilities. These simulations set the direction for the study reported in this paper.
1.4 LITERATURE CITED


PAN, W., 2009 Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genetic Epidemiology 33: 497-507.


VISSCHER, P., T. ANDREW and D. NYHOLT, 2008 Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. Eur J Hum Genet 16: 387-390.


CHAPTER 2. LITERATURE REVIEW

2.1 LINKAGE ANALYSIS

Genetic mapping is the localization of genes contributing to phenotypes based on the correlation with DNA variation, without a prior knowledge about their biological function (ALTSHULER et al. 2008). In its simplest form it was started by Sturtevant for fruit flies in 1913 (ALTSHULER et al. 2008). Linkage analysis (LA) involves crosses between parents that differ at a Mendelian trait and many segregating markers (ALTSHULER et al. 2008). In this set up, any marker that shows co-segregation ("linkage") with the trait is inferred to be linked to the gene contributing to the trait (ALTSHULER et al. 2008). Based on the Mendel’s law of independent assortment two independent loci show a recombination rate ($\theta$) of $\frac{1}{2}$, hence in LA two loci are said to be linked if their $\theta < \frac{1}{2}$ (LAIRED and LANGE 2011). LA traces transmission of genetic material from parents to offspring across a few generations. With such time frame, only a limited number of recombinations can occur between linked loci, and as a result only a few markers are needed to cover a large region. This property was a major advantage for LA in the early days of genetic mapping when genotyping of more than 20-40 marker loci per chromosome would have been very costly and practically infeasible for most of research groups. LA approach proved to be successful for localizing genes affecting simple Mendelian disorders such as cystic fibrosis and Huntington’s disease in human (MORRIS and CARDON 2008). Thus, the number of disorders tied to a specific gene grew from nearly 100 in the late 1980s to more than 22,000 in 2008 (ALTSHULER et al. 2008). Genes conferring susceptibility to such diseases, so called monogenic disorders (ORR and CHANOCK 2008), are typically classified as major genes. These genes have low population
frequencies but are highly penetrant and their severe phenotypic consequences make them good targets for LA (COLLINS 2007).

2.1.1 LA of complex traits

Although LA of complex traits successfully was conducted in experimental organisms in the late 1980s there was not such success in human populations (ALTSHULER et al. 2008). In fact, LA yielded equivocal results when it was applied for genetic mapping of complex diseases like type 2 diabetes, cancer and heart disease in humans (ALTSHULER et al. 2008; MORRIS and CARDON 2008). For such diseases it is difficult to define disease status from multiple intermediate phenotypes and hence there is not a one-to-one relation between phenotype and the underlying causative mutation(s) (DARVASI 1998). In addition, individuals with complex diseases are less concentrated within families and affected family members are less likely to share the same variants at the underlying functional polymorphisms than that for Mendelian disorders (MORRIS and CARDON 2008).

In most domesticated animals, crosses were made between pairs of divergent breeds due to the lack of purebred lines (ANDERSSON et al. 1994). For complex traits like growth rate and fatness in pigs (e.g., ANDERSSON et al. 1994), milk production in dairy cattle (e.g., GEORGES et al. 1995), carcass traits in beef cattle (e.g., KEELE et al. 1999) and growth and carcass traits in chicken (e.g., VAN KAAM et al. 1999) hundreds of QTL were identified. An excellent review of the state of QTL mapping in different farm animals can be found in COCKETT and KOLE (2008). Complex multi-factorial or quantitative phenotypes result from the collective action of numerous, possibly interacting, genes and environment (DEKKERS and HOSPITAL 2002). An important lesson from extensive research on quantitative traits of importance in animal breeding was that to identify QTL underlying variation of such traits
often requires a series of experiments (Hayes et al. 2004). The detection of a mutation responsible for a large proportion of the genetic variation in milk-fat percentage in dairy cattle exemplifies the required steps (Grisart et al. 2002). Grisart et al. (2002) identified a point mutation in the DGAT1 gene, which is responsible for nearly 43% of the genetic variation of fat percentage. The first step in identifying the mutation was a genome-wide LA undertaken in 1995 by Georges et al. They found that a region of chromosome 14 contains a QTL with a large effect on fat percentage. In the next step, location of the QTL was narrowed down from a large confidence interval of 20-40 cM to a 3 cM region by taking advantage of association mapping (Riquet et al. 1999). The DGAT1 gene was identified as a strong candidate in this region, and subsequent sequencing revealed a single base pair mutation in the gene (Hayes et al. 2004). Eventually, the study of Thaller et al. (2003) proved that the mutation is associated with major effects on milk yield and composition.

### 2.1.2 LA mapping resolution

The precision of positioning of a putative QTL is the accuracy with which a QTL is mapped along a chromosome and typically is expressed as a confidence interval (CI) with a certain significance level (Hayes et al. 2004). CI resulted from LA typically are of the order of 20-40 cM long, which is equivalent to ~20-40 million base pairs or 200 to 400 genes in mammals (Georges 2007). Thus, the task of narrowing down the predicted QTL location from a few hundred to a short list of several genes still is formidable (Georges 2007). Increasing of marker density is a natural, and by today’s technology a trivial way of improving the QTL position resolution (e.g., Georges 2007; Hayes et al. 2004) but the effectiveness of this strategy is limited with linkage mapping because a very large number of recombinations between closely spaced markers is needed to get a refined QTL position.
(HAYES et al. 2004). Another method of getting a higher resolution for QTL position is increasing the crossover density (GEORGES 2007; HAYES et al. 2004). Recombinant chromosomes are the only source of mapping information (GEORGES 2007) and the most straightforward way of increasing crossover density is to generate more progeny (GEORGES 2007; HAYES et al. 2004). However, for a linkage mapping study that needs to limit the QTL position to a 1-3 cM interval, the number of progeny required is often beyond the reproductive capacity of the target species or practically infeasible (HAYES et al. 2004). For example, in an F₂ or backcross (BC) linkage mapping experiment nearly 5000 progeny are needed to provide a mapping resolution of 5 cM or less (GEORGES 2007). Another possibility for increasing the crossover density is to employ advanced intercross lines (AIL), i.e., F₃, F₄, ..., Fₙ generations (DARVASI and SOLLER 1995). AIL are produced by random crossing of progeny produced from an F₂ or BC experiment. AIL differs from recombinant inbred lines (RIL) in that the subsequent generations following F₂ or BC are not created via selfing or sib-breeding, rather they are produced via semi-random crossing to avoid inbreeding (DARVASI 1998). The accumulation of crossovers in RILs is limited due to the fact that each generation inbreeding makes the recombining chromosomes more and more similar to each other, as a result meiosis stops generating new recombinant haplotypes (FLINT-GARCIA et al. 2003; ROCKMAN and KRUGLYAK 2008). The CI for the QTL is reduced by a factor of nearly \( \frac{2}{n} \) when an Fₙ AIL is compared with that in an F₂ population of the same size, where n is the number of generations of intercrossing (GEORGES 2007). Nonetheless, given the long generation interval and housing cost, AIL remains a costly alternative way of increasing the QTL location resolution for most domestic animal species (GEORGES 2007). A natural alternative for fine-mapping of a QTL, which is now taking the lead is GWA mapping
(Georges 2007). GWA builds on of the historical recombinations which have accumulated over numerous generations preceding the genotyped generation (e.g., Hayes et al. 2004).

2.2 POPULATION-BASED ASSOCIATION STUDY

Compared to conventional bi-parental mapping populations, population-based association (PBA) studies have several advantages: (1) PBA studies are usually conducted on a much larger population and hence are more powerful than studies based on the LA, given that the causative mutations underlying the phenotype are not very rare (e.g., Morris and Cardon 2008). In addition, there is less chance of an overestimation of QTL effects and more refined estimates of QTL locations, (2) Evaluation of the genotype by environment interactions via PBA study is more precise than that in a designed mapping experiment conducted in limited number of locations and/or environmental conditions and exposures, so the results of the PBA studies are applicable to a wider range of conditions, (3) Germplasm diversity and genetic variability is much higher in a PBA study than that in a bi-parental population such as an F2. The amount of segregating genetic variance within traditional crossbred mapping populations is limited, because per locus at most two alleles can segregate in a diploid species, (4) Unlike LA which requires specific pedigree relationships, an attractive property of PBA studies is that they do not need specially designed crosses, instead they can be applied to any collection of genotypes with arbitrary and even unknown relationship between them (Georges 2007; Maloletti et al. 2007; Parisseaux and Bernardo 2004), (5) PBA study takes advantage of the effect of historical recombination, i.e., the cumulative effects of tens or hundreds past generations of recombination, to achieve fine-scale gene localization. However, a major difficulty is that past historical events, like admixture, random drift, multiple mutations and selection can
disturb the relationship between LD and physical distance (JORDE 2000). While linkage refers to the correlated inheritance of alleles at different loci due to their physical proximity, LD only refers to dependence of alleles at different loci in a population and does not convey any information about their physical linkage (FLINT-GARCIA et al. 2003; JANNINK and WALSH 2002; MALOSETTI et al. 2007).

**2.2.1 Linkage disequilibrium**

Linkage disequilibrium refers to the non-random association or the correlation of alleles at two or more loci due to shared ancestry among individuals in a population (e.g., FLINT-GARCIA et al. 2003; MORRIS and CARDON 2008). Alternatively, it can be said that two markers $A$ and $B$ with alleles $A/a$ and $B/b$, are in linkage disequilibrium (LD) if:

$$p(A|B) \neq p(A|b)$$

where $p(A|B)$ is the frequency of allele $A$ among gametes that contain allele $B$. This means that two gametes that are alike at marker $B$, have a higher chance of being identical at marker $A$ compared to two random gametes (GODDARD and MEUWISSEN 2005). This is only possible if the two gametes share a ‘recent’ common ancestor at both markers (GODDARD and MEUWISSEN 2005). Note that by ‘recent’ we mean more recent than expected for markers $A$ and $B$ in two random gametes, as based on coalescence theory there is always a common ancestor for two loci (GODDARD and MEUWISSEN 2005). If this is not the case, then alleles at the two loci assort independently in accordance with Mendel’s second law (GODDARD and MEUWISSEN 2005).
2.2.2 Measures of LD

Consider the pair of markers $A$ and $B$ on haplotype $h$ and let $X = 1$ if $h$ carries allele $A$ at the first locus ($X = 0$, otherwise) and $Y = 1$ if $h$ carries allele $B$ at the second locus ($Y = 0$, otherwise). Then, following Chapman and Thompson (2001):

$$D = \sigma_{XY} = p_{AB} - p_A p_B$$

where $p_{AB}$, $p_A$ and $p_B$ refer to the relative frequencies of haplotype $AB$, alleles $A$ and $B$ in the population, respectively. Under gametic-phase equilibrium (LE), $p_{AB}$ is equal to $p_A p_B$ and hence $D = 0$. Now, if we standardize $D$ we will get the standard measure of correlation of allelic states at locus $A$ and $B$ as (Chapman and Thompson 2001):

$$r = corr(X,Y) = \frac{D}{\sqrt{p_A p_a p_B p_b}}$$

which is Pearson’s correlation between two binary variables. There are two commonly used measures of LD (e.g., Morris and Cardon 2008):

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

And

$$
D' = \begin{cases} 
\frac{D}{\max (-p_A p_B, p_a p_b)}, & \text{if } D < 0 \\
\frac{D}{\min (p_A p_b, p_a p_B)}, & \text{if } D \geq 0 
\end{cases}
$$

Both measures range between 0 and 1. When there is complete LD between two loci, i.e., at least one of the four possible haplotype has a population frequency of zero, $D'$ will be equal to 1. This means that no recombination has occurred between the two loci since the mutations generating the polymorphisms occurred (Morris and Cardon 2008). On the other hand, in presence of perfect LD, i.e., when genotypes at one locus can be used as proxies for the
genotypes at the second locus, $r^2$ will be equal to 1 (Morris and Cardon 2008). An observed association between a marker and the phenotype of interest suggests that there is LD between the marker and a QTL contributing to the trait. This simple idea is the basis of association mapping (Laird and Lange 2011). It is important to note that both linkage mapping and association mapping strategies rely on LD between marker and QTL. Linkage mapping only considers LD that exists within a finite pedigree, whereas association mapping is based on the LD that exists between a marker and a QTL at the level of population (Hayes et al. 2004). Genetic markers that are very close to each other have either the same or similar ancestral origin, and this induces the correlation of allele states at different loci (Pritchard and Przeworski 2001). Whereas, loci that are more distantly apart might have different ancestral origin due to recombination. Hence, the strength of LD between a pair of marker depends on the genetic distance between them (Pritchard and Przeworski 2001). The power of an association study is directly related to the strength of marker-QTL LD as measured by $r^2$ (Pritchard and Przeworski 2001).

### 2.2.3 Linkage, LD and Hardy-Weinberg disequilibrium

LD always exists between linked loci within a family (Dekkers and Hospital 2002), but is usually not referred so because the term LD is a population concept (Goddard and Meuwissen 2005). In a randomly mating population, LD originates in the same way as linkage, i.e., through a recent common ancestor except that the common ancestor is before the recorded pedigree (if the common ancestor is within the recorded pedigree we would call it linkage rather than LD) (Goddard and Meuwissen 2005). LD measures the population probability that two alleles at two different loci appear together on the same parental haplotype. In this respect, LD is comparable to Hardy-Weinberg disequilibrium (HWD)
which is the population probability of two alleles at the same locus appear together in an individual’s genotype. Similar to HWD, LD can be originated from different sources, including mutation, close linkage and population structure (Laird and Lange 2011). We describe each of these factors in detail in next section.

2.3 FACTORS AFFECTING LD

In a randomly mating population and in the absence of forces that change gene frequency (i.e., mutation, selection, migration and drift) polymorphic loci will be in LE. The distribution of LD, i.e., the level and extent of LD, varies across populations and genomic regions and it might change dramatically between different pairs of adjacent loci. Some of the factors influencing LD variance are population specific, like random drift, inbreeding and admixture. Some others are genomic regions specific, e.g., recombination rate, gene conversion and selection (Shifman et al. 2003). LD might exist between pairs of loci on the same chromosome (intra-chromosomal LD) or on different chromosomes (inter-chromosomal LD). Recombination is the driving factor that weakens the former, whereas independent assortment breaks down the later (Flint-Garcia et al. 2003). In a closed random mating population, LD between unlinked loci is halved every generation, but LD between linked loci dissipates much more slowly per generation and it might not disappear even after hundreds of generations (Falconer and Mackay 1996). Several evolutionary forces can generate LD between alleles at different loci of which only recombination is correlated with physical distance between loci (Kaplan et al. 1995). High levels of LD might be result of tight linkage (Flint-Garcia et al. 2003). The number of markers and the experimental design needed for an association study is determined by the rate of decay of LD, or the distance over which LD persists (Jorde 1995). The mapping of genes based on
LD might not be successful if the process involved in creation of LD is ignored. Because, a strong LD detected between a pair of loci may be result of a recent occurrence of LD rather than the close proximity of the two loci (Wu and Zeng 2001).

### 2.3.1 Mutation

LD is the result of mutation and transmission of the mutant allele in subsequent generations (Zöllner 2001). With introduction of a new mutation in a population, it necessarily resides on a single chromosome (the founder chromosome) and thus on a single haplotype. As a result, all of the loci on the founder chromosome are in complete LD with the new allele, i.e., there are only two haplotypes present in the population, one carrying the mutant and another one carrying the wild-type allele. Over generations, LD between the mutant allele and the set of linked markers decays gradually (Jorde 1995). With the spread of the new allele in the population, the halpotype carrying the new allele will became more frequent. In the absence of recombination, all haplotypes carrying the new allele are identical by descent (IBD). However, in reality over time the IBD region of the founder chromosome reduces in size until it gets narrowed down to a tiny region around the locus carrying the new allele (Zöllner 2001).

### 2.3.2 Selection

Selection changes allele frequencies of favorable QTL and hence creates LD between the selected allele and linked loci, in a process called hitchhiking (e.g., Mackay and Powell 2007). In addition, it may result in LD between unlinked loci (Flint-Garcia et al. 2003; Goddard and Meuwissen 2005; Kaplan et al. 1998; Lande and Thompson 1990; Ochieng et al. 2007; Remington et al. 2001). If there is a favorable (or unfavorable) epistatic interaction between alleles at two loci, then selection can changes the frequency of
the favorable haplotypes and hence generate new LD despite the fact that the two loci might not physically be linked (e.g., FLINT-GARCIA et al. 2003; LANDE and THOMPSON 1990; XU 2010). As a result of the Bulmer effect (BULMER 1971), negative LD will occur between loci affecting a phenotype under stabilizing or directional selection. In contrast, positive LD will be generated between loci affecting a trait under disruptive selection (Xu 2010). If mutation creates a new allele that has some selective advantages, then its frequency in the population will increase rapidly (selection sweep). In this case, the LD between the new allele and loci linked to it can extend over large distances, because there has been a short amount of time for recombination to break down this LD (ZÖLLNER 2001). It is important to note that LD resulting from a selection sweep is localized to the parts of the genome that contain QTL for the trait under selection, whereas population structure affects the pattern of LD over the whole genome (BRESEGHELLO and SORRELLS 2006). Selection creates correlation between the QTL affecting the trait subject to selection, and hence a marker that is linked with one of the QTL is likely to be in LD with all others (PLATT et al. 2010).

2.3.3 Random genetic drift

Random drift refers to stochastic changes in allele and haplotype frequencies over time. In small populations, drift results in the consistent loss of rare haplotypes which increase levels of LD (FLINT-GARCIA et al. 2003; GODDARD and MEUWISSEN 2005; LANDE and THOMPSON 1990; LAZZERONI 2001; REMINGTON et al. 2001). The larger the population, the smaller will be the variance of allele frequency across generation. Thus, in small populations drift can have a large impact on allele frequencies and hence plays a major role in generating LD. It can create LD, even if all loci start in mutual LE (TERWILLIGER et al. 1998). Moreover, drift can generate LD through sudden population bottlenecks. This is because only
limited allelic combinations are passed on to future generation due to rapid changes of population size (FLINT-GARCIA et al. 2003; OCHIENG et al. 2007). It has been hypothesized that the extensive LD between micro-satellite markers in the Dutch black and white dairy cattle population is due to the bottlenecks resulted from the globalization of semen trading (FLINT-GARCIA et al. 2003). In a randomly mating population with effective size of $N_e$, it is expected that random drift to produce substantial association between polymorphic loci with recombination rates of $\theta < \frac{1}{4N_e}$ (HILL and ROBERTSON 1968). The expected value of LD, as measured based on $r^2$, under equilibrium between random drift and recombination is:

$$E(r^2) = \frac{1}{1 + 4N_e c}$$  \[6\]

where $c$ is the recombination rate between the two sites (SVED 1971).

2.3.4 Non-random mating

Non-random mating is mating of individuals who are more (or less) closely related than those drawn by chance from a random mating population (HEDRICK 2005). For example, assortative mating is mate selection based on the phenotypic characteristics. Non-random mating can strongly affect LD. In general, LD decays faster in outcrossing species as compared to self-mating species. Individuals are more likely to be homozygous in a self-mating population; hence recombination is less effective in such populations (FLINT-GARCIA et al. 2003). When a hybrid population is formed by crossing of two large genetically diverged populations, substantial LD may exist between selectively neutral loci with recombination rates $\theta < \frac{1}{T}$, where $T$ is the number of generations past parental populations mating (KIMURA and OHTA 1971). In such populations, substantial LD between a pair of loci is expected if recombination rate between them is less than
\[
\theta^* = \max \left( \frac{1}{T}, \frac{1}{4N_e} \right)
\]  

[7]

In domesticated populations, T is usually smaller than 4N_e. Therefore, crossbreeding is more effective in creating LD than random drift (LANDE and THOMPSON 1990). We will discuss in more detail about the consequences of non-random mating when we introduce some basic concepts of population genetics.

2.4 POPULATION STRATIFICATION

Natural populations usually are not represented by a single panmictic entity where individuals mate at random across the entire population, rather they are subdivided into smaller units, which can be arranged in space, time (e.g., age groups (LI 1972)), ecology or otherwise (EXCOFFIER 2007). There might be different levels of population divisions, sometimes arranged in a hierarchical order. Therefore, a population can be divided into groups or regions and each of them again may be subdivided into smaller units and so on until they reach to a basic unit, called ‘deme’, which might be regarded as a homogenous group. However, even within a deme often individual show inbreeding and there exits excess of homozygosity due to some shared ancestry and finite deme size (EXCOFFIER 2007). A natural population almost always shows differences in allele and genotype frequencies from one geographical region to another (HARTL and CLARK 1989). Several factors, including genetic bottlenecks, random genetic drift, natural or artificial selection and mating system can lead to diversified subpopulations (SNELLER et al. 2009; YU et al. 2006). Population stratification (PS) exits when a sample of individuals is subdivided into several mutually exclusive strata with different allele frequencies. Typically these strata represent different racial, ethnic and/or geographical subgroups in human populations (LAIRD and LANGE 2011).
Broadly speaking, PS can be defined as any deviation from random mating (McVean 2001; Pritchard and Rosenberg 1999). This agrees with (Redden and Allison 2006) who found that assortative mating in a homogeneous population can lead to PS. Even in a randomly mating population, which is expected to be of devoid of any PS (Wright 1951), PS could exist if some recent ancestors are heavily represented in the current population (Goddard and Meuwissen 2005). Perhaps, it makes more sense if we define PS as the variation in the degree of relationship among individuals in the entire population. In a complex population, whenever some individuals are more closely related than what is expected under random mating PS exists (Sneller et al. 2009). In Wright’s (1951) point of view PS referrers to the entire pattern of relatedness among individuals in a population, both within and across population subdivisions (Sneller et al. 2009).

2.4.1 Measures of PS

It is clear that when there is population subdivision, individuals do not mate at random and mating between individuals from the same strata is more likely than mating between two individuals from different strata. This departure from panmixia creates a correlation between homologous genes of uniting gametes relative to a pair of genes taken at random from the population (Excoffier 2007). Wright’s F statistic is a measure of such a correlation. If we define ‘inbreeding’ as any type of mating that increases homozygosity, there are two distinct processes that are able to do so: mating among relatives and PS (Devlin and Roeder 1999). Hence, it is convenient to measure the extent of PS using F statistics (Hartl and Clark 1989). A subdivided population has three distinct levels: individual organisms (I), subpopulations (S), and the total population (T) (Hartl and Clark 1989). Before we show
how to calculate Wright’s fixation index ($F_{ST}$) we need to define different levels of heterozygosity in a population as below (HARTL and CLARK 1989):

$H_I$: is the heterozygosity of an individual in a subpopulation, and can be interpreted as the average heterozygosity of all genes in an individual. If $H_i$ is the heterozygosity in subpopulation $i$, then for the case of $k$ subpopulations,

$$H_I = \sum_{i=1}^{k} \frac{H_i}{k} \quad [8]$$

$H_S$: is the expected heterozygosity of an individual in an equivalent randomly mating subpopulation. It represents the expected level of heterozygosity that would be found in a subpopulation if it were undergoing random mating. If $p_{i,k}$ is the frequency of the $i^{th}$ allele in population $s$, then the expected HW heterozygosity in subpopulation $s$ is:

$$H_s = 1 - \sum_{i=1}^{k} p_{i,s}^2 \quad [9]$$

$H_T$: is the expected heterozygosity of an individual in an equivalent randomly mating total population. It is the level of heterozygosity at the whole population if the subpopulations were pooled together and mated at random. If $\bar{p}_s$ is the frequency of allele $i$ averaged over the subpopulations, then

$$H_T = 1 - \sum_{i=1}^{k} \bar{p}_i^2 \quad [10]$$

The effects of population subdivision are measured by Wright’s fixation index (EXCOFFIER 2007; HARTL and CLARK 1989) defined as below:

$$F_{ST} = \frac{H_T - \bar{H}_S}{H_T} \quad [11]$$
Wright defined two other F statistics as:

\[ F_{IS} = \frac{H_s - H_I}{H_s} \]  

[12]

And

\[ F_{IT} = \frac{H_T - H_I}{H_T} \]  

[13]

All three statistics are measures of inbreeding but they differ according to the reference populations. \( F_{IS} \) measures the reduction in an individual’s heterozygosity within population and is affected by non-random mating within subpopulations. Whereas, \( F_{IT} \) measures individual’s inbreeding relative to total population. \( F_{IT} \) is the overall inbreeding of an individual and includes a contribution due to non-random mating within subpopulations (\( F_{IS} \)) and another (\( F_{ST} \)) due to population subdivision itself (Hartl and Clark 1989).

### 2.4.2 Impact of PS on statistical inferences in population genetics

In the presence of undetected PS almost any test of hypothesis on population’s characteristics would be invalidated. For example, the goodness-of-fit test of Hardy-Weinberg proportions is dramatically affected by inclusion of individuals from different strata (Hartl and Clark 1989). PS affects both the variance of the allele frequency estimates and the distribution of test statistics that are based on the allele frequencies. In addition, PS can affect the distribution of genotypes in the population (Laird and Lange 2011). We start with a randomly mating population and consider locus \( A \) with alleles \( A/a \).

Let’s consider X as the count of the favorable allele in an individual. Then, let’s take \( p = P(A) \) and define \( P(X = 0) = p_{aa}, P(X = 1) = p_{Aa}, \) and \( P(X = 2) = p_{AA} \).

It follows by definition that:

\[ E(X) = 2p_{AA} + p_{Aa} = 2p \]  

[14]
and

\[ \text{Var}(X) = 4p_{AA} + p_{Aa} - 4p^2 \]  \[15\]

Assuming that the population is at HWE with respect to this locus, then following (LAIRD and LANGE 2011):

\[ \text{Var}(X) = 2pq \]  \[16\]

Now, suppose the population is subdivided into \( t \) strata and locus \( A \) is under HWE in all strata. The frequency of heterozygotes in the population as a whole will be (LAIRD and LANGE 2011):

\[ P(X = 1) = 2 \sum_{k=1}^{t} s_k p_k q_k \]

\[ = 2 \sum_{k=1}^{t} s_k p_k (1 - p_k) \]

\[ = 2p - 2E(p_k^2) + 2p^2 - 2p^2 = 2pq - 2\sigma_p^2 \]

\[ = 2pq(1 - F) \]  \[17\]

where \( s_k \) is the proportion of individuals from stratum \( k \) in the whole population, \( F \) is the inbreeding coefficient and

\[ p = \sum_{k=1}^{t} s_k p_k \]  \[18\]

In the same way the frequency of homozygote genotypes will be:

\[ P(X = 0) = p_{aa} + \sigma_p^2 = Fq + (1 - F)q^2 \]  \[19\]

And

\[ P(X = 2) = p_{AA} + \sigma_p^2 = Fp + (1 - F)p^2 \]  \[20\]

Similarly following LAIRD and LANGE (2011) and based on equation \([14]\), we have:
$$\text{Var}(X) = 2pq + 2\sigma^2_p = 2pq(1 + F)$$  \[21\]

which shows that in a stratified population the $\text{Var} (X)$ is inflated relative to that in a population under HWE. Obviously, if $\sigma^2_p = 0$, i.e., the allele frequencies stays stable across strata, $\text{Var} (X)$ reduces to the binomial variance and is equal to that in the population at HWE. The HWE defines the relationship between genotype- and allele-frequencies in an infinitely large random mating population (FOULKES 2009). As it can be seen from equations \[17\], \[19\] and \[20\] when undetected PS exists, the pooled population departs from HWE because there is an excess of homozygote and a deficiency of heterozygote genotypes in the population (CHAPMAN and THOMPSON 2001; EXCOFFIER 2007; FOULKES 2009). The deficit of heterozygotes relative to a population at HWE, due to undetected PS is called Wahlund effect. Statistical tests or models that are rely on the assumption of HWE are invalid in the presence of the Wahlund effect (HARTL and CLARK 1989; LAIRD and LANGE 2011).

Departure from HWE for the locus under consideration implies associations between alleles at the locus. Let’s see how the concept of HWE is related to the correlation of allelic state of two uniting gametes at the same locus. Considering locus $A$, let $X = 1$ if the maternal gamete carries allele $A$ ($X = 0$, otherwise) and similarly let $Y = 1$ if the paternal gamete carries allele $A$ ($Y = 0$, otherwise). Then in an infinitely large population where matings occur between individuals whose coefficient of kinship is $f$, following CHAPMAN and THOMPSON (2001) we have:

$$\rho = corr(X,Y) = \frac{p_{AA} - p_A^2}{p_{AP}p_a} = \frac{p_Af + p_A^2(1-f) - p_A^2}{p_{AP}p_a} = f$$  \[22\]

Note that how the above equation relates $\rho$, a measure of identical by state (IBS), to $f$ a measure of identical by descent (IBD). Under HWE the correlation of allelic states at one
locus ($\rho$) is zero, in other words there is no association between the two alleles at the same locus. Now, let’s see how PS can lead to allelic association. Based on equations [17] and [19] we can calculate the measure of allelic association for a stratified population as follows (CHAPMAN and THOMPSON 2001):

$$\rho = \frac{p_{AA} - p_A^2}{p_A(1 - p_a)} = \frac{p^2 + \sigma_p^2 - p^2}{p(1 - p)} = \frac{\sigma_p^2}{p(1 - p)}$$  \[23\]

We already noticed that the whole population is not in HWE as a result of the Wahlund effect. In addition, the above equation demonstrate that with PS and allele frequency differences over strata, that is $\sigma_p^2 \neq 0$, there is an association between alleles at a single locus, even though for each of the $k$ subpopulations HWE conditions might hold.

2.4.3 Population admixture

Population admixture refers to a situation where individuals in a population have different ancestries as a result of mixing of several genetically distinct populations (LAIRD and LANGE 2011; REDDEN and ALLISON 2006). Typically, admixed populations are the result of migration (LAIRD and LANGE 2011). African-American or European-American populations are the well-known examples of admixed populations (PARRA et al. 1998; PFAFF et al. 2001). A famous example in animal breeding literature is the crosses made between the two cattle subspecies, *Bos Taurus* (Taurine) and *B. indicus* (FREEMAN et al. 2006). The former were domesticated in the Middle East, Anatolia, and perhaps Africa ~10,000 years ago and now can be found across northern Eurasia and some parts of Africa. The *B. indicus* or Zebu cattle, originated from a different variant of wild cattle ancestors, were domesticated on the Indian subcontinent (FREEMAN et al. 2006). Given the allele frequency differences
among the original ancestral populations, the probability that a given individual carries a specific allele depends on the individual’s mixture of ancestry (Laird and Lange 2011).

2.4.4 Difference between PS and population admixture

PS and admixture are usually occurring together and the same statistical models are used to describe these two phenomena. However, this does not mean that they are necessarily the same (Foulkes 2009; McKiege 2008). Admixture between genetically divergent populations creates gametes consisting of a mosaic of segments inherited from each of the parental populations. The key point about admixture is that it generates LD that decays with map distance, whereas PS creates LD between loci irrespective of their linkage status (McKeige 2008; Pfaff et al. 2001). Consider two loci A and B being carried on a gamete generated by a parent with mixed ancestry. The likelihood of the loci A and B being originated from the same ancestral sub-population is proportional to their map distance. Therefore, the age of the admixture process might be inferred based on marker data from a sample of admixed individuals (Chakraborty and Weiss 1988; McKeige 2008).

Admixed populations might be used for linkage mapping by conducting tests of associations conditional on the parental sub-populations admixture proportions to eliminate bias resulting from admixture (e.g., Parra et al. 1998; Pfaff et al. 2001). This approach, known as Mapping by Admixture Linkage Disequilibrium (MALD), was first introduced by (Chakraborty and Weiss 1988) and was extended by others (see Parra et al. 1998; and references there in). “MALDsoft” (Montana and Pritchard 2004) is one of the software specifically developed for QTL mapping in admixed populations.

