

COMPARISON OF METHODS FOR PARTITIONING TRAINING AND VALIDATION  
POPULATIONS TO OPTIMIZE PREDICTION ACCURACY AND ENABLE ACROSS-  
BREED GENOMIC SELECTION IN BEEF CATTLE

---

A Dissertation presented to  
the Faculty of the Graduate School  
at the University of Missouri-Columbia

---

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy

---

By  
MEGAN M. ROLF

Dr. Jeremy F. Taylor, Dissertation Supervisor

JULY 2012

The undersigned, appointed by the deal of the Graduate School, have examined the dissertation entitled

COMPARISON OF METHODS FOR PARTITIONING TRAINING AND VALIDATION POPULATIONS TO OPTIMIZE PREDICTION ACCURACY AND ENABLE ACROSS-BREED GENOMIC SELECTION IN BEEF CATTLE

presented by Megan M. Rolf,

a candidate for the degree of doctor of Philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

---

Dr. Jeremy F. Taylor

---

Dr. William L. Lamberson

---

Dr. Robert D. Schnabel

---

Dr. Robert L. Weaber

---

Dr. Gary S. Johnson

*To my parents, Walt and Jackie Rolf:*

*You taught me to believe in myself  
and to work hard to achieve my dreams-  
I love you and thank you for everything*

## ACKNOWLEDGEMENTS

As with anything in science, the experiments and studies contained in this dissertation could not have been accomplished without the help of my colleagues. I have had the privilege to work with and study under one of the preeminent scientists in his field and I want to thank Dr. Jerry Taylor for the time he has spent mentoring me and for accepting me into his program. I also want to thank my other committee members, Dr. Lamberson, Dr. Schnabel, Dr. Weaber and Dr. Johnson for their help and guidance on this, and many other endeavors. Working with each of them has opened up new areas of insight and helped me to build skills and knowledge that will be immensely valuable in my career.

I would be remiss not to extend a huge thank you to the other members in my lab, both current and former, including Stephanie McKay for all of her genotyping skills, Matt McClure for his help extracting DNA samples, JaeWoo Kim for providing support and knowledge, Jared Decker for all his scientific insight, Rich Chapple for answering programming questions, and Holly Ramey for her support and friendship. I would also like to thank my friends, both inside and outside the lab that have become my “Mizzou family” over the past few years including Holly Ramey, Elizabeth Benavides, Mark Benavides, and Veronica Negron.

None of the research presented in this dissertation would be possible without the support of breeders and semen distributors whom have provided samples for these projects. I want to extend a special thank you to Wulf Limousin in Morris, MN, Kent Abele of the Green Springs Bull Test in Nevada, MO, and all of the breed associations I

have been fortunate enough to work with including the American Angus Association, North American Limousin Foundation, American International Charolais Association, American Simmental Association, and the American Hereford Association. They have been an invaluable source of information for both this project as well as others, and I appreciate all of their support.

No degree is complete without spending time as a teaching assistant and I am grateful for Drs. Randy Prather, Anne McKendry, William Lamberson, Bryon Wiegand and George Jesse for providing me the opportunity to be a teaching assistant in their classes, and in some cases, teach some lectures and provide guidance during lab activities. I value these opportunities and believe they have helped me to be a more effective teacher and scientist.

I would like to extend a special thank you to the administration, faculty, and staff at Oklahoma State University for giving me the wonderful opportunity to be a member of their faculty and extension team. I am looking forward to serving the producers in the state and being a member of the department, and I appreciate everyone's efforts to make me feel welcome.

Finally, I would like to thank my family and friends for all of their support and for making life worth living. My parents deserve a special thank you for putting up with me for 28 years. Their love, support, and sacrifice have enabled me to make it to where I am today and I hope that I have, and continue to, make them proud.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	ii
LIST OF FIGURES .....	vi
LIST OF TABLES .....	xi
LIST OF ABBREVIATIONS.....	xiii
ABSTRACT.....	xv

### Chapter

1. REVIEW OF LITERATURE.....	1
Introduction .....	1
Principles and applications of genomic selection .....	4
Genomic selection methodology .....	8
Genomic relationship matrices .....	8
Genomic BLUP .....	9
Bayesian modeling .....	10
Within-breed Analyses .....	12
Across-breed genomic selection.....	14
Implementation of genomic selection in the beef industry.....	20
Future of genomic selection in the beef industry .....	22
2. SUMMARY OF OBJECTIVES.....	24
Goals of the study	
3. EXPLORATION OF METHODS FOR SEPARATION OF TRAINING AND VALIDATION POPULATIONS TO IMPROVE GENOMIC SELECTION IN CROSSBRED AND PUREBRED BEEF CATTLE POPULATIONS .....	27
Summary .....	27
Introduction .....	28
Materials and Methods .....	35
Population.....	35
Phenotypic data .....	35
Genotypic data.....	37
Approach .....	37

Genetic distance measures.....	44
Random allocation measures.....	49
Genomic relationship matrix measures .....	49
Simmental external validation.....	50
BayesB0 and BayesC0 comparisons .....	52
Results and Discussion.....	53
Informed priors.....	53
Genetic distance analyses .....	53
GRM analyses .....	56
Random allocation.....	58
Model comparisons .....	61
External validation.....	70
Conclusions .....	73
4. GENOME-WIDE SCAN FOR QUANTITATIVE TRAIT LOCI AFFECTING CARCASS TRAITS IN FIVE CROSSBRED AND PUREBRED TAURINE BEEF BREEDS.....	74
Summary .....	74
Introduction .....	75
Materials and methods.....	79
Population.....	79
Phenotypic data .....	79
Genotypic data.....	80
Analysis .....	81
QTL delineation.....	83
Results and Discussion.....	84
Conclusions .....	113
5. SUMMARY AND CONCLUSIONS .....	114
Summary of conclusions from these studies	
APPENDIX.....	117
REFERENCES .....	148
VITA.....	154

## LIST OF FIGURES

Figure	Page
<p>3.1 Influence of priors on BayesB analyses. Comparison of the Bayesian analyses used with original uninformed parameters for BayesB0 (<math>B0_u</math>) and informed parameters for BayesB0 (<math>B0_i</math>). All analyses were completed on WBSF using a random sample of animals from the total sample for training and the remainder of animals for validation. Panel A shows realized accuracies calculated using a constant heritability from the BayesC0 best fit analysis. Panel B shows realized accuracies calculated using heritability estimates within each respective analysis. ....</p>	43
<p>3.2 Comparison of allocation using measures of genetic distance between breeds compared to random allocation of animals regardless of breed composition using BayesC<math>\pi</math> analyses. The single random allocation category is random allocation of 2/3 of the dataset into training and 1/3 into allocation. Each distance analysis was paired with an analysis using random allocation with the same sized training population for an equivalent comparison. Panel A shows WBSF analyses and panel B shows REA analyses. Significant differences between paired random and distance analyses (<math>p &lt; 0.05</math>) are denoted by an asterisk .....</p>	47
<p>3.3 Phylogenetic tree for all CMP samples .....</p>	48
<p>3.4 Schematic explanation of methods to separate animals into training and validation groups based on genomic relationship matrix (GRM) coefficients.....</p>	51
<p>3.5 WBSF distance analyses with extra bootstraps. Set 1 and set 2 consist of 20 bootstrap analyses each for every method tested. No significant differences were detected in any of the analyses .....</p>	55
<p>3.6 Comparison of GRM methods for separation of training and validation populations. Blue (Green) bars represent those analyses that maximize (minimize) the relationship between training and validation sets, thus it is expected that green bars will be lower than blue bars for the most effective methods. Panels A, B and C show methods 1, 2, and 3 from Figure 3.5, respectively.....</p>	57

3.7	Comparison of realized accuracies calculated for WBSF BayesC $\pi$ analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.....	60
3.8	Comparison of BayesC $\pi$ within-breed realized accuracies for WBSF using random allocation.....	62
3.9	Comparison of BayesC $\pi$ within-breed realized accuracies for REA using random allocation.....	63
3.10	Comparisons of mean realized accuracies over 20 bootstraps for BayesB0 (red), BayesC $\pi$ (blue), and BayesC0 (green) WBSF analyses. Across-breed realized accuracies are presented in Panel A and realized accuracies within breeds are presented in Panel B.....	65
3.11	Comparisons of mean realized accuracies over 20 bootstraps for BayesB0 (red), BayesC $\pi$ (blue), and BayesC0 (green) FT analyses. Across-breed realized accuracies are presented in Panel A and realized accuracies within breeds are presented in Panel B.....	66
3.12	Comparisons of mean realized accuracies over 20 bootstraps for BayesB0 (red), BayesC $\pi$ (blue), and BayesC0 (green) REA analyses. Across-breed realized accuracies are presented in Panel A and realized accuracies within breeds are presented.....	68
3.13	Comparisons of mean realized accuracies over 20 bootstraps for BayesB0 (red), BayesC $\pi$ (blue), and BayesC0 (green) MARB analyses. Across-breed realized accuracies are presented in Panel A and realized accuracies within breeds are presented.....	69
3.14	Results from realized accuracies for WBSF in an external validation using Simmental animals separated by about 2 generations and largely unrelated to the training set. Panel A shows across breed means for random allocation in BayesB0, Bayes C $\pi$ , and dBayesC0 analyses and Panel B shows these means compared to the original mean realized accuracy in the internal validation set.....	71
4.1	Association analysis of WBSF using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	86
4.2	Association analysis of REA using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	87
4.3	Association analysis of WBSF using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	88

4.4	Association analysis of REA using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	89
4.5	Association analysis of WBSF using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	90
4.6	Association analysis of REA using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	91
4.7	Association analysis of WBSF using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	92
4.8	Association analysis of REA using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	93
A.1	Comparison of realized accuracies calculated for REA BayesC $\pi$ analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.....	117
A.2	Comparison of realized accuracies calculated for MARB BayesC $\pi$ analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.....	118
A.3	Comparison of realized accuracies calculated for FT BayesC $\pi$ analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.....	119
A.4	Comparison of realized accuracies calculated for HCW BayesC $\pi$ analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.....	120
A.5	Comparison of realized accuracies calculated for YG BayesC $\pi$ analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.....	121
A.6	Comparison of BayesC $\pi$ within-breed realized accuracies for MARB using random allocation .....	122
A.7	Comparison of BayesC $\pi$ within-breed realized accuracies for FT using random allocation .....	123
A.8	Comparison of BayesC $\pi$ within-breed realized accuracies for HCW using random allocation .....	124

A.9	Comparison of BayesC $\pi$ within-breed realized accuracies for YG using random allocation .....	125
A.10	Comparisons of mean realized accuracies over 20 bootstraps for BayesB0 (red), BayesC $\pi$ (blue), and BayesC0 (green) YG analyses. Across-breed realized accuracies are presented in Panel A and realized accuracies within breeds are presented in Panel B.....	126
A.11	Comparisons of mean realized accuracies over 20 bootstraps for BayesB0 (red), BayesC $\pi$ (blue), and BayesC0 (green) HCW analyses. Across-breed realized accuracies are presented in Panel A and realized accuracies within breeds are presented in Panel B.....	127
A.12	Association analysis of MARB using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	128
A.13	Association analysis of FT using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	129
A.14	Association analysis of HCW using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	130
A.15	Association analysis of YG using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	131
A.16	Association analysis of MARB using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	132
A.17	Association analysis of MARB using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome .....	133
A.18	Association analysis of FT using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	134
A.19	Association analysis of FT using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	135
A.20	Association analysis of HCW using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	136

A.21	Association analysis of HCW using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	137
A.22	Association analysis of YG using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	138
A.23	Association analysis of YG using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	139
A.24	Association analysis of FT using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	140
A.25	Association analysis of YG using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	141
A.26	Association analysis of MARB using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	142
A.27	Association analysis of HCW using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.....	143
A.28	Association analyses of Tender. Panel A depicts individual SNP estimates from BayesB95, Panel B individual SNP estimates from BayesB0, and Panel C individual SNP estimates from BayesC0.....	144
A.29	Association analyses of Juicy. Panel A depicts individual SNP estimates from BayesC95, Panel B individual SNP estimates from BayesC0, and Panel C individual SNP estimates from BayesB0.....	145
A.30	Association analyses of Flavor. Panel A depicts individual SNP estimates from BayesC0, Panel B individual SNP estimates from BayesC95, and Panel C individual SNP estimates from BayesB0.....	146
A.31	Association analyses of ConnTiss. Panel A depicts individual SNP estimates from BayesC95, Panel B individual SNP estimates from BayesC0, and Panel C individual SNP estimates from BayesB95.....	147

## LIST OF TABLES

Table	Page
3.1 Number of phenotypes available for analysis in each breed and trait .....	38
3.2 Starting priors for BayesC $\pi$ , BayesC0, and BayesB0 analyses .....	40
3.3 Within breed heritabilities derived from GBLUP.....	45
3.4 Parameters ( $\pi$ , $h^2$ , $r_{\hat{g},y}$ ) derived from the mean of posterior distributions for the best-fit analyses for each trait along with calculated measures of realized accuracy .....	59
4.1 Summary of association analyses for each trait and model combination .....	95
4.2 Number of genes concordant between analyses for each individual trait. Total number of genes in each analysis is listed on the diagonal. Percentages in the lower triangular section are percentages of total genes implicated for each trait that are shared between analyses.....	98
4.3 Number of genes concordant between analyses for each individual trait. Total number of gene names in each analysis is listed on the diagonal. Percentages in the lower triangular part of the table are percentages of the total genes implicated for each trait that are shared between any two pairs of analyses.....	100
4.4 Number of genes concordant for each trait. Expected relationships are highlighted in blue. Total numbers of genes from each cumulative gene list are listed on the diagonal .....	101
4.5 Number of genes in each gene list that overlap between different analysis models. DAVID analysis included all genes that were concordant in 3 of 4 or 2 of 3 analyses. Percentages are a reflection of the number of genes in each category divided by the total number of genes in each cumulative gene lists for an individual trait .....	103
4.6 Functional annotation clustering results from DAVID analyses of cumulative concordant gene lists for each trait .....	104

4.7 QTL regions detected by McClure et al. (2012) using within-breed analysis approaches that were also concordant with across-breed analysis methods used in this study for WBSF and Tender ..... 110

## LIST OF ABBREVIATIONS

ERT	Economically relevant trait
BLUP	Best Linear Unbiased Prediction
NCE	National Cattle Evaluation
EPD	Expected progeny difference
MAS	Marker assisted selection
SNP	Single nucleotide polymorphism
GS	Genomic selection
LD	Linkage disequilibrium
QTL	Quantitative trait loci
NRM	Numerator relationship matrix
GRM	Genomic relationship matrix
ASE	Allele substitution effect
$\pi$	Proportion of markers that do not influence a trait
MCMC	Markov Chain Monte Carlo
GE-EPD	Genomic-enhanced EPDs
GBLUP	Genomic best linear unbiased prediction

NCBA	National Cattlemen's Beef Association
CMP	Carcass merit Project
WBSF	Warner-Bratzler Shear Force
REA	Ribeye area
MARB	Marbling score
FT	Fat thickness at the 12 <sup>th</sup> and 13 <sup>th</sup> rib interface
HCW	Hot carcass weight
YG	Yield grade
Tender	Tenderness
Juicy	Juiciness
ConnTiss	Connective tissue

## ABSTRACT

Across-breed genomic selection practices have the potential to revolutionize national genetic evaluation systems in the United States by including commercial cattle and increasing prediction power for hybrid animals. We used a population of 3,240 animals from the Carcass Merit Project to build across-breed genomic selection models for six carcass and four sensory panel traits across five breeds of commercially relevant beef cattle (Angus, Charolais, Hereford, Limousin, and Simmental). Allocation of these animals to training or validation populations based on genetic distance measures or genomic relationships coefficients proved to be no more effective than random allocation of animals. Realized accuracies in these populations showed that the prediction models were very effective when used on animals within the same project and short time span (0.41-0.78). When used in an external validation in animals separated by approximately 10 years, prediction accuracies showed severe reductions (from ~0.6 down to 0.05), indicating that retraining of prediction models will have to be done frequently (possibly annually) in commercial populations. We also identified numerous regions of the genome which showed evidence of harboring causal mutations for these traits of economic importance. These regions will serve as independent validations in the literature as well as a guidepost for researchers looking for causal mutations within the bovine genome.

## CHAPTER I: REVIEW OF LITERATURE

### **Introduction**

Within the beef industry, annual cow inventories have continued to decline over the past decade, while the amount of saleable product available for consumers has continued to increase (Figure 1.1). Along with improved management practices, genetic selection within the industry has been at the forefront of providing consumers with a healthy, affordable and palatable product. Many of the earliest improvements were in easily measured economically relevant traits (ERT) such as birth, weaning, and yearling weight, where selection could be practiced relatively effectively on phenotypic data. With the advent of Best Linear Unbiased Prediction (BLUP; Henderson 1963), National Cattle Evaluation (NCE) was made possible and implemented within the beef industry (Willham 1993). This development created a selection tool in the form of expected progeny differences (EPDs) for maternal traits (calving ease, milking ability, maternal effects on weaning weight) and carcass value (ribeye area, marbling, fat thickness). These traits were historically difficult to select using phenotypic data due to small numbers of records (carcass value) and low heritabilities (maternal traits). Because of the widespread availability of effective selection tools for producers, genetic trends within most of the major purebred breeds in the US have been rapid and advantageous over

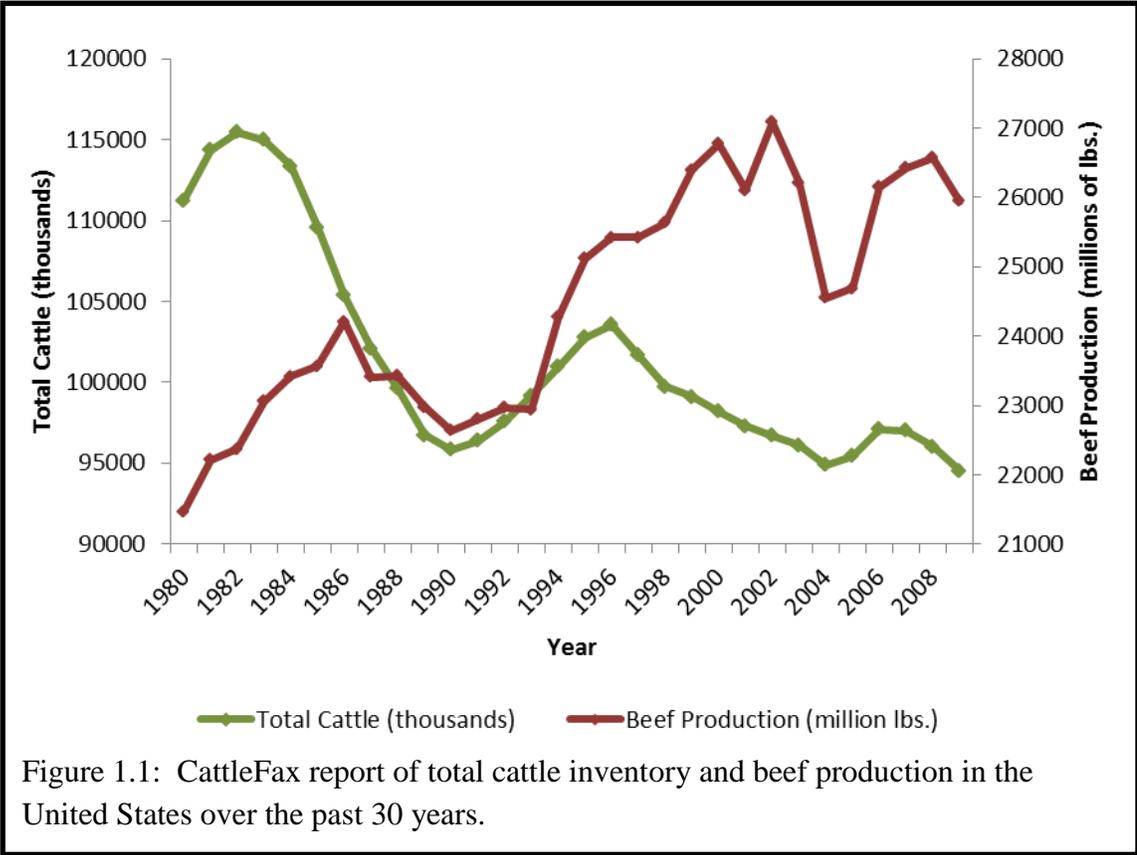


Figure 1.1: CattleFax report of total cattle inventory and beef production in the United States over the past 30 years.

the last four decades, allowing producers to generate more than enough product to offset the decline in cow inventory.

Genetic marker data in the form of high density single nucleotide polymorphism (SNP) genotypes is poised to define the next frontier for the development of selection tools for beef producers. Genotype information was first used as a selection tool using an approach known as marker assisted selection (MAS). MAS involves the use of a relatively small number of genotypes which provide information on only a small portion of the genome in selection decisions. Early approaches using MAS for quantitative traits were unsuccessful due to lack of information on the majority of the genome which harbored the majority of the genetic variation influencing the traits. However, MAS approaches have been very beneficial for traits with Mendelian inheritance such as the detection and testing of cattle to determine their carrier status for recessive autosomal defects (Arthrogyrosis Multiplex, Neuropathic Hydrocephalus, Fawn Calf Syndrome, Osteopetrosis [Meyers *et al.* 2010], Protoporphyrin, dwarfism [Takeda *et al.* 2002], hypotrichosis [Drogemuller *et al.* 2002], idiopathic epilepsy, Tibial Hemimelia, and Pulmonary Hypoplasia with Anasarca) as well as for determining the number of black coat color and polled alleles possessed by an individual.

With the advent of large SNP panels which can genotype anywhere from 9K to 700K genotypes per animal for under \$200, new approaches which incorporate this information into selection decisions and the development of genotype-enhanced selection tools for producers must be explored. The term genomic selection (GS) was coined by

Meuwissen *et al.* (2001) and refers to using high-density SNP panels to simultaneously select for all genes in the genome which affect an economically relevant trait. GS has the advantage over marker assisted selection in that it explains more of the genetic variance in a trait than does the small number of markers used for MAS, making it more expedient as a selection tool for quantitative (growth, carcass and fertility) traits.

### **Principles and Applications of Genomic Selection**

Traditional pedigree-based analyses include only the probabilities that genes are identical by descent, whereas genomic selection has the ability to trace the inheritance of individual genes identical in state within a population (VanRaden *et al.* 2009). Genetic variation in quantitative traits can be modeled using GS via exploiting the linkage disequilibrium (LD) present between the causal mutations that underlie a particular phenotype and large numbers of well-spaced SNP markers spread throughout the genome. SNP markers are superior to other forms of markers for this purpose, because they are inexpensive to genotype, vary in allele frequency and they are plentiful. SNPs can be found approximately every 700 bases in *Bos taurus* cattle and about every 300 bases in *Bos indicus* cattle (The Bovine Hap Map Consortium 2009). Any SNP in high LD with a causal mutation can be exploited as a selection tool for the underlying quantitative trait loci (QTL) without knowing the true causal mutation, theoretically allowing the simultaneous selection for all QTL in an animal's genome using only a panel of high-density SNP markers (Hayes *et al.* 2009).

As evidenced by the historic genetic trends that have occurred within breeds (i.e., <http://www.angus.org/Nce/GeneticTrends.aspx>, <http://www.nalf.org/pdf/2010/aug19/tackletemperament.pdf>, <http://simmental.org/site/index.php/simgenetics/genetic-trends>), selection using traditional EPDs within the beef industry has been very successful. However, GS has the opportunity to further increase the efficacy of selection by providing genetic predictions with high accuracies at a young age, provided that the computational infrastructure can be developed to process the data. More intense selection can be practiced on young animals before they have produced progeny (de Roos *et al.* 2007), which dramatically shortens generation interval and increases genetic progress (Seidel 2009). It has been estimated that GS could increase response to selection by a factor of two, and save approximately 92% of the costs of “proving” bulls in the dairy industry (Schaeffer 2006). In contrast, the beef industry does not typically bear the expense of proving bulls due to a much larger use of natural service sires. However, more accurate EPDs at a younger age could reduce expenses in the beef industry by eliminating inferior animals and their progeny from the breeding pool before the expense of raising the animal to breeding age. Additionally, it is likely that there are substantially more genetically superior females than males due to their larger numbers in the population, but females rarely achieve high-accuracy EPDs due to limited numbers of progeny. Using GS, EPD accuracies are equivalent for both males and females (VanRaden *et al.* 2009), thus allowing the identification of superior females in large populations of animals which have been genotyped.

Many of the most important applications of GS in the beef industry will undoubtedly pertain to traits that are:

- a. Novel – Traits which do not have many phenotypic records and are not suited to routine collection outside of discovery and research populations (i.e., greenhouse gas emissions)
- b. Measured late in life – By the time that sufficient data exists to enable selection on these traits (i.e., stayability and longevity), the animal’s reproductive life has nearly ended
- c. Difficult and/or expensive to record – Cost of data collection for these traits is prohibitive for routine collection due to the necessity of special equipment (i.e., feed efficiency) or loss of life/productivity (i.e., disease resistance)
- d. Collected at harvest – These phenotypes can only be collected on progeny after they reach adulthood, and animals used in breeding herds will never possess their own individual record (i.e., carcass traits)
- e. Sex-limited – Observations that can only be recorded in one gender require many progeny records and accuracies will tend to be lower in the sex in which no individual observations are recorded
- f. Associated with decreases in animal welfare – The collection of useful phenotypes may require a reduction in welfare for the animals in the discovery population, which necessitates the minimization of record collection to limit exposure to disease or invasive techniques (Solberg *et al.* 2009).

GS may also be valuable for the prediction of merit for traits with low heritabilities, such as heifer pregnancy rate. Currently, the best methods for achieving gains in these traits are through better management and crossbreeding schemes; however, GS would more easily allow selection on the genes underlying variability in these traits. Increases in additive genetic merit can be paired with improved management to make expeditious gains in lowly heritable traits.

The principles of GS dictate that a large population, called the training population, is needed to estimate the marker effects that are used to predict genetic merit in future generations of animals. These marker effects are tested for their predictive power and accuracy in an independent, mutually exclusive validation population. The success of this approach in the validation population is believed to be dictated by the LD present between the markers and underlying QTL, as well as the number of animals and phenotypic records available in the training population (Toosi *et al.* 2010). Additional value can be captured from the incorporation of GS methodologies into selection decisions by the generation of large phenotypic and genotypic databases (Sosnicki and Newman 2010). These databases are a valuable asset for the validation of DNA markers identified in independent experimental populations and for testing candidate markers for their causality (Sosnicki and Newman 2010). When GS was first proposed, it was suggested that phenotypes would be non-essential when DNA information was routinely used for genetic prediction (Habier *et al.* 2007). However, it is now well accepted that it will continue to be necessary to collect high-quality phenotypes to re-train the prediction models (Seidel 2009) to account for changes in population structure and LD (Schaeffer

2006). As the selection of beneficial mutations causes alleles to move towards fixation in a population, these resources and DNA repositories will allow the re-estimation of marker effects (re-training) and will be critical to the success of GS programs.

### **Genomic Selection Methodology**

Many of the earliest GS studies were performed using simulated data because the data for actual populations did not yet exist to test methodologies. Recently, studies have been published in dairy, but few studies have been published that explore GS within beef cattle populations. Because these methodologies can also concurrently serve to perform genome wide association analyses (GWAS; i.e., McClure *et al.* 2012), these methodologies should be attractive for researchers working in gene discovery in beef cattle populations. However, further work must be performed to test these concepts within the different population structures in beef cattle, while taking into account industry considerations unique to beef production systems. Simulation results suggest that the underlying distribution of genetic variation within the genome determines the predictive ability of any particular GS model (Clark *et al.* 2011), thus the analysis of traits important in beef production, as opposed to dairy, may yield different results.

### *Genomic Relationship Matrices*

NCE uses pedigrees to define the proportion of genomic information shared in common between individuals with shared ancestry via the generation of a numerator relationship matrix (NRM). NRMs assume that an animal has inherited a random sample

of genes from each of its parents and that the extent of identity between two individuals can be unbiasedly estimated from the number of meioses which separate them (i.e., half-sibs share, on average,  $\frac{1}{4}$  of their genetic material). In reality, animals may share slightly more or less of their genomes identical by descent than expected based upon pedigree due to Mendelian sampling and selection, and, with genomic data, we can estimate the exact proportion of the genome that is shared (or alike in state) between relatives using a genomic relationship matrix (GRM). This matrix can be generated using only genomic data (i.e., methods 1 and 2 of VanRaden 2008) or by blending both pedigree and genomic data (i.e., Legarra *et al.* 2009, method 3 of VanRaden 2008). Villanueva *et al.* (2005) suggested that when no QTL have been mapped or when no QTL have large effects on the trait of interest, the use of a GRM in genomic selection will produce higher EBV accuracies by precisely explaining the additive relationships between individuals due to the marker genotypes that they share. One advantage of this computational strategy is that it does not require training and validation populations - all animals can be combined into one analytical dataset (i.e., Rolf *et al.* 2010) and predictions can be performed in a single step.

### *Genomic BLUP*

Genomic BLUP (GBLUP) is a linear mixed model methodology that uses an animal model which incorporates a GRM and is functionally equivalent to fitting individual SNP effects in a mixed model analysis (Garrick 2007). With this methodology, SNP effects are directly fit as a random effect and rather than being used to

form the relationships between animals. Alternatively, blocks of adjacent markers can be phased and formed into haplotype blocks, which can be used as a substitute to single SNP analyses. At times, a QTL may not be in significant LD with any single marker, but may be in complete LD with a haplotype block (Goddard and Hayes 2007). Calus *et al.* (2008) observed that the advantages of using haplotype blocks decreased as the LD between adjacent markers increased and was roughly equivalent when the average  $r^2$  value was 0.21. The disadvantages of haplotyping markers are most apparent when the exact genomic positions of markers are not well known due to mistakes and inaccuracies in genome assemblies, rendering the haplotype blocks unusable. Marker effects are assumed to be drawn from a normal distribution with a constant variance for all loci. Typically, all genotyped SNPs are fit in the model. GBLUP is particularly appealing because the only prior information that is required is the additive genetic variance for the trait of interest (Hayes *et al.* 2009). As a consequence, this method may ultimately be preferred for implementation when the genotyping of all animals is ubiquitous (Goddard and Hayes 2007). However, it is very likely that for most traits, some markers have zero effect (meaning that they are not predictive of a QTL); therefore the main drawback of GBLUP is that all markers are used in the prediction and have associated effects estimated. Thus, the markers that are not in LD with a causal mutation and that actually have no effect on the trait will obtain a small (but non-zero) estimated effect, and when breeding values at individual loci are summed, add noise to the breeding value predictions (Goddard and Hayes 2007).

### *Bayesian Modeling*

Recent studies suggest that Bayesian methodologies may ultimately be superior to the above described methods where there are a few mutations with very large effects (Clark *et al.* 2011), as they allow for the modeling of different allele substitution effect (ASE) variances among loci. Bayesian methods allow the inclusion of information about the distribution of QTL effects through use of priors. These priors (which can be normal or otherwise) may allow a better characterization of the true distribution of QTL effects across the genome. As shown by Clark *et al.* (2011) in a simulation study, under an infinitesimal model, GBLUP and BayesB perform similarly; however more accurate breeding values were obtained using BayesB when a QTL or rare variant model was utilized to simulate the underlying data. Similar advantages for Bayesian methodologies were noted by VanRaden *et al.* (2009) who reported results from a study in dairy cattle where the coefficients of determination obtained for non-linear predictions were 0.05 to 0.38 greater than parent average predictions. Most significantly, the largest gain in  $R^2$  was obtained for fat percentage, which has a known gene of large effect (*DGATI*). While VanRaden *et al.* (2009) reported similar results between Bayesian estimation and linear modeling, Hayes *et al.* (2009) found the increase in reliability of Bayesian over linear methods to be 2-7%. Another study conducted by Harris *et al.* (2008) reported approximately 2-3% greater reliabilities using Bayesian analysis compared to BLUP.

With Bayesian modeling, the specification of the priors is particularly important, and reflects assumptions made about the distribution of effects within the genome. For example, most models require an *a priori* specification of the proportion of markers ( $\pi$ ) that do not influence a particular trait. For BayesA, all markers are fit, as they are in a

GBLUP analysis, however, the ASE variances are assumed to be heteroscedastic (unlike GBLUP) and these are estimated in the analysis using Markov chain Monte Carlo (MCMC) methodology. ASE variances for BayesB are treated identically to those for BayesA; however when  $\pi > 0$ , BayesB assumes that a known proportion of markers, which is specified *a priori*, have no effect ( $0 \leq \pi < 1$ ) on the trait. This approach allows for non-influential marker effects to be shrunk to zero, theoretically minimizing a source of noise in the analysis that is caused by having many more markers than data points with which to estimate them. In the special case of BayesB where  $\pi = 0$ , the analysis is equivalent to BayesA. A third type of Bayesian analysis, BayesC, assumes that all markers have a constant variance (like GBLUP, except that when  $\pi > 0$ , ASE for non significant SNP effects are again shrunk towards zero), but assumes that only a fraction of markers influence a trait, similar to BayesB. The special case of BayesC where  $\pi = 0$  is roughly equivalent to GBLUP, because the ASE variances are constant for all SNPs included in the model. Certain Bayesian methodologies, such as BayesC $\pi$ , allow the estimation of the proportion of informative markers ( $1-\pi$ ) from the analyzed data.

#### *Within-Breed Analyses*

One of the first studies performed on field captured data was performed in Holstein bulls by de Roos *et al.* (2007). EBVs and genotypes for 32 loci across chromosome 14 (including the causal mutation for DGAT) were obtained for 1,300 bulls. The analysis compared results from BLUP (pedigree-based) and GS using a haplotype-based Bayesian approach to evaluate molecular breeding value (MBV) accuracies for fat

percentage. The MBV accuracies were 0.75 compared to the BLUP accuracies of 0.51, which was concordant with the results of simulation studies that noted higher accuracies when employing GS methodologies as compared to pedigree-based BLUP.

Su *et al.* (2010) reported results from a GS study comprised of 3,330 Danish Holstein bulls genotyped for 38,134 SNPs. Bayesian estimation methods resulted in reliabilities of approximately 0.49 to 0.73 based upon cross-validation. For the 18 traits analyzed in the study, reliabilities were 0.13 higher than the reliabilities of parent average EBVs.

Luan *et al.* (2009) used 18,991 SNPs genotyped in 500 Norwegian Red bulls to evaluate differences in GS using GBLUP, BayesB and a mixture model. A high correlation was observed between trait heritability and the corresponding trait average EBV accuracy. They concluded that GBLUP was superior in this population and found that accuracies from all analyses ranged from between 0.12 and 0.62.

One of the most recent studies (Saatchi *et al.* 2011) examined accuracies of MBVs in 3,570 American Angus bulls using a K-means methodology to cluster animals for cross-validation. K-means was used to maximize the within-group relatedness and minimize the relationships between animals in different clusters to evaluate accuracies with minimal modeling of linkage effects. Five clusters were formed and training was performed in four clusters and validation in the fifth, recursively, until all possible combinations were exhausted. Accuracies averaged 0.44 with a range of 0.22 to 0.69 for all 16 analyzed traits. Random allocation of animals to training and validation

populations yielded predictions with accuracies ranging from 0.38 to 0.85 (mean = 0.65), which reflects an increase in the extent of pedigree relatedness between members of the two populations. Training on ancestral animals and validating in the youngest animals yielded accuracies that were intermediate to K-means clustering and random allocation.

Regardless of the method, one of the largest barriers to the adoption of GS within an individual breed association in the beef industry is the lack of high-density SNP data on large numbers of animals. For the best results, a large training population of animals with high-accuracy EPDs (or other high-quality phenotypic data) must be assembled and genotyped. This is a difficult proposition in the beef industry, both in terms of expense and lack of sufficient numbers of animals within a breed. One solution to this problem is to pool animals across breeds (deRoos *et al.* 2009) to estimate marker effects and applying those prediction models to diverse breeds of animals.

### **Across-Breed Genomic Selection**

While pooling animals across breeds to train prediction models may solve issues related to inadequate numbers of samples, a number of other considerations arise when utilizing this approach. First, patterns of LD vary across breeds and influence the ability to detect the same QTLs in multiple breeds. To estimate marker effects across breeds, a high level of LD must exist between any particular causal mutation and a SNP in the marker panel; thus the predictive power of this marker panel is limited by the population history and the design of the SNP assay. This is largely influenced by the degree of difference in allele frequencies of the SNPs included in the analysis and the causal

mutations. Most SNPs in commercialized SNP panels are selected because they are common variants within the populations used in the SNP discovery process. The amount of LD between any particular locus and a QTL is limited by the difference in minor allele frequency between the loci (Kizilkaya *et al.* 2010). As a result, common variants can only exhibit very high levels of LD with other common variants and individually cannot detect rare QTL. In order to model rare variants, more low frequency SNPs must be included in the SNP assays. Additionally, the linkage phase must persist across all breeds in the analysis (Goddard and Hayes 2007) so that the directionality of the ASE is the same for all breeds. The Bovine HapMap Consortium (2009) showed that SNP allele phase relationships are preserved over only approximately 10 Kb across different breeds of cattle, indicating the need for at least one SNP every 10 Kb to tag every LD block in the genome within the analysis. Realistically, higher density SNP chips such as the 600-800K HD assays now being marketed by Affymetrix and Illumina will probably be needed for across-breed analyses to be effective across many breeds.

An additional complication arises when considering the predictive power of the model over time. Since LD is directly influenced by the distance between markers via the frequency of recombination events between loci, only markers that are in high LD and in close proximity to a causal mutation can be expected to preserve their informativeness due to minimizing the likelihood of crossover events within the genome as generations advance. Predictive power of the assay is also determined by the extent of familial relationships between animals in the training and validation populations (VanRaden *et al.* 2009). VanRaden *et al.* (2009) partitioned bulls according to their year

of birth (older animals were used in the training set and younger in the validation) which maximized the extent and magnitude of pedigree relationships between the training and validation populations. They discovered that this approach increased the genetic variance explained by the models in the validation population, which suggests that the SNPs are not only capturing LD relationships between markers and QTL, but are predicting breeding value based upon the extent of identity by descent to animals in the training population. Thus, if we cannot detect rare variants with the current SNP panels utilizing LD, it might be possible to capture their effects based on linkage effects which detect the extent of shared haplotypes among individuals. However, effects that are estimated due to familial relationships will decay more rapidly over time than that the break down in LD in an entire population (Habier *et al.* 2007, Solberg *et al.* 2009) due to the halving in identity by descent that occurs with each meiosis that separates individuals in a pedigree. The extent of these effects on the accuracy of genomic prediction can be estimated, but requires the analysis of several generations of animals to determine the contribution that is due to LD alone (Habier *et al.* 2007). With 50K SNP data, it is likely that routine re-training to include newly genotyped sires will be necessary to produce MBVs with the highest accuracies by increasing the relatedness between the training and validation populations (Saatchi *et al.* 2011).

When across-breed GS models are attempted, one important consideration is the method used to partition the training and validation populations. Studies have found that training in one breed and predicting in another breed is mostly ineffective with accuracies in the validation population tending to be quite low (Harris *et al.* 2008). If SNP effects

cannot be accurately estimated in one breed and applied to another, other strategies for the partitioning of training and validation populations must be considered and tested for their efficacy.

A study performed by Kizilkaya *et al.* (2010) showed that training in multibreed (encompassing both purebred and crossbred animals) and prediction in purebred populations is less effective than training in purebred and predicting in multibreed populations. This effect is most likely due to the greater persistence of LD in purebred populations, compared to the shorter LD blocks seen in crossbred and multibreed populations.

A study from Iowa State University (Toosi *et al.* 2010) evaluated different methods of partitioning populations for training and validating GS models using purebred, admixed, and crossbred populations using simulated data. They demonstrated that training and validating within the same breed produced the highest accuracies of all tested strategies, without exception. While training in admixed populations produced similar accuracies to training and validating in the same purebred populations, a 46% reduction in accuracy was observed when validating in a breed that was not included in the training set, which supports the results from Harris *et al.* (2008). A smaller decrease (35%) in accuracy was observed when training occurred in a crossbred (F<sub>1</sub>) population and validation was performed in a purebred population that was not represented in the breed composition of the crossbred training animals. Decreases in accuracy from training in three- and four-way crosses were approximately 10%. One important result is that

training and validating in a crossbred population increased accuracy by 11% compared to training in the purebreds and validating in the crossbred populations.

Toosi *et al.* (2010) also examined the effect of increasing the marker density in known QTL regions on the accuracy of prediction. They deduced that high marker densities were essential when the training population had a small contribution from the breed used for the validation population. They also noted that a larger sample size may be needed for a multi-breed training population to obtain an accuracy that is comparable to that of a purebred population due to the fact that a larger number of effects (both across-breed and breed-specific effects) need to be estimated.

It has been suggested that utilizing the time since divergence, or the genetic distance, between breeds might be an effective method to partition training and validation populations for across-breed GS. In this approach, animals would be selected from different populations to increase the amount of genetic variation present in the training population and animals from all breeds would be proportionally represented in the training population according to their distinctness. Toosi *et al.* (2010) examined the effect breed divergence on the accuracy of GS predictions in a simulated dataset. When different populations were used for training and validation, reducing the time since divergence greatly increased the accuracy of the prediction. They also noted that training in admixed (simulated to contain many breed combinations similar to commercial cattle where sires are purebreds mated to dams with a heterogeneous breed composition) rather than crossbred populations resulted in a greater accuracy regardless of time since

divergence. This is likely due to a larger number of recombination events occurring between markers and QTL as the time since divergence increases, which can break down and, in some cases, even reverse the phase of prevalent alleles present in the ancestral haplotypes that predate breed formation.

These results were mirrored in a study published by de Roos *et al.* (2009) who simulated two cattle populations with different divergence times and examined the reliabilities of GS models. Their training sets comprised 1,000 individuals from population A and different subsets of a population B. When individuals from population B were omitted from training, the reliabilities of the predictions for population B were up to 0.77 lower than those for population A. Moreover, this effect was most severe when the divergence time between breeds was greatest. Adding individuals from population B into the training set resulted in reliabilities close to the same level as for population A, provided that the marker density was sufficient for marker-QTL phase relationships to be preserved across both populations.

Few attempts at building across-breed GS models have been documented in the scientific literature. McClure *et al.* (2012) estimated allele substitution effects for Warner-Bratzler Shear Force (WBSF) both within and across five breeds of cattle in the National Cattlemen's Beef Association (NCBA) Carcass Merit Project (CMP) using GBLUP incorporating a GRM. The SNPs in this study were from the Illumina BovineSNP50 Beadchip and but were augmented with a custom panel of 96 SNPs around  $\mu$ -calpain and calpastatin which are known to be large-effect QTL influencing WBSF.

The GBLUP estimates of ASE across-breeds were moderately correlated with results from the within-breed analyses (0.31-0.66), indicating that this across-breed model should be reasonably effective for selection within the individual breeds used in this study.

Simulation studies have shown that divergence between breeds is a factor in the predictive ability of GS models, but that effective prediction models can be built when all of the breeds of animals present in the validation population are represented in the training population in sufficient numbers. Even so, the challenges of building across-breed GS models are daunting and very few studies have examined the principles and methods discussed above using real-world cattle populations. Haplotypes with strong LD ( $r^2 \geq 0.7$ ) among pairs of loci are significantly shorter in length in admixed and crossbred populations compared to purebred populations (Toosi *et al.* 2010) which makes across-breed modeling difficult. However, simulation studies have indicated that while it may be harder to achieve superior across-breed GS model predictions, when good models are obtained, they are more effective across many breeds and accuracies persist over a longer period of time. While the building of across-breed GS models is fraught with difficulty, they would provide invaluable tools for selection by both purebred and commercial cattle producers and should be ardently explored.

