MULTIVARIATE GENOME-WIDE ASSOCIATION STUDIES AND GENOMIC PREDICTIONS IN MULTIPLE BREEDS AND CROSSBRED ANIMALS

A Thesis Presented to the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment of the Requirements

for the Degree

Master of Science

---

by

MIRANDA WILSON

Dr. Jared E. Decker, Thesis Advisor

May 2016

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

MULTIVARIATE GENOME-WIDE ASSOCIATION STUDIES AND GENOMIC PREDICTIONS IN MULTIPLE BREEDS AND CROSSBRED ANIMALS

Presented by Miranda Wilson, a candidate for the degree of Master in Science, and hereby certify that in their opinion it is worthy of acceptance.

_____

Dr. Jared E. Decker, Animal Science, UMC

_____

Dr. Jeremy F. Taylor, Animal Science, UMC

_____

Dr. Robert D. Schnabel, Animal Science, UMC

_____

Dr. Ian R. Gizer, Psychological Sciences, UMC

DEDICATION

This thesis is dedicated to my family, for their unwavering support throughout my life. Thank you for all your encouragement as I underwent this wonderful journey.

ACKNOWLEDGEMENTS

I would like to acknowledge those people that have made a tremendous impact on not only making my master's degree an exceptional learning experience but also on my professional development as a researcher. I would like to thank my advisor, Dr. Jared Decker, for all of the guidance and encouragement during my Master's program. Dr. Decker has helped me to improve my ability to present research by providing me with chances to speak at extension meetings. With his direction, I went from having absolutely no prior experience with programming to being able to create custom scripts to identify QTL regions for the selection of haplotypes. I am extremely grateful for the guidance and opportunities that Dr. Decker has provided me with during my time here.

I appreciate all of the guidance and help provided by Dr. Jerry Taylor and Dr. Robert Schnabel with my research projects. I also appreciate all of the advice and help from Dr. Taylor on deciding what the next step should be after the completion of my Master's. I would like to thank them along with Dr. Ian Gizer for being on my thesis committee and taking the time out of their busy schedules to meet and determine that my class work and research were on the right path. I would also like to thank Dr. Lamberson and Dr. Smith, along with Dr. Decker and Dr. Taylor, for writing me letters of recommendation for my application for the Veterinary Pathobiology Ph.D program at Texas A&M. I have to thank Dr. Lamberson for introducing me to Dr. Taylor when I was an undergraduate and interested in genetics and research, which lead to a semester of undergraduate

research for credit in Dr. Taylor's lab. I am thankful to Dr. Taylor for hiring me after that semester as an undergraduate researcher, where I was later introduced to Dr. Decker.

I would like to thank Lynsey Whitacre for being such a wonderful lab mate. I appreciate all of the help, guidance, and friendship that you have given me over the years. I would also like to thank the Taylor lab for their encouragement and for making lab meeting a great learning environment. I would like to thank Helen in Dr. Taylor's lab for teaching my valuable skills in the lab, such as, extracting DNA (from hair, blood, tissue, semen, and saliva) and PCR. I am thankful to both the Animal Science professors and graduate students for providing such an excellent collaborative environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF APPENDIX FIGURES

Figure **Page**

CHAPTER 1

REVIEW OF LITERATURE

Introduction to Genome-Wide Association Studies

In recent years, the price of genome-wide genotyping has become more affordable, leading to researchers having a plethora of genotypes, from multiple species, with tens of thousands of Single Nucleotide Polymorphism (SNP) markers per individual. With this increase in availability of markers, there has been a dramatic increase in genome-wide association studies (GWAS). These studies seek to identify common DNA variants that are associated with variation in the trait being studied (Yang et al. 2010). These studies are a standard approach used for identifying genomic regions associated with economically important production traits in agricultural species, variation among wild populations, and candidate regions associated with complex genetic diseases (Gondro et. al. 2013). Genome-wide association studies are incorporated across the fields from medical to evolutionary genetics.

Introduction to Linear Mixed Models

Although GWAS have the potential to identify candidate regions or polymorphisms underlying genetic diseases and traits, false discoveries are a major concern. False discoveries can be caused by multiple testing, but are

exacerbated by population structure and unequal relatedness among individuals in a given population (Zhang et al. 2010).

Population structure can arise naturally in all organisms, caused by geography, natural selection or artificial selection. Species that have both larger sample sizes and broader allelic diversity often include familial relationships. Even though family-based samples have been taken advantage of to avoid the effect of population structure, they are usually limited by sample size and allelic diversity.

General linear model (GLM) methods incorporating techniques such as genomic control, structured association, and principal-components analysis were initially used to address population stratification. Genomic control uniformly scales the test statistics; such that the observed median test statistic equals the expected statistic (Segura et al. 2012). This approach applied to statistics generated by GLM methods reduces the global inflation of test statistics. Unfortunately, all polymorphisms are subjected to the same correction because the approach does not change the ranking of statistics. When using structured association and principal component analysis, population structure is accounted for by including covariates in the model that represent both the cluster memberships and the principal component loadings of the individuals. These approaches perform well when the population structure is simple but have been shown to perform poorly when the population structure is complex (Segura et al. 2012). Accounting only for population structure and not for relatedness among

individuals, could lead to either loss of control for false positives or a decrease in power. (Zhang et al. 2010)

Linear mixed models (LMM) were introduced to simultaneously account for unequal relatedness among individuals and population structure (Zhang et al. 2010; Kang et al. 2010). In the LMM based methods, population structure is explained by fitting a random effect representing the additive genetic merit of each individual. Kinship among individuals is incorporated as the variance-covariance structure of the random effect for individuals (Zhang et al. 2010). This approach performs well in studies using plant, animal, and human models (Segura et al. 2012). Recently, there has been considerable attention on LMMs as a powerful and effective tool for accounting for population stratification and relatedness in genomic association studies (Zhou & Stephens 2014). One limitation that researchers have found when using LMM methods is that they can be computationally inefficient, especially when handling large samples. However, several algorithms have recently been introduced to address this issue.

The majority of published association studies that utilize LMMs have mostly been univariate applications. These univariate analyses, consider each phenotype independently even when multiple phenotypes are available for each individual (Stephens 2013). “Univariate linear mixed models (LMM) are fit for marker association tests with a single phenotype to account for population stratification and sample structure, and for estimating the proportion of variance in phenotypes explained (PVE) by typed genotypes (chip heritability).” (Zhou &

Stephens 2014). As explained by Zhou and Stephens, the statistical equation for Univariate Linear Mixed Models is:

$$\boldsymbol{y} = \boldsymbol{W\alpha} + \boldsymbol{x\beta} + \boldsymbol{u} + \boldsymbol{\epsilon}; \; \boldsymbol{u} \sim \boldsymbol{MVN}_n(\boldsymbol{0}, \boldsymbol{\lambda\tau^{-1}K}), \; \boldsymbol{\epsilon} \sim MVN_n(0, \tau^{-1}I_n),$$

where $\boldsymbol{y}$ is an n-vector of quantitative traits (or binary disease labels) for n individuals; $\boldsymbol{W} = (\boldsymbol{w_1}, \cdots, \boldsymbol{w_c})$ is an $n \times c$ matrix of covariates (fixed effects) including a column of 1s; $\boldsymbol{\alpha}$ is a c-vector of the corresponding coeffcients including the intercept; $\boldsymbol{x}$ is an n-vector of marker genotypes; $\boldsymbol{\beta}$ is the effect size of the marker; $\boldsymbol{u}$ is an n-vector of random effects; $\boldsymbol{\epsilon}$ is an n-vector of errors; $\boldsymbol{\tau^{-1}}$ is the variance of the residual errors; $\boldsymbol{\lambda}$ is the ratio of the individual to error variance components; $\boldsymbol{K}$ is a known $n \times n$ relatedness matrix and $\boldsymbol{I_n}$ is an $n \times n$ identity matrix. MVN_n denotes the n-dimensional multivariate normal distribution.

The alternative hypothesis H_1: $\beta \neq 0$ is tested against the null hypothesis H_0: $\beta = 0$ for each SNP in turn, using one of three commonly used test statistics; the Wald, likelihood ratio or score statistics. Either the maximum likelihood estimate (MLE) or the restricted maximum likelihood estimate (REML) of λ and β are obtained and a p-value is calculated under the null hypothesis. The PVE by typed genotypes or “chip heritability” is also estimated (Zhou & Stephens 2014). When the traits being studied are not genetically correlated, univariate models can have an increase in power over multivariate models to detect associated loci (Korte et al. 2012). However, univariate models are unable to detect pleiotropic

loci, making these models less undesirable for studying associations when there are multiple complex traits.

An under-utilized approach that has the power to detect pleiotropic loci is multivariate linear mixed model (mvLMMs). Multivariate models are commonly used in genetic analyses, with applications in estimating cross-tissue heritability of gene expression, calculating genetic correlation between complex phenotypes, detecting pleiotropic quantitative trait loci, and enabling animal breeding programs. Multivariate linear mixed models are very effective in controlling for population stratification while testing SNP associations with multiple correlated phenotypes (Zhou & Stephens 2014).

As explained by Zhou and Stephens the statistical equation for mvLMMs is:

$\tilde{Y} = A\widetilde{W} + \beta\tilde{x}^{T} + \tilde{G} + \tilde{E}$; $\tilde{G} \sim MN_{d\times n}(0, V_g, K)$, $\tilde{E} \sim MN_{d\times n}(0, V_e, I_{n\times n})$,

where *n* is the number of individuals, *d* is the number of phenotypes, $\tilde{Y}$ is a d by *n* matrix of phenotypes, $\widetilde{W}$ is a *c* by *n* matrix of covariates including a row of 1s as intercept and A is a *d* by *c* matrix of corresponding coefficients, $\tilde{x}$ is a *n*-vector of genotype for a particular marker and β is a *d*-vector of its effect sizes for the *d* phenotypes, $\tilde{G}$ is a *d* by *n* matrix of random effects, $\tilde{E}$ is a *d* by *n* matrix of residual errors, K is a known *n* by *n* relatedness matrix, $I_{n\times n}$ is a *n* by *n* identity matrix, V_g is a *d* by *d* symmetric matrix of genetic variance component, V_e is a *d* by *d* symmetric matrix of environmental variance component and $MN_{d\times n}(0,V1,V2)$ denotes the *d*×*n* matrix normal distribution with mean 0, row covariance matrix V1 (*d* by *d*), and column covariance matrix V2 (*n* by *n*). The null hypothesis (the

marker effect sizes for all of the phenotypes are zero) H_0: $\beta = 0$, where 0 is a d- vector of zeros, is compared against the general alternative H_1: $\beta \neq 0$. A maximum likelihood estimate (MLE) or the restricted maximum likelihood estimate (REML) of V_g and V_e is calculated. The *p*-value is reported for each SNP. (Zhou & Stephens 2014).

There has been a growing appreciation of the power gains from mvLMMs. Zhou and Stephens (2014) conducted a study comparing results from univariate LMM analyses and mvLMM analyses using four blood lipid phenotypes from the Hybrid Mouse Diversity Panel. There were 16 significant SNPs in the four-phenotype multivariate analysis; however, none of these 16 SNPs were found to be significant in the univariate analyses. This study shows that mvLMMs can increase the power to detect pleiotropic genetic variants. Power gains can still be achieved even when genetic variants affect only one of the multiple correlated phenotypes (Zhou & Stephens 2014). Given that univariate models are unable to detect pleiotropic loci, mvLMMs are more desirable for studying associations when the recorded data involve multiple complex traits.