Mapping in an admixed population is both a challenge and an opportunity (Montana and Pritchard 2004). It is an opportunity because of the existence of long-range LD across
the genome, of the order of 10-20 cM in human populations for example (KAPLAN et al. 1998; PARRA et al. 1998), only a fraction of the markers needed for a typical whole-genome association study would be enough (MONTANA and PRITCHARD 2004; PARRA et al. 1998). The caveat with MALD is that it requires populations that are highly diverged from each other. If this is the case, then the power of MALD is larger than the LA and is comparable to the PBA study (MONTANA and PRITCHARD 2004). However, the admixed populations are prone to the cofounding due to variation in ancestry (MONTANA and PRITCHARD 2004).

2.4.5 Confounding due to PS and admixture

In the previous section we showed how PS leads to a deficit in the number of heterozygote genotypes relative to their expected number under HWE and also noticed that the variance of the genotypic distribution being inflated. If an association study is conducted in such populations without properly accounting for the effect of PS on the genotypic distribution, the variance of association tests might be underestimated. This would lead to a remarkable increase in the number of false-positive findings relative to the number that is expected based on the specified significance level (LAIRD and LANGE 2011).

A well-known example that clearly shows the confounding effect of PS on association study results is the study of KNOWLER et al. (1988) who conducted an association study in American Indians with genetic admixture. In their study, the authors observed a very strong negative association between the Gm haplotype Gm$^{3,5,13,14}$, which is located on a locus of the human immunoglobulin G gene, and type-2 diabetes mellitus. The population under study had different degrees of American Indian and European Hispanic history. Figure 2.1 shows the distribution of the Gm haplotype and the prevalence of diabetes disease across different levels of Indian ancestry. As it can be seen from the figure, for persons with the highest level
of Indian ancestry the frequency of the Gm haplotype is nearly zero, whereas the opposite is true for those with no Indian ancestry in their pedigree. The risk of the disease varies inversely with the degree of Indian admixture, even though that the direct role of the Gm haplotype on diabetes susceptibility is known to be very unlikely (Knowler et al. 1988). Markers like Gm haplotype, which are indicator of population subtypes, are called Ancestry Informative Markers (AIM) (e.g., Montana and Pritchard 2004). In the study of the European genetic contribution to several populations of African descent in the United States, Parra et al. (1998) identified several population-specific markers that were ancestry informative. This example represents a common problem in genetic epidemiology studies known as confounding due to population stratification (e.g., Cordell and Clayton 2005; Laird and Lange 2011; McKeigue 2008). As an example, consider a study aiming at determining whether there is an association between a single genetic polymorphism and lung cancer. If the polymorphism being tested increases the chance of an individual to smoke, and in turn it is known that smoking causes lung cancer then the marker is in the causal pathway to the disease or is an effect mediator (See Figure 2.2). It is important to know that a confounding factor like PS is associated with both the marker and the trait but is not within the causal pathway of the trait (Foulkes 2009).

As we noticed, in the study of Knowler et al. (1988) the confounding factor, i.e., the genetic ancestry was associated with both the genotypes at the marker and disease risk. Figure 2.3 depicts confounding due to PS in a case-control association study (DeWan et al. 2007). As it can be seen, there is an undetected population structure in the sample and the two sub-populations vary in the tested marker genotype frequencies. The frequency of alleles A and B are higher in sub-populations 1 and 2, respectively. Owing to the fact that
more cases have been selected from sub-population 2, the disease status is misleadingly associated with the genotypes at this marker. This indicates that the cases and controls have not been matched with respect to their ancestry carefully, although they should (DeWan et al. 2007).

The coupling of allele frequency heterogeneity to the disease prevalence heterogeneity (or, in case of a quantitative phenotype, the trait mean) across the sub-populations, confounds the disease-marker association. To avoid such confounding, the case-control association studies are often conducted in a single race or ethnicity group, or at least the data is analyzed within each stratum (Epstein et al. 2007). Campbell et al. (2005) demonstrated the effect of population stratification on human height in a sample of European-American individuals. They created a case-control study in a sample of adult individuals ranked as tall and short in height. All the individuals were US born and were self-described as “white” or “Caucasian”, but their grandparents were either born in the US or Europe. The authors knew that height varies across European populations. Thus, their purpose was that to determine whether matching of the cases and controls by age, country of birth (US) and self-described ethnicity is sufficient to protect their findings against PS or not. A highly significant ($P < 10^{-6}$) association was found between the tall/short status and a marker located within the lactase genet (LCT), which had a frequency that varied considerably from northern to southern Europe. However, when they repeated the association test conditional on the grandparental ancestry a much weaker signal was found ($P < 0.005$). Additional dataset from Poland and Scandinavia ($P > 0.05$) proved that the LCT association was a false-positive (Campbell et al. 2005).
2.5 TESTS OF GENETIC ASSOCIATIONS IN PBA STUDIES

Like any other type of statistical association analysis, the purpose of a genetic association test is to establish an association between, or examine independence of, two variables: a trait of interest and a genetic marker (FOULKES 2009; LAIRD and LANGE 2011). If the marker being tested is known to be a neutral locus without any known effects on DNA coding, then the LD between the marker and a QTL affecting the trait could be a valid reason for the observed association (LAIRD and LANGE 2011). Before we start discussing different tests of association, first we show how LD between a marker and a QTL induces marker-trait association. Following LAIRD and LANGE (2011) we consider a disease susceptibility locus (DSL) $D$ with alleles $D$ and $d$ (disease allele and non-disease allele, respectively), in a case-control study that consists of equal number of individuals in the two groups. Let $P(D|Cases)$ and $P(D|Controls)$ be the frequency of allele D in cases and controls, respectively. Then a test of association between the DSL and the disease can be stated as below:

$$H_0: \Delta_D = P(D|Cases) - P(D|Controls) = 0$$  \[24\]

Since we usually do not observe the DSL, we genotype all individuals at a marker locus $B$ with alleles $B/b$, instead. In the same way, we have

$$\Delta_B = P(B|Cases) - P(B|Controls) = 0$$  \[26\]

It can be argued that conditional of the DSL genotype, $P(Disease)$ is independent of the marker genotype. Therefore, according to PRITCHARD and PRZEWORSKI (2001):

$$\Delta_B = \Delta_D (P(B|D) - P(B|d))$$  \[27\]

It is clear that in the absence of LD between the two loci, $P(B|D) = P(B|d) = P(B)$ and hence $\Delta_B = 0$ (LAIRD and LANGE 2011). This simple derivation shows that unless the marker
and the causative mutation alleles are in LD with each other, there will be no association between the marker being tested and the trait of interest (Laird and Lange 2011).

Now we briefly introduce some simple and popular statistical measures and tests of associations for binary and quantitative traits, respectively. The choice of the proper statistical procedure to examine the marker-trait association depends on the nature of the data and the hypothesis being tested (Foulkes 2009). In addition, the parameter estimation and the hypothesis testing might be based on pre-adjusted or un-adjusted data for the nuisance variables, e.g., confounding effects, and involves either univariate or multivariate statistical methods (Foulkes 2009).

2.5.1 Tests of independence or contingency tables for binomial traits

There are several well-known tests commonly used in the case-control association studies, such as the Pearson’s $\chi^2$-test and the Fisher’s exact test (e.g., Foulkes 2009). The former sometimes is referred to the score test as well (Clayton 2008; Morris and Cardon 2008; Zhu and Zhang 2010).

**Pearson’s $\chi^2$-test**

Fisher (1925) developed test of independence or contingency table to examine the association of two variables each categorized into several levels. To start, consider a sample of unrelated cases, showing the trait of interest and their corresponding controls with the opposite form of the trait. Having all individuals typed at marker $A$, we could establish the contingency table as Table 2.1. Then the Pearson’s $\chi^2$ statistic is calculated as (e.g., Morris and Cardon 2008; Siegmund and Yakir 2007):

$$\chi^2 = \sum_{i=0,1,2} \left( \frac{(r_i - E[r_i])^2}{E[r_i]} + \frac{(s_i - E[s_i])^2}{E[s_i]} \right)$$ [28]
where \( E[r_i] = \frac{r \cdot n}{n} \) and \( E[s_i] = \frac{s \cdot n}{n} \)

Under the null hypothesis of independence of the rows and columns (i.e., no association between marker genotypes and the binary trait) the \( \chi^2 \) statistic has an approximate \( \chi^2_{(r-1)(c-1)} \) distribution, where \( r \) and \( c \) are number of rows and columns of the contingency table (MORRIS and CARDON 2008; SIEGMUND and YAKIR 2007).

2.5.2 **Alleles test**

If the genotype frequencies satisfy the HWE condition, the test of marker genotypes and the trait association can be replaced by the alleles test (DEWAN et al. 2007; LAIRD and LANGE 2011). To run the test the number of \( A_1 \) alleles for the marker being tested is counted and then the association test is conducted based on a 2x2 contingency table. LAIRD and LANGE (2011) derived an equivalent expression of the alleles test which is very similar to the trend test statistics (see below). The null hypothesis of no marker-trait association can be formulated as

\[
H_0: P(A_1|\text{Cases}) = P(A_1|\text{Controls}) = p_{\text{Cases}} - p_{\text{Controls}} = 0
\]  

[29]

and the test statistic is:

\[
Z_{AT} = \frac{2\sqrt{rs(\hat{p}_{\text{Cases}} - \hat{p}_{\text{Controls}})}}{\sqrt{2np(1 - p)}}
\]  

[30]

where \( p \) is the frequency of allele \( A_1 \) in the whole sample. Under the null hypothesis, \( Z_{AT} \sim N(0,1) \) and \( Z_{AT}^2 \sim \chi^2_1 \) (LAIRD and LANGE 2011)

2.5.3 **Fisher’s exact test**

When at least 20% of the expected cell counts in the contingency table are small,
(E[n_{ij}]< 5), the asymptotic assumptions of the $\chi^2$-test are violated (Foulkes 2009; Vittinghoff et al. 2005). Fisher (1925) derived the required exact p-values for a 2x2 contingency table based on the hyper-geometric distribution. Fisher’s exact test is preferred over the $\chi^2$-test although it is computationally more demanding, but it has been implemented in R and other software (Balding 2006; Vittinghoff et al. 2005).

### 2.5.4 Cochran-Armitage trend test

The C-A trend test is a common test of association in case-control studies, especially when the additive gene action is assumed as the mode of inheritance for the trait under study (Foulkes 2009; Laird and Lange 2011). Considering the bi-allelic marker $A$, and the same null hypothesis as in [29] the test statistic can be formulated as:

$$Z_T = \frac{2\sqrt{F_s(p_{\text{Cases}} - p_{\text{Controls}})}}{\sqrt{4n_2 + n_1 - 4np^2}}$$

[31]

Under $H_0$ of equal frequency of allele $A_1$ in both cases and controls, the $Z_T$ is approximately $N(0,1)$ and $Z_T^2$ is approximately $\chi^2_1$ (Laird and Lange 2011). If the trait of interest is disease status then the trend test is basically evaluating whether the probability of disease changes linearly with the marker genotype (Foulkes 2009; Laird and Lange 2011). The C-A trend test has the advantage of not relying on the HWE assumption compared to the alleles test and hence it is preferred (Laird and Lange 2011).

Comparing the $Z_{AT}$ and $Z_T$ defined above reveals that: the test of hypothesis of no association is essentially proportional to the allele frequency differences between the cases and control and more importantly shows that the two tests are only differ in their denominators, i.e., variances (Devlin and Roeder 1999; Laird and Lange 2011). In fact, the two tests are different in their assumption regarding independence of the uniting gametes. The alleles test
is based on the allele counts and hence relies on the assumption of HWE (e.g., BALDING 2006), in other words it assumes that the uniting gametes of an individual are independent (LAIRED and LANGE 2011). DEVLIN and ROEDER (1999) showed that the ratio of \( Z_{AT} \) to \( Z_T \) is approximately equal to \( (1+F) \), which is the variance of allele frequencies across sub-populations relative to that in a population in HWE (see equations [16] and [21]). They concluded that unlike the alleles test, the trend test automatically accounts for the extra binomial variance induced by correlation of the uniting gametes. Nevertheless, the trend test does not account for the correlation of uniting gametes across individuals; instead it assumes that the genotypes of individuals are independent. This assumption is not valid when population stratification exists or if individuals within the strata are related and leads to spurious false-positives associations (DEVLIN and ROEDER 1999).

2.5.5 Log-Likelihood ratio test

Similar to the \( \chi^2 \)-test, the LRT is another standard asymptotic test of hypothesis of no association (CLAYTON 2008; LIN and ZHAO 2010). For our example marker shown in Table 2.1, the LRT test statistics can be calculated as:

\[
LRT = 2 \sum_{i=0,1,2} (r_i \log \frac{r_i}{E[r_i]} + s_i \log \frac{s_i}{E[s_i]})
\]

Under the null hypothesis LRT is approximately distributed as \( \chi^2 \). The LRT is more difficult to calculate than the score test or the \( \chi^2 \)-test because it is based on the iterative computation of the ML estimates of the common odd-ratio (CLAYTON 2008).

2.5.6 Test of hypothesis of no association in the presence of PS

For simplicity, consider a population consisting of two distinct but undetected sub-populations and a Mendelian disorder, scored as case or control with prevalence \( v_1 \) and \( v_2 \) in
the sub-populations, respectively. Also, consider marker $A$, with alleles $A/a$ and frequency of $p_1$ and $p_2$ in the sub-populations, that is not in LD with any trait locus. Further, let’s assume that within each stratum there is random mating (SIEGMUND and YAKIR 2007). Again, our null hypothesis is that the frequency of $A_1$ is the same in cases and controls. As we demonstrate below, when $p_1 \neq p_2$ and $\nu_1 \neq \nu_2$ our test of hypothesis may result in a positive association (CLAYTON 2008; LAIRD and LANGE 2011; SIEGMUND and YAKIR 2007), despite the fact that marker $A$ is not in LD with any trait locus. Let $c$ designate the proportion of the cases and $d$ the proportion of controls falling into the first sub-population. Then, following SIEGMUND and YAKIR (2007):

$$p_{\text{Cases}} = cp_1 + (1 - c)p_2 \quad [33]$$

and

$$p_{\text{Controls}} = dp_1 + (1 - d)p_2 \quad [34]$$

Then,

$$p_{\text{Cases}} - p_{\text{Controls}} = p_1(c - d) + p_2[(1 - c) - (1 - d)] = (p_1 - p_2)(c - d) \quad [35]$$

LAIRD and LANGE (2011) showed that

$$E(c - d) = \frac{S_1S_2(\nu_1 - \nu_2)}{\nu(1 - \nu)} \quad [36]$$

where $\nu$ is the disease prevalence in the whole population and $S_1$ and $S_2$ are the proportion of subjects in sup-populations 1 and 2, respectively.

Remember that both the alleles test and the C-A trend test (equations [30] and [31]) had $p_{\text{Cases}} - p_{\text{Controls}}$ in their numerator. Thus, the numerator of both test statistic will have a non-zero expectation unless either $p_1 = p_2$ or $\nu_1 = \nu_2$. One should note the problem of spurious association gets worse as the sample size increases (SIEGMUND and YAKIR 2007).
Obviously, if the sub-populations are identified, a natural remedy to avoid spurious association will be a stratified association analysis (CLAYTON 2008; LAIRD and LANGE 2011). As the term suggest, by stratified analysis we mean analyzing the data within each level of the PS (Figure 2.4). A common caveat to this approach is the lack of sufficient power to detect association signals (CLAYTON 2008).

2.5.7 Tests of associations for quantitative traits

It is interesting that the vast majority of QTL have been identified using simple statistical methods (GEORGES 2007). For a quantitative trait, usually the hypothesis can be formulated as whether the genotypic variation at a randomly selected neutral marker in a population could account for some of the phenotypic variation of the trait. A positive answer to such question might imply linkage between the marker and a QTL affecting the trait (HAYES et al. 2004). Common statistical methods like linear regression of phenotypes on marker genotypes or the analysis of variance (ANOVA) can be used for test of association for quantitative traits (e.g., BALDING 2006; BROMAN and SEN 2009; GEORGES 2007; JANNINK et al. 2001; Wu et al. 2007).

2.5.8 M-sample and non-parametric tests of association for a quantitative trait

A two-sample t-test might be used to test the hypothesis of $H_0: \mu_1 - \mu_2$, where $\mu_1$ and $\mu_2$ are the trait population means for the two genotype groups of interest (e.g., AA and Aa vs. aa). The t-statistic, assuming equal variance between the two genotype groups is calculated as:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_p^2 \left[ 1/n_1 + 1/n_2 \right]}}$$

[37]
where \( \bar{y}_1 \) and \( \bar{y}_2 \) are the sample means of the quantitative trait for the genotypes groups with samples sizes of \( n_1 \) and \( n_2 \), respectively and \( s_p^2 \) is the pooled estimate of the residual variance. Under the null hypothesis, the t-statistic has a t-distribution with \( n_1 + n_2 - 2 \) degrees of freedom (FOULKES 2009; Wu et al. 2007). A problem with using the t-statistic is that it does not have the usual Student t-distribution under \( H_0 \) because the underlying distribution of the response variable is a mixture distribution (Wu et al. 2007). One way to get around with this issue is to use permutation test which gives the distribution of the test statistic given the null hypothesis is true (Wu et al. 2007).

Alternatively, if the trait is not normally distributed or the sample size is small a non-parametric test known as the Wilcoxon rank-sum test (also called the Mann-Whitney U-test) can be applied to the data (FOULKES 2009). ANOVA and Kruskal-Wallis (K-W) tests are the M-sample extensions of the two-sample t-test and the Wilcoxon rank test, respectively (FOULKES 2009). If sample size is small to the extent that the assumption of normality of the response variable is not reasonable, the K-W test is a better choice than the ANOVA (FOULKES 2009).

### 2.5.9 Generalized linear model

A difficulty with the regular C-A trend test or the \( \chi^2 \)-test is that consideration of additional covariates, such as age or sex, in the model is not straightforward. Alternatively, regression models have the advantage of allowing to simultaneously controlling for several confounding factors (FOULKES 2009). The general linear model (GLM) can be expressed in matrix notations as:

\[
g(\mathbb{E}(Y|X)) = X\beta
\]  

[38]
where $E(\mathbf{y})$ is the expected value of $\mathbf{Y}$, $g(.)$ is the link function and $\mathbf{X}$ is the design matrix. In case of a quantitative trait, the link function is the identity link, for example (Foulkes 2009).

Thus, for a normally distributed phenotype the GLM is:

$$g(E(Y|\mathbf{X})) = E(Y) = \mathbf{X}\beta, \text{ with } \beta = (\beta_0, \beta_1)'$$

or equivalently

$$Y = \mathbf{X}\beta + \epsilon$$

For a single marker association analysis, the linear model can be written in scalar form as:

$$y_i = \beta_0 + \beta_1 + \epsilon_i$$

Then, the null hypothesis of no association between the marker locus and the trait of interest will be:

$$H_0: \beta_1 = 0$$

The above model assumes that the residuals are independently and identically distributed with 0 mean (Foulkes 2009).

### 2.5.10 Logistic regression

Logistic regression can be viewed as an extension of the standard regression analysis (Christensen 1997). Consider the random variable $\mathbf{X}$ as the number of $A_1$ alleles in an individual’s genotype, and take $\mathbf{Y}$ as a dichotomous phenotype (affected and non-affected, for example). Unlike the case of quantitative trait, to build a model that relates the disease status to the marker genotypes we need to establish a link function between the two (Andersen and Skovgaard 2010). To do this, let’s first define the penetrance function as the conditional probability of an individual’s phenotype given its genotype as $\pi = P(Y|\mathbf{X})$.

If the marker locus under the test is not associated with the phenotype, then the penetrance probabilities of all genotypes at the marker locus will be equal irrespective of the individual’s
genotype at the marker (LAIRD and LANGE 2011). The penetrance function can be equivalently expressed as $E(Y|X)$. We may think of modeling the relation between the outcome and the predictor linearly as:

$$E(Y|X) = \beta_0 + \beta_1 x + \varepsilon$$  \[43\]

Note that the right-hand side of the above equation is bounded between 0 and 1. The ordinary least square (OLS) is not suitable for a bounded response variable because it will produce fitted values outside of the permitted range (WEISBERG 2005). Therefore, for a binary outcome we typically assume that some function of the basic parameter (here $E(Y|X)$) depends linearly on $X$, rather than the parameter itself. This function is called the *link function* (ANDERSEN and SKOVGAARD 2010). For binary traits, the link function is typically *log* or *logistic* link (LAIRD and LANGE 2011). The *logit* link is defined as:

$$\text{logit}(\pi) = \ln \left( \frac{\pi}{1-\pi} \right) = \beta_0 + \beta_1 x + \varepsilon$$  \[44\]

where $\pi = (\pi_1, \pi_2, \ldots, \pi_n)'$, for a sample of size $n$ individuals, and $\pi_i = \Pr (y_i = 1|x_i) = E(Y|X)$ (ANDERSEN and SKOVGAARD 2010; FOULKES 2009). Note that the ratio $\frac{\pi_i}{1-\pi_i}$ is the odds of the success (affection, in our example), hence the above model assumes that the log-odds are a function of the predictor $x$ (marker genotype) (CASELLA and BERGER 2002). The above equation can be rewritten in scalar formulation as:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$  \[45\]

Parameters of the model can be estimated using maximum likelihood. As the OLS here $\beta_1$ is the change in the log-odds of success corresponding to a one-unit increase in $x$ (CASELLA and BERGER 2002).
2.6 ANALYTICAL CHALLENGES OF HIGH-DIMENSIONAL GWAS DATA

Based on the large-scale sequencing and genotyping projects conducted over the last few years, it has been estimated that there are 10-15 SNPs with MAF > 1% in the human genome (DeWan et al. 2007). The high-throughput genotyping technology that generate a huge amount of high-dimensional data poses a series of statistical challenges as well (DeWan et al. 2007). Studies involving a large number of genetic markers that are potentially informative are faced with two analytical difficulties: (1) the multiplicity of statistical tests, i.e., the inflation of the Type I Error due to multiple-testing. Under this condition, significance of association of a single marker is meaningless without appropriately adjusting for multiple comparisons (Foulkes 2009; Freimer and Sabatti 2005), (2) the complex, generally unknown relationships among the genetic markers under consideration (Foulkes 2009).

2.6.1 Control of false-positives in GWAS

We briefly describe some of the common methods of multiple-testing adjustments in this section. A comprehensive discussion on the current methods with implementation in R and SAS can be found in Dudoit and van der Laan (2008). Following Sabatti (2007) here we assume that false-positives are potentially due to random chance, i.e., all confounding factors have been properly accounted for by study design or statistical model. To start with, let’s assume that we have conducted a whole-genome association study and are testing the association of the \( m \) markers with a phenotype of interest. Let \( m \) be the number of tests, with \( T = (T_1, T_2, ..., T_m) \) test statistic calculated to examine whether any of the null hypotheses of \( H_0 = (H_0^1, H_0^2, ..., H_0^m) \) is true. In such a study, we are often deal with two questions (Lehmann and Romano 2005; Sabatti 2007): (1) Can \( H_0 \) be rejected? In other words,
Based on the sample data we collected can we reject the hypothesis that there is no genetic basis for the trait? and (2) If $H_0$ is not rejected, nothing needs to be done but if it was rejected one would ask which of the $H_i$ should be rejected? This is more relevant to the identification of the genomic regions declared as associated with the trait. The statistical technique used to answer the first question is called global test, whereas the one addressing the second question are titled the multiple-testing procedure (Sabatti 2007). One might think of ANOVA procedure a global test and the step following the rejection of null hypothesis, i.e., finding which treatment means are different from the other, as the multiple-testing step, for example (Foulkes 2009). To deal with the multiple-testing issue, a new measure of the Type I Error needs to be defined. Let’s consider Table 2.2. Much of the literature on methods for adjusting for multiple-testing describes controlling one or two error rates: the family-wise error rates and the false discovery rate (Foulkes 2009; Sabatti 2007).

2.6.1.1 Family-wise error rate

The FWER is the probability of committing at least one false rejection:

$$FWER = Pr (V > 0)$$  \[46\]

This is a quite stringent threshold and is typically applicable when the costs of making a wrong decision are really high (Sabatti 2007).

2.6.1.2 False discovery rate

The FDR was introduced by Benjamini and Hochberg (1995). It is defined as the expected proportion of null hypotheses that are true (V) among those that are declared significant (V+S). Thus, defining $Q$ as
\[ Q = \begin{cases} \frac{V}{V + S}, & \text{if } V + S > 0 \\ 0, & \text{otherwise} \end{cases} \]  

\[ FDR = E(Q) \]

For \( R = 0 \), we defined \( Q = 0 \) since no false rejection has been made. Equivalently, the FDR can be expressed as (FOULKES 2009):

\[
FDR = E\left( \frac{V}{R} \middle| R > 0 \right) \cdot Pr(R > 0) + E\left( \frac{V}{R} \middle| R = 0 \right) \cdot Pr(R = 0)
\]

\[ = E \left( \frac{V}{R} \middle| R > 0 \right) \cdot Pr \left( R > 0 \right) \]  

Note that FDR will be equal to FWER if all null hypotheses are true, but in general for a certain threshold the FDR is more liberal, i.e., \( FDR \leq FWER \) (DUDOIT and VAN DER LAAN 2008; FOULKES 2009; LEHMANN and ROMANO 2005). Therefore, the procedures controlling the FWER are typically more conservative than those controlling the FDR (FOULKES 2009; HUBER et al. 2008). In general, any approach that controls FWER will control the FDR as well, but the opposite of this is not true (FOULKES 2009). For the whole-genome association studies (GWAS) of a quantitative trait, where one expects a large number of tests to be false, the choice between controlling of FWER or FDR is important. In fact, as the number of false null hypotheses increases, i.e., \( m - m_0 \) in table 2.2, the number of true-positives, \( S \), will also increase leading to a smaller \( V/R \) ratio. This is the scenario where there will be a large difference between the FDR and FWER (FOULKES 2009). As a general rule, for discovery step of a whole-genome association study where one needs a higher power to detect associations using the FDR is preferred to the FWER (FOULKES 2009).
2.6.1.3 **Bonferroni correction for multiple comparisons**

*Bonferroni* correction is a single-step procedure and perhaps the most straightforward adjustment to apply. The procedure suggests that testing each hypothesis at the level $\alpha/m$ controls FWER at the level $\alpha$ (Dudoit and van der Laan 2008; Sabatti 2007). The *Bonferroni* procedure does not take into account the dependency between tests. Consider the scenario that someone has conducted $k$ identical tests and wants to control FWER at the level $\alpha$ for each test, in this case he simply can use level $\alpha$ for each single test (Sabatti 2007). This shortcoming of the test makes it overly conservative, because in GWAS we expect that adjacent markers to be correlated to each other due to the background LD across the genome (Gao et al. 2008; Li and Ji 2005; Morris and Cardon 2008; Sabatti and Risch 2002).

2.6.1.4 **Effective number of tests**

Proper modeling of the correlation structure of markers in a large-scale GWAS that examines tens of thousands markers is important because the appropriate correction for multiple-testing should be based on the possible number of independent tests, rather than the actual number of tests being carried out (Freimer and Sabatti 2004). Simply stated, if $H^C = (H^1_0, H^2_0, \ldots, H^m_0)$ is the complete set of tests and $H^S_0$ is a subset of it comprising of $S$ independent tests, then the *Bonferroni* adjustment would have been $\alpha/S$ rather than $\alpha/m$. This has motivated development of the idea of *the effective number of tests* ($M_{\text{eff}}$) (Cheverud 2001; Han et al. 2009; Li and Ji 2005; Moskvina and Schmidt 2008; Nyholt 2004; Pe'Er et al. 2008).

Consider a set of $M$ traits and their corresponding correlation matrix. If one calculates the eigenvalues ($\lambda$s) of the correlation matrix, then the variance of these eigenvalues ($V_{\lambda\text{obs}}$) can be used to measure the total correlation among the traits (Cheverud 2001). The higher the
correlation among the traits the larger is $V_{\lambda_{obs}}$. For instance, if the correlation between all pairs of traits is zero, then all the $\lambda$'s are equal to one and thus $V_{\lambda_{obs}} = 0$. On the other hand, if all traits are maximally correlated the first $\lambda$ is equal to the number of traits and the remaining $\lambda$'s are all zero. This situation generates the maximum $V_{\lambda_{obs}}$ that is equal to $M$ (Cheverud 2001). Thus, $V_{\lambda_{obs}}$ varies between 0 and $M$ depending on the strength of the correlation among the variables represented in the correlation matrix (Cheverud 2001). The proportional reduction in the number of tests due to correlation is $V_{\lambda_{obs}}/M$ (Cheverud 2001). Thus the effective number of tests can be calculated as (Cheverud 2001; Nyholt 2004):

$$M_{\text{eff}} = 1 + (M - 1) \left[ 1 - \frac{V_{\lambda_{obs}}}{M} \right]$$  \hspace{1cm} [50]

Note that if all markers are perfectly correlated then $V_{\lambda_{obs}} = M$ and $M_{\text{eff}} = 1$. Whereas, when they are all independent, $V_{\lambda_{obs}} = 0$ and $M_{\text{eff}} = M$. Once the effective number of tests has been determined, then the Bonferroni adjusted significance threshold will be (Foulkes 2009):

$$\alpha'' = 1 - (1 - \alpha)^{1/M_{\text{eff}}} \approx \frac{\alpha}{M_{\text{eff}}}$$  \hspace{1cm} [51]

Several different alternatives have been proposed for the correlation matrix among the set of markers being tested, ranging from the matrix of the Pearson’s correlation between markers genotypes (coded 0, 1, and 2) (Cheverud 2001), to the one based on the pair-wise LD between markers ($r^2$) (Nyholt 2004), for example (Foulkes 2009).
2.6.2 Model selection in GWAS

The primary objective of QTL mapping is to identify a set of loci that contribute most to the observed phenotype. Therefore, QTL mapping is best viewed as a model selection or variable selection problem (Broman and Sen 2009; Wang et al. 2005). A model selection procedure for QTL mapping has at least two main parts: a model search algorithm and a model comparison criterion (Broman and Sen 2009). For model searching, we need a method that explores the model space to identify the good ones, recognizing the fact that we are only able to visit a very small fraction of the model space. On the other hand, model comparison is a procedure that balances between model complexity and quality of fit. Considering an additive QTL model and assuming that markers are on the top of QTL, then for a GWAS with only 100 markers genotyped the number of possible models is \(2^{100} \approx 10^{30}\) (Broman and Sen 2009). Hence, for a typical GWAS consisting of a couple of thousands subjects and hundreds of thousands of markers genotyped, it is impossible to analyze the data using traditional multivariate regression (Li et al. 2010).

Alternative approaches, therefore, have been developed, including ridge regression (Hoerl and Kennard 1970), least absolute shrinkage and selection operator (LASSO) (Tibshirani 1994), and the Bayesian variable selection (e.g., George and McCulloch 1993). Ridge regression is classified as a model-selection-free approach or shrinkage estimator, because it keeps all potential model effects in the model but the estimated effects are forced to shrink toward zero (Wang et al. 2005). Whittaker et al. (2000) implemented the ridge regression in the context of MAS to improve the accuracy of selection. The method has been applied to genomic selection as well (Habier et al. 2007; Meuwissen et al. 2001; Xu 2003). The difficulty of ridge regression is its simplifying assumption of equal and fixed
marker effect variances which can lead to over-shrinkage of large effects (HEFFNER et al. 2009). Bayesian methods have relaxed this assumption and allow a marker-specific variance is estimated for each marker. Different aspects of these methods have been fully discussed elsewhere (e.g., GIANOLA et al. 2009; HABIER et al. 2011; HEFFNER et al. 2009; HOGGART et al. 2008).