### **Implementation of Genomic Selection in the Beef Industry**

The dairy industry in the United States has adopted the use of GRMs in their genetic evaluations to include DNA marker information into their NCE. In the beef

industry, however, savings in progeny testing schemes do not offset the added cost of genotyping assays; therefore, a strategy was devised to reduce cost of technology adoption for beef producers. The aforementioned methods are only feasible with large numbers of markers evenly spread throughout the genome, which is not economically sensible for the majority of beef producers until the costs of high-density genotyping assays considerably decrease. Most DNA marker panels commercialized for the beef industry to date are small (i.e., 384 markers in a multi-trait panel) and consist of SNPs deemed to be the most predictive in research populations. Regardless of the panel's size, in the beef industry, the results from these marker panels are summarized into a single marker score or molecular breeding value. These MBVs provide additional information on animals with low EPD accuracies, but, unlike EPDs, are not inclusive of information on all of the genetic variation underlying a trait. Therefore, the most constructive solution is to incorporate the genotypes or MBVs into existing NCE and publish a single EPD which combines both sources of information to increase the EPD's accuracy. Because MBVs are reported in lieu of making the SNP genotypes available to the breed associations, a method was developed (Kachman 2008) to include MBV information into NCE in the form of genomic-enhanced EPDs (GE-EPD). This method treats the MBV and the observed data as genetically correlated traits in a multi-trait animal model.

The American Angus Association implemented this methodology for their carcass EPD NCE in fall 2009. GE-EPDs for carcass weight, marbling score, ribeye area and fat thickness are calculated for approximately 2 million animals and are updated on a weekly basis to account for incoming genomic information (Northcutt, 2010). This allows the

rapid accumulation of data from marker panels, carcass phenotypes, and ultrasound data as well as the generation of carcass EPDs for dams, which previously had no records, as soon as the MBV is processed for their progeny (Northcutt 2010). Young animals with MBVs and no ultrasound scan records will achieve EPD accuracies of 0.28 to 0.38 depending on the trait, compared to a parent average EPD accuracy of 0.05 for animals without an MBV or scan record (Northcutt 2010). Genomic information has now been included into Angus evaluations for docility, residual gain, and growth traits in addition to the carcass evaluation (American Angus Association 2011). The American Simmental Association recently released their first run of GE-EPDs and the American Hereford Association will follow suit in June 2012. Both of these breed associations are using a blended index approach. Most of the other high-profile breed associations in the US are developing the infrastructure for handling genomic data and plan to implement these data into their evaluations within the next few years using either the correlated trait method (Kachman 2008) or through a blended index approach.

### **Future of Genomic Selection in the Beef Industry**

The use of genomics within the US beef industry will precipitate rapid advancements in methodologies over the next few years. Scientists will likely develop new methods for the inclusion of genome sequence data, the addition of causal mutations discovered from genomic and transcriptomic sequencing as well as epigenetic studies. As the costs of genotyping and sequencing continue to decrease, it is expected that breed associations will move towards the generation of their own data and the assembly of

breed-specific training populations with high density SNP and sequence data. As the methodology for implementing across-breed GS improves and more breed associations include genomic information into NCE, these technologies will have a transformational impact on beef production in the US.

## CHAPTER II: SUMMARY OF OBJECTIVES

The studies contained in this dissertation were structured around three main objectives. The first objective was to use real-world commercial cattle populations to evaluate results from simulation studies in the literature which examine several methods for the partitioning of training and validation populations in groups of multi-breed cattle (Toosi *et al.* 2010, Kizilkaya *et al.* 2010, de Roos *et al.* 2009). Commercial cattle, which are largely crossbred and admixed populations, make up the largest sector of the beef cattle industry and are the largest source of genetic variation. Currently, no system exists to generate EPDs or EBVs which would allow commercial cattlemen to make informed selection decisions on crossbred animals within their herds. Their selection decisions are primarily driven only by data available on the herd bulls or replacement females that they purchase from breeders of registered seedstock and the phenotypic performance of progeny in their herds. Providing a useful selection tool for this sector of the industry, which directly provides the largest proportions of beef product to consumers and where genetic variation is the greatest, would be valuable.

The second objective was to evaluate the predictive power of the BovineSNP50 BeadChip for use in building across-breed genomic selection models for the estimation of

MBVs in the beef industry and to evaluate the persistence of MBV accuracies over time. This was evaluated using two separate populations, one which contained purebred and crossbred commercial cattle from the Carcass Merit Project, and a second population which consisted of a small group of Simmental and SimAngus calves which were born approximately 10 years (~2 generations) after the CMP population was created. First, we evaluated the predictive power of three different types of Bayesian models in the CMP population for a variety of economically relevant carcass traits. This approach allowed us to compare and contrast the predictive ability of models both in the presence and absence of genes of large effect, which helped determine which models were most appropriate for specific genomic architectures as they relate to the size of gene effects for any particular trait. In addition to the value to the commercial beef industry as previously discussed, this may provide an incentive and the ability of smaller purebred breed associations to combine efforts and more affordably generate genomic predictions for their members. It will also certainly be informative for breed associations currently performing NCE for hybrid animals. Our second population of animals allowed us to validate the prediction models that were built in the CMP population for one trait (Warner-Bratzler Shear Force) in the second independent population which was stratified in time. Quantification of the decay in prediction accuracy over time will be especially valuable for determining how often prediction models will need to be retrained. The inclusion of new data from more contemporary populations in retraining should prevent the decay in prediction accuracy of these models over time.

This third objective was to identify areas of the genome which harbor mutations influencing ERTs in beef cattle populations. The CMP population allowed the discovery of QTL for a variety of novel traits (WBSF and sensory panel traits in the CMP) as well as for established traits for which EPDs exist, but for which relatively few QTL have been identified (HCW, FT, YG, MARB). These results not only provide a basis for QTL validation for future QTL studies, but concurrently serve as a discovery population for future studies targeting the identification of causal mutations for ERT in beef cattle.

CHAPTER III: EXPLORATION OF METHODS FOR PARTITIONING  
TRAINING AND VALIDATION POPULATIONS TO IMPROVE GENOMIC  
SELECTION IN CROSSBRED AND PUREBRED BEEF CATTLE

**Summary**

There have been numerous studies which have examined the accuracy of genomic selection both within and across purebred populations in the beef and dairy industries. However, the examination of the accuracies of MBV prediction models has been less prevalent in crossbred and admixed cattle populations. We used a population of 3,240 crossbred steers and heifers of mixed ancestry which included five different breeds from the National Cattlemen's Beef Association (NCBA) Carcass Merit Project (CMP) to predict molecular breeding values (MBVs) for five economically important traits in beef cattle. Realized accuracies ranged from 0.4 to 0.77, depending on the trait, number of records, and utilized model (BayesB, BayesC, or BayesC $\pi$ ). Advantages of mixture models (BayesC $\pi$ ) and models which allow for unequal allele substitution effect variances (BayesB) were observed for some traits which are affected by genes of relatively large effect, such as Warner-Bratzler Shear Force. When the architecture of the trait fit the infinitesimal model, the difference in prediction accuracy between models was

not significant. External validations of WBSF predictions in populations separated by ~10 years showed an approximately 90% reduction in prediction power.

Key Words: beef cattle, across-breed genomic selection, BovineSNP50, accuracy

## **Introduction**

National Cattle Evaluation (NCE) has been used within the US beef industry for over four decades (Willham 1993) and is based upon mixed model methodologies outlined by Henderson (1963). To generate effective genetic predictions, these models require extensive ancestral pedigree information which is used to form a relationship matrix to account for the extent of identity by descent among phenotyped individuals. NCE has provided an invaluable tool for purebred breeders to increase genetic gains in many economically relevant traits (ERT).

Meuwissen *et al.* (2001) proposed a methodology called genomic selection (GS), which had the potential to revolutionize NCE. Theoretically, GS uses many markers evenly spaced throughout the genome to model QTL causal variants and predict the genetic merit of individuals for a variety of ERT, even if only a small subsection of the population has been both genotyped and phenotyped. With the advent of high-density genotyping platforms which can affordably generate many thousands of genotypes and because the cost of beef production is rapidly rising every year, GS has become increasingly attractive. Many studies in the literature have examined genomic best linear unbiased prediction (GBLUP) which utilizes a genomic relationship matrix in lieu of a pedigree relationship matrix in the mixed model analysis to estimate the merit of

genotyped animals that do not have phenotypes. This approach is equivalent to fitting each individual SNP as a random effect (Garrick 2007) in a mixed model analysis and then estimating the MBV of an individual as the sum of contributions from each fitted SNP.

Recent studies (i.e., Saatchi *et al.* 2011; Habier *et al.* 2011) have utilized a Bayesian framework to build MBV prediction models. While some studies have found a slight advantage of nonlinear Bayesian models over GBLUP (de Roos *et al.* 2007, Harris *et al.* 2008, Hayes *et al.* 2009), many of these studies were conducted in dairy populations, for which traits may differ for their underlying QTL architecture. Some results suggest that the advantage conferred by using Bayesian analyses may be due to its increased ability to model the architecture of QTL effects within the genome, especially for traits which have a gene of large effect (VanRaden *et al.* 2009). One advantage of Bayesian methodologies is that we can use prior knowledge about the distribution of QTL effects in the genome for each trait to define prior values for  $\pi$ , the proportion of markers in the analysis that do not influence the trait, which will assist in characterizing variation within the genome more accurately.

All of the previously described methodologies have been shown to be relatively effective when analyses are constrained to one breed of cattle (i.e., Saatchi *et al.* 2011). Within the US dairy industry, where an overwhelming majority of animals are of Holstein lineage, this constraint has little bearing on the efficacy of GS methodologies as long as the resulting predictions are effective for Holsteins. Within the US beef industry,

the majority of purebred animals are of Angus descent; however, there are dozens of beef breeds that are admixed within the commercial population, which makes within-breed GS models ineffective outside of the purebred sector. Currently, genetic evaluation is only performed in the purebred sector, but increasingly breed associations are beginning to perform NCE on British-Continental hybrids (i.e., LimFlex, Balancer, SimAngus, etc.), whose population architectures will more closely resemble crossbred and admixed populations than their purebred analogues. There is increasing interest by breed associations and the AI companies in performing genetic evaluation in the commercial sector due to the increased amounts of genetic variation and greater number of phenotypes available for analysis.

Pursuing across-breed GS has many theoretical advantages, including an increased availability of samples and the ability to use animals in commercial and non-pedigreed populations. Simultaneously, it also has a few drawbacks that must be considered. First, consideration must be given to the patterns of LD which vary across breeds and may influence the ability to detect the same QTLs within multiple breeds. To estimate marker effects across breeds, a high level of LD must exist between any particular causal mutation and a SNP in the marker panel, thus the predictive power of a marker panel is limited by the population history and the design of the SNP assay. LD is primarily influenced by the extent of difference in allele frequencies between the SNPs and the causal mutations they are attempting to detect. Additionally, the linkage phase between SNP and QTL alleles must persist across all breeds in the analysis (Goddard and Hayes 2007) so that the directionality of the effect is the same for all breeds.

An additional complication arises when considering the prediction accuracy of the model over time. Since LD is influenced by the distance between markers via the frequency at which recombination will occur, only markers in high LD and in close proximity to a causal mutation can be expected to retain their informativeness in advanced generations. Prediction accuracy is also determined by the amount of linkage, or familial relationships between animals, that is also being detected by the genotypes, as demonstrated by VanRaden *et al.* (2009). Thus, while we are unlikely to detect rare causal variants with the current SNP panels utilizing LD due to the fact that these SNP panels primarily consist of common variants, it might be possible to model them using linkage effects which capture extensive haplotype sharing among close relatives. However, effects that are modeled based upon on linkage will decay more rapidly in time than those modeled strictly due to LD because the rate of decay of LD for an entire population is less than that of linkage (Habier *et al.* 2007, Solberg *et al.* 2009). The extent of impact on the accuracy of genomic predictions due to these effects can be estimated, but requires the analysis of several generations of animals to partition the effects (Habier *et al.* 2007). With 50K SNP data, it is likely that routine re-training to include newly genotyped sires and progeny will be necessary to produce MBVs with the highest accuracies to ensure that the relatedness between the training and validation populations is maximized (Saatchi *et al.* 2011).

Various methods for improving the accuracy of across-breed genomic selection models have been proposed, but few have been tested in existing beef cattle populations. Studies in the literature have illustrated that training in one breed and predicting in

another breed is ineffective and resulting accuracies in the validation population tend to be quite low (Harris *et al.* 2008). A study performed by Kizilkaya *et al.* (2010) used simulated data to demonstrate that training in multibreed populations (encompassing both purebred and crossbred animals) followed by prediction in purebred populations is less effective than training in purebred and predicting in multibreed populations. This effect is presumably due to the greater persistence of LD in purebred populations, compared to the shorter LD blocks that are found in crossbred and multibreed populations.

A study performed by Toosi *et al.* (2010) evaluated different methods of partitioning populations for training and validating GS models using purebred, admixed, and crossbred populations using simulated data. Their research confirmed that training and validating within the same breed produced the highest accuracies. However, training in admixed populations performed similarly to training and validating in the same purebred populations. However, a 46% decrease in accuracy was observed when validating in a different breed to that in the training set, which supports the results of Harris *et al.* (2008). A smaller decrease (35%) in accuracy was observed when training was in a crossbred ( $F_1$ ) population and validating occurred in a purebred population that was not represented in the breed composition of the crossbred animals used for training. One important result to note is that training and validating in a crossbred population increased accuracy by 11% compared to training in the purebreds and validating in the crossbred populations. Toosi *et al.* (2010) also noted that a larger sample size may be needed in a multi-breed training population to achieve an accuracy that was comparable

to that of a purebred population due to the estimation of a larger number of effects (both across-breed and breed-specific effects need to be estimated).

It has been suggested that the time since divergence, or the genetic distance between breeds, might be effectively used to partition training and validation populations in multibreed GS applications. In this approach, animals would be selected from different populations to increase the amount of genetic variation present in the training population, and animals from all breeds would be represented in the training population. Toosi *et al.* (2010) examined the effect of time since divergence on the accuracy of GS predictions in a simulated dataset. When different populations were used for training and validation, reducing the time since divergence greatly increased the accuracy of the prediction achieved in the validation population. This is likely due to the increase in the number of recombination events with increasing time which can break down and, in some cases, reverse the phase relationship between SNP and QTL alleles relative to that in the ancestral haplotypes.

These results supported the findings of de Roos *et al.* (2009) who simulated two cattle populations (A and B) with different divergence times. This study determined that when individuals from population B were omitted from training, the reliabilities of the predictions for population B were up to 0.77 lower than for population A. Additionally, this effect was most severe when divergence times between the populations were greater. Including individuals from population B into the training set resulted in reliabilities that

were close to those for population A, provided that the utilized marker density was sufficient for phase relationships to be preserved in both populations.

Haplotypes with strong LD ( $r^2 \geq 0.7$ ) are significantly shorter in admixed and crossbred populations as compared to purebred populations (Toosi *et al.* 2010) which reduces the probability that SNPs in the assay will be in strong LD and with preserved phase relationships with QTL alleles in the crossbred animals. However, simulation studies have suggested that while it may be difficult to achieve highly accurate across-breed GS model predictions, when these models are achieved, they are effective across many breeds and prediction accuracy persists over a longer period of time. While the development of across-breed GS models appears to be difficult, their value as breeding tools for both purebred and commercial cattle producers justifies their pursuit.

Nevertheless, very few studies have examined the development of multibreed MBV prediction models using field data. Consequently, the objectives of this study were to use an admixed population of commercial steers and heifers from the CMP to evaluate the effectiveness of across-breed MBV prediction models in beef cattle for a variety of economically relevant carcass traits. Alternative methods for partitioning samples into training and validation populations including genetic distance, genomic relationships and random allocation, were examined using three different Bayesian prediction models to characterize differences in MBV accuracy. Additionally, WBSF phenotypes and whole genome SNP genotypes were obtained for a contemporary population of Simmental

animals developed approximately 10 years after the CMP animals and the developed MBV prediction models for WBSF were evaluated for their persistence in accuracy.

## **Materials and Methods**

### *Population*

The National Cattleman's Beef Association instituted the Carcass Merit Project (CMP) in 1998 to address issues of consumer dissatisfaction with beef eating experiences. Samples (n = 3,360) were chosen from the CMP comprising five different breeds of taurine cattle (Angus n = 660, Charolais n = 702, Hereford n = 1192, Limousin n = 285, and Simmental n = 521). Animals were preferentially chosen for genotyping from all available samples based on two factors: availability of observations for Warner-Bratzler Shear Force and completeness of carcass records for all other traits. The design of the CMP and data collection process was described by Minick *et al.* (2004). All of the animals enrolled in the CMP were sired by bulls from the respective breed classifications (Angus, Charolais, Hereford, Limousin, and Simmental) and dams were from commercial herds. Angus- and Hereford-sired CMP progeny were assumed to be purebred due to the sires being mated to commercial Angus and Hereford dams, however, the Continental-sired progeny are largely crossbred due to the mating of Limousin, Simmental and Charolais bulls to commercial Angus cows (McClure *et al.* 2012; Minick 2004).

### *Phenotypic Data*

Phenotypic data collection procedures were also detailed by Minick *et al.* (2004). Briefly, USDA personnel recorded marbling score (MARB), hot carcass weight (HCW), fat thickness at the 12<sup>th</sup> and 13<sup>th</sup> rib interface (FAT), and ribeye area (REA) between 24 and 48 hours postmortem. Steaks were vacuum packaged and aged for 14 days before the collection of Warner-Bratzler Shear Force (WBSF) records at Kansas State University. Muscle, DNA, and white blood cells (WBC) were obtained for each animal from Texas A&M University with the permission of the sample owners (American Angus Association, American Hereford Association, American Simmental Association, American International Charolais Association, and the North American Limousin Foundation). WBC samples were obtained at weaning and muscle samples were obtained at harvest as carcass data was recorded and steaks were harvested for WBSF analysis. Paternity and identification (matching DNA profiles from WBC and muscle samples) was previously performed as part of the CMP protocol for all samples for which we received DNA. Any animals found to have pedigree or identification errors were removed and DNA was re-extracted from muscle samples at the University of Missouri to insure the greatest probability that the produced genotypes and phenotypes would originate from the same animal, regardless of paternity, which was not considered in our analysis. Genomic DNA was extracted from 2,940 muscle samples by proteinase K digestion followed by Phenol:Chloroform:Isoamyl alcohol extraction and ethanol precipitation (Sambrook *et al.* 1989). The remaining 420 samples were not re-extracted and used DNA provided by Texas A&M University since these samples had successfully

passed identification and paternity verification. The number of phenotypes available for analysis for each breed and trait is in Table 3.1.

### *Genotypic Data*

All CMP samples were genotyped using the Illumina BovineSNP50 BeadArray (Matukumalli *et al.* 2009), which assays 54,790 SNPs. A custom Illumina GoldenGate assay (additional details are in McClure *et al.* 2012) was used to generate genotypes for an additional 96 putative SNPs located within 186 kb of *CAST* and *CAPNI*. All genotypes were called in the Illumina GenomeStudio software. SNP locations were obtained using the UMD3.1 build coordinates (Zimin *et al.* 2009) and were filtered as a quality control measure. Filtering criteria included call rate  $< 0.89$  (to include all commercialized tenderness SNPs), and minor allele frequency (MAF)  $> 0.01$ , leaving 40,645 SNPs for analysis on 3,240 animals (Angus  $n = 651$ , Charolais  $n = 695$ , Hereford  $n = 1,095$ , Limousin  $n = 283$ , and Simmental  $n = 516$ ). FastPHASE v1.2.3 (Scheet and Stephens 2006) was used to phase all genotypes and impute the 0.89% of missing genotypes.

### *Approach*

We built across-breed genomic selection models for traits recorded as part of the CMP using three Bayesian methodologies (BayesB0, BayesC0, and BayesC $\pi$ ) that are implemented in the GenSel software package (Fernando and Garrick 2009), which was developed at Iowa State University and has previously been used for several studies in the literature (Saatchi *et al.* 2011; Habier *et al.* 2011). Each trait (WBSF; Ribeye Area,

Table 3.1: Number of phenotypes available for analysis for each breed and trait.

Trait <sup>1</sup>	Angus	Charolais	Hereford	Limousin	Simmental	Total
WBSF	651	695	1095	283	516	3240
REA	644	693	1090	276	510	3213
MARB	644	695	1095	276	53	2763
FT	611	693	1057	276	509	3146
HCW	644	695	1095	276	509	3219
%CL	644	695	1079	282	468	3168
YG	627	689	1095	249	510	3170

<sup>1</sup>WBSF, Warner-Bratzler Shear Force; REA, Ribeye Muscle Area; MARB, Marbling score; FT, Backfat Thickness; HCW, Hot Carcass Weight; %CL, Percent Cook Loss; YG, Yield Grade

REA; marbling score, MARB; fat thickness, FT; hot carcass weight, HCW, percent cook loss, %CL, and yield grade, YG) was analyzed using 160,000 iterations of Markov chain Monte Carlo (MCMC) and each model was parameterized using priors estimated as weighted means of the within-breed residual and additive genetic variances derived from a genomic best linear unbiased prediction (GBLUP) analysis (presented in McClure *et al.* 2012), except where otherwise noted. Because we possessed prior knowledge about the distribution of QTL effects within the genome, Bayesian methodologies were especially appropriate and this knowledge was used to define prior values for  $\pi$ , the proportion of markers in the analysis that do not influence a trait, which assisted in more accurately characterizing variation within the genome.

Under these models, marker allele substitution effects have normal priors which are conditional on the allele substitution effect variances. When these variances are individually estimated for each locus (BayesB), they have scaled inverse  $\chi^2$  priors (Habier *et al.* 2011). When  $\pi$  is estimated from the data (BayesC $\pi$ ), it is considered to have a noninformative uniform(0,1) prior distribution. For all other analyses,  $\pi$  was considered known and was specified within the analysis. Complete parameter prior values are presented in Table 3.2. Because there are many different Bayesian methodologies described in the literature, we will briefly detail the assumptions inherent to the models used in our analysis. First, BayesC where  $\pi = 0$  (BayesC0) was used due to its similarity to GBLUP (McClure *et al.* 2012). Like GBLUP, BayesC0 assumes that all allele substitution effect variances are equal and that all SNPs contribute towards predicting MBVs. BayesB where  $\pi = 0$  (BayesB0) was also utilized, and, like BayesC0,

Table 3.2: Starting priors for BayesC $\pi$ , BayesC0, and BayesB0 analyses.

Trait	Analysis	$\pi$	$V_a$	$V_e$
WBSF	BayesC $\pi$	0.99	0.416	0.624
	BayesC0	0	0.416	0.624
	BayesB0	0	0.16	0.55
REA	BayesC $\pi$	0.95	25.629	40.170
	BayesC0	0	25.629	40.170
	BayesB0	0	21	44
MARB	BayesC $\pi$	0.9	3500	2600
	BayesC0	0	3500	2600
	BayesB0	0	3500	2600
FT	BayesC $\pi$	0.95	0.092	0.063
	BayesC0	0	0.092	0.063
	BayesB0	0	0.011	0.136
HCW	BayesC $\pi$	0.9	373.06	571.19
	BayesC0	0	373.06	571.19
	BayesB0	0	610	615
YG	BayesC $\pi$	0.9	0.2049	0.2049
	BayesC0	0	0.2049	0.2049
	BayesB0	0	0.035	0.358

includes all markers into the prediction model. BayesB0 is identical to BayesA (Meuwissen *et al.* 2001), but differs from BayesC0 due to the fact that individual allele substitution effect variances are estimated for each locus using Metropolis-Hastings (MH) steps. Two MH iterations were used in each step of the Markov chain Monte Carlo sampling to estimate these variances. Finally, we performed a BayesC $\pi$  analysis. Like BayesC, BayesC $\pi$  fits a constant allele substitution effect variance, which is estimated using Gibbs steps (Habier *et al.* 2011). However, this variance is shrunk according to the frequency of each SNP's inclusion into the model for each MCMC chain, which results in allele substitution effect variances that are unique to each SNP. In the BayesB0 and BayesC0 analyses,  $\pi$  was treated as a known constant (0), but in BayesC $\pi$ ,  $\pi$  was estimated from the data.

The general model fit to the data was as follows:

$$y_i = \sum_{j=1}^k z_{ij}u_j + e_i$$

where:

$y$  = a  $n \times 1$  vector of phenotypes for a given trait with the mean and contemporary group effects removed,

$k$  = the number of marker loci used in the analysis,

$z_{ij}$  = the allelic state of animal  $i$  at marker  $j$

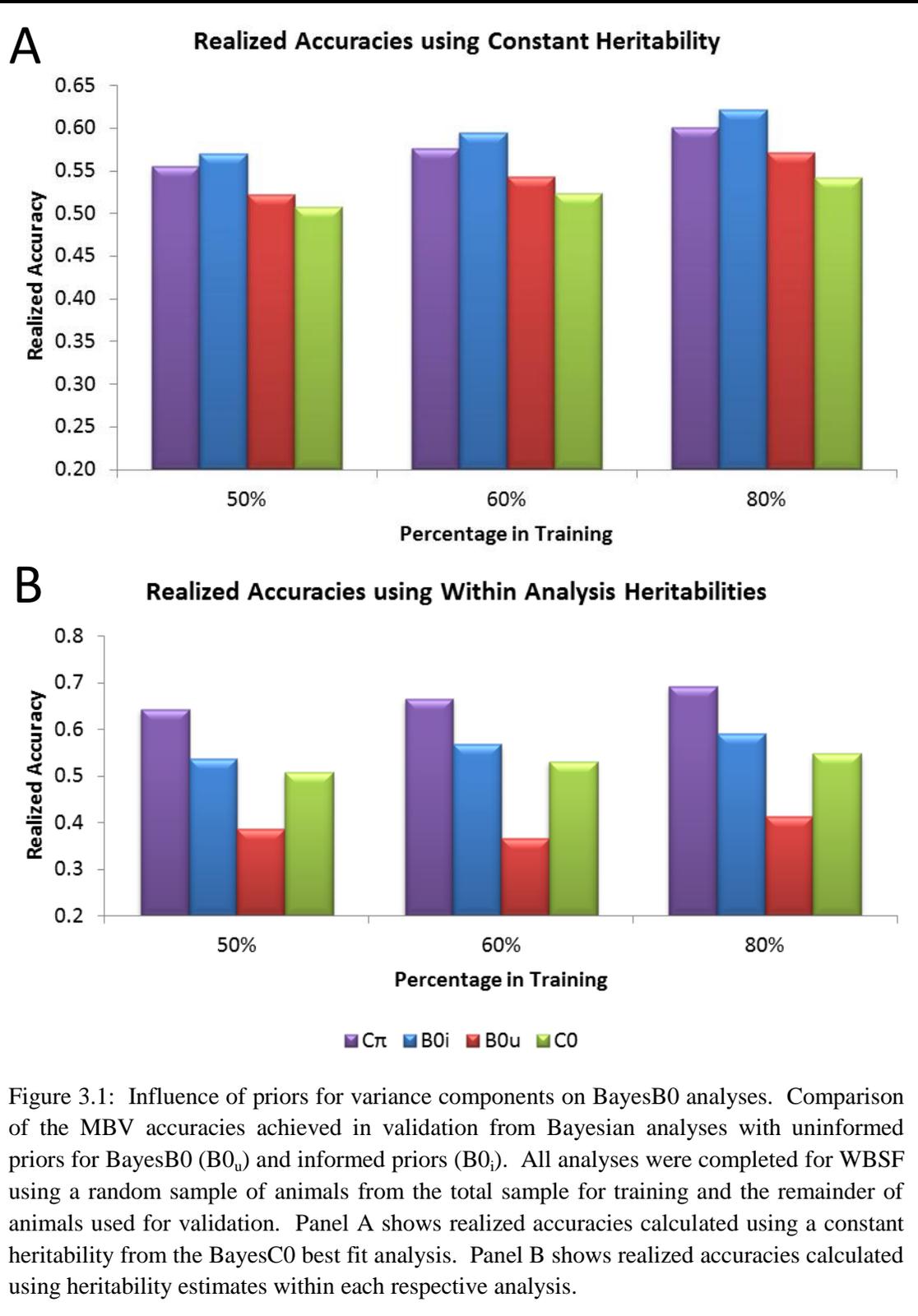
$u_i$  = the random effect for marker  $j$ , and

$e_i$  = the residual.

For these models, the random effect estimated for each marker was influenced by its likelihood of inclusion in the model ( $1-\pi$ ) thus  $u_i \sim N(0, \sigma_u^2)$  with probability  $1 - \pi$  or  $u_i = 0$  with probability  $\pi$ . The MBV for each animal in the validation set was obtained as the sum of all SNP allele effects (SNP posterior means over all post burn-in samples) over all  $k$  markers estimated during training.

Due to the increased sensitivity of BayesB analyses to priors for variance components (Habier *et al.* 2011, Gianola *et al.* 2009), prior values for the additive genetic and residual variances for all BayesB analyses were updated from weighted averages from a within-breed GBLUP analysis ( $B0_u$ ) to means from the posterior distributions of BayesC analyses ( $B0_i$ ) to compensate for this sensitivity. In the BayesC analyses, the data overwhelmed the priors allowing convergence of the variance component estimates, however, the BayesB0 analyses were sensitive to the prior values which necessitated the use of values obtained from the BayesC $\pi$  analyses (Figure 3.1).

Contemporary groups were defined by herd of origin, breed, sex, and harvest date and were modeled as a fixed effect in a single BayesC $\pi$  analysis including all animals, and observations were then pre-corrected for these effects to generate the phenotypes used for analysis. The inclusion of all records in a single analysis provided the maximum



amount of data with which to estimate these effects, however, the adjustment was made at a heritability that represented the pool of breeds and not any one breed.

Results from GenSel are reported as correlations between the validation population phenotypes and the molecular breeding values estimated using the model developed in the training set. Because our input variables were phenotypes and not deregressed breeding values, the reported correlation is not a direct measure of the accuracy of prediction as would have been the case in a purebred population using deregressed breeding values. To account for this difference, and the fact that estimated heritabilities varied between breeds and also the across-breed heritability varies due to the breed composition of the training set, realized accuracies were reported in most cases. Realized accuracies were calculated as  $\frac{r_{\hat{g},y}}{\sqrt{h^2}}$  due to the fact that  $\sqrt{h^2}$  is the largest value that the correlation between phenotypes and breeding values can achieve. Within breed realized accuracies were estimated using the correlation from the GenSel analysis and the estimated heritabilities from a within-breed GBLUP analysis utilizing procedures outlined in McClure *et al.* (2012). Within-breed heritabilities are reported in Table 3.3. Estimated heritabilities for percent cook loss, percent kidney, pelvic, and heart fat, and internal fat were not different from zero, so no further analyses were conducted with these traits.

#### *Genetic Distance Measures*

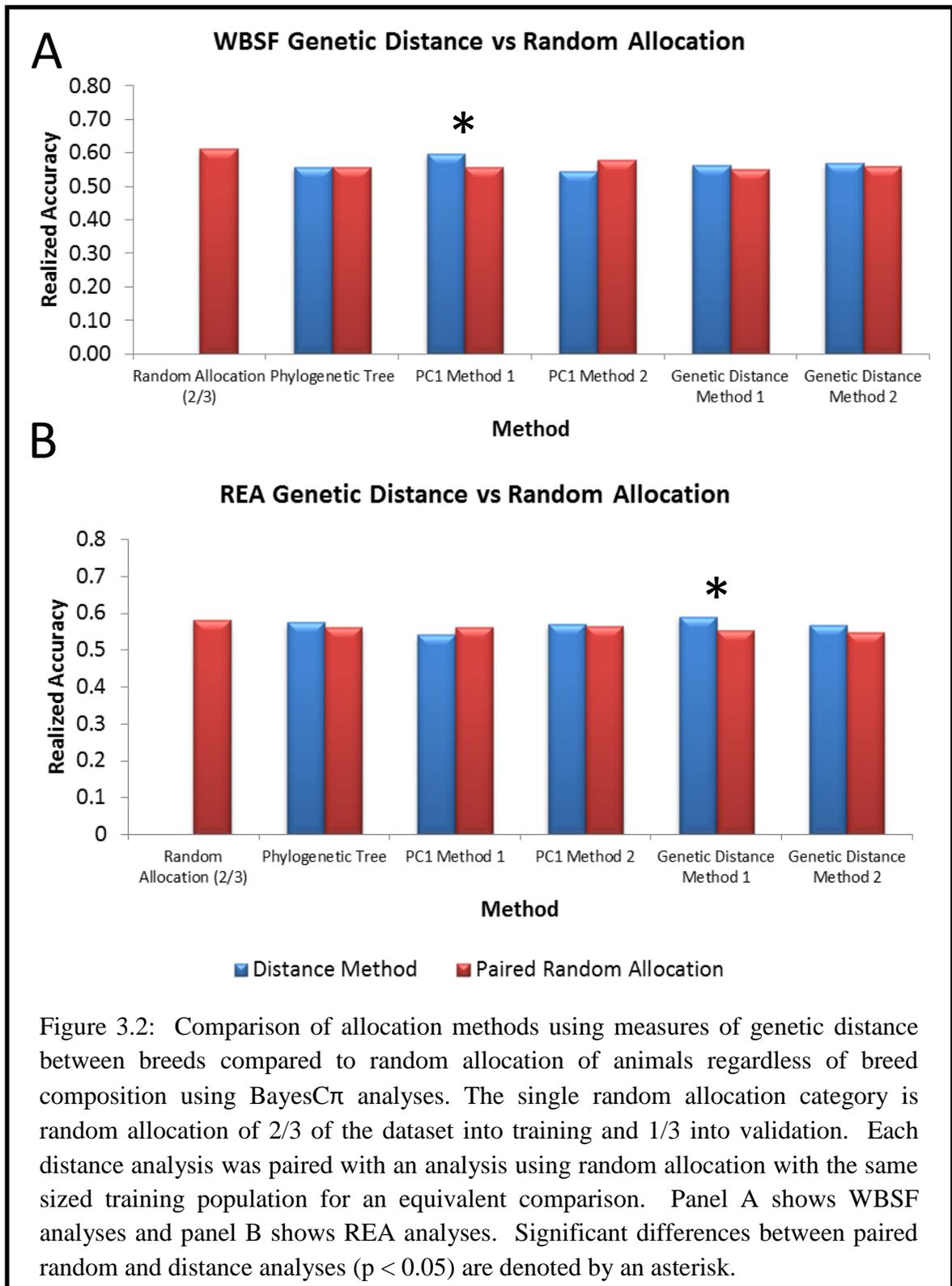
In simulation studies, animals are typically assigned to a population and mated for multiple generations to establish LD within the advanced generation population of

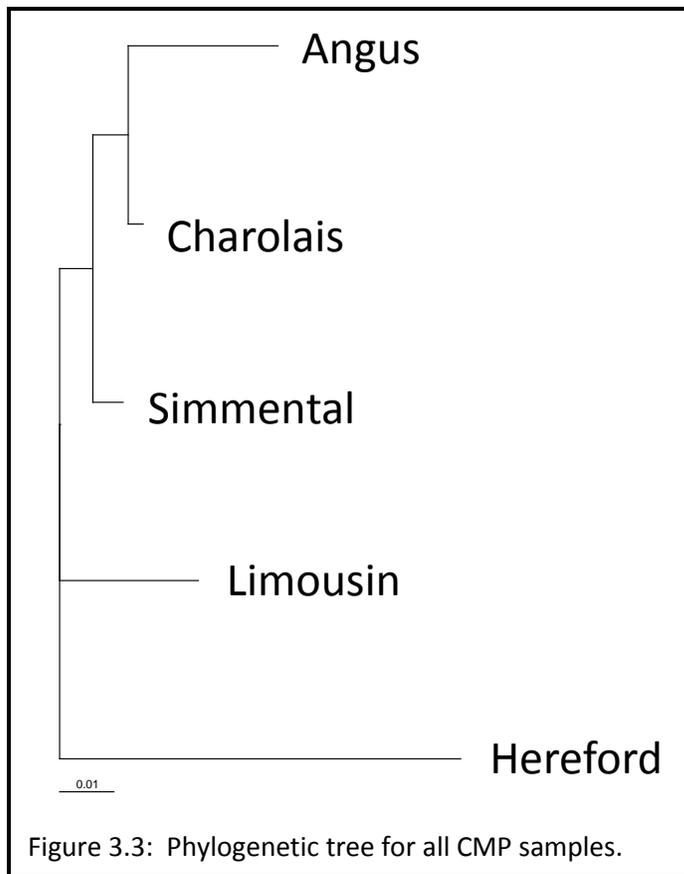
Table 3.3: Within breed heritability estimates derived from GBLUP.

Trait	Angus	Charolais	Hereford	Limousin	Simmental
WBSF	0.52	0.46	0.17	0.09	0.08
REA	0.61	0.21	0.23	0.55	0.29
MARB	0.51	0.57	0.49	0.41	0.87
FT	0.40	0.50	0.28	0.94	0.65
HCW	0.31	0.65	0.37	0.51	0.11
YG	0.39	0.50	0.23	0.79	0.45

animals. In these analyses, the exact time since the populations diverged is controlled and can be precisely determined. When using real-world data on commercial cattle, the time since breed divergence can only be estimated using an unascertained sample of genetic variation – which is not available on the BovineSNP50 assay. In light of these complications, we utilized several different methods to estimate the time since divergence of the CMP breed groups. The estimated divergence between breeds was then used to determine the proportion sampled from each breed to include within the training set. For each specified proportion, 20 bootstrap replicates were performed to obtain samples of animals representing these breed proportions. While larger training sets are usually desirable, the small sample size for Limousin ( $n = 283$ ) limited the size of the training set for some analyses. In all cases where a genetic distance partitioning methodology was utilized, a second analysis using the same number of randomly allocated animals in the training set without regard for breed composition was used as a “control.”

Breed allele frequency data was used to generate a genetic distance matrix using GeneDist (part of the Phylip package) employing Reynold’s distance (Felsenstein 2005). To test the need for an outgroup species, allele frequencies were obtained from a Brahman population and were utilized in a second analysis. Method 1 is used to denote the analyses which excluded Brahman, and method 2 denotes those analyses that included Brahman (Figure 3.2). Additionally, genetic distance matrices were used to generate a neighbor-joining tree using Phylip (Felsenstein 2005; Figure 3.3). The addition of genotype frequencies for 99 purebred Brahman cattle did not change the tree topology and all further analyses using the phylogenetic tree excluded the Brahman.





Principal component decompositions of the genetic distance matrices with Brahman (method 2) and without Brahman (method 1) were performed in Matlab (Natick, MA).

#### *Random Allocation Measures*

When random allocation was used, animals were partitioned into training and validation sets at random regardless of breed. Because random allocation creates sampling variation in breed representation in the training population, 20 bootstrap replicates were performed for each training population. Animals were randomly assigned to either the training or validation populations using Matlab's (Natick, MA) random number generator, which was seeded by the CPU clock time to prevent identical assignments across bootstrap replicates. Because a substantial number of different assignment proportions were tested in numerous analyses, an empirical method was required to identify the "best-fit" analysis. Best-fit analyses were determined by finding the smallest coefficient of variation  $\left(\frac{\sigma}{\mu}\right)$  for the BayesC $\pi$  correlations over the 20 bootstrap replicate analyses only for analyses which utilized greater than 50% of the data in the training set (to ensure parameters were estimated accurately).

#### *Genomic Relationship Matrix Measures*

A genomic relationship matrix (GRM) was generated using the allele frequency method outlined by VanRaden *et al.* (2008). Allele frequencies varied considerably across the breeds used in this analysis, which partitioned the GRM into blocks according to the animal's breed and the off diagonal elements between breed blocks were negative. Thus,

GRMs formed from data within a single breed were used in this study. Animals were selected based on their within-breed genomic relationships so that the number of animals from any given breed in the final training set was proportional to the composition of the entire dataset. For example, the original dataset was comprised of 20% Angus animals, so the proportion of Angus animals in any GRM training set would be 20%, regardless of the size of the specific training set. Because these partitioning methods resulted in a single set of animals that best fit the specified criteria, bootstrap sampling was not performed. We attempted to maximize the relationships between animals in the training and validation populations (maxbw) while minimizing the within-group relationships within either the training (maxbwminwintrain) or validating populations (maxbwminwintest). Likewise, we minimized the relationships between the training and validation populations (minbw) while also maximizing the relationships within the training (minbwmaxwintrain) or validation (minbwmaxwintest) populations. Additional details pertaining to the partitioning of these animals based on GRM coefficients can be found in Figure 3.4.

#### *Simmental External Validation*

The American Simmental Association (ASA) has continued an annual evaluation of meat tenderness through their own CMP following the same protocols established for the NCBA CMP. Genotypes and phenotypes were obtained on a small, contemporary population of these animals (n = 443) to perform an external validation for WBSF in a population of animals somewhat distantly related and at least two generations more

### Method 1:

1	.50	.45	.40	.53
.50	1	.49	.41	.43
.45	.49	1	.59	.51
.40	.41	.59	1	.48
.53	.43	.51	.48	1

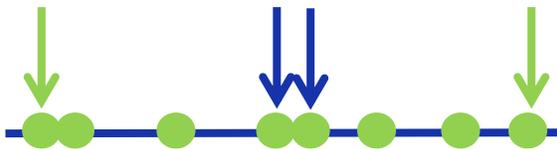
Identify the largest (smallest) relationship between pairs of animals and place one animal from the pair in the training and the other in the validation group. Clear the row and column. Continue until the desired training or validation group size (whichever is smaller) has been achieved.

### Method 2:

1	.50	.45	.40	.53	=.47
.50	1	.49	.43	.43	=.46
.45	.49	1	.59	.51	=.51
.40	.43	.59	1	.48	=.48
.53	.43	.51	.48	1	=.49

Method 2: Identify the animal with the largest (smallest) mean relationship with all other animals (animal<sub>1</sub>) and then the animal (animal<sub>2</sub>) with the largest (smallest) relationship to animal<sub>1</sub>. Place one animal in the training group and the other in the validation group and clear the corresponding row and column. Continue until the desired training or validation group size (whichever is smaller) has been achieved.

### Method 3:



Method 3: Perform a principal component decomposition on the GRM. Using PC1, rank the values in the eigenvector and sample the individuals with the highest and lowest values to minimize the relatedness or closest to the mean to maximize relatedness.

Figure 3.4: Schematic explanation of methods to separate animals into training and validation groups based on genomic relationship matrix (GRM) coefficients.

contemporary than the animals in the CMP training population. SNP data for these animals was also obtained using the BovineSNP50 assay in the same manner as for the CMP animals and missing data was imputed using FastPHASE (Scheet and Stephens 2006). After quality filtering using the same criteria listed above, 42,588 SNPs remained for analysis. To build prediction models with utility across both populations, the intersection set of quality filtered SNPs ( $n=37,334$ ) was obtained. As in previous analyses, 20 bootstrap samples were used for each analysis to average sampling effects in the training population. To ensure unbiased comparisons across analyses and validation sets, the original training datasets were employed for all analyses. The entire external Simmental population ( $n = 443$ ) was used for validation. A single training run was used to obtain prediction models using all 3,240 CMP phenotypes as well as training within only the CMP Simmental population ( $n = 516$ ), as no sampling variation exists with these datasets.

#### *BayesB0 and BayesC0 Comparisons*

Priors for parameters for the BayesC0 analyses were identical to those for the BayesC $\pi$  analyses with the exception of  $\pi$ . Initial parameter values are in Table 3.2. BayesB0 has been found to be especially sensitive to the priors (Habier *et al.* 2011, Gianola *et al.* 2009), and therefore priors for the BayesB0 analyses had to be adjusted to optimize the analysis (Figure 3.1). Adjustments were made to utilize the means of the posterior distributions for the genetic and residual variances from the BayesC analyses. To ensure fair comparisons across analytical models, identical training and validation datasets were

used for every model across all 20 bootstraps, thus the only difference between these analyses was the model used.