Introduction to Genomic Prediction

Just as researchers continue to discover SNP x trait associations through GWAS, genomic predictions of complex traits from genotypic data for individuals in plant, animal, and human populations are becoming increasingly popular (Wray et al. 2013). Genomic prediction refers to the “prediction of genetic merit of selection candidates based on genome-wide marker genotypes using information

from a reference population of individuals with both phenotypes and genotypes" (Moghaddar et al. 2014).

Genomic prediction models (i.e., whole-genome regression methods) can be used in a variety of ways, but two noteworthy uses are in heritability estimation (also known as chip heritability or PVE) and in prediction of breeding values. Estimates of the PVE (the total proportion of variance in phenotype explained by the allele substitution effects) for large-scale genotyped marker sets can be informative relative to sources of "missing heritability" of common diseases (Yang et al. 2010; de Los Campos, Vazquez, et al. 2013) and underlying genetic architecture of these diseases (Zhou et al. 2013; Hayes et al. 2010). Missing heritability is the difference between the PVE for all markers versus the markers that are found to be statistically significant. Part of this missing heritability could be from the lack or of linkage disequilibrium between SNPs and casual variants (Boichard et al. 2012).Genomic prediction has already had a major impact in agricultural production (de Los Campos, Hickey, et al. 2013; Decker 2015), for example, over 1 million Holstein animals have now been genotyped with genome-wide SNP arrays. Applications of whole-genome regression models continue to increase in human and model organism research. GCTA (Yang et al. 2011), a popular software package used in human whole-genome regression research, currently has over 950 citations.

Introduction to Bayesian Methods

Bayesian procedures are known to better handle data structures in which the number of markers largely exceeds the number of observations which is

known as the “small n, large p” situation (Gianola et al. 2009). Marker-specific variances are allowed to vary across the loci fit in these models. Bayesian methods naturally take into account uncertainty about all unknown paramters in a model (Gianola et al. 2009).

(Meuwissen et al. 2001) presented a hierarchical Bayesian model, termed BayesB, that has extensively been used in animal breeding. BayesB is popular among researchers because of its straightforward approach to single site locus sampling and because it has reasonable computation times (Meuwissen et al. 2001). Not only is LD between individual markers and QTLs exploited better in Bayes B than in least-squares or ridge-regression methods but this method also obtains higher accuracies in genomic prediction analyses (Habier et al., 2011.). However, this method possesses statistical drawbacks. Gianola et al. (2009) pointed to the fact that BayesB (or other variants that assume marker-specific variances) does not allow for “Bayesian learning about marker-specific variances so the extent of shrinkage will always be dictated strongly by the prior, which negates the objective of introducing marker-specific variances into the model.” With changes in degrees of freedom and scale parameters describing the prior distribution, estimates of marker effects can be made smaller or larger at the researcher’s discretion (Gianola et al. 2009; Habier et al. 2011). BayesB treats the paramter π (proportion of markers not associated with the trait) as known (Gianola et al. 2009; Habier et al. 2011) . Shrinkage of SNP effects is affected by π, therefore, it should be treated as an unknown and should be inferred from the data (Habier et al. 2011).

One method to overcome the drawbacks of BayesB is the use of a single effect variance that is common to all SNPs instead of the locus-specific variance used in BayesB. This method is called BayesC. Since shrinkage is affected by the magnitude of π, π is treated as an unknown, and then method is then referred to as BayesCπ (Habier et al. 2011). In BayesB models, each SNP has its own variance, while in BayesCπ models, the priors for all SNP effects have a common variance (Habier et al. 2011). Habier et al. compared accuracies of Genetically Estimated Breeding Values (GEBVs) for four traits in Holstein cattle produced using a BayesB method and a BayesCπ method and found that estimates produced by BayesCπ were sensitive to both the training data size and the SNP architecture of the quantitative trait while Bayes B were not. The BayesCπ method did not result in a significant increase in accuracy relative to the Bayes B method used in this analysis (Habier et al. n.d.). However, a study conducted by Hulsman Hanna in a Nellore-Angus crossbred population found that BayesCπ yielded higher accuracies but a larger bias in GEBV compared to the BayesB analysis (Hulsman Hanna et al. 2014).

Another Bayesian method, Bayesian Sparse Linear Mixed Models (BSLMM) performs similarly to BayesCπ in scenarios that involve small numbers of large effect variants but outperforms BayesCπ in scenarios when trait variance involves a large number of SNPs that with small effects. The Bayesian sparse linear mixed model (BSLMM) methodology is a hybrid between linear mixed models (LMMs) and Bayesian sparse regression developed by Zhou, Carbonetto, and Stephens (2013). Linear mixed models are used in association

studies to control for population stratification and relatednesswhereas, sparse regression models are often used in expression QTL analyses (Zhou et al. 2013).

These models are both used for estimating breeding values in genomic studies. While having considerable overlap in their applications, LMMs and Bayesian sparse regression models are based on diametrically opposed assumptions (Zhou et al. 2013). LMM approaches are effective at assuming that with effect sizes normally distributed, every genetic variant affects the phenotype. Bayesian sparse regression on the other hand, assumes that a relatively small proportion of all variants are associated with the phenotype. The performance of these two methods would be expected to vary depending on the true underlying genetic architecture of the phenotype, which is usually unknown while studying a trait (Zhou et al. 2013). This presents a challenge in determining which of the two models is more appropriate for a given analysis. This dilemma motivated Zhou, Carbonetto, and Stephens to develop the hybrid method. “BSLMM consists of a standard linear mixed model, with one random effect term, and with sparsity inducing priors on the regression coefficients.” (Zhou, Carbonetto, and Stephens 2013). As shown by Zhou, Carbonetto, and Stephans the statistical equation for BSLMM is:

$$\boldsymbol{y} = \boldsymbol{1}_n\mu + \boldsymbol{X\beta} + \boldsymbol{u} + \boldsymbol{\epsilon};\ \beta_i \sim \pi N(0, \sigma^2 \underset{a}{\tau^{-1}}) + (1-\pi)^{\delta}0,\ \boldsymbol{u} \sim MVN_n(0, \sigma^2 \underset{b}{\tau^{-1}}\boldsymbol{K}),\ \epsilon \sim MVN_n(0, \boldsymbol{\tau^{-1}I_n}),$$

where $\mathbf{1_n}$ is an *n*-vector of 1s, $\mathbf{u}$ is an n-vector of random effects, **X** is an *n*×*p* matrix of genotypes measured on n individuals at *p* genetic markers, β is the corresponding *p*-vector of the genetic marker effects, $\boldsymbol{\epsilon}$ is an *n*-vector of errors, $\boldsymbol{\lambda}$ is the ratio of the genetic to error variance components, $\boldsymbol{K}$ is a known *n*×*n* relatedness matrix and $\boldsymbol{I}_n$ is an *n*×*n* identity matrix, *and* MVN_n denotes the n-dimensional multivariate normal distribution. BSLMM uses MCMC to estimate β, u and all other hyper-parameters including μ and τ^{-1} control the phenotype mean and residual variance, π controls the proportion of $\tilde{\beta}$ values in (6) that are non-zero, σ_a controls the expected magnitude of the (non-zero) $\tilde{\beta}$, and σ_b controls the expected magnitude of the random effects **u**.

This model is capable of "learning" the genetic architecture of the analyzed trait from the data to a certain extent, and this could increase model performance across a large range of scenarios (Zhou et al. 2013). BSLMM estimates the hyper-parameters (μ, τ^{-1}, π, σ_a and σ_b) from the data instead of fixing them to pre-specified values. This achieves the adaptive behavior of the model. Since BSLMM is therefore somewhat adaptive to the data, it eliminates the need to select one model over the other. Models like BSLMM have been used to predict breeding values to assist with genomic selection in plants and animals, controlling for batch effects while selecting for genes in expression analysis, prediction of phenotypes for complex traits, and mapping complex traits by jointly modeling all SNPs in structured population (Zhou et al. 2013).

A difference between BSLMM and other Bayesian models is that BSLMM is a novel computational algorithm that avoids ad hoc approximations that can be

made when fitting these models and presents reliable results for large populations that contain hundreds of thousands of markers (Zhou et al. 2013). Overall, BSLMM frequently outperforms LMMs, Bayesian variable selection regression models (BVSR), and several related models in predicting phenotypes in complex traits (Zhou et al. 2013; Berger et al. 2015).

Performing Genomic Prediction Fitting SNPs as Effects

In genomic prediction methods, single nucleotide polymorphisms (SNPs) are most commonly used, each as an individual explanatory variable. SNPs are fit as variables in linear and non-linear Bayesian methods previously discussed and the effects of each individual marker are simultaneously estimated (Taylor 2014). While SNPs have been shown to perform well for estimating breeding values, there are several limitations to using these markers as predictors (Wray et al. 2013; Hulsman Hanna et al. 2014; Kachman et al. 2013).

One of these limitations are predicting phenotypes. Variation in complex traits is caused by both genetic and environmental factors (Wray et al. 2013). One of the best ways to quantify the importance of additive genetic factors is the narrow sense heritability (h^2), or the proportion of phenotypic variation in a trait that is explained by additive genetic factors (Wray et al. 2013; Villumsen et al. 2009; de los Campos et al. 2015; Makowsky et al. 2011). It is assumed that the model-estimated h^2 is an unbiased estimate of the population parameter, meaning that h^2 is the upper limit for the proportion of the phenotypic variance that can be explained by a linear predictor of additive genetic merit based on the

genomic markers (Makowsky et al. 2011; de los Campos et al. 2015; Wray et al. 2013; Villumsen et al. 2009). If there are non-additive genetic or permanent environmental effects, such linear predictors of genetic merit can never fully explain all of the phenotypic variation (Wray et al. 2013; de Los Campos, Hickey, et al. 2013; de Los Campos, Vazquez, et al. 2013). Researchers can only achieve this upper limit if all of the genetic variants that directly affect each trait are known and if the effects of these variants are estimated without error (Wray et al. 2013).

The second limitation is the variance that is explainable by markers as opposed to the causal variants that underlie trait variation. Most of the SNPs that are on a SNP chip are typically not causal variants for any phenotype (Wray et al. 2013). However, they can be associated with a trait because they are in linkage disequilibrium (LD) with one or more causal variants (Wray et al. 2013; Makowsky et al. 2011). This is only true for causal variants with common alleles because SNPs that are included on most commercially used SNP chips are common variants, therefore they cannot be in strong LD with rare causal variants (Makowsky et al. 2011; Wray et al. 2013). One In both livestock and humans, some casual variants are rare and in poor LD with SNPs, thus are unable to be predicted by SNPs on SNP chips (Wray et al. 2013; Boichard et al. 2012; Cuyabano et al. 2015; Cuyabano et al. 2014).

A third limitation of using SNPs for genomic prediction is error in the estimated effects of the markers. Some traits are affected only by a few loci, making it possible to estimate their effects quite accurately and with low sampling

errors. However, most complex traits are controlled by a large number of unknown loci (Wray et al. 2013; de los Campos et al. 2015). Selecting for the most significant SNPs and using those to estimate SNP effects can lead to bias and the variance explained will be inflated (Powell & Zietsch 2011). A fourth limitation is the statistical methods used in the training sample, discussed in the Bayesian methods section.