2.7 LITERATURE CITED

CHRISTENSEN, R., 1997 Log-Linear Models and Logistic Regression. Springer-Verlag, USA.


DUDOIT, S., and M. J. VAN DER LAAN, 2008 Multiple Testing Procedures with Applications to Genomics. Springer Science+Business Media LLC, USA.


HEDRICK, P. W., 2005 Genetics of populations. Jones and Bartlett Publishers, Sudbury, MA USA.


XU, Y., 2010 Molecular Plant Breeding. CAB International, Oxfordshire, UK.


Zöllner, S., 2001 Population History and Linkage Disequilibrium in eLS. John Wiley & Sons, Ltd.
2.8 FIGURES

**Figure 2.1** Percentage of individuals with diabetes disease in Gila Indian River population along with frequency of Gm haplotype in subpopulations with different Indian heritage ancestry. The X axis is the number of great grandparents with Indian heritage in an individual’s pedigree. Data adopted from (KNOWLER et al. 1988).

1. The X axis is the number of great grandparents with Indian heritage in an individual’s pedigree. Data adopted from (KNOWLER et al. 1988).
**Figure 2.2** An effect mediator is in the causal pathway between the genotype and the trait under study. Consider a study aiming at determining whether there is an association between a single genetic polymorphism and lung cancer. If the polymorphism being tested increases the chance of an individual to smoke, and in turn it is known that smoking causes lung cancer then the marker is in the causal pathway to the disease or is an *effect mediator*.

1. Adopted from (FOULKES 2009)
Figure 2.3 False association at a marker locus due to population stratification in a case–control study\(^1\). The schematic shows the underlying problem of PS in a case-control study. The allele frequency of the marker of interest is not the same in the two sub-populations. Because more cases are falling into the sub-population 2, it would appear as if there is an association between the disease and the marker. However, this happens due to the fact that the marker has different frequencies in the two sub-populations.

1. Adopted from (DEWAN et al. 2007)
Figure 2.4 Stratified analysis of case-control studies, based on either genotypes or allele counts.

1. Adopted from (CLAYTON 2008)
### 2.9 TABLES

**Table 2.1** A 2x3 contingency table for marker-trait association test in a case-control study

<table>
<thead>
<tr>
<th>Phenotype</th>
<th># A&lt;sub&gt;1&lt;/sub&gt; alleles</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Case</td>
<td>r&lt;sub&gt;0&lt;/sub&gt;</td>
<td>r&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
<tr>
<td>Control</td>
<td>s&lt;sub&gt;0&lt;/sub&gt;</td>
<td>s&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
<tr>
<td>Total</td>
<td>n&lt;sub&gt;0&lt;/sub&gt;</td>
<td>n&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
</tbody>
</table>
Table 2.1 Possible outcomes of multiple hypotheses testing of an experiment with m markers

<table>
<thead>
<tr>
<th></th>
<th>No. Not Rejected</th>
<th>No. Rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>H$_0$ is True</td>
<td>U</td>
<td>V</td>
<td>m$_0$</td>
</tr>
<tr>
<td>H$_0$ is false</td>
<td>T</td>
<td>S</td>
<td>m$_1$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$m-R$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>
CHAPTER 3. GENOMIC SELECTION IN ADMIXED AND CROSSBRED POPULATIONS


Ali Toosi*, Rohan L. Fernando*1, and Jack C.M. Dekkers*

*Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, IA 50011 USA

3.1 SUMMARY

In livestock, genomic selection (GS) has primarily been investigated by simulation of purebred populations. Traits of interest are, however, often measured in crossbred or mixed populations with uncertain breed composition. If such data are used as the training data for GS without accounting for breed composition, estimates of marker effects may be biased due to population stratification and admixture. To investigate this, a genome of 100 cM was simulated with varying marker densities (5 to 40 segregating markers per cM). After 1000 generations of random mating in a population of effective size 500, four lines with effective size 100 were isolated and mated for another 50 generations to create 4 pure breeds. These breeds were used to generate combined, F1, F2, three- and four-way crosses, and admixed training data sets of 1000 individuals with phenotypes for an additive trait controlled by 100 segregating QTL and heritability of 0.30. The validation data set was a sample of 1000

1 Corresponding author: Rohan Fernando, 237 Kildee Hall, Department of Animal Science, Iowa State University, Ames IA 50011, USA
Phone: +1 515 294 5348, Fax: +1 515 294 9150, e-mail: rohan@iastate.edu
genotyped individuals from one pure breed. Method Bayes-B was used to simultaneously estimate the effects of all markers for breeding value estimation. With 5 (40) markers per cM, the correlation of true with estimated breeding value of selection candidates (accuracy) was highest, 0.79 (0.85), when data from the same pure breed was used for training. When the training dataset consisted of crossbreds, the accuracy ranged from 0.66 (0.79) to 0.74 (0.83) for the two marker densities, respectively. The admixed training dataset resulted in nearly the same accuracies as when training was in the breed to which selection candidates belonged. However, accuracy was greatly reduced when genes from the target pure breed were not included in the admixed or crossbred population. This implies that, with high-density markers, admixed and crossbred populations can be used to develop GS prediction equations for all pure breeds that contributed to the population, without a substantial loss of accuracy compared to training on purebred data, even if breed origin has not been implicitly taken into account. In addition, using GS based on high-density marker data, purebreds can be accurately selected for crossbred performance without the need for pedigree or breed information. Results also showed that haplotype segments with strong linkage disequilibrium are shorter in crossbred and admixed populations than in purebreds, providing opportunities for QTL fine mapping.

**Key words:** Genomic Selection, Marker-Assisted Selection, Admixture, Crossbreeding

### 3.2 INTRODUCTION

Genomic selection (GS) (Meuwissen *et al.* 2001) is a form of marker-assisted selection that uses marker genotypes and phenotypes in a training population to simultaneously estimate effects of a large number of markers across the genome for the purpose of predicting
breeding values (BV) of selection candidates based on their marker genotypes. The accuracy of GS depends on the amount of linkage disequilibrium (LD) between QTL and markers and the number of records available to estimate marker effects. Most commercial beef cattle populations consist of animals with different and often unknown breed compositions. Presence of unknown population structure have raised concerns about using admixed or crossbred populations as training data for GS, yet these are the populations that are most relevant as the target for genetic improvement of purebreds (DEKKERS 2007). Admixture is the presence of multiple genetically distinct subgroups within a population (WANG et al. 2005). Numerous studies (GODDARD and MEUWISSEN 2005; HIRSCHHORN and DALY 2005; PFAFF et al. 2001; RABINOWITZ 1997) have reported that admixture can produce spurious associations and seriously elevate false discovery rates in QTL detection. Several methods have been proposed to address this problem (e.g., KENNEDY et al. 1992; MEUWISSEN et al. 2002; PRICE et al. 2006; PRITCHARD et al. 2000; SPIELMAN et al. 1993; YU et al. 2006).

Ideally, however, if all QTL that explain genetic variation in the trait of interest were included in the model, it would not be necessary to explicitly account for pedigree or breed composition in the analysis. Thus, provided high-density SNPs are used and analyzed simultaneously, as in GS, pedigree and breed composition need not be explicitly modeled. The objective of this work, therefore, was to evaluate accuracy of GS with high-density markers for predicting BV of purebred animals based on estimates of marker effects in a crossbred or admixed population, without explicitly accounting for pedigree or breed composition.

3.3 MATERIALS AND METHODS

No live animals were used for this study.
3.3.1 Population

A base population of unrelated individuals was stochastically simulated and used as the ancestral population of four pure breeds that were used to create admixed and crossbred populations (Figure 3.1). The base population was randomly mated for 1000 generations, including selfing, with an effective size ($N_e$) of 500. To simulate the four purebred populations (referred to as breeds A, B, C and D hereafter), at generation 0 four independent random samples of 100 animals were drawn from the base population, and each was randomly mated (including selfing) for another 53 generations, with $N_e$ of 100. Breeds A and B were then crossed to produce an F$_1$ (AB) population. Mating (AB) with breeds A and B created backcrosses (AB)A and (AB)B, which were again backcrossed to either A or B to create crossbred populations ((AB)A)A, ((AB)A)B, ((AB)B)A, and ((AB)B)B. An F$_2$ population, (AB)$^2$, was created by inter-mating the (AB) F$_1$s. Similarly, breeds C and D were used to create corresponding crossbred populations.

In beef cattle, commercial animals are often produced by mating purebred sires to crossbred dams of heterogeneous breed composition. In order to simulate such a population, an admixed population of two breeds ((Adm_AB)B) was created by first putting dams from breed A and all crossbreds involving breeds A and B in the same group and then mating them to sires from breed B. An admixed population of four breeds ((Adm_ABCD)B) was formed by mating dams from all breeds, excluding breed B, and all of the crosses involving the four breeds to sires from breed B. Further, three-way and four-way crosses were made by crossing (AB) with C and (AB) with (CD). To create the combined population, two random samples of equal size from purebreds A and B were put together into a single population. In the remainder, this latter population will be referred to as the combined_AB population, in
contrast to the admixed populations described in the previous, which include purebreds and their crosses.

Each of the purebreds A and B, the (AB), \((AB)^2\), (AC), (AB)C, (AB)(CD) crossbreds, combined_AB and admixed datasets was used as a training population consisting of 1000 animals for estimating marker effects. These animals were created by mating randomly sampled individuals from the appropriate parental lines, each of size 100. Thus, on the average each sire (and dam) had 10 offspring in the training data set. Because the objective was to determine how well marker effects estimated in the various training populations predicted BVs of purebred individuals, a separate generation of purebred population B was used for validation. To generate the validation population of size 1000, in generation 50 a sample of 1000 animals was drawn from breed B and randomly mated for another 4 generations (B_{Val} in Figure 3.1).

For evaluating the impact of breed differences on the accuracy of GS, a second scenario was also considered, in which breeds were separated in generation 25 rather than generation 0 (Figure 3.1), such that breeds were diverged for only 25 instead of 50 generations. In order to maintain the same level of LD, effective population size was reduced from 500 to 100 in generation 0, as before.

### 3.3.2 Genome

To make the simulation computationally feasible, a genome consisting of one chromosome of 100 cM with 100 segregating QTL and different marker densities was simulated (Table 3.1). To end up with the required number of segregating loci after 1000 generations, about three times as many biallelic loci were simulated with starting allele frequencies of 0.5 and a reversible random mutation rate of \(2.5 \times 10^{-5}\). Each locus was either a
marker locus or a QTL. A binomial map function was used to simulate recombination and interference was allowed by setting the maximum number of uniformly and independently distributed crossovers on the chromosome to be 4 (Karlin 1984). To make a marker panel, 500, 1000, 2000, or 4000 marker loci were drawn at random from segregating loci, MAF$^1 \geq 0.10$, at generation 53 after pooling all four breeds into a single cohort. As a result, markers were not evenly dispersed along the chromosome and some of them may not be segregating in all training populations.

### 3.3.3 Phenotypes

To create phenotypic values for each training population, 100 QTL were randomly picked from the set of segregating QTL in that population. This was done in generation 54 for breed B, (AB), and combined_AB, in generation 56 for (AB)$^2$, three- and four-way cross datasets and in generation 57 for the admixed training data sets (Figure 3.1). Note that QTL with MAF$^1 < 0.10$ in each breed will have an intermediate allele frequency in the combined_AB, crossbreds or admixed training populations. The QTL were additive and their effects were sampled from a gamma distribution with shape and scale parameters of 0.4 and 1/1.66, respectively. This provides us with an L-shaped distribution of QTL effects which (Meuwissen et al. 2001) suggest is close to the real distribution of the QTL effects. With equal probability, one of the two alleles was chosen to be positive or negative. To keep the genetic variance constant across training populations, the effect of each QTL was scaled in each replicate. This was done to ensure that each training dataset had the same genetic variance, such that this could not contribute to differences between training datasets.

The scaled QTL effects then were summed over all QTL genotypes for each individual to

---

$^1$ Minor allele frequency
compute its true BV. With this setting, each training population received a different set of QTL affecting phenotypes, although the number of segregating QTL and the genetic variance were the same for all training populations. Finally, a standard normal deviate was added to each true BV to provide the phenotype of an individual for a quantitative trait with heritability 0.30.

It should be mentioned that here the whole genetic variance is assigned to a single chromosome, whereas in reality the total genetic variance is distributed to all chromosomes (30 chromosomes in case of cattle, for instance). In this study, a short genome was chosen to reduce computational costs. An additional data set with a total of 5000 markers and 100 QTL on five chromosomes, each of length 1 M, was simulated to examine the impact of genome size on our results. The analysis was run for the training populations of purebred B, (((Adm_ABCD)B), combined_AB and four-way crossbred only, with 96 replicates.

### 3.3.4 Estimation of marker effects

Method Bayes-B of (MEUWISSEN et al. 2001) was used to estimate effects of markers in the training data using the model:

\[
y = \mathbf{1}\mu + \sum_i \mathbf{x}_i g_i + e
\]

where \(y\) is the vector of phenotypic values of individuals in the training data, \(\mu\) is a single unknown population mean, \(\mathbf{1}\) is a vector of ones, \(\mathbf{x}_i\) is a column vector containing the genotypes (0,1 or 2) of each individual at locus \(i\), \(g_i\) is the random unknown allele effect for marker \(i\), with \(g_i \sim N(0, \sigma^2_{g_i})\) and \(e\) is a random vector of unknown residuals with \(e_i \sim N(0, \sigma^2_e)\). In method Bayes-B the informative prior for locus specific genetic variance \((\sigma^2_{g_i})\) has a mixture distribution (SOLBERG et al. 2008). Following (MEUWISSEN et al. 2001),
it was postulated that with probability $\pi$, $\sigma^2_{g_i} = 0$, and with probability $(1 - \pi)$, it follows an inverted chi-square distribution with known parameters $\nu = 4.234$ and $S = 0.0429$. In practice, when $\sigma^2_{g_i} = 0$, $g_i$ is also set to zero.

The probability $\pi$ is assumed known. Based on preliminary analyses with several different $\pi$ values, $\pi$ was set equal to 0.95, except for a density of 40 markers per cM, for which $\pi$ was increased to 0.975. An MCMC chain of length 10,000 cycles with a burn in period of 1000 cycles was conducted. Convergence of the MCMC chain was examined using the R package CODA (Plummer et al. 2006).

### 3.3.5 Validation of genomic prediction

Once estimates of marker effects were obtained from the training dataset (posterior means from the MCMC chain), the estimated BV of individual $k$ ($GEBV_k$) in the validation dataset (generation 54 of population $B_{val}$) was computed as

$$GEBV_k = \sum_{i=1}^{m} x_{ik} g_i$$  \[53\]

where $x_{ik}$ and $g_i$ are the genotype and the estimated effect of genotype at locus $i$, respectively, and $m$ is the total number of markers. Accuracy was calculated as the correlation between the estimated and true BV of individuals in the validation data. This accuracy was used to compare performance of the different scenarios and training populations. All scenarios were replicated 160 times and results were averaged across replicates. Mean accuracies from alternate training populations were compared by the LSD test using the JMP software package (JMP, Version 7.0.1, SAS Institute Inc., Cary, NC, 1989-2007).
3.3.6 Linkage disequilibrium and between breed diversity

To evaluate the extent and magnitude of linkage disequilibrium in the training populations and its impact on accuracy, LD between pairs of SNP markers were estimated using $r^2$ (HILL and ROBERTSON 1968). Only markers with a MAF $\geq 0.1$ were considered in this analysis. The power to detect LD between two loci is minimum when at least one of them has an extreme allele frequency (GODDARD et al. 2000). Further, to evaluate the persistence of LD phase across training and validation populations, the correlations of $r$ between the two populations were calculated for different distances between loci (GODDARD et al. 2006).

To assess and compare the decline of LD with distance in different training populations, a non-linear regression model was fitted to the observed $r^2$ between marker pairs in each training population. The model used was based on the (SVED 1971) equation:

$$r_{ij}^2 = \frac{1}{1 + 4bd_{ij}} + e_{ij} \quad [54]$$

where $r_{ij}^2$ is the observed LD between markers i and j in the training data, b is a coefficient that describes the decline of LD with distance in the training data, $d_{ij}$ is distance in Morgans between markers i and j, and $e_{ij}$ is the random residual that was assumed normally distributed.

The level of genetic diversity present in the simulated breeds was investigated using Wright’s F-statistics (WRIGHT 1965b) $F_{IT}$, $F_{ST}$ and $F_{IS}$, as implemented in the program Fstat (GOUDET 2001). Genotypes at 200 loci from a random sample of 100 individuals from each of the four simulated breeds from generation 53 were used to estimate F-statistics. Significance levels for the F-statistics and related variance components were obtained from
20,000 permutations and from jackknife over loci (with different loci as re-sampling units), as provided by the Fstat program.

3.4 RESULTS

3.4.1 Accuracy of genomic selection

Correlations between estimated and true BVs for individuals in the validation dataset (population B_{val}, Figure 3.1) for different training populations and marker densities are shown in Table 3.2. Training in the same breed as the validation population (B_{54}) resulted in the highest accuracy in all cases. Accuracies tended to be lower if populations other than the validation breed were used for training but reductions in accuracy were not significant in some cases and depended on the breed composition of the training population. Based on differences in accuracies, the training populations can be divided into four groups: (1) purebred B and admixed populations, (2) two-breed combined_AB and crossed populations, (3) three- and four-way crosses, and (4) purebred A and AC. Training in the admixed populations resulted in similar accuracies as training in the purebred B population (group 1). The largest drop in accuracy compared to group 1 was for training populations in group 4, which included no contribution from the validation breed B. Averaged over all marker densities, relative to the accuracy of training and validating in the same breed, accuracy dropped by 46% when validation was in a different breed, whereas training in crossbred AC and validating in breed B resulted in a drop in accuracy of 35%. Populations in groups 2 and 3 had accuracies intermediate to those of groups 1 and 4. Comparing accuracies for populations in groups 2 and 3, as the number of breeds contributing to the training population increased, the accuracy dropped more. While training in group 2 populations resulted in a 6% decrease of accuracy, the decrease in accuracy when using the three- and four-way crosses
for training was on average about 10%. Interestingly, within group 2, differences in accuracy were not practically significant. Comparing groups 1, 2 and 3, the three- and four-way crossbred training populations showed the lowest accuracy of prediction.

3.4.2 Marker density

Accuracy generally increased with marker density (Table 3.2). The increase was most noticeable when training in A, AC and in the three-way and four-way crosses. Thus, the effect of marker density was more pronounced when the training population had a lower contribution of the breed that comprised the validation population (breed B in this case). Increasing marker density from 5 to 20 markers per cM, improved accuracy by 35, 20, 10 and 10 percent for groups 4, 3, 2 and 1, respectively. However, increasing marker density from 20 to 40 per cM did not improve accuracy as much; except for group 4 training datasets that showed an additional increase of 11% in the accuracy. Table 3.2 shows that with the level of LD and the amount of breed divergence simulated in this study, marker densities as low as 5 markers per cM were sufficient for accurate prediction of BV without explicitly accounting for pedigree or breed composition in the crossbred and admixed training populations, as long as the target breed contributed to the training population.

3.4.3 The effect of time since divergence of breeds

Table 3.3 shows the effect of the number of generations of random mating after isolation (referred to as Time Since Divergence, TSD) on the accuracy of GEBV. As expected, TSD did not significantly affect accuracy when the purebred B population was used as training data, but accuracies significantly increased for all other training datasets when TSD was reduced from 50 to 25. The maximum increase in accuracy was observed when using group 4 populations (A or (AC)) for training to predict breed B, for which accuracies increased by
62% and 42%, respectively. Considering both Tables 3.2 and 3.3, for both values of TSD, training in the admixed populations (group 1) resulted in a higher accuracy than training in the crossbred populations (groups 2 and 3).

The above results were based on one chromosome of 1 M to make the simulation computationally feasible. To determine if the simulated genome size affects the main conclusions, an additional data set with a total of 5000 markers and 100 QTL on five chromosomes, each of length 1 M, was simulated. The analysis was run for the training populations of purebred B, ((Adm_ABCD)B), combined_AB and four-way crossbred only. The corresponding accuracies for this scenario were 0.80, 0.77, 0.72 and 0.71, respectively. These values are comparable to the accuracies presented in the Table 3.2 for the scenario of 10 markers per cM.

### 3.4.4 Extent of linkage disequilibrium and differentiation between breeds

To explain the differences in accuracy between the groups described earlier and shown in Table 3.2, LD in the different training populations was examined by comparing the average distances between flanking markers at different levels of $r^2$ (Table 3.4). There were significant differences in the extent of LD between the training populations. Figure 3.2 depicts how LD decayed with distance in different training populations and shows significant differences in the rate of breakdown of LD between the populations. Note that for all training datasets average LD was between the expected LD based on Sved’s (1971) formula for effective population sizes of 100 and 500. The slowest and the steepest rates of decline of LD were in the purebred and the four-way cross training populations, respectively. The combined_AB and the two-way crosses had a slower rate of decay of LD than the three-way crosses. Non-linear regression was used to estimate a coefficient that describes the rate of
decay of LD with distance in each training population, based on the Sved (1971) formula. Resulting estimates of rate of decay of LD are shown in Table 3.4. In Sved (1971), the rate of decay constant was an estimate of the effective population size but this assumes a closed random mating population with constant historical effective population size. The interpretation is not valid for the populations analyzed here, because effective population size was not constant for the purebred population and the other populations also represented crosses and admixtures. Nevertheless, the estimated constants give a good indication that the effective number of founders, and therefore, the extent and decline of LD with distance differed substantially between populations.

Wright’s F-statistics were used to quantify the amount of divergence between the simulated breeds in generation 53. The estimated \( F_{IT} \), \( F_{ST} \) and \( F_{IS} \) were 0.240 (SE = 0.011), 0.236 (SE = 0.011) and 0.005 (SE = 0.004), respectively, when breed separation was in generation 0.

### 3.5 DISCUSSION

Training datasets consisting of a purebred, a two-breed combined, several crossbreds and admixed populations were compared for their ability to accurately predict true BVs of selection candidates in a purebred population using genomic selection. The main focus will be on crossbred, combined and admixed populations.

#### 3.5.1 Accuracy of genomic selection

In Table 3.2, all types of training populations performed remarkably well, except when another pure breed was used for training, or when the training data consisted of a cross that did not include the target breed (AC). The admixed training populations resulted in nearly the
same accuracies as when training was in the breed to which selection candidates belonged (breed B).

3.5.2 Extent of linkage disequilibrium

Based on results presented in Table 3.4, when considering the average distance between pairs of adjacent markers that had $r^2 \geq 0.1$, there was more extensive LD in the crossbred, combined_AB and admixed populations than in the purebred population. The extent of LD is proportional to the age of the LD generating event (Reich et al. 2001). In a subdivided or a crossbred population, like the combined_AB and the $F_1$ training populations, LD is composed of two parts (Lo et al. 1993; Nei and Li 1973). The first part is the average LD that existed within the parental populations, referred to as ‘old’ LD, and the second part is LD generated in the cross as a result of difference in gene frequencies between the parental breeds, referred to as ‘new’ LD, because it is created by a recent phenomenon. While the ‘old’ LD is confined to shorter distances due to the accumulation of recombination events, the ‘new’ LD extends over longer intervals. The combined_AB training population was composed of breeds A and B with equal proportion; as a result, the distribution of LD in this population was the same as in the $F_1$ training population. Figure 3.3, depicts the average distance between pairs of markers at various levels of LD.

Table 3.4, on the other hand, shows that haplotype with strong LD ($r^2 \geq 0.70$) are significantly shorter in the admixed and the crossbred populations compared to the purebred population. The average distance between pairs of markers with strong LD is three times larger in the purebred than in the admixed and four-way crossbred populations (Table 3.4 and Figure 3.3c). This very narrow region of strong LD in the crossbred training populations might explain the high accuracy obtained with these populations (Table 3.2). In the same way
that LD limited to short distances is beneficial in QTL fine mapping by providing a more accurate estimate of the QTL position (AERTS et al. 2007), it can also result in higher accuracy of genomic selection because only markers that are very close to the QTL will explain a high proportion of the QTL variance and this association will not rapidly erode over generations by recombination.

In a human genetics study, (SHIFMAN and DARVASI 2001) compared the average level of LD between SNP markers for distances below and over 200 kb in an outbred population (CEPH) with that in several isolated populations (Finnish, Ashkenazi, and Sardinian). Their findings showed that at short intervals the amount of LD in the outbred population was comparable to that in the isolated populations, whereas at long intervals (>200 kb) there was up to six times more LD in the isolated populations than that in the outbred population. In another study, (SHIFMAN et al. 2003) compared an admixed, an outbred and an isolated population (African Americans, Caucasians and Ashkenazi Jews, respectively). They found that the average LD declined with distance between loci more rapidly in the admixed population. This is in accordance with our results, which also showed that the average level of LD was higher and extended over longer intervals in the purebred population compared to crossbred and admixed populations (see Figures 3.2, 3.3c and 3.4). In a crossbred population, individuals are more “distantly” related to each other, i.e., the mean time to a common ancestor is longer, thus LD haplotypes in the population are narrower than that in a purebred population. Results of the canine genome project have also shown that haplotype blocks are several Mb long within a breed but they are much shorter across breeds, extending only to tens of kb (LINDBLAD-TOH et al. 2005). The ancient domestic dog diverged from wolves 15,000-100,000 years ago, while most of the new breeds of dog were formed within the past
few hundred years (Lindblad-Toh et al. 2005).

Figure 3.2, illustrates that LD in all training populations fell between the expected LD based on the Sved (1971) formula for effective population sizes of 100 and 500. As can be seen from the figure, LD at short distances followed the expectation based on $N_e = 500$, whereas at larger distances it tended towards its expectation based on $N_e = 100$. LD at short intervals is a function of $N_e$ in the distant past, whereas LD at longer intervals reflects $N_e$ in the recent past (Hayes et al. 2003).

3.5.3 Marker density

As it is evident from Table 3.4, the frequency with which strong LD haplotypes occur differs substantially between the training populations. Consider the difference of accuracies (Table 3.2) when the training population is purebred B or a four-way crossbred population, for example. While the high LD signals are restricted to very short distances in the four-way crossbred population, there are about three times as many markers with strong LD in the purebred B population (Table 3.4). This might describe why the three- and four-way crossbred populations were much more affected by marker density compared to the purebred training population. Increasing marker density is expected to raise the level of LD between markers and QTL because the average distance between adjacent loci is inversely related to marker density and recombination is less likely to erode associations between tightly linked loci. Figure 3.4 illustrates the relationship between marker density and the level of LD between markers. The higher the LD between a pair of loci (in fact a marker and QTL), the larger is the variance that is associated with the marker (Luo et al. 1997).

The effect of marker density on the accuracy of GS has been discussed in some recent studies (Calus et al. 2008; Muir 2007; Solberg et al. 2008). Solberg et al. (2008)
simulated SNP markers with several densities. In their study, where both training and validation populations were purebreds, accuracies of GS using SNP markers with densities of 1, 2, 4 and 8 markers per cM, were found 0.69, 0.79, 0.84 and 0.86, respectively (with 1000 QTL on the genome). For the density of 4 SNP per cM, their accuracy of 0.84 is roughly comparable to our estimate of 0.79 (Table 3.2), keeping in mind that we had 5 SNP per cM. The higher heritability of the trait (0.5 vs. 0.30) considered in Solberg et al. (2008) study might be a reason for the difference in accuracies obtained in the two studies. With a lower heritability, markers explain a lower proportion of additive genetic variance and a lower accuracy will be obtained (GODDARD and HAYES 2007). Thus, to get accurate estimates of marker effects a larger sample size is required. The level of LD between markers and QTL and the sample size used to estimate the QTL effects are the two factors driving the accuracy of marker assisted selection and the power of QTL detection (HAYES et al. 2007; LANDE and THOMPSON 1990).

3.5.4 Persistence of linkage disequilibrium phase

Markers in LD with putative QTL are valuable for marker-assisted selection if the marker-QTL linkage-phase and extent of LD is consistent between the population used for estimation and the population where selection is to be practiced (DEKKERS and HOSPITAL 2002; GODDARD et al. 2006). Figure 3.5 shows the persistence of LD phase between adjacent markers in the training and validation populations, as measured by the correlation of r between the two populations. A higher correlation implies that the marker-marker (and most probably the marker-QTL) linkage phase is more consistent between the two populations. This figure shows that the correlation of r increased with marker density and was lower if the training and validation populations were more different, e.g. when training
and validation was in different breeds versus in the same breed. This relationship between persistence of LD phase and divergence between breeds agrees with other reports (Andreeescu et al. 2007; de Roos et al. 2008; Gautier et al. 2007). Obviously, the shorter the length of a haplotype, the higher is the chance of its similarity across populations. In the same way, as distance in time between two sub-populations increases, there is a higher chance for recombination to break down the LD that was present in the ancestral population and drift to create new LD within each sub-population (Goddard et al. 2006; Hill and Robertson 1968).

In a study of extent and persistence of LD phase in Holstein-Friesian, Jersey and Angus cattle, de Roos et al. (2008) reported a correlation close to 1 of r between two breeds for pairs of markers that were <10 kb apart and a decline of this correlation as distance between markers or divergence between breeds increased. Considering marker loci that were less than 10 kb apart, Gautier et al. (2007) reported correlations of 0.54 to 0.93 (0.77, on the average) of r for pairs of European cattle breeds, which again reflects how the degree of relationship between two breeds changes the correlation of r. In our simulations, marker densities were not greater than 40 per cM (a distance of ≈ 25 kb between adjacent markers). For pairs of markers with such average distance, the correlation of r for across breeds GS, was 0.81 (see Figure 3.5). In order for the LD correlation between two breeds to be high, tight LD between a pair of loci should exist in the ancestral population before divergence of the two breeds such that recombination cannot erode it (Goddard et al. 2006).

The high correlation of r for the admixed training populations compared to the crossbred training populations, reveals a closer relationship between these populations and breed B, as the admixed populations had a higher proportion of breed B genes. Correlation of r between
two populations can be used as a good estimate of relationship between the two populations (ANDREESCU et al. 2007). This could explain why the admixed populations resulted in higher accuracy of selection than the crossbred training populations. The greater accuracy obtained when training in the AC population versus in the A purebred population (Table 3.2) might be explained as follows. The use of AC cross forces the model to look only at ancestral LD that was already present at the time of separation of breeds, rather than using the ‘new’ LD. Ancestral LD is more likely to be present in breed B as well. This explains why the correlation of r between the AC and breed B was higher than the correlation of r between breeds A and B (data not shown). The correlation of r between two populations may be used as an indication of the required marker density to ensure marker-QTL linkage-phase persists across the populations (GODDARD et al. 2006).