## **Results and Discussion**

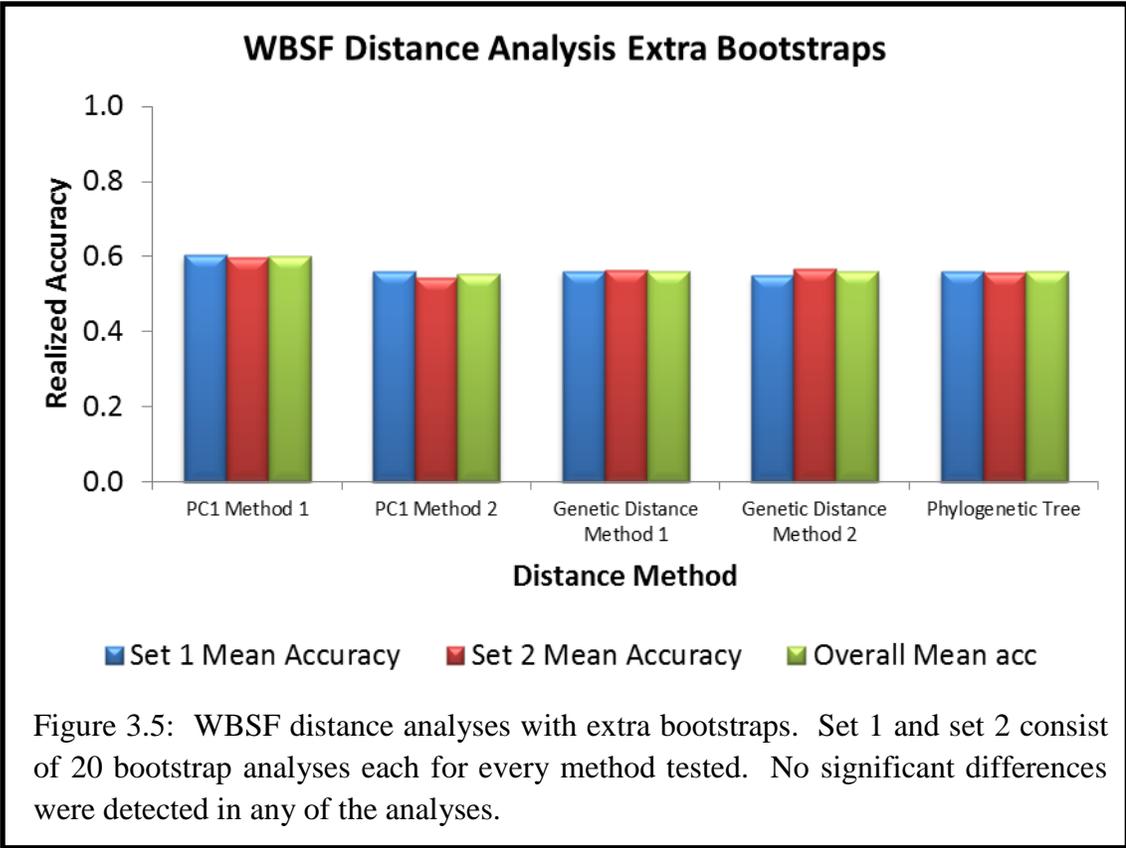
### *Informed Priors*

Panel A in Figure 3.1 illustrates the decrease in realized accuracy from analyses when comparing two different sets of priors employed in the BayesB0 analyses.  $B0_u$  (red) represents the standard (uninformed) priors that were used in all other analyses and  $B0_i$  (blue) represents priors that were informed by the means of the posterior distributions for additive and residual variance from the BayesC analyses. In Panel A, realized accuracies were calculated using a constant heritability estimate derived from the BayesC0 analyses, where the decrease in realized accuracy simply reflects a decrease in the correlation between the phenotype and the MBV. Differences between using uninformed ( $B0_u$ ) and informed ( $B0_i$ ) priors on realized accuracies were greater for these analyses when the realized accuracies were estimated within each analysis using the  $h^2$  estimate obtained for that model (Figure 3.1, Panel B).

### *Genetic Distance Analyses*

Results for analyses that accounted for the genetic distances between breeds are presented in Figure 3.2. For WBSF (Panel A), a significant difference ( $p < 0.05$ ) between distance methods and random allocation was found only for the analysis based upon partitioning using PC1. These analyses were also performed using REA phenotypes (Panel B),

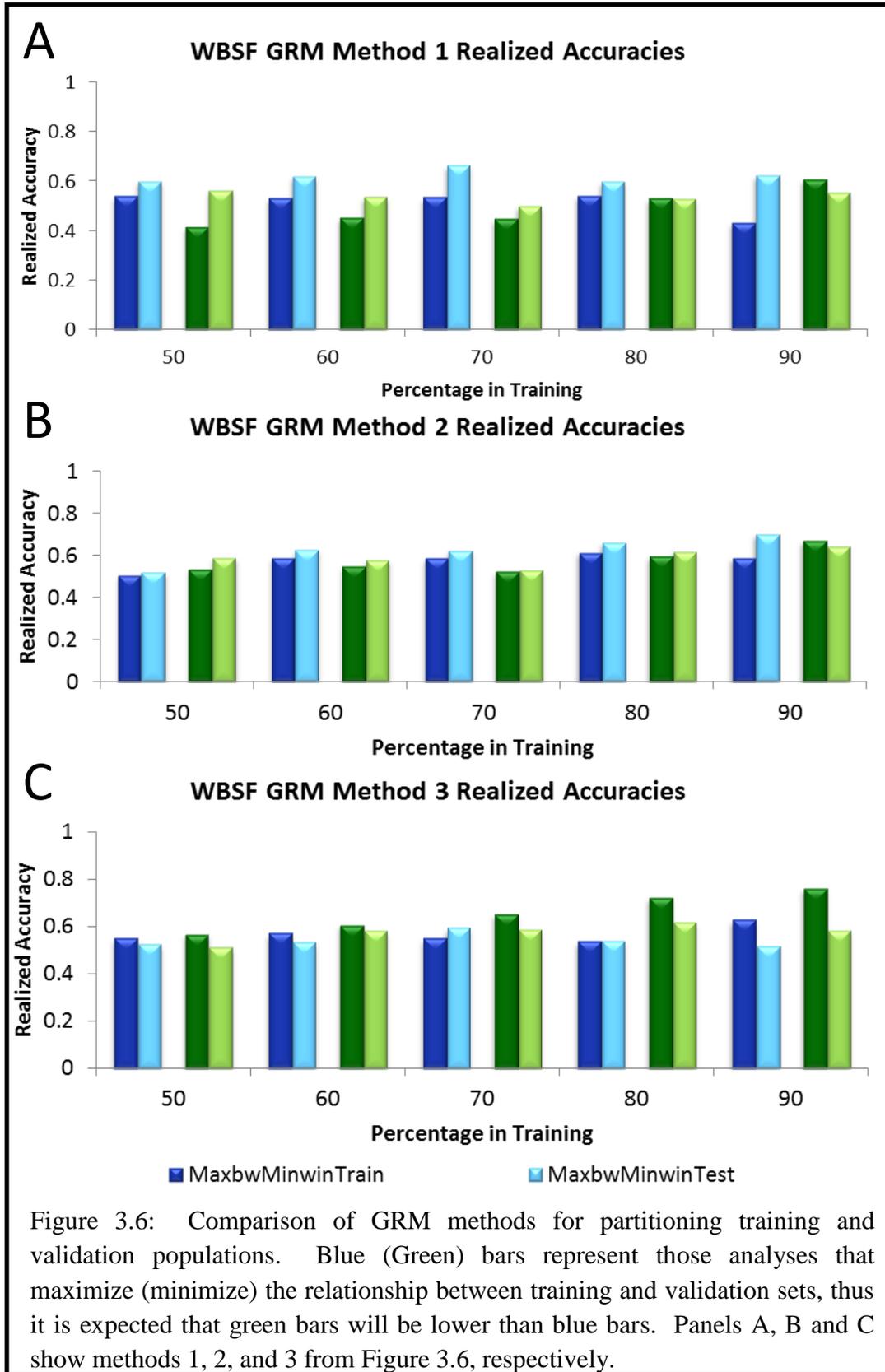
however, only partitioning based upon Genetic Distance Method 1 was significantly different ( $p < 0.05$ ) between the distance methods and random allocation. Because these differences were inconsistent among traits, we hypothesized that this result could be due to insufficient numbers of bootstrap replications and that the generation of additional replicate samples may elucidate differences between these methods. To test this, we compared the results from the first set of 20 bootstraps (Set 1) to a second analysis conducted using WBSF phenotypes which included an additional 20 bootstraps (Set 2). Results of the mean realized accuracies over the two sets of 20 bootstraps are presented in Figure 3.5. No significant differences ( $p < 0.05$ ) were found between any of the bootstrap sets for any analyses, which lead to the conclusion that 20 bootstraps were sufficient to estimate the mean realized accuracy for all of the tested methods. Because the inconsistencies between the genetic distance analyses could not be reconciled by running additional bootstrap replicates and without evidence to support a consistent advantage such as that noted by Toosi *et al.* (2010) and de Roos *et al.* (2009) in simulated data, we concluded that our results were most likely artifactual and influenced by the pedigree structure of our sample. In simulation studies such as those previously discussed, the exact number of generations to coalescence of two populations is known, however, with populations of commercial cattle, this information is unknown. It is also possible that the extensive admixture and the crossbred structure of these populations rendered our methods incapable of effectively partitioning animals into groups or of correctly estimating the genetic distances between the populations in this study. While pursuing these approaches may provide a short term benefit when generating prediction



models, the advantage does not appear to be consistent across traits and these approaches appear to not merit the computational time necessary to perform the additional distance analyses for each trait.

### *GRM Analyses*

Results for all GRM partitioning methods are presented in Figure 3.6. It has been demonstrated that maximizing the relationships between the training and validation populations (i.e., VanRaden *et al.* 2009), such as training in sires and validating in sons, increases the accuracy of genomic predictions. To determine whether this effect would be observed in the CMP population and to evaluate genomic predictions in the presence or absence of close pedigree relationships, we endeavored to partition animals into training and validation populations by either maximizing (Figure 3.6, blue bars) or minimizing (Figure 3.6, green bars) the relationships between individuals in the training and validation populations. If these methodologies are performing as expected, the green bars in Figure 3.6 should display lower realized accuracies than the blue bars. While method 1 (Figure 3.4) tended to lead to higher accuracies when individuals were partitioned to maximize the relationship between training and validation population individuals, none of the tested GRM methods led to consistent differences in MBV accuracy. Because genomic relationship coefficients cannot be accurately estimated between individuals in different breeds, the analyses were performed using proportions of animals sampled from each breed based on within-breed genomic relationships. It is likely that the extensive half-sib structure present in each of these breed-specific



populations has prevented effective partitioning of these groups. Without large differences in the average genomic relationship coefficients between the training and validation groups due to the animals within each breed all being highly related to each other, the desired genetic distance between training and validation sets could not be achieved.

### *Random Allocation*

Random allocation of animals was the most computationally tractable method used in this study and provided accuracies that were equivalent or superior to the other tested methods. Estimated heritabilities,  $\pi$ , correlations, and realized accuracies are reported for all best-fit analyses in Table 3.4. Best-fit analyses were selected based on the set of 20 bootstraps for each model that exhibited the lowest coefficient of variation when greater than 50% of the data was used in training. When less than 50% of the data were used in training, parameters were not always accurately estimated; therefore realized accuracies calculated using these estimates can be biased upward due to their reliance on heritability estimates for their calculation (Figure 3.7, Panel A). When constant heritability estimates are used to estimate the realized accuracies, results are more stable across the analyses with less data (Figure 3.7, Panel B). Additional results for the other traits are provided in the appendix (Figures A.1-A.5). Estimates of  $\pi$  were very large for all traits with the exception of MARB (Table 3.4). It should be noted that BayesC $\pi$  tended to underestimate the heritability compared to the other tested methods, especially for traits with large effect gene, and the most accurate way to compare realized accuracies across

Table 3.4: Parameters ( $\pi, h^2, r_{\hat{g},y}$ ) derived from the mean of posterior distributions for the best-fit analyses for each trait along with calculated measures of realized accuracy.

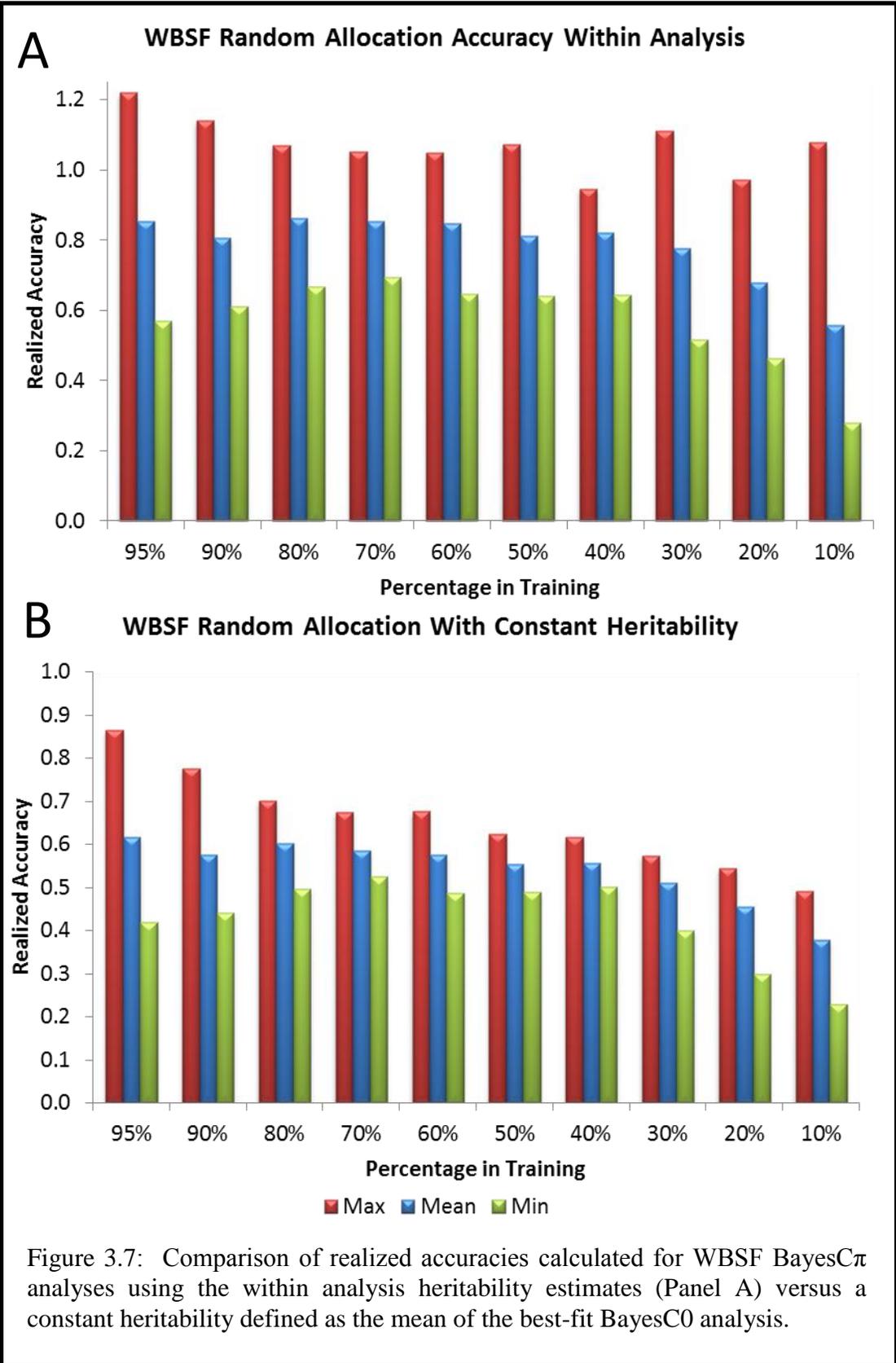
Trait	n Train	Analysis	$\pi$	$h^2$	$r_{\hat{g},y}$	Realized Accuracy <sup>1</sup>	Realized Accuracy <sup>2</sup>	Realized Accuracy <sup>3</sup>
WBSF	2268	BayesC $\pi$	0.9998	0.12	0.298	0.854	0.862	0.585
		BayesC0	0	0.26	0.276	0.547	0.796	0.541
		BayesB0	0	0.29	0.314	0.586	0.906	0.616
REA	1927	BayesC $\pi$	0.9931	0.32	0.336	0.600	0.594	0.585
		BayesC0	0	0.33	0.345	0.599	0.609	0.600
		BayesB0	0	0.38	0.343	0.560	0.606	0.597
MARB	1657	BayesC $\pi$	0.7432	0.62	0.595	0.757	0.756	0.762
		BayesC0	0	0.62	0.595	0.759	0.756	0.762
		BayesB0	0	0.72	0.590	0.697	0.750	0.756
FT	1887	BayesC $\pi$	0.9999	0.06	0.206	0.815	0.842	0.397
		BayesC0	0	0.27	0.267	0.517	1.091	0.514
		BayesB0	0	0.11	0.242	0.746	0.990	0.467
HCW	1931	BayesC $\pi$	0.9539	0.49	0.536	0.763	0.763	0.766
		BayesC0	0	0.48	0.543	0.785	0.772	0.776
		BayesB0	0	0.63	0.536	0.677	0.762	0.765
YG	2219	BayesC $\pi$	0.9998	0.08	0.217	0.757	0.753	0.412
		BayesC0	0	0.28	0.264	0.502	0.914	0.500
		BayesB0	0	0.13	0.256	0.714	0.888	0.486

\* BayesC $\pi$  estimates are not provided for %CL due to a heritability estimate of 0

<sup>1</sup> Mean of realized accuracies calculated from the within-analysis heritability for each bootstrap

<sup>2</sup> Mean of realized accuracies calculated from a standardized heritability from the best-fit BayesC $\pi$  analysis

<sup>3</sup> Mean of realized accuracies calculated from a standardized heritability from the best-fit BayesC0 analysis

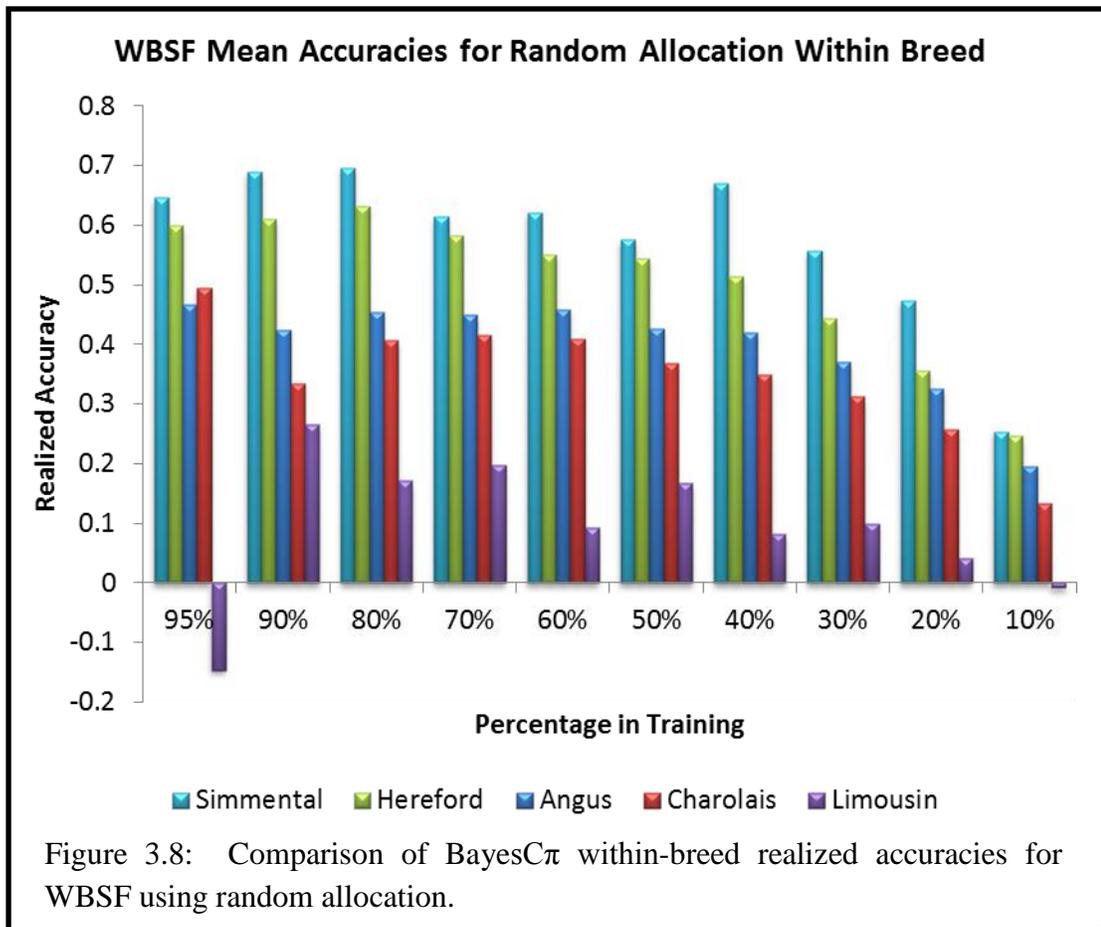


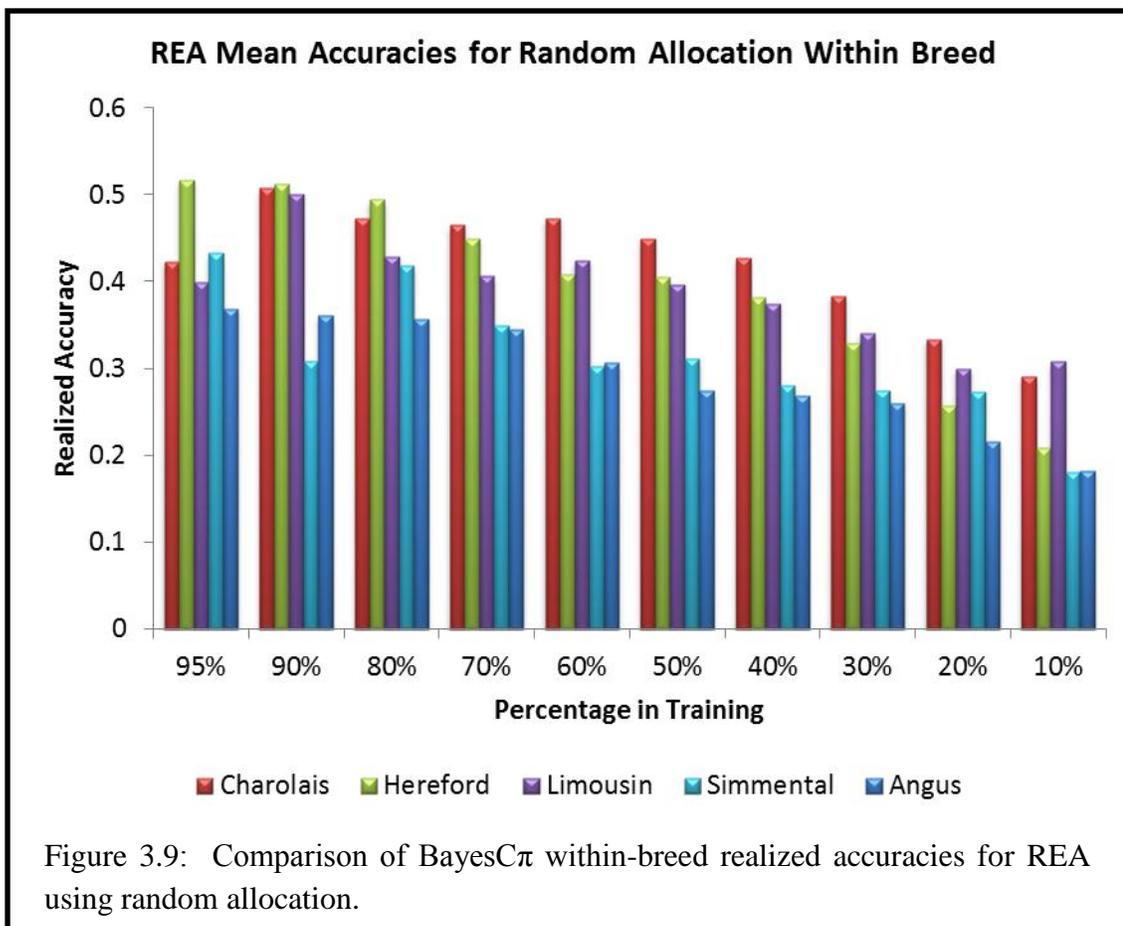
all analyses was by using a constant heritability from the BayesC0 analysis in the calculation of the realized accuracies (Table 3.4, last column). Estimates of  $\pi$  tended to be the highest for those traits which had genes of large effect (WBSF and FT). Realized accuracies ranged from 0.41 (YG BayesC $\pi$ ) to 0.78 (HCW BayesC0) in this study. The highest accuracies were obtained for those traits where  $\pi$  was estimated to be smaller (more markers influenced the trait). Several SNPs detecting the largest effects on WBSF within the *CAPNI* and *CAST* genes were not found on the 50K SNP chip, but were additional genotypes provided by a custom GoldenGate assay (McClure *et al.* 2012). Additional fine-mapping around other genes of large effect in these traits (WBSF and FT) and the inclusion of additional markers in those regions may significantly improve predictive power and accuracy for those traits.

Realized accuracies within each breed used in the analysis varied widely depending on the predictive power in the analysis and the breed's heritability for a given trait. Within breed accuracies are shown for all traits (WBSF Figure 3.8, REA Figure 3.9, and all other traits in Figures A.6-A.9). Presumably because of limited sample size, Limousin realized accuracies tended to be among the lowest for a variety of traits, but this was not universally true (i.e., FT Figure A.7).

### *Model Comparisons*

Within the literature, differences have been noted between linear modeling approaches as compared to Bayesian approaches for analyses of traits with large gene effects (i.e., VanRaden *et al.* 2009). Few studies have examined differences between different

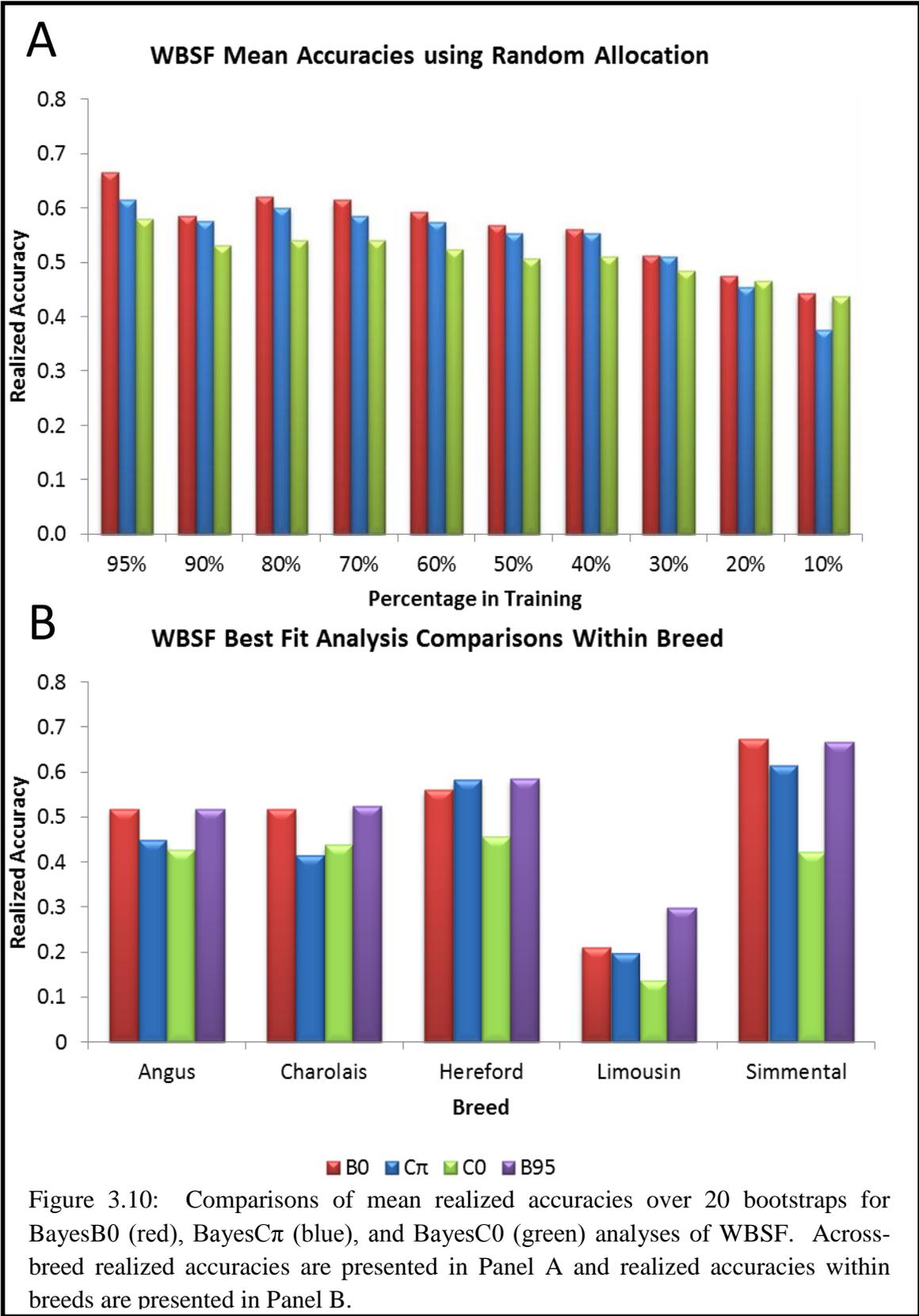


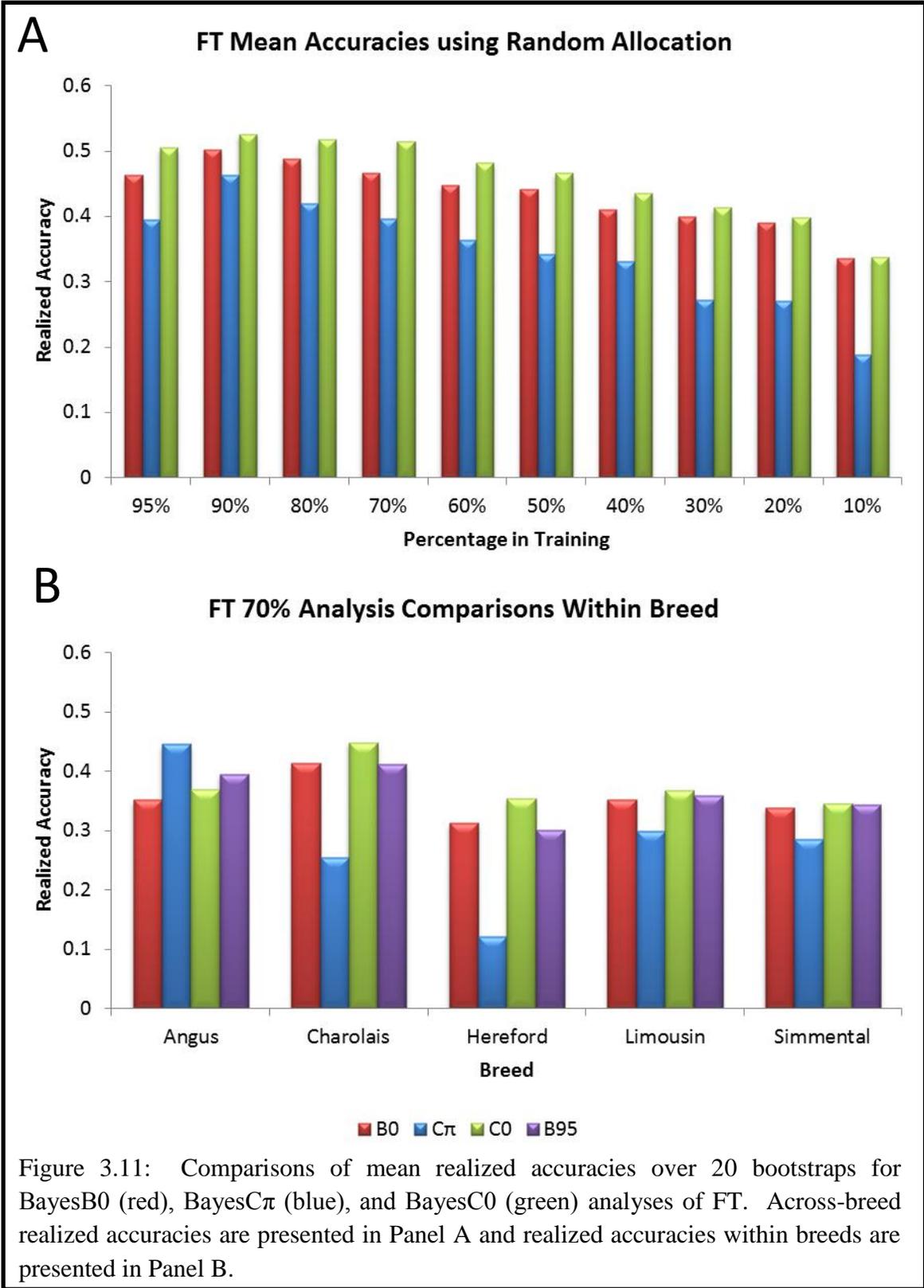


Bayesian modeling approaches for traits which either follow the infinitesimal model or those which have large effect genes. We compared the performance of BayesB0, BayesC0, and BayesC $\pi$  for all of the traits in this analysis to test whether these relationships held within a spectrum of populations and traits. For an unbiased comparison, the same animals were used in both the training and validation populations for all model comparisons.

Results from the WBSF analyses (Figure 3.10) show a consistent advantage for models that include unequal allele substitution effect (ASE) variances, which is consistent with results of VanRaden *et al.* (2009), and is undoubtedly due to the large influence of *CAPNI* and *CAST* on WBSF in beef cattle. BayesB ultimately achieved the highest realized accuracies and there appeared to be no penalty for the inclusion of all the markers, because the unequal variances allowed the small effect loci to be shrunk appropriately. Similarly, BayesC $\pi$  accounted for differences in ASE variances by utilizing a mixture model approach ( $\pi > 0$ ) which facilitates regression of the ASE to zero for the large majority of markers ( $\pi$ ) that are not associated with the trait variation. BayesC0 performed the poorest for this trait and the inclusion of all the markers without accounting for differences in ASE variances likely added noise to the MBV estimation in addition to over-regressing the effects of *CAPNI* and *CAST* towards zero.

Although the results for WBSF were consistent with previously reported results in the literature, those for FT did not follow the same pattern (Figure 3.11), even though we found several genes of large effect for this trait (See Chapter 4). For FT, BayesB0 and





BayesC0 performed similarly and BayesC $\pi$  was vastly inferior. This discrepancy is most likely due to the fact that BayesC $\pi$  performed poorly in Hereford (Figure 3.11 Panel B), and to a lesser extent, in Charolais, which together comprised over 55% of the total FT dataset (Table 3.1). Results for YG were similar (Figure A.10), presumably due to the heavy reliance on FT in the calculation of YG. Because of the population structure in this study (Angus and Hereford being largely purebred with the Continental breeds being Continental x Angus), the superiority of BayesC $\pi$  predictions in Angus and the drastic reduction in prediction accuracy for Hereford with all of the Continental breeds being intermediate suggests that the large effect gene discovered for this trait is Angus-specific. The decrease in accuracy in the Continental breeds likely reflects the degree of Angus chromosome admixture within the respective purebred populations, resulting in segregation of this QTL in the crossbred progeny of those breeds with the most Angus admixture (Limousin and Simmental) which would lead to superior BayesC $\pi$  predictions over those crossbreds with lower Angus admixture (Charolais). BayesC $\pi$  models appear to perform best in traits with a large effect gene, with the stipulation that these effects must consistently segregate in all breeds used in the prediction.

For traits which did not appear to possess large effect genes, such as REA (Figure 3.12), MARB (Figure 3.13) and HCW (Figure A.11), no differences were detected between the analytical models. This indicates that when QTL effects are small and distributed throughout the genome, any of the tested models perform equivalently. Whether  $\pi$  was fixed at 0 or fixed at 0.95 resulted in very few differences in the realized accuracies across all of the breeds and traits analyzed in this study (Panel B, Figures 3.9-

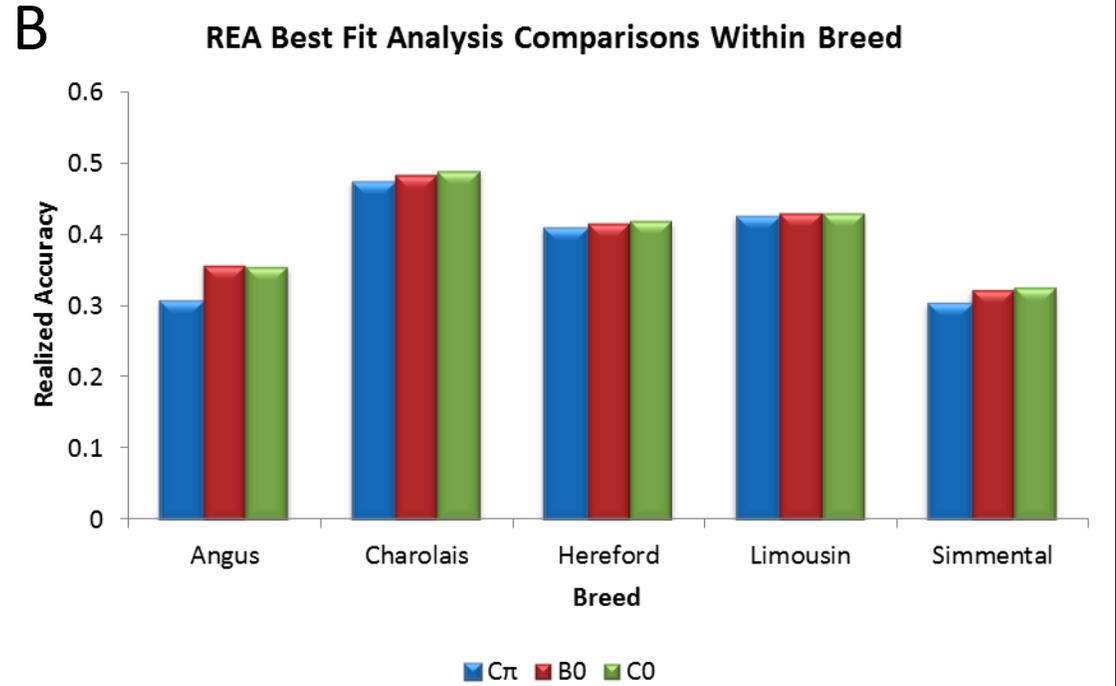
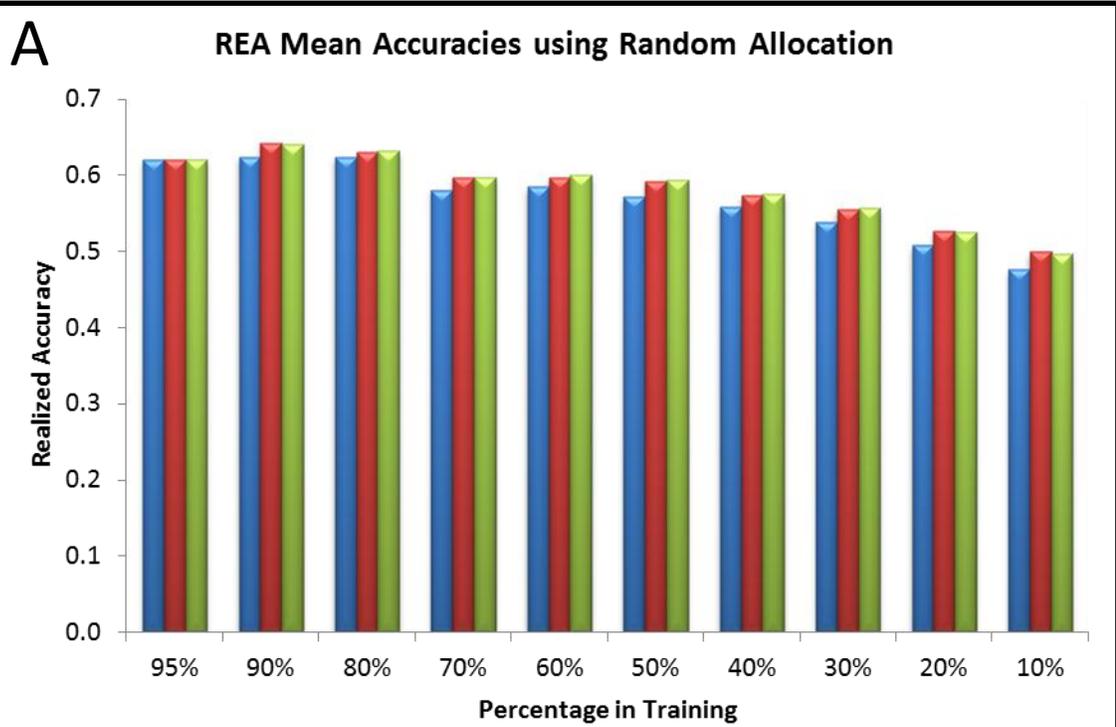
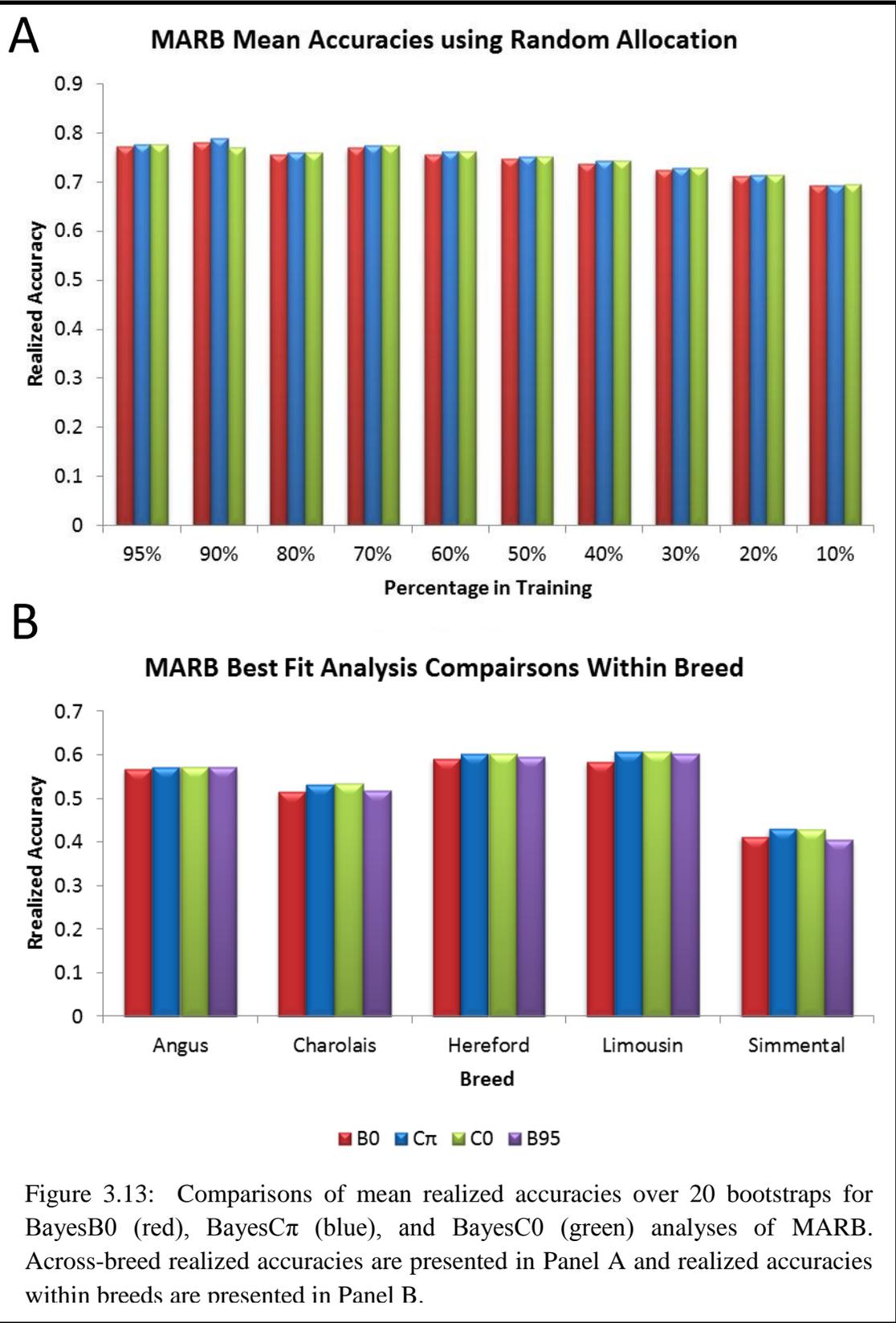


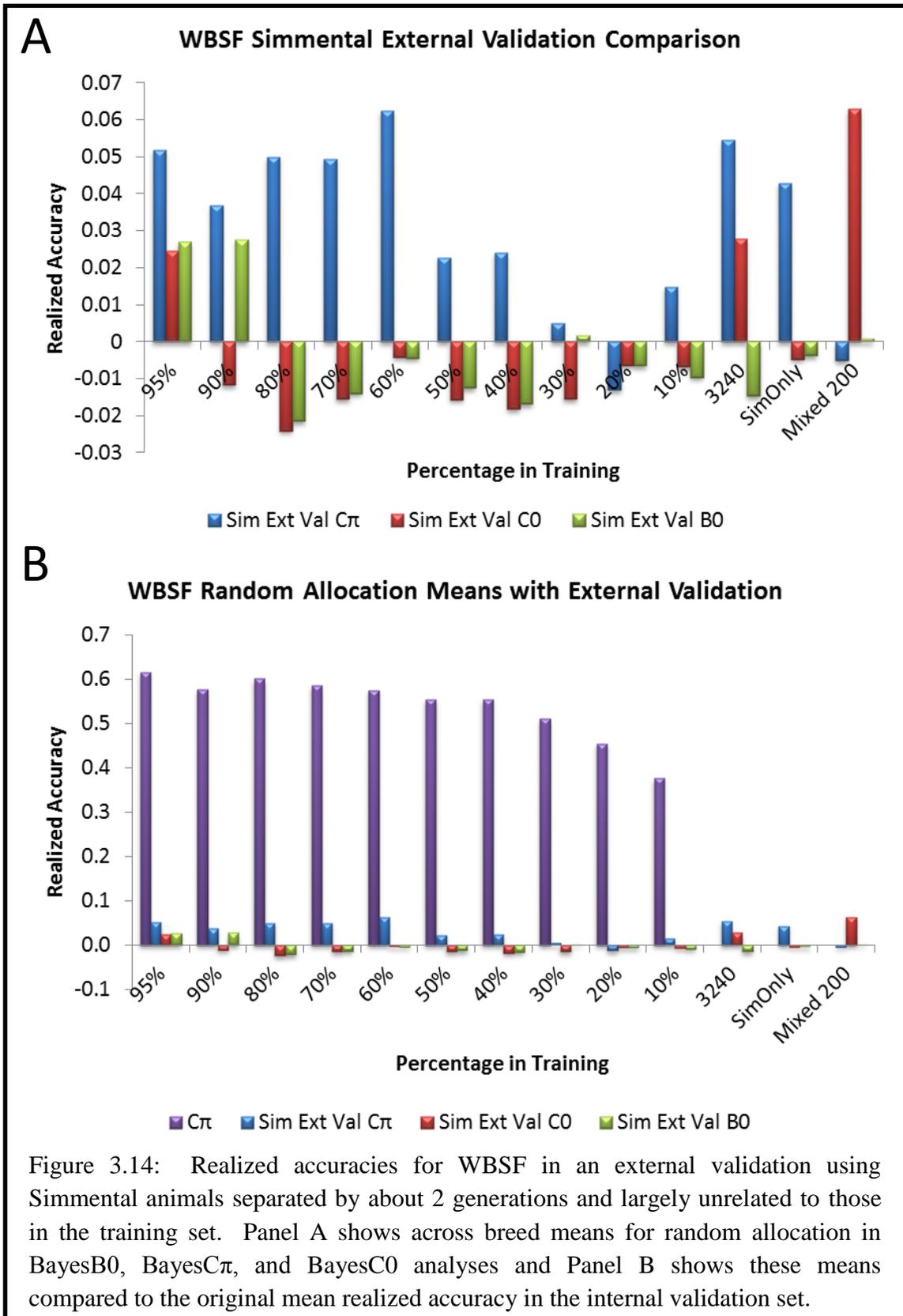
Figure 3.12: Comparisons of mean realized accuracies over 20 bootstraps for BayesB0 (red), BayesC $\pi$  (blue), and BayesC0 (green) analyses of REA. Across-breed realized accuracies are presented in Panel A and realized accuracies within breeds are presented in Panel B.