Performing Genomic Prediction Fitting Haplotypes as effects

An alternative approach to performing genomic prediction is to use haplotypes as explanatory variables instead of SNPs. In 2001, Meuwissen, Hayes, and Goddard (2001) introduced this idea of using haplotypes for genomic prediction. However, researchers subsequently tended to fit SNPs in Bayesian models after the BovineSNP50 assay was introduced in 2008.

As discussed above, an individual marker effect is probably not the best predictor of a QTL effect. It is hypothesized that haplotypes may be in higher linkage disequilibrium (LD) with the causal variants than individual SNPs are. Therefore, haplotypes are estimated to improve accuracy when used as predictors to estimate breeding value. Previous studies have shown that the reliability of predictions fit using haplotypes is greater than predictions fit using individual markers (Villumsen et al. 2009). For example, comparisons of predictions in a Nordic Holstein population showed an increase in reliability of 3% when predictions were made using haplotypes versus individual SNPs. (Cuyabano et al. 2014).

The literature on using haplotypes as effects for genomic predictions has shown two different methods for constructing haplotypes. The most efficient number of markers that a haploblock should contain, the most advantageous number of haplotypes to be used across the genome, and in which genomic regions the haploblocks should be defined are three questions that should be asked before constructing haploblocks (Moghaddar, Swan, and Hj Van Der Werf 2014). Some researchers are concerned that using haplotypes would drastically increase the number of variants to be estimated because the number of explanatory variables is increased.(Cuyabano et al. 2014).

The first method for constructing haplotypes is to use local patterns of LD to define where haploblocks start and end throughout the genome. Markers can be grouped into haploblocks that are not of fixed length (either number of SNPs or physical distance), by defining the interval in which to haplotype SNPs according to the minimum amount of LD between pairs of markers (usually estimated using r^2). The number of SNPs per haploblock may be substantially reduced in genomic regions with relatively strong LD (Cuyabano et al. 2014). The study conducted in Nordic Holstein cattle by Cuyabano, in which haploblocks were created using LD, found that predictions made from haplotypes achieved reliabilities equal to or higher than the predictions using SNPs. An increase of up to 1.3% in accuracy was achieved for the trait mastitis when using haploblocks over SNPs (Cuyabano et al. 2014).

An alternative method used to construct haplotypes is to form haploblocks from a fixed number of SNPs. Boichard et al. (2012) constructed haplotypes by

selecting the SNPs found in the same centimorgan region of the genome. If only a single marker was present in any such region, two flanking SNPs were also selected, so that a haplotype contained at least three markers. Biochard et al. (2012) found better predictions in dairy cattle traits can be obtained by using haplotypes compared to predictions based on individual markers.

It has been shown by the studies above that haplotypes can increase the accuracy of genomic predictors of additive genetic merit compared to those in obtained using SNPs fit as effects in the model. Again, this is due to the fact that haplotypes increase the extent of linkage disequilibrium between prediction loci and causal variants. Therefore, the use of haplotypes in genomic prediction should be preferred over the use of individual markers.

Issues Concerning the Prediction of Breeding Values in Crossbred Animals

A major advantage of genomic prediction is that predictors can increase the accuracy of predicted breeding values in animals that have no or few progeny (Boichard et al. 2012; Kachman et al. 2013; Snelling et al. 2015). Several U.S. breed associations have implemented genomic predictions in their National Cattle evaluations for routinely recorded traits (Snelling et al. 2015; Kachman et al. 2013). Breeds have routinely gained increases in accuracy from within-breed genetic predictions. However, using predictions that have been trained in one breed to predict breeding values in a different breed has not been successful (Weber et al. 2012; Kachman et al. 2013). Models using genomic predictions that

were trained in multiple breeds have been able to increase accuracy in crossbred animals, but mainly when the candidate breeds were included in the reference population (Boichard et al. 2012; Weber et al. 2012). When the candidate breeds are distantly related or were not included in the reference population, accuracy tends to be low or zero (Kachman et al. 2013). This suggests that among distantly related breeds, phase relationships between quantitative trait loci (QTL) and SNP markers is not preserved (Kachman et al. 2013). This is especially problematic for breed associations that cannot afford to collect data to genotype the large reference populations required for effective genomic prediction. These breeds can overcome this problem by performing single step best linear unbiased prediction (BULP) analysis (Misztal et al. 2013; Legarra et al. 2009). However, reliability of genomic prediction in commercial crossbred cattle continues to be problematically low. Across-breed evaluations may offer new opportunities for these breeds and is therefore, a necessary evolution in genomic prediction (Boichard et al. 2012).

Summary

Multivariate linear mixed models are able to account for population structure that frequently exists in GWAS analyses, while simultaneously considering within and between trait variances and covariances for two or more phenotypes (Stephens 2013; Kachman 2008). Even though multivariate analyses are known to be more powerful than univariate analyses, no single test will always be the most powerful in a GWAS study.(Zhou & Stephens 2014).

Therefore, no single test will always be the most powerful in a GWAS study. Multivariate and univariate analyses should be seen as being complementary instead of competing approaches (Zhou & Stephens 2014; Stephens 2013). The advantages of increased accuracy in predicting breeding values using haplotypes as effects instead of individual markers in a BSLMM model, compared to other Bayesian models and linear mixed model (LMM) methods, could overcome the issues of low accuracies for genomic estimates of genetic merit in breeds differing from the training set breeds and in crossbred cattle.

Publication Outline

The following manuscripts are under preparation for publication and are presented as chapters in this thesis:

1. Miranda L. Wilson, Robert D. Schnabel, Jeremy F. Taylor, Jared E. Decker. Multivariate Genome-Wide Association Study Across Six Breeds for Tenderness and Other Carcass Traits. Genetics, Selection, and Evolution.

2. Miranda L. Wilson, Robert D. Schnabel, Robert L. Weaber, Jeremy F. Taylor, Jared E. Decker. Using Haplotype Based Models for Genomic Predictions in Crossbred Animals. Genetics, Selection, and Evolution.

CHAPTER 2

MULTIVARIATE GENOME-WIDE ASSOCIATION STUDY OF CARCASS TRAITS ACROSS SIX BREEDS OF CATTLE

Background

Genome-wide association studies aim to assess associations between genetic variants and one or more phenotypes. Sample size is an important component of a GWAS design (Charan & Kantharia 2013). It is frequently the case in livestock research that we measure a large number of phenotypes in a relatively small number of animals. When there are small sample sizes or small QTL effects, analyses can have low statistical power. Low power can negatively affect the likelihood that a significant finding is actually reflective of a true effect (Button et al. 2013). Therefore, even if biologically large variant effects are present in a population, they could be missed in the sample set.

Another common concern in GWAS is accounting for the complicated structure within the data, between loci as well as between individuals (Zhang et al. 2012). Linear mixed models (LMM) are known to be a flexible approach for accounting for population structure in GWAS analyses. Mixed models handle population structure by partitioning the phenotypic covariance into genetic variance due to the genetic kinship among individuals and residual environmental variance (Stephens 2013). The majority of published association studies that utilize LMMs use univariate models. These univariate analyses, consider each phenotype independently even when multiple phenotypes are available for each individual (Stephens 2013). Univariate models can have greater power over multivariate models to detect QTL when they are phenotype specific, or

when the trait studied has no underlying genetic correlation with all other traits and thus all QTL are expected to be trait specific (Korte et al. 2012). However, univariate models are unable to directly detect pleiotropic loci, making these models undesirable for studying associations in complex traits.

When multiple measurements are taken from an individual, the resulting phenotypes can be correlated because of pleiotropy or gametic phase disequilibrium (Lynch et al. 1998). Researchers have expanded on the LMM approach to handle correlated phenotypes by utilizing a multivariate linear mixed model approach (mvLMM) (Stephens 2013). Multivariate analyses consider within-trait and between-trait (co)variances simultaneously for multiple traits. Multivariate analyses have the potential to increase power compared to univariate mixed models since information on correlated traits contributes to the estimation of effects for pleiotropic QTL (Zhou & Stephens 2014). This is essential for studies that are based on small sample sizes.

The National Cattlemen's Beef Association (NCBA) checkoff funded Carcass Merit Project (CMP) was initiated to validate markers for economically important traits, such as tenderness and marbling, and to improve consumer satisfaction with beef products (Devon 1998). Unfortunately, because of budget constraints, the project was unable to phenotype a large number of animals. However, many phenotypes were collected from the individuals in the study. In this study, we analyzed a data set comprised of 8 carcass traits: Warner-Bratzler shear force (WBSF), marbling (MB), rib-eye area (REA), hot carcass weight (HCW), fat thickness (FT), internal fat (IF), cooking loss (CL), and kidney, pelvic, and heart fat (KPH). Warner-Bratzler shear force is a measure of meat tenderness. We hypothesize that multivariate linear mixed models will

not only increase power over univariate analyses but will also focus on associations of genetic variants that are not seen in univariate LMMs because of pleiotropy.

Materials and Methods

Animals

A total of 3,504 animals representing six breeds from the Carcass Merit Project were used for this research (Table 2.1). These six breeds included Angus, Charolais, Hereford, Limousin, Maine-Anjou, and Simmental. Hereford sires were mated to commercial Hereford dams. Angus, Limousin, Charolais, Simmental, and Maine-Anjou sires were mated to predominantly commercial Angus cows. In additional to the Carcass Merit Project samples, we used 3,993 Angus, 101 Charolais, 1,255 Hereford, 2,366 Limousin, 11 Maine-Anjou, and 1,913 Simmental registered animals with Illumina BovineSNP50 data to assist in the phasing of genotypes. We created a multi-breed population by including phenotypes and genotypes of every individual from all six breeds.

Genotypes

All samples were genotyped using the Illumina BovineSNP50 BeadArray for 54,790 single-nucleotide polymorphisms (SNPs), previously described in (McClure et al. 2012). Genotypes were filtered using PLINK v.109 with a call rate of >0.90 and minor allele frequency of >0.01 for each SNP within each breed. Animals were excluded from the data set if their overall genotype call rate was <0.9. After filtering, the data set contained 38,686 SNPs assayed in 3,504 animals.

Analysis

BEAGLE v3.3.2 (Browning & Browning 2007) was used to phase all genotypes and to impute missing genotypes. A standardized genomic relationship matrix was calculated in GEMMA v0.94. GEMMA estimates the relationship matrix from standardized genotypes. A standardized genotype has a genetic mean of zero and a genetic variance of one. Zhou and Stephens (Zhou & Stephens 2014) explain how GEMMA estimates the relationship matrix by denoting X as the $n \times p$ matrix of genotypes, x_i as its ith column representing genotypes for the ith SNP, $\bar{x}_i$ as the sample mean and v_{xi} as the sample variance of ith SNP, and 1_n as a $n \times 1$ vector of 1's.

To account for population structure (families and breeds) in our data, a linear mixed model (LMM) was used to perform association studies. Implemented in GEMMA software, fixed effects (overall mean, covariates and SNP effect) and random effects (individual effect and residual errors) were included in the mixed model. Contemporary groups defined as breed, herd of origin, gender of calf, and slaughter date, were fit as factors. The model was defined by (Zhou and Stephens) as $y = W\alpha + x\beta + u + \varepsilon$ where y is an n-vector of quantitative traits for n individuals; W is an h x c matrix of covariates (fixed effects) including a column of ones, α is a c-vector of the corresponding coefficients including the intercept; x is an n-vector of marker genotypes; β is the effect size of the marker; u is an n-vector of random effects; ε is an n-vector of errors. Univariate analyses were performed for each trait for both the individual breed populations and the multiple-breed data set (genotypes from every individual from all six breeds were pooled).