In a simulation study, (IBÁNEZ-ESCRICHE et al. 2009) applied genomic selection to select purebreds for crossbred performance. In their work, they compared the performance of a model with breed-specific effects to a model with the same effects across breeds. It was shown that for two unrelated breeds, where correlation of r between the two breeds was zero, the across-breed model was as accurate as the breed-specific model in prediction of BVs. The breed-specific model resulted in a lower accuracy of prediction when the marker density increased compared to the across-breed model (IBÁNEZ-ESCRICHE et al. 2009). More effects might need to be estimated in a multi-breed population. When alternative alleles are fixed in different breeds, there are almost twice as many effects to be estimated compared to the purebred population. Thus, one may need a larger sample size in a multi-breed population to get an accuracy that is comparable to the accuracy in a purebred population.
3.5.5 Divergence of the breeds

Wright’s F-statistics are inbreeding coefficients that differ in the “reference” population that is used (HARTL and CLARK 1989). $F_{IT}$ is the broadest measure of inbreeding in that it takes into account both the effects of nonrandom mating within the sub-populations ($F_{IS}$) and the effects of population subdivision ($F_{ST}$) (HARTL and CLARK 1989). The estimate of $F_{IS}$ of 0.005 implies only a minor deficit of heterozygosity within breeds. Because individuals in each breed were randomly mated, a significant divergence from the Hardy-Weinberg proportions within each breed is not expected. The expected value of $F_{IS}$ is $1/(2N_e)$, which with $N_e = 100$ in each breed agrees with the results and indicates negligible levels of inbreeding within the breeds. The expected value of $F_{ST}$ under the conditions of an idealized population with subdivision (FALCONER and MACKAY 1996) is $[1 - \left(1 - \frac{1}{2N_e}\right)^t]$, which with $t = 53$ generations is equal to 0.233 and is in close agreement with our estimate based on marker data of 0.236. This value of $F_{ST}$ shows that about 24 percent of the total genetic variability in the whole population can be attributed to the difference among breeds (e.g., CANON et al. 2001), or that about 24% of shared allelic diversity was lost within each breed since they were separated. Thus, the breeds had significantly diverged from each other. With $F_{IS} = 0$, $F_{ST}$ and $F_{IT}$ are expected to be equal, because $(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT})$ (WRIGHT 1965a). Recently, (MCKAY et al. 2008) published estimated pairwise and global $F_{ST}$ values for several cattle breeds, based on a panel of 2641 SNPs. Their estimated global $F_{ST}$ when they considered both *Bos taurus* and *Bos indicus* breeds was 0.29. However, the estimated global $F_{ST}$ reduced to 0.17 when they excluded the *Bos indicus* breeds from their analysis.
(McKay et al. 2008). Therefore, our simulated breeds had enough divergence to represent current breeds of beef cattle.

### 3.5.6 The effect of time since divergence of breeds

The accuracies for the scenarios of TSD = 25 and 50 (Table 3.3) were compared. As expected, no effect of TSD on accuracy was observed for the scenario of training and validating in the same breed (B), because there was no divergence within the same breed. The minimum (4%) and the maximum (62%) increase in accuracies were observed for the admixed training populations and when training and validating in different breeds, respectively, when TSD changed from 50 to 25 generations. Again, this reflects the fact that the more distantly two populations are related, the higher is the chance of recombination to break down the shared ancestral haplotypes (and even reverse the LD phase) across the populations. This might explain why accuracy of GS was reduced for training populations other than breed B when TSD changed from 25 to 50 (Table 3.3). The more time elapsed since separation of two subpopulations, the higher is the loss of shared allelic diversity between them (McKeigue 2005).

### 3.5.7 Effect of selection

In this simulation only LD generated by mutation and drift was considered. In reality, livestock populations have been under selection for a long time and breeds may have been under varying intensities and directions of selection. To assess the impact of differential selection of breeds on the validity of our results, the relationship, across replicates, between the accuracy of GS and the difference in the mean true breeding value of the breeds that were crossed or admixed was evaluated. Although any breed differences in the simulations were the result of mutation and drift, rather than directed selection, selection can be viewed as
‘directed random drift’. Therefore, the accuracy of genomic selection against the variance of the mean true BV of the breeds that are crossed was plotted for the 160 replicates of our simulation (Figure 3.6). This did not reveal any significant association of accuracy with the extent of the diversity of breeds contributing to the cross. Considering figure 3.6a, for example, although breeds A and B showed quite a range of different true genetic means for the trait of interest, this difference did not affect the accuracy of selection.

Traditionally, GS studies by simulation have only considered additive QTL effects (Meuwissen et al. 2001). However, in reality QTL contribute to total genetic variation either by themselves or by interacting with other QTL (Carlberg et al. 2006). Interaction among loci might result in a biased estimate of the effect of each locus (Carlberg and Haley 2004). In a recent study, Carlborg et al. (2006) identified a genetic network of several interacting loci that significantly contributed to body weight at 56 days of age in chicken. Their results showed that the power of QTL mapping experiment in identifying loci whose effect is dependent on the genotype at another locus improves when the inter- and intra-locus interactions are included in their statistical model. Thus, dominance and epistatic QTL effects might need to be considered in GS studies where the objective is improvement of purebreds for their crossbred progeny performance.

In this study our focus was on purely additive gene effects, however, there might be the question of how accurate will be GS predictions in the presence of heterosis. In an F<sub>1</sub> population, where the population is homogenous in terms of breed composition we think ignoring the effects of heterosis does not bias the prediction of marker effects. However, in an admixed population, because individuals have different breed compositions the dominance effects must be explicitly accounted for.
Another question might be the choice of training population when the selection candidates are crossbred themselves. In a recent simulation study, (ODEGARD et al. 2009) investigated introgression of favorable alleles from an inferior donor line into a superior recipient line using dense marker genotyping and genomic selection. Their proposed method of combining backcrossing and GS increased the frequency of favorable QTL alleles at the expense of unfavorable ones (irrespective of origin) across the entire genome, without any specific effort to reduce the linkage drag from the donor line (ODEGARD et al. 2009).

In a recent study with real data, (HARRIS et al. 2008) compared the accuracy of genomic selection of purebred Jersey (J), purebred Holstein-Friesian (HF), and crossbred J-HF bulls using the BovineSNP50 BeadChip. The training data sets were either one of the two breeds (J or HF) or a combined data set of both breeds. Training in one breed and validating in another breed resulted in an accuracy of -0.10 to 0.3. Accuracy of genomic selection of crossbred J-HF bulls was 5-10% higher when training was done in the combined data set compared to when training was in J or HF breeds (HARRIS et al. 2008). Assuming that the validation population is a crossbred, we compared the accuracy of GS when the training population was either purebred or crossbred (F1). Training in the crossbred population increased accuracy of GS in crossbred population by 11% compared to training in the purebred (data not shown).

3.6 CONCLUSIONS

A population that is a crossbred or an admixture of different breeds can be used as a training data set for GS and can provide reasonably accurate estimates of true BVs of purebred selection candidates. This also implies that, with GS using high-density SNP markers, marker estimates obtained from crossbred populations can be used to select
purebreds for crossbred performance, as suggested by (DeKkers 2007), and examined by (Ibáñez-Escríche et al. 2009). Our results showed that in crossbred and admixed populations haplotypes with strong LD are much shorter than in purebred populations. Thus, crossbred or admixed populations are more suitable for QTL fine mapping than purebred populations, provided marker density is sufficient.

Further, as haplotype segments with strong LD in crossbred and admixed populations are narrower; markers in such segments are expected to have more consistent associations with QTL across the training and validation populations. Therefore, the decline of accuracy of GS over generations that has been observed in simulation studies (e.g., Habier et al. 2007) might be slower when admixed or crossbred populations are used for training than when purebred populations are used. By combining two pure breeds into a single training population, one can take advantage of a larger sample size for simultaneous estimation of marker effects and thus improve the accuracy of GS. In our simulation, when the size of the training population for the combined_AB training population was doubled, a 7% increase of the accuracy resulted (data not shown). In addition, by combining breeds into a single training population, versus making certain crosses like an F1, a lot of time and effort can be saved. More importantly, there is a higher chance of segregation of breed-specific QTL in a multi-breed training population.

In the present study, while dealing with admixed populations, the population structure or additive genetic relationships were not explicitly modeled, which might be regarded as the standard method to limit the false discoveries due to population admixture in marker-phenotype association studies. Nevertheless, GS using high-density markers proved to be efficient enough to distinguish between true signals of association from spurious signals, at
least under the idealized population structures that were used in the simulations. Whether this could provide an alternative methodology for association studies in populations with cryptic structures or extensive genealogical relationships requires further research.

3.7 ACKNOWLEDGEMENTS

This research was motivated by a question by R.L. Quaas at a meeting of the statistical methods group of the National Beef Cattle Evaluation Consortium. Funding for this research was provided by Newsham Choice Genetics, the United States Department of Agriculture, National Research Initiative grant USDA-NRI-2007-35205-17862, and the Iowa Agricultural and Home Economics Experiment Station, Ames, IA.

3.8 LITERATURE CITED


### 3.9 TABLES

**Table 3.1** – The parameters used for the simulation program

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size</td>
<td>100 cM</td>
</tr>
<tr>
<td>Number of chromosomes</td>
<td>1</td>
</tr>
<tr>
<td>Marker density per cM</td>
<td>5, 10, 20 or 40</td>
</tr>
<tr>
<td>Number of segregating QTL</td>
<td>100</td>
</tr>
<tr>
<td>Mutation rate of QTL or marker locus</td>
<td>$2.5 \times 10^{-5}$</td>
</tr>
<tr>
<td>Minor allele frequency</td>
<td>$\geq 0.10$</td>
</tr>
<tr>
<td>Distribution of additive QTL effects</td>
<td>Gamma (Shape = 0.4; Scale = 1 / 1.66)</td>
</tr>
<tr>
<td>Prior distribution for $\sigma^2_{gi}$, when $d_i = 1$</td>
<td>$\chi^2 (v = 4.234, S = 0.0429)$</td>
</tr>
<tr>
<td></td>
<td>0.950 for marker density &lt;=20 and 0.975 for marker density of 40 per cM</td>
</tr>
<tr>
<td>Population size</td>
<td>$N_e^1 = 500$</td>
</tr>
<tr>
<td>Generations -1000 to 0</td>
<td>$N_e = 100$</td>
</tr>
<tr>
<td>Generations 0 to 53</td>
<td>$N^2 = 1000$</td>
</tr>
<tr>
<td>Generation &gt;= 54</td>
<td>0.30</td>
</tr>
<tr>
<td>Heritability</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1. Effective population size  
2. Number of phenotypic and or genotypic records  
3. Variance of the marker genotype effects ($\sigma^2_{gi}$)  
4. $\pi$ is Prob. ($\sigma^2_{gi} = 0$) for Bayes-B method
Table 3.2- Average accuracy of estimated breeding values in the validation dataset (pure breed B) from genomic selection with different training datasets and marker densities (number of SNPs per cM)\(^1,2\)

<table>
<thead>
<tr>
<th>Marker density</th>
<th>Training dataset</th>
<th>5/cM</th>
<th>10/cM</th>
<th>20/cM</th>
<th>40/cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>0.79(^a)</td>
<td>0.83(^a)</td>
<td>0.84(^a)</td>
<td>0.85(^a)</td>
</tr>
<tr>
<td>1</td>
<td>(Adm_AB)B</td>
<td>0.77(^ {ab})</td>
<td>0.81(^ {ab})</td>
<td>0.84(^ {a})</td>
<td>0.84(^ {ab})</td>
</tr>
<tr>
<td>1</td>
<td>(Adm_ABCD)B</td>
<td>0.76(^ {bc})</td>
<td>0.80(^ {ab})</td>
<td>0.84(^ {a})</td>
<td>0.84(^ {ab})</td>
</tr>
<tr>
<td>2</td>
<td>(A + B)</td>
<td>0.71(^f)</td>
<td>0.76(^ {de})</td>
<td>0.80(^ {cd})</td>
<td>0.79(^d)</td>
</tr>
<tr>
<td>2</td>
<td>(AB)</td>
<td>0.74(^ {cd})</td>
<td>0.78(^ {cd})</td>
<td>0.82(^b)</td>
<td>0.82(^c)</td>
</tr>
<tr>
<td>2</td>
<td>(AB)(^2)</td>
<td>0.72(^ {de})</td>
<td>0.77(^d)</td>
<td>0.81(^ {bc})</td>
<td>0.83(^ {bc})</td>
</tr>
<tr>
<td>3</td>
<td>(AB)C</td>
<td>0.66(^f)</td>
<td>0.75(^ {cf})</td>
<td>0.77(^e)</td>
<td>0.79(^d)</td>
</tr>
<tr>
<td>3</td>
<td>(AB)(CD)</td>
<td>0.67(^f)</td>
<td>0.72(^f)</td>
<td>0.79(^ {de})</td>
<td>0.79(^d)</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>0.34(^h)</td>
<td>0.45(^h)</td>
<td>0.48(^g)</td>
<td>0.54(^f)</td>
</tr>
<tr>
<td>4</td>
<td>(AC)</td>
<td>0.43(^g)</td>
<td>0.50(^g)</td>
<td>0.58(^f)</td>
<td>0.64(^e)</td>
</tr>
</tbody>
</table>

\(^1\) Values with different letters within a column are significantly different (P<0.05). Based on 160 replicates
\(^2\) B is the purebred B; (Adm_AB)B and (Adm_ABCD)B are admixture of 2 and 4 breeds; (A+B) is the combined AB; (AB) is the F\(_1\); (AB)\(^2\) is the F\(_2\); (AB)C is the three-way cross; (AB)(CD) is the four-way crossbred; A is purebred A; and (AC) is cross of breeds A and C
Table 3.3 - The impact of time since divergence of breeds on the accuracy of genomic selection when training in different datasets with 5 markers per cM on a 1 M genome.

<table>
<thead>
<tr>
<th>Training data set</th>
<th>Time since divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25</td>
</tr>
<tr>
<td>B</td>
<td>0.80&lt;sup&gt;ab&lt;/sup&gt;</td>
</tr>
<tr>
<td>(Adm_AB)B</td>
<td>0.80&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>(Adm_ABCD)B</td>
<td>0.79&lt;sup&gt;abc&lt;/sup&gt;</td>
</tr>
<tr>
<td>(A + B)</td>
<td>0.76&lt;sup&gt;cde&lt;/sup&gt;</td>
</tr>
<tr>
<td>(AB)</td>
<td>0.77&lt;sup&gt;bcd&lt;/sup&gt;</td>
</tr>
<tr>
<td>(AB)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.78&lt;sup&gt;abcd&lt;/sup&gt;</td>
</tr>
<tr>
<td>(AB)C</td>
<td>0.73&lt;sup&gt;efg&lt;/sup&gt;</td>
</tr>
<tr>
<td>(AB)(CD)</td>
<td>0.75&lt;sup&gt;def&lt;/sup&gt;</td>
</tr>
<tr>
<td>A</td>
<td>0.55&lt;sup&gt;jk&lt;/sup&gt;</td>
</tr>
<tr>
<td>(AC)</td>
<td>0.61&lt;sup&gt;l&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>1</sup> Values with different letters within and across columns are significantly different (P<0.05).

<sup>2</sup> B is the purebred B; (Adm_AB)B and (Adm_ABCD)B are admixture of 2 and 4 breeds; (A+B) is the combined_AB; (AB) is the F<sub>1</sub>; (AB)<sup>2</sup> is the F<sub>2</sub>; (AB)C is the three-way cross; (AB)(CD) is the four-way crossbred; A is purebred A; and (AC) is cross of breeds A and C.

Based on 160 replicates.
Table 3.4- Average distance (in cM) between adjacent markers with $r^2$ greater than 0.1, 0.4 or 0.7 in different training datasets, with the percentage of such marker pairs out of all adjacent pairs with $r^2$ greater than 0 in brackets. Estimated coefficients of LD decline (Beta estimate) are shown in the third row of the table. 1, 2, 3, 4

<table>
<thead>
<tr>
<th>Training population</th>
<th>B</th>
<th>(AB)</th>
<th>(A+B)</th>
<th>(AB)$^2$</th>
<th>(AB)C</th>
<th>Adm_4b</th>
<th>(AB)(CD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta estimate</td>
<td>151</td>
<td>262</td>
<td>263</td>
<td>269</td>
<td>349</td>
<td>355</td>
<td>440</td>
</tr>
<tr>
<td>Minimum $r^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.31$^a$</td>
<td>0.21$^b$</td>
<td>0.21$^b$</td>
<td>0.16$^c$</td>
<td>0.12$^d$</td>
<td>0.10$^e$</td>
<td>0.09$^f$</td>
</tr>
<tr>
<td>(0.42)</td>
<td>(0.21)</td>
<td>(0.21)</td>
<td>(0.22)</td>
<td>(0.17)</td>
<td>(0.17)</td>
<td>(0.14)</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.75$^c$</td>
<td>2.21$^a$</td>
<td>2.17$^b$</td>
<td>0.50$^d$</td>
<td>0.56$^e$</td>
<td>0.32$^f$</td>
<td>0.21$^g$</td>
</tr>
<tr>
<td>(1.0)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.55)</td>
<td>(0.41)</td>
<td>(0.41)</td>
<td>(0.33)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>3.22$^g$</td>
<td>15.12$^a$</td>
<td>15.05$^b$</td>
<td>4.34$^e$</td>
<td>10.5$^c$</td>
<td>7.39$^d$</td>
<td>4.41$^f$</td>
</tr>
<tr>
<td>(5.7)</td>
<td>(7.2)</td>
<td>(7.2)</td>
<td>(4.4)</td>
<td>(4.1)</td>
<td>(3.4)</td>
<td>(2.3)</td>
<td></td>
</tr>
</tbody>
</table>

1 Values with different letters within each row are significantly different (P<0.05)
2 B is the purebred B training population; (AB) is the F1; (A+B) is the combined_AB, (AB)C is the three-way cross, Adm4br is the admixture of four breeds and (AB)(CD) is the four-way crossbred training population
3 Based on 60 replicates; in each replicate distances were averaged across adjacent pairs that met the minimum $r^2$ value, resulting in at least 100,000 pairs per replicate
4 All Beta estimates had SE of less than 1
3.10 FIGURES

Figure 3.1 – Schematic representation of the simulated population history ($N_e =$ effective population size) and the different types of crossbred and admixed populations that were simulated\(^1,2\).

Generation -1000 Random mating ($N_e = 500$)

Generation 0

- Line A\(_0\)
- Line B\(_0\)
- Line C\(_0\)
- Line D\(_0\)

Line separation ($N_e = 100$)

Generation 0

- Line A\(_0\)
- Line B\(_0\)
- Line C\(_0\)
- Line D\(_0\)

Generation 50

- Breed A
- Breed B
- Breed C
- Breed D

Generation 53

- A\(_{53}\)
- B\(_{53}\)
- C\(_{53}\)
- D\(_{53}\)

Generation 54

- A\(_{54}\), AB, (A+B), B\(_{54}\), B\(_{Val}\), C\(_{54}\), CD , D\(_{54}\)

Generation 55

- (AB)\(^2\), (AB)A, (AB)B, (AB)C , (AB)(CD) , (CD)\(^2\)

Generation 56

- ((AB)A)A , ((AB)A)B , ((AB)B)A , ((AB)B)B

Generation 57

- ((Adm_AB)B) , ((Adm_ABCD)B)

\(^1\) A\(_{54}\) and B\(_{54}\) represent pure breed training populations; AB is an F\(_1\) training population; A+B is a training population consisting of individuals from breeds A and B; (AB)\(^2\) is an F\(_2\) training population; (AB)A, (AB)B, ((AB)A)A, ((AB)A)B, ((AB)B)A and ((AB)B)B are different back-cross populations; (Adm_AB)B and (Adm_ABCD)B are admixed training populations of two and four breeds; B\(_{Val}\) is the validation population;

\(^2\) As of generation 54, arrows are not shown for simplicity of the picture.
Figure 3.2 - Average linkage disequilibrium as measured by $r^2$ against distance (cM) in different training populations.\(^1,2,3,4\)

\(r^2 \approx \frac{1}{1 + 4N_e c}\), where \(N_e = 100\) or \(500\) and \(c\) is recombination rate calculated as \(0.5(1-\exp(-2*\text{map distance}))\); Pure breed is the purebred B training population; (AB) is the F\(_1\); Adm4br is the admixture of four breeds and (AB)(CD) is the four-way crossbred training population.

The graph is based on 60 replicates; average \(r^2\) over all replicates are plotted against distance.

Lines in the graph are in the order as shown in the legend. The top line is Sved Ne100, second line is purebred, third line is (AB), fourth line is Adm4br, fifth line is (AB)(CD) and the last line is Sved Ne500.

Other training populations are not shown for clarity of the picture.
Figure 3.3 - Average distance (cM) between adjacent markers in different training populations at various levels of linkage disequilibrium ($r^2$) $^{1,2,3}$

a) $B$ is the purebred $B$, $F_1$ is the (AB), $F_2$ is the (AB)$^2$, $ABC$ is the three-way crossbred, $ADMX$ is the admixture of 4 breeds and $ABCD$ is the four-way cross training population.

b) Based on 60 replicates, results were averaged over distances with certain amount of LD and over replicates. Note to the different scales of the graphs.

c) a) Minimum $r^2$ between 0.10 and 0.30, b) Minimum $r^2$ between 0.40 and 0.60, c) Minimum $r^2$ between 0.70 and 0.90
**Figure 3.4** - The average level of linkage disequilibrium as a function of marker density (# of markers per cM)\(^1,2\) and type of training population

\(^1\) B is the purebred B, F1 is the (AB), F2 is the (AB)\(^2\), ABC is the three-way crossbred, ADMX is the admixture of 4 breeds, and ABCD is the four-way crossbred training population. Other training populations are not shown, because they showed the same trend and magnitude of LD.

\(^2\) The graph is based on 60 replicates, for each training population average \(r^2\) of all replicates for marker distances of 0.025, 0.05, 0.10 and 0.20 cM was calculated.
Figure 3.5 - Correlation of $r$ between each pair of training and validation populations$^2$, as a function of marker density (in 1 cM)$^3$

$^1$ $r$ is the correlation coefficient

$^2$ B is training and validating in breed B; A is training in breed A and validating in breed B; A+B is training in combined AB and validating in B; AB is training in AB and validating in breed B; F$_2$ is training in $\text{(AB)}^2$ and validating in B; (AB)C is training in the three-way cross and validating in B; (AB)(CD) is training in a four-way cross and validating in B; and Adm4Br is training in the admixture of 4 breeds and validating in breed B.

$^3$ Based on 60 replicates
**Figure 3.6 - Plot of accuracy against between breeds variance of true breeding values**

a) Plot of accuracy when training in an F1 versus between breeds [A and B] variance of true breeding values.
b) Plot of accuracy when training in a three-way cross [(AB)C] versus between breeds (A, B and C) variance of true breeding values.
c) Plot of accuracy when training in a four-way cross [(AB)(CD)] versus between breeds (A, B, C and D) variance of true breeding values.

In all plots, the black line shows the regression of accuracy on between breeds variance of true breeding values.
CHAPTER 4. GENOME-WIDE MAPPING OF QUANTITATIVE TRAITS

LOCi IN ADMIXED POPULATIONS

In preparation to be submitted to *Genetics*

Ali Toosi³, Rohan L. Fernando¹,⁴ and Jack C.M. Dekkers¹

4.1 ABSTRACT

Population stratification and cryptic relationship have been the main sources of excessive false-positives and false-negatives in population-based association studies. A plethora of methods have been developed to model these confounding factors and minimize their impact on the results of the genome-wide association studies. In majority of these studies, a two-stage approach has been applied where: (1) methods are used to determine if there is a population structure in the sample dataset and (2) correcting for the effects of population structure either by properly modeling it or by running a separate analysis within each sub-population.

We conducted genome-wide QTL mapping study in a stochastically simulated admixed population. Genome was composed of six chromosomes each with 1000 equally spaced markers. Fifteen segregating QTL contributed to the genetic variation of a quantitative trait with heritability of 0.30. To model kinship and breed composition three approaches were evaluated: (1) Single marker simple regression, (2) Single marker mixed linear model and (3)

³ Dept. of Animal Science, Iowa State University, Ames IA USA
⁴ Corresponding author, rohan@iastate.edu
Bayesian multiple-regression. Each of the methods was fitted with and without breed composition. Accuracy, power, false-positive rate and positive predictive value of each method was calculated and used for comparison. Models (1) and (3), both without breed composition, were ranked as the worst and the best performing approaches, respectively. Our results showed that while implicit modeling of kinship and breed composition is a must for models (1) and (2), model (3) can disregard them without a significant effect on its power and false-positive rate. Modeling of breed composition as a fixed effect resulted in some loss of power of QTL detection when model (3) was employed. In conclusion, this study showed that Bayesian multiple-regression method is robust to population structure and kinship among study subjects and its performance is comparable to that of a linear mixed model approach.

**Keywords:** Population structure, kinship, Bayesian multiple regression, Association study

**4.2 INTRODUCTION**

The ideal condition in a genome-wide association study (GWAS) is that the correlation, i.e., linkage disequilibrium (LD), between a polymorphism and the trait of interest to be high if and only if the polymorphism itself is a causative mutation (Quantitative Trait Loci, QTL) or it is closely linked to a QTL (Platt et al. 2010). This requires a panmictic population to be used for the study. Unfortunately, except in population genetics theory, this type of populations probably does not exist (Tiwari et al. 2008; Zhang et al. 2006). Widespread prevalence of non-random mating (e.g., assortative mating) in livestock and crop populations has resulted in complex patterns of population stratification (PS) and kinship in them (Crossa et al. 2007; Flint-Garcia et al. 2003; Goddard and Hayes 2009; Nordborg et
al. 2005; Yu et al. 2006). Without properly accounting for these factors, GWAS in such populations could lead to spurious false-positives (FP) (markers declared as significant but not closely linked to a QTL) and false-negatives (FN) (markers closely linked to a QTL but not being declared as significant) due to extensive LD between systemic and non-systemic loci (Aranzana et al. 2005; Atwell et al. 2010; Brachi et al. 2010; Dadd et al. 2009; Goddard and Hayes 2009; Gu et al. 2009; Hoggart et al. 2003; Iwata et al. 2007; Kang et al. 2008; Lander and Schork 1994; Miclaus et al. 2009; Palmer and Cardon 2005; Pritchard et al. 2000; Pryce et al. 2010; Rosenberg et al. 2010; Setakis et al. 2006; Yu et al. 2006; Zhao et al. 2007). Compared to simple monogenic traits, complex polygenic phenotypes are more vulnerable to the elevated FP rates in GWAS, where the magnitude of signals from multiple QTL may be comparable to those resulting from PS (Teo et al. 2009).

4.3 LITERATURE REVIEW

Plenty of approaches have been developed to account for PS and relatedness in a population-based GWAS. In the following section we briefly explain properties of some of the most popular ones, and then discuss how the power of a GWAS might be improved by combining data from different population and simultaneous analysis of all markers.

4.3.1 Genomic Control

GC (Bacanu et al. 2000; and 2002; Devlin and Roeder 1999; Devlin et al. 2001) is concerned with the effect of PS and admixture on the variance of the $\chi^2$ test statistic. At the first step, the variance of the $\chi^2$ test statistic is empirically estimated using a set of null markers, assuming that the effect of PS is uniform across the genome. The null markers are non-functional random markers scattered across the genome that are known to be independent of the phenotype. Then this variance is compared with that of the $\chi^2$ distribution
to estimate an inflation factor ($\lambda$). Finally, the inflation factor is used to scale the observed variance of the test statistics at the marker loci of interest (Laird and Lange 2011). Assuming a set of L null markers has been genotyped, $\lambda$ can be estimated as (Laird and Lange 2011):

$$\hat{\lambda} = \frac{0.4549}{\text{median}(\chi^2_1, \ldots, \chi^2_L)}$$

where the value of 0.4549 corresponds to the median of the chi-square distribution with 1 df. Then the scaled test statistic for the marker of interest $k$ can be calculated as: $\chi^2_{GC} = \hat{\lambda} \cdot \chi^2_k$ which is asymptotically distributed as $\chi^2$ with 1 df. With no doubt, the method is simple and fast and even is applicable to pooled DNA samples (Zhu and Zhang 2010). Although GC conserves the type I error rate in the presence of PS, but this comes at the price of loosing its power for detection of true associations (Clayton 2008). Estimate of $\lambda$ is based on the assumption of ‘exchangeability’ of markers and hence applying GC for an association study in a selected population, where the distribution of the test statistic could be different for loci linked to a favorable QTL compared to the rest of the genome, might lead to high FP rates (Clayton 2008; Zhang et al. 2003). GC has lower power if compared to its rivals especially in samples with complex pattern of kinship and extensive PS (Astle and Balding 2009; Bacanu et al. 2000; Balding 2006; Marchini et al. 2004; Price et al. 2010; Setakis et al. 2006; Wang et al. 2012; Yu et al. 2006; Zhang et al. 2008).

### 4.3.2 Structured Associations

In SA approach (Alexander et al. 2009; Falush et al. 2003; Hoggart et al. 2003; Pritchard and Rosenberg 1999; Pritchard et al. 2000; Thornsberry et al. 2001) a set of null markers are first used to infer PS information of each individual in the sample and
then association tests are done within each stratum or the stratification information are modeled as covariates in the association models (Gao and Edwards 2001).

The idea behind the approach is to use a set of null markers to quantify the background level of PS and then proceed to test marker-trait association conditional on the estimated ancestries (Cardon and Palmer 2003). Most popular SA methods assume that the ancestry of each individual is drawn from one or more discrete sub-populations (the so-called “islands” model) (Astle and Balding 2009; Myles et al. 2009), an assumption that is not supported with real data (see Novembre et al. 2008, for an example). By default, these methods assume that all allelic associations between markers has been originated from an admixture or stratification event, therefore, they require a minimum spacing between markers (Cardon and Palmer 2003), for example around 1 cM within the European-ancestry populations (McKeigue 2008). STRUCTURE, the popular software developed based on a Bayesian model-based clustering algorithm (Pritchard et al. 2000) assigns individuals to different sub-populations \( (k) \) based on their genome-wide marker data. Of the three different models implemented in the program (Falush et al. 2003), the ‘linkage’ model seems to be more applicable for animal and plant breeding samples, because it has fewer restrictive assumptions compared to other models. This model accounts for ‘mixture LD’ (LD due to variation in ancestry) and ‘admixture LD’ (LD due to admixture), but ignores ‘background LD’ (LD due to drift) (Falush et al. 2003). With continuous selection applied to livestock and plant populations there is always a chance that some allelic combinations to be over-represented in a sample; this could create misleading results if ‘linkage model’ is used. Therefore, one need to make sure that the marker panel is dense enough to allow for admixture LD and yet is sparse enough that allows ignoring of background LD (Falush et
STRUCTURE results are sensitive to the prior assumption on $k$ (Gao and Edwards 2001; Garnier-Géré and Chikhi 2001; Zhang et al. 2010). Even when the program successfully detects PS it might not fully protect against it (Atwell et al. 2010; Zhao et al. 2007) and its success rate may vary depending on the trait (Aranzana et al. 2005; Zhao et al. 2007). SA methods have been shown to be suboptimal in protecting against FP in commercial crop or model organism populations (Aranzana et al. 2005; Atwell et al. 2010; Kang et al. 2008; Maloletti et al. 2007; Myles et al. 2009; Wang et al. 2012; Wei et al. 2010; Yu et al. 2006; Zhang et al. 2008; Zhao et al. 2007). Lastly, the algorithm is computationally intensive and not applicable for high-density marker data (e.g., Zhu and Zhang 2010).