3.12, and Figures A.10-A.11) which demonstrates that the mixture proportion was less important than the ability of this model to assign individual locus ASE variances to allow the regression of small effect SNPs strongly towards zero. These results and the fact that the BayesB model analyses (both BayesB where  $\pi = 0$  or  $\pi = 0.95$ ) tended to perform consistently well for traits with and without genes of large effect suggests that BayesB analyses should strongly be considered as the “gold standard” for GS training models.

#### *External Validation*

A study by Habier *et al.* (2007) noted that predictions based on models which include large amounts of linkage (familial relationships between animals) will decay in accuracy more rapidly over time than those based on models which identify markers that are in high LD with causal mutations. Because haplotype blocks in which all markers are in strong LD are much shorter in crossbred and admixed populations than in their purebred counterparts (Kizilkaya *et al.* 2010), models trained and used for MBV prediction in crossbred populations require testing to determine the amount of linkage that is being modeled and the rate of decay of prediction. Accuracy decreases as generations advance away from the training population. To achieve this goal, we performed an additional external validation for the WBSF prediction models. A population of largely unrelated Simmental, Angus, and Sim-Angus steers and heifers produced at the University of Illinois was utilized to determine the decrease in prediction accuracy observed over approximately 2 generations (~8-10 years). Results from this analysis are presented in Figure 3.14.



When validation was performed in a completely independent population stratified in time relative to the training population, prediction accuracies dropped approximately 90% from  $\sim 0.6$  to 0.05. BayesC $\pi$  predictions were consistently superior across all analyses. Predictions based upon BayesB0 estimation of SNP effects tended to increase in accuracy as more data was included in the analyses (up to utilizing all of the original 3,240 records for training). It is likely that the addition of more records, especially those on animals of similar breed composition to the external validation set would improve prediction accuracies. The severity of the erosion of prediction accuracy in a population separated from the training population by at least two generations (the sires of each population were not genotyped and so the extent of relatedness among the sires could not be established) reinforces the fact that retraining of prediction equations and the continued collection of high-quality phenotype data will be necessary to capitalize on the opportunities afforded by genomic selection.

A similar result was found by de Roos *et al.* (2009) using simulated data from two different populations with different divergence times. In their study, the erosion of prediction accuracy could be corrected by the addition of animals from the new population into the training set. When we added approximately half of the ( $n = 243$ ) animals in the external validation set to the full complement of 3,240 phenotypes (total  $n=3,483$ ) and retrained the analyses for BayesC $\pi$ , BayesB0, and BayesC0 we did not observe the expected restoration of prediction accuracies in the 200 animals remaining in the validation set (Figure 3.14, Mixed200). The extremely small sample size and the small proportion of animals used in the training set (5% of the total available number of

animals) likely explains this failure to reestablish prediction accuracy. If a larger number of animals were available to include in the retraining, we would expect to see a significant restoration of the size of the realized accuracies in the validation set.

## **Conclusions**

Genomic predictions utilizing 50K SNP data for 3,240 animals with carcass quality phenotypes yielded MBV predictions that ranged in accuracy from 0.41 to 0.78 across all populations and traits examined in this analysis. Models built using BayesB0 tended to consistently produce the highest prediction accuracies across all traits in the analysis, however, the best model varied depending on the genomic distribution of QTL size effects and the breed architecture of the populations which influenced whether large effect QTL were segregating. Within an integrated program such as GenSel, tailoring the analysis to the trait's QTL effect distribution is feasible. It appears that considerable care must be taken when applying genomic selection models to diverged populations within the same breed, even those which coalesce within a few generations. Frequent retraining with the inclusion of newly genotyped samples will probably mitigate this effect; however, it is possible that retraining may need to capture much of the diversity within each breed to allow GS to be effective breed-wide.

CHAPTER IV: GENOME-WIDE SCAN FOR QUANTITATIVE TRAIT LOCI  
AFFECTING CARCASS TRAITS IN FIVE CROSSBRED AND PUREBRED  
TAURINE BEEF BREEDS

**Summary**

Association analyses have formed the backbone of the search for regions of the genome harboring mutations that affect economically relevant traits in beef cattle and other species. These regions can become candidates for further study to identify the causal mutations and also for the formation of small targeted trait-specific SNP panels used to predict genetic merit. We performed a genome wide association analysis for carcass traits spanning five breeds of taurine cattle (n = 615 Angus, n = 695 Charolais, n = 1,065 Hereford, n = 283 Limousin, and n = 516 Simmental) from the National Cattleman's Beef Association Carcass Merit Project (CMP). These animals were genotyped using the Illumina BovineSNP50 chip. After quality filtering, 40,645 SNPs were available for analysis. GenSel was used to implement Bayesian analyses of these data to estimate SNP effects for each trait. The QTL associated with the top 1% of SNP

effects explained from 2-18% (9-88% after correction for region dependency) of the additive genetic variance in the trait. Different models for any given trait identified 10-78% concordant genes across all analyses and showed 32-85% concordance when a gene was allowed to be missing from one analysis.

Key Words: BovineSNP50, quantitative trait loci, beef cattle, carcass traits, Carcass Merit Project

## **Introduction**

Because of the structure of the beef industry, and the fact that most cattle are sold at weaning rather than participating in various retained ownership schemes, the cow/calf sector has historically lacked the data and selection tools to efficiently practice selection for carcass traits. The generation of expected progeny differences (EPDs), which were implemented in the 1970s (Willham 1993), has made efficient selection possible, and indeed, genetic progress has been made in these traits for most beef cattle breeds. However, for many producers, the generation of these metrics is largely dependent on the use of ultrasound data as correlated traits because carcass data collected from packing plants has been difficult to gather in the absence of retained ownership and subsequent marketing of calves on quality and yield grids. Ultrasound data is widely accepted as providing good estimates of body composition; however, small systematic biases (over- or under-estimation of body fat or longissimus muscle area depending on the fatness of the animal) have been noted in the literature (Greiner *et al.* 2003). Because of these challenges, genomic selection approaches based on small, targeted panels of markers

which assay regions of the genome harboring mutations important for variation in carcass traits could serve as an affordable method to achieve greater accuracy for carcass trait EPDs and improve the response to selection in the beef industry.

In an effort to address issues of consumer dissatisfaction with beef tenderness, the CMP was begun in 1998 (Thallman *et al.* 2003) with the fundamental goal of providing tools to identify genetically superior animals by providing breed associations with the relevant data with which to generate EPDs for carcass, tenderness, and sensory traits. This project was a massive effort which combined expertise from four universities, the USDA Agricultural Research Service, 13 breed associations, and the National Cattlemen's Beef Association (NCBA). A unique aspect of this project was its incorporation of trained sensory panel evaluations of tenderness, juiciness, flavor, and connective tissue on steaks derived from several different breeds and sires involved in the project. Extensive analyses of these data were presented by Thallman *et al.* (2003) and Minick *et al.* (2004) and exploration of the utility of these data for the development of genomic selection models was performed by Rolf *et al.* (found in Chapter 3).

Numerous studies in the literature have focused on building knowledge pertaining to regions of the genome which harbor mutations which affect variation in economically important carcass traits. Leptin (*LEP*; Fitzsimmons *et al.* 1998, Nkrumah *et al.* 2005, Schenkel *et al.* 2005), Calpain (*CAPNI*; Page *et al.* 2002), Calpastatin (*CAST*; Schenkel *et al.* 2006), muscular hypertrophy (*MSTN*; Casas *et al.* 1998), neuropeptide Y (*NPY*; Sherman *et al.* 2008), uncoupling protein 2 (*UCP2*; Sherman *et al.* 2008), and insulin-

like growth factor 2 (*IGF2*; Sherman *et al.* 2008) mutations, to name a few, have all been associated with carcass quality traits. It is widely acknowledged that association analyses can falsely identify regions of the genome as being associated with a phenotype. Consequently, all associations must be verified in independent populations (Schenkel *et al.* 2005). More recently, McClure *et al.* (2012) utilized genomic BLUP (GBLUP) to identify regions of the genome affecting variation in tenderness. These data were also used to fine-map the causal mutations underlying *CAPNI* and *CAST*, which have long been known to have large effects on beef tenderness but for which the causal mutations remain unknown. Our study, spanning five breeds of taurine cattle, used the same data but alternative models to perform association analyses of several carcass traits in the CMP population. In addition to the identification of important regions in the bovine genome regulating variation in carcass traits, this study serves as an independent validation of previously reported associations. An advantage of our study design is the utilization of an admixed population. While it can be more difficult to detect regions of the genome harboring causal mutations with this approach due to the shorter range of linkage disequilibrium (LD) in admixed populations (Toosi *et al.* 2010), the regions that are detected are likely to be relevant and maintain the same phase between marker and QTL alleles in a large number of breeds, making these associations of wider interest than those identified in only a single breed. Finally, we employed a Bayesian framework for these analyses, which has been shown in some studies (de Roos *et al.* 2007, Harris *et al.* 2008, Hayes *et al.* 2009) to better reflect the architecture of QTL effects within the genome, especially in the presence of large effect QTL (VanRaden *et al.* 2009, Chapter

3). Simulation studies have shown these Bayesian methodologies to be effective in detecting QTL effects within the genome (Habier *et al.* 2011). In particular, Habier *et al.* (2011) described a relatively new Bayesian analysis named BayesC $\pi$ , which was able to distinguish SNPs with non-zero normally distributed effects describing QTL from those predicted to have no effect on the trait. This methodology may provide an accurate way to predict the number of QTL effects within the genome, provided that QTL effects are normally distributed and the sample size is large (Habier *et al.* 2011).

Carcass quality and yield traits are the culmination of a long-term selection program that begins in the registered and concludes in the cow/calf sector. Improvements in selection for these traits must be initiated within both sectors, but the availability of selection tools has historically been limited. The best application of these tools is a genomic-enhanced EPD, such as those produced in the seedstock sector by the American Angus Association, American Hereford Association and American Simmental Association. Even small single carcass trait SNP panels have been shown to yield positive net returns for the computation and deployment of MBVs (DeVuyst *et al.* 2011), however the use of panels which do not contain causal variants must be cautious due to the expected decay in prediction accuracy that will occur over time (Chapter 3). In an effort to encourage the adoption of genomic selection within the beef industry, we may need to identify small, targeted panels of markers that are predictive of genetic merit for carcass traits and that are cost-effective within both the commercial and feedlot sectors. This approach may assist in bridging the gap until the causal mutations are identified and reduce the cost of testing relative to the use of larger genotyping panels. For this

purpose, we performed an association analysis using 3,240 animals with carcass records produced in the National Cattleman's Beef Association Carcass Merit Project (CMP) and representing five commercially relevant crossbred and purebred taurine breeds.

## **Materials and Methods**

### *Population*

Samples were chosen from the CMP comprising five different breeds of taurine cattle. A total of 3,360 animals (Angus n = 660, Charolais n = 702, Hereford n = 1192, Limousin n = 285, and Simmental n = 521) were chosen to be genotyped from all available samples based on the preferential selection of animals with observations for Warner-Bratzler Shear Force and the most complete carcass data. As noted in Minick *et al.* (2004), all steers and heifers in the CMP were sired by bulls of the respective breed classifications but the dams were from commercial herds. While the Angus and Hereford CMP progeny were largely purebred, the Continental breeds were largely crossbred due to Limousin, Simmental and Charolais bulls being mated to commercial Angus cows (McClure *et al.* 2012; Minick *et al.* 2004).

### *Phenotypic Data*

Data collection procedures were described by Minick *et al.* (2004). Briefly, USDA personnel recorded marbling score (MARB), hot carcass weight (HCW), fat thickness at the 12<sup>th</sup> and 13<sup>th</sup> rib interface (FAT), and ribeye area (REA) between 24 and 48 hours postmortem. For reference, a MARB score of 500 corresponds to Small<sup>00</sup>. Kansas State

University personnel recorded WBSF observations using steaks which had been aged in vacuum packaging for 14 days (further details can be obtained in Minick *et al.* 2004). Muscle, DNA, and white blood cells (WBC) were obtained from Texas A&M University with the permission of the sample owners (American Angus Association, American Hereford Association, American Simmental Association, American International Charolais Association, and the North American Limousin Foundation). WBC samples were obtained at weaning and muscle samples were obtained at harvest. For the limited number of samples for which DNA was provided, paternity and individual identification (DNA profiles matched between WBC and muscle samples) had previously been tested as part of the CMP protocol. Any DNA samples found to have pedigree or identification errors were removed and DNA was re-extracted from muscle samples at the University of Missouri to ensure the greatest probability that the DNA and phenotypes would be concordant, regardless of paternity, which was not considered in our analysis. Genomic DNA was re-extracted from 2,940 muscle samples by proteinase K digestion followed by Phenol:Chloroform:Isoamyl alcohol extraction and ethanol precipitation (Sambrook *et al.* 1989). The remaining 420 samples were DNA samples that had successfully passed identification and paternity verification.

#### *Genotypic Data*

All CMP samples were genotyped for 54,790 SNPs using the Illumina BovineSNP50 BeadArray (Matukumalli *et al.* 2009). An additional 96 putative SNPs located within 186 kb of *CAST* and *CAPNI* were also assayed using a custom Illumina GoldenGate assay (additional details are in McClure *et al.* 2012). All genotypes were called in the

Illumina GenomeStudio software. After filtering for SNP assignment to the genome using UMD3.1 coordinates (Zimin *et al.* 2009), call rate  $<0.89$  (to include all commercialized tenderness SNPs), and minor allele frequency (MAF)  $>0.01$ , 40,645 SNPs remained for analysis on 3,240 animals (Angus  $n = 651$ , Charolais  $n = 695$ , Hereford  $n = 1,095$ , Limousin  $n = 283$ , and Simmental  $n = 516$ ). FastPHASE v1.2.3 (Scheet and Stephens 2006) was used to phase all genotypes and impute the 0.89% of missing genotypes.

### *Analysis*

Several studies have documented a slight advantage to Bayesian methodologies for modeling SNP effects predicting molecular breeding values (de Roos *et al.* 2007, Harris *et al.* 2008, Hayes *et al.* 2009), especially in the presence of large effect genes (VanRaden *et al.* 2009, Chapter 3). Consequently, we employed four Bayesian models for genome-wide association analysis, that are implemented in the GenSel software package (Fernando and Garrick 2009) developed at Iowa State University and that has successfully been used to generate within-breed MBVs for genomic selection (Saatchi *et al.* 2011). We utilized the capabilities of GenSel to implement four different models for each trait, each using 160,000 iterations of Markov chain Monte Carlo (MCMC) and parameterized using priors estimated from weighted means of within-breed residual and additive genetic variance component estimates derived from a previous genomic best linear unbiased prediction (GBLUP) analysis (McClure *et al.* 2012). The exception was for the BayesB analyses, where the selection of priors was previously described in

Chapter 3. Based on assumptions about the distribution of QTL effects in the genome, we set prior values for  $\pi$ , the proportion of markers in the analysis that do not influence a trait, to be large in an effort to better describe the trait-associated variation within the genome. Additional details about each type of analysis are in Chapter 3.

Briefly, BayesC with  $\pi = 0$  (BayesC0) was used to estimate parameters in a Bayesian framework, due to its similarity to GBLUP (McClure *et al.* 2012). BayesB with  $\pi = 0$  (BayesB0) was also utilized, and, like BayesC0, included all markers into the prediction model. BayesB with  $\pi=95$  (BayesB95) assumed that only 5% of the markers were associated with each trait and were sampled into the model on each MCMC iteration. Two Metropolis-Hastings iterations were used in each Markov-chain Monte Carlo step to estimate the allele substitution effect variances in each of the BayesB analyses. Finally, we performed a BayesC $\pi$  analysis and  $\pi$  was estimated from the data. BayesC $\pi$  may be an important tool for association analysis, because  $\pi$  provides information about the genetic architecture of the analyzed traits. In particular,  $\pi$  provides an estimate of the number of SNPs in the analysis that are significantly associated with the trait provided that the SNP effects are normally distributed (Habier *et al.* 2011). Specific details concerning model assumptions were presented in Chapter 3. Each analysis for the traits that were not part of the sensory panel evaluation (WBSF, REA, MARB, FT, HCW, and YG) was performed according to the procedures outlined in Chapter 3 for random allocation of animals to independent training and validation sets, which included 20 bootstrap replicates for each sampling procedure. For the sensory

panel traits, a single analysis was performed for each model due to insufficient numbers of records for bootstrapping or independent validation.

Due to the study design, all animals produced records that were subject to significant contemporary group effects which required the adjustment of the observations to phenotypes. Contemporary groups were defined by the combination of herd of origin, breed, sex, and harvest date and were modeled as a fixed effect in a single BayesC $\pi$  analysis encompassing all animals for each trait. Observations were then pre-corrected for these effect estimates before additional analyses were performed. For the sensory panel traits (Tender, Juicy, Flavor, and ConnTiss), there were insufficient numbers of records for the estimation of  $\pi$  and every record was used in all of the association analyses. To account for sampling effects on the estimated heritabilities due to varied sample breed compositions and the need to standardize comparisons across analyses, realized accuracies were calculated as  $\frac{r_{\theta,y}}{\sqrt{h^2}}$  where  $h^2$  was the heritability estimated in the BayesC0 analysis of the complete data.

### *QTL Delineation*

During the course of the analysis, SNP effects were estimated in the training population and window variances were generated in the training population for each five SNP sliding window within the genome. In the case of the sensory panel traits (Tender, Juicy, Flavor, ConnTiss), no animals were assigned to the validation population due to an insufficient numbers of records, so the training population was also used to generate the window variances. Five SNP window variances and individual SNP effects were ranked

for each bootstrap analysis. Means were calculated for the window variances, SNP effects, and the SNP rankings for each SNP and window. The top 1% of mean SNP effects was retained for further analysis. The number of likely QTL regions was then determined from a visual examination of the window variance plots (see Figures 4.1, 4.2 and Figures A.12-A.15) for the BayesC0 analysis. BayesC0 analyses were chosen for this purpose because they tended to be the analyses with the largest number of QTL effects, regardless of the size of the detected QTL effects. This approach allowed for a more liberal selection of QTL regions which could then be further narrowed by independent validation using the literature, through functional analyses, or by concordance between different analyses performed in this study. A QTL region was established by identifying SNPs with effects ranked in the top 1% within 3.5 Mb of other top 1% ranked SNPs. This approach generated numbers of predicted QTLs which approximately matched the number of QTLs that were detected by visual examination. Analyses were next conducted to ascertain the annotated genes within these QTL regions ( $\pm 0.25$  Mb) using the UMD 3.1 genome assembly. Following the identification of these gene lists for each analysis (BayesC $\pi$ , BayesC0, BayesB0, and BayesB95), a cumulative list of unique gene symbols was generated. Gene symbols present in at least three out of four (all non-sensory traits) or two out of three (all sensory traits) models were identified and analyzed within the Database for the Annotation, Visualization, Integration and Discovery (DAVID; Huang *et al.* 2009a, Huang *et al.* 2009b).

## **Results and Discussion**

Bayesian modeling has the ability to capture the genetic architecture underlying complex quantitative traits as was well evidenced in this study. BayesC0 is similar to a traditional mixed linear model analysis, and the Manhattan plots for these analyses reflected this similarity (Figures 4.1, 4.2 and Figures A.12-A.15). Consistent with the fact that the BayesC0 analyses estimate allele substitution effects for all markers, traits with large QTL effects such as WBSF had aggressive regression of SNP effects towards the mean (Figure 4.1) as compared to BayesC $\pi$  analyses, which attempted to identify the largest effect markers in the analysis and regress these less severely (Figure 4.3). This was less noticeable when the trait more closely followed the infinitesimal model and demonstrated no evidence for genes of large effect (Figure 4.2 vs. Figure 4.4). We expected that the mixture models (BayesC $\pi$  and BayesB95) would produce the largest estimated gene effects, which is evidenced in Figures 4.3-4.6 and Figures A.16-A.23. Notably, BayesB0 was able to accurately model regions harboring genes of large effect by modeling individual allele substitution effect variances leading to the strong shrinkage of uninformative markers, even though all markers were included in the model predictions (Figure 4.7 and Figures A.24-A.25). When no genes of large effect were present, the BayesB0 analyses produced estimates of SNP effects that were intermediate to the BayesC0 and the mixture models (Figure 4.8 and Figures A.26-A.28). Supporting the prediction accuracy results from Chapter 3, BayesB0 appears to be an effective “one-size-fits-all” analysis regardless of whether a trait exhibited evidence for genes of large effect, or followed the infinitesimal model, and may be the most appropriate analysis when little is known about the QTL architecture for a particular trait.

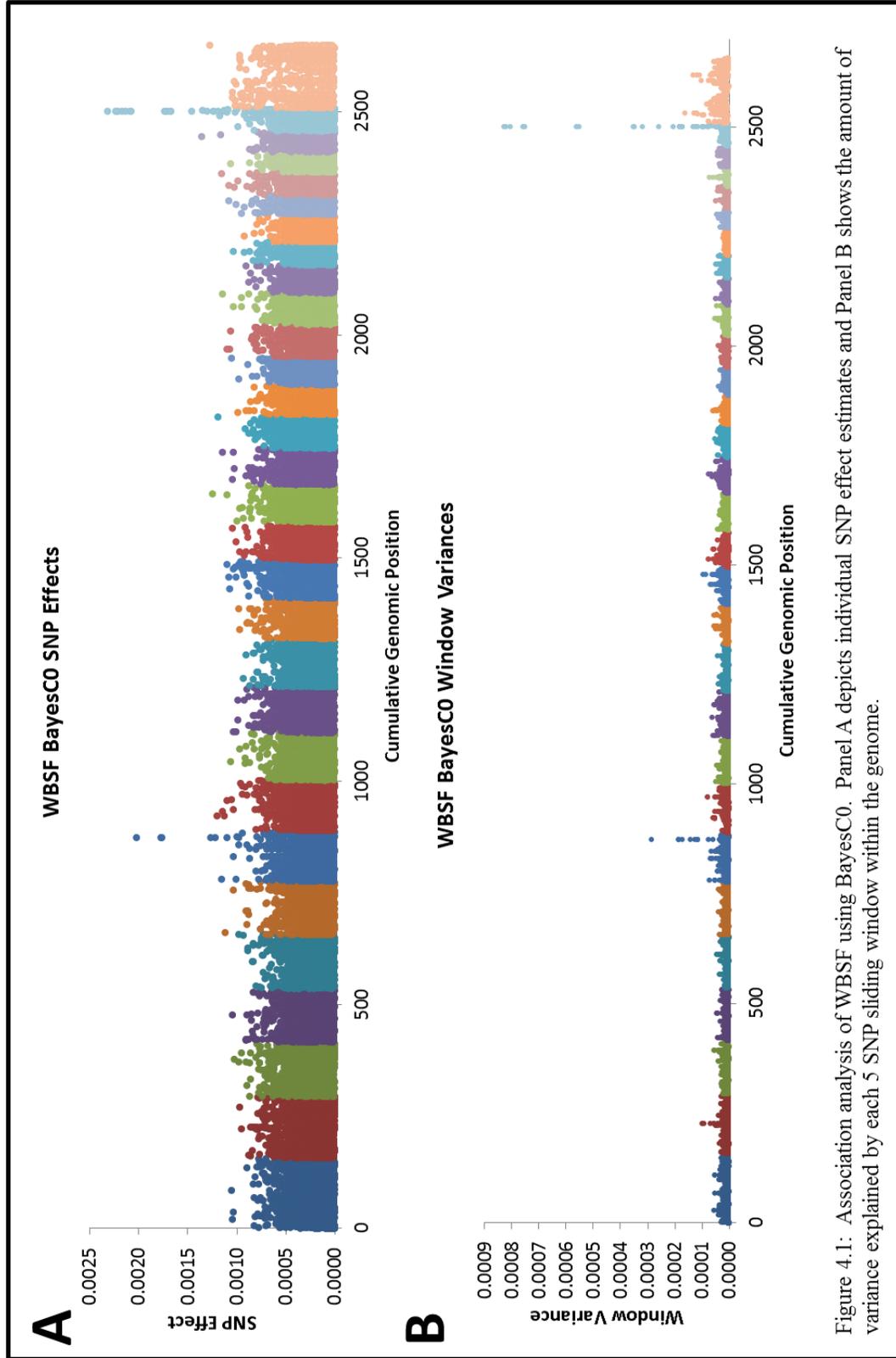


Figure 4.1: Association analysis of WBSF using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

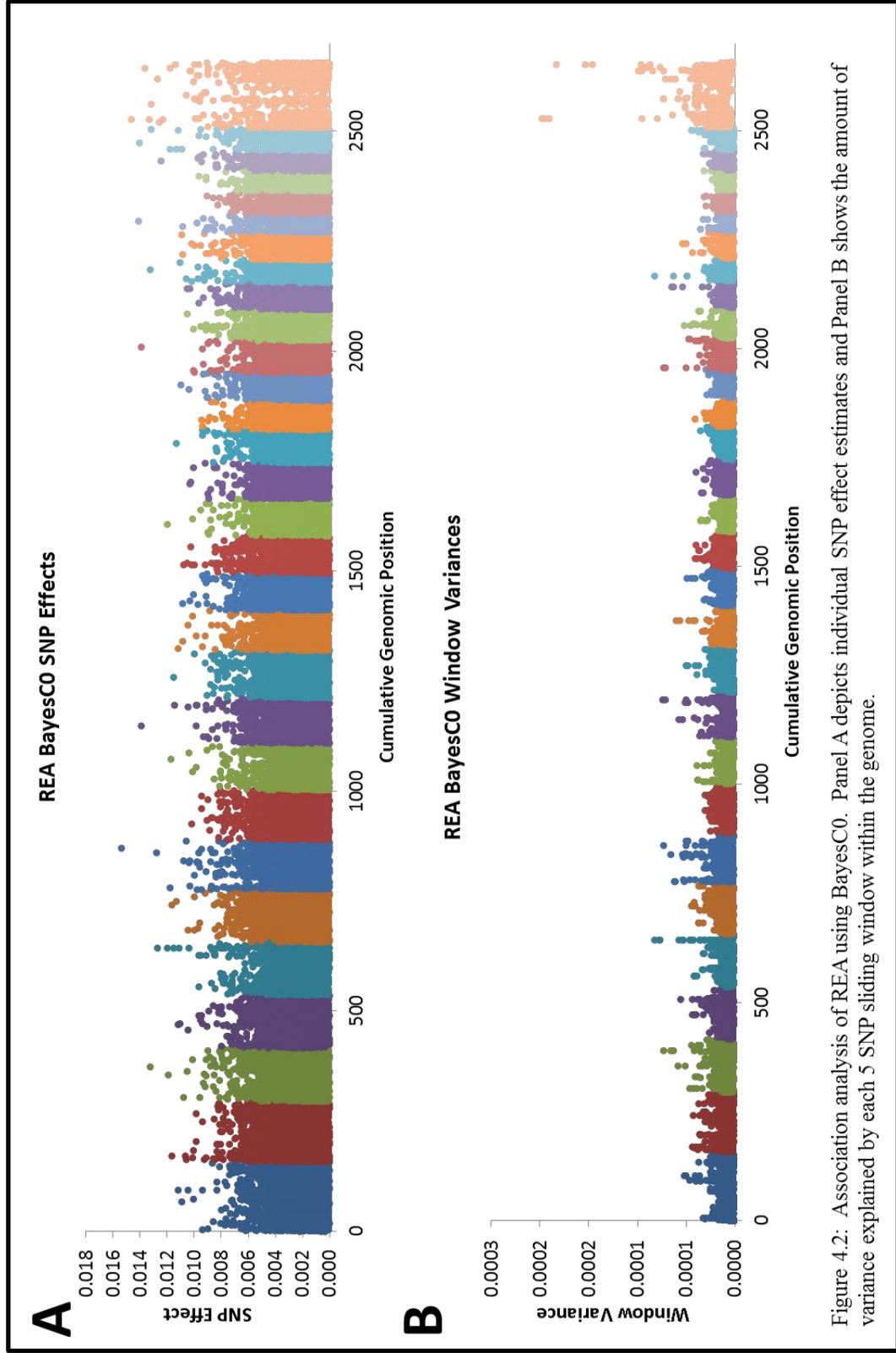


Figure 4.2: Association analysis of REA using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

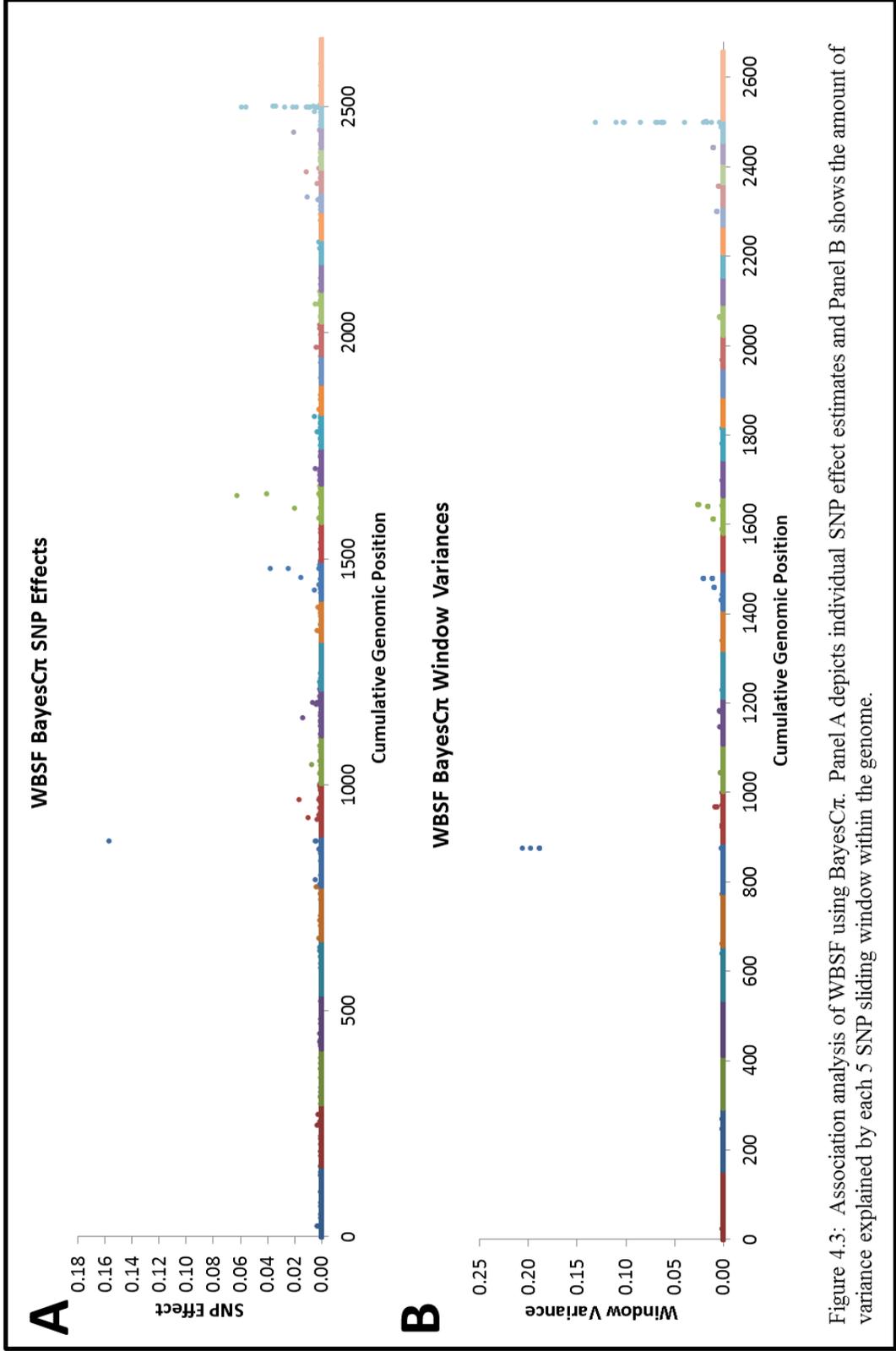


Figure 4.3: Association analysis of WBSF using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

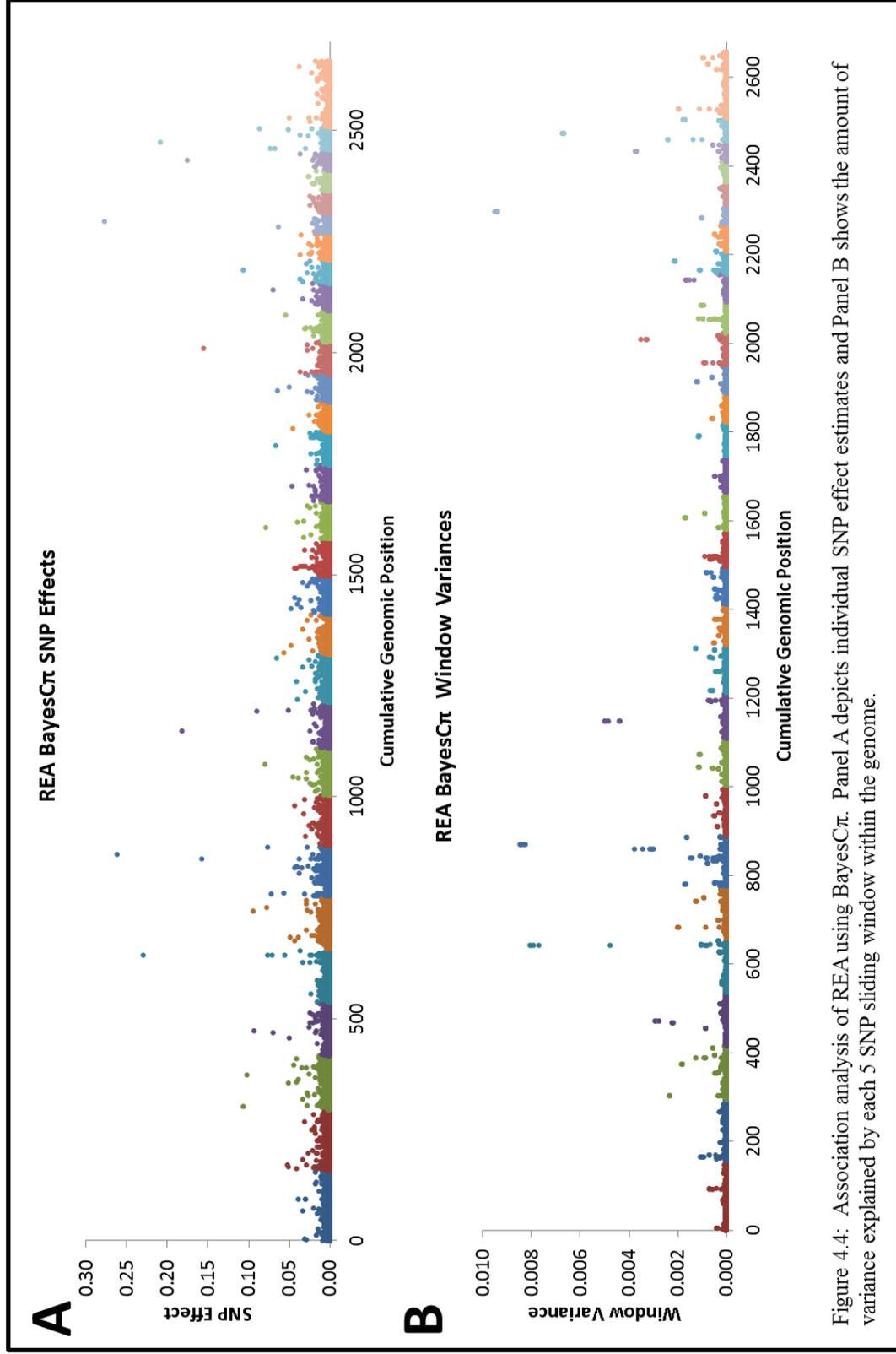


Figure 4.4: Association analysis of REA using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

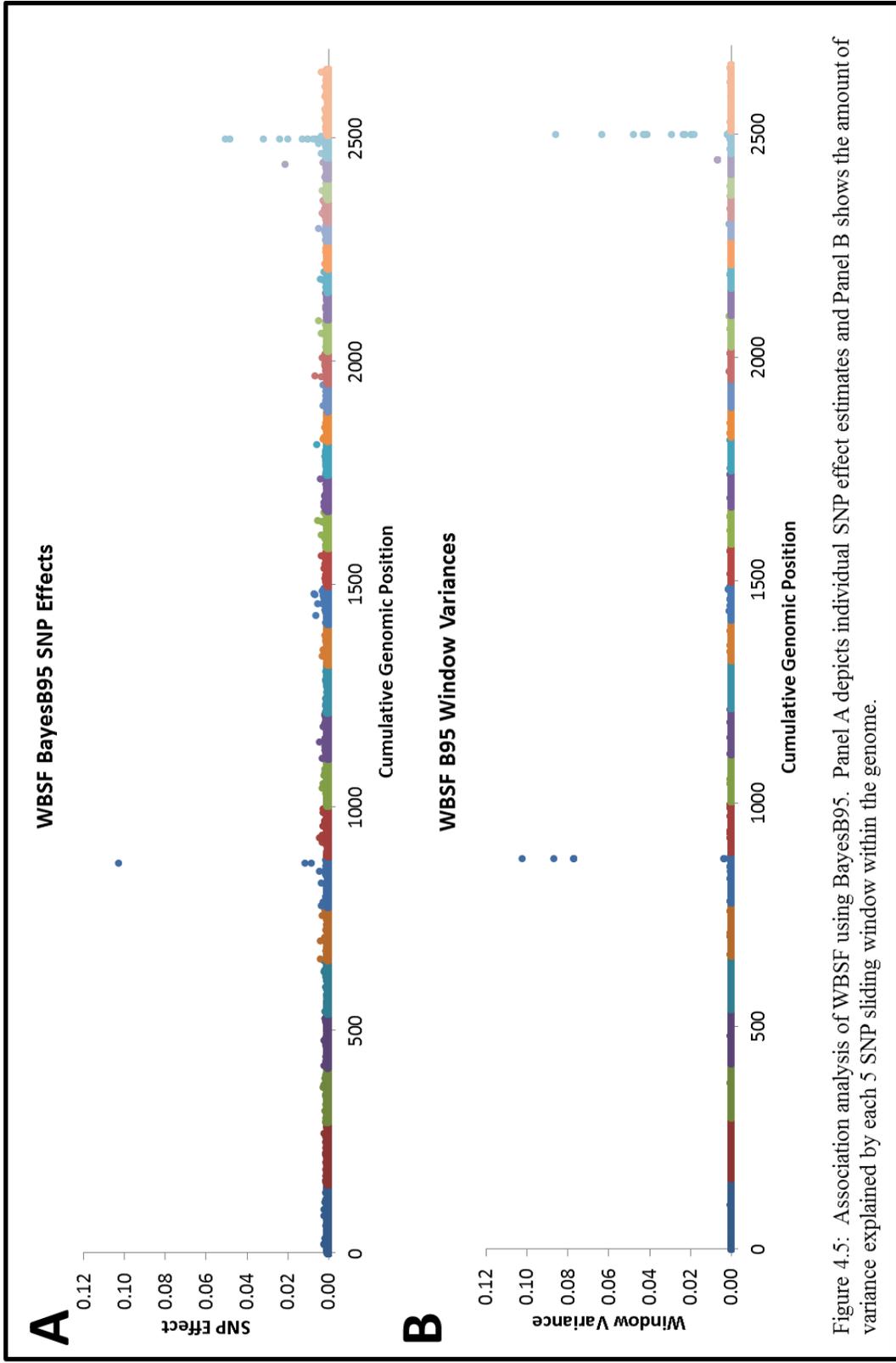
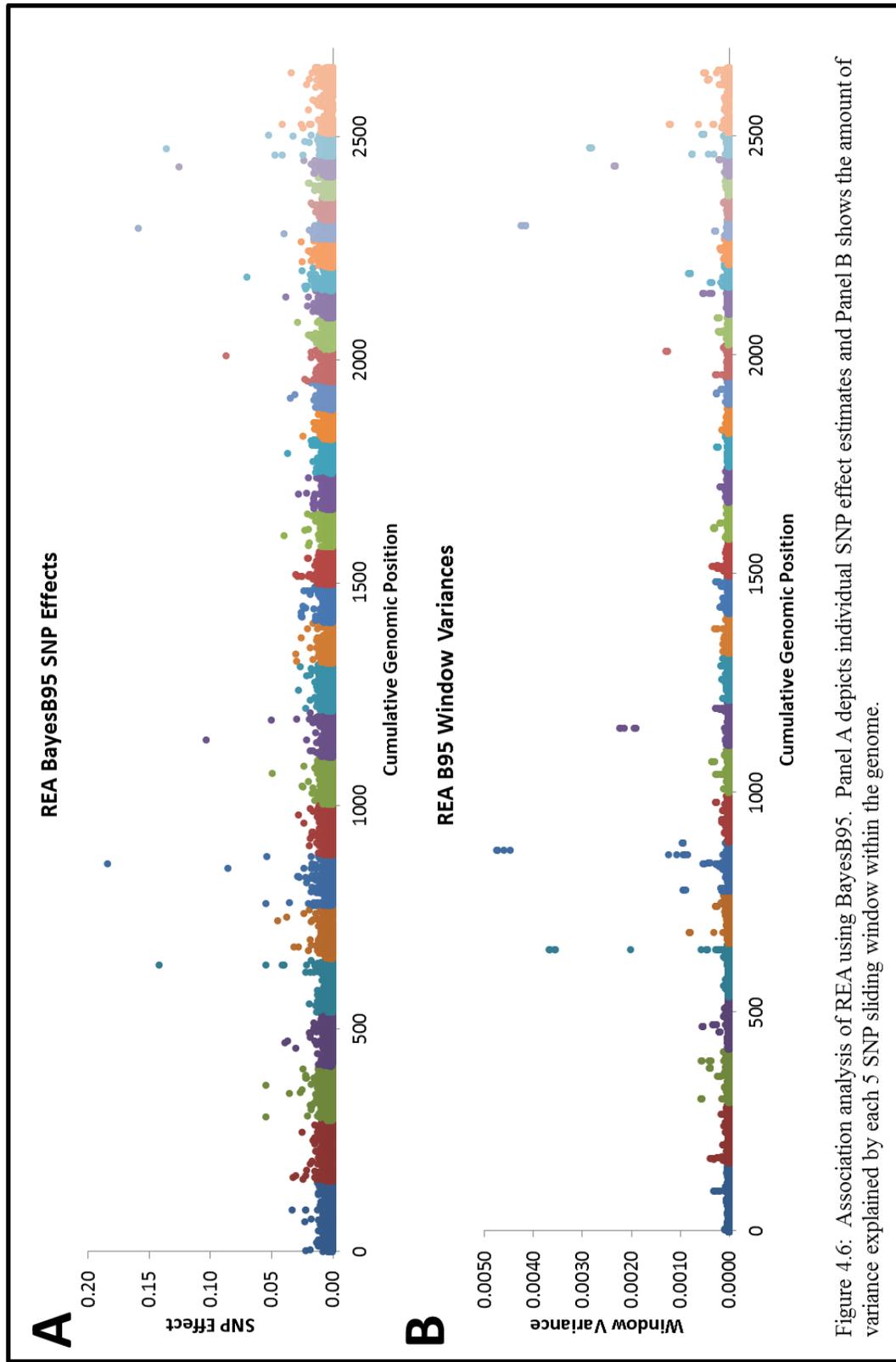


Figure 4.5: Association analysis of WBSF using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.