To capture information due to the underlying genetic correlations between the traits, we also utilized a multivariate linear mixed model approach in GEMMA. The

model was defined by Zhou and Stephens (2014) as $\boldsymbol{Y} = \boldsymbol{WA} + x\boldsymbol{\beta}^T + \boldsymbol{U} + \boldsymbol{E}$ where $\boldsymbol{Y}$ is an n x d matrix of d phenotypes for n individuals; $\boldsymbol{W}$ is an n x c matrix of covariates (fixed effects) including a column of ones; $\boldsymbol{A}$ is a c by d matrix of the corresponding coefficients including the intercept; x is a n-vector of marker genotypes; $\boldsymbol{\beta}$ is a d vector of marker effect sizes for the d phenotypes; $\boldsymbol{U}$ is an n by d matrix of random effects; $\boldsymbol{E}$ is an n by d matrix of errors (Zhou & Stephens 2014). GEMMA uses Restricted Maximum Likelihood (REML) to estimate the variance components σ^2_A and σ^2_E and the genetic and environmental covariance components in the multivariate analysis. We applied eight-phenotype, three-phenotype, and two-phenotype analyses to the CMP data within each breed and in the multiple breed population. We applied the three-phenotype and two-phenotype analyses to the CMP data in GEMMA because the software will only include individuals that have complete values for all of the phenotypes being fit in the model. Therefore, the eight-phenotype analyses uses a smaller proportion of the available individuals because of missing phenotypes, leading to a decrease in power. This is especially true for analyses were marbling is being fit in the model because almost every Simmental animal has a missing value for marbling. Therefore, when fitting all eight-phenotypes in the model using the multi-breed dataset, Simmental is not represented in the analyses. Traits that had been previously shown to be strongly correlated based on published estimates of correlations (Devon 1998; Rolf et al. 2015), were fit in the multiple trait analyses. The p-values were corrected for multiple testing by computing Benjamini-Hochberg q-values using the GenABEL package (Aulchenko et al. 2007) in R. To check for unaccounted population structure or kinship and to assess power, Quantile-Quantile plots were generated using the qqman package (Turner 2014)

in R. Manhattan plots were created using the qqman package to visualize significant SNP associations.

Results and Discussion

Heritability, phenotypic correlations, and genetic correlations were estimated from the multiple breed, eight-phenotype multivariate analysis. (Table 2.2). The heritabilities of each trait were found to be similar to previously published heritabilities (Rolf et al. 2015; Miar et al. 2013; Dikeman et al. 2005). When comparing pedigree-based phenotypic and genetic correlations (Devon 1998) to our genomic-based correlations we found that most of the traits had similar estimates, except that the correlations between WBSF and KPH, MB and REA, and WBSF and MB varied greatly. Genetic correlations of REA with both HCW and FT are similar to those found by (Miar et al. 2013). The Manhattan plots for all of the univariate and multivariate analyses performed are in Figures 2.1, 2.3 and a.1 through a.117. The Manhattan and Q-Q plots for the univariate and the three-phenotype GWAS for hot carcass weight, ribeye area, and fat thickness in the multiple breed population are in Fig. 2.1. In the univariate analyses, four SNPs located on chromosomes 5, 14, and 20 were significantly associated with HCW, one SNP on chromosome 7 was significantly associated with REA, and there were no significant associations with FT. The multivariate GWAS analysis for the three traits resulted in ten association signals located on the same chromosomes as found for the three univariate analyses. In addition, three other significant SNPs located on chromosomes 13 and 26 were identified. As expected, almost all of the SNPs show greater signal in the three-phenotype multivariate analyses (Figure 2.1 b and f). The multivariate analysis had a two-fold increase in significance

compared to the REA univariate analysis and a six-fold increase in significance compared to the MB univariate analysis. Multivariate analyses performed in most of the breeds show a gain in signal compared to the univariate analyses. Maine-Anjou and Simmental did not show any increase in signal for the multivariate over the univariate analyses (Figures a.96 through a.117). This may be attributed to the small sample sizes for both breeds and the lack of MB and IF phenotypes recorded in Simmental.

MSTN

In the Limousin sample, a two-phenotype analysis for MB and REA resulted in a 2.6X increase in the strength of the association on chromosome 2, compared to the univariate analysis of REA. There were not any significant associations on chromosome 2 in the univariate analysis of MB. The Myostatin gene (*MSTN*) is located on chromosome 2 (6,213,566-6,220,196) and harbors mutations that are responsible for the doubling-muscling phenomenon seen in Limousin and other cattle breeds (Sellick et al. 2007). One known *MSTN* mutation, F94L, is responsible for an increase in muscle growth in Limousin cattle (Alexander et al. 2009). The multivariate analysis identified two SNPs that flank the myostatin locus. One SNP is 410,828 kb upstream of myostatin, while the other SNP is 454,849 kb downstream. Several other multivariate analyses fitting either MB or REA in Limousin detected association of SNPs in this region (Figures a.89, a.91, a.93, and a.94). While the multivariate analyses increased the significance of associations in this region, resolution to pinpoint the location of the causal variant was not increased due to the linkage disequilibrium in the region.

Chromosome 7

Multivariate analyses can be informative about the biological processes causing variation in traits from the associations of SNPs and causative variants throughout the genome. (Saatchi et al. (2014) found a pleiotropic QTL on chromosome 7 at 93 Mb that was associated with HCW and REA in Angus, Hereford and Simmental. Several multivariate analyses detected a significant SNP in analyzed traits in Hereford and the multiple-breed population (Figures 2.1, a.6, a.7, a.9, a.11, a.14, a.16, a.20, a.21, a.22, a.23, a.65, a.68, a.71, a.73, and a.74).

The most strongly associated SNP in this region, rs110059753, is located at 93.22 Mb. A promising candidate gene, *ARRDC3* in the arrestin superfamily, is located less than 30 kb away from this SNP at 93.24 – 93.25 Mb (Saatchi et al. 2014). It has previously been shown that obesity is regulated by *ARRDC3* in mice and humans (Patwari & Lee 2012; Patwari et al. 2011). Arrestins are signaling proteins that are known to control metabolism through desensitization of the beta-agrenergic receptors. These receptors are found on the surface of almost every type of mammalian cell and are stimulated physiologically by norepinephrine and epinephrine (Mersmann 1998). Published studies have shown that administering beta-androgenergic agonists to cattle, pigs, sheep, and poultry increases muscle and decreases fat (Beermann 2002; Mersmann 1998). In cattle, if causal mutations that are responsible for variation in carcass traits can be determined as influencing *ARRDC3* transcriptional activity, it could be thought of as a natural beta-adrenergic agonist. Unfortunately, there is no information available in beef cattle on the physiological role of *ARRDC3*.

Chromosome 14

A pleiotropic QTL on chromosome 14 at 24 - 25 Mb is known to segregate in many cattle breeds (Karim et al. 2011; Fortes et al. 2013; Bolormaa et al. 2014; Saatchi et al. 2014). Four SNPs spanning the region from 24.5 – 24.6 Mb were identified in the multivariate analyses performed in Charolais and in the multiple-breed sample, and in every multivariate GWAS that detected this QTL, HCW was included in the model (Figures 2.1, a.6, a.14, a.15, a.16, a.20, a.21, a.22, a.24, a.48, a.49, a.53, and a.55). This QTL region is known to house the gene *PLAG1* (pleiomorphic adenoma gene 1) (Nishimura et al. 2012; Saatchi et al. 2014). Stature in Holstein x Jersey F_2 cross cattle (Karim et al. 2011) and carcass traits in Japanese black cattle (Nishimura et al. 2012) have been shown to be associated with *PLAG1*. Saatchi et al. (2014) detected SNP associations in this region in Simmental, but we did not find any association for Simmental in this region (Figures a.108 through a.117).

Several genes (*LYN*, *TGS1*, and *THEM68*) that are involved in lipid and carbohydrate metabolism are also found in this region (Ramayo-Caldas et al. 2014). *HAS2* is also involved in lipid and carbohydrate metabolism but is upstream 5 Mb (19.7 Mb) from the other chromosome 14 candidate genes (Ramayo-Caldas et al. 2014). When the phenotypes HCW, FT, and KPH were analyzed in the multiple-breed sample in a multivariate GWAS analysis (Figure a.20), we detected an association on chromosome 14 at 21 Mb; 1.3 Mb away from *HAS2*.

Chromosome 20

Three SNPs spanning the region 4.5 – 4.7 Mb on chromosome 20 were identified in Hereford and the multiple-breed sample as being associated with HCW. As for the

QTL on chromosome 14, HCW was included in the model for every multivariate analysis that detected an association in this region of chromosome 20 (Figures 2.1, a.6, a.14, a.16, a.20, a.24, and a.73). This region harbors the gene *ERGICI* (endoplasmic reticulum Golgi intermediate compartment protein 1), which is cycling membrane protein (Saatchi et al. 2014) that is involved in membrane traffic, including transportation between the Golgi apparatus, endoplasmic reticulum, and the intermediate compartment. This gene has important roles in the function of the early secretory pathway (Breuza et al. 2004) and is likely to have other functions besides protein sorting, but these are unclear (Breuza et al. 2004).

Chromosome 5

A Pleiotropic QTL on chromosome 5 located at 106 Mb in Hereford influencing mature weight was detected by (Saatchi et al. (2014). We detected an associated SNP at 106.27 Mb on chromosome 5 in several of the multivariate analyses of the multiple-breed data set, but no significant SNPs were detected in Hereford (Figures 2.1 and a.16). However, Hereford showed suggestive association for this SNP in several of the multivariate GWAS (Figures a.65 and a.75). The detection of the association by this SNP in the multiple-breed sample may be explained by the increase in sample size. This SNP is located in the intron of *TIGAR* (TP53-induced glycolysis and apoptosis regulator) which is a fructose-2, 6-bisphosphatase that aids in the regulation of glucose metabolism (Bensaad et al. 2006).

There were several other significant SNPs found across the genome in the multivariate analyses in the different populations. Two of these SNPs have previously been shown to be associated with trait × breed specific QTL. There is a QTL located on

chromosome 5 at 34 Mb that was found to be associated with REA in Shorthorn (Saatchi et al. 2014). We found a SNP at 37.1 Mb on chromosome 5 that was associated with HCW in the multiple-breed sample (Figures 2.1 and a.16). A QTL located on chromosome 26 at 42 Mb has been associated with yield grade in Red Angus (Saatchi et al. 2014). We found a SNP at 45.8 Mb on chromosome 6 in the multivariate analysis that included all 8 traits (Figure a.24) and the multivariate analysis of HCW, FT, and REA in the multiple-breed sample (Figure 2.1).

Conclusions

Our results show that multivariate analyses can increase the statistical power to detect associations, although if SNP density is not sufficient, the chromosomal position of the identified QTL may not be precise. Multivariate analyses also help to identify QTL that are associated with different traits and help to understand some of the biological processes underlying these traits. We also identify a larger number of trait associated QTL in the multiple-breed sample, reflecting the increased power of GWAS with increased sample size.

Table 2.1 Number of Animals and Sires per breed

Breed	Animals[a]	Sires
Angus	651 (644)	36
Charolais	695 (690)	14
Hereford	1095 (1072)	24
Limousin	283 (281)	13
Maine-Anjou	301 (301)	13
Simmental	516 (516)	40
All Breeds	3,541 (3,504)	140

a. Numbers of animals with genotype call rate ≥ 0.9 in parentheses.