### 4.3.3 Principal Components Analysis

Based on a set of null markers PCA, a multivariate dimension reduction technique, derives axes of variation (eigenvalues) that can be used as covariates to calculate ancestry adjusted phenotype and genotypes (Patterson et al. 2006; Price et al. 2006). These ancestries adjusted phenotypes and genotypes then are used to calculate association test statistic (Price et al. 2006). Alternatively, the top eigenvalues or principal components (PCs) can be used either as covariates in a multiple regression model or used to match the cases and controls in a case-control association study (Laird and Lange 2011; Zhu and Zhang 2010). The method is fast, makes none of the assumptions made by SA methods and also unlike SA (which is sensitive to the number of inferred strata) is robust to the number of modeled PCs (Sneller et al. 2009). The idea of PC-adjustment for protecting against PS is valid under an island model; however, this might not be the case when there is a sample with complex PS. PCA may produce artifactual PCs in the presence of outliers (Zhang 2010), long-range LD
on the genome (ASTLE and BALDING 2009; LAURIE et al. 2010; NOVEMBRE et al. 2008; PRICE et al. 2010; TIAN et al. 2008) and family structure or cryptic relatedness in the sample (ASTLE and BALDING 2009; PATTERSON et al. 2006). The success of PC adjustment to control FP is conditional on whether sufficient number of PCs are included in the model or not (ASTLE and BALDING 2009). While inclusion of insufficient number of PCs may reduce the chance of controlling FP, inclusion of extra PCs could undermine power of the association study (ZHANG 2010; ZHU and ZHANG 2010). Pruning of markers has been suggested (ASTLE and BALDING 2009) and applied as an ad-hoc procedure for reducing the correlation between adjacent markers (e.g., PAUSCH et al. 2011) before applying PCA. However, this could lead to loss of some subpopulation differences, as the partially independent markers might not optimally reflect the underlying genetic heterogeneity among individuals in a population (MICLAUS et al. 2009). On the other hand, if some of the SNPs that are truly associated with the trait of interest fall in the pruned regions then adjusting for PCs is counterproductive (LAURIE et al. 2010). Overall, still there is an uncertainty on the proper method of PC adjustment and the optimal criteria for selecting PCs to be retained in the model (PELOSO and LUNETTA 2011; SHRINER 2011).

4.3.4 Mixed Linear Models

Approaches described so far might be identified as two-stage methods of correcting for PS. These methods could eliminate the true association signals whenever the strength of association due to PS is comparable to that of a QTL. In such cases, simultaneously inferring PS and testing association has the advantage of being able to separate the true and false signals from each other (ARANZANA et al. 2005; PLATT et al. 2010). Methods like MLM (ATWELL et al. 2010; KANG et al. 2008; KENNEDY et al. 1992; LISTGARTEN et al. 2010;
WANG et al. 2012; Yu et al. 2006; Zhang et al. 2010) that explicitly model population and pedigree structure in a unified way have been shown to perform better than the others in controlling FP rates (Aranzana et al. 2005; Astle and Balding 2009; Atwell et al. 2010; Kang et al. 2008; Listgarten et al. 2010; Malosetti et al. 2007; Wang et al. 2012; Yu et al. 2006; Zhang et al. 2008; Zhang et al. 2010). The MLM approach, however, is computationally expensive if applied to large dataset (Zhang et al. 2010). On the other hand, the success of MLM in finding associations might depend on markers MAF\(^5\). In fact using MLM strong phenotypic associations are easier to detect when the MAF is low (Bergelson and Roux 2010).

4.3.5 Multiple regression methods

In principle, PS simply can be adjusted for by including in the model a set of ancestry-informative or null markers as covariates. These markers or a function of them can effectively serve as proxies for the underlying PS (Balding 2006; Laird and Lange 2011; Setakis et al. 2006; Wang et al. 2005; Zhu and Zhang 2010). Valdar et al. (2009) suggested modeling PS explicitly in a multi-marker association analysis (MMA) framework. Comparing single-marker association (SMA) with MMA model, they showed that family structure needed to be considered in the former model to get a reasonable power, whereas the later model could safely ignore pedigree relationship without losing power. However, in their study when the sample was highly structured the MMA model had the risk of more FP. Pikkukokana and Sillanpaa (2009) compared two scenarios of correcting for and ignoring of pedigree relationship in a Bayesian multiple regression (BMR) using empirical simulation and real data. For the empirical simulation they used marker genotypes of 52

\(^5\) Minor Allele Frequency
SNPs located on 8 chromosomes from 210 individual from 15 different families and for the real dataset they used 58 individuals from 4 families. The authors found that Bayesian MMA analysis without correction for kinship was capable of self-correcting for residual dependencies and did not produce spurious associations. In a comprehensive simulation study, SETAKIS et al. (2006) using logistic regression for association study of a binary trait were able to account for PS without explicitly modeling it.

**4.3.6 Multi-population association studies**

Recently, PUNIYANI (2010) proposed a multi-task regression-based technique for GWAS in a multi-population dataset. The joint analysis allowed them to filter out weak signals shown up in some populations, thus reducing the overall FP rates. In another recent study, ZAITLEN et al. (2010) showed that by taking advantage of the between-populations genetic variance the success rate of mapping QTL could be improved relative to a single-population association study. Their study indicated that the average power of QTL detection would be higher in combined analyses compared to single population analyses. Pooling of data from several populations may intensify the QTL signal because markers strongly linked to a QTL in one population might not be so in another population (EASTON et al. 2007; ZAITLEN et al. 2010). The study of ZAITLEN et al. (2010) demonstrated that a sample from a single homogeneous population with the maximum genetic diversity is not always the optimal choice, especially at the fine-mapping stage.

The significance of implicit modeling of PS when MMA is used has been subject of several studies (GU et al. 2009; IWATA et al. 2007 and 2009; PIKKUHOOKANA and SILLANPAA 2009; VALDAR et al. 2009) with somewhat contradictory findings regarding the proper way of accounting for PS (see discussion for more details). In the last few years, genomic
selection (GS) (MEUWISSEN et al. 2001) has shown promising results for predicting breeding values of selection candidates (see GODDARD and HAYES (2009); HAYES et al. (2009) and references therein). In this approach, the whole-genome marker effects are first estimated in a reference population (training dataset) and then these estimates are used to predict the breeding values of individuals in an independent dataset (validation dataset). Simulation studies of GS in multi-breed admixed populations (IBÁNEZ-ESCRICHE et al. 2009; KIZILKAYA et al. 2010; TOOSI et al. 2008; TOOSI et al. 2010) showed that the estimated marker effects in such samples might accurately predict the BV of purebred animals in a validation dataset, provided that marker density is sufficient to capture the shared ancestral LD across breeds. These results suggest that the performance of QTL mapping in an admixed population, using a Bayesian multiple-regression approach with high-density markers, may not be hampered by the spurious FP when breed composition and relatedness have not been explicitly accounted for. Therefore, the purpose of this study was to evaluate the performance of genome-wide QTL mapping in a highly structured admixed population typical of animal and plant breeding datasets using the BMR and compare that with the performance of a MLM approach, which has been the current method of choice for many recent GWAS.

4.4 METHODS

4.4.1 Population

A base population of unrelated individuals was stochastically simulated and used as the ancestral population of four pure breeds that were used to create admixed and crossbred populations. The base population was randomly mated for 1000 generations, including selfing, with an effective size ($N_e$) of 1000. To simulate the four purebred populations (referred to as breeds A, B, C and D hereafter), at generation 0 four independent random
samples of 100 animals were drawn from the base population, and each was randomly mated (including selfing) for another 50 generations, with $N_e$ of 100. A previous study (TOOSI et al. 2010) showed that this setting is successful for creating genetically diversified breeds.

In generation 1051, pure breed population sizes were increased to $N = 1000$. Each population was composed of 50 half-sib families with average size of 20 offspring per family created by random mating of sires and dams from the previous generation. No attempt was made to keep family sizes equal. These breeds were then crossed to create (AB), (AB)A, (AB)C and (AB)(CD) populations. This created 8 different populations (including the four pure breeds) of size 1000. To create an admixed population, first a quarter of each of the populations was sampled and then all samples were pooled together. The final population was a random sample of size 1000 from such population. This was done to ensure unequal representation of different breeds in the admixed population. The pure breed A and admixed datasets were used as the resource populations for QTL mapping, at generation 1053 (referred to as the training generation here after). In this study PCA based on the whole-genome marker genotypes was used to verify the population structure in the simulated admixed dataset.

4.4.2 Genome

A genome of size 600 cM composed of six chromosomes with 1000 equally spaced segregating markers on each was simulated. Markers were bi-allelic with starting allele frequency of 0.5 and a reversible random mutation rate of $2.5 \times 10^{-5}$. A binomial map function was used to simulate recombination and interference was allowed by setting the maximum number of uniformly and independently distributed crossovers on the chromosome to be 4 (KARLIN 1984).
4.4.3 Phenotypes

At the training generation, 15 segregating markers (MAF > 0.02) that were the closest to certain positions (on chromosomes 1 to 3, Table 4.1) and also had at least one segregating marker adjacent to them were chosen to represent the QTL with an assigned effect. Fifteen other markers with similar positions on chromosomes 4 to 6 (to be referred as null chromosomes herein) were also chosen to represent QTL but they did not contributed to the simulated phenotype. Markers that surrogate for QTL were removed from the marker panel before association analysis. To keep the genetic variance constant across the simulated datasets, the allelic substitution effects of the QTL were standardized such that each of them explained a predefined percentage of the total genetic variance of the admixed population. Only additive effects were simulated. With equal probability, one of the two alleles was chosen to be positive or negative. The scaled QTL effects then were summed over all QTL genotypes for each individual to compute an individual’s true BV. Finally, a standard normal deviate was added to each true BV to provide the phenotype of an individual for a quantitative trait with heritability 0.30. Following some preliminary analyses, each dataset was replicated 20 times to minimize the variation of the results from a given dataset. Overall, 32 different datasets were simulated that differed in their QTL minor allele frequencies and, to a lesser extent, QTL positions.

4.4.4 Association mapping methods

The following models were used to analyze the simulated datasets.

4.4.4.1 Single marker association analysis (SMA) - Simple regression analysis was used to examine association of each marker’s genotype with individual’s phenotypic value. Markers were fitted one-at-a-time using the following linear model:
where $\mathbf{y}$ is the vector of phenotypic values of size $n$, $\mathbf{1}$ is a vector of ones of length $n$, $\mu$ is the population mean, $\mathbf{w}$ is a vector of the genotypic values at a marker locus (0, 1 or 2; number of copies of an arbitrary allele at the locus being tested), $a$ is the fixed marker genotype effect and $\mathbf{e}$ is the vector of random residual errors. The model assumes that $\mathbf{y} \sim \mathcal{N}(\mathbf{1}\mu + \mathbf{w}a, \mathbf{I}\sigma^2_e)$. Obviously, with our admixed population data, the residuals are not independently, identically distributed (iid). This model was only used for comparison purpose and to access how the type I error rate is affected due to the violation of the underlying assumptions of association tests. Association analysis was done using the PLINK software package with its assoc option (Purcell et al. 2007). To examine $H_0: \beta_i = 0$ versus $H_1: \beta_i \neq 0$ the Wald test as implemented in PLINK was used.

### 4.4.4.2 Single marker association analysis with breed composition (SMABC)

The second model explicitly considered breed composition in the admixed population.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{\beta} + \mathbf{w}a + \mathbf{e}$$

where $\mathbf{X}$ is the $n \times q$ incidence matrix relating observations to breed composition, and $\mathbf{\beta}$ is the fixed vector of breed composition. It was assumed that breed composition is known without error. Further, it is assumed that $\mathbf{y} \sim \mathcal{N}(\mathbf{1}\mu + \mathbf{X}\mathbf{\beta} + \mathbf{w}a, \mathbf{I}\sigma^2_e)$. All other parameters and assumptions were the same as model [1]. ASReml (Gilmour et al. 2009) was used for association analysis. The Wald test as implemented in the software was used for significance tests of the marker-trait association and breed composition effects. Both models mentioned above are inadequate at least in the sense that they do not account for additive relationship in
the population. As a result, the assumptions under which the null hypothesis is being tested might not be valid.

4.4.4.3 Single marker mixed linear model (MLM) - Conventional mixed model analysis with fitting one marker at a time was applied as below:

\[ y = 1\mu + Zu + wa + e \]  \[3\]

where \( Z \) is the incidence matrix relating observations to the corresponding random effect and \( u \) is the vector of random additive genetic effects or BVs. It was assumed that

\[ y \sim N(1\mu + wa, ZGZ' + R) \] and \( u|\sigma_u^2 \sim N(0, G) \), where \( G = A\sigma_u^2 \) and \( R = I\sigma_e^2 \). Here \( A \) is the matrix of additive genetic relationships, where \( a_{ij} \) is twice the coefficient of coancestry between individuals \( i \) and \( j \), and \( \sigma_u^2 \) is the additive genetic variance. Other parameters were as introduced before. ASReml was used for analysis and testing of marker effect was based on the Wald test implemented in the software.

4.4.4.4 Single marker mixed linear model with breed composition (MLMBC) - The following mixed linear model was used for association analysis with fitting markers one at a time:

\[ y = 1\mu + X\beta + Zu + wa + e \]  \[4\]

It was assumed that \( y \sim N(1\mu + X\beta + wa, ZGZ' + R) \). All other assumptions and parameters used were as mentioned before.

4.4.4.5 Bayesian multiple regression (BMR) – Stochastic search variable selection (SSVS) is a hierarchical Bayesian model which stochastically searches for ‘promising’ subsets of predictors (George and McCulloch 1993). Properties of such models have been
discussed in detail elsewhere (GILKS et al. 1996; O'HARA and SILLANPAA 2009). We used the BayesC\pi method of HABIER et al. (2011).

\[
y = 1\mu + \sum_k \gamma_k w_k \alpha_k + e
\]  

[5]

where \( w_k \) is a column vector of marker genotypes at locus \( k \) and \( \gamma_k \) is a latent 0/1 variable showing absence or presence of marker \( k \) in the model. Note that \( \pi \) determines the sparsity of the model. Here \( \alpha_k \) is the random substitution effect of marker \( k \) and is priori assumed independently distributed as

\[
\alpha_k | \pi, \sigma_{\alpha_k}^2 = \begin{cases} 
0, & \text{with probability } \pi \\
\sim N \left( 0, \sigma_{\alpha_j}^2 \right), & \text{with probability } (1 - \pi)
\end{cases}
\]

where \( \sigma_{\alpha_j}^2 \) are a priori assumed iid scaled inverted chi-square variables with scale and shape parameters of \( S_{\alpha}^2 \) and \( \nu_{\alpha} \), respectively. In addition, it was assumed that the residuals are iid and \( e \sim N(0, \sigma_e^2) \). Further, a priori it was assumed that \( \sigma_e^2 \) follows a scaled inverted chi-square distribution with parameters \( S_e^2 \) and \( \nu_e \), respectively. A deterministic approach was used to find the hyperparameters of the prior distribution of the \( \sigma_{\alpha_j}^2 \). A Gibbs sampler was used to generate a MCMC chain of 100,000 samples with a burn-in period of length 10,000. Convergence of the chain was examined using the R software package CODA (PLUMMER et al. 2006) and visual inspection of the chain plots. The posterior inclusion probability (PIP) of a marker, \( Pr (\gamma_k = 1 | y) \), was calculated as the average of all post burn-in values of \( \gamma_k \).

4.4.4.6 Bayesian multiple-regression with breed composition (BMRBC) - This model was similar to model [5] except that breed composition was also included as a fixed effect.

4.4.5 Estimation of thresholds
To estimate the thresholds required for hypothesis testing each of the null chromosomes was divided to non-overlapping bins of 40 markers. For each bin, the minimum P-values for the non-Bayesian approaches (or the maximum PIP values for the Bayesian approaches) were recorded. For each dataset that was composed of 20 replicates, the set of these values for all bins on the null chromosomes were combined and then were used to determine the 5% (or the 95% for the Bayesian approaches) quantile of their distribution using the quantile function in R (R DEVELOPMENT CORE TEAM 2011). We refer to this approach that is based on the empirical distribution of P- or PIP-values on the null chromosomes, as NCHR method of finding threshold. This method might not be applicable to real datasets, but used here to make comparison between the non-Bayesian and the Bayesian approaches feasible (SAHANA et al. 2010). Therefore, for the non-Bayesian approaches an alternative method for finding thresholds was also used. SLIDE, (a Sliding-window approach for Locally Inter-correlated markers with asymptotic Distribution Errors corrected) was developed by HAN et al. (2009) as an alternative to permutation testing. The program first estimates the effective number of tests ($M_{eff}$) using a sliding window Monte-Carlo approach. The Bonferroni threshold then can be calculated by dividing the nominal p-values by the $M_{eff}$. The sliding window MCMC approach approximates the asymptotic multivariate normal distribution of test statistic and accounts for all correlation among markers within a sliding window. We ran SLIDE with a window of size 40 markers and applied 100k cycles of MCMC chain. The program estimated the $M_{eff}$ and then this number was used as the actual number of markers (rather than 5970 markers that were actually on the panel) for calculating the Bonferroni adjusted P-values. Quantile-Quantile plots (WILK and Gnanadesikan 1968) were used to characterize the
extent to which the distribution of P-values on the null chromosomes deviates from their expected distribution, when using different association analysis methods.

### 4.4.6 QTL detection and power calculation

For each replicate of a dataset, a QTL was declared detected if any of the markers within an interval of ±2 cM of it (40 markers) had a P-value smaller than the 5% threshold P-value (for the non-Bayesian analysis), or a PIP value larger than the 95% threshold PIP value (for the Bayesian analysis). Power was defined as the proportion of times that a QTL was detected out of 20 replicates in that dataset.

**Type I error rates (FPR), Accuracy and positive prediction values (PPV)** – Excluding the ±2 cM intervals harboring the QTL, the remaining parts of the chromosomes 1 to 3 were divided into segments of length 4 cM as intervals where the null hypothesis was correct. In any of these regions if a marker was declared significant it was regarded a false-positive. FPR was the proportion of false-positives across the genome and then averaged over all replicates of a single dataset. Comparing power of methods that have different FPR could be misleading because positive results might be due to PS as well. Therefore, two other measurements, Accuracy and PPV were also calculated for evaluating the performance of different models (Zhang et al. 2008). Positive and negative results falling in the H₁ regions (interval where H₀ has to be rejected) were counted as true-positive (TP) and false-negatives (FN), respectively. The same way, positive and negative results in H₀ regions were regarded as false-positives (FP) and true-negatives (TN). Then accuracy and PPV were defined as below.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]
\[ PPV = \frac{TP}{TP + FP} \]

All performance measures, i.e., accuracy, power, FPR and PPV, were calculated on a per dataset basis and then averaged across all datasets. Also, instead of reporting three separate measurements for each QTL size, i.e., large, medium and small QTL, we decided to combine all three into one single statistic by taking a weighted average over all QTL (1 large, 4 medium and 10 small QTL).

4.5 RESULTS

4.5.1 Population stratification

PCA of the marker data for the purebred and admixed populations revealed that there are distinct clusters of related animals within the admixed population as opposed to the purebred one (Figure 4.1). In addition, the effect of breed composition was highly significant in all of our non-Bayesian analyses where this term was in the model. This makes proper modeling of population structure a must in order to conserve the type I error rate.

4.5.2 Distribution of P-values on the null chromosomes

Examination of the Q-Q plots of P-values of the markers on null chromosomes showed spurious FP in association analysis of both the purebred and the admixed populations when relationship and/or breed composition were not modeled properly (Figure 4.2). As it was expected, the SMA of the admixed (ADMX) population showed an incredibly high rate of spurious associations. For the purebred (PB) population, MLM was able to control FPR, whereas for the ADMX population it failed to do so (see MLM ADMX and MLM BC ADMX Q-Q plots in Figure 4.2).
4.5.3 Single marker regression

Results of SMA and MLM analysis of the ADMX population are shown in Tables 4.2 and 4.3. As it was expected, the SMA model ended up with the lowest accuracy and PPV and the highest FPR among the four tested models. With the NCHR method of finding thresholds (Table 4.2), modeling of breed composition increased power of QTL detection and PPV by 60% and 20-30%, respectively, at the cost of inflation of FPR by nearly 20%. Whereas, accuracy of QTL detection did not change regardless of the method used for the analysis. On the other hand, modeling breed composition dramatically improved accuracy, FPR and PPV when the SLIDE method was used for finding thresholds (Table 4.3). The accuracies of models accounting for breed composition were 30-60% higher than those that did not. Also, modeling of breed composition improved PPV by 300%. However, modeling of breed composition resulted in loss of power by more than 50% when SLIDE method was used.

4.5.4 Bayesian Multiple regression

Table 4.4, shows the results of the Bayesian association analysis in the ADMX population. With density of 10 markers per cM adding breed composition as a fixed effect to the model reduced both power and FPR by 11% and 13%, respectively. However, accuracy and PPV of QTL detection remained nearly unchanged.

4.6 DISCUSSION

In this simulation study, we compared three different methods of GWAS in an admixed population: single marker simple regression, single marker mixed model and Bayesian multiple regression models. In each of the methods, two scenarios were compared: ignoring and fitting of breed composition. Our results showed that with the traditional single marker association analysis, modeling of both the pedigree relationship and breed composition
improves the performance of the model. Whereas, with the Bayesian multiple-regression association analysis, it was shown that modeling of breed composition not only does not improve the accuracy and the positive predictive values- but also it reduces the power of GWAS. Overall, the multi-marker association models outperformed the single marker association models.

The PCA showed that there is a distinct PS in the ADMX dataset. In a study by Toosi et al. (2010) who simulated different breeds using the same scenario as described here, the between breeds genetic distance was nearly 24% as measured by the Wright’s FST statistic. In other words, the breed differences could explain nearly 24% of the total genetic variance of the ADMX population. PS will be a source of spurious associations if both allele frequencies and mean phenotypic values of a certain trait vary across the sub-populations (Hirschhorn and Daly 2005). The effect of breed composition was found to be highly significant (P<0.001) for both SMA and MLM analysis (data not shown).

The effectiveness of accounting for breed composition can be seen from Figure 4.2. The PB dataset showed spurious FP due to the pedigree relationship that existed in the sample, when all the sample individuals regarded as unrelated to each other. Unequal relatedness within a sample can result in increased FP rates in two ways: first, regions where QTL are residing may be co-inherited with regions completely devoid of QTL (Payseur and Place 2007) and second, genotype correlations within larger families can have a larger impact on the association results compared to the smaller families (Peirce et al. 2008). Kennedy et al. (1992) showed that for both randomly mated and selected populations with complex pedigrees the use of MLM approach provides unbiased estimates and exact tests of associations, whereas the ordinary least squares does not. If dependencies among study
subjects are not accounted for many statistical tests of association are not strictly valid (Newman et al. 2001). For the PB dataset, the MLM approach was able to control FPR at the nominal level on the null chromosomes, as it is evident from Figure 4.2, but it failed to do so for the ADMX dataset when breed composition was ignored. The extent of FPR is a function of the extent to which the population is structured (Aranzana et al. 2005), therefore, for the highly divergent breeds simulated in our study modeling of breed composition is necessary for controlling of FPR.

In a recent GWAS in a massively structured population consisted of 1800 bulls of the German Fleckvieh breed, Pausch et al. (2011) applied the same SMA model as we used here and observed extensive significant association signals possibly due to the variation of the relatedness between and within the families in the sample. Very recently, Wang et al. (2012) conducted a GWAS of several morphological and agronomic traits in a highly structured population of barley cultivars. When they used a similar SMA model an excessive number of significant associations were found.

Comparing the different analytical models for the ADMX population, the best model in term of conserving FPR was the MLM approach that explicitly modeled breed composition. This has been the preferred method in GWAS with multi-breed datasets (Mujibi et al. 2011; Snelling et al. 2010; Snelling et al. 2011). Even though the MLM approach without modeling breed composition performed better than its corresponding SMA approach, it showed spurious associations, implying that some of the markers that were not associated with any QTL were picking up breed differences. In this situation, an arbitrary marker that has different allele frequencies across different breeds shows association with the phenotype under study.
The changes in the performance of SMA models when fitting and not fitting breed composition were most evident with the SLIDE method of finding thresholds (Tables 4.2 and 4.3). While modeling of breed composition with the NCHR method improved the power of QTL detection, this was not the case when SLIDE method was applied. On the other hand, for SMA_{BC} model reduced FPR dramatically compared to the SMA model. This result agrees with the result of IWATA et al. (2007) who made a similar comparison. Care should be taken when comparing the power of two methods that have different FPR. Positive results could be due to both true QTL signals and PS (ZHANG et al. 2008). This is evidenced by the high FPR of models not fitting breed composition (compared to those fitting it), when SLIDE method was used for hypothesis testing. Further, modeling of BC sharply improved both the accuracy and the PPV of QTL detection. The SMA_{BC} models performed similar to the MLM_{BC}, even thought that SMA_{BC} did not fully account for the relatedness in the sample. It might be argued that correcting for the breed composition effects (or, PS in general) is an indirect way of correcting for relatedness among individuals in the sample and hence less spurious associations are expected (MALOSETTI et al. 2007).

While there were no difference between accuracies of the SMA and MLM or the SMA and the SMA_{BC} when the NCHR method was used, there was a noticeable difference between them when SLIDE method was applied. As an example for the SMA and SMA_{BC} methods, consider their accuracy (0.58 and 0.92, respectively), power (0.72 and 0.30, respectively) and FPR (0.44 and 0.007, respectively). It is evident that many of the significant results of the SMA have been FP. Also, the difference between their accuracy implies that modeling of BC has dramatically increased the number of TN. Preventing the confounding effect of PS comes at a cost, however. Explicit modeling of PS introduces some
FN (Andersen et al. 2005; Bergelson and Roux 2010; Brachi et al. 2010; Ingvarsson and Street 2011). Adjusting for PS may cancel out the effect of QTL that contribute to phenotypic differences between breeds (Sneller et al. 2009). Anderson et al. (2007) conducted a GWAS on 32 lines of European inbred maize with different line origins. Comparing a model that adjusted for line origin versus the one that did not, they showed that several true QTL remained undetected when the former model was used, because these polymorphisms were confounded with the line origin. This confounding is especially important for traits that have experienced adaptive selection and thus their variation may coincide with PS (Brachi et al. 2010; Veyrieras et al. 2007).

When the sample under study is not heavily structured (e.g., PS is due to a very recent genetic drift) but still is composed of related individuals, using of MLM without considering breed composition (by fitting PC covariates, for example) has been suggested as the preferred analytical approach (Price et al. 2010). Obviously, this was not the case in our simulation, as evidenced by the comparison of the performances of MLM methods with and without fitting BC. In our study, fitting of BC resulted in a considerable drop of power of QTL detection when SLIDE method was used, but this was compensated for with a significant drop in FPR. Further inspection showed that in most instances it was the smallest QTL effects that were missed. This finding agrees with Iwata et al. (2009) who showed that smaller QTL effects have larger FN rates.

To control family-wise type I error rate, SMA requires methods like Bonferroni correction for multiple testing. Such adjustments are usually too conservative, especially in large scale SMA with a lot of LD between linked markers, and thus they may cause true associations to be missed (Gu et al. 2009; Stacey et al. 2010; Udler et al. 2009; Zhang et
That is why for most complex polygenic traits SMA only detects a very small proportion of genetic variation (Cho et al. 2010; Hoggart et al. 2008). Comparison of the performance of the BMR models with and without fitting BC (Table 4.4) indicates that in the MMA framework controlling the confounding effects of PS and kinship might be unnecessary. The BMR model performed much better than the MLM and MLM_{BC}. While modeling of BC in the MLM approach improved power of association (Table 4.2), it resulted in loss of power when BMR was used.

As it was expected, the FPR of the MMA methods were less than that of the SMA methods. One major concern with SMA is that it ignores the information contained in the joint distribution of all markers (Balancing 2006; Parker et al. 2007; Zhang et al. 2011). A marker’s marginal effect might be different with its effect when it is considered jointly with some other markers. MMA models that decide whether to add or skip adding a marker to the set of pre-existing markers in the model have the advantage of lowering FPR over a SMA model. In fact, once the marker with the strongest marginal correlation with the phenotype is in the model, other markers that are in LD with this marker but do not provide additional information about the phenotype are automatically discarded (Gu et al. 2009; He and Lin 2011; Puniani et al. 2010). MMA analysis also improves performance over SMA tests, first because a weak signal may be more apparent when other QTL are already accounted for, and second because a false signal may be weakened by inclusion in the model of a stronger signal from a real QTL (Hoggart et al. 2008).

For complex traits there are possibly multiple genes across the genome that each have a small effect picked up by markers in LD with them; therefore, a MMA model better explains the true underlying genetic architecture of the trait than a SMA model (Chapman and
Whittaker 2008; Fridley 2009; He and Lin 2011). Atwell et al. (2010) in a GWAS of more than 100 phenotypes in inbred lines of Arabidopsis thaliana showed that GWAS yields unambiguous results for monogenic (or mostly controlled by a single major gene) characters, regardless of whether they correct for PS or not. As the authors concluded, the reason was not because there were no confounding effects, but it was because the true signals were showing the strongest associations. Platt et al. (2010) showed that for a simply inherited trait with sufficient sample size, the most significant results are expected to be real QTL, or if they are not present on the marker panel, the marker with the strongest LD with them. Therefore, Atwell et al. (2010) suggested the problem of confounding due to PS in GWAS of quantitative traits might be better thought as a model misspecification. That is, modeling of a polygenic trait using a SMA that ignores the multi-factorial background of the trait and implicitly assumes that a single QTL is causing all the phenotypic variation (Atwell et al. 2010). When we attempt to model traits that are in fact result of the collective action of multiple factors, that might be correlated and/or epistatically interacting with each other, using single-locus models our results could be severely biased. Not only we do get spurious false-positives across the genome, but also we may find the strongest associations on chromosomes that are completely devoid of QTL (Platt et al. 2010). Therefore, Platt et al. (2010) suggested that the real goal of GWAS in controlling of PS effects should be accounting for the confounding effects of multiple QTL rather than modeling of PS per se.

Our results agree with Setakis et al. (2006), Iwata et al. (2007), Iwata et al. (2009), Pikkukookana and Sillanpaa (2009) and Valdar et al. (2009) who demonstrated that unlike SMA models, the MMA models are able to self-correct for family structure. However, with a highly structured sample MMA model of Valdar et al. (2009) showed more FPR.
Therefore, they suggested that the MMA model should be applied after major structural variations in the population have been corrected for (VALDAR et al. 2009). IWATA et al. (2007) proposed a Bayesian MMA for an empirical GWA analysis in a rice germplasm collection. Analysis of simulated data based on real marker genotypes revealed that their model could successfully conserve both FP and FN over SMA models. Comparing a MMA model that fitted PS with a model that did not; they noticed that the former model was slightly better in controlling FN and FP (IWATA et al. 2007). In a SMA framework, Yu et al. (2006) and ZHAO et al. (2007) compared FPR of a MLM model that fitted both PS (Q) and polygenic effects (based on a marker-derived relationship matrix, K) versus a model that fitted only polygenic effects. Their results showed a small advantage of the Q + K model over the K only model. In a very recent GWAS in a highly structured population of barely cultivars, WANG et al. (2012) compared the performance of the Q + K model with the K only model, GC, SA and PCA models. The K only model outperformed all the other rivals. In fact, the K matrix already contains all the information on PS and hence the strict modeling of PS might not be necessary, given high-density marker data (ASTLE and BALDING 2009; GODDARD and MEUWISSEN 2005). Nonetheless, the presence of PS term in the model might provide a low dimensional summary of the key features of the relationship in the population (ASTLE and BALDING 2009). This might describe why the Q + K model, in case of SMA analysis, or the MMA models of IWATA et al. (2007) and VALDAR et al. (2009) had some advantage over models that did not fit PS implicitly.