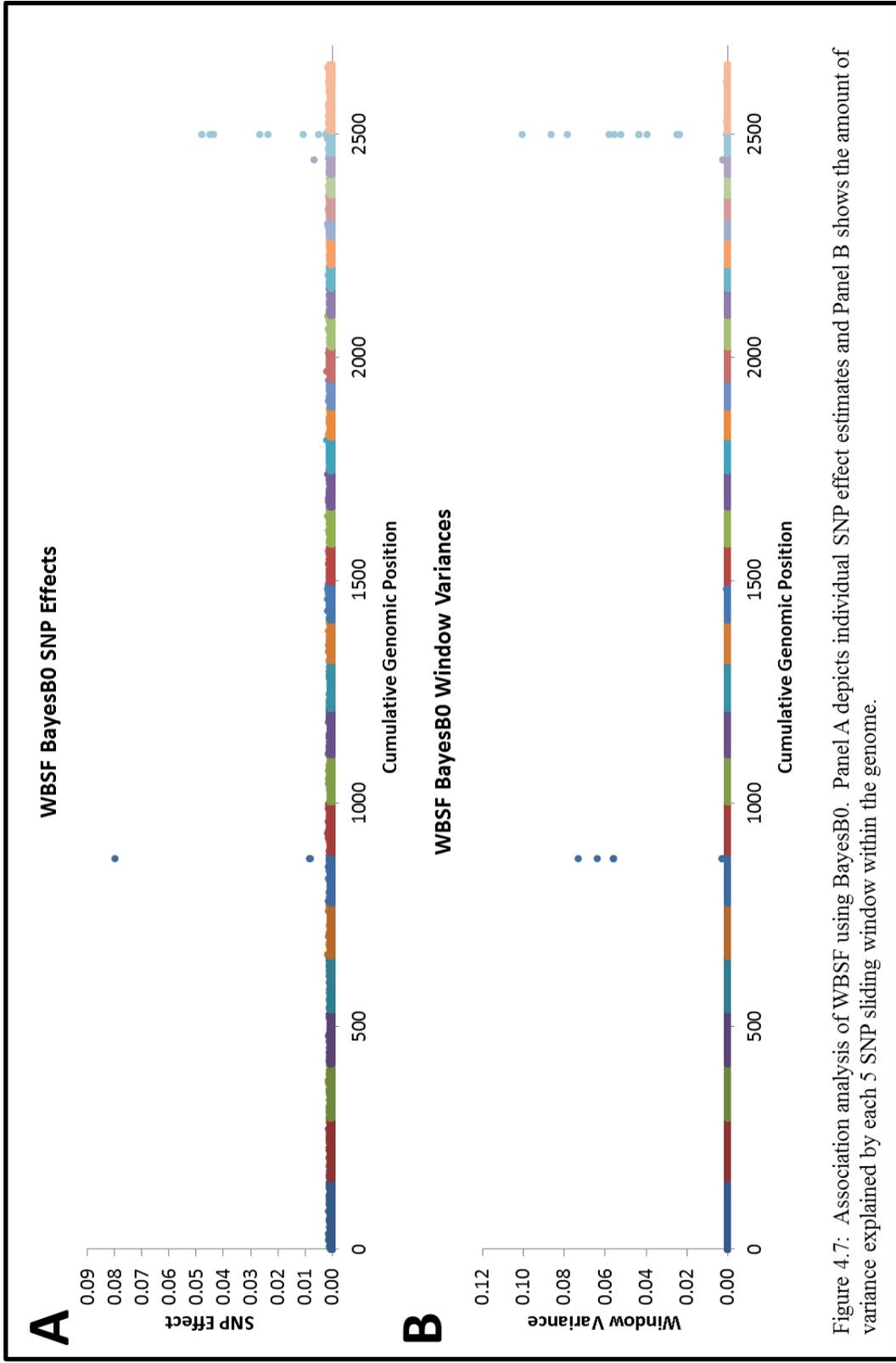


Figure 4.7: Association analysis of WBSF using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

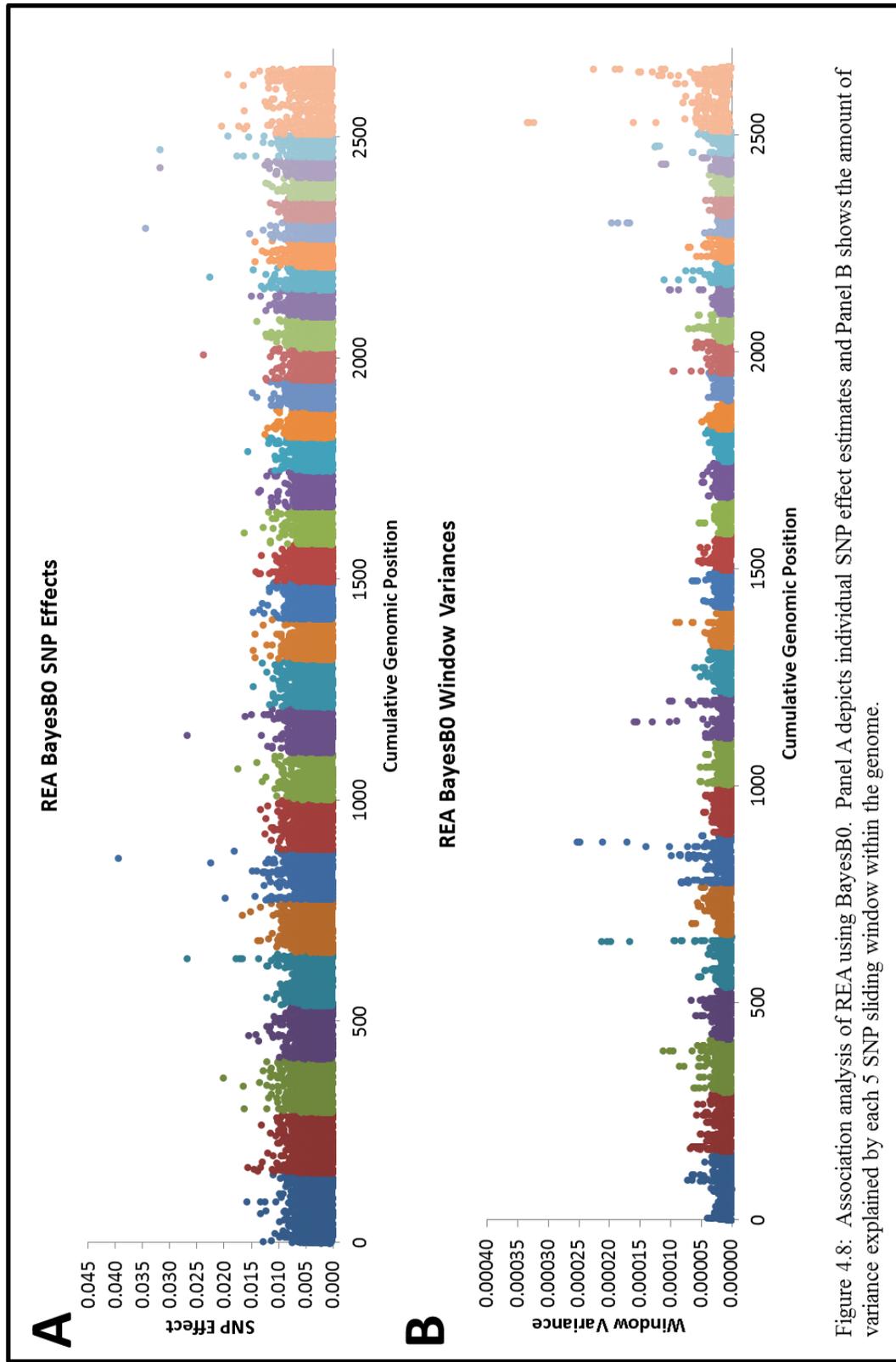


Figure 4.8: Association analysis of REA using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

The amount of variance explained by the largest effect QTL for each trait and the QTL regions identified using the top 1% of SNP effects are in Table 4.1. The effect of individual QTL regions ranged from 0.04% (Flavor) up to 9.9% (FT) of the total additive genetic variance for a trait. After correction for the bias due to dependent window variance estimates (because of the sliding five SNP window), ranges for single QTL regions were between .22% and 49.8%. Correction was performed by summing the largest variance estimates out of each group of five windows and using that as the denominator in the calculation of variance explained. The top 1% of SNP effects was collapsed to from 174 to 212 QTL regions that cumulatively explained from 1.88% to 17.8% of the variance for each trait. After correction, these ranges increased to 9.42% to 88.94%. Evidence for the extreme regression of SNP effects in the BayesC0 analyses can also be observed in Table 4.1. Window variances and percent variances explained (both cumulative and individual regions) were consistently much lower for the BayesC0 analyses for all traits in comparison to the other three models. As expected, the largest overall reduction in cumulative variance explained by the largest QTL regions was for BayesC0 analyses. This reduction in variance explained was a result of shrinkage applied to the genes of large effect noted for WBSF, FT, and YG.

Large effect genes for WBSF (*CAPNI* and *CAST*) are well documented. However, an extremely large effect FT QTL has not been previously reported in the literature, though it is likely Angus-specific (Chapter 3). The region of largest effect on FT in our study was on chromosome 13 between 59.02 Mb and 67.88 Mb. Two QTL are known to occur in this region, one of which has a most likely position within our region

Table 4.1: Summary of association analyses for each trait and model combination.

Trait	Analysis Type	Largest Effect				Top 1% of SNP effects		
		Window Variance	% of Total Variance	% of Total (Corrected)*	2q $\alpha$	No QTL	Cum % of Total Variance	Cum % of Total (Corrected)*
WBSF	C $\pi$	0.188293	6.89739	35.4221	0.01570	174	15.9943	82.1401
	C0	0.000826	0.30098	1.5063	0.00051	189	2.2669	11.3449
	B0	0.100457	8.47814	44.9514	0.00335	200	13.7533	72.9204
	B95	0.085720	7.12671	38.5796	0.00351	196	15.5547	84.2035
REA	C $\pi$	0.009370	1.02108	5.1158	0.13683	209	13.7629	68.9558
	C0	0.000197	0.07961	0.4002	0.00484	199	2.0191	10.1500
	B0	0.000333	0.10749	0.5393	0.00624	212	2.6529	13.3121
	B95	0.004725	0.97234	4.8747	0.03173	204	10.0537	50.4033
MARB	C $\pi$	0.000091	0.05349	0.2677	0.07731	194	2.4538	12.2808
	C0	0.000068	0.04137	0.2069	0.06328	190	1.8843	9.4250
	B0	0.000193	0.06959	0.3480	0.13516	199	2.3698	11.8529
	B95	0.002657	0.71927	3.5883	1.23457	203	8.5923	42.8653
FT	C $\pi$	0.317903	9.27293	46.3146	0.00391	190	17.8079	88.9435
	C0	0.000330	0.08036	0.4008	0.00021	196	2.2208	11.0774
	B0	0.031019	9.08337	45.4901	0.00061	202	9.7712	48.9349
	B95	0.060341	9.92956	49.8323	0.00158	201	11.6569	58.5013
HCW	C $\pi$	0.000643	0.42084	2.1253	0.06180	193	5.8413	29.4999
	C0	0.000093	0.07981	0.4000	0.02245	201	2.0410	10.2286
	B0	0.000323	0.11499	0.5768	0.04584	202	2.3592	11.8339
	B95	0.001684	0.48589	2.4599	0.06821	197	8.0824	40.9198
YG	C $\pi$	0.114073	5.04925	24.9732	0.00926	187	17.3186	85.6569

	C0	0.000345	0.09354	0.4677	0.00038	194	2.0938	10.4688
	B0	0.000894	0.76951	3.8373	0.00063	186	3.4859	17.3833
	B95	0.016696	5.29025	25.8724	0.00304	187	12.5133	61.1976
	C0	0.000071	0.07817	0.3909	0.00012	195	2.6732	13.3676
Tender	B0	0.000075	0.08076	0.4037	0.00015	194	2.7738	13.8660
	B95	0.000338	0.34554	1.6427	0.00108	199	4.4773	21.2851
	C0	0.000071	0.08457	0.4221	0.00016	196	2.6882	13.4193
Juicy	C95	0.000113	0.14672	0.7328	0.00029	193	3.2694	16.3301
	B0	0.000072	0.08516	0.4250	0.00026	191	2.8585	14.2659
	C0	0.000046	0.04558	0.2277	0.00009	186	2.2626	11.3057
Flavor	C95	0.000067	0.06427	0.3212	0.00021	193	3.0216	15.1027
	B0	0.000046	0.04615	0.2305	0.00009	189	2.4595	12.2854
	C0	0.000072	0.07080	0.3553	0.00014	191	2.4577	12.3334
ConnTiss	C95	0.000248	0.20370	1.0212	0.00042	191	3.0299	15.1911
	B95	0.007398	2.66657	13.2883	0.01039	198	9.6944	48.3103

\*Corrected for bias due to dependent window variances from a sliding 5 SNP window by summing the largest window from each set of 5 window variances

of interest, and were noted by McClure *et al.* (2010) using a linkage analysis in Angus cattle, but the size of the additive effects were not reported. No other QTL influencing FT within this region have been added to the Cattle QTL Database (<http://www.animalgenome.org/cgi-bin/QTLdb/BT/chromap?chromos=13&showQTL=y&showEQTL=y&showAsso=y&showwmap=y&showlmap=y&optqtl=FATTH>, accessed May 11, 2012). The SNP within this region with the largest estimated ASE is *ss86337652* (located at 63.69 Mb) and is directly within *NECAB3* (N-terminal EF-hand calcium binding protein 3). The SNP with the second largest effect is *ss86295572* (located at 63.32 Mb) is directly within the *BPIFB5* (BPI fold containing family B, member 5) gene. *SNTA1* (Syntrophin, alpha1) is one of the few reasonably well-annotated genes which lies between these two SNPs and is associated with muscle tissue growth in animals. Syntrophin also forms a complex with dystrophin and utrophin and mutations in any of these proteins are associated with muscular dystrophy. It has been suggested that *SNTA1* is upregulated in cattle with myostatin loss of function mutations (Chelh *et al.* 2011), which is consistent with its association with muscular growth and disease phenotypes. It is unclear how mutations in this protein would cause changes in fat deposition, other than if a mutation in this gene causes an energy deficiency due to changes in muscle growth or repair that has downstream effects on fat thickness.

After formation of QTL regions, the UMD3.1 genome assembly annotation database was queried to determine the number of genes residing within  $\pm 0.25$  Mb of each QTL region (Table 4.2). Because we performed four analyses per trait (three analyses

Table 4.2: Number of genes concordant between analyses for each individual trait. Total number of genes in each analysis is listed on the diagonal. Percentages in the lower triangular section are percentages of total genes implicated for each trait that are shared between analyses.

WBSF (n=3556)					
WBSF		Сπ	С0	В0	В95
	Сπ	1861	720	777	941
	С0	20%	1944	1531	1289
	В0	22%	43%	1996	1588
	В95	27%	36%	45%	1927
REA (n=2836)					
REA		Сπ	С0	В0	В95
	Сπ	1919	1136	1274	1663
	С0	40%	1979	1598	1466
	В0	45%	56%	1809	1553
	В95	59%	52%	55%	2017
MARB (n=3503)					
MARB		Сπ	С0	В0	В95
	Сπ	2791	2623	2431	2365
	С0	75%	2797	2355	2283
	В0	69%	67%	2736	2319
	В95	68%	65%	66%	2763
FT (n=3880)					
FT		Сπ	С0	В0	В95
	Сπ	2239	618	809	837
	С0	16%	1668	1095	1170
	В0	21%	28%	1811	1488
	В95	22%	30%	38%	1834
HCW (n=3004)					
HCW		Сπ	С0	В0	В95
	Сπ	1948	1422	1641	1666
	С0	47%	1723	1395	1270
	В0	55%	46%	2261	1872
	В95	55%	42%	62%	2316
YG (n=4654)					
YG		Сπ	С0	В0	В95
	Сπ	2726	664	1054	1261
	С0	14%	2225	1456	1487
	В0	23%	31%	2218	1940
	В95	27%	32%	42%	2484

per trait for all sensory panel traits), we examined the number of concordant genes identified in Table 4.2 for all four analyses for each carcass trait and in Table 4.3 for all three analyses for each sensory panel trait. Generally, the traits that were influenced by genes of large effect tended to possess a lower concordance among the genes that were detected by the different analysis models. Because the largest 1% SNP effect threshold was chosen arbitrarily, this may reflect the fact that there are actually a lower number of QTL influencing these traits (as suggested by the  $\pi$  parameter estimates, Chapter 3). In the case of FT and YG, this may also reflect the Angus breed specificity of the detected large effect.

After gene lists were generated for each analysis, a cumulative gene list was formed for each trait. Table 4.4 examines the number of genes that were identified in the cumulative trait lists that were concordant across different traits. Relationships between traits for which extensive pleiotropy was anticipated due to strong genetic correlations are highlighted in blue in Table 4.4. In general, these relationships can also be identified from our gene lists. The largest concordance between traits was between YG and FT, which was expected due to the reliance of YG on FT values for its calculation. There was also a strong concordance between YG and REA and HCW, also presumably due to the use of REA and HCW values in the calculation of YG. A strong relationship was also noted between the Tender and ConnTiss sensory panel traits. Because WBSF is a quantitative measure of steak tenderness, we detected a significant concordance between the WBSF and both Tender and ConnTiss gene lists.

Table 4.3: Number of genes concordant between analyses for each individual trait. Total number of gene names in each analysis is listed on the diagonal. Percentages in the lower triangular part of the table are percentages of the total genes implicated for each trait that are shared between any two pairs of analyses.

Tender (n=3264)				
Tender		B0	C0	B95
	B0	2849	2084	2141
	C0	64%	2265	1785
	B95	66%	55%	2447
Flavor (n=2852)				
Flavor		B0	C0	C95
	B0	2526	2097	2038
	C0	74%	2318	2075
	C95	72%	73%	2267
Juicy (n=2630)				
Juicy		B0	C0	C95
	B0	2226	2080	2073
	C0	79%	2293	2183
	C95	79%	83%	2404
ConnTiss (n=3651)				
ConnTiss		B95	C0	C95
	B95	2312	1081	1247
	C0	30%	2303	2002
	C95	34%	55%	2285

Table 4.4: Number of genes concordant for each trait. Strongly correlated traits for which extensive pleiotropy was expected are highlighted in blue. Total numbers of genes from each cumulative gene list are listed on the diagonal.

	WBSF	REA	MARB	FT	HCW	%CL	YG	Tender	Juicy	Flavor	ConnTiss
WBSF	3556	548	403	665	498	504	375	947	294	477	983
REA		2836	453	723	461	485	835	410	304	364	577
MARB			3503	576	525	441	715	579	656	654	386
FT				3880	595	378	1629	375	319	645	470
HCW					3004	324	705	269	263	400	357
%CL						3138	420	408	442	526	428
YG							4654	689	444	496	543
Tender								3264	513	440	1267
Juicy									2630	594	472
Flavor										2852	412
ConnTiss											3651

We performed functional annotation clustering on the gene lists produced for each trait using DAVID. Because of the variation in concordance of QTL regions detected across the analyses for traits with large gene effects, we analyzed in DAVID only genes which were detected in at least three of the four analyses for the carcass traits and at least two of the three analyses for the sensory traits. The number of genes used in each analysis is summarized in Table 4.5. As  $\pi$  decreased (Chapter 3, lowest for MARB, intermediate for REA and HCW, and highest for WBSF, FT and YG), the concordance of the genes identified within QTL regions by all analyses tended to increase for all of the carcass traits, indicating that estimates of  $\pi$  accurately describe the true architecture of the traits, even if the exact numbers of “significant” QTL were not established (they are likely underestimated since the estimate of  $\pi$  must depend on the size of the analyzed sample). Gene concordance for the lists submitted to DAVID analysis ranged from 32.6% to 85.6%. Results from the functional annotation clustering by DAVID are presented in Table 4.6. Results were reported only for those clusters which had an enrichment score  $\geq 1$ .

The CMP data were previously used in a within-breed analysis of WBSF published by McClure *et al.* (2012). They utilized a GBLUP model incorporating a genomic relationship matrix to examine WBSF SNP effects within each of the five breeds used in this analysis. A summary of the concordance of QTL detected between the two analyses is in Table 4.7. The SNP with the largest effect in each QTL window reported by McClure *et al.* (2012) was compared to the list of SNPs for each analysis containing the top 1% largest SNP effects. Of all QTL detected in McClure *et al.* (2012),

Table 4.5: Number of genes in each gene list that overlap between different analysis models. DAVID analysis included all genes that were concordant in 3 of 4 or 2 of 3 analyses. Percentages are a reflection of the number of genes in each category divided by the total number of genes in each cumulative gene lists for an individual trait.

Trait	2 of 4	3 of 4	2 of 3	All
WBSF	2140/60.2%	1390/39.1%		642/18.1%
REA	2200/77.6%	1574/55.5%		1114/39.3%
MARB	2918/83.3%	2540/72.5%		2126/60.7%
FT	1866/48.1%	1267/32.6%		539/13.9%
HCW	2368/78.8%	1730/57.6%		1146/38.2%
YG	2607/56.0%	1921/41.3%		471/10.1%
Tender			2584/79.2%	1713/52.5%
Juicy			2250/85.6%	2043/77.7%
Flavor			2309/81.0%	1952/68.4%
ConnTiss			2169/59.4%	1081/29.6%

Table 4.6: Functional annotation clustering results from DAVID analyses of cumulative concordant gene lists for each trait.

Trait	Enrichment Score	Count	Functional Annotation Cluster	p-value	Benjamini
WBSF	3.11	18	Enzyme inhibitor activity	5.60E-03	6.30E-01
	1.92	22	Cell cycle	3.80E-02	9.60E-01
	1.46	3	Lipid-binding serum glycoprotein, conserved site	8.50E-03	9.00E-01
	1.42	7	Transition metal ion transport	3.70E-02	9.70E-01
	1.28	7	Winged helix repressor DNA-binding	1.80E-01	1.00E+00
	1.26	9	Anchored to membrane	4.50E-02	9.60E-01
	1.23	8	Ectoderm development	1.90E-02	9.80E-01
	1.19	26	Macromolecular complex subunit organization	2.30E-02	9.60E-01
	1.17	24	Structural molecule activity	2.80E-01	9.90E-01
	1.12	17	Zinc finger, C2H2-type	5.40E-02	9.70E-01
	1.10	4	p53 signaling pathway	6.80E-01	9.60E-01
	1.05	5	Ubiquitin conserved site	1.20E-02	9.10E-01
	1.02	7	Copper ion binding	1.90E-02	8.20E-01
	1.02	89	Nucleus	1.20E-02	4.70E-01
	1.01	27	Cation transport	6.40E-02	9.80E-01
1.00	145	Ion binding	5.40E-02	9.70E-01	
REA	1.77	7	Glycolysis / Gluconeogenesis	1.40E-01	6.70E-01
	1.49	52	Plasma membrane part	3.70E-02	7.80E-01
	1.46	5	Nucleobase, nucleoside, nucleotide and nucleic acid transport	1.50E-01	9.90E-01
	1.42	6	Regulation of muscle contraction	1.40E-02	9.60E-01

	1.41	11	Metabolism of xenobiotics by cytochrome P450	2.40E-04	3.70E-02
	1.38	9	Extrinsic to membrane	6.10E-01	9.70E-01
	1.37	20	Lipoprotein	1.40E-01	9.70E-01
	1.15	10	Negative regulation of signal transduction	5.60E-02	9.90E-01
	1.11	28	Ribonucleoprotein complex	4.90E-02	7.40E-01
	1.08	8	Adult behavior	1.80E-02	9.60E-01
	1.06	7	Transferase activity, transferring alkyl or aryl (other than methyl) groups	4.30E-02	8.80E-01
MARB	5.35	71	Calcium ion binding	3.80E-03	8.00E-01
	1.48	4	Longevity assurance, LAG1/LAC1	9.10E-03	9.60E-01
	1.42	28	Cofactor binding	3.10E-02	9.80E-01
	1.39	23	NAD(P)-binding domain	7.60E-03	9.80E-01
	1.37	23	ncRNA metabolic process	1.60E-02	1.00E+00
	1.29	11	Cadherin	2.80E-04	1.80E-01
	1.21	43	Ribonucleoprotein complex	1.80E-01	9.30E-01
	1.17	4	Endoglin/CD105 antigen	4.00E-02	1.00E+00
	1.17	136	Regulation of transcription	5.40E-02	1.00E+00
	1.09	15	Regulation of cellular protein metabolic process	8.00E-01	1.00E+00
	1.06	7	Glycosyltransferase	9.70E-01	1.00E+00
	1.05	23	Ubl conjugation	2.30E-01	9.00E-01
FT	6.96	24	Peptidase inhibitor activity	1.00E-08	1.80E-06
	4.71	20	Lipid binding	5.20E-03	4.20E-01
	1.99	6	Serine protease inhibitor	1.30E-02	9.80E-01
	1.85	5	Alcohol dehydrogenase GroES-like	4.50E-03	5.30E-01
	1.6	23	Cytoplasmic vesicle	4.20E-03	6.80E-01

	1.39	8	Wnt receptor signaling pathway	1.30E-02	1.00E+00
	1.16	4	Transferase activity, transferring amino-acyl groups	3.10E-02	8.70E-01
	1.12	63	Intracellular non-membrane-bounded organelle	7.00E-02	9.60E-01
	1.07	5	Ubiquitin-conjugating enzyme/RWD-like	1.20E-01	9.90E-01
HCW	2.99	18	Lipid binding	2.60E-01	9.90E-01
	1.69	31	Ribonucleoprotein complex	6.10E-02	8.80E-01
	1.54	8	Duplication	1.90E-01	9.60E-01
	1.5	8	Peroxisome	3.30E-02	7.60E-01
	1.46	6	Glycosyltransferase	8.60E-01	9.90E-01
	1.27	9	Arachidonic acid metabolism	2.20E-02	6.10E-01
	1.21	28	Cytoplasmic vesicle	1.20E-02	5.40E-01
	1.19	21	Phosphatase activity	3.10E-02	7.90E-01
	1.13	41	Ion transport	1.90E-01	1.00E+00
	1.11	8	ATPase, AAA+ type, core	1.40E-01	1.00E+00
	1.01	9	Embryonic morphogenesis	5.70E-01	1.00E+00
YG	2.80	30	Immunoglobulin-like	6.10E-04	1.60E-01
	1.57	51	Proteolysis	2.20E-01	1.00E+00
	1.40	6	Cysteine and methionine metabolism	8.50E-02	9.50E-01
	1.24	8	Proteasome complex	6.00E-02	1.00E+00
	1.19	5	Transferase activity, transferring amino-acyl groups	1.80E-02	9.80E-01
	1.13	100	Cell surface receptor linked signal transduction	7.80E-01	1.00E+00
	1.11	54	Membrane-enclosed lumen	9.60E-02	9.50E-01
	1.01	23	Ubl conjugation	1.10E-02	8.70E-01

	1.01	8	PDZ/DHR/GLGF	1.10E-01	1.00E+00
Tender	5.34	13	Somatotropin hormone	1.90E-10	2.40E-07
	3.18	190	Cell surface receptor linked signal transduction	4.70E-07	5.30E-04
	2.25	18	Defense response	8.50E-01	1.00E+00
	2.20	44	Macromolecular complex subunit organization	5.40E-03	6.40E-01
	2.00	13	CLECT	1.70E-03	1.30E-01
	1.75	11	Metallopeptidase activity	8.40E-01	1.00E+00
	1.73	25	Ubl conjugation	5.90E-02	7.20E-01
	1.63	22	Actin binding	1.80E-02	9.20E-01
	1.43	9	Nucleobase, nucleoside, nucleotide and nucleic acid transport	1.40E-02	8.90E-01
	1.39	4	Protein synthesis factor, GTP-binding	1.00E-01	9.90E-01
	1.12	8	Regulation of actin filament-based process	2.70E-02	9.50E-01
	1.06	6	Cellular aldehyde metabolic process	1.70E-02	8.90E-01
	1.03	6	Zinc finger, CCCH-type	8.40E-02	9.80E-01
Juicy	2.34	29	Ubl conjugation	8.00E-03	4.90E-01
	1.98	47	Ribonucleoprotein complex	5.30E-03	5.00E-01
	1.60	37	Homeostatic process	2.90E-01	1.00E+00
	1.58	36	Vesicle	2.10E-02	6.10E-01
	1.47	4	Regulation of phosphoprotein phosphatase activity	2.10E-02	1.00E+00
	1.38	3	Alpha(1,2)-fucosyltransferase activity	2.30E-02	1.00E+00
	1.28	125	Non-membrane-bounded organelle	1.50E-02	6.40E-01
	1.27	10	Ribosomal subunit	3.80E-02	6.30E-01
	1.25	15	Chromatin modification	5.10E-02	1.00E+00

	1.18	13	Blood circulation	2.90E-02	1.00E+00
	1.16	31	Chromatin organization	6.90E-03	9.50E-01
Flavor	2.62	176	Nucleotide binding	6.40E-05	1.60E-02
	2.60	28	Cellular homeostasis	1.50E-02	7.90E-01
	2.45	4	Fibroblast growth factor receptor antagonist activity	2.20E-03	1.90E-01
	2.39	44	Ribonucleoprotein complex	3.00E-03	3.00E-01
	1.90	4	Interleukin-1	6.30E-03	9.80E-01
	1.56	74	Plasma membrane part	1.90E-02	5.80E-01
	1.55	44	Mitochondrial part	1.50E-02	6.10E-01
	1.55	30	Membrane-bounded vesicle	6.90E-03	4.60E-01
	1.37	17	Organic acid biosynthetic process	4.40E-03	6.60E-01
	1.35	6	DEATH	1.00E-02	5.20E-01
	1.29	8	Metal cluster binding	5.70E-02	8.30E-01
	1.21	87	Mitochondrion	1.50E-04	5.40E-02
	1.14	12	Arginine and proline metabolism	2.30E-03	3.30E-01
	1.13	12	Growth factor activity	1.40E-01	9.30E-01
	1.13	17	Response to abiotic stimulus	8.60E-02	9.50E-01
	1.07	12	FAD	3.90E-03	4.20E-01
	1.05	9	Carboxylic acid binding	1.30E-01	9.50E-01
	1.04	110	Non-membrane-bounded organelle	3.30E-02	6.40E-01
	1.01	24	Neurological system process	3.50E-01	9.90E-01
Conn	7.75	55	immune response	1.40E-08	1.60E-05
Tiss	3.74	30	carbohydrate binding	3.70E-06	2.60E-03
	3.05	11	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	1.70E-07	1.30E-04

2.63	10	MHC class I protein complex	1.40E-02	6.50E-01
2.26	16	C-type lectin-like	1.30E-04	1.80E-02
2.14	29	negative regulation of cellular biosynthetic process	1.80E-03	3.70E-01
1.84	26	defense response	7.50E-02	9.40E-01
1.79	7	cartilage development	5.90E-02	9.10E-01
1.65	24	Cognition	1.20E-02	5.50E-01
1.55	8	SPla/RYanodine receptor SPRY	1.20E-02	5.80E-01
1.34	34	Translation	2.90E-02	8.20E-01
1.28	19	Immunoglobulin subtype	2.90E-02	7.80E-01
1.18	9	lung development	4.20E-02	8.80E-01
1.17	13	Tetratricopeptide-like helical	8.90E-02	9.10E-01
1.14	35	cell adhesion	4.80E-02	8.80E-01
1.09	8	Thrombospondin, type 1 repeat	6.40E-03	4.90E-01
1.08	12	skeletal system morphogenesis	2.20E-03	4.00E-01
1.07	6	cellular aldehyde metabolic process	1.00E-02	5.60E-01
1.05	21	chordate embryonic development	4.70E-02	8.80E-01
1.04	13	Vacuole	2.20E-01	8.90E-01
1.01	10	Thioredoxin fold	3.40E-01	1.00E+00

---

Table 4.7: QTL regions detected by McClure *et al.* (2012) using within-breed analysis approaches that were also concordant with across-breed analysis methods used in this study for WBSF and Tender.

BTA	SNP	Tender Bayes B0*	Tender Bayes B95*	Tender Bayes C0*	WBSF Bayes B0*	WBSF Bayes B95*	WBSF Bayes C $\pi$ *	WBSF Bayes C0*
3	<i>ss86301348</i>				18	18		15
4	<i>rs43403458</i>				31	30		27
5	<i>rs29014779</i>	42	43	38				
5	<i>rs41654473</i>				39	38		34
6	<i>rs42756258</i>				42	41		37
6	<i>ss117968229</i>				45	44		40
7	<i>rs29012174</i>				51	50	51	45
7	<i>ss86318554</i>				52	51	52	46
7	<i>rs43527386</i>				52	51	53	46
7	<i>rs41255587</i>	60	61	55	53	53	54	47
7	<i>rs43531510</i>				54	54		48
8	<i>rs41618019</i>				55			
8	<i>rs42312419</i>				57	56	57	51
8	<i>ss117969253</i>				60	59		53
8	<i>ss86319219</i>				62	61	63	55
8	<i>ss86338099</i>				63	61		
9	<i>rs41623216</i>				65	63		56
10	<i>ss86317957</i>				72	69		63
10	<i>ss86305679</i>				73	70	75	65
10	<i>rs42412333</i>							66
10	<i>rs41590854</i>				78	77	85	69

10	<i>rs41596899</i>				79			70
11	<i>rs41606137</i>				82	81		73
12	<i>ss117970656</i>				88	86	94	79
12	<i>rs43699567</i>				89	87	95	80
13	<i>rs42862024</i>				92	90		83
13	<i>rs29011158</i>				95	93	101	87
13	<i>rs41631563</i>				99	98	103	91
13	<i>ss86338902</i>				100	99		92
13	<i>ss86289318</i>				100	99	104	92
13	<i>rs42630433</i>				101	99		93
14	<i>rs41633333</i>				103	101		95
14	<i>ss86297726</i>				107	105		99
15	<i>ss86291817</i>				112	111	112	104
15	<i>rs41757680</i>							104
15	<i>rs41582705</i>							105
15	<i>rs41621125</i>				114	112	113	107
15	<i>ss86314348</i>	113	119	109	114	112	113	107
15	<i>ss86296417</i>				116	114	115	109
16	<i>rs41623175</i>				119	117		111
16	<i>ss86290236</i>				120	118		112
16	<i>ss86329907</i>				120	118	118	112
16	<i>ss86291490</i>				120	118		112
16	<i>rs41824081</i>				123	121	122	115
17	<i>rs41626299</i>							117
17	<i>ss86317522</i>				131			120
18	<i>ss86336538</i>				133	131	129	122

20	<i>rs41933103</i>				142	139	135	130
20	<i>ss86335963</i>				148	144		136
21	<i>rs29015146</i>				149	146		137
21	<i>rs42503056</i>				150	147	141	138
21	<i>rs41585245</i>				152	148		139
21	<i>ss86312849</i>				153	149	145	140
23	<i>rs41617911</i>							148
25	<i>ss117973580</i>				161	159	153	152
25	<i>ss86336453</i>				162	160		153
25	<i>rs41572366</i>				163	161	155	154
25	<i>ss86283327</i>				164	162	156	155
26	<i>ss86273489</i>				166	165		157
26	<i>ss86287439</i>				167	166	157	158
26	<i>rs41646897</i>				168	167		159
26	<i>ss86282954</i>							160
27	<i>rs42118878</i>				172	171	161	162
27	<i>ss86310277</i>				173	173		
28	<i>rs41612729</i>				175	174		164
28	<i>ss86337100</i>	175	182	175	177	177	166	166
28	<i>rs29013966</i>				178	178	167	167
28	<i>ss86283362</i>				179	179		168
29	<i>rs29022154</i>				184	184		172
29	<i>rs42192103</i>				185	185	172	173

---

\*Denotes the QTL region ID from this study was concordant with those detected in at least 3 breeds in McClure *et al.* 2012

70 of 79 (88.6%) were detected in at least one analysis type for Tender or WBSF. Of the WBSF analyses, the greatest concordance was found between regions detected using GBLUP and BayesC0 (66/79, 83.5%), as expected, and the lowest concordance was between BayesC $\pi$  and GBLUP. If a direct match is not a requisite, all but three of the regions between the two studies overlap for at least one analysis of WBSF or Tender.

### *Conclusions*

This study identified a large number of putative QTL regions for economically important carcass traits for future validation. In addition, it is one of the first studies to report QTL for sensory panel traits in beef cattle. Our analysis identified QTL regions that explain approximately 2-18% (9-88% corrected) of the total additive genetic variance in these traits. Gene concordance within QTL regions identified by different types of analyses was generally high, and tended to be highest for those analyses where  $\pi$  was smallest (i.e., for those traits that had the largest number of influential SNPs). Concordance of gene lists across traits showed evidence for the greatest pleiotropy between those traits which have the strongest genetic correlations, as expected. The overall concordance with the within-breed analyses validated our QTL discovery approach across multiple breeds of beef cattle. Analysis using several Bayesian models also afforded a deeper understanding of the genomic architecture of these traits in beef cattle.

## CHAPTER V: SUMMARY AND CONCLUSIONS

The use of genetic distance between breeds to partition training and validation populations for across-breed genomic selection was ineffective in this study. It is possible that the underlying relatedness of the animals within each of the breeds represented in these data and the common Angus background among the Continental crossbred calves prevented an accurate assessment of genetic distance or partitioning into discrete validation populations, in contrast to what can be achieved in a simulated population. The use of coefficients of relationship from a genomic relationship matrix was also found to be ineffective for partitioning animals into training and validation populations. It is plausible that this result may be due to the extensive paternal half-sib structure of the pedigree for the animals within this study, which prevented significant differentiation of the animals within the training and validation groups.

Random allocation of animals into training and validation sets proved to be very effective and MBV prediction models were developed across five commercially relevant beef cattle breeds for six economically relevant carcass traits. Realized accuracies ranged from 0.40 (FT) up to 0.78 (HCW) across all traits and analyses. The underlying architecture of the trait greatly influenced prediction accuracy and model superiority.

BayesC $\pi$  analyses were found to be superior when genes of large effect were present for a particular trait and these effects were not breed-specific. All models performed similarly when the architecture of the traits were consistent with the infinitesimal model. BayesB0 analyses performed consistently well regardless of the underlying architecture of the trait provided priors for variance components were first estimated from a BayesC analysis and should be considered the “gold standard” analysis when little is known about the architecture of a trait.