Table 2.2 Heritablilities, phenotypic correlations, and genetic correlations. Heritabilities are on the diagonals in bold, phenotypic correlations are above the diagonals, and genetic correlations are below the diagonals.

Trait Name	MB	WBSF	CL	HCW	FT	REA	KPH	IF
Marbling (MB)	**0.49**	-0.21	-0.14	0.15	0.25	-0.038	0.002	0.07
Shear Force (WBSF)	-0.27	**0.32**	0.27	-0.01	-0.06	0.109	-0.012	-0.01
Cooking Loss (CL)	-0.50	0.32	**0.06**	0.068	-0.02	0.092	-0.013	-0.06
Hot Carcass Weight (HCW)	0.19	0.10	0.31	**0.43**	0.24	0.39	-0.017	0.14
Fat Thick (FT)	0.38	-0.12	-0.35	0.014	**0.45**	-0.17	0.0159	0.1
Ribeye Area (REA)	0.09	0.10	0.34	0.43	-0.59	**0.38**	-0.045	0.04
Kidney, Pelvic, and Heart Fat (KPH)	0.27	0.01	-0.8	-0.032	0.13	-0.26	**0.031**	-0.51
Internal Fat (IF)	0.17	0.03	0.18	0.43	0.27	0.113	0.025	**0.08**

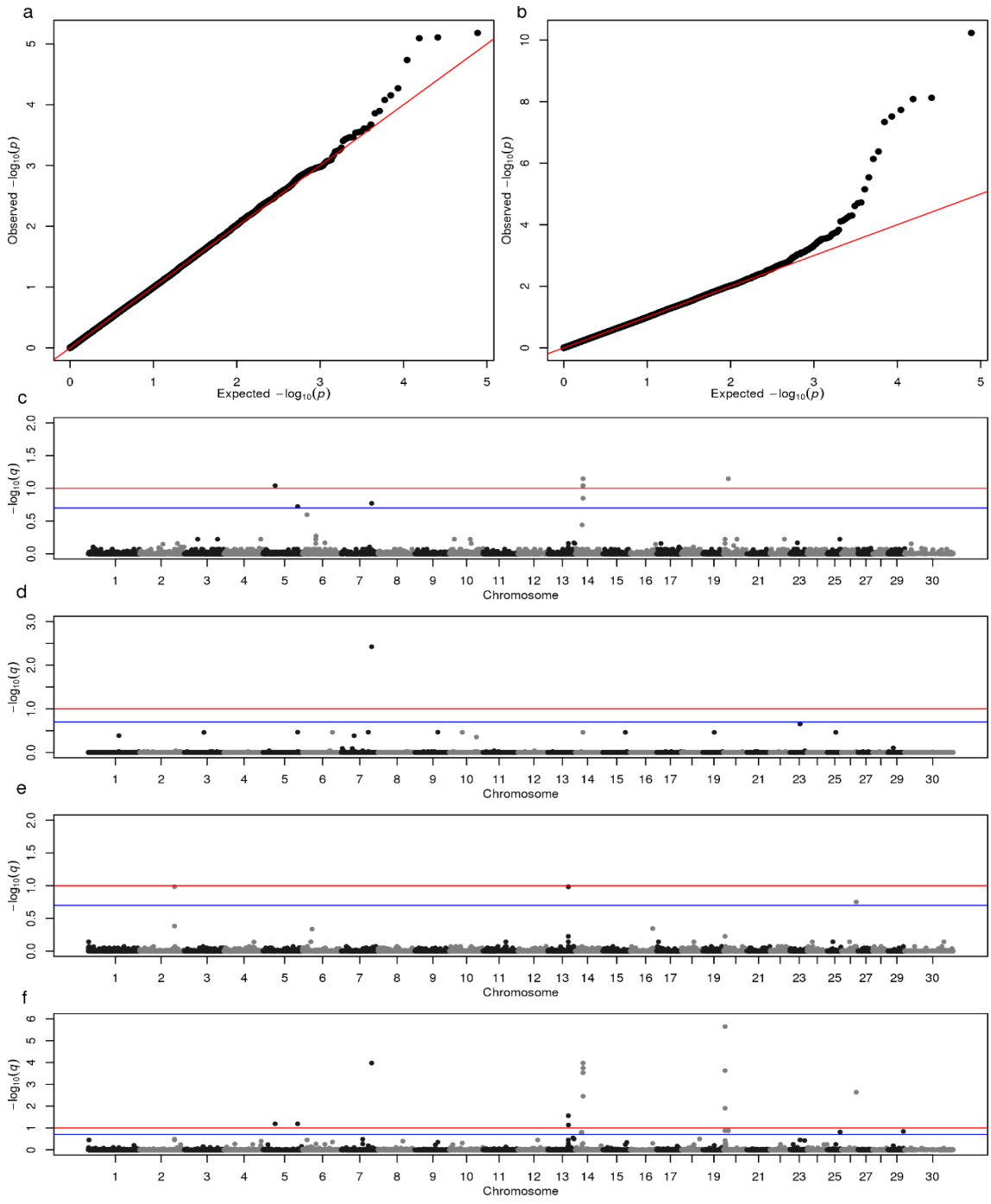

Figure 2.1 Comparison of univariate and multivariate GWAS results of HCW, FT and REA in multiple-breed data set. a Q-Q plot using the *p*-values of the univariate analysis of FT. **b** Q-Q plot using the *p*-values of the multivariate analysis of three phenotypes (HCW, REA and FT) from the multiple-breed sample set. **c** Manhattan plot of SNP *q*-values estimated in the univariate analysis of HCW. **d** Manhattan plot of SNP *q*-values estimated in the univariate analysis of REA. **e** Manhattan plot of SNP *q*-values estimated in the univariate analysis of FT. **f** Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, REA and FT.

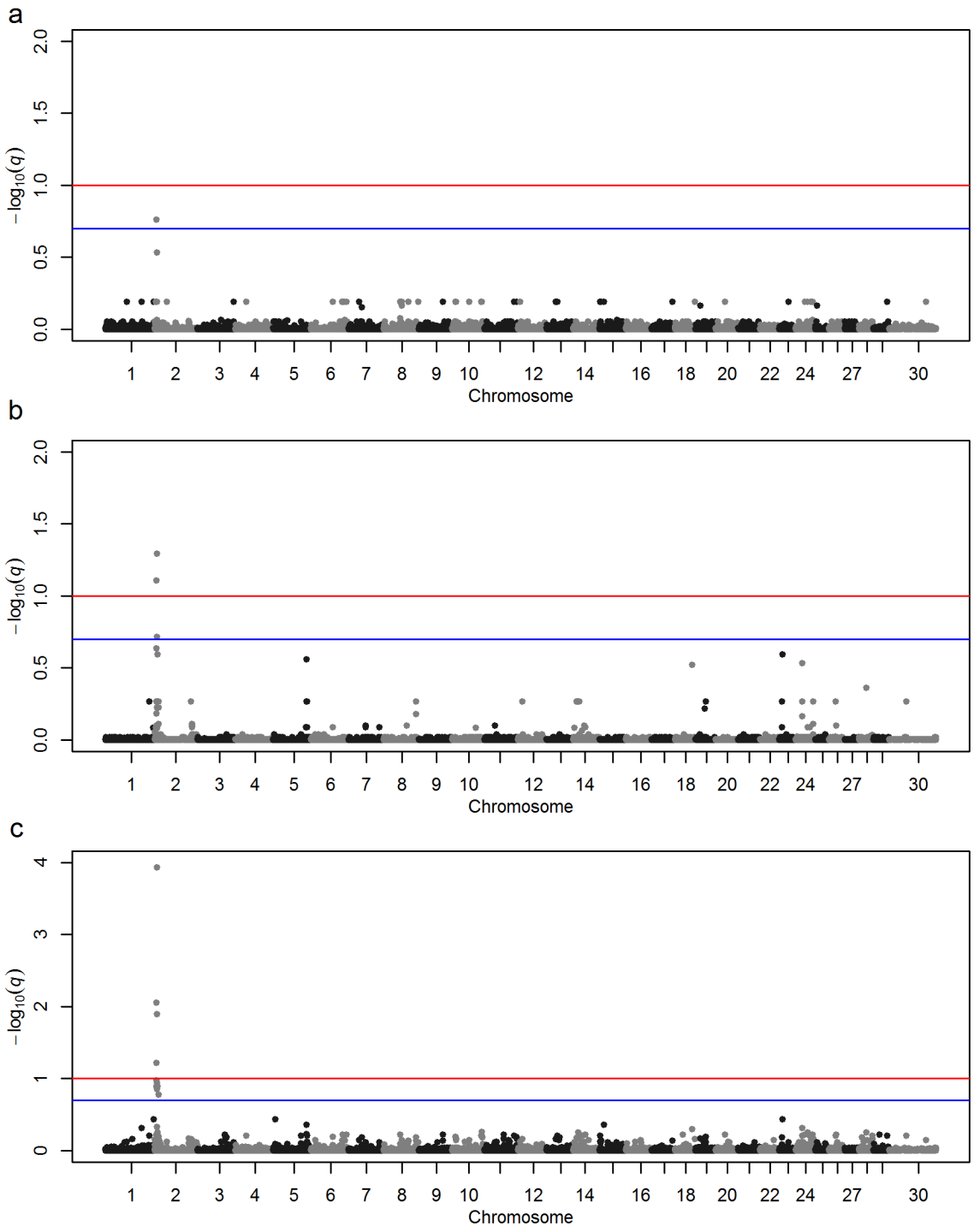

Figure 2.2 Comparison of associations obtained from the multivariate and univariate analyses in Limousin of MB and REA. a Manhattan plot of SNP q-values estimated in the univariate analysis of MB **b** Manhattan plot of SNP q-values estimated in the univariate analysis of REA. **c** Manhattan plot of SNP q-values estimated in the multivariate analysis of MB and REA.

CHAPTER THREE

USING HAPLOTYPE-BASED MODELS FOR GENOMIC PREDICTIONS IN MULTIPLE BEEF BREEDS AND CROSSBRED ANIMALS

Background

Genomic predictions are dependent on the linkage disequilibrium (LD) between individual markers and a causal variant (known or unknown) that affects phenotype (Snelling et al. 2015) and relationships between animals in the training set and selection candidates (Habier et al. 2007). Genomic predictions can increase reliability of predicted breeding values in animals that have no or few progeny. These predictions are mostly breed specific and can be effective, if LD between the single nucleotide polymorphism (SNP) and causal loci remains constant (Snelling et al. 2015). However, when predicting breeding values in breeds that are not included in the reference population or are distantly related, accuracy and correlation tend to be low, suggesting that LD between quantitative trait loci (QTL) and SNPs is low (Kachman 2008). Breed associations that do not possess the funding to collect sufficient data and genotype the number of animals required to have a reliable reference population can overcome this problem by utilizing a one-step best linear unbiased prediction (BULP) model. (Misztal et al. 2013; Legarra et al. 2009). But, developing across-breed models may allow these groups to more effectively predict breeding values (Boichard et

al. 2012). Across-breed genomic predictions will also enable predicting genetic merit in commercial, crossbreed animals.

In genomic prediction methods, SNPs are most commonly used, each as an individual explanatory variable fit simultaneously as random effects (Cuyabano et al. 2015). Alternatively, haplotypes can be used as the predictive variables. In fact, Meuwissen, Hayes and Goddard (2001) originally proposed using haplotypes as predictors in genomic selection. The greatest benefit of using haplotypes is that haplotypes maximize linkage disequilibrium between markers and QTLs (Boichard et al. 2012). Phase relationships between individual SNP markers and causal variants may not be consistent across breeds, but the relationship between SNP marker haplotypes and causal variants should be stronger.