The studies of VALDAR et al. (2009) and IWATA et al. (2007) had barely enough markers. It seems that when marker density is sufficient, each marker explains a part of the effects of kinship and PS such that the overall effects of variation of relatedness between and within
sub-populations are diluted. This argument agrees with Sillanpää (2011) point of view that states in MMA models variable selection is done simultaneously with the effect estimation and thus, the large number of markers considered jointly might account for many types of variations.

Perhaps, our results are most comparable to those of Gu et al. (2009) who applied a modified forward multiple regression (MFMR) approach based on maximum order statistics in an empirical GWAS. Their simulation was based on a 115k Affymetrix SNP marker panel and a dataset that was mainly composed of Caucasians, Blacks, and Hispanics races. They picked up three independent SNPs that were significantly correlated with race to surrogate for QTL. With comparing the results of SMA and MFMR analysis they showed that FPR of MFMR approach was not influenced by PS. This implies that once the QTL that is correlated with PS is included in the multiple regression model, the effect of PS has been accounted for (Gu et al. 2009). Pikkuhookana and Sillanpaa (2009) who used a BMR model for a clinical QTL study in a sample with family structure, showed that regardless of having a correction term for PS in the model the MMA fits a few extra small effects markers. As a result, the MMA model was able to conserve both FP- and FN-rates. Another interesting finding in Gu et al. (2009) that agrees with our result, was that fitting of PS in the MFMR model reduced power without changing of the FPR. Whenever we attempt to control FPR in a GWAS some FN are inevitable (Bergelson and Roux 2010; Ingvarsson and Street 2011; Zhao et al. 2007). If the distribution of a QTL is highly correlated with PS, the effect of the allele may be absorbed in the population effects and the QTL will be obscured (Andersen et al. 2007; Brachi et al. 2010; Ingvarsson and Street 2011; Iwata et al. 2009; Kang et al. 2010; Zhao et al. 2007).
An important feature of using a multi-population sample is that it makes the detection of QTL causing between-population differences possible. A combined analysis of data from several populations (or breeds, races) takes the advantage of using the between-population genetic variability and hence is more powerful than single-population association study (Guo et al. 2008; Parker et al. 2007; Rosenberg et al. 2010; Zaitlen et al. 2010). A pooled sample of populations has potentially more informative recombination events and shorter haplotype length due to narrower LD distances across breeds (Goddard et al. 2006; Parker et al. 2007; Toosi et al. 2010). Hence, association mapping in such populations is much more accurate, although this comes at a cost: a much higher maker density is needed (Goddard and Hayes 2009; Hamblin et al. 2011; Hayes et al. 2009; Toosi et al. 2010). In addition, inferring GxE interactions without a multi-population dataset is difficult, if not impossible (Guo et al. 2008). The advantages of using a multi-population sample might get offset when a SMA is used for GWAS. As we showed in this study and in agreement with studies of Gu et al. (2009) and Zhao et al. (2007), explicit accounting for PS results in some FN. In fact, any method that effectively eliminates confounding due to PS will also effectively removes QTL that are highly correlated with PS (Zhao et al. 2007). This might be more of a problem with the more subtle QTL effects typical of quantitative traits and the small sample sizes usually available for GWAS in animal breeding. Whereas, the BMR method as shown here not only is capable of reducing FN due to explicit modeling of PS but also it reduces FN resulted from the highly conservative multiple-test correction methods like Bonferroni.

Teo et al. (2009) showed that in the presence of opposing LD between populations, i.e., change of LD phase between a marker and a QTL across populations, might have a negative
impact on power in case-control or family-trio association studies. We think by fitting all markers simultaneously we might overcome this problem. It is unlikely that all markers that are in LD with a specific QTL change their LD phase especially if they are close enough to the QTL. In a multi-population sample, markers in strong LD with QTL have less distance to the QTL compared to that in a single-population sample (Toosi et al. 2010). In contrast, the SMA models that use GC, SA or PCA for controlling the confounding effect of PS, might not be able to correct for the LD differences that reduces the power in a multi-population association study (Teo et al. 2009). Several studies have shown that leveraging the LD differences across populations may amplify the signal of QTL, because markers strongly liked to a QTL in one population may not be even segregating in another population (Easton et al. 2007; Udlér et al. 2009; Zaitlen et al. 2010).

As the GS approaches are gaining popularity, some recent GWAS in animal breeding dealing with multi-breed datasets have used methods similar to the BMR method we used here (Mujibi et al. 2011; Snelling et al. 2011). In these studies, breed composition has been added as a fixed effect to the model. If breed composition is fully confounded with contemporary group effects (e.g., slaughter date or geographical region) adding it as a fixed effect in the BMR model might be necessary. This will suppress association signals that are due to the correlation of the phenotype with the contemporary group effect and thus reduces FP. However, if breed composition is not confounded with any other effect and a MMA model like what we used is applied, then the cost of implicit accounting for it might be increase of FN.

In conclusion, our results showed the superiority of MMA models over SMA models. More specifically, our study confirms that MMA models are capable of automatically
suppressing the confounding effects of relationship and population structure in genome-wide association studies without compromising their power of QTL detection.

4.7 LITERATURE CITED


ANDERSEN, J., I. ZEIN, G. WENZEL, B. KRÜTZFELDT, J. EDER et al., 2007 High levels of linkage disequilibrium and associations with forage quality at a Phenylalanine Ammonia-Lyase locus in European maize (Zea mays L.) inbreds. TAG Theoretical and Applied Genetics 114: 307-319.


SILLANPÄÄ, M. J., 2011 Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity 106: 511-519.


WEI, X., P. A. JACKSON, S. HERMANN, A. KILIAN, K. HELLER-USZYNSSKA et al., 2010 Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker–trait associations in sugarcane. This article is one of a selection of papers from the conference “Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming”. Genome 53: 973-981.


4.8 FIGURES

Figure 4.1 – Scatter plots of the first two principal components of the genome-wide markers in the admixed (left) and the purebred (right) populations. Numbers in brackets show the percentage of variances explained by corresponding PCs. Different colors represent various breed compositions.
Figure 4.2- Q-Q plots of the observed distribution of $-\log_{10}(P)$ values on the null chromosomes, with different analysis approach, vs. their expected distribution; PB = Purebred population, ADMX = Admixed population, SMA = Single Marker Association, SMA_BC: SMA with Breed Composition, MLM = Mixed Linear Model association, MLM_BC = MLM with Breed Composition.
4.9 TABLES

Table 4.1 - QTL positions (Morgam), Mean, Standard deviation, Minimum and Maximum of $h^2_{QTL}$ across 32 simulated datasets

<table>
<thead>
<tr>
<th>Chrom.</th>
<th>QTL</th>
<th>Position</th>
<th>Mean ($h^2_{QTL}$)</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.150</td>
<td>0.06</td>
<td>0.011</td>
<td>0.030</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.600</td>
<td>0.01</td>
<td>0.002</td>
<td>0.010</td>
<td>0.02</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.610</td>
<td>0.01</td>
<td>0.003</td>
<td>0.004</td>
<td>0.02</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.750</td>
<td>0.03</td>
<td>0.005</td>
<td>0.020</td>
<td>0.04</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.950</td>
<td>0.01</td>
<td>0.003</td>
<td>0.010</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1.200</td>
<td>0.03</td>
<td>0.008</td>
<td>0.010</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1.210</td>
<td>0.01</td>
<td>0.003</td>
<td>0.003</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>1.550</td>
<td>0.01</td>
<td>0.003</td>
<td>0.004</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>1.600</td>
<td>0.01</td>
<td>0.002</td>
<td>0.010</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1.800</td>
<td>0.03</td>
<td>0.007</td>
<td>0.010</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>2.050</td>
<td>0.01</td>
<td>0.003</td>
<td>0.003</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>2.150</td>
<td>0.01</td>
<td>0.003</td>
<td>0.003</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>2.250</td>
<td>0.01</td>
<td>0.002</td>
<td>0.010</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>2.400</td>
<td>0.01</td>
<td>0.002</td>
<td>0.010</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>2.700</td>
<td>0.03</td>
<td>0.006</td>
<td>0.010</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table 4.2- Accuracy, Power, Prob. Of Type I Error and Positive Predictive Value (PPV) for SMA and MLM analysis with NCHR method of finding thresholds in the ADMX population

<table>
<thead>
<tr>
<th></th>
<th>SMA</th>
<th>SMA&lt;sub&gt;BC&lt;/sub&gt;</th>
<th>MLM</th>
<th>MLM&lt;sub&gt;BC&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.86 (0.003)</td>
<td>0.87 (0.005)</td>
<td>0.86 (0.005)</td>
<td>0.87 (0.004)</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>0.40 (0.027)</td>
<td>0.63 (0.038)</td>
<td>0.40 (0.049)</td>
<td>0.64 (0.035)</td>
</tr>
<tr>
<td><strong>Type I error rate</strong></td>
<td>0.08 (0.003)</td>
<td>0.10 (0.007)</td>
<td>0.08 (0.006)</td>
<td>0.11 (0.006)</td>
</tr>
<tr>
<td><strong>PPV</strong></td>
<td>0.34 (0.016)</td>
<td>0.43 (0.014)</td>
<td>0.36 (0.031)</td>
<td>0.43 (0.012)</td>
</tr>
</tbody>
</table>

1. Numbers in brackets are SE of means.
Table 4.3- Accuracy, Power, Prob. Of Type I Error and Positive Predictive Value (PPV) for SMA and MLM analysis with SLIDE method of finding thresholds in the ADMX population\(^1\)

<table>
<thead>
<tr>
<th></th>
<th>SMA</th>
<th>SMA(_{BC})</th>
<th>MLM</th>
<th>MLM(_{BC})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.58 (0.047)</td>
<td>0.92 (0.003)</td>
<td>0.69 (0.078)</td>
<td>0.92 (0.002)</td>
</tr>
<tr>
<td>Power</td>
<td>0.72 (0.034)</td>
<td>0.30 (0.026)</td>
<td>0.63 (0.063)</td>
<td>0.27 (0.021)</td>
</tr>
<tr>
<td>Type I error rate</td>
<td>0.44 (0.056)</td>
<td>0.007 (0.001)</td>
<td>0.30 (0.096)</td>
<td>0.004 (0.001)</td>
</tr>
<tr>
<td>PPV</td>
<td>0.25 (0.030)</td>
<td>0.85 (0.018)</td>
<td>0.34 (0.057)</td>
<td>0.90 (0.017)</td>
</tr>
</tbody>
</table>

1. Numbers in brackets are SE of means.
Table 4.4- Accuracy, Power, Prob. Of Type I Error (FPR) and Positive Predictive Value (PPV) for BMR analysis in the ADMX population

<table>
<thead>
<tr>
<th></th>
<th>BMR</th>
<th>BMR&lt;sub&gt;BC&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.89 (0.002)</td>
<td>0.89 (0.002)</td>
</tr>
<tr>
<td>Power</td>
<td>0.65 (0.016)</td>
<td>0.58 (0.017)</td>
</tr>
<tr>
<td>Type I error rate</td>
<td>0.08 (0.001)</td>
<td>0.07 (0.001)</td>
</tr>
<tr>
<td>PPV</td>
<td>0.51 (0.007)</td>
<td>0.51 (0.008)</td>
</tr>
</tbody>
</table>

1. Numbers in brackets are SE of means.
CHAPTER 5. GENERAL DISCUSSION AND CONCLUSION

This thesis presents an investigation of how the nature of linkage disequilibrium in purebred, crossbred and admixed populations affects the accuracy of genomic prediction and the performance of genome-wide association studies. It was shown that the rate of decay of LD with distance between loci varies significantly across these populations, with the steepest rate of decay of LD observed in a four-way crossbred and the slowest in a purebred population. In addition, it was observed that weak LD ($r^2 \geq 0.10$) extends farther in admixed and crossbreds (up to 15 cM in an F$_1$ population) than in purebred populations (less than 4 cM). This is in agreement with the practice of using crossbred populations for QTL mapping when the marker density is low. In contrast, an important finding of this research is that strong LD ($r^2 \geq 0.70$, for example) extends farther in purebreds than in crossbred and admixed populations. It was found that the length of segments with strong LD is inversely related to the number of breeds contributing to the cross. Results of the canine genome project have also shown that haplotype blocks are several Mb long within a breed but they are much shorter across breeds, extending only to tens of kb (LINDBLAD-TOH et al. 2005). The shared haplotype between several breeds that have been isolated for many generations traces back to a common haplotype in their ancestral population. These haplotypes must be very narrow in order to remain intact following the large number of meiosis events that separates the breeds. This agrees with our finding that a training population of crossbred AC, predicts breeding values of breed B more accurately than a population of the same size but consisting of only breed A or C. That is the use of crossbred AC forced the model to pick the shared haplotype between A and C, rather than the breed-specific haplotypes present in A or C. The shared haplotype between A and C is more likely to be present in breed B as well.
Obviously, the shorter the length of a haplotype, the higher is the chance of its similarity across populations. In the same way, as distance in time between two sub-populations increases, there is a higher chance for recombination to break down the LD that was present in the ancestral population and drift to create new LD within each sub-population (Goddard et al. 2006; Hill and Robertson 1968).

In summary the above results confirm that while the purebreds are suitable for fine-mapping, multi-breed populations, which have larger effective population sizes, are preferred for an ultra fine-mapping step (Easton et al. 2007; Hayes et al. 2009a; Hayes et al. 2009b; Parker et al. 2007; Zaitlen et al. 2010).

Another important finding of this research is that training in crossbred or admixed populations can be used for GS of purebreds with little or no drop in accuracy compared to training in purebreds. When training is in a cross between breeds A and B, for example, the results can be applied for GS in both breeds with a high accuracy. However, when results of training in breed A are applied for GS of breed B the accuracy would be much lower. This is an advantage of training in a crossbred or admixed population. The high accuracy of GS of purebreds with training in crossbred or admixed populations implies that this accuracy stems from the short segments with strong LD rather than from the longer segments with weak LD. The longer segments with weak LD are not expected to contribute to the accuracy of GS because meioses between the training and validation populations are expected to break up these longer segments. However, a higher marker density is needed for training in a crossbred population compared to training in a purebred population. The reason for this is, as observed in here, that short haplotypes with strong LD are more frequent in purebreds than in crossbred populations.
In this thesis we have also demonstrated that with Bayesian multiple-regression explicit modeling of breed composition and pedigree structure is not necessary. Numerous studies have shown that population structure such as breed composition or pedigree relationship, are sources of cofounding in GWAS. Therefore, at the first glance the need for accounting for breed composition and pedigree relationships in the analysis seems to be necessary. However, when a Bayesian multiple-regression model was used, fitting of breed composition did not have an impact on the false-positive rate and slightly reduced the power of QTL mapping. In contrast, when each marker was fitted separately using ordinary least square or a mixed model approach, the results showed that fitting of population structure was required for controlling of false-positives as it has been emphasized in numerous studies.

It seems that when fitting all markers simultaneously, each maker absorbs part of the effects of population structure and relatedness, but because these effects are distributed across many markers none of them might be individually significant (Setakis et al. 2006). Alternatively, if all QTL were included in the model then breed differences do not provide any additional information. Thus, in this case population structure does not have to be explicitly accommodated in the model. This would also be the case, when all QTL are not in the model but the marker density is sufficiently high to account for all genetic variability due to QTL.
LITERATURE CITED


APPENDIX 1. GENOMIC SELECTION OF PUREBREDS FOR CROSSBRED PERFORMANCE

A paper published in Genetics Selection Evolution (2009) 41:12

Noelia Ibánéz-Escrive¹, Rohan L Fernando², Ali Toosi² and Jack CM Dekkers²

¹Genètica i Millora Animal- Centre IRTA Lleida, 25198 Lleida, Spain
²Department of Animal Science, Iowa State University, Ames 50011-3150, USA

A1.1 ABSTRACT

Background

One of the main limitations of many livestock breeding programs is that selection is in pure breeds housed in high-health environments but the aim is to improve crossbred performance under field conditions. Genomic selection (GS) using high-density genotyping could be used to address this. However in crossbred populations, 1) effects of SNPs may be breed specific, and 2) linkage disequilibrium may not be restricted to markers that are tightly linked to the QTL. In this study we apply GS to select for commercial crossbred performance and compare a model with breed-specific effects of SNP alleles (BSAM) to a model where SNP effects are assumed the same across breeds (ASGM). The impact of breed relatedness (generations since separation), size of the population used for training, and marker density were evaluated. Trait phenotype was controlled by 30 QTL and had a heritability of 0.30 for crossbred individuals. A Bayesian method (Bayes-B) was used to estimate the SNP effects in the crossbred training population and the accuracy of resulting GS breeding values for commercial crossbred performance was validated in the purebred population.
**Results**

Results demonstrate that crossbred data can be used to evaluate purebreds for commercial crossbred performance. Accuracies based on crossbred data were generally not much lower than accuracies based on pure breed data and almost identical when the breeds crossed were closely related breeds. The accuracy of both models (ASGM and BSAM) increased with marker density and size of the training data. Accuracies of both models also tended to decrease with increasing distance between breeds. However the effect of marker density, training data size and distance between breeds differed between the two models. BSAM only performed better than AGSM when the number of markers was small (500), the number of records used for training was large (4000), and when breeds were distantly related or unrelated.

**Conclusion**

In conclusion, GS can be conducted in crossbred population and models that fit breed-specific effects of SNP alleles may not be necessary, especially with high marker density. This opens great opportunities for genetic improvement of purebreds for performance of their crossbred descendents in the field, without the need to track pedigrees through the system.

A1.2 INTRODUCTION

One of the main limitations of many livestock breeding programs is that selection is in purebred nucleus lines or breeds that are housed in high-health environments but the goal of selection is to improve crossbred performance under field conditions. Due to genetic differences between purebreds and crossbreds and environmental differences between nucleus and field conditions, performance of purebred parents can be a poor predictor of performance of their crossbred descendents [1]. Furthermore, some important traits such as
disease resistance cannot be measured in nucleus lines. In order to avoid these problems, it has been proposed to select purebred relatives based on crossbred performance using combined crossbred and purebred selection or CCPS [2-6]. This approach can increase response to selection for crossbred performance relative to the classical method of selection on purebred performance [7]. It has, however, not been extensively implemented in livestock due mainly to the difficulty and cost of routine collection of phenotypic and pedigree data from crossbreds in the field [1]. In addition, using CCPS increases the rate of inbreeding [8] and makes it difficult to accommodate non-additive gene action [6]. As an alternative to CCPS, Dekkers [1] proposed to select purebreds for commercial crossbred performance using genomic selection.

In livestock, genomic selection is becoming increasingly feasible because of the availability of massive numbers of single nucleotide polymorphism (SNP) markers. This approach consists of predicting breeding values on the basis of a larger number of SNPs [9-11], utilizing linkage disequilibrium (LD) between SNPs and the QTL. Genomic selection of purebreds for crossbred performance involves estimating effects of SNPs on crossbred performance, using phenotypes and SNP genotypes evaluated on crossbreds, and applying the resulting estimates to SNP genotypes obtained on purebreds (Dekkers 2007). Genomic selection for crossbred performance has three main advantages over CCPS: 1) it does not require pedigree information on crossbreds, 2) after estimates of SNP effects are obtained using genotype and phenotype data, prediction can continue for several generations without additional phenotypes [9], 3) it reduces the rate of inbreeding [12], and 4) it makes accommodating non-additive gene action easier [1]. The success of genomic selection depends mainly on the prediction accuracy of the estimated breeding values (GEBVs).
Several authors have studied the accuracy of these predictions by computer simulation [9,13,14]. However, these studies have focused on pure breeds. In crossbred populations, effects of SNPs may be breed specific because the extent of LD between SNPs and QTL can differ between breeds. Moreover, the LD may not be restricted to markers that are tightly linked to the QTL. Both these problems could be addressed by using a model with breed-specific effects of SNP alleles. Toosi et al. [15] evaluated simulated training populations consisting of crosses or mixtures of breeds and found the accuracy of genomic selection to be lower compared to using purebred data for training, but not by a large degree. They, however, used a genomic selection model in which SNP allele effects were assumed the same in all breeds. Thus, the objective of this study was to compare by computer simulation the accuracy of genomic selection of purebreds for commercial crossbred performance, using either the classical genomic selection model with across-breed effects of SNP genotypes (ASGM) or a model with breed-specific effects of SNP alleles (BSAM).

**A1.3 METHODS**

* **A1.3.1 Simulation**

In all simulations, the genome consisted of one chromosome of 1 Morgan with 6000 SNPs and 30 biallelic QTL. A gamma distribution with shape and scale parameters equal to 0.4 and 1/1.66 was used to sample the absolute value of effects of the QTL. The sign of the QTL effect was sampled to be positive or negative with probability 0.5. Effects were rescaled to result in a genetic variance equal to 1.0. The phenotypic trait was simulated under additive gene action. Dominance and epistatic effects were not simulated but would be captured to the extent that they are incorporated in allele substitution effects (see discussion).
In the base population, SNP and QTL alleles were sampled from a Bernoulli distribution with frequency 0.5. A mutation rate of $2.5 \times 10^{-5}$ per generation was applied in the following generations for all loci, where mutations switched the allele state from 1 to 2 or from 2 to 1. Recombinations on a chromosome were modeled according to a binomial map function [16].

Three scenarios for breed history were considered in this study. In the first two scenarios, the breeds were assumed to have a common origin either 50 or 550 generations ago. In the third scenario, the breeds did not have a common origin. These scenarios will be referred to as having closely related breeds, distantly related breeds, and unrelated breeds, respectively. In all cases LD was simulated by drift and mutation in two periods. In the first period of 1000 generations, random mating was simulated in an effective population of size 500. In the second period of 50 generations, random mating continued after reducing the effective population size to 100. In generation 1051 the population size was expanded to 1000 or 4000 individuals simulating more matings and seven more generations of random mating with the expanded population size were produced. Also, in generation 1051 three different commercial crossbred lines were generated with 1000 or 4000 individuals. These crossbred lines were an AxB two-breed cross, an ABxC three-breed cross, and an ABxCD four-breed cross. The crossbred lines in generation 1051 were used for "training" with phenotype and genotype data, and the purebred lines in generation 1058 for validation with only genotype data. Either 500 or 2000 segregating SNPs (minor allele frequency > 0.05) from the crossbred population were chosen for analysis. Some of these segregating SNPs in the crossbred populations were fixed in the purebred populations. Heritability of the quantitative trait was set to 0.3 by rescaling QTL effects in the training population. The method to estimate SNP effects was Bayes-B [9], which is described further in the following. The
criterion to compare models was the accuracy of estimated breeding values for the purebred validation population, calculated as the correlation between true and estimated breeding values. Each simulated data set and analysis was replicated 40 times.

**A1.3.2 Statistical Models**

The statistical models used for the analyses are described here. The across-breed SNP genotype model (ASGM) is:

\[
\gamma_i = \mu + \sum_j X_{ij}\beta_j\delta_j + e_i
\] [1]

where \(\gamma_i\) is the phenotype of \(i\), \(\mu\) is the overall mean, \(X_{ij}\) (0, 1, or 2) is the genotype of \(i\) at marker locus \(j\), \(\beta_j\) is the across-breed allele substitution effect of locus \(j\) in the training population, \(\delta_j\) is a 0/1 indicator variable that specifies if locus \(j\) is included in the model or not, and \(e_i\) is the residual of \(i\). The breed-specific SNP allele model (BSAM) is:

\[
\gamma_i = \mu + \sum_j (A_{ijk}^S \beta_{jk}^S \delta_j^S + A_{ijkl}^D \beta_{ji}^D \beta_{ji}^D) + e_i
\] [3]

where \(A_{ijk}^S\) (0,1) is the SNP allele at locus \(j\), of breed origin \(k\) that \(i\) received from its sire, \(\beta_{jk}^S\) is the breed-specific substitution effect for allele \(A_{ijk}^S\). If the sire of \(i\) is a purebred, \(k\) takes the same value for all alleles, e.g. \(k = 1\) if the sire is purebred A. On the other hand, if the sire is crossbred, AxB for example, \(k\) can take values 1 or 2, indicating whether the SNP allele received for the sire originated from breed A or B.

The variable, \(\delta_j^S\), is a 0/1 indicator that specifies if the sire allele is included in the model for locus \(j\).
Similarly, $A_{jl}^D$, $\beta_{jl}^D$, and $\beta_{jl}^D$ are defined for the SNP allele at locus $j$, of origin $l$ that $i$ received from its dam. Breed origin of alleles was assumed to be known without error in the analyses.

The Bayes-B method described by Meuwissen et al. [9] was used to estimate the across-breed additive effects in ASGM and the breed-specific additive effects in BSAM. The prior probability for a locus to be included in the model was set to 0.05, i.e., $\Pr(\delta_j = 1) = 0.05$. A previous study of prior sensitivity was performed to validate that it did not influence in the model results. For loci in the model, the locus effects were assumed to be normal with null mean and locus specific variance $\sigma^2\beta_j$ in ASGM, and locus and breed-origin specific variance $\sigma^2\beta_{Sjk}$ and $\sigma^2\beta_{Djl}$ for BSAM. Following Meuwissen et al. [9], the prior for these variance components was an inverse chi-square with 4.234 degrees of freedom and scale parameter $S = 0.0429$. The prior for the $\sigma^2e$ was an inverse chi-square distribution with four degrees of freedom and scale parameter $S = 0.4$, and a flat prior was used for $\mu$. A difference between the Bayes-B implementation of Meuwissen et al.[9] and that used here is that we fitted effects of SNP genotypes and alleles rather than of haplotypes. After some exploratory analyses, a single chain of 100,000 samples was used, with a burn-in period of 1000. Convergence was tested for all dispersion parameters separately using the Raftery and Lewis [17] method and a visual check of the chain plots.

### A1.4 RESULTS

Accuracy of prediction of breeding values in the purebred lines using ASGM and BSAM are in Tables 1, 2 and 3. Results when the AxB two-breed cross was used as training population are in Table A1.1. In this table, the accuracy of both models (ASGM and BSAM) increased with marker density and size of the training data. Accuracies of both models also
tended to decrease with increasing distance between breeds. The effect of marker density, training data size and distance between breeds, however, differed between the two models, which resulted in the model with the highest accuracy to differ between scenarios. Given the differences in marker-QTL LD, we would have expected the model that fitted breed-specific SNP allele effects (BSAM) to have greater accuracy. However, that was the case only when the number of markers was small (500), the number of records used for training was large (4000), and when breeds were distantly related or unrelated. When the number of markers was increased to 2000, ASGM gave better results when breeds were closely related, and the difference in accuracy was significant in the simulation with 1000 records. For distant or unrelated breeds, BSAM had accuracies that were equal to or better than those with ASGM.

Accuracies of the best model in the cross are, however, lower for distant and unrelated breeds. Results when the ABxC three-breed cross was used as the training population are in Table A1.2. In this scenario, 50% of the alleles in the training population are from breed C but only 25% are from either breed A or B. Thus, accuracies are given in this table for predicting breeding values of B and C purebred animals. In all cases and for both models, accuracies were lower for breed B than for breed C, as expected. Also, general trends in accuracies for a given model with changes in marker density, data size, and breed distance were similar as observed for the two-way cross in Table A1.1. The relative performance of the two models, however, differed from what was observed for the two-way cross. For the three-way cross (Table A1.2), with closely related breeds, ASGM gave better results when 1000 records were used, and with the exception of predicting purebred B animals using 500 markers, all these differences were significant. For close breeds, when 4000 records were used for training, ASGM was significantly better only for predicting purebred C animals.
using 2000 markers. For distant or unrelated breeds, ASGM was significantly better than BSAM for predicting purebred C animals using 2000 markers and 1000 records for training. When the number of records for training was increased to 4000, BSAM was significantly better for predicting purebred B animals using 500 markers in scenario 2, but ASGM was better for predicting purebred B animals using 500 markers in scenario 3 and for predicting purebred C animals using 2000 markers in scenarios 2 and 3.

Results when the ABxCD four-breed cross was used as the training population are in Table A1.3. Because the same accuracy is expected for all breeds, since all contribute 25% to the cross, only accuracy for one breed is shown. Here, BSAM was significantly better when 500 markers were used with 4000 records for training for distant or unrelated breeds. However, ASGM was significantly better when 2000 markers were used with 1000 records for training for close breeds and with 4000 records for training for unrelated breeds. Figure A1.1 shows the frequency of SNP alleles for purebreds A and B in generation 1050 for unrelated breeds. This figure shows that a large number of loci that were segregating in one of the purebred lines were fixed in the other purebred line. For these loci that are fixed in one of the purebred lines, ASGM and BSAM are equivalent. This partially explains why differences between ASGM and BSAM were small for unrelated breeds.

To further investigate the impact of the genetic difference between breeds on the accuracy of genomic selection based on crossbred data, Figure A1.2 plots the difference in average genotypic values of the two breeds against the accuracy of breeding values predicted based on their crossbred data. Each point represents one replicate for the scenario with distantly related breeds, 2000 SNPs, and 1000 records. Although in general high accuracies were obtained for genotype differences smaller than 4 sd, the small number of samples with
breed differences greater than 4 sd was not enough to disclose a clear relationship between breed difference and accuracy.

The results presented above were based on a simulated genome consisting of only 1 chromosome of 1 M. To compare these results with a more realistic situation, we simulated the scenario for closely related breeds with a genome of 10 chromosomes with a total genome size of 10 M, 60,000 SNPs and 1,000 QTL. For the statistical analysis we chose 20,000 segregating SNPs from the crossbred population. The analysis of this data showed a 25% drop in of accuracy relative to the results with 1 chromosome. However, the relationship between training in a purebred line or crossbred line did not change (Table A1.5).

A1.5 DISCUSSION

The objective of this study was to compare the accuracy of genomic selection of purebreds for commercial crossbred performance using either ASGM or BSAM. Alleles in a crossbred line originate from one of the purebred parental lines. If these purebred lines are not closely related, the effect of SNP alleles will depend on their line of origin. Thus, a model with breed-specific effects of SNP alleles (BSAM) was used to estimate the effects of alleles in purebreds for crossbred performance. These estimated effects and the SNP genotypes of purebred candidates for selection were then used to predict their breeding values for crossbred performance. The accuracy of prediction was quantified by the correlation of the predicted and true breeding values. This accuracy was compared to that obtained using the classical model with across breed effects of SNP genotypes (ASGM).

Due to the genetic differences among the pure lines, BSAM with breed-specific effects of SNP alleles was expected to perform better. Contrary to expectation, however, accuracy of prediction with ASGM often was equal to or higher than with BSAM. In addition to the
relationship between the purebred parental lines, there are two other factors that contribute to the
difference in accuracy of prediction using ASGM and BSAM in our simulations. Marker
density is one of these, and the other is the number of records used in training. Marker
density affects the difference between ASGM and BSAM in two ways. The first is that as
marker density increases the model will include markers that are closer to the QTL. In a
finite population, marker alleles that are closer to the QTL will more accurately reflect the
state of the QTL alleles. Thus, as the marker density increases the need for BSAM is
reduced. The second is that BSAM has, relative to ASGM, twice as many effects that need to be estimated in a two-breed cross, three times as many in a three-breed cross, and four times as many in a four-breed cross. Thus, due to the greater number of effects that need to be estimated, BSAM is at a disadvantage over ASGM, and this disadvantage increases with marker density. On the other hand, as the number of records used for training increases more information becomes available to estimate the effects of markers and, given sufficient records for training, even small differences in breeds will make BSAM advantageous. So, BSAM will give better results only when breed differences are big enough to compensate for the additional breed-specific effects in the model, given the number of records used for training. Note that in the absence of epistasis, there are no breed differences for effects at the QTL. Thus, as the marker density increases, breed differences of markers effects decreases while the number of extra parameters in BSAM increases.