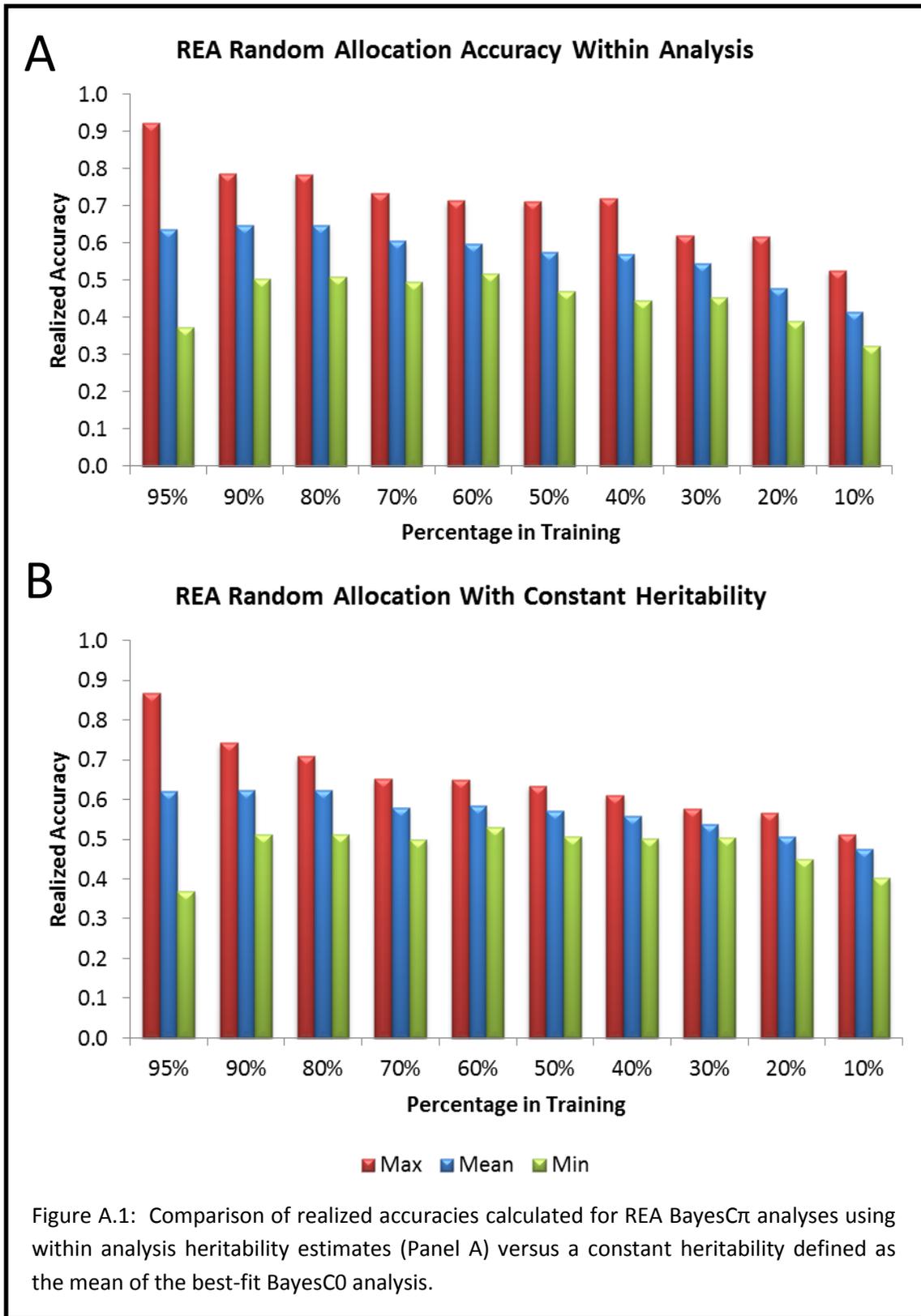
External validation of the WBSF models resulted in massive decreases in realized accuracies, even when the populations were separated by only approximately 10 years (~2 generations) in time. Inclusion of approximately one half of the animals from the external population into the training population did not alleviate the problem, but this result is possibly due to the extremely small sample size for the animals in the external validation population. These results suggest that the genomic selection models currently being utilized in the beef industry may have to be retrained more often than previously planned, possibly with every new calf crop.

Across-breed genomic selection should be further studied for use in breed association national cattle evaluation. This could result in more efficient and accurate analyses of hybrid populations in the breed registry, as well as enhancing the ability of smaller breed associations to perform joint genomic analyses, thus saving money and duplication of effort within the industry. The utilization of across-breed genomic selection models for genetic prediction in the commercial beef industry is also a viable

opportunity. While the beef industry does not yet have the infrastructure to support these types of analyses, its implementation should be further explored. Further studies of genomic technologies in crossbred and admixed commercial populations will be essential to gain the knowledge needed to implement these technologies in the industry and capitalize on the increased genetic variation and number of carcass phenotypes present in the commercial sector.

Finally, we discovered QTL regions which explain large proportions of variation in several economically important carcass traits in beef cattle and, more uniquely, for four sensory panel traits vital to increasing consumer satisfaction with beef products. These results will serve as an independent validation of previous and future studies. Moreover, a large-effect QTL for FT that appears to be Angus-specific was discovered and we have proposed a possible candidate gene for further study towards the identification of the causal mutation.

APPENDIX



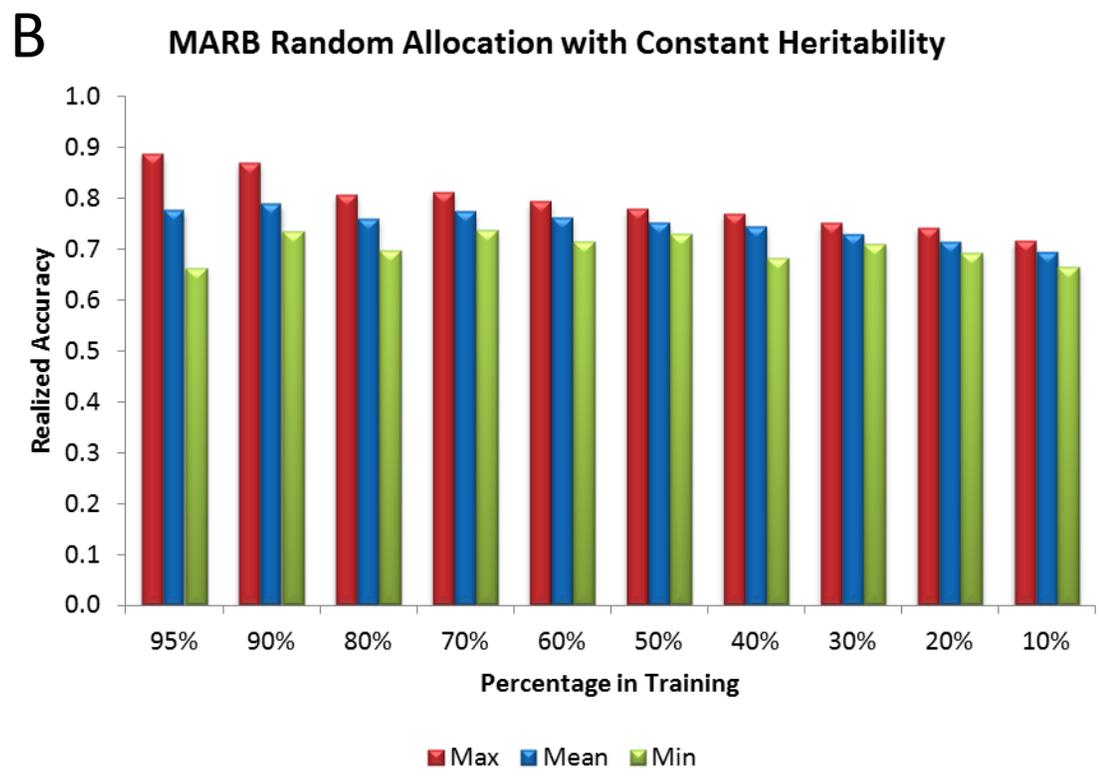
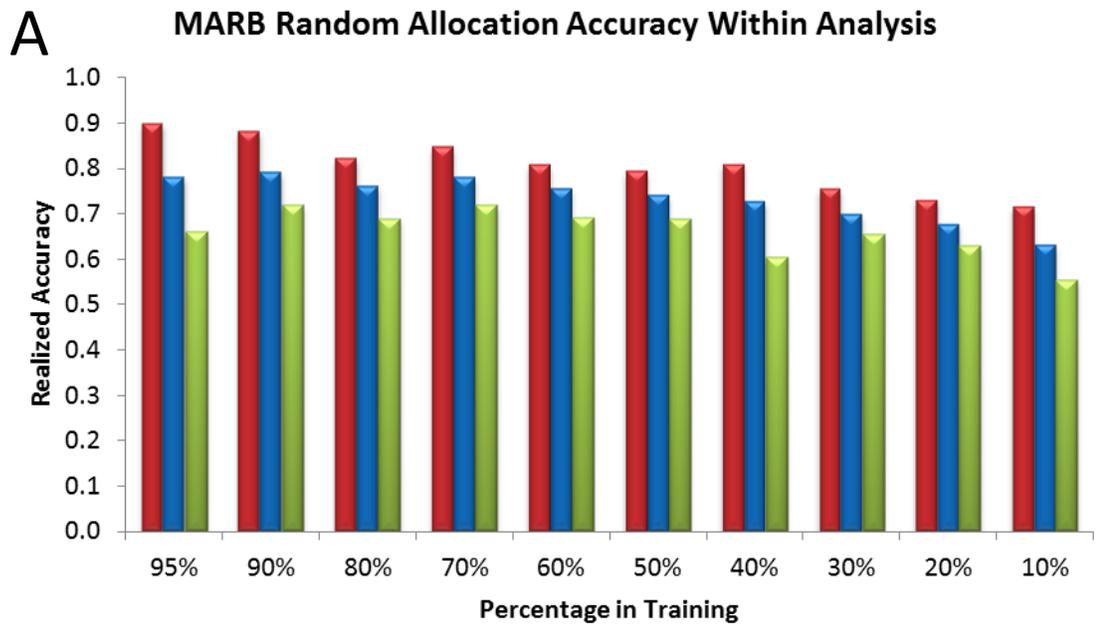


Figure A.2: Comparison of realized accuracies calculated for MARB BayesC $\pi$  analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.

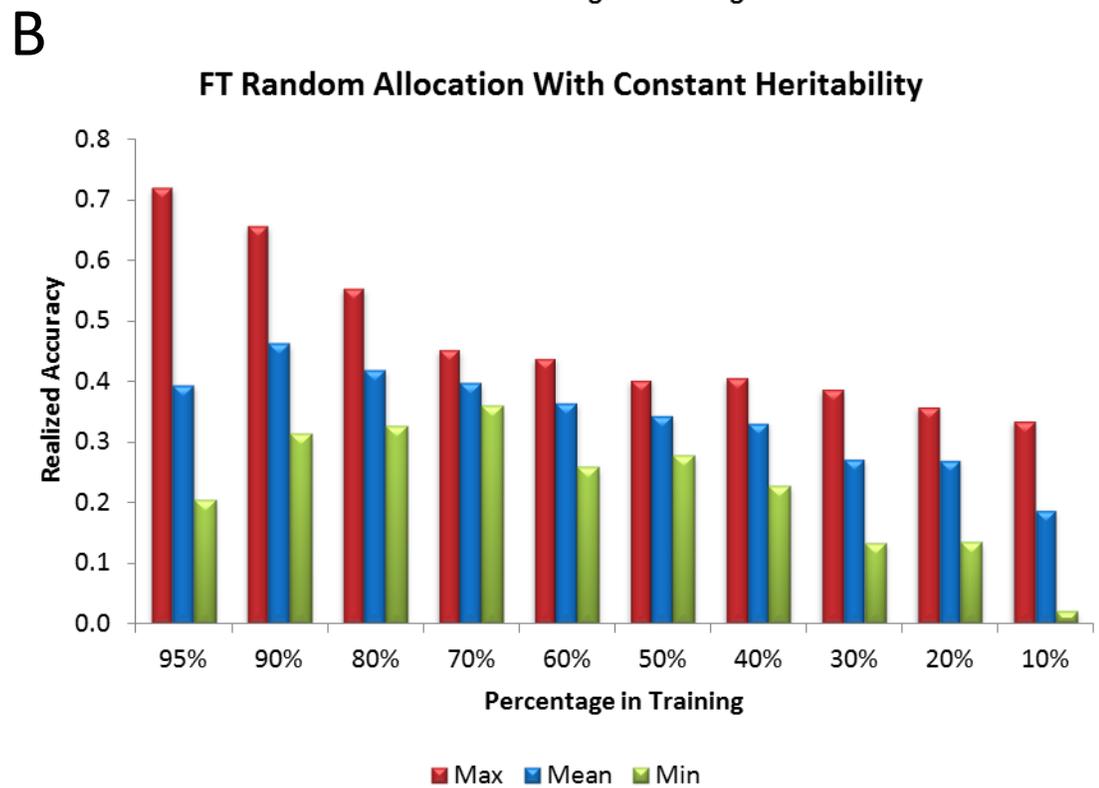
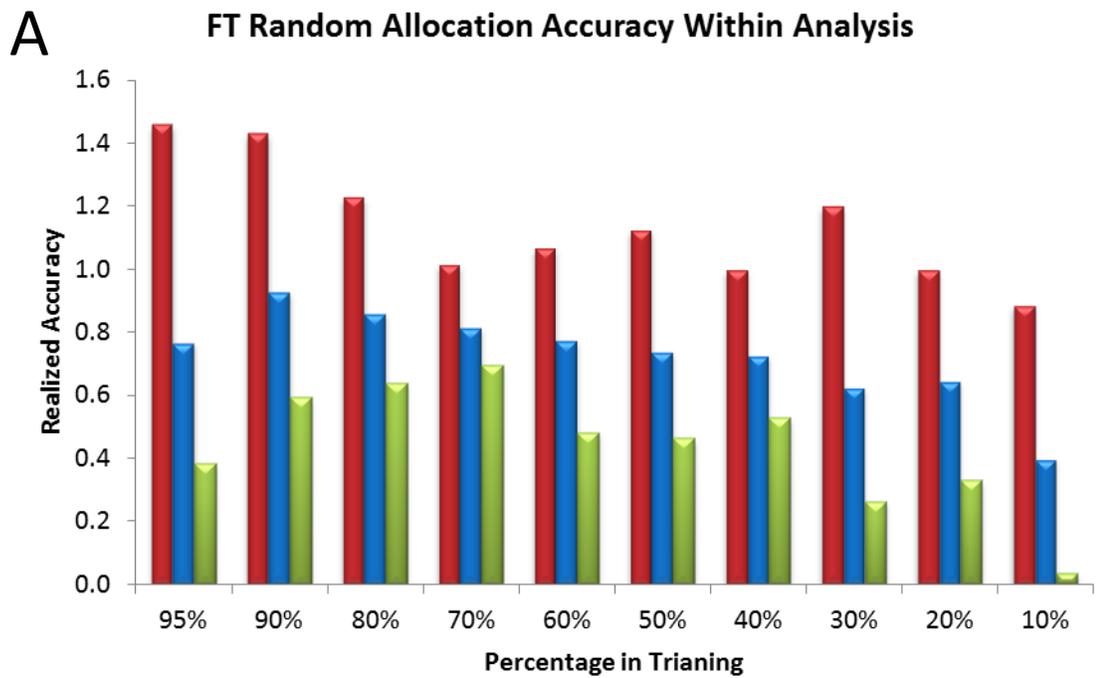
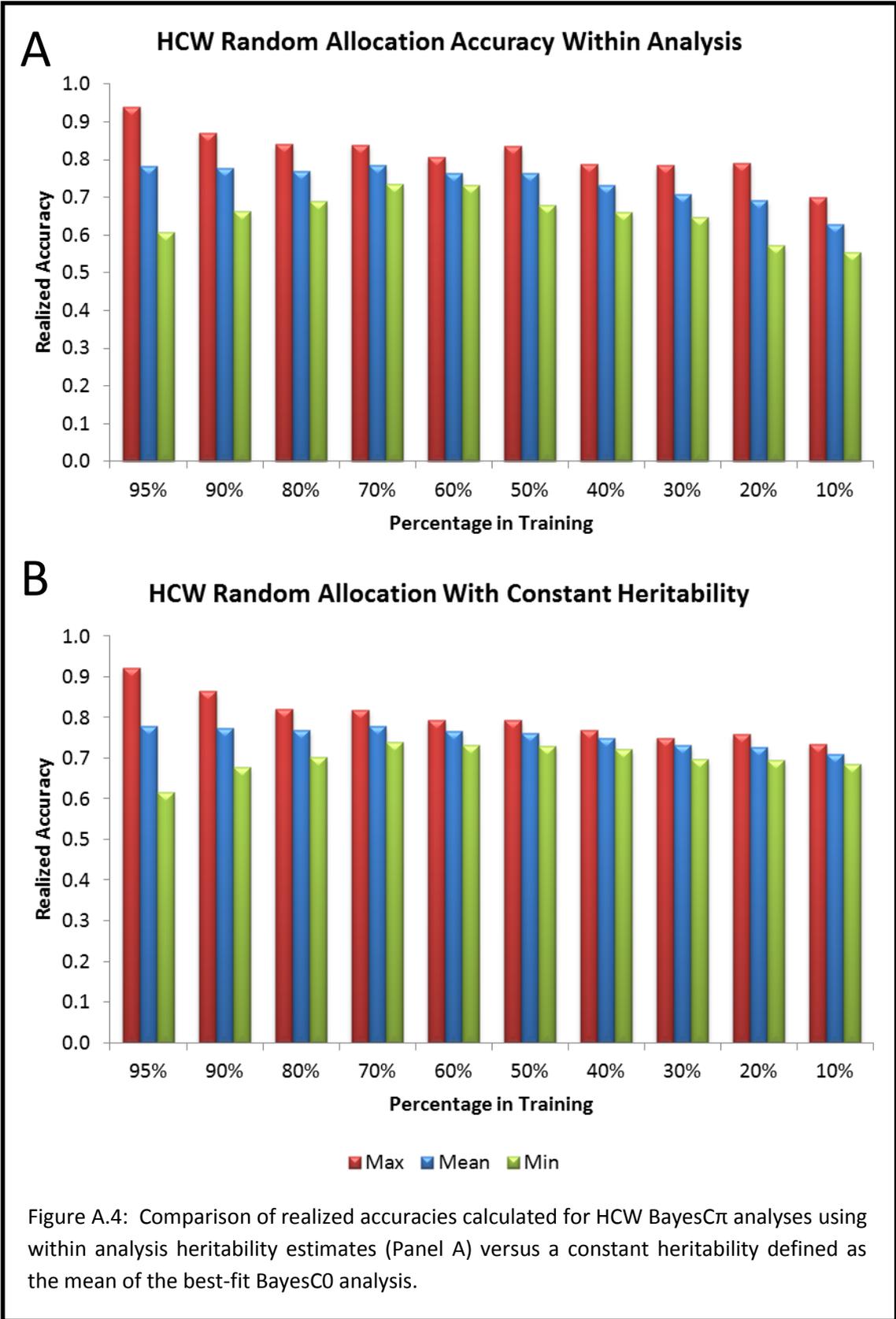


Figure A.3: Comparison of realized accuracies calculated for FT BayesC $\pi$  analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.



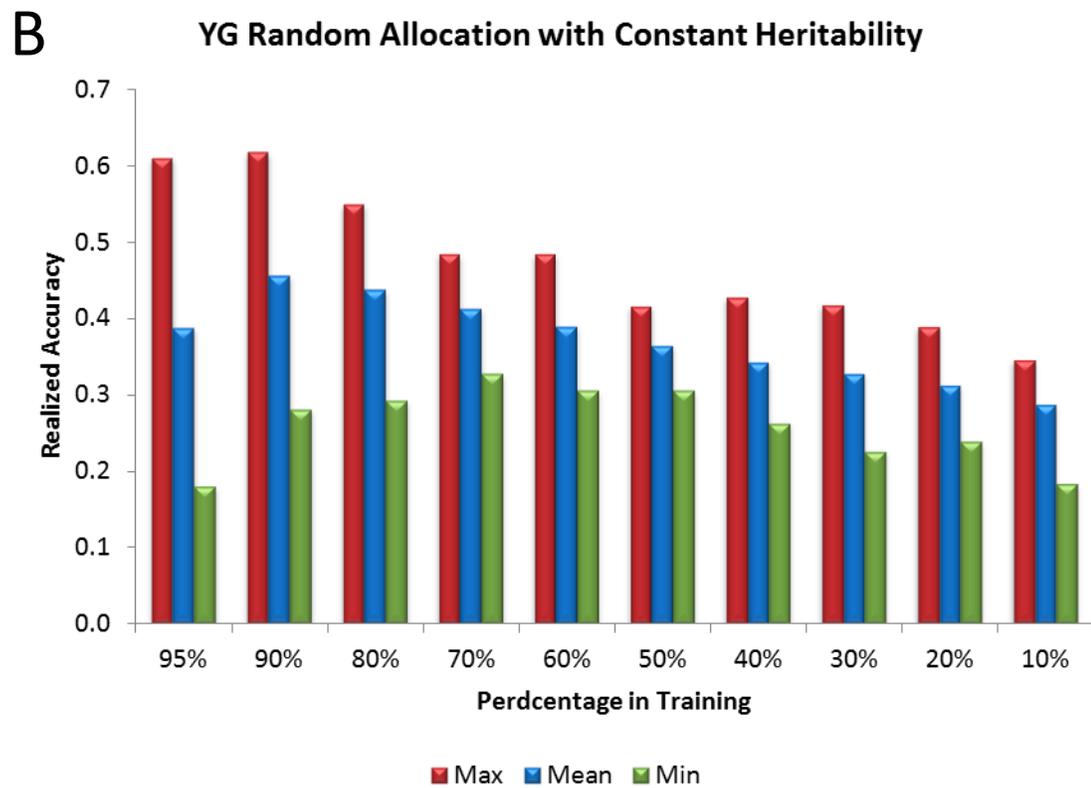
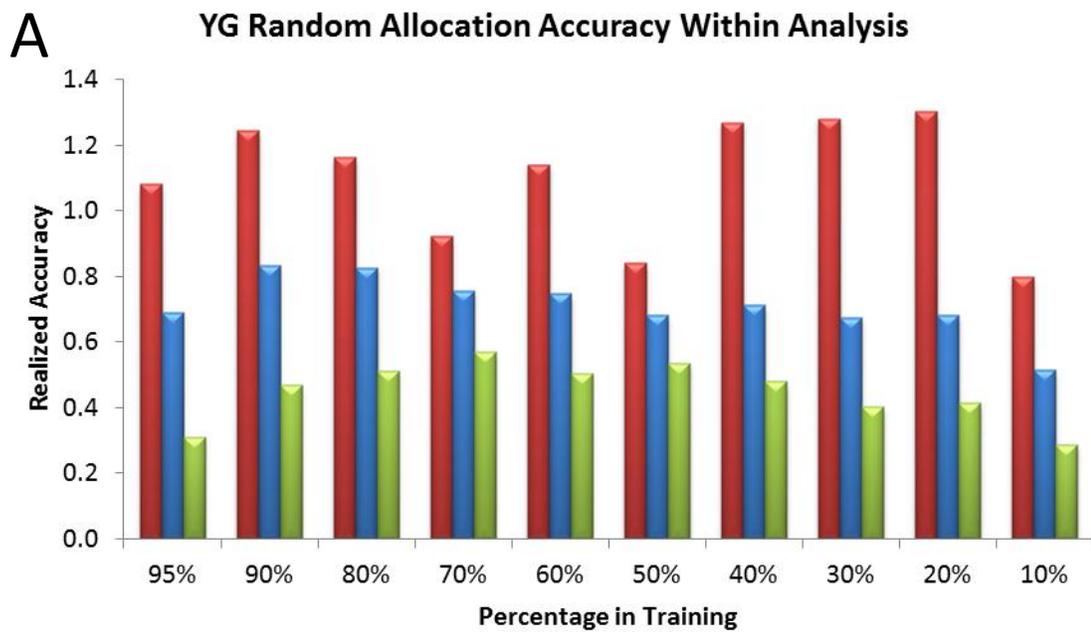
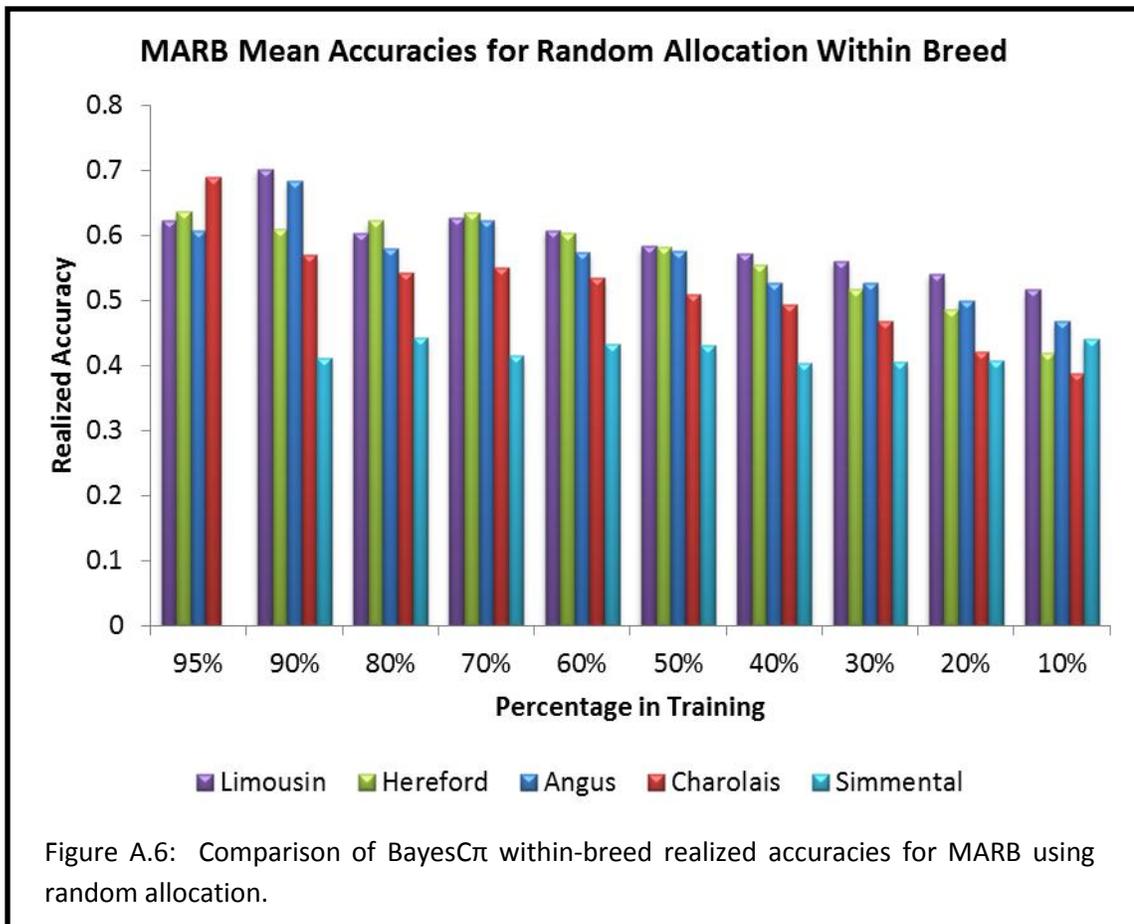
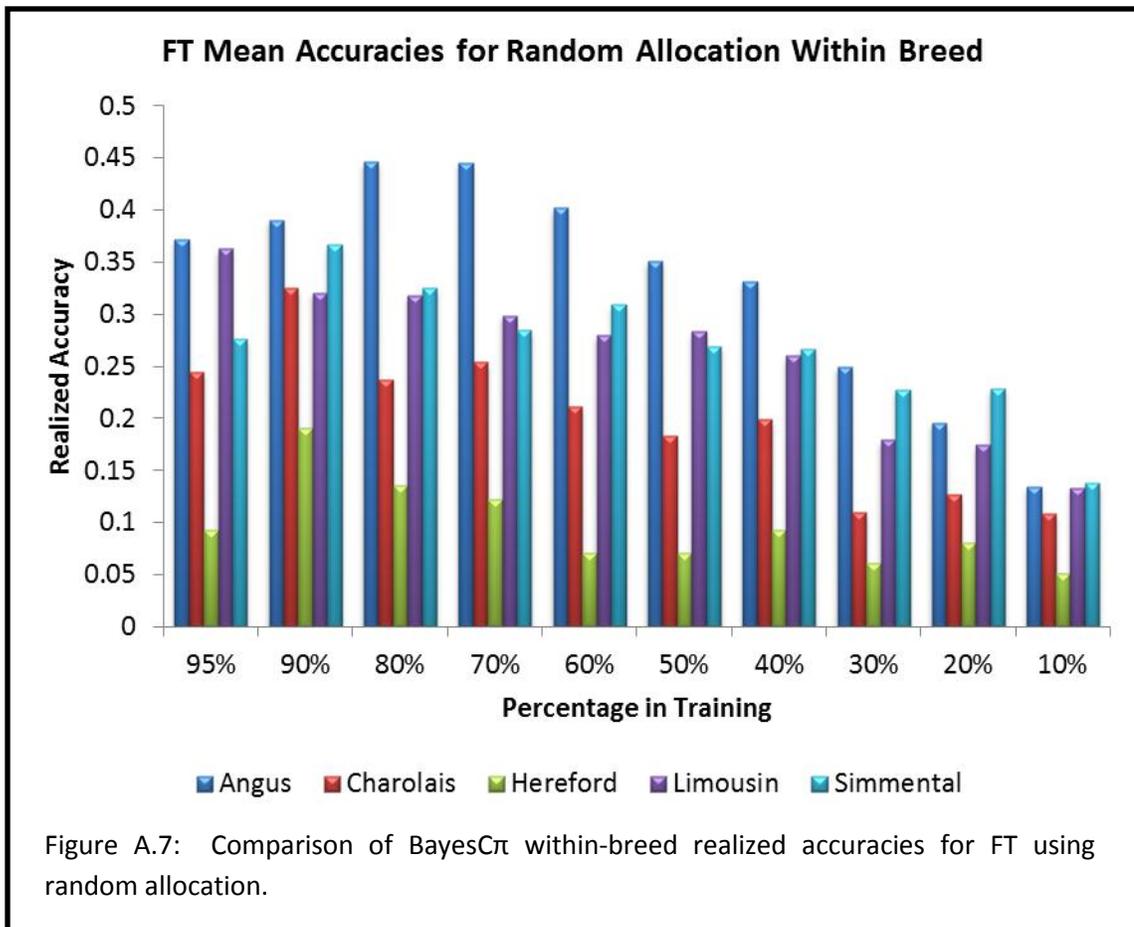
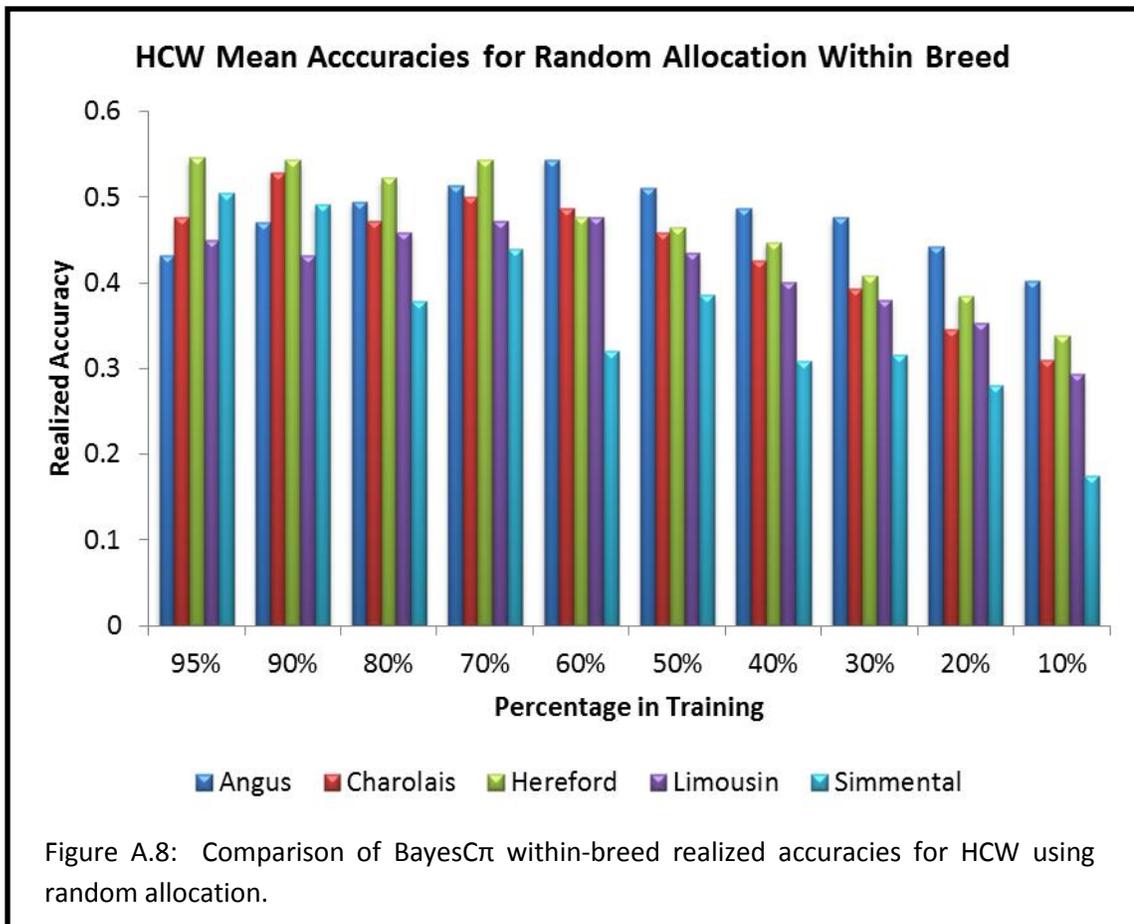
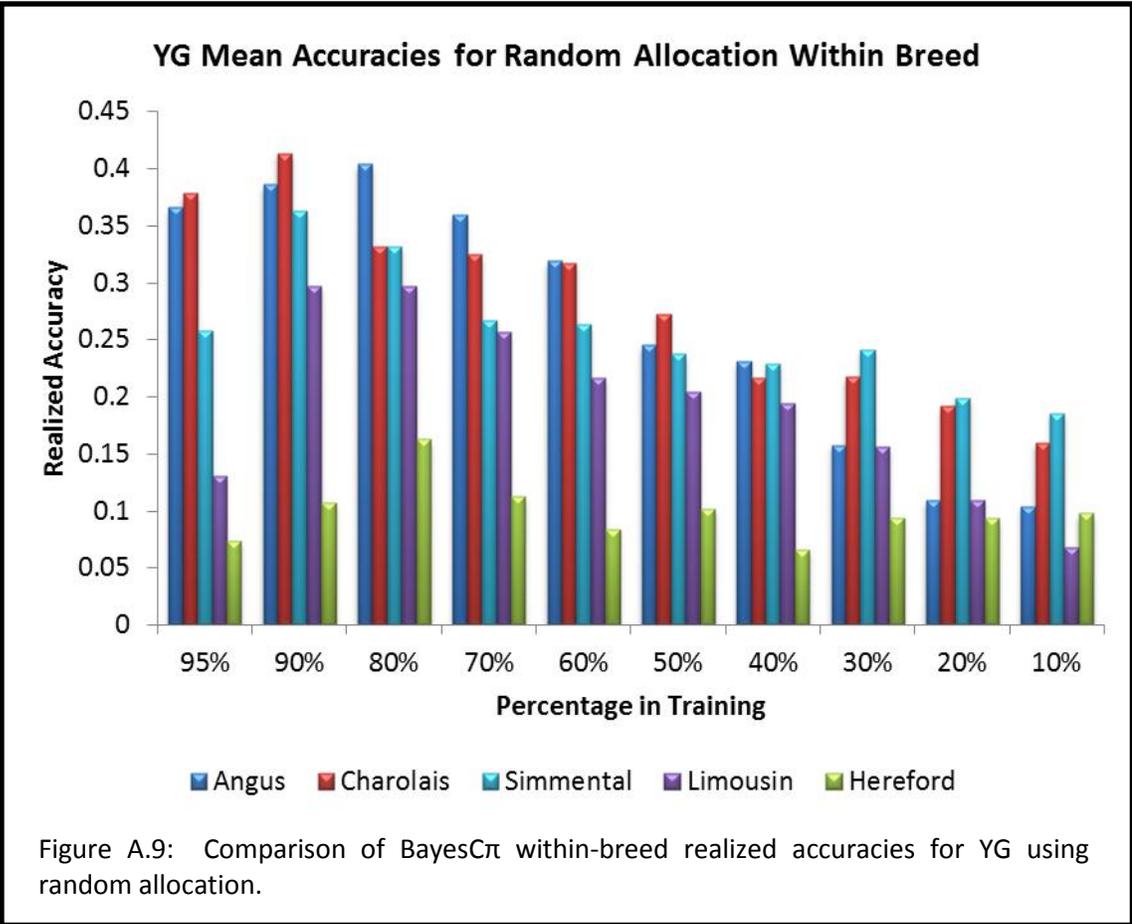


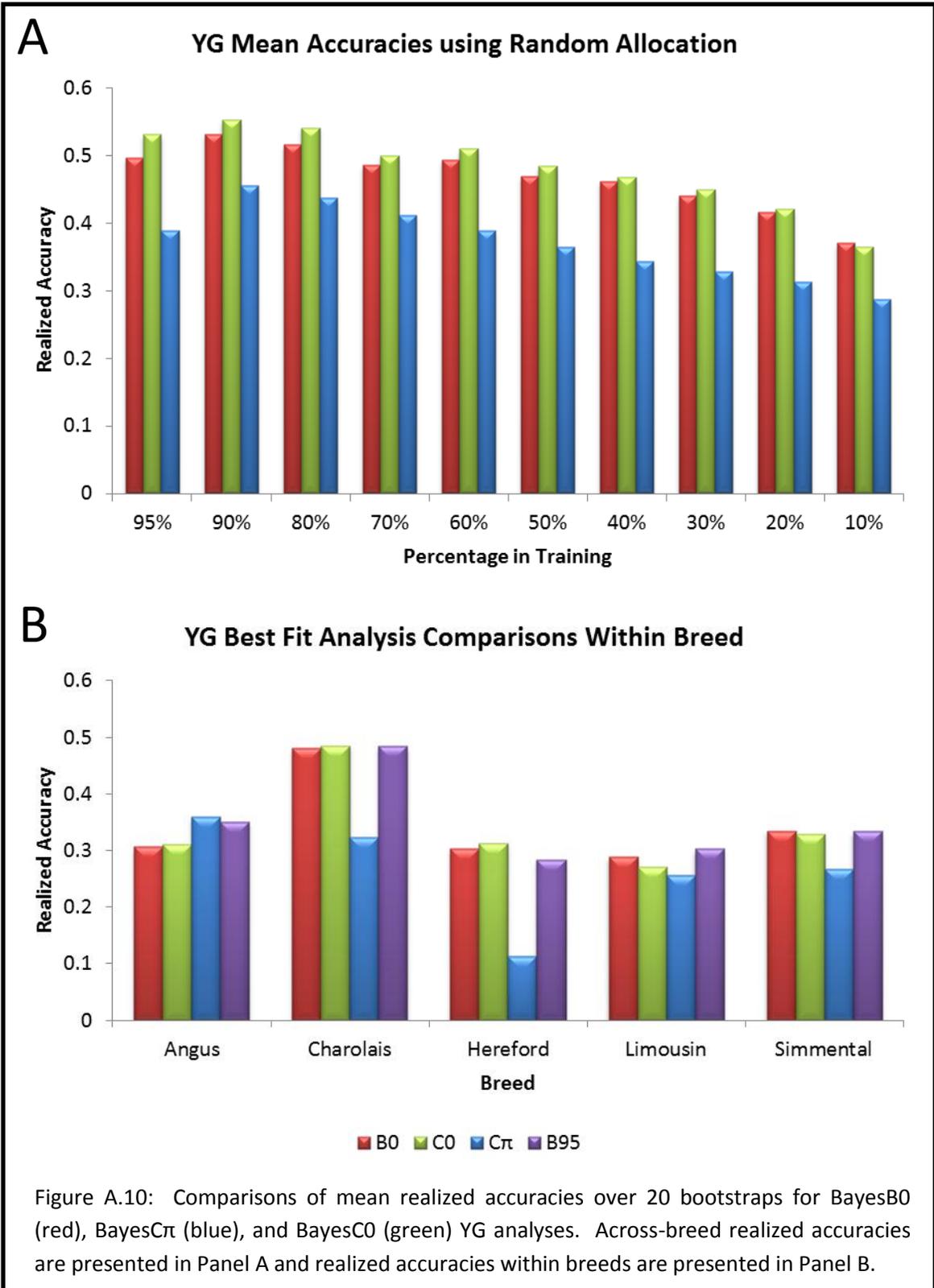
Figure A.5: Comparison of realized accuracies calculated for YG BayesCπ analyses using within analysis heritability estimates (Panel A) versus a constant heritability defined as the mean of the best-fit BayesC0 analysis.