Previous studies show higher reliability of predictions fit using haplotypes than predictions fit using individual markers (Villumsen et al. 2009). For example, comparisons of predictions in a Nordic Holstein population showed an increase in reliability by 3% when predictions were made using haplotypes versus individual SNPs (Cuyabano et al. 2014). Our hypothesis is that using haplotypes from QTLs with the highest absolute substitution effect estimated from the Bayesian Sparse Linear Mixed Model (BSLMM) will allow us to achieve higher correlations than genomic predictions using SNPs in multiple-breed populations. Using these haplotypes should give greater prediction accuracy, especially in crossbred populations.

This study compared genomic predictions using an individual marker approach of all available SNPs and haplotype approaches using different numbers of QTLs to construct haplotypes. To contrast whether haplotypes as effects or feature selection was influencing changes in correlations between genomic predictions and phenotypes, we ran three feature selection analyses. The first analysis used the 5,000 SNPs that were used to construct the 1,000 QTL-haplotype matrix. The second analysis used the 1,000 tagging SNPs that were used to identify the 1,000 QTL-regions. The final analysis used 5,000 haplotypes that contained the highest effects from the haplotype-based predictions. The analyses were performed using data from the Carcass Merit Project dataset, containing two purebred populations, Angus and Hereford, and four crossbred populations, in which continental sires (Charolais, Limousin, Maine-Anjou, and Simmental) were mated to Angus dams. The analyses were also performed in a multiple-breed population that contained all six breed populations. The objective was to fit haplotypes as predictive variables and to compare correlations to genomic predictions using SNPs.

Materials and Methods

Animals

A total of 3,504 animals representing six breeds from the Carcass Merit Project were used for this research (Table 1). These six breeds included Angus, Charolais, Hereford, Limousin, Maine-Anjou, and Simmental. Hereford sires were mated to commercial Hereford dams. Angus, Limousin, Charolais, Simmental, and Maine-Anjou sires were mated to predominantly commercial Angus cows. In

additional to the Carcass Merit Project samples, we used 3,993 Angus, 101 Charolais, 1,255 Hereford, 2,366 Limousin, 11 Maine-Anjou, and 1,913 Simmental purebred animals when phasing genotypes. We also used a multiple-breed data set that contained all of the genotypes and phenotypes from all six breeds. To train and validate our genomic predictions, we used the traits, marbling (Marb), Warner-Bratzler shear force (WBSF), hot carcass weight (HCW), and ribeye area (REA), that had the most phenotypic records in the genotyped samples.

Genotypes

All samples were genotyped using the Illumina BovineSNP50 BeadArray for 54,790 SNPs, as described in (McClure et al. 2012). Genotypes were filtered using PLINK v.109 with a call rate of >0.90 and minor allele frequency of >0.01 within each breed. Animals were excluded from the data set if their individual genotype call rate was <0.9. After filtering, the data set contained 38,686 SNPs assayed in 3,504 animals. BEAGLE v3.3.2 (Browning & Browning 2007) was used to impute missing genotypes and phase all genotypes. The genotypes were then used to generate a genomic relationship matrix across all breeds in GEMMA v0.94. This genomic relationship matrix was used in the rrBLUP package (Endelman, 2011) in R v2.1-7 to pre-adjust the phenotypes for contemporary group effects, which were defined as herd of origin, gender, and slaughter date. The effect of breed was not removed from the phenotype.

Analysis

We used two methods to evaluate the predictive accuracy of our genomic predictions. First, we used 3-fold cross validation (Saatchi et al. 2011) in which k-means clustering, using the PamK function of the fpc package in R, was used to identify three clusters within each of the breeds. We calculated a dissimilarity matrix from the genomic relationship matrix, and identified clusters in which relatedness was maximized within a cluster but minimized between clusters. The individuals in two groups are used as the training set, while the individuals in the other group are the validation set; this is repeated three times with each cluster used as a validation set. We also validate the model by having the individuals of one breed as a validation set and the individuals in the other 5 breeds in the training set. We repeated this for each of the six breeds.

Bayesian Sparse Linear Mixed Models

Genomic prediction was performed for four traits mentioned previously using a Bayesian Sparse Linear Mixed Model (BSLMM) including both individual SNPs and haplotypes. The model is defined by (Zhou et al. 2013) in the equation:

$$\boldsymbol{y} = \boldsymbol{1_n}\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{\epsilon};\ \beta_i \sim \pi N(0, \sigma^2 \tau^{-1}_a) + (1-\pi)\delta 0,\ \boldsymbol{u} \sim MVN_n(0, \sigma^2 \tau^{-1}_b \boldsymbol{K}),\ \boldsymbol{\epsilon} \sim MVN_n(0, \boldsymbol{\tau^{-1}} \boldsymbol{I_n}),$$

where $\mathbf{1_n}$ is an n-vector of 1s used to fit the mean $\boldsymbol{\mu}$, $\mathbf{u}$ is an n-vector of random effects, $\mathbf{X}$ is an $n \times p$ matrix of genotypes measured on n individuals at p genetic markers, β is the corresponding p-vector of the genetic marker effects, $\boldsymbol{\epsilon}$ is an n-vector of errors, $\boldsymbol{\lambda}$ is the ratio of the genetic to error variance components, $\boldsymbol{K}$ is a

pre-calculated $n \times n$ relatedness matrix and $\boldsymbol{I}_n$ is an $n \times n$ identity matrix*, and* MVN_n denotes the n-dimensional multivariate normal distribution. BSLMM uses MCMC to estimate β, u and all other hyper-parameters, including μ the phenotype mean, τ^{-1} control and residual variance, π controls the proportion of $\widetilde{B}$ values in (6) that are non-zero, σ_a controls the expected magnitude of the (non-zero) $\tilde{B}$ and σ_b controls the expected magnitude of the random effects **u**.

Identification of QTLs and construction of haplotype variables

Quantitative trait loci were identified based on estimated SNP effects obtained from a BSLMM analysis of 38,686 SNPs (Figure 3.1a). For each individual trait, the absolute value of each allele substitution effect was calculated based on the estimated SNP effects from the BSLMM SNP model. These absolute values were ranked and a set number of lead SNPs with the highest allele substitution effects were identified (Figure 3.1b). Lead SNPs that tagged QTLs were defined to be separated by more than 4 SNPs, because we did not want QTL haplotypes based on 5 contiguous SNPs to overlap (Figure 3.1 c, d, and e). We chose to select haplotypes based on 5 SNP blocks, because if used in an industry setting 3 SNP blocks would be difficult to phase (especially with missing genotypes).However, models using 7-SNP blocks were evaluated.

To select an optimal strategy for constructing haplotypes with which to predict genetic merit across breeds we compared correlations between adjusted phenotypes and molecular breeding values from BSLMM, obtained from the 3-fold cross-validation, between the three different QTL selection strategies (Figure

3.2 a and b). We tested these strategies using a haplotype-based BSLMM analyses fit to WBSF.

First, we selected 27 lead SNPs from each of the six breeds from a breed-specific SNP-based BSLMM analysis. To avoid overlapping haplotypes, we selected QTLs from the breeds in order of largest sample size to smallest sample size; in other words, 27 QTLs were selected first from Hereford, then Charolais, followed by Angus, Simmental, Maine-Anjou, and finally Limousin. This approach identified 810 SNPs (162 tag-SNPs and 648 flanking SNPs) that were used to construct haplotypes.

Second, we identified 160 lead SNPs from the multiple-breed SNP-based BSLMM analysis of WBSF.

Third, we identified 160 lead SNPs from the multiple-breed analysis of a trait different from the one being predicted. This third strategy, to test whether genomic predictions based on trait-specific QTLs would be more accurate than selecting a defined number of regions that were not necessarily QTLs for that trait, we fit BSLMM analyses with WBSF phenotypes but used haplotypes for QTL that had been selected from a multiple-breed analysis of marbling. This same process was followed in the analysis of MB, HCW, and REA using the 160 QTL identified from the SNP effects estimated for WBSF.

The number of QTLs used to construct haplotypes ranged from 160 to 3000 for the analysis of WBSF and MB. 1,000 QTLs were used to construct haplotypes for the analyses of REA and HCW (Figure 3.2c). These QTLs were identified by lead-SNPs which were the most strongly associated SNPs with the trait being predicted.

When selecting the QTLs with the highest allele substitution effects, efforts were made to ensure that each QTL was separated by four or more SNPs so that haplotypes would not overlap. We refer to these analyses as n QTL-haplotypes where n is the number of QTLs that were used in analysis and each QTL was included in the model using 5-SNP haplotypes. These QTL-haplotype effects were estimated using a coefficient matrix comprising values of 0,1, or 2 for each haplotype according to whether an individual possessed zero, one, or two copies of each haplotype (Figure 3.1e). Thus, each QTL could be represented by up to 32 haplotypes ($2^5 = 32$). SNP effects differ according to the trait; therefore, each trait had different selected tag-SNPs and this lead to different QTLs and 5-SNP haplotypes being used for genomic prediction for each trait.

To test whether the haplotypes or feature selection was improving correlations, we predicted phenotypes using three feature selection models. The first analysis employed the 5,000 SNPs that were used to construct the 1,000 QTL-haplotype matrix (Figure 3.2d). The second analysis reduced the number of overall haplotypes in the 1,000 QTL analysis to the 5,000 haplotypes that with the largest absolute allele substitution effects (Figure 3.2e); another haplotype-based genomic prediction was trained using only these 5,000 haplotypes (not the

complete set of haplotypes from the 1,000 QTLs). The third analysis used the 1,000 tag-SNPs (that had the largest allele substitution effects) to construct the 1,000 QTL-haplotypes in a SNP-based BSLMM (Figure 3.2f).

To evaluate any bias in the predictions, we obtained the intercept and slope from regressing the phenotype on the predicted breeding value (obtained from the output of the BSLMM analyses) for each trait in R. When the slope is closer to one, there is less bias in the model. We performed this regression for the all SNPs analysis, the 5,000 SNP analysis, the 1,000 QTL-haplotypes analysis, and the 5,000 haplotypes analysis.

Results and Discussion

Figure 3.1 represents a heat map generated from the genomic relationship matrix for all animals organized by breed. Each cluster represents a breed and the size of each cluster is proportional to the size of the sample from each breed. Dark purple off-diagonals (smaller relationship coefficients) between Hereford animals and the animals from the other breeds reflect that Hereford sires were breed to Hereford dams, whereas sires from all other breeds were breed to Angus dams.

The correlations from the analyses using all available SNPs to predict phenotypes for each trait are presented in Table 3.1. These provide a set of base correlations that we used to compare to the haplotype and feature selection analyses.

Finding the optimal strategy to identify QTLs

The correlations obtained from the 3-fold cross-validation of predicted phenotypes for the three QTL selection strategies are presented in Table 3.2. In general, when identifying 27 QTLs from the breed-specific SNP-based analysis, we identified different QTLs in each of the breeds. Charolais shared 2 of its highest ranked effects with Hereford, Angus shared 1 with Hereford and Charolais, Simmental shared 3 with Hereford, Charolais and Angus, Maine-Anjou shared 2, and Limousin shared 3. Regions selected from a different trait had lower predictive ability than QTLs selected for the trait being predicted. The 160 QTLs identified from allele substitution effects obtained from the multiple-breed analysis was clearly the best strategy for constructing haplotypes. This strategy of selecting QTLs from a multiple-breed analysis was used for all additional predictions.