In Table A1.1, BSAM had greater accuracy when 500 markers were used, but when the number of markers was increased to 2000, this advantage disappeared except when breeds were unrelated and, thus, breed differences were greatest. The effect of increasing the number of records used for training can be seen from Tables A1.1, A1.2, A1.3, where given
the same number of markers, increasing the number of records used tended to favor BSAM. In Table A1.2, for example, the difference in accuracy between ASGM and BSAM was not significant with 500 markers for distantly related breeds when 1000 records were used for training, but when the number of records for training was increased to 4000, BSAM was significantly more accurate, with the difference in accuracies between ASGM and BSAM changing from 0.02 to -0.11. Our results include several such examples where increasing the number of records favors BSAM (Tables 1, 2, 3), but none that goes in the opposite direction. This demonstrates that BSAM will have an advantage provided sufficient information is available for estimating the additional breed-specific effects.

In livestock, production animals often are either from a three-breed or four-breed cross. When an ABxC three-breed cross was used for training, the accuracy of prediction of purebred C animals was about the same as the accuracy of prediction of purebred B animals with training in an AxB two-breed cross. This is because 50% of the alleles in the ABxC cross are from purebred line C. On the other hand, only 25% of the alleles in ABxC are from purebred line B. Thus, the accuracy of prediction for line B animals was significantly lower. The same was true in a four-breed cross, where only 25% of the alleles in the crossbreds are from any particular parental line. Thus, the accuracy of prediction of purebred B animals with training in an ABxCD cross was similar in accuracy to that for purebred B animals with training in an ABxC cross (Tables 2 and 3). It is interesting that the accuracy of prediction with training in a four-breed cross using 4000 records was about the same as that with training in a purebred line with 1000 records (Tables 3 and 4).

The results in Table A1.5 show that, given the same number of records used for training, when marker effects from 10 chromosomes were included in the model, the accuracy of
prediction dropped. Table A1.1 showed that when the model included 2000 markers from one chromosome, ASMG was significantly more accurate than BSAM. When the model includes 20,000 markers from 10 chromosomes, the difference in accuracy became smaller but remained significant (Table A1.5).

Dominance and epistatic effects were not considered in the present study. However, the genomic selection methods for crossbred performance do not require absence of non-additive effects. If non-additive effects are present, the marker effects estimated by the genomic selection methods are allele substitution effects, which incorporate the additive components of dominance and epistatic effects [18]. Thus, by estimating allele substitution effects based on crossbred phenotypes, the effects of purebred alleles will be estimated against the genetic background that they will be expressed in. Thus, genomic selection on SNP effects estimated on crossbred data is equivalent to practicing reciprocal recurrent selection.

The simulation model also assumed absence of genotype by environment interactions. Such interactions could, however, be present when comparing performance in nucleus and field environments and contribute to the low genetic correlations between purebred and crossbred performance that have been estimated in literature. However, similar to non-additive effects, allele substitution effects estimated based on phenotypes collected in the field would allow the effects of purebred alleles to be estimated under the environment in which they will be expressed.

Although genomic selection models accommodate non-additive effects to the extent that they are captured by allele substitution effects, presence of non-additive effects can reduce the accuracy of GEBV compared to those obtained here, and also affect the comparison between the ASGM and BSAM models. The reason is that non-additive effects will increase
differences in breed-specific allele substitution effects because breeds are expected to differ in allele frequencies at QTL. Specifically, with dominance, the QTL allele substitution effect for breed A on performance of AxB crossbreds is equal to \( a + d(1-2p_B) \), where \( p_B \) is the QTL allele frequency in breed B and \( a \) and \( d \) are the additive and dominance effects at the QTL [19]. Thus, if breeds that are being crossed have different QTL allele frequencies, they will have different allele substitution effects at the QTL and, therefore also at markers that are in LD with the QTL. Epistatic effects also contribute to allele substitution effects, depending on allele frequencies. Thus, if epistatic effects are present, allele substitution effects will further differ between breeds. These additional differences in breed-specific SNP effects compared to what was simulated here will likely increase the accuracy of the BSAM model that includes breed-specific allele effects compared to the ASGM model. The accuracy of the ASGM model will likely decrease slightly, as the average allele effects across breeds will tend to be reduced when differences in breed-specific allele effects are greater. Further work is needed to investigate these scenarios. Presence of genotype by environment interactions for the nucleus versus field environment are not expected to affect the accuracy of either the ASGM or BSAM model because allele effects are evaluated in the target environment for both models.

In this study, divergence between breeds was created by drift only. In practice, in addition to drift, breeds will have diverged as a result of different selection pressures imposed upon them through either artificial or natural selection. The potential impact of differential artificial selection on the trait being evaluated is indirectly evaluated in Figure A1.2 by considering breed pairs that have drifted apart to differing degrees for average genotypic values for the trait. As shown in the Figure, this did not have a discernible effect
on the accuracy of genomic selection. The same is expected to hold for breeds that have been
differentially selected for other characteristics.

Results from this study show the potential for genomic selection of purebreds for
commercial crossbred performance. This would enable genetic improvement of purebreds for
performance of their crossbred descendents in the field, without the need to track pedigrees
through the system. Further, these results indicate that a model with breed-specific effects of
alleles may not be necessary, especially when the marker density is high. It is obvious that
ASGM would be better when breeds are not very different. However, in some cases ASGM
was significantly better even when the breeds did not have any common origin (Table A1.3).
The reason for this can be seen from figure A1.1. There are three types of loci in this figure:
1) those that are segregating in both lines, 2) those that are segregating only in one line, and
those that are fixed in both lines. Loci of the first type would favor BSAM, those of the
second type would contribute equally to both models, and those of the third type would not
contribute to either. Crosses of highly inbred lines that were separated in the distant past will
have only a few loci of the first type and thus, would not favor BSAM over ASGM. So, even
in this extreme case, ASGM can do well. Using ASGM has the advantage that it does not
require tracing alleles from crossbreds in the field to their purebred ancestors in nucleus
lines. In this study, we assumed that alleles could be traced from the crossbreds to the
purebred parents without error. Given very high density marker information, it may be
possible to trace alleles to ancestors very accurately [20], but some errors may be inevitable.
Thus, in practice, ASGM may even perform relatively better than in this study.
Authors' contributions

NIE participated in the design of the study, carried out the simulation studies, performed the statistical analyses, and drafted the manuscript. AT participated in the design of the study and helped with the simulation studies. RLF and JCMD conceived of the study, oversaw its design and execution, and helped to revise and finalize the manuscript. RLF also assisted with development of the simulation and analysis programs. All authors read and approved the final manuscript.

A1.6 ACKNOWLEDGEMENTS

Financial support from Spain's Ministerio de Educacion y Ciencia (Programa movilidad Jose Castillejo) for NEI, and from Newsham Choice Genetics for AT is gratefully acknowledge. RLF and JCMD are supported by the United States Department of Agriculture, National Research Initiative grant USDA-NRI-2007-35205-17862 and by Hatch and State of Iowa funds through the Iowa Agricultural and Home Economics Experiment Station, Ames, IA.

A1.7 LITERATURE CITED

9. Meuwissen T, Hayes B, Goddard M: Prediction of Total Genetic Value Using Genome-
163(2):789-801.
13. Habier D, Fernando R, Dekkers J: The Impact of Genetic Relationship Information on
15. Toosi A, Fernando R, Dekkers J, RL Q: Genomic selection of purebreds using data from
admixed populations. ASAS/ADSA annual meeting, 407, Indianapolis 2008.
17. Raftery A, Lewis S: The number of iterations, convergence diagnostics and generic
18. Falconer DS, Mackay TFC: Values and Means. In An Introduction to Quantitative
19. Dekkers JCM, Chakraborty R: Optimizing purebred selection for crossbred performance
markers to map quantitative trait loci and estimate effective population size. Genetics 2007,
176(4):2551-60.
A1.8 FIGURES

Figure A.1.1 Frequency of SNP alleles for purebreds A and B in generation 1050 for unrelated breeds
Figure A1.2 Difference in average genotypic values of two breeds against the accuracy of breeding values predicted based on their crossbred data. Each point represents one replicate for the scenario with distantly related breeds, 2000 SNPs, and 1000 records.
### Table A1.1

Accuracy (se) of breeding values in pure breed predicted based on two-breed cross data using ASGM or BSAM for three different scenarios (40 replicates)

<table>
<thead>
<tr>
<th>Markers</th>
<th>VP</th>
<th>ASGM</th>
<th>BSAM</th>
<th>Diff $^a$</th>
<th>ASGM</th>
<th>BSAM</th>
<th>Diff $^a$</th>
<th>ASGM</th>
<th>BSAM</th>
<th>Diff $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 records</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>B</td>
<td>0.78</td>
<td>0.79</td>
<td>-0.01</td>
<td>0.72</td>
<td>0.76</td>
<td>-0.04</td>
<td>0.72</td>
<td>0.73</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>2000</td>
<td>B</td>
<td>0.87</td>
<td>0.81</td>
<td>0.06</td>
<td>0.81</td>
<td>0.81</td>
<td>0.00</td>
<td>0.80</td>
<td>0.81</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>4000 records</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>B</td>
<td>0.83</td>
<td>0.85</td>
<td>-0.02</td>
<td>0.78</td>
<td>0.82</td>
<td>-0.04</td>
<td>0.77</td>
<td>0.80</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>2000</td>
<td>B</td>
<td>0.92</td>
<td>0.91</td>
<td>0.01</td>
<td>0.91</td>
<td>0.91</td>
<td>0.01</td>
<td>0.88</td>
<td>0.91</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>

$^a$ Pure breed used as validation population.

$^b$ Difference (se) of accuracy between ASGM and BSAM.
### Table A1.2 Accuracy (se) of breeding values in pure breed predicted based on three-breed cross data using ASGM or BSAM for three different scenarios (40 replicates)

<table>
<thead>
<tr>
<th>Markers</th>
<th>VP</th>
<th>ASGM</th>
<th>BSAM</th>
<th>Diff↑</th>
<th>ASGM</th>
<th>BSAM</th>
<th>Diff↑</th>
<th>ASGM</th>
<th>BSAM</th>
<th>Diff↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 records</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>B</td>
<td>0.68</td>
<td>0.63</td>
<td>0.05</td>
<td>0.57</td>
<td>0.59</td>
<td>-0.02</td>
<td>0.44</td>
<td>0.42</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.03)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.79</td>
<td>0.74</td>
<td>0.05</td>
<td>0.64</td>
<td>0.63</td>
<td>0.01</td>
<td>0.56</td>
<td>0.57</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>2000</td>
<td>B</td>
<td>0.82</td>
<td>0.74</td>
<td>0.08</td>
<td>0.66</td>
<td>0.63</td>
<td>0.04</td>
<td>0.63</td>
<td>0.63</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.85</td>
<td>0.73</td>
<td>0.11</td>
<td>0.77</td>
<td>0.68</td>
<td>0.09</td>
<td>0.71</td>
<td>0.67</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>4000 records</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>B</td>
<td>0.79</td>
<td>0.81</td>
<td>-0.02</td>
<td>0.68</td>
<td>0.75</td>
<td>-0.07</td>
<td>0.63</td>
<td>0.71</td>
<td>-0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.06)</td>
<td>(0.03)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.82</td>
<td>0.79</td>
<td>0.02</td>
<td>0.74</td>
<td>0.74</td>
<td>0.00</td>
<td>0.76</td>
<td>0.77</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>2000</td>
<td>B</td>
<td>0.87</td>
<td>0.86</td>
<td>0.01</td>
<td>0.85</td>
<td>0.87</td>
<td>-0.02</td>
<td>0.79</td>
<td>0.67</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.04)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.02)</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.92</td>
<td>0.86</td>
<td>0.06</td>
<td>0.83</td>
<td>0.80</td>
<td>0.03</td>
<td>0.79</td>
<td>0.72</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.05)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

* Pure breed used as validation population.

* Difference (se) of accuracy between ASGM and BSAM.
Table A1.3  Accuracy (se) of breeding values in pure breed predicted based on four-breed cross data using ASGM or BSAM for three different scenarios (40 replicates)

| Marker | VP | 1000 records |  |  |  |  |  |  |  |  |  |
|--------|----|--------------|----|-------------|----|-------------|----|-------------|----|-------------|----|-------------|
|        |    | ASGM | BSAM | Diff<sup>a</sup> | ASGM | BSAM | Diff<sup>a</sup> | ASGM | BSAM | Diff<sup>a</sup> | ASGM | BSAM | Diff<sup>a</sup> |
| 500    | B  | 0.65 | 0.60 | 0.05 | 0.46 | 0.48 | 0.02 | 0.46 | 0.50 | -0.03 | (0.03) | (0.03) | (0.03) |
| 2000   | B  | 0.84 | 0.75 | 0.09 | 0.62 | 0.59 | 0.04 | 0.52 | 0.54 | -0.02 | (0.02) | (0.02) | (0.02) |

<table>
<thead>
<tr>
<th>Marker</th>
<th>VP</th>
<th>4000 records</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ASGM</td>
<td>BSAM</td>
<td>Diff&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ASGM</td>
<td>BSAM</td>
<td>Diff&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ASGM</td>
<td>BSAM</td>
<td>Diff&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ASGM</td>
<td>BSAM</td>
</tr>
<tr>
<td>500</td>
<td>B</td>
<td>0.78</td>
<td>0.80</td>
<td>-0.02</td>
<td>0.62</td>
<td>0.72</td>
<td>-0.11</td>
<td>0.55</td>
<td>0.70</td>
<td>-0.14</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>2000</td>
<td>B</td>
<td>0.87</td>
<td>0.85</td>
<td>0.01</td>
<td>0.85</td>
<td>0.86</td>
<td>-0.01</td>
<td>0.72</td>
<td>0.62</td>
<td>0.10</td>
<td>(0.01)</td>
<td>(0.04)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Pure breed used as validation population.
<sup>b</sup> Difference (se) of accuracy between ASGM and BSAM.
Table A1.4 Accuracy of breeding values in pure breed predicted based on performance in the same pure breed using ASGM (40 replicates)

<table>
<thead>
<tr>
<th>Marker</th>
<th>% PB&lt;sup&gt;a&lt;/sup&gt;</th>
<th>1000 records</th>
<th>4000 records</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>100%</td>
<td>0.79 (0.02)</td>
<td>0.83 (0.03)</td>
</tr>
<tr>
<td>2000</td>
<td>100%</td>
<td>0.91 (0.01)</td>
<td>0.94 (0.01)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Percentage in the training population of the breed evaluated.
Table A1.5 Accuracy of breeding values in pure breed predicted based on crossbred data when the breeds are closely related for a simulated genome of 10 chromosomes of 1 M each (40 replicates)

<table>
<thead>
<tr>
<th>1000 records</th>
<th>Training population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Two-breed cross (AxB)</td>
</tr>
<tr>
<td>Marker</td>
<td>VP²</td>
</tr>
<tr>
<td>20000</td>
<td>B</td>
</tr>
</tbody>
</table>

² Validation population.
APPENDIX 2. APPLICATION OF WHOLE-GENOME PREDICTION METHODS FOR GENOME-WIDE ASSOCIATION STUDIES: A BAYESIAN APPROACH

A paper submitted to Animal Genetics

Rohan Fernando, Ali Toosi, Dorian Garrick, Jack Dekkers
Department of Animal Science and Center for Integrated Animal Genomics,
Iowa State University, Ames, Iowa, 50011-3150, USA

A2.1 SUMMARY

Data that are collected for whole-genome prediction can also be used for genome-wide association studies (GWAS). This paper discusses how Bayesian multiple-regression methods that are used for whole-genome prediction can be adapted for GWAS. It is argued here that controlling the posterior type I error rate (PER) is more suitable than controlling the genome-wise error rate (GER) for controlling false positives in GWAS. It is shown here that under ideal conditions, PER can be controlled by using Bayesian posterior probabilities that are easy to obtain. Computer simulation was used to examine the properties of this Bayesian approach when the ideal conditions were not met. Results indicate that even then useful inferences can be made.
A2.2 INTRODUCTION

High-density SNP genotypes are currently being used in livestock for whole-genome prediction (VanRaden et al., 2009; Hayes et al., 2009; Habier et al., 2010b; Wolc et al., 2011). This requires obtaining genotypes and phenotypes on several thousand animals in a training population to estimate effects of the SNP genotypes on the traits of interest. These estimated SNP effects are then used to predict the breeding values of selection candidates that may not have any phenotypes recorded but have been genotyped (Meuwissen et al., 2001). The genotype and phenotype data obtained for whole-genome prediction can also be used for genome-wide association studies (GWAS) to locate causal variants (QTL) for traits of economic importance.

Many GWAS for quantitative traits are based on testing one SNP at a time using simple regression models or using a mixed model with a fixed substitution effect of the SNP genotype included, along with a polygenic effect correlated according to pedigree relationships to capture the effects of all the other genes. Such GWAS have been successful in detecting many associations, but the established associations typically explain only a small fraction of the genetic variability of quantitative traits (Maher, 2008; Manolio et al., 2009; Visscher et al., 2010). On the other hand, when whole-genome selection models that simultaneously fit all SNPs as random effects are used, the SNPs jointly explain a large proportion of the genetic variance (Onteru et al., 2011; Hayes et al., 2010; Fan et al., 2011).

In these analyses, however, any given SNP may have only a weak association even with a close QTL. The reason for this is that in multiple-regression, the association of a SNP with the phenotype is a partial association conditional on all the other SNPs. Further, in a high-density SNP panel, many SNP genotypes within a narrow genomic region are expected to be
highly correlated with each other and with any QTL that are close to them. So, any one SNP may contribute only a little more to explain the variability of the QTL in addition to the other SNPs in the neighborhood. On the other hand, even if each SNP in a neighborhood is only weakly associated with a QTL, the SNPs in the neighborhood may jointly explain much more of the variability of a QTL than any SNP by itself. Therefore in multiple-regression models, SNPs in a genomic window should be used to locate QTL (Onteru et al., 2011; Sahana et al., 2010; Hayes et al., 2010; Fan et al., 2011). Inferences on genomic windows by frequentist methods, however, are computationally very challenging because they require repeated analyses of the data with bootstrap or permuted samples to obtain significance levels for tests (Onteru et al., 2011; Hayes et al., 2010; Fan et al., 2011). It can be shown that Bayesian posterior probabilities obtained from a single analysis can be used to make inferences on genomic windows. This approach to inference is related to the frequentist approach of controlling the posterior type I error rate (PER), which is the conditional probability of a false positive (type I error) given a positive (significant) result (Morton, 1955). It is PER that traditionally has been used in human genetics to control false positives in linkage analyses of monogenic traits (Elston, 1997) rather than the usual type I error rate, which is the conditional probability of a false positive result given that the null hypothesis is true.

A requirement for controlling PER is knowledge of the distribution of the test statistic under the null hypothesis of no association, which is also required to control the usual type I error rate. In addition to this requirement, controlling PER requires knowing the proportion $\pi$ of SNPs for which the null hypothesis is true and the average power of the test, which is the average probability of rejecting the null hypothesis when it is not true. These quantities are
almost never known in a GWAS of a quantitative trait, and thus PER cannot be controlled in the sense that the usual type I error rate can be controlled (Elston, 1997).

In a QTL mapping study, Soller and colleagues were the first to show that \( \pi \) can be estimated from a histogram of p-values (Mosig et al., 2001). In this multiple test setting, they showed how the estimate of \( \pi \) can be used to obtain an adjusted false discovery rate (FDR) (Benjamini & Hochberg, 1995) that results in increased power to detect QTL. Their seminal paper also showed how to estimate the average power of tests (Mosig et al., 2001). Such estimates of \( \pi \) and average power can also be used to estimate PER for a test with a specified significance level (type I error rate). However, when the resulting estimates of PER are used for inference, the estimates of \( \pi \) and average power are treated as known values because it is not straightforward to incorporate the errors in these estimates into the calculations. In contrast to frequentist methods, when the Bayesian multiple-regression models (Meuwissen et al., 2001) that are used for whole-genome prediction are applied to GWAS, posterior probabilities that are similar to PER can be obtained even without the requirement of knowing the null distribution of any test statistic. Further, in Bayesian analyses, \( \pi \) and the partial regression coefficients of markers, which determine average power, can be formally treated as unknowns such that their uncertainty is incorporated in the inference.

It can be shown that using PER or related posterior probabilities for inference has the following advantage. Suppose PER is controlled, for example, at 0.05 for each of many tests, which may be dependent. Then, among the positive results that accumulate over many independent experiments the proportion of false positives will converge to 0.05 (Section 2.2). As a result, regardless of the number of tests in an experiment and their interdependence, the proportion of false positives that accumulate in the literature can be controlled by controlling
the PER for individual tests. It has been shown that controlling PER for individual tests in an experiment is related to controlling FDR or its close relatives pFDR (Storey, 2002) or PFP (Southey & Fernando, 1998; Fernando et al., 2004) for the collection of tests from the experiment.

Posterior probabilities that are related to PER are still not widely used for inference in GWAS by animal geneticists. One reason for this is that to manage false positives in a GWAS controlling the genome-wise error rate (GER) is better accepted than controlling the PER for individual tests or FDR for the experiment. A second reason is that the relationship between PER and the Bayesian posterior probabilities for inference in GWAS has not been studied in a realistic setting. A third reason for preferring GER over PER is given by Chen & Storey (2006). Their reasoning can be interpreted as follows. In linkage analysis, a QTL is in principle linked to every marker on the same chromosome, and this makes the null hypothesis that a marker is not linked to a QTL false for every marker on a chromosome that has even a single QTL. Then, a marker with a significant linkage signal on a chromosome that has a QTL will be a true positive even if the marker is not close to the QTL. This makes controlling the proportion of false positives among significant results meaningless (Chen & Storey, 2006). They refer to this problem as signal dependence. Thus, the objective of this paper are to: 1) address the problem of signal dependence and review the advantages of managing false positives in GWAS by controlling PER or related measures such as FDR or PFP rather than by controlling the GER, 2) discuss the theoretical relationship between Bayesian posterior probabilities and PER, and 3) use computer simulation to examine this relationship in a realistic GWAS setting.
In Section 2.1 the problem of signal dependence are addressed and the advantages of managing false positives by controlling PER rather than GER are reviewed. Section 2.2 provides the justification for using PER in a multiple test setting and gives its relationship to FDR and related measures. Section 2.3 describes the calculation of Bayesian posterior probabilities that are useful for inference in GWAS, and Section 2.4 shows how these posterior probabilities are related to PER. A computer simulation that was used to study the properties of the Bayesian approach is described in Section 2.5. Results are discussed and summarized in Section 3.

A2.3 METHODS

A2.3.1 Controlling False Positives

Many papers have discussed the advantages of controlling FDR and related measures rather than controlling GER (Benjamini & Hochberg, 1995; Storey, 2002; Fernando et al., 2004; Stephens & Balding, 2009). Chen & Storey (2006), however, have argued that measures such as FDR that control the proportion of false positives among significant results are not suitable for controlling false positives in QTL studies. The essence of their argument stems from the implicit assumption that the linkage signal from a QTL spans the whole chromosome. Thus, the null hypothesis that a marker is not linked to a QTL is equivalent to that of no QTL on the same chromosome as the marker. Then, rejection of the null hypothesis does not imply that the marker is close to a QTL. To address this problem others have employed multiple-regression models where the signal from a QTL is localized to a relatively narrow segment of the chromosome (Zeng, 1993, 1994; Southey & Fernando, 1998).
Suppose data are available from a backcross population derived from two completely homozygous lines. In such a population, the following multiple-regression model can be used to test the null hypothesis that a chromosomal segment with a central marker $C$ bracketed by two markers $L$ on the left and $R$ on the right does not contain any QTL:

\[ Y_i = \beta_0 + X_{Li} \beta_L + X_{Ci} \beta_C + X_{Ri} \beta_R + \epsilon_i, \]

where $Y_i$ is the trait value, $\beta_0$ is the intercept, $\beta_L$, $\beta_C$, and $\beta_R$ are regression coefficients for three markers $L$, $C$ and $R$, respectively, and $X_{Li}$, $X_{Ci}$ and $X_{Ri}$ are 0 or 1 depending on whether the marker genotype is heterozygous or homozygous. It can be shown that $\beta_C$ is non-null if only if the chromosomal segment bracketed by $L$ and $R$ contains a QTL (Zeng, 1993). Thus with this multiple-regression model, the signal from the trait gene is localized to a chromosomal segment of, for example, 10 cM in length, and testing the null hypothesis that such a segment does not contain a QTL is meaningful in a QTL study. In this setting, we will consider the difference between controlling GER and PER.

A hypothetical example is used next to show why controlling PER is more meaningful than controlling GER in a QTL study. Suppose only one chromosomal segment is tested. Then, GER is identical to the type I error rate for the single test, but PER can be much higher, depending on the probability $\pi$ that the null hypothesis is true for a randomly chosen segment and on the power of the test. To calculate PER, assume the trait is controlled by a single gene, the genome consists of 30 chromosomes each of length 1 Morgan, and the segment length is 10 cM. Then, the probability the null hypothesis is true that the chromosomal segment does not contain a QTL for a randomly chosen segment is $\pi = \frac{299}{300}$. 
Further, we assume that if the null hypothesis is not true, the power of the test is \( p = 90\% \) when a significance level of \( \alpha = 0.05 \) is used. Now, with GER at the usual 0.05, PER is

\[
\text{PER} = \frac{\alpha \pi}{\alpha \pi + p(1 - \pi)}
\]

\[
= \frac{0.05 \times \frac{299}{300}}{0.05 \times \frac{299}{300} + 0.9 \times \frac{1}{300}} \quad [2]
\]

\[
= 0.94,
\]

where the numerator of (2) is the joint probability that the null hypothesis is true and it is rejected, and the denominator is the marginal probability of rejecting the null hypothesis. Thus, although we have controlled GER at 0.05, a significant result will be a false positive with probability of over 90%! In other words, among significant results over 90% will be false positives. If we want PER to be 0.05, \( \alpha \) must be reduced to 0.0001584 even for a single test. Thus, controlling GER does little to control PER.

Suppose now a chromosomal segment of 10 cM is chosen at random from each of the 30 chromosomes. For each of these segments, the probability \( \pi \) that the null hypothesis is true remains \( \frac{299}{300} \). So, if the significance level is kept at \( \alpha = 0.0001584 \), the PER will be 0.05 for the results from each chromosome. In other words, if results are accumulated over repetitions of the experiment, the proportion of false positives among the significant results would remain 0.05. Thus, the proportion of false positives in the collection of significant results from all chromosomes will also be 0.05. A proof of this result is given in the next section.

Given that the 30 tests are independent, the GER can be calculated (Sidak, 1967) as

\[
\text{GER} = 1 - (1 - 0.0001584)^{30}
\]

\[
= 0.004742. \quad [3]
\]
Note that as the number of tests increased GER also increased from 0.0001584 for one test to 0.004742 for 30 tests while PER stayed constant at 0.05. So, if GER is used, as the number of tests increases, the significance level for individual tests has to be decreased to keep GER at 0.05, and this will result in lower power. This is not the case with PER. Thus, it is more suitable for GWAS that involve very large numbers of tests.

A2.3.2 Relationship of PER to FDR, pFDR and PFP

In this section we provide the justification for using PER to control false positives in a multiple test setting where the tests may be dependent such as in a GWAS. Furthermore, we will show how controlling PER for individual tests is related to measures such as FDR that attempt to control false positives in the collection of tests from an experiment.

In order to describe this relationship between PER that is defined for an individual test and measures such as FDR, pFDR and PFP that are defined for a collection of tests, for test $j$ of experiment $i$, $V_{ij}$ and $R_{ij}$ are defined as follows: $V_{ij} = 1$ if $H_0$ is falsely rejected (false positive) and else $V_{ij} = 0$; $R_{ij} = 1$ if $H_0$ is rejected (positive) and else $R_{ij} = 0$. Now using this notation, PER for test $j$ in experiment $i$ is

$$\text{PER} = \Pr(V_{ij} = 1|R_{ij} = 1),$$

which according the frequentist definition of probability is the limiting value of the relative frequency:

$$\frac{\sum_i^N V_{ij}}{\sum_i^N R_{ij}}$$

as $N$ the number of independent experiments tends to infinity. Thus, controlling PER for a test is equivalent to controlling the proportion of false positive results that accumulate in...
the literature for that test. Now, the relationship of PER with PFP is used to justify its use to control false positives in a multiple test setting.

In Fernando et al. (2004), PFP was defined as

$$\text{PFP} = \frac{E(V_i)}{E(R_i)}$$

where $V_i = \sum_j V_{ij}$ and $R_i = \sum_j R_{ij}$. Let $S_{ij} = 1$ denote that $H_0$ is true and $S_{ij} = 0$ that it is false for test $j$ in experiment $i$. When $S_{ij}$ is considered fixed, they showed that PER for a randomly chosen test is equal to PFP for the collection of tests that may be dependent (Fernando et al., 2004). Further, they showed that the limiting value of the frequency of false positives relative to the total number of positives across tests, which may be dependent, and across $N$ independent experiments:

$$\frac{\sum_i \sum_j V_{ij}}{\sum_i \sum_j R_{ij}}$$

is equal to PFP as $N$ tends to infinity [property 2 of PFP in Fernando et al. (2004)]. Thus, in this case, PER for a random test is equal to the limiting value of (6).

In Bayesian analyses, $S_{ij}$ is considered random, and for this case we show here that if $\text{PER} \leq (1 - t)$ for each test, then for the collection of tests that may be dependent $\text{PFP} \leq (1 - t)$. First note that PER can be written as

$$\text{PER} = \text{Pr}(V_{ij} = 1 | R_{ij} = 1)$$

$$= \frac{\text{Pr}(V_{ij} = 1)}{\text{Pr}(R_{ij} = 1)}$$

$$= \frac{E(V_{ij})}{E(R_{ij})}$$

[7]
Now, if \( \text{PER} \leq (1 - t) \) for each test, using (7), PFP for the collection of tests from experiment \( i \) becomes

\[
PFP = \frac{\mathbb{E}(\sum_j V_{ij})}{\mathbb{E}(\sum_j R_{ij})}
\]

\[
= \frac{\sum_j \mathbb{E}(V_{ij})}{\sum_j \mathbb{E}(R_{ij})}
\]

\[
= \frac{\sum_j \mathbb{E}(R_{ij})\mathbb{E}(V_{ij})/\mathbb{E}(R_{ij})}{\sum_j \mathbb{E}(V_{ij})}
\]

\[
\leq \frac{\sum_j \mathbb{E}(R_{ij})(1 - t)}{\sum_j \mathbb{E}(V_{ij})}
\]

\[
= (1 - t), \quad [8]
\]

where in going from line three to four of (8) we have used the result (7) that \( \mathbb{E}(V_{ij})/\mathbb{E}(R_{ij}) = \text{PER} \) and the condition that \( \text{PER} \leq (1 - t) \). This proves that if \( \text{PER} \leq (1 - t) \) for each test then \( \text{PFP} \leq (1 - t) \), for the collection of tests. Combining this result with property 2 of PFP from Fernando et al. (2004), we have that if \( \text{PER} \leq (1 - t) \) for each test, the limiting value of (6) \( \leq (1 - t) \).

The above result justifies using PER to control false positives in a multiple test setting where the tests may be dependent as in a GWAS, because as we have shown above, if \( \text{PER} \leq (1 - t) \) for each test in an experiment, then as results accumulate over \( N \) independent experiments, among all positive results the proportion of false positives (6) will converge to a value that is \( \leq (1 - t) \).