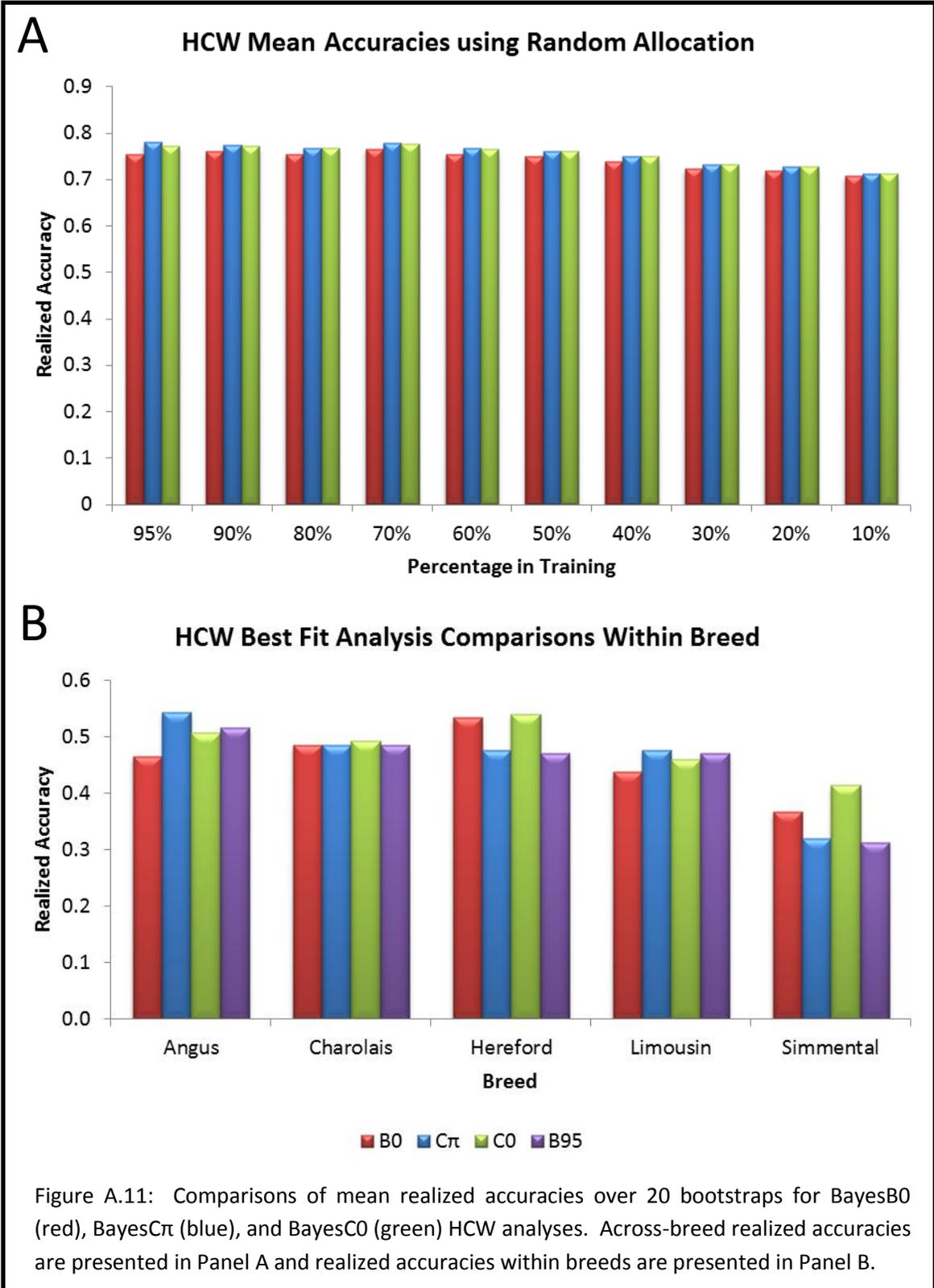


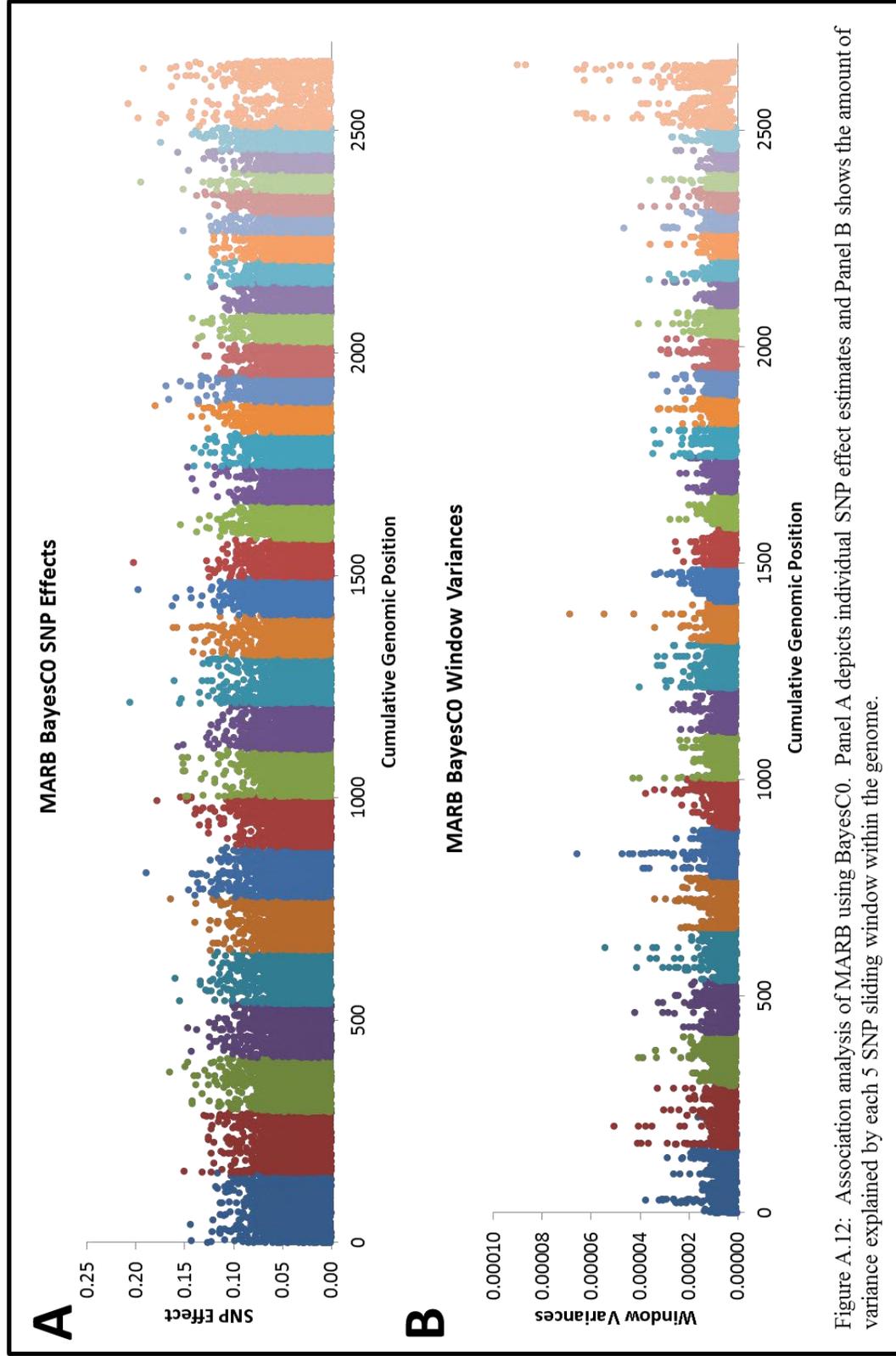












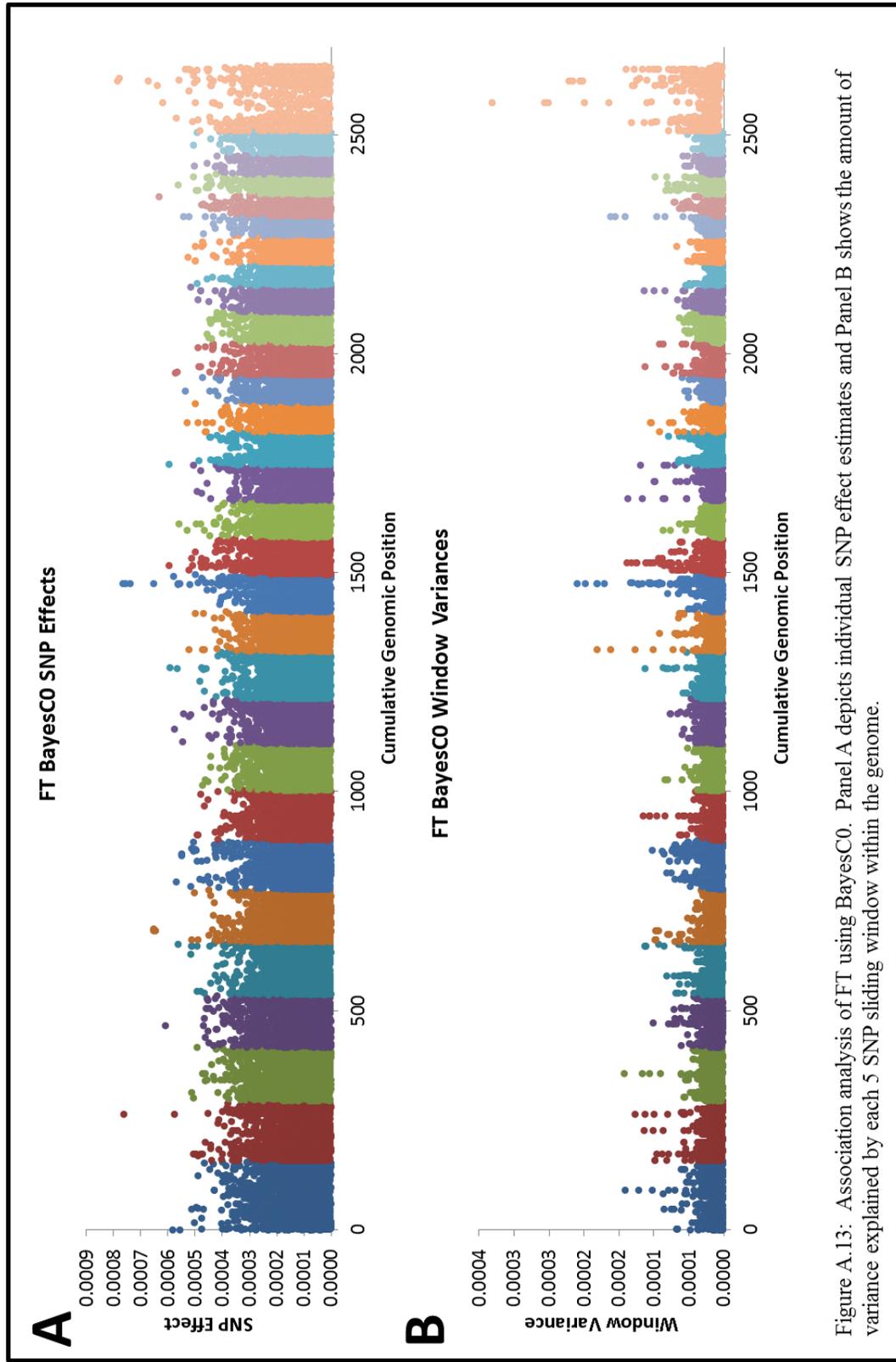


Figure A.13: Association analysis of FT using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

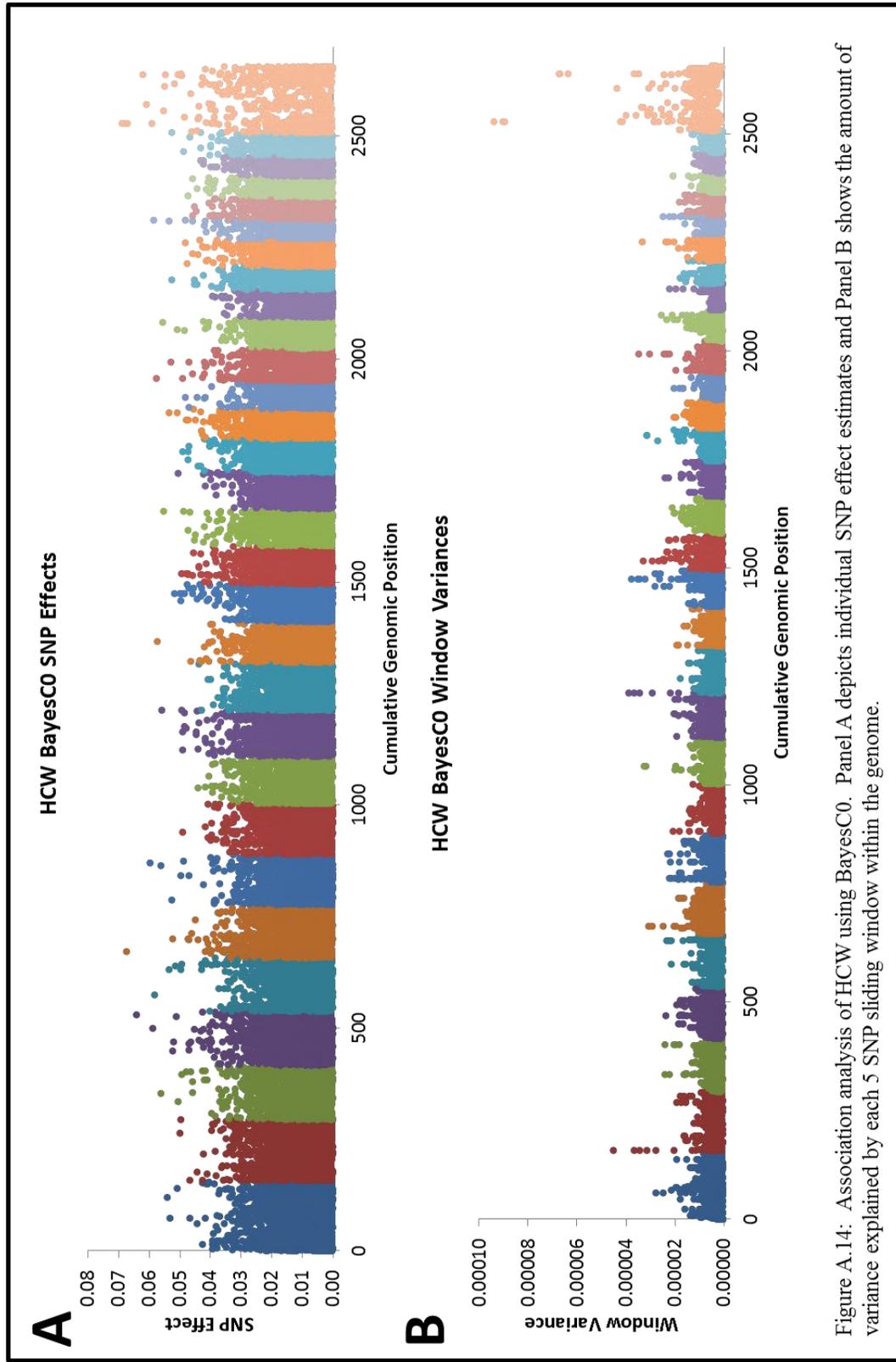


Figure A.14: Association analysis of HCW using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

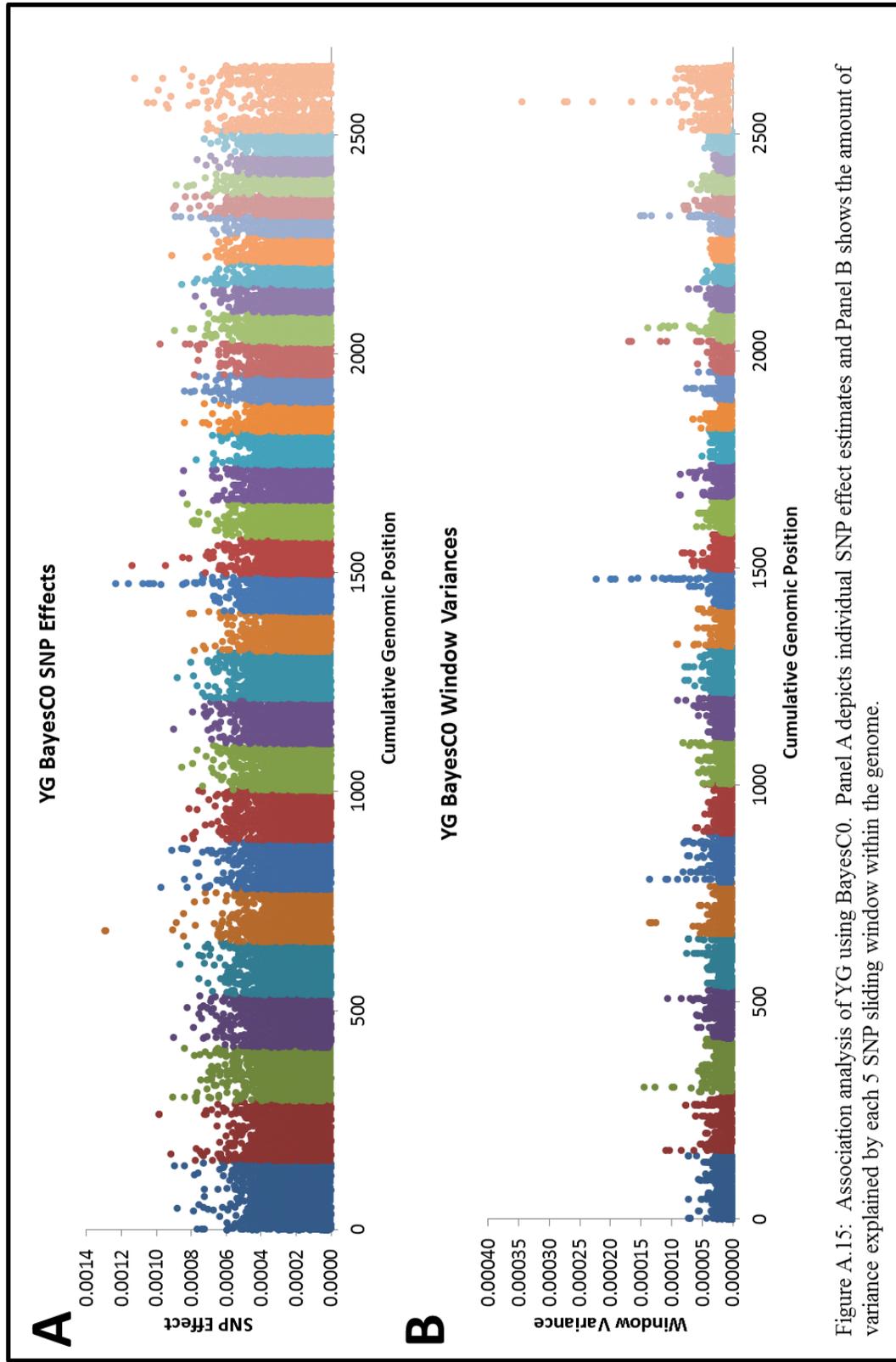
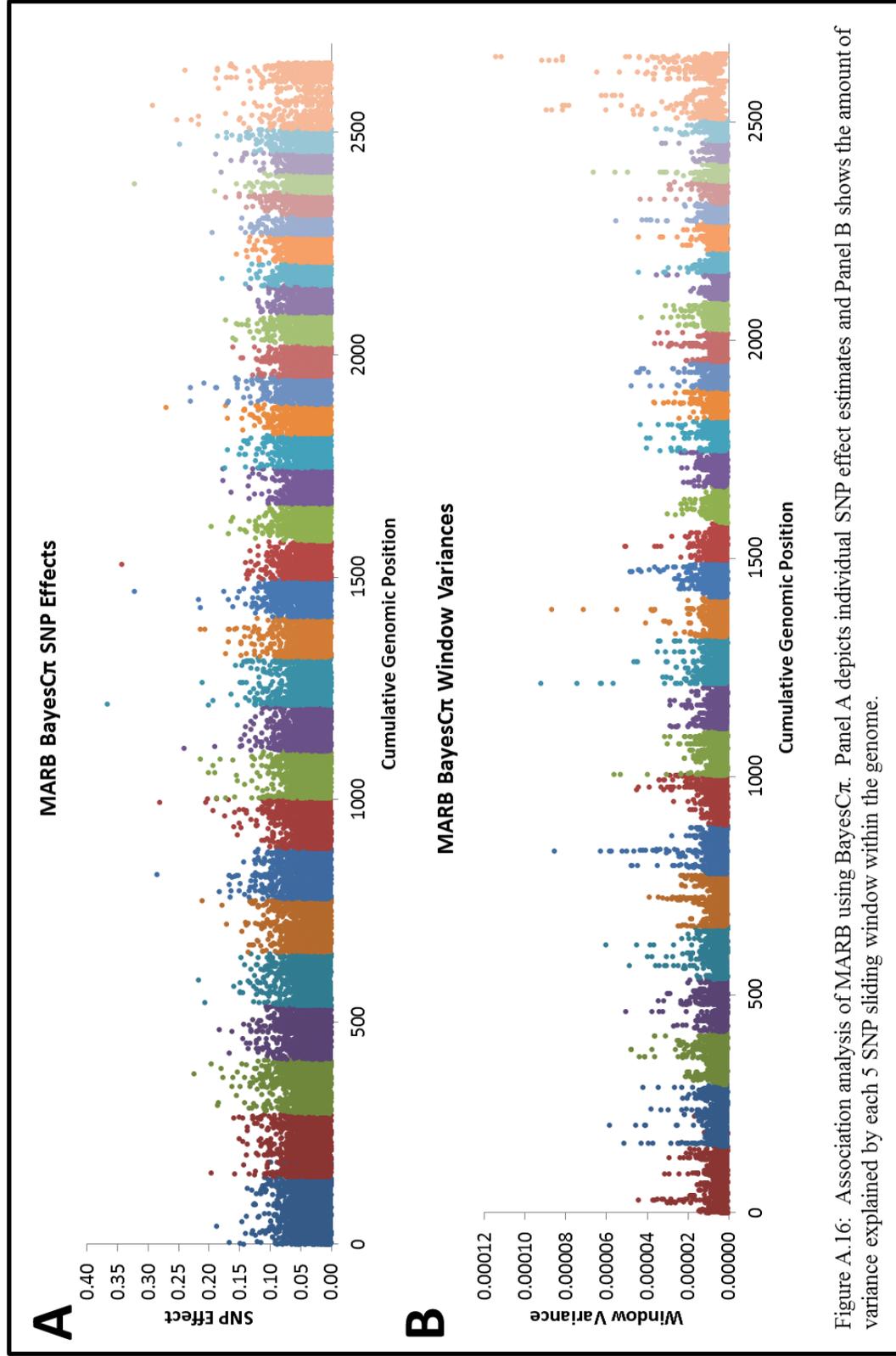
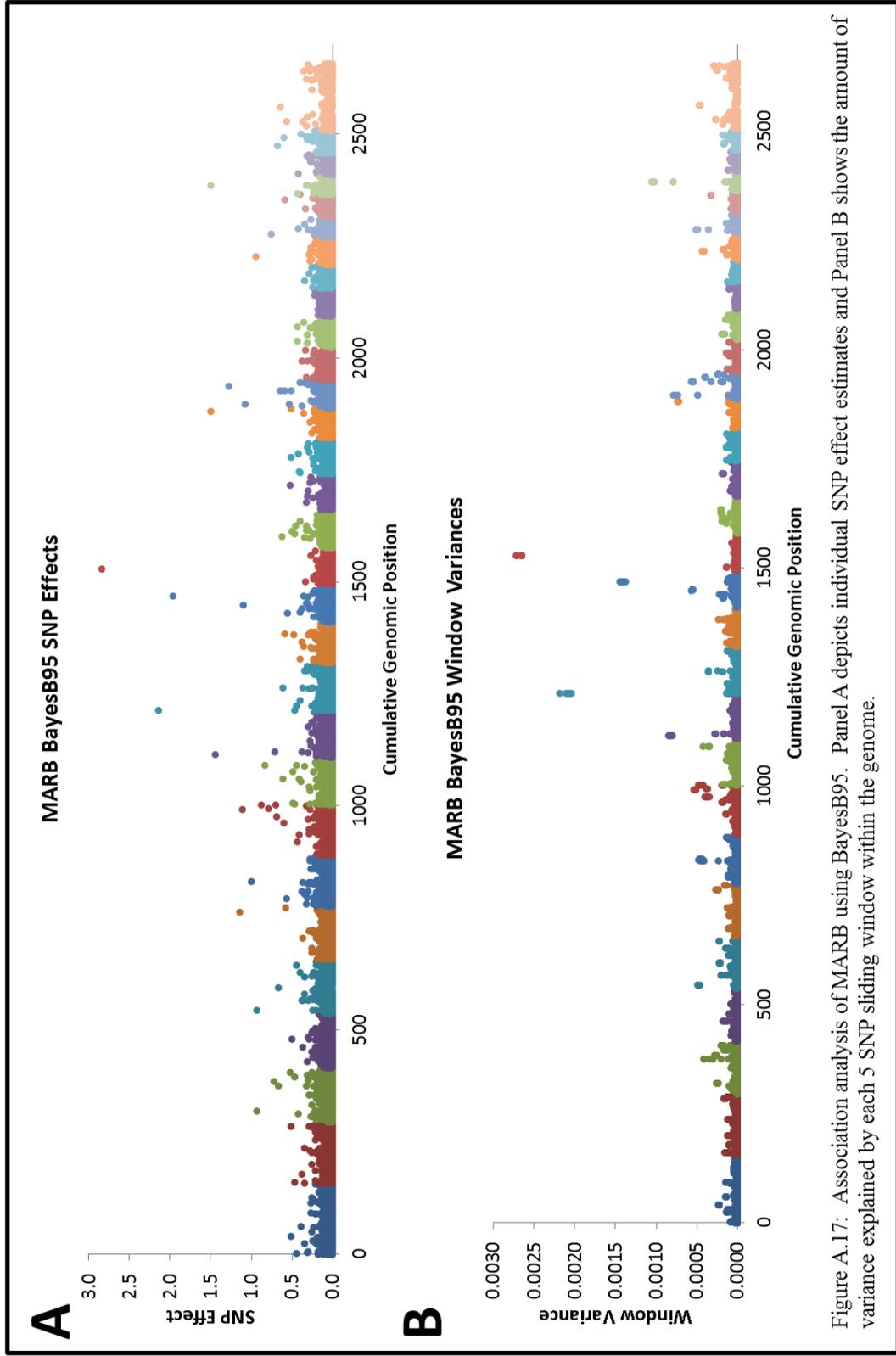


Figure A.15: Association analysis of YG using BayesC0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.





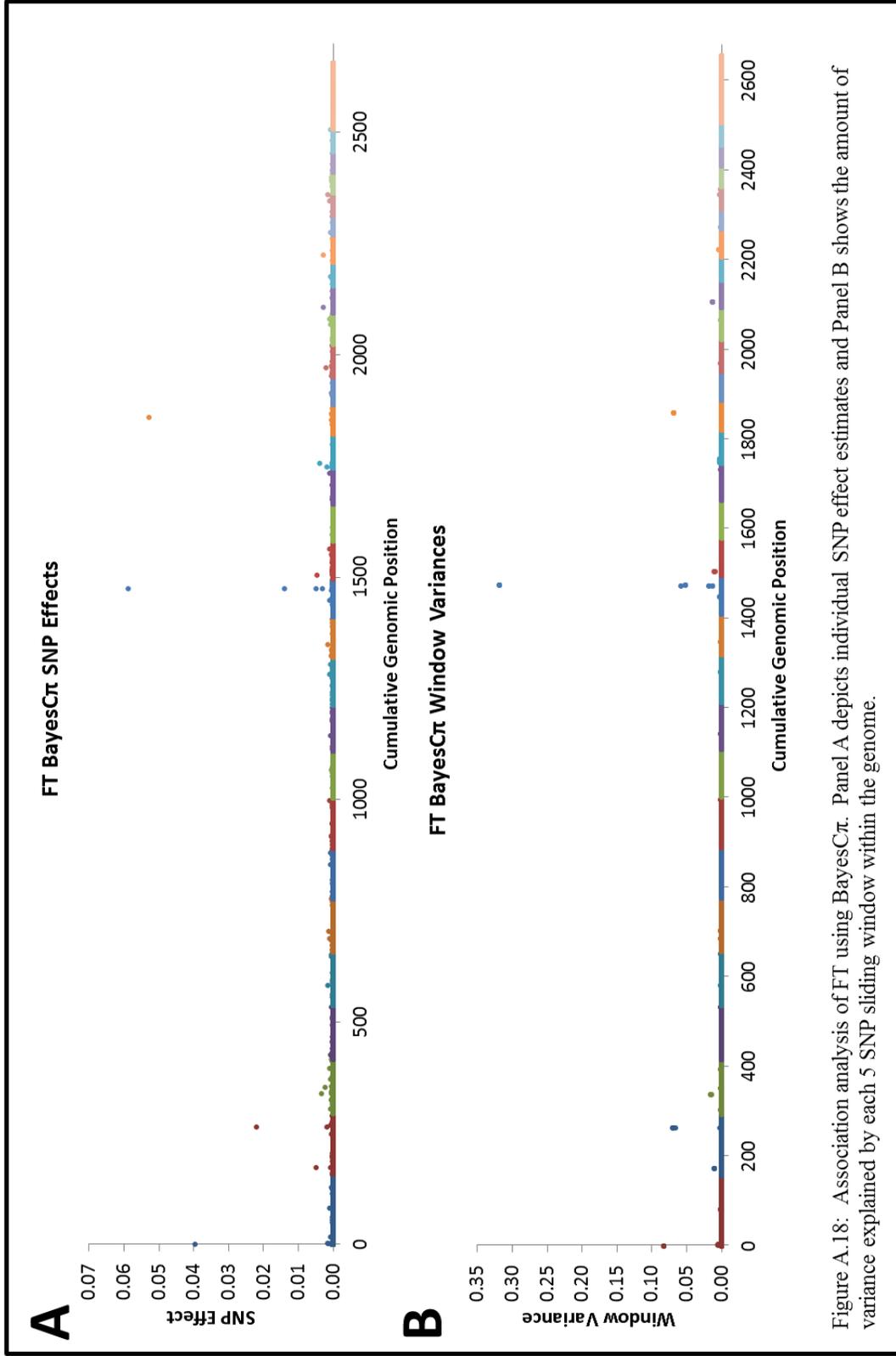


Figure A.18: Association analysis of FT using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

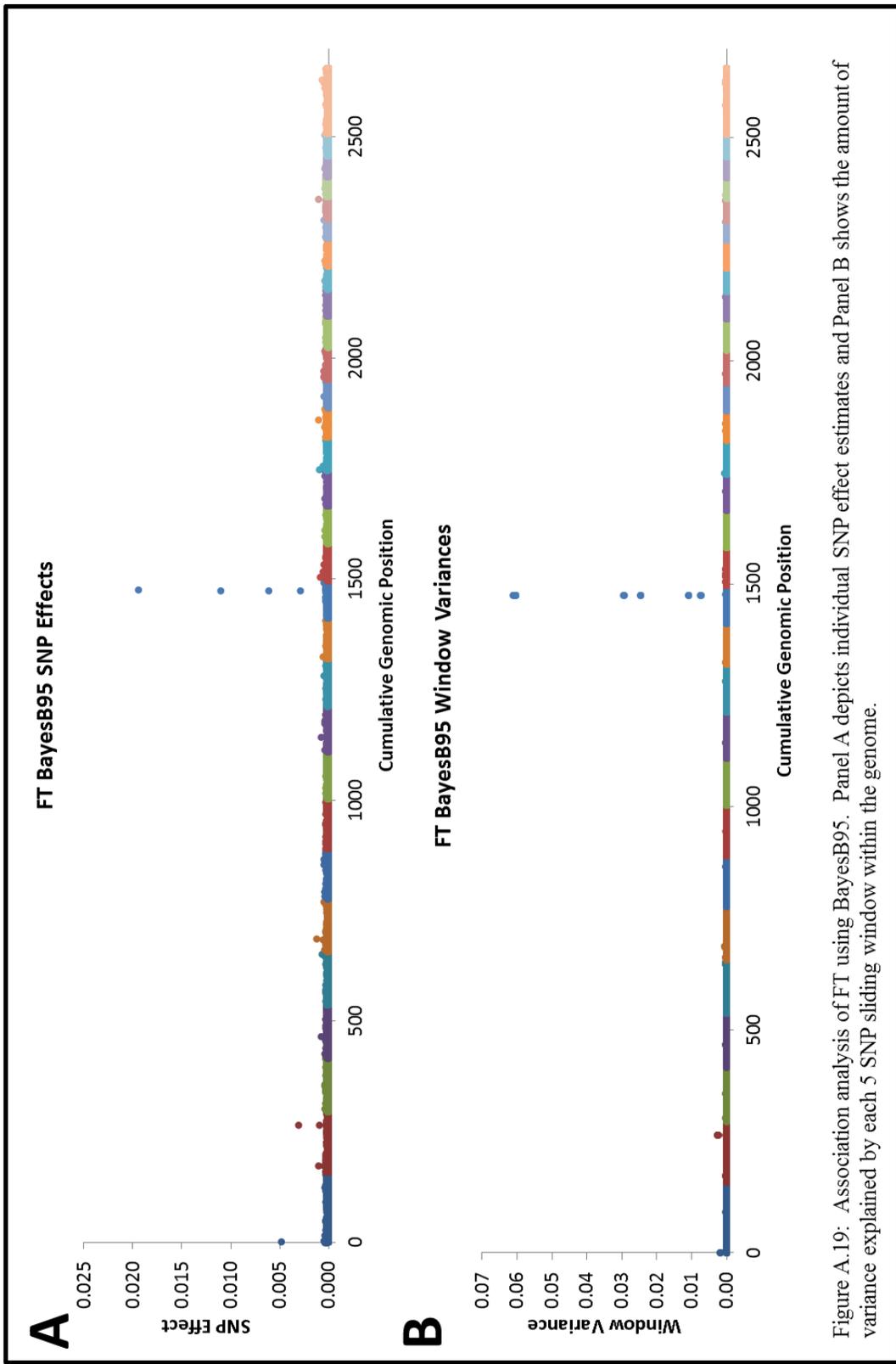
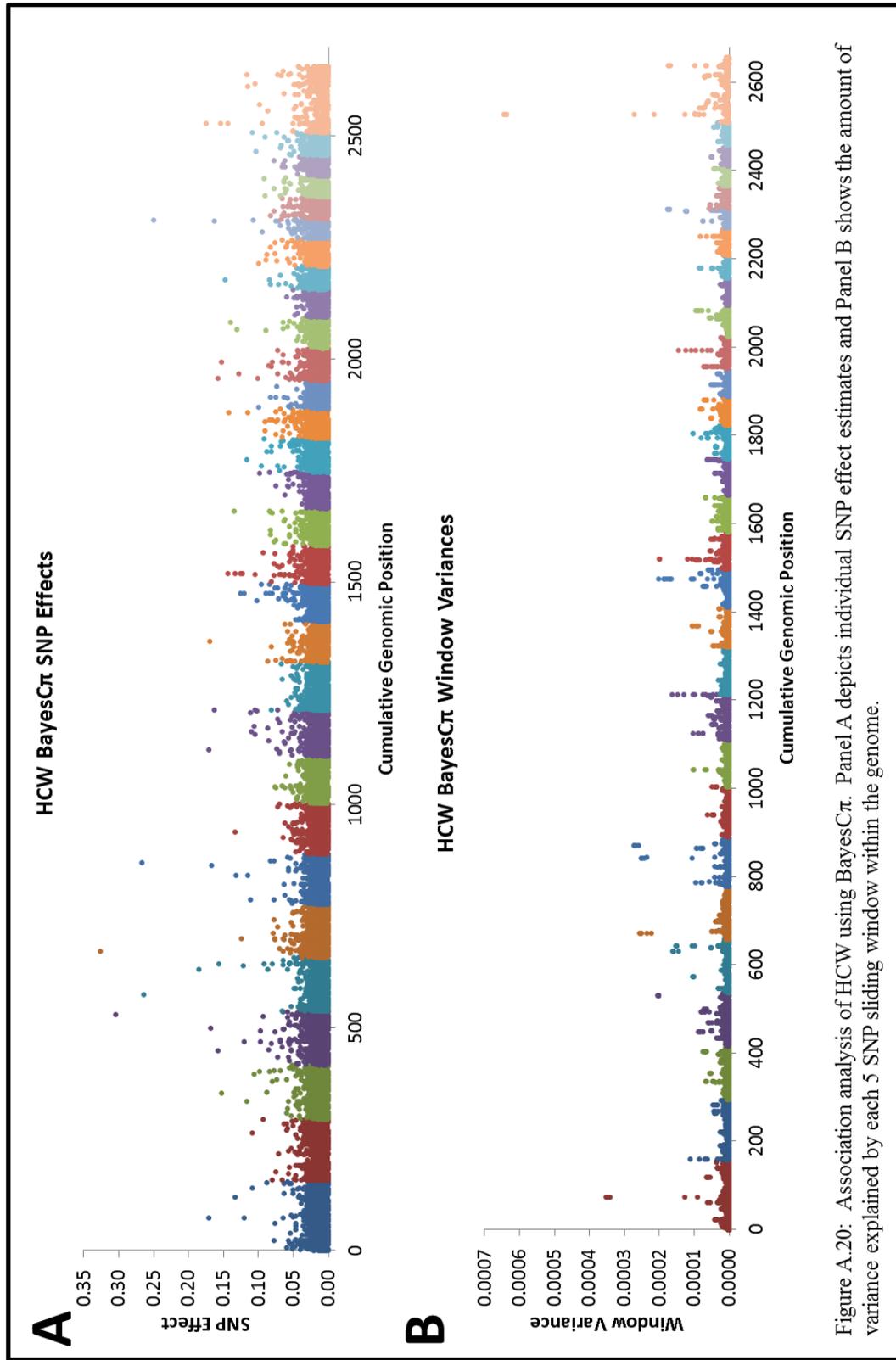
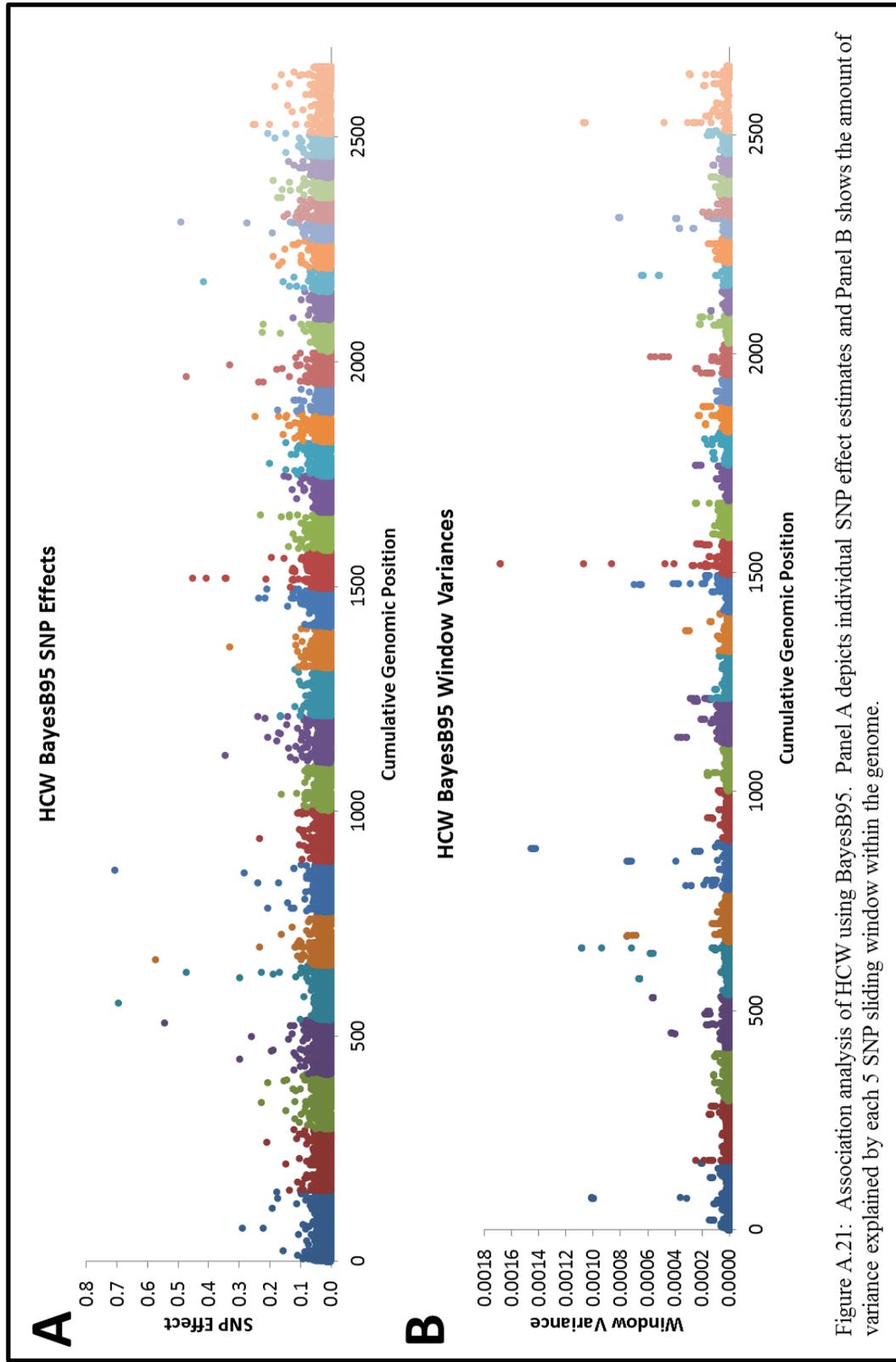


Figure A.19: Association analysis of FT using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.





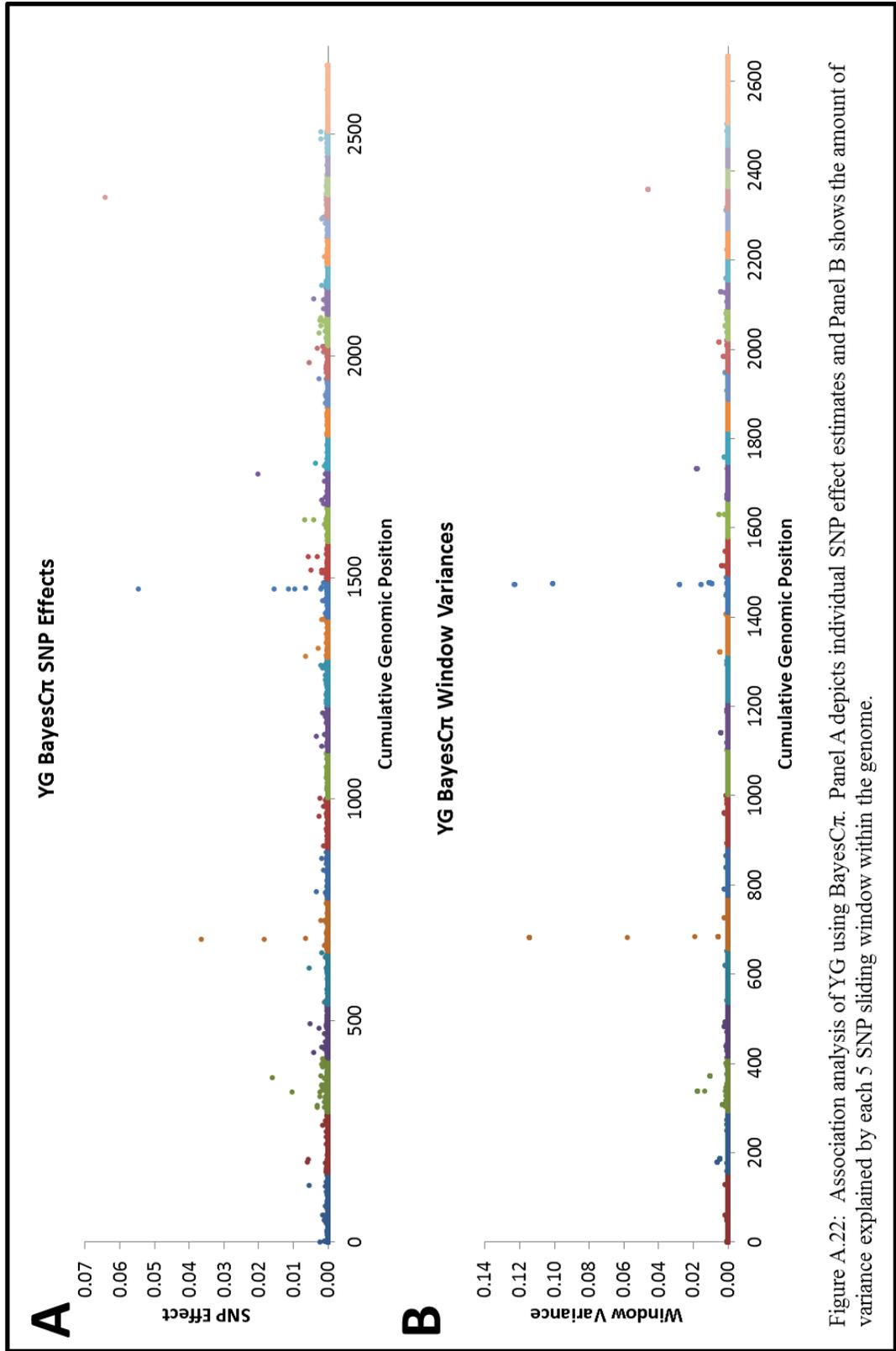


Figure A.22: Association analysis of YG using BayesC $\pi$ . Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

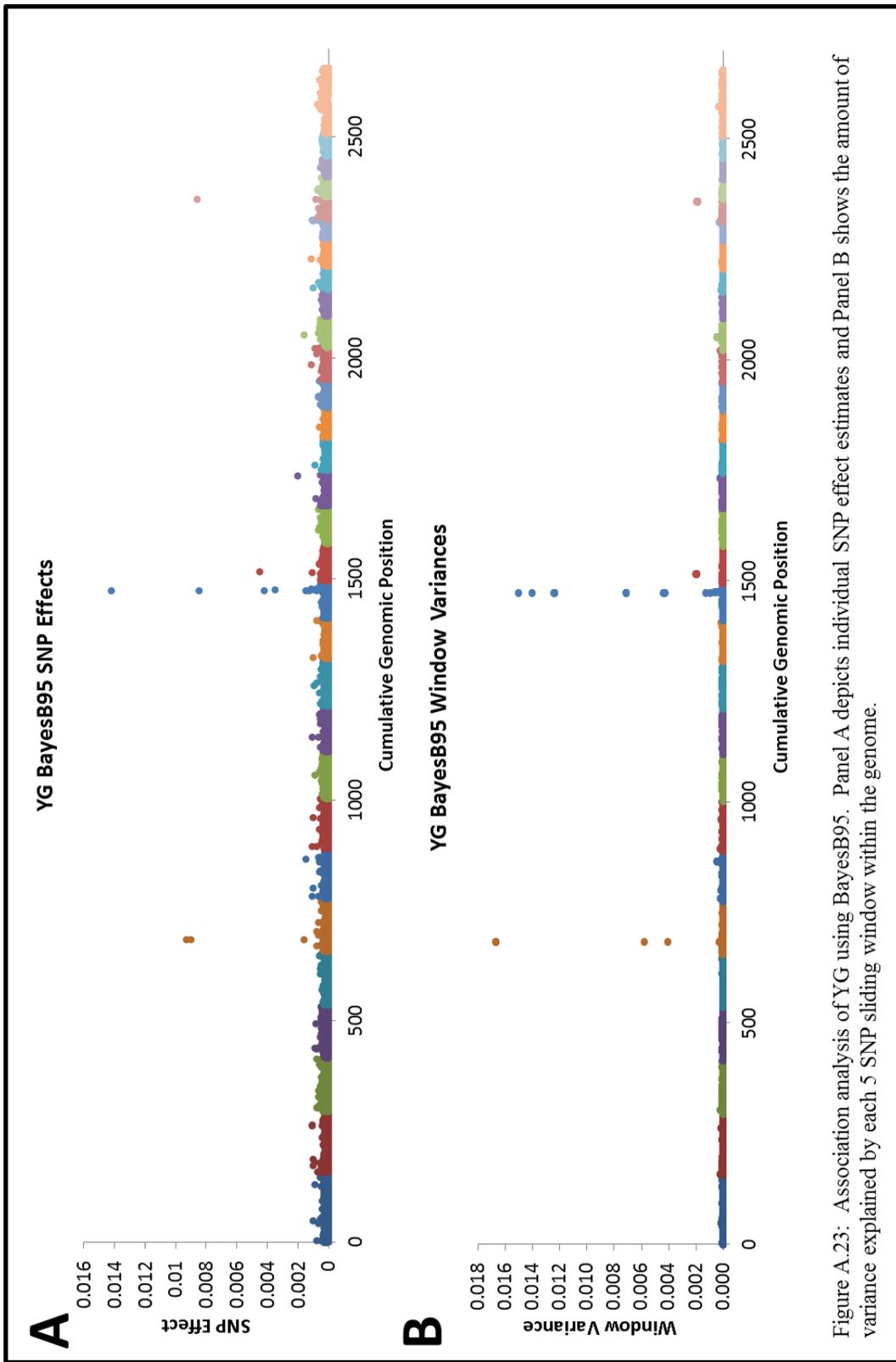


Figure A.23: Association analysis of YG using BayesB95. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