Number of SNPs per haplotype

Models fit using 7-SNP haplotypes had smaller correlations between molecular breeding values and phenotypes than the correlations obtained using 5-SNP blocks. When SNPs are separated by more than 250 kb there is little haplotype-sharing between breeds (Gibbs et al. 2009), thus favoring the use of smaller SNP blocks to construct haplotypes.

Optimal number of QTLs

Figure 3.2 presents the correlations obtained from the different number of QTLs used to predict WBSF. We tested haplotypes constructed for 160, 250, 500, 1,000, 2,000, and 3,000 QTLs. The majority of the correlations, calculated

using 3-fold cross-validation, were best from genomic predictions fit with the 1,000 QTL-haplotypes. The 500 QTL-haplotypes analyses had very similar correlations to those from the 1,000 QTL-haplotypes analyses, even surpassing them in the cross-validation when training in the multiple-breed data set and validating in Simmental. Correlations began to decline at and beyond the 2000 QTL-haplotype analyses. We saw the same trend in correlations from genomic predictions for MB (Fig. 3.3).

Feature Selection

To determine if using haplotypes for genomic prediction produced improved correlations of the prediction of phenotypes in a crossbred population or if feature selecting was resulting in increased correlations, we ran three additional analyses predicting WBSF. The first analysis fit a SNP-based genomic prediction using the 5,000 SNPs that were used to construct the 1,000 QTL-haplotype matrix. The second analysis fit a genomic prediction using the 1,000 lead SNPs used to identify the 1,000 QTLs in the haplotype-based model. The third analysis used a subset of 5,000 haplotypes from the 24,188 haplotypes generated for the 1,000 QTL-haplotypes. The 5,000 haplotypes that were fit in this model were the haplotypes with the highest allele substitution effects obtained from 1,000 QTL-haplotypes BSLMM analysis.

Table 3.3 presents a comparison of the correlations (from the 3-fold cross-validation and from training in the 5 breeds and validating in the left out breed) from the analyses using all available SNPs, the 1,000 QTL-haplotype analyses, the analyses using the 5,000 SNPs used to construct the 1,000-QTL haplotype

matrix, the analyses that fit the 1,000 tag-SNP used to select the 1,000 QTLs, and the analyses that used the 5,000 haplotypes with the largest allele substitution effects. All of these analyses predicted phenotypes for WBSF. Figures 3.4, 3.5, and 3.6 present similar results from comparing the correlations calculated from the 3-fold cross-validation performed for MB, HCW, and REA. These results, along with the results of fitting QTLs selected from a different trait from the one being predicted, indicate that feature selection has a strong impact on prediction accuracy. Feature selection has previously been shown to improve predictions (Sarup et al. 2016), this may be due to increasing the signal to noise ratio. Most industry applications use all available SNPs for all traits, and feature selection may be an untapped area to improve genomic predictions.

Bias

We find when regressing the phenotype on the breeding value for each trait that the analysis using all available SNPs tended to have higher bias than any of the other models, with WBSF having the largest bias of the traits. The 5,000 SNP model and the two haplotype-based models decreased the slope to be closer to one, but there was still bias shown in every model for every trait. (Figures 3.9 through 3.24) The regression plot (Figure 3.11) fitting all of the SNPs as effects in a BSLMM analysis for HCW shows the separation of the British breeds (Angus and Hereford) and the continental breeds (Charolais, Limousin, Maine-Anjou, and Simmental), suggesting that breed effects are strongly influencing the model.

Conclusions

The correlations between phenotypes and genomic predictions fit with the 5,000 SNP matrix were greater than for the genomic predictions fit with the 1,000 QTL-haplotypes matrix, thus we conclude that feature selection is a valuable approach to improve the accuracy of genomic predictions. However, correlations were greater for the 5,000 haplotype matrix analyses than either the 5,000 SNP analyses or the analyses using the 1,000 tag-SNPs, suggesting that the use of haplotypes also improves genomic predictions.

Haplotype-based models incorporating feature selection resulted in strong correlations when validation was performed in breeds that were not included in the training set. Consequently, the use of haplotype-based models has utility for the prediction of genetic merit in crossbred animals and in animals of breeds that were not included in the reference population. The next step is to validate that these haplotype and feature selection models are effective at predicting genetic merit in unrelated breeds that were not used in the process of selecting haplotypes.

Table 3.1 Correlations of predictions using all available SNPs.

Train in/Validate in	Tenderness	Marbling	HCW	REA
Multiple-breed/ANG	0.29	0.29	0.28	0.25
Multiple-breed/CHA	0.26	0.3	0.41	0.22
Multiple-breed/HFD	0.17	0.38	0.22	0.16
Multiple-breed/LM	0.04	0.28	0.25	0.34
Multiple-breed/MAAN	0.21	0.31	0.22	0.2
Multiple-breed/SIM	0.18	-	0.09	0.08
Multiple-breed/multiple-breed	0.253	0.54	0.51	0.34
ANG/ANG	0.16	-	-	-
CHA/CHA	0.28	-	-	-
HFD/HFD	0.12	-	-	-
LM/LM	0.08	-	-	-
MAAN/MAAN	0.19	-	-	-
SIM/SIM	0.07	-	-	-

Table 3.2 Comparisons of strategies used to select QTLs for haplotype-based genomic predictions of WBSF.

Train in/Validate in	162 QTLs (27 QTLs from 6 single-breed analyses)	160 QTLs from multiple-breed analysis	160 Marbling QTLs to predict WBSF breeding values
Multiple-breed/ANG	0.28	0.38	0.11
Multiple-breed/CHA	0.22	0.37	0.21
Multiple-breed/HFD	0.26	0.41	0.08
Multiple-breed/LM	0.25	0.34	0.07
Multiple-breed/MAAN	0.24	0.51	0.21
Multiple-breed/SIM	0.17	0.37	0.079
Multiple-breed/multiple-breed	0.19	0.41	0.19

Table 3.3 Comparisons of models to predict phenotypes.

Train in	Validate in	Analyses fitting all SNPs	Analyses fitting haplotypes selected from 1,000 QTL regions	Analyses fitting the 5,000 SNPs that were used to select haplotypes in 1,000 QTL regions	Analyses fitting 1,000 tagging SNPs used to select haplotypes	Analyses using top 5000 haplotypes with largest effects from 22,000 hap analysis
Multiple-breed	ang	0.29	0.45	0.5	0.6	0.64
Multiple-breed w/o ang	ang	0.32	0.5	0.54	-	0.68
Multiple-breed	cha	0.26	0.42	0.48	0.59	0.61
Multiple-breed w/o cha	cha	0.19	0.47	0.53	-	0.69
Multiple-breed	hfd	0.17	0.41	0.52	0.65	0.65
Multiple-breed w/o hfd	hfd	0.16	0.33	0.54	-	0.65
Multiple-breed	lm	0.04	0.37	0.51	0.65	0.67

Multiple-breed w/o lm	lm	-0.01	0.4	0.57	-	0.72
Multiple-breed	maan	0.21	0.48	0.54	0.65	0.72
Multiple-breed w/o maan	maan	0.17	0.54	0.57	-	0.76
Multiple-breed	sim	0.18	0.48	0.56	0.7	0.66
Multiple-breed w/o sim	sim	0.17	0.32	0.57	-	0.71
Multiple-breed	Multiple-breed	0.25	0.45	0.53	0.65	0.66

Table 3.4 Regression coefficients from the different models for WBSF, MB, HCW, and REA.

	Analysis using all available SNPs		**Analysis using 1,000 QTL region haplotypes**		**Analysis using 5,000 SNPs used to select haplotypes**		**Analysis using the top 5,000 haplotypes from the 1,000 QTL region haplotype analysis**	
	intercept	slope	intercept	slope	intercept	slope	intercept	slope
WBSF	-2.40E-09	3.103	-2.86E-10	1.577	-2.15E-09	1.455	-1.40E-09	1.198
MB	0.294	1.508	0.294	1.189	0.294	1.155	0.294	1.069
HCW	-0.079	1.554	-0.079	1.251	-0.079	1.214	-0.079	1.106
REA	-0.007	1.999	-0.007	1.353	-0.007	1.234	-0.007	1.127