When tests are independent or have a certain type of dependence, Storey (2003) has shown that pFDR, which is defined as
\[ pFDR = E \left( \frac{V_i}{R_i} \middle| R_i > 0 \right), \]  

is equal to PFP and to PER for an individual test. FDR (Benjamini & Hochberg, 1995) on the other hand, defined as

\[ pFDR = E \left( \frac{V_i}{R_i} \middle| R_i > 0 \right) \Pr (R_i > 0), \]  

can be quite different from pFDR, PFP or PER even when the tests are independent if Pr \((R_i > 0)\) is not close to one (Zaykin et al., 2000; Storey, 2003). Further, both pFDR and FDR can be quite different from PFP when tests are not independent as demonstrated by a hypothetical example in Fernando et al. (2004).

**A2.3.3 Bayesian Multiple Regression**

Here we will explain how multiple-regression models that were developed for whole-genome prediction can be used for GWAS. In these models, inference is based on posterior probabilities that, as will be shown in Section 2.4, are similar to PER.

Following Meuwissen et al. (2001), consider the mixed linear model

\[ y = X\beta + \sum_{j=1}^{k} z_j \alpha_j + e, \]  

where \( y \) is the vector of trait phenotypes, \( X \) is an incidence matrix relating the vector of non-genetic, fixed effects \( \beta \) to \( y \), \( z_j \) is a vector of genotype covariates (coded as 0, 1 or 2) for SNP \( j \), \( \alpha_j \) is the random, partial regression coefficient for SNP \( j \), and \( e \) is a vector of residuals. In this model, the fixed effects are assumed to have a constant prior, the \( \alpha_j \) are a priori assumed independently distributed as

\[ \alpha_j | \pi, \sigma_{\alpha_j}^2 = \begin{cases} 0, & \text{with probability } \pi \\ \sim N \left( 0, \sigma_{\alpha_j}^2 \right), & \text{with probability } (1 - \pi) \end{cases}, \]
where the $\sigma^2_{a_j}$ are a priori assumed independently and identically distributed (iid) scaled inverse chi-square variables with scale $S^2_\alpha$ and degrees of freedom $\nu_\alpha$. The residuals are assumed iid normal with null mean and variance $\sigma^2_e$, with a scaled inverse chi-square prior for $\sigma^2_e$ with scale $S^2_e$ and degrees of freedom $\nu_e$. Inferences on the unknowns in the model are made from their marginal posterior distributions, using Markov Chain Monte-Carlo (MCMC) methods (Meuwissen et al., 2001; Habier et al., 2010a).

Although this model was first proposed for whole-genome prediction (Meuwissen et al., 2001), it can also be used to locate genomic regions that contain QTL (Yi et al., 2003; Sun et al., 2011; Fan et al., 2011). Consider a model where $\pi$ is close to one, i.e., a model where most regions of the genome do not have markers that are associated with the trait. This is the model called BayesB by Meuwissen et al. (2001). Given such a model, the posterior probability that $\alpha_j$ is non-zero for at least one SNP $j$ in a window or region can be used to make inferences on the presence of QTL in that region. We will refer to this probability as the window posterior probability of association (WPPA). The underlying assumption here is that if a genomic window contains a QTL, one or more SNPs in that window will have non-zero $\alpha_j$. Thus, WPPA, which is estimated by counting the number of MCMC samples in which $\alpha_j$ is non-zero for at least one SNP $j$ in the window, can be used as a proxy for the posterior probability that the genomic window contains a QTL.

It is possible, however, that SNPs in a window that does not contain any QTL are in association with a QTL outside the window, which is what has been called signal dependence (Chen & Storey, 2006). Fortunately, the linkage signal from LD extends over only short distances compared to that from cosegregation, and as described below, when all SNPs are
fitted simultaneously, signal dependence is further reduced. Let $W_C$ denote the window for which WPPA is estimated. Let $W_L$ and $W_R$ be windows of length $k$ cM to the left and right of $W_C$ as illustrated in Figure A2.1. A high window WPPA for $W_C$ is taken as evidence of a QTL in the “composite” window $W$ comprising of $W_L$, $W_C$, and $W_R$. Because WPPA for $W_C$ is a partial association conditional on all other SNPs in the model, including those in the flanking windows $W_L$ and $W_R$, the influence of QTL from outside the composite window on the WPPA signal for $W_C$ will be inversely related to the length $k$ of the flanking windows. In other words, as the number of markers between a QTL and $W_C$ increases, the influence of the QTL on the WPPA signal for $W_C$ will decrease. Thus, as shown in more detail in the next section, WPPA computed for $W_C$ can be used to locate QTL.

A2.3.4 Relationships of Posterior Probabilities to PER

Here, we show that Bayesian posterior probabilities such as WPPA are related to PER, and therefore they can be used to control false positives in a multiple test setting. The relationship of WPPA to PER is most straightforward when the priors used in the Bayesian analysis have a frequentist interpretation. This would be the case in a simulation study where data are generated and analyzed as follows. Suppose SNP genotypes are available on $n$ individuals at $K$ markers. Then, phenotypes for these $n$ individuals can be generated according to model (11) by taking the matrix $X$ to be a vector of ones and the vector $\beta$ to be any constant (for example 0), sampling the partial regression coefficients, $\alpha_j$, according to (12), and sampling the residuals from a normal distribution with null mean and variance $\sigma^2_e$. Now, in the Bayesian analysis of the simulated data, suppose the distributions used to simulate the partial regression coefficients and the residuals are used as priors. Then, the
WPPA calculated in the analysis for a genomic window, $W_C$, is the conditional probability of a true association (i.e., a QTL) within that window (in this case, where the QTL are included in the marker panel, the QTL signal should stay within its own window) given the data. The frequentist interpretation of this probability is as follows. Suppose the simulation and analysis described above are repeated many times. Then, among all genomic windows with WPPA equal to $q$, a proportion $q$ is expected to contain one or more SNPs with a true associations with the trait.

Recall that PER is associated with the test of a hypothesis. The null hypothesis $H_0$ in this case is that the genomic window $W_C$ does not contain any SNPs associated with the trait. Using this notation, WPPA is the conditional probability that $H_0$ is false given the observed data, while PER is the conditional probability that $H_0$ is true given that $H_0$ has been rejected based on some statistical test. Suppose the test is based on WPPA and $H_0$ is rejected whenever WPPA is larger than some value $t$. Then, PER is the probability that $H_0$ is true given WPPA is larger than $t$, and it can be written as:

$$
\text{PER} = \Pr(H_0 \text{ is true} | \text{WPPA} > t)
= \frac{\int_t^1 (1 - q)f(q) dq}{\int_t^1 f(q) dq}
= \mathbb{E}[(1 - \text{WPPA}) | \text{WPPA} > t]
\leq 1 - t,
$$

where $f(q)$ is the density function of WPPA. Thus in hypothetical repetitions of the analysis, for any interval with WPPA $> t$ the proportion of false positives among significant results will be $\leq 1 - t$. 

One of the advantages of inference based on posterior probabilities such as WPPA is that derivation of the distribution of the test statistic is not required. Further, the posterior probabilities of interest are easily obtained from the MCMC samples from the Bayesian analysis. Consider for example the situation where interest is in detecting regions that explain more than some proportion $v$ of the total genetic variance ($\sigma^2_g$) (Hayes et al., 2010; Fan et al., 2011). The variance that is attributed to a genomic region or window is defined as follows. First, the component of the genotypic value that corresponds to genomic window $W_C$ is defined as

$$g_w = Z_w \alpha_w,$$  \hspace{1cm} [14]

where $Z_w$ is the matrix of genotype covariates and $\alpha_w$ is the vector of partial regression coefficients for SNPs in the window $W_C$. In other words, $g_w$ is the sum of the genotypic values for SNPs in window $W_C$. Then, the window variance is defined as the sample variance of the components of $g_w$. In most Bayesian multiple-regression analyses, samples of the partial regression coefficients are drawn from their posterior distribution, and these samples can be used to compute window variances. The posterior probability that the window variance is larger than $v\sigma^2_g$ can be estimated by counting the number of MCMC samples where the window variance is larger than $v\sigma^2_g$.

In this section, the relationship between WPPA and PER was discussed in the context of Bayesian analysis of simulated data, where the distributions used in the generation of the data were used as priors in the Bayesian analysis. In analysis on real data, the distribution of unobservables quantities such as the partial regression coefficients and residuals are not known. Thus, based on available knowledge of the problem, priors that lead to
computationally efficient algorithms are used. Computer simulation will be used here to study the impact of such priors on the relationship between WPPA and PER in a realistic GWAS setting. As the amount of data that are combined with the priors increases, the impact of the priors on the posterior distributions is expected to decrease.

**A2.3.5 Computer Simulation**

The simulation described here was used to test if WPPA can be used to control false positives in GWAS, where the tests are dependent. Actual SNP genotypes of purebred Angus bulls were used to simulate QTL and phenotypes as described in Kizilkaya et al. (2010). Exactly 100 data sets with 1,000 observations and another 100 with 3,570 observations were simulated, using genotypes at 52,910 SNP loci on 3,570 purebred Angus bulls. The 1,000 bulls were randomly sampled without replacement for inclusion in the data sets with 1,000 observations, whereas all bulls were used in the data sets with 3,570 observations.

In each of the 200 data sets, SNP effects of markers were sampled according to the prior of the BayesC model of Habier et al. (Habier et al., 2011) with $\pi = 0.995$, where a proportion $\pi$ of the loci have null effects and the remaining loci have normally distributed effects with null mean and common variance $\sigma^2_\pi$ of SNP effects. The value of the common variance of SNP effects was chosen as in Kizilkaya et al. (2010) such that the additive genetic variance for the trait was 0.9. The average number of QTL in the data sets was about 260. The residual variance for the trait was set at 0.1 to give a heritability of 0.9.

The data sets with 1,000 observations were analyzed with and without including the SNPs that represent the QTL in the marker panel. The data sets with 3,570 observations were only analyzed without including the QTL in the marker panel. Posterior inferences were based on 10,000 MCMC samples after a burn-in of 1,000 samples.
In all analyses, the genome was divided into 2,676 one cM intervals according to the bovine map. The WPPA and window variance was computed for each such window, \( W_c \), as explained previously. In order to study the relationship between WPPA and the true probability of a QTL, each genomic window, \( W_c \), was classified into one of 10 WPPA classes of length 0.1 between 0 and 1. For example, all windows with WPPA between 0 and 0.1 were be classified into the first class, and those with WPA between 0.1 and 0.2 to the second class. The true probability of a QTL for a WPPA class was estimated by the observed frequency as follows. Let \( Q_{ij} \) denote the observed number of windows \( W_c \) belonging to WPPA class \( j \) or windows \( W_L \) or \( W_R \) flanking \( W_c \) with one or more QTL from data set \( i \), and let \( R_{ij} \) denote the total number of genomic windows that belong to WPPA class \( j \) from data set \( i \). Then, the actual frequency of a simulated QTL for WPPA class \( j \) was computed as

\[
f_j = \frac{\sum_{i=1}^{100} Q_{ij}}{\sum_{i=1}^{100} R_{ij}}.
\]  

BayesC with \( \pi = 0.995 \) or \( \pi = 0.999 \) was used for the analyses where the QTL were included in the marker panel. Recall that the prior of BayesC with \( \pi = 0.995 \) was used in the simulation of SNP effects. Thus, with \( \pi = 0.995 \), WPPA is expected to agree well with the actual frequency of the QTL. The relationship between WPPA and the actual frequency of QTL for these analyses are given for flanking windows of length \( k = 0, 1, \) or 2 cM in Figure A2.2.

Further, for these analyses with QTL included in the marker panel, each marker was classified into ten posterior probability of association (PPA) classes as described above for WPPA. Figure A2.3 shows the relationship between the PPA classes and the actual frequency of QTL among markers that belong to each class.
BayesC and BayesB with $\pi = 0.995$, and BayesC$\pi$, where $\pi$ is treated as an unknown, were used to analyze the data sets with 1,000 observations without including the QTL in the marker panel. The relationship between WPPA and the actual frequency of QTL for these analyses are given in Figure A2.4.

BayesC$\pi$ was used to analyze the data sets with 3,570 observations without including the QTL in the marker panel (Figure A2.5). In addition to computing WPPA for each genomic window, $W_c$, the posterior probability that the window variance ($\sigma^2$) exceeds 1/1,000 of the posterior mean of the total variance ($\sigma^2_{\alpha}$) was computed and grouped into 10 equally spaced posterior probability classes. For each such class, the actual frequency of a QTL in $W_c$ or in the flanking windows having a variance ($\nu_i$) that exceeds 1/1,000 of the total QTL variance ($\nu_{\theta}$) was also computed. The relationships between these probabilities and corresponding actual frequencies are given in Figure A2.6 using BayesC$\pi$.

**A2.4 DISCUSSION**

There are two main approaches to control false positives in genome-wide association studies. The most widely used approach is based on controlling the genomewise type I error rate (GER). The other is controlling the proportion of false positives among significant results, which as we have shown here, is equivalent to controlling the posterior type I error rate (PER) (Morton, 1955) for each test.

The longstanding practice of using a LOD score of three for declaring linkage between a monogenic disease locus and a random marker is based on control of PER to about 0.05 (Elston, 1997). The examples given in section 2.1 show that a test can have very different GER and PER values. If only one marker is tested, controlling GER to 0.05 will result in a
much larger value for PER, i.e., among significant results a large proportion would be false positives (94% in the example). Thus, to control PER to 0.05, a more stringent significance threshold has to be used (0.0001584 in the example). Now, suppose several markers are tested with the same stringent significance threshold each with the same prior probability of linkage to a QTL and power of test. Then for each test, PER would be 0.05, even when the tests are not independent. Thus, provided the prior probability of linkage and power are constant, the same significance threshold can be used to control PER regardless of the number of tests. This is not true for GER, which for a given significance threshold increases with the number of tests.

To control PER, however, it is required to know the value of \( \pi \), which is the probability that the null hypothesis is true for a test, and the power of the test. Mosig et al. (2001) have shown how to estimate these quantities, and these estimates can be used to estimate PER. On the other hand, in Bayesian analyses, these quantities can be treated as unknowns and an upper bound for PER can be obtained from Bayesian posterior probabilities as described in Section 2.4. For example, the PER for the test of a QTL in a genomic window, \( W \), is obtained as 1 – WPPA, where WPPA is estimated by counting the number of MCMC samples in which \( \alpha_j \) is non-zero for any SNP \( j \) in \( W_c \), the central window of \( W \) (Figure A2.1).

In Section 2.2 we discussed the relationship of PER to FDR, pFDR and PFP. The relationship between PER and PFP was used there to show that controlling PER for individual tests is equivalent to controlling the relative frequency given by (6) of false positives among all positives that accumulate across tests, which may be dependent, and across independent experiments. It is this property of PER, which does not depend on the
number of tests or their dependence structure, that makes it a desirable measure for controlling false positives in a GWAS and not its relationship to FDR, pFDR or PFP. Further, when PER is used for inference there is no multiple-test penalty (Stephens & Balding, 2009).

We have argued here that by using multiple-regression models the linkage signal can be localized to relatively short segments of the genome and thereby avoid the problem of signal dependence raised by Chen & Storey (2006). In model (1), assuming no interference, segregation of alleles at locus C are conditionally independent of the segregation of alleles at loci to the left of locus L and to the right of locus R given the segregation events at L and R. Thus, this model can be used to test the null hypothesis of no QTL between loci L and R in a linkage analysis, where the signal for QTL detection comes from the cosegregation of alleles at markers and QTL. In an association analysis, the signal for QTL detection comes from LD between markers and QTL, which is the non-independence between the allele states at these loci. Unfortunately, even under the assumption of no interference, allele states at locus C may not be conditionally independent of allele states at loci to the left of locus L and to the right of locus R given the allele states at loci L and R. However, the LD signal, depending on the effective population size, decays much faster with distance between loci than that due to cosegregation. Further, in the Bayesian multiple-regression models that we have used here, all SNPs are used in the analysis. Thus, in association analyses signal dependence is less of a problem than in linkage analysis, and when multiple regression models are used with high-density SNP markers, the signal from a QTL is expected to be almost completely explained by the markers within a narrow genomic window containing it. If the QTL is at the edge of a window, however, its signal may bleed into the next window. Thus, we considered using a
composite window (Figure A2.1). The signal is measured only in the central window to test the null hypothesis of no QTL in the composite window.

In this study, posterior probabilities of association were computed for genomic windows of 1cM (WPPA). When the QTL were included in the marker panel used in the analysis and the distribution used to simulate the QTL effects was used as the prior for marker effects in the Bayesian analysis, Figure A2.2 (plot A) shows good agreement between WPPA for a 1cM window and the frequency of QTL in that window ($k = 0$). For example, among genomic windows with WPPA between 0 and 0.1, about 5% contained one or more QTL, and among windows with WPPA between 0.9 and 1.0, about 95% contained QTL.

Figure A2.2A also shows the frequency of QTL in composite windows consisting of a 1cM central window and left and right flanking windows of length $k = 0, 1,$ or 2 cM. In all three curves, WPPA was computed for the central window, but the actual QTL frequency was computed for either a 1, 3, or 5 cM window centered at the window for which WPPA was computed. Windows with WPPA between 0 and 0.1 had actual QTL frequencies of 0.06, 0.22, or 0.35 for windows of length 1, 3 or 5 cM. Given that about 260 QTL were simulated and uniformly placed in 2,676 genomic windows, the prior probability that a composite window would contain one or more QTL is 0.09, 0.25 or 0.38 for $k = 0, 1$ or 2. Thus, the actual frequencies were reduced from the prior values towards the WPPA of 0.05. This is most evident when $k = 0$, where the actual frequency was 0.06, indicating that in this case, the WPPA is mainly influenced by the presence or absence of QTL in the central window. In windows where WPPA was between 0.45 and 0.85, the actual frequency of QTL for $k = 0$ was about 0.05 lower than WPPA. This indicates that the WPPA is slightly inflated by the presence of QTL outside $W_C$ whose signal bleeds into the signal observed for $W_C$. 
Figure A2.2 (plot B) shows the same relationships when the Bayesian analysis used $\pi = 0.999$ although $\pi = 0.995$ was used in the simulation (QTL were included in the marker panel). The high value for $\pi$ makes it difficult for a locus to be included in the model, and this can explain why WPPA was lower than the actual QTL frequencies for $k = 0$ and WPPA between 0.15 and 0.55. On the other hand, for values of WPPA between 0.65 and 0.85, the agreement between WPPA and the actual QTL frequencies for $k = 0$ was better in plot B of Figure A2.2 with $\pi = 0.999$ than in plot A with $\pi = 0.995$, which was used in the simulation. A possible explanation is that when $\pi = 0.999$, loci with a strong signal are favored for entering the model relative to those with weaker signals. Thus, in this analysis where QTL were included in the marker panel, QTL with large effects will be favored relative to markers that are in high LD with these QTL. When $\pi = 0.995$ was used, markers in high LD with the QTL may have entered the model and prevented the actual QTL from coming in. This indicates a mixing problem given the 10,000 MCMC samples used for inference in the simulation. If that is the case, use of a longer chain would give better results with $\pi = 0.995$ for the entire range of WPPA. Also, it may indicate that alternative sampling strategies to improve mixing should be investigated.

Figure A2.3 shows the relationships between PPA for each marker and the actual frequency that the corresponding marker is a QTL (QTL were included in the marker panel). Here, the actual frequencies of QTL were even lower than in plot A of Figure A2.2 for WPPA between 0.45 and 0.85. Again, this may be due to a mixing problem. In plot A of Figure A2.2, the overestimation of the QTL frequency for $k = 0$ was thought to be due to markers from adjacent windows in high LD with the QTL entering the model and preventing the actual QTL from coming in. In Figure A2.3, this can even happen with markers in the
window that contains the QTL. Thus, the observation that QTL frequencies are lower in Figure A2.3 than in plot A of Figure A2.2 is consistent with the expectation that markers within the QTL window would have higher LD with the QTL than markers from adjacent windows.

In the remaining analyses, the QTL were not included in the marker panel. Figure A2.4 presents results from three analyses of the 100 data sets with 1,000 observations, and Figures A2.5 and A2.6 gives results for the 100 data sets with 3,570 observations. In these analyses that did not have the QTL included in the marker panels, in genomic windows of 1cM and \( k = 0 \) WPPA for \( W_c \) substantially overestimated the frequency of QTL in the window \( W_c \) when WPPA greater than about 0.15. For example, in plot B of Figure A2.4, which shows the relationship between WPPA and the frequency of QTL for BayesC with \( \pi = 0.995 \), in genomic windows of 1cM with WPPA between 0.9 and 1.0 the frequency of QTL was about 0.72 and in genomic windows of 1cM with WPPA between 0.8 and 0.9 the frequency of QTL was only about 0.5. When the QTL were included in the analysis, the comparable QTL frequencies were 0.97 and 0.81 (plot A of Figure A2.2). Thus, when the QTL were not in the panel, WPPA overestimated the frequency of QTL in \( W_c \). Following are two possible reasons for this. The first is that the prior used for marker effects does not agree with the actual distribution of effects. When the QTL are not included in the marker panel, only markers that are in complete LD with the QTL will have effects that are distributed as the QTL. In Angus, the average LD between adjacent markers for the 50k SNP panel is about 0.2 (Goddard & Hayes, 2009). Thus, the distribution of marker effects may be quite different from that of the QTL and this may have an impact on the relationship between WPPA for a genomic interval and the frequency of QTL in that interval even when the distribution used
to generate the QTL effects is used as the prior for marker effects as in the BayesC analysis with $\pi = 0.995$. The second reason is violation of the assumption that WPPA is equivalent to the posterior probability that $W_C$ contains a QTL (WPPQ). Recall that WPPA is the posterior probability that a marker in window $W_C$ has a non-zero effect on the trait. When the QTL are included in the panel, WPPA is also the posterior probability of a QTL in $W_C$ because QTL by definition have non-zero effects on the trait. However, when the QTL are not included in the panel, WPPA is not equivalent to probability of a QTL in $W_C$. A marker in $W_C$ may have a non-zero effect even when $W_C$ does not contain any QTL due to it being in LD with a QTL in an adjacent window. This would cause WPPA to be higher than the QTL frequency, which is consistent with our results.

It can be argued that both of the reasons given above played a role in the observed overestimation of QTL frequencies in $W_C$ by WPPA. Violation of the assumption that WPPA is equivalent to WPPQ, however, seems to have played a greater role. The three plots in Figure A2.4 were obtained using three different priors. Plot A is from a BayesB analysis with $\pi = 0.995$, where a central $t$ distribution with four degrees of freedom was used as the prior for marker effects. Plot B is from BayesC with $\pi = 0.995$, where a normal distribution is used for marker effects, and plot C is from BayesC$\pi$, where $\pi$ is treated as unknown with a uniform prior between 0 and 1 and a normal prior for marker effects. The results from these three analyses being very similar indicates that with 1,000 observations these differences in priors had a negligible effect on the relationships between WPPA and QTL frequencies. Further, if the overestimation of QTL frequencies by WPPA was due to the prior for marker effects not being appropriate, then better results would be expected in the data sets with 3,570 observations. However, this was not the case. Overestimation was even greater with the
bigger data sets (Figure A2.5). On the other hand, if the observed overestimation of QTL frequencies was due to markers in $W_C$ being in LD with QTL in adjacent windows, it is possible that with more data associations with even more distant QTL could further inflate WPPA. Comparison of QTL frequencies in plot C of Figure A2.4 with those in Figure A2.5 for genomic windows with WPPA between 0.8 and 0.9 and $k = 0, 1, \text{ and } 2$ suggests that with the bigger data sets more distant QTL contributed to the WPPA value calculated for $W_C$.

In these analyses that did not include the QTL in the marker panel, there was good agreement between WPPA and the actual frequency of the QTL in the composite window $W$ with $k = 2$ when WPPA was larger than 0.8. At lower values of WPPA, WPPA underestimated the QTL frequency for $k = 2$. In genomic windows with WPPA between 0 and 0.1, the QTL frequency with $k = 0$ was almost 0.05, agreeing very well with WPPA. This is because the QTL in these windows have only very small effects. Thus, only the QTL from $W_C$ contribute to WPPA. As mentioned previously, with $k = 2$, the prior probability of a QTL in $W$ is 0.38. The actual value for windows with WPPA between 0 and 0.1 was 0.3 for $k = 2$, which is lower than the prior value of 0.38. Genomic windows with higher values of WPPA, for example, between 0.2 and 0.3, consists of a mixture of windows containing QTL of moderate size in the flanking windows which affect WPPA computed for $W_C$ and smaller QTL that do not affect WPPA computed in the central window $W_C$. As WPPA gets higher, most windows contain large QTL that contribute to the high value of WPPA.

Figure A2.6 shows the relationship between the posterior probability that the window variance (PPWV) exceeds 1/1,000 of the total variance and the corresponding actual frequency for the QTL variance. PPWV are especially useful for GWAS using models such
as BayesA (Meuwissen et al., 2001) and Bayesian Lasso (de los Campos et al., 2009) where all markers are assumed to have non-null effects and, thus WPPA is always 1.

Here, we will use window variances to examine the signal from the flanking windows. In the BayesC\(\pi\) analyses (Figures A2.4 and A2.5), the actual frequencies of QTL were in good agreement with WPPA for WPPA values larger than 0.85 and \(k = 2\). However, when \(k = 0\) the actual frequencies were much lower than WPPA. This indicates that the high values of WPPA are due partly to strong signals from the flanking windows. This can be tested by examining the QTL signal in the central and flanking windows for segments with \(PPWV=0.95\) in comparison to the corresponding signal in segments with \(PPWV=0.05\). In segments with \(PPWV=0.95\), the QTL in the central window \((k = 0)\) had a mean that was 1.1% of the total variance, and those in the flanking windows \((k = 2)\) had a mean that was 1.3% of the total variance. In segments with \(PPWV=0.05\), the QTL in the central window had a mean that was 0.1% of the total variance, and those in the flanking windows had a mean that was 0.3% of the total variance. Thus, in segments with \(PPWV=0.95\), there was a QTL with a strong signal in the central window or one with even a stronger signal in the flanking windows. These simulation results show that there is good agreement between posterior probabilities and the actual frequencies for the corresponding events when the priors used for the analysis represent the actual distribution of the marker effects. When the QTL are not included in the marker panel, the distribution of marker effects is not known even for simulated data. In analysis of real data, this will be even more of a problem. When the “correct” prior was not used in the analysis, there was good agreement between the posterior probabilities and the actual frequencies for low values of the posterior probabilities and \(k = 0\) and at high values of the posterior probabilities and \(k = 2\). The width of the
genomic interval that gives good agreement between the posterior probabilities and the actual frequencies may depend on the distribution of the QTL effects, the LD structure between the markers and the QTL, and the amount of data. However, based on these simulation results it is expected that genomic windows with a high values for WPPA or PPWV would have equally high frequencies of large QTL in \( W_c \) of close to it.

The simulation of QTL genotypes and trait phenotypes, and the Bayesian analyses presented here were based on version 4.0 of the GenSel program (Fernando & Garrick). A web version of this program is available at: http://bigs.anisci.iastate.edu/bigsgui/login.html.

In summary, we have argued here that PER is more suitable for controlling false positives in GWAS than GER. Controlling PER at individual tests results in control of false positives for the collection of tests that may be dependent. Further, we have shown the relationship between PER and WPPA under the ideal situation where the “correct” prior is known. Computer simulation was used to examine the impact of not knowing the “correct” prior. If a high value of WPPA or PPWV is used to detect QTL, among the positive results that accumulate over many experiments the proportion of false positives can be expected to be low. Further, use of multiple regression models allows inference on the presence or absence of QTL to be specific to relatively narrow regions of the genome. Thus, the problem of signal dependence (Chen & Storey, 2006) is to a large degree avoided. Further research is needed to determine the optimum size of the central and flanking window sizes, which may depend on the nature of LD, the number of observations, heritability of the trait and possibly other factors such as the number of QTL which may be unknown.
A2.5 ACKNOWLEDGEMENTS

Authors are grateful to Dr. Soller for his instruction and research that has been a source of inspiration for much of our work.

A2.6 LITERATURE CITED


Figure A2.1 Illustration of composite genomic window $W$ consisting of central window $W_C$ and flanking windows $W_L$ and $W_R$. To test the null hypothesis of no QTL in $W$, window PPA (WPPA) is computed by counting the number of MCMC samples in which $a_j$ is non-zero for at least one SNP in the central window $W_C$. 
Figure A2.2 Relationship between window posterior probability of association (WPPA) and the actual frequency of simulated QTL in analyses where the QTL were included in the marker panel. WPPA was computed for each 1cM window ($W_c$) of the genome and grouped into 10 WPPA classes (x-axis). For each WPPA class, the actual frequency of simulated QTL in the composite window consisting of $W_c$ and the flanking windows of $k$ cM ($k = 0, 1, \text{ or } 2$) in length is given in the y-axis. Results are for BayesC with $\pi = 0.995$ (plot A) and BayesC with $\pi = 0.999$ (plot B) from 100 data sets of 1,000 observations.
Figure A2.3 Relationship between posterior probability of association (PPA) of individual markers and the actual frequency of simulated QTL in analyses where the QTL were included in the marker panel. PPA was computed for each marker in the panel and grouped into 10 PPA classes (x-axis). For each PPA class, the actual frequency of markers in that class being a QTL is given in the y-axis. Results are for BayesC with $\pi = 0.995$ from 100 data sets each with 1,000 observations.
Figure A2.4 Relationship between window posterior probability of association (WPPA) and the actual frequency of simulated QTL. WPPA was computed for each 1cM window ($W_C$) of the genome and grouped into 10 WPPA classes (x-axis). For each WPPA class, the actual frequency of simulated QTL in the composite window consisting of $W_C$ and the flanking windows of $k$ cM ($k = 0, 1, 2$) in length is given in the y-axis. Results are for BayesB with $\pi = 0.995$ (plot A), BayesC with $\pi = 0.995$ (plot B), and BayesC$\pi$ (plot C) from 100 data sets each with 1,000 observations. The QTL were not included in the marker panel.
Figure A2.5 Relationship between window posterior probability of association (WPPA) and the actual frequency of simulated QTL. WPPA was computed for each 1cM window ($W_C$) of the genome and grouped into 10 WPPA classes (x-axis). For each WPPA class, the actual frequency of simulated QTL in the composite window consisting of $W_C$ and the flanking windows of $k$ cM ($k = 0, 1, 2$) in length is given in the y-axis. Results are for BayesC-$\pi$ from 100 data sets each with 3,570 observations. The QTL were not included in the marker panel.
Figure A2.6 Relationship between posterior probability that variance of the central window $W_C$ exceeds 1/1,000 of total variance and corresponding actual frequency of the simulated QTL variance for the composite window $W$. The posterior probability that the window variance ($\sigma^2_l$) exceeded 1/1,000 of the posterior mean of the total variance ($\sigma^2_g$) was computed for each 1cM window ($W_C$) of the genome and grouped into 10 posterior probability classes (x-axis). For each such class, the actual frequency of simulated QTL in the composite window consisting of $W_C$ and the flanking windows of $k$ cM ($k = 0, 1, or 2$) in length having a variance ($\sigma^2_l$) that exceeds 1/1,000 of the total QTL variance ($\sigma^2_g$) is given in the y-axis. Results are for BayesC$\pi$ from 100 data sets each with 3,570 observations. The QTL were not included in the marker panel.