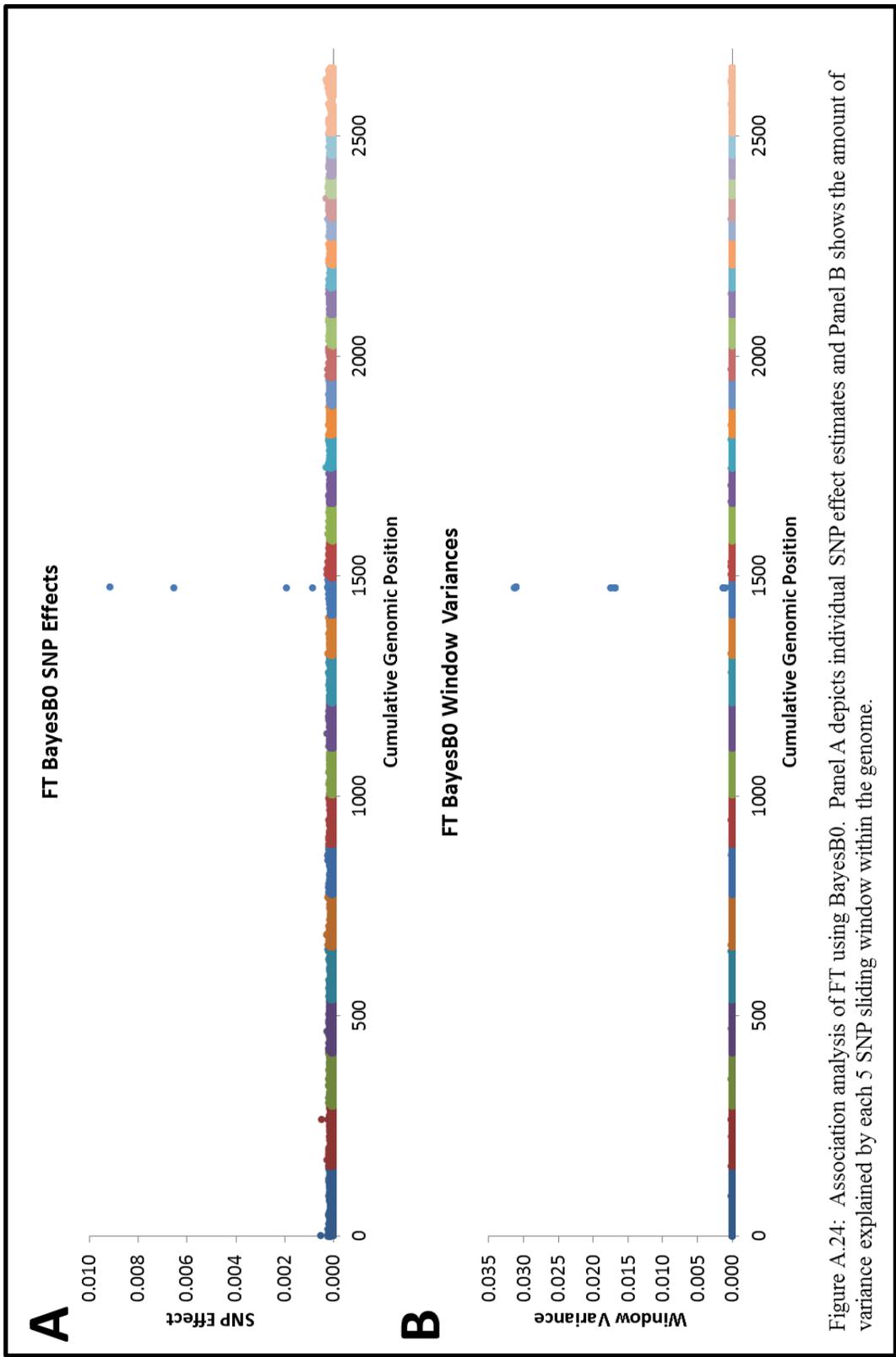


Figure A.24: Association analysis of FT using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

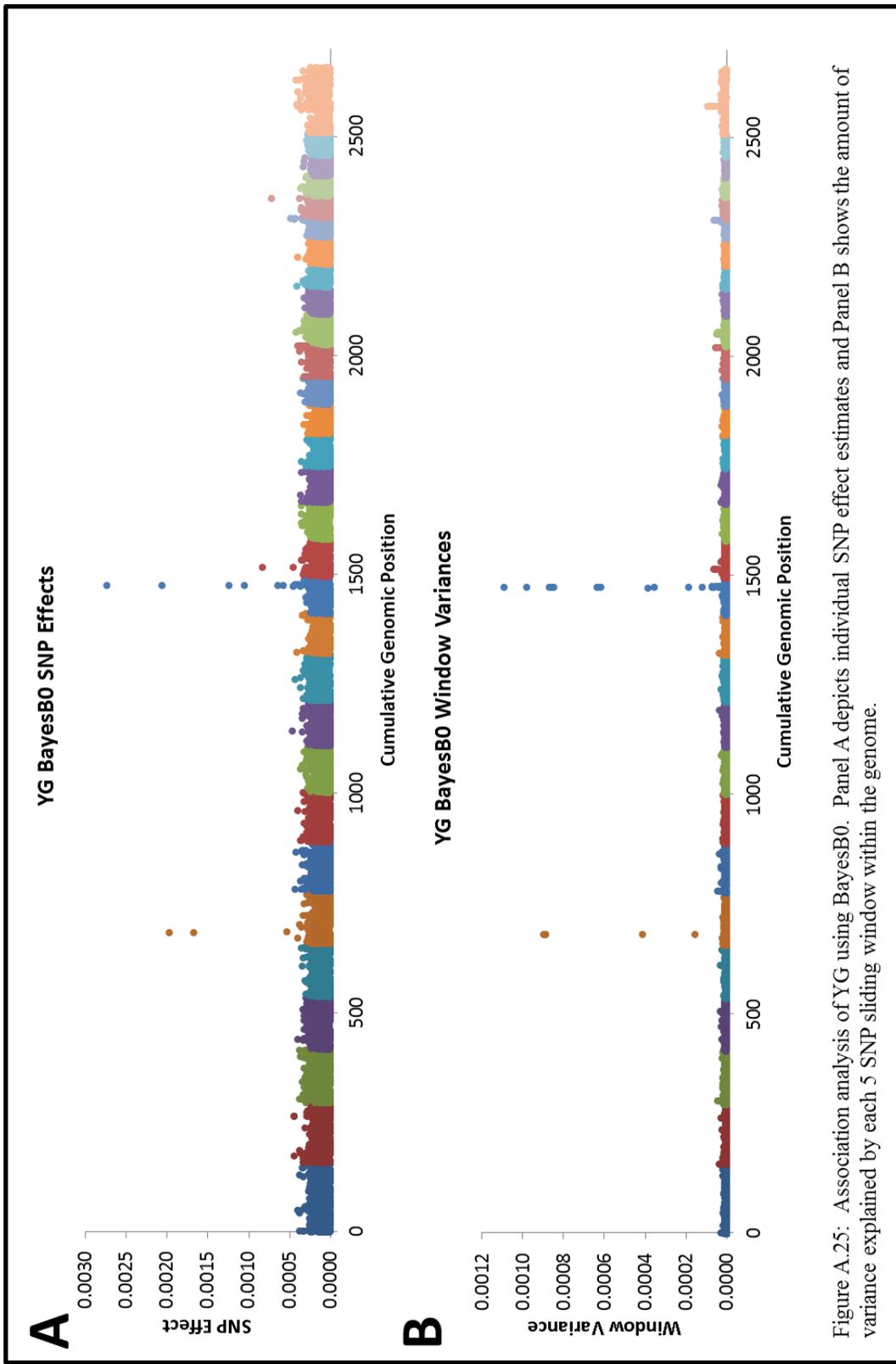
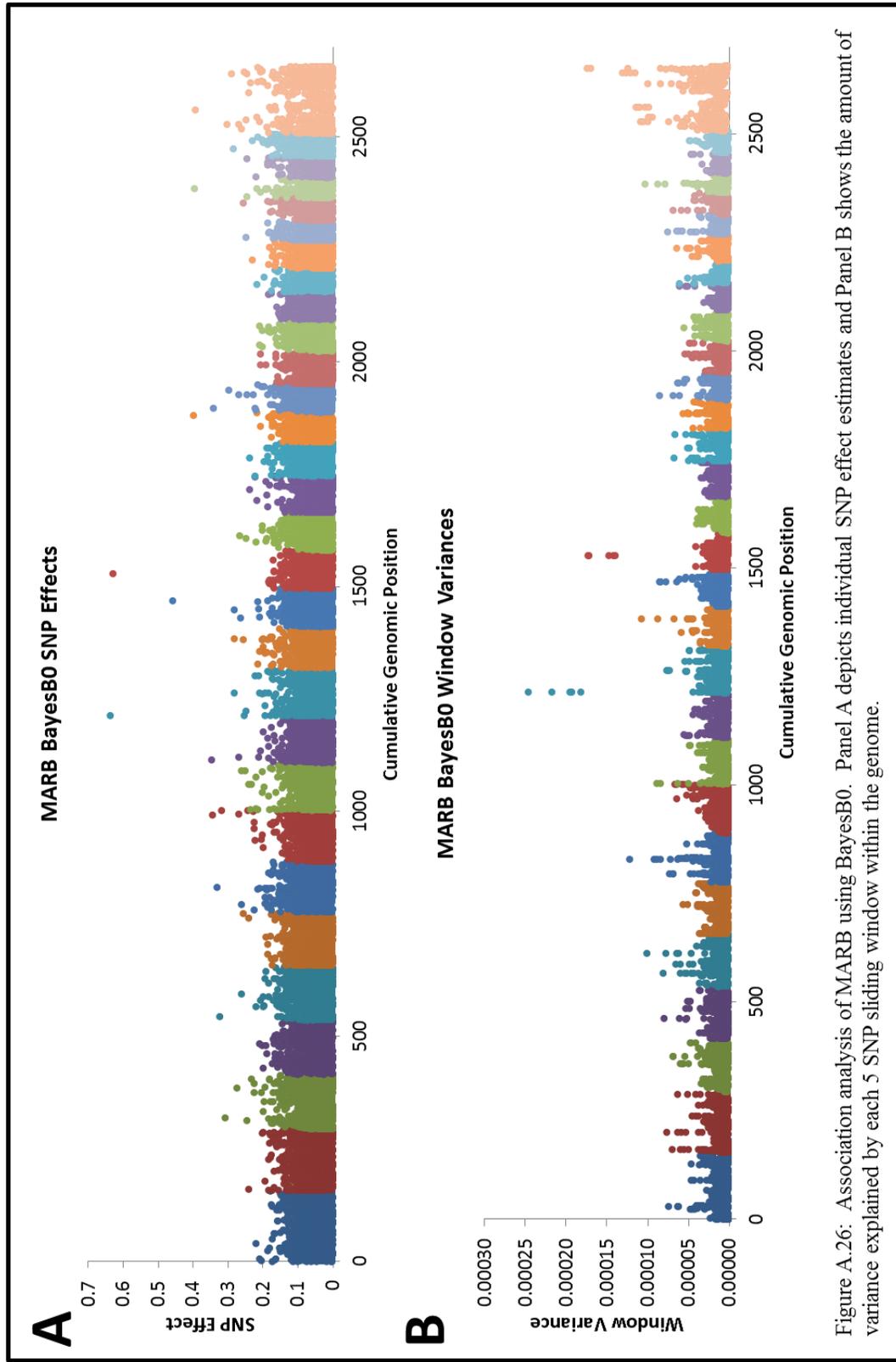


Figure A.25: Association analysis of YG using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.



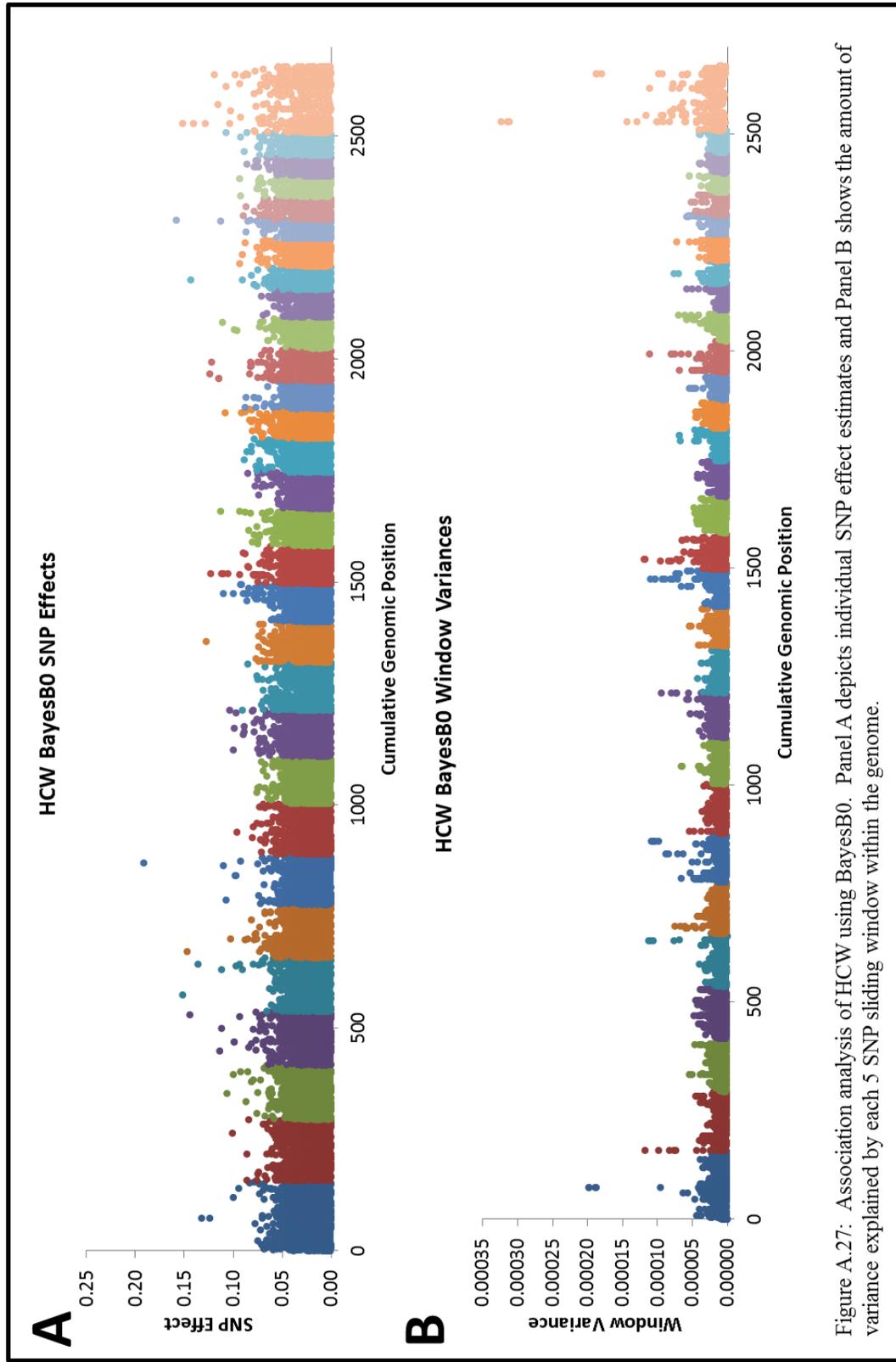


Figure A.27: Association analysis of HCW using BayesB0. Panel A depicts individual SNP effect estimates and Panel B shows the amount of variance explained by each 5 SNP sliding window within the genome.

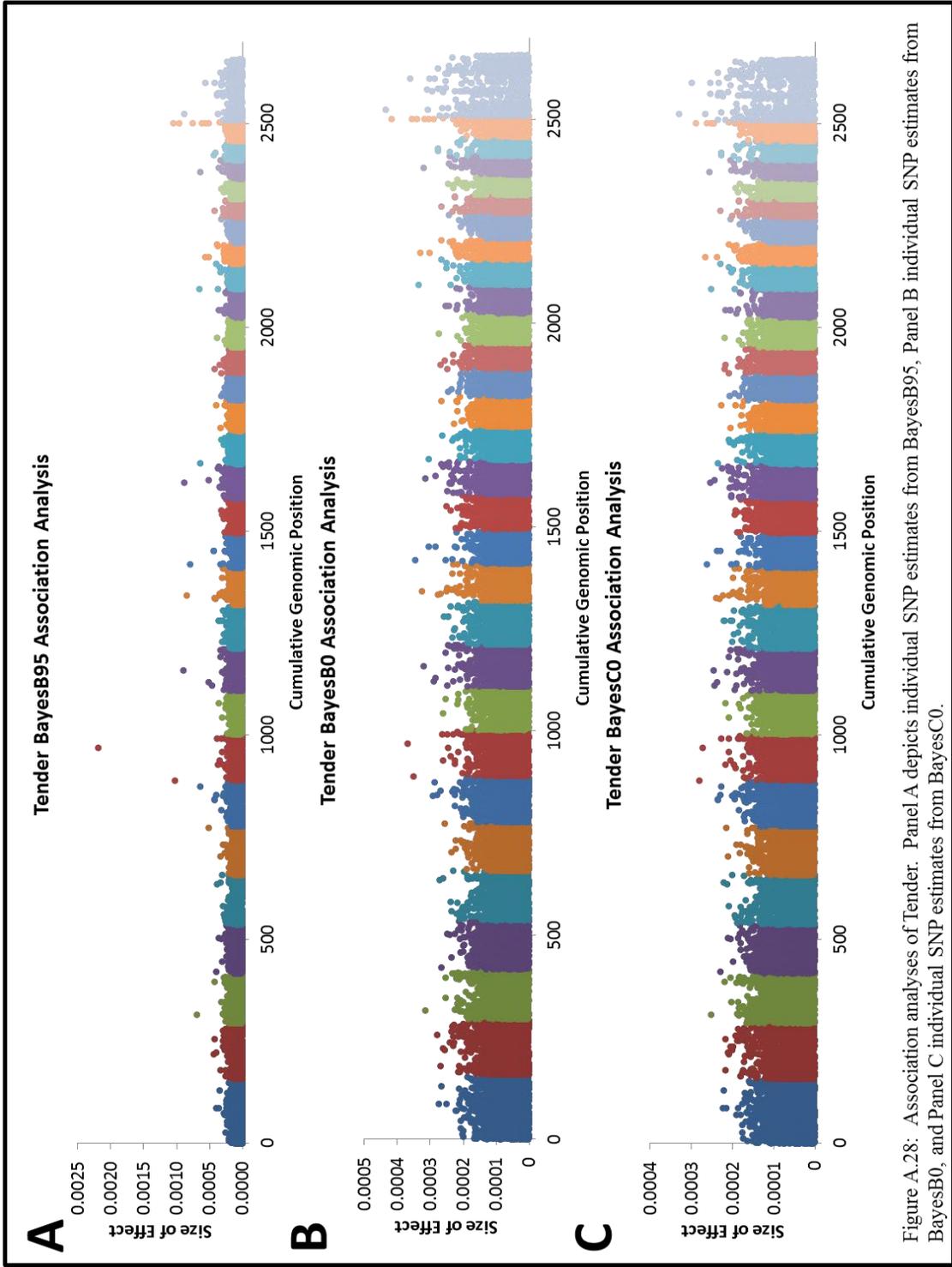
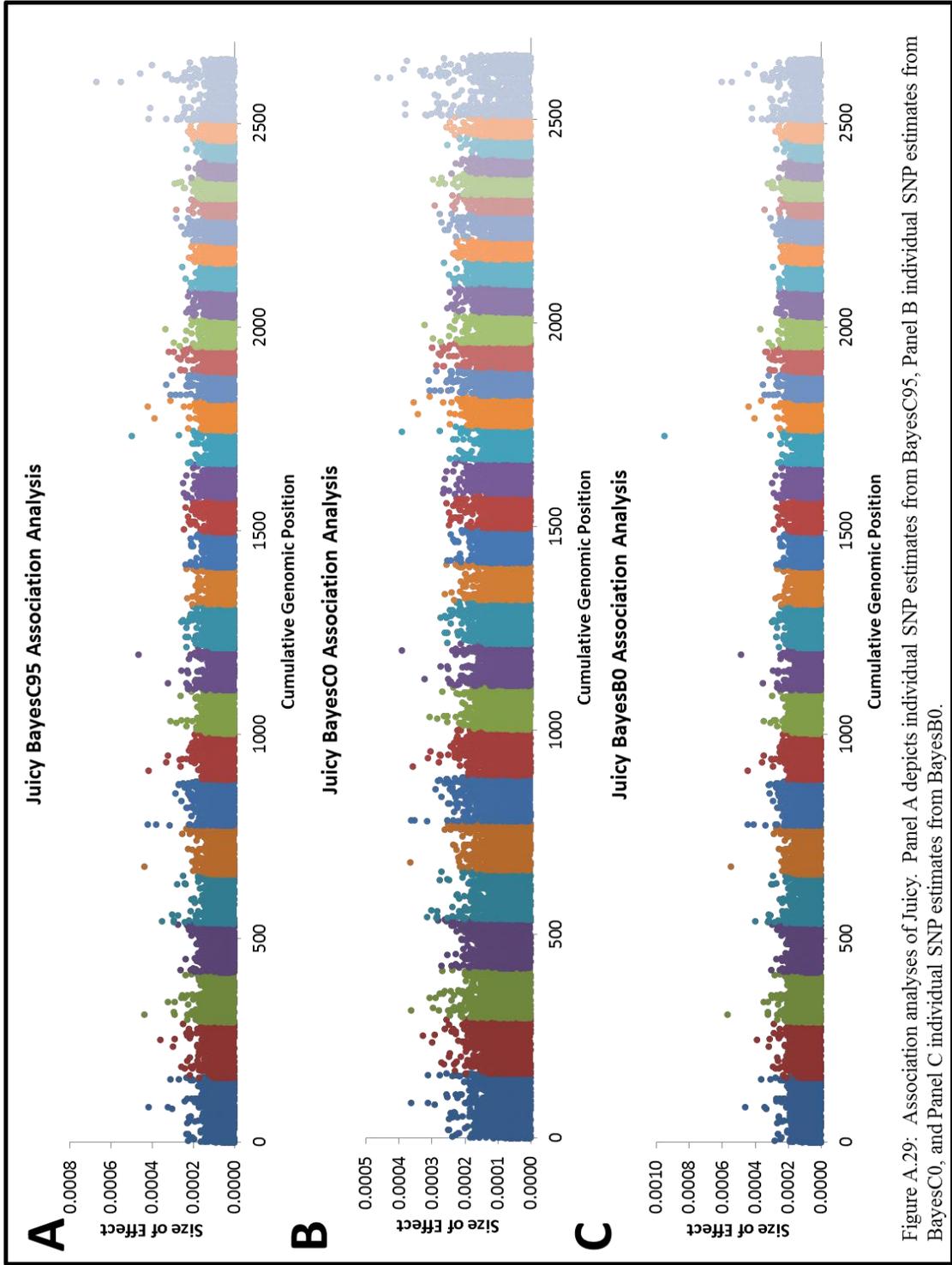


Figure A.28: Association analyses of Tender. Panel A depicts individual SNP estimates from BayesB95, Panel B individual SNP estimates from BayesB0, and Panel C individual SNP estimates from BayesC0.



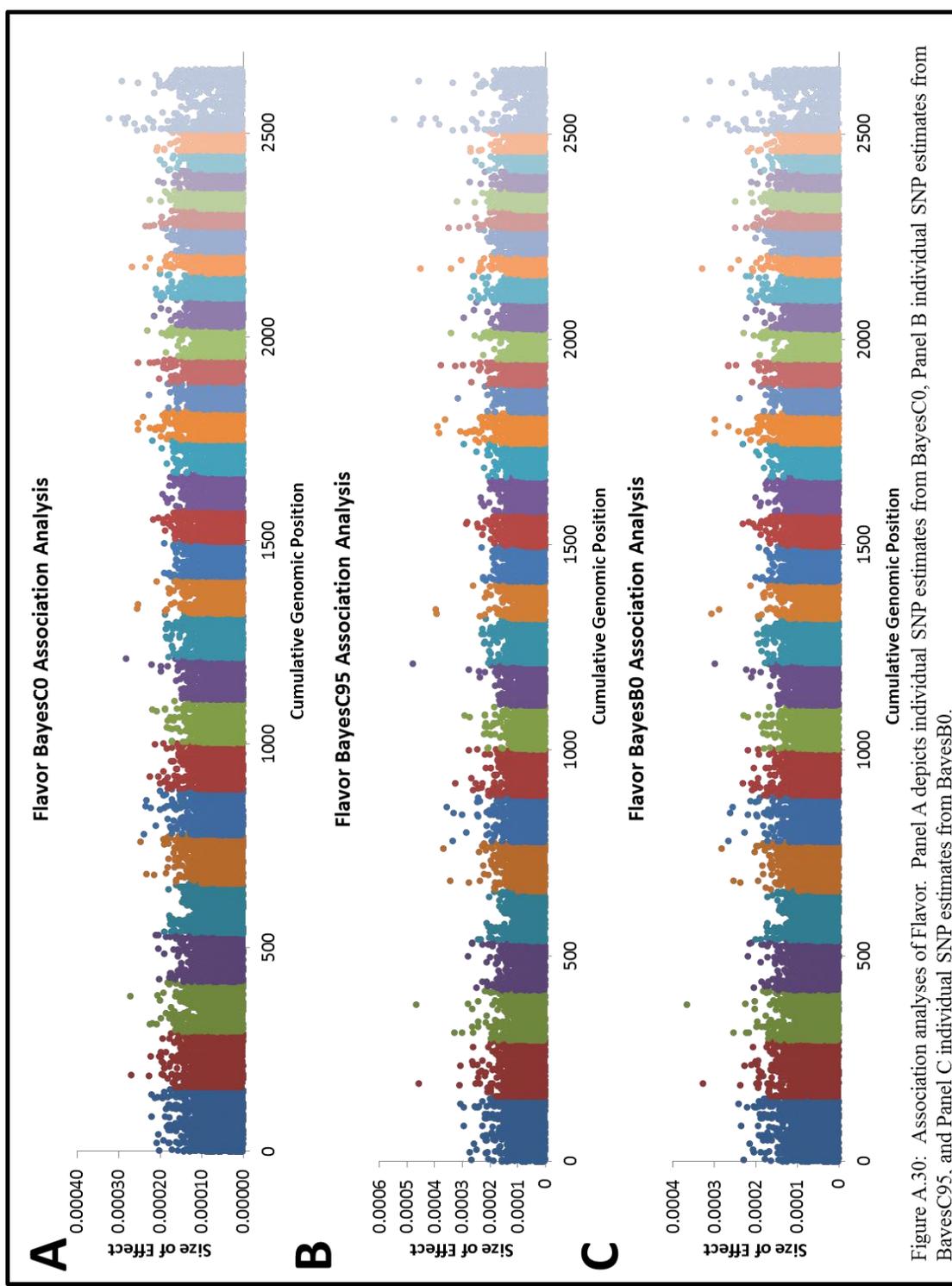
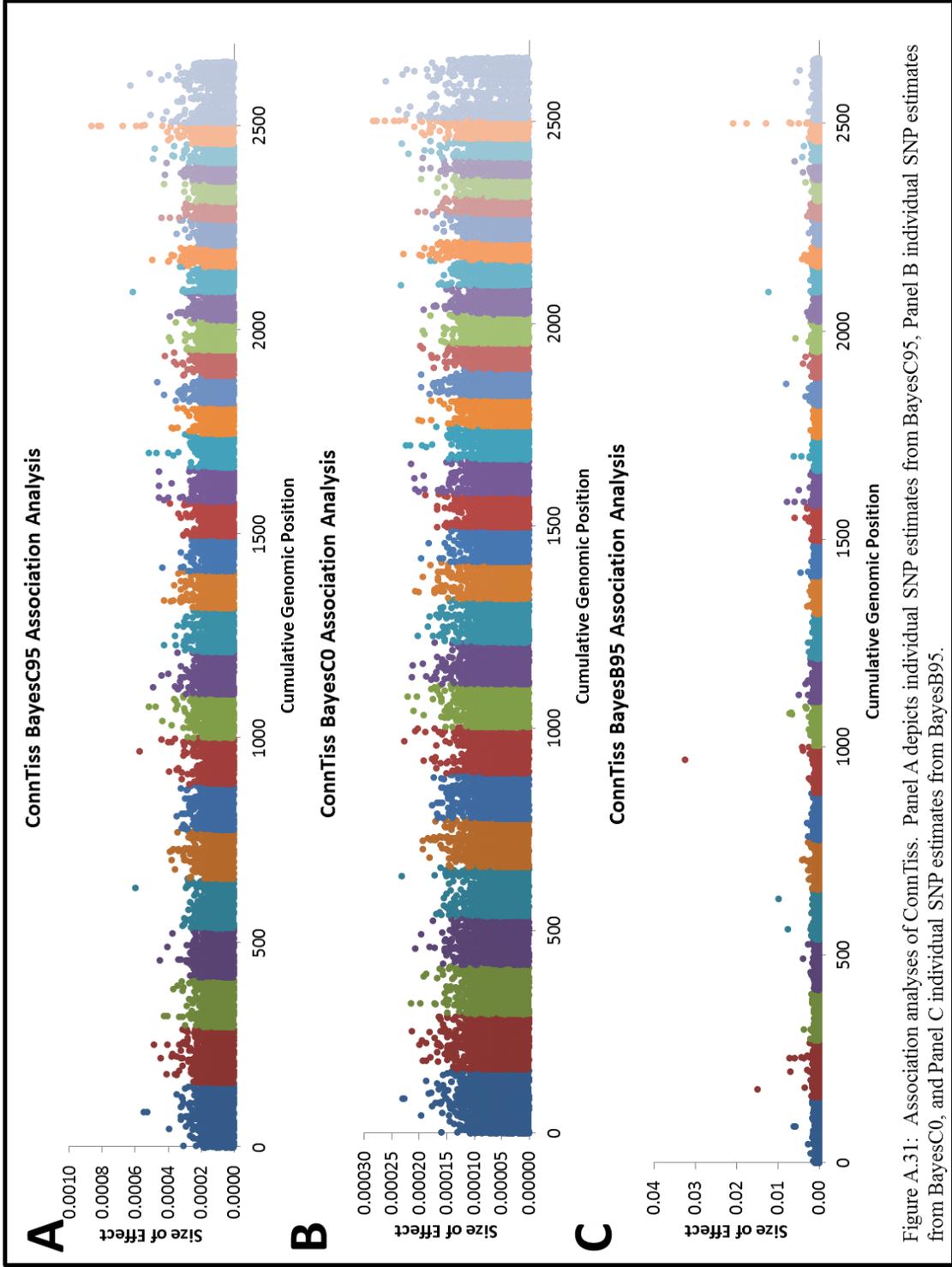


Figure A.30: Association analyses of Flavor. Panel A depicts individual SNP estimates from BayesC0, Panel B individual SNP estimates from BayesC95, and Panel C individual SNP estimates from BayesB0.



## REFERENCES

- American Angus Association. (2011). Weekly evaluation traits with genomic data. <http://www.angus.org/Nce/WeeklyEvalGenomicData.aspx>. Accessed 1/26/2012.
- The Bovine HapMap Consortium. (2009). Genome wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324:528-532.
- Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178:553-61.
- Casas, E., J.W. Keele, S.D. Shackelford, M. Koohmaraie, T.S. Sonstegard, T.P. Smith, S.M. Kappes and R.T. Stone. (1998). Association of the muscle hypertrophy locus with carcass traits in beef cattle. *Journal of Animal Science* 76:468-473.
- Chelh, I., B. Picard, J-F. Hocquette, and I. Cassar-Malek (2011). Myostatin inactivation induces a similar muscle molecular signature in double-musled cattle as in mice. *Animal* 5:278-286.
- Clark, S.A., J.M. Hickey and J.H.J. van der Werf. (2011). Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution*. 43:18.
- de Roos, A.P.W., C. Schrooten, E. Mullaart, M.P.L. Calus and R.F. Veerkamp. (2007). Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. *Journal of Dairy Science*. 90:4821-4829.
- de Roos, A.P.W., B.J. Hayes and M.E. Goddard. (2009). Reliability of genomic predictions across multiple populations. *Genetics*. 183:1545-1553.
- DeVuyst, E.A., J.T. Biermacher, J.L. Lusk, R.G. Mateescu, J.B. Blanton Jr., J.W. Swigert, B.J. Cook, and R.R. Reuter. (2011). Relationships between fed cattle traits and Igenity panel scores. *Journal of Animal Science* 89:1260-1269.
- Drogemuller, C., M. Peters, J. Pohlenz, O. Distl, and T. Leeb. (2002) A single point mutation within the ED1 gene disrupts correct splicing at two different splice sites and leads to anhidrotic ectodermal dysplasia in cattle. *Journal of Molecular Medicine* 80(5):319-323.
- Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

- Fernando, R.L. and D.J. Garrick. (2009). GenSel-user manual. [http://taurus.ansci.iastate.edu/Site/Welcome\\_files/GenSel%20Manual%20v2.pdf](http://taurus.ansci.iastate.edu/Site/Welcome_files/GenSel%20Manual%20v2.pdf).
- Fitzsimmons, C.J., S.M. Schmutz, R.D. Bergen, and J.J. McKinnon. (1998). A potential association between the BM1500 microsatellite and fat deposition in beef cattle. *Mammalian Genome* 9:432-434.
- Garrick, D.J. (2007). Equivalent mixed model equations for genomic selection. *Journal of Dairy Science* 90, 376 (Abstr.).
- Gianola, D., G. de los Campos, W.G. Hill, E. Manfredi, and R. Fernando. (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183:347-363.
- Goddard, M.E. and B.J. Hayes. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*. 124:323-330.
- Greiner, S.P., G.H. Rouse, D.E. Wilson, L.V. Cundiff, and T.L. Wheeler. (2003). The relationship between ultrasound measurements and carcass fat thickness and longissimus muscle area in beef cattle. *Journal of Animal Science* 81:676-682.
- Habier, D., R.L. Fernando and J.C.M. Dekkers. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. 177:2389-2397.
- Habier, K., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186.
- Harris, B.L., D.L. Johnson and R.J. Spelman. (2008). Genomic selection in New Zealand and the implications for national genetic evaluation. *Proceedings ICAR 36<sup>th</sup> Session*. 325-330.
- Hayes, B.J., P.J. Bowman, A.J. Chamberlain and M.E. Goddard. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* 92:433-443.
- Henderson, C.R. (1963). Selection index and expected genetic advance. In: W.D. Hanson and H.F. Robinson (editors). *Statistical Genetics and Plant Breeding*. NAS-NRC 982:141-163..
- Huang, D.W, B.T. Sherman, and R.A. Lempicki. (2009a). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols* 4:44-57.

- Huang, D.W, B.T. Sherman, and R.A. Lempicki. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* 37:1-13.
- Kachman, S.D. (2008). Incorporation of marker scores into national genetic evaluations. *Proceedings of the 9<sup>th</sup> Beef Improvement Federation Genetic Prediction Workshop: Prediction of Genetic Merit of Animals for Selection*. 92-98.
- Kizilkaya, K., R.L. Fernando and D.J. Garrick. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of Animal Science*. 88:544-551.
- Legarra, A., I. Aguilar and I. Misztal. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*. 92:4656-4663.
- Luan, T., J.A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T.H.E. Meuwissen. (2009). The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. *Genetics* 183:1119-1126.
- Matukumalli, L.K., C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, M.P. Heaton, J. O'Connell, S.S. Moore, T.P. Smith, T.S. Sonstegard, and C.P. Van Tassell. (2009). Development and characterization of a high-density SNP genotyping assay for cattle. *PLoS ONE* 4:e5350.
- McClure, M.C., N.S. Morsci, R.D. Schnabel, J.W. Kim, P. Yao, M.M. Rolf, S.D. McKay, S.J. Gregg, R.H. Chapple, S.L. Northcutt, and J.F. Taylor. (2010). A genome scan for quantitative trait loci influencing carcass, post-natal growth, and reproductive traits in commercial Angus cattle. *Animal Genetics* 41:597-607.
- McClure, M.C., H.R. Ramey, M.M. Rolf, S.D. McKay, J.E. Decker, R.H. Chapple, J.W. Kim, T.M. Taxis, R.L. Weaber, R.D. Schnabel and J.F. Taylor. (2012). Genome wide association analysis for quantitative trait loci influencing Warner Bratzler shear force in five taurine cattle breeds. *Animal Genetics* published 27 February 2012, doi: 10.1111/j.1365-2052.2012.02323.x..
- Meuwissen, T.H.E., B.J. Hayes and M.E. Goddard. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819-1829.
- Meyers, S.N., T.G. McDanel, S.L. Swist, B.M. Marron, D.J. Steffen, D. O'Toole, J.R. O'Connell, J.E. Beever, T.S. Sonstegard, and T.P.L. Smith. (2010) A deletion mutation in bovine SLC4A2 is associated with osteopetrosis in Red Angus cattle. *BMC Genomics* 11:337.
- Minick, J.A., M.E. Dikeman, E.J. Pollak, and D.E. Wilson. (2004). Heritability and correlation estimates of Warner-Bratzler shear force and carcass traits from

- Angus-, Charolais-, Hereford-, and Simmental-sired cattle. *Canadian Journal of Animal Science* 84:599-609.
- Nkrumah, J.D., C. Li, J. Yu, C. Hansen, D.H. Keisler, and S.S. Moore. (2005). Polymorphisms in the bovine leptin promoter associated with serum leptin concentration, growth, feed intake, feeding behavior, and measures of carcass merit. *Journal of Animal Science* 83:20-28.
- Northcutt, S.L. (2010). Implementation and deployment of genomically enhanced EPDs: Challenges and opportunities. *Proceedings of the Beef Improvement Federation 42nd Annual Research Symposium and Annual Meeting* 42:57-61.
- Page, B.T., E. Casas, M.P. Heaton, N.G. Cullen, D.L. Hyndman, C.A. Morris, A.M. Crawford, T.L. Wheeler, M. Koohmaraie, J. W. Keele, and T.P.L. Smith. (2002). Evaluation of single-nucleotide polymorphisms in CAPN1 for association with meat tenderness in cattle. *Journal of Animal Science* 80:3077-3085.
- Rolf, M.M., J.F. Taylor, R.D. Schnabel, S.D. McKay, M.C. McClure, S.L. Northcutt, M.S. Kerley and R.L. Weaber. (2010). Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. *BMC Genetics*. 11:24.
- Saatchi, M., M.C. McClure, S.D. McKay, M.M. Rolf, J.W. Kim, J.E. Decker, T.M. Taxis, R.H. Chapple, H.R. Ramey, S.L. Northcutt, S. Bauck, B. Woodward, J.C.M. Dekkers, R.L. Fernando, R.D. Schnabel, D.J. Garrick, and J.F. Taylor. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution* 43:40.
- Sambrook, J., E.F. Fritsch, and T. Maniatis. (1989). *Molecular Cloning: A laboratory manual*. Plainview, Cold Spring Harbor Laboratory Press.
- Schaeffer, L.R. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* 123:218-223.
- Scheet, P., and M. Stephens. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* 78:629-644.
- Schenkel, F.S., S.P. Miller, X. Ye, S.S. Moore, J.D. Nkrumah, C. Li, I.B. Mandell, J.W. Wilton, and J.L. Williams. (2005). Association of single nucleotide polymorphisms in the leptin gene with carcass and meat quality traits of beef cattle. *Journal of Animal Science* 83:2009-2020.
- Schenkel, F.S. S.P. Miller, Z. Jiang, I.B. Mandell, X. Ye, H. Li, and J.W. Wilton. (2006). Association of a single nucleotide polymorphism in the calpastatin gene with

- carcass and meat quality traits of beef cattle. *Journal of Animal Science* 84:291-299.
- Seidel, G.E., Jr. (2009). Brief introduction to whole-genome selection in cattle using single nucleotide polymorphisms. *Reproduction, Fertility and Development* 22:138-144.
- Sherman, E.L., J.D. Nkrumah, B.M. Murdoch, C. Li, Z. Wang, A. Fu, and S.S. Moore. (2008). Polymorphisms and haplotypes in the bovine neuropeptide Y, growth hormone receptor, ghrelin, insulin-like growth factor 2 and uncoupling proteins 2 and 3 genes and their associations with measures of growth, performance, feed efficiency, and carcass merit in beef cattle. *Journal of Animal Science* 86:1-16.
- Solberg, T.R., A.K. Sonesson, J.A. Woolliams, J. Odegard and T.H.E. Meuwissen. (2009). Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genetics Selection Evolution* 41:53.
- Sosnicki, A.A. and S. Newman. (2010). The support of meat value chains by genetic technologies. *Meat Science* 86:129-137.
- Su, G., B. Guldbbrandtsen, V.R. Gregersen, and M.S. Lund. (2010). Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *Journal of Dairy Science* 93:1175-1183.
- Takeda, H., M. Takami, T. Oguni, T. Tsuji, K. Yoneda, H. Sato, N. Ihara, T. Itoh, S.R. Kata, Y. Mishina, J.E. Womack, Y. Moritomo, Y. Sugimoto, and T. Kunieda. (2002) Positional cloning of the gene LIMBIN responsible for bovine chondrodysplastic dwarfism. *Proceedings of the National Academies of Science* 99(16):10549-10554.
- Thallman, R.M., D.W. Moser, E.W. Dressler, R.L. Totir, R.L. Fernando, S.D. Kachman, J.M. Rumph, M.E. Dikeman, and J.E. Pollak. (2003). Carcass merit project: DNA marker validation. Proceedings of the Beef Improvement Federation 8<sup>th</sup> Genetic Prediction Workshop 8:70-90.
- Toosi, A., R.L. Fernando and J.C.M. Dekkers. (2010). Genomic selection in admixed and crossbred populations. *Journal of Animal Science* 88:32-46.
- VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*. 91:4414-4423.
- VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor and F.S. Schenkel. (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* 92:16-24.

Villanueva, B., R. Pong-Wong, J. Fernandez and M.A. Toro. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *Journal of Animal Science*. 83:1747-1752.

Willham, R.L. (1993). Ideas into action: a celebration of the first 25 years of the Beef Improvement Federation. University Printing Services, Oklahoma State University, Stillwater, OK.

Zimin, A.V., A.L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C.P. Van Tassell, T.S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J.A. Yorke, and S.L. Salzberg. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology* 10:R42.

## VITA

Megan was born on February 9, 1984 in Iola, Kansas. She grew up on a commercial cow/calf operation near LeRoy, KS. After using Limousin bulls and grading up cows for several years, the remainder of the commercial cows were sold in 2003 and the cow/calf operation was completely converted to a Limousin seedstock operation, Rolf Limousin, which is still in operation today. Megan graduated from Burlington High School in Burlington, Kansas in 2002 and began attending school at Kansas State University in fall of 2002. She graduated from Kansas State University in December of 2005 with a Bachelors of Science in Animal Science with a science/pre-vet option. Megan joined the University of Missouri Division of Animal Science M.S. program in January of 2006. She was awarded her M.S. degree in Animal Sciences in May of 2009 and joined the Genetics Area Program Ph.D. program shortly thereafter. After completion of her Ph.D. in June 2012, Megan moved to Stillwater, OK to begin work as an Assistant Professor of Beef Cattle Management at Oklahoma State University, where she serves as a State Beef Extension Specialist focusing on the use of genetic and genomic technologies in the beef industry.