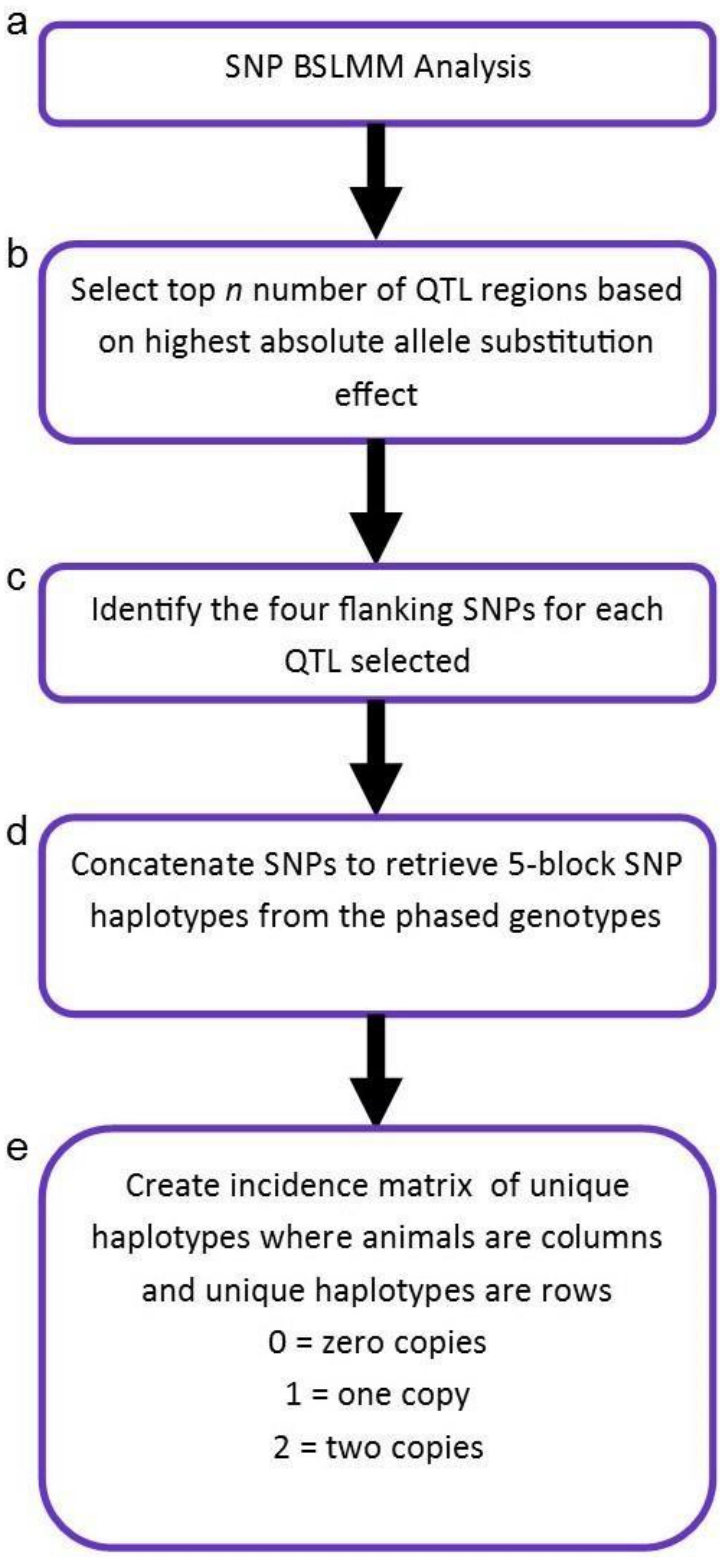

Figure 3.1 Haplotype Selection Process

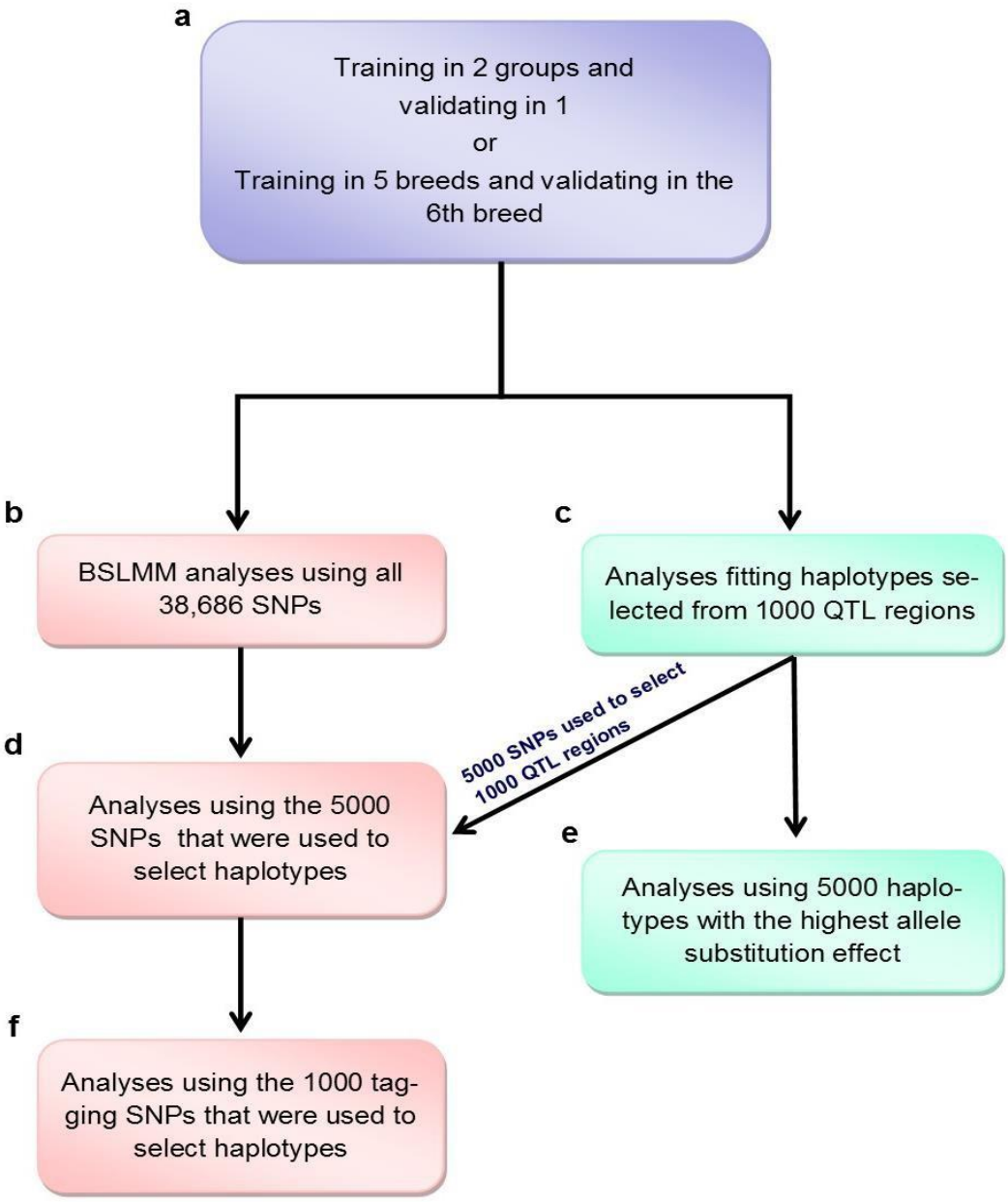

Figure 3.2 SNP, haplotype, and feature selection models used for genomic prediction.

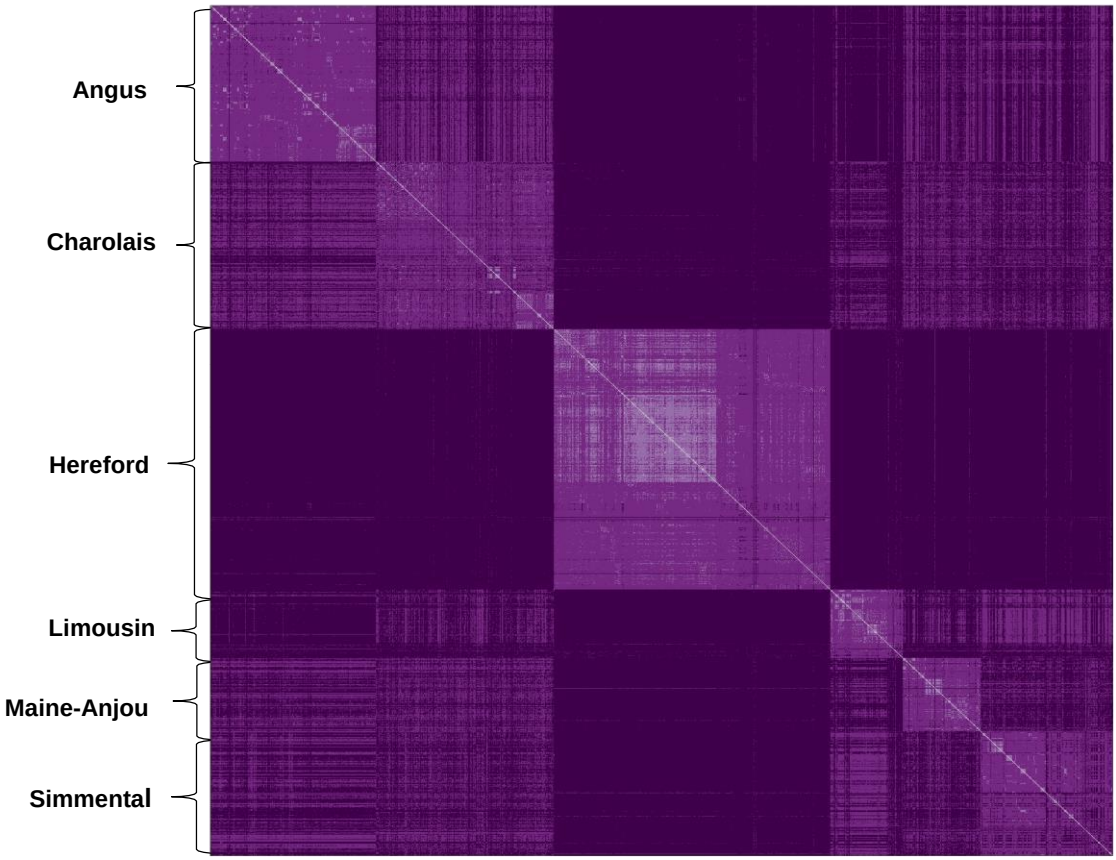

Figure 3.3 Heat map of the genomic relationship matrix (GRM).

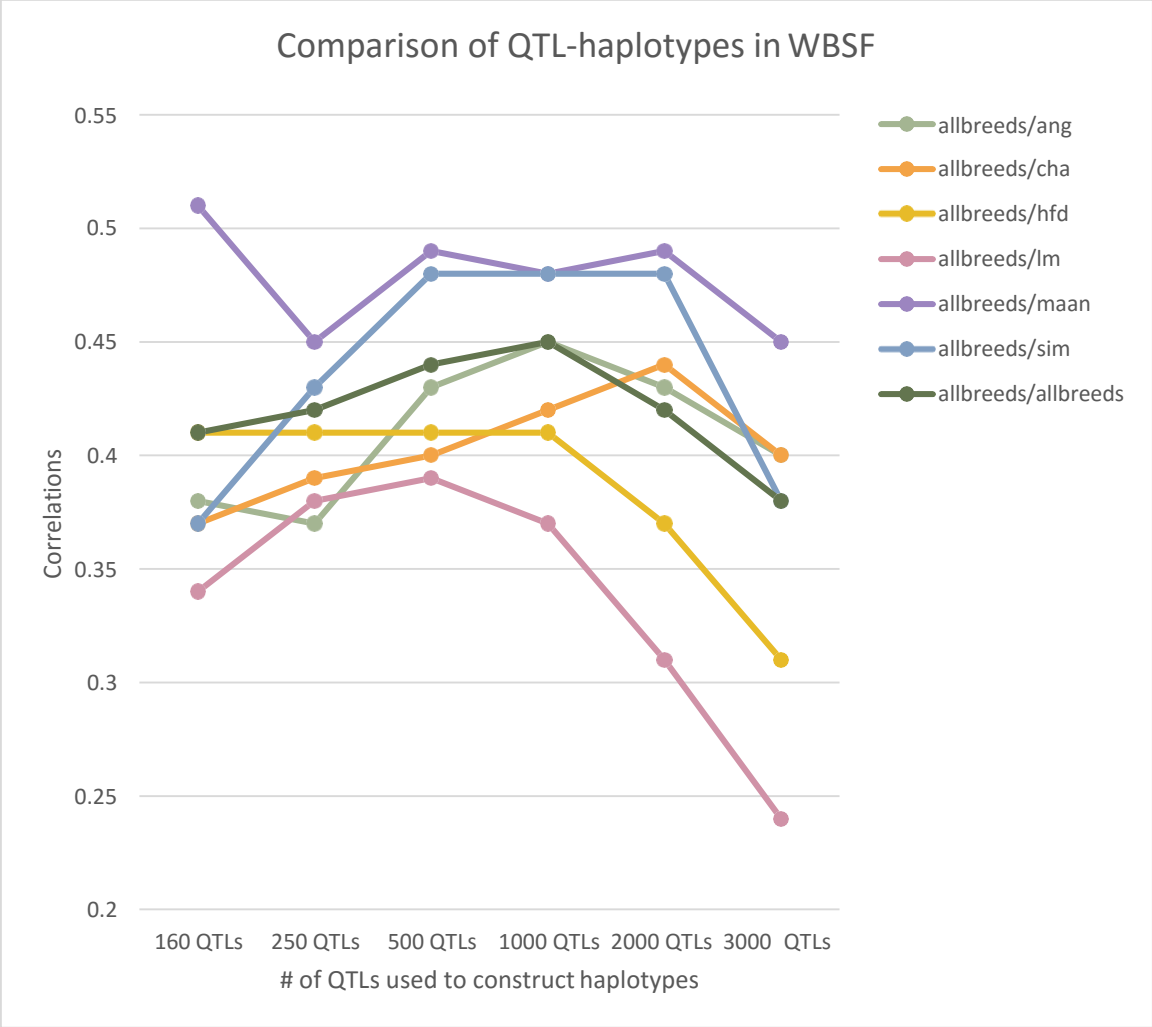

Figure 3.4 Comparison of WBSF correlations obtained from the different QTL-haplotypes predictions. The legend states the training breed/validation breed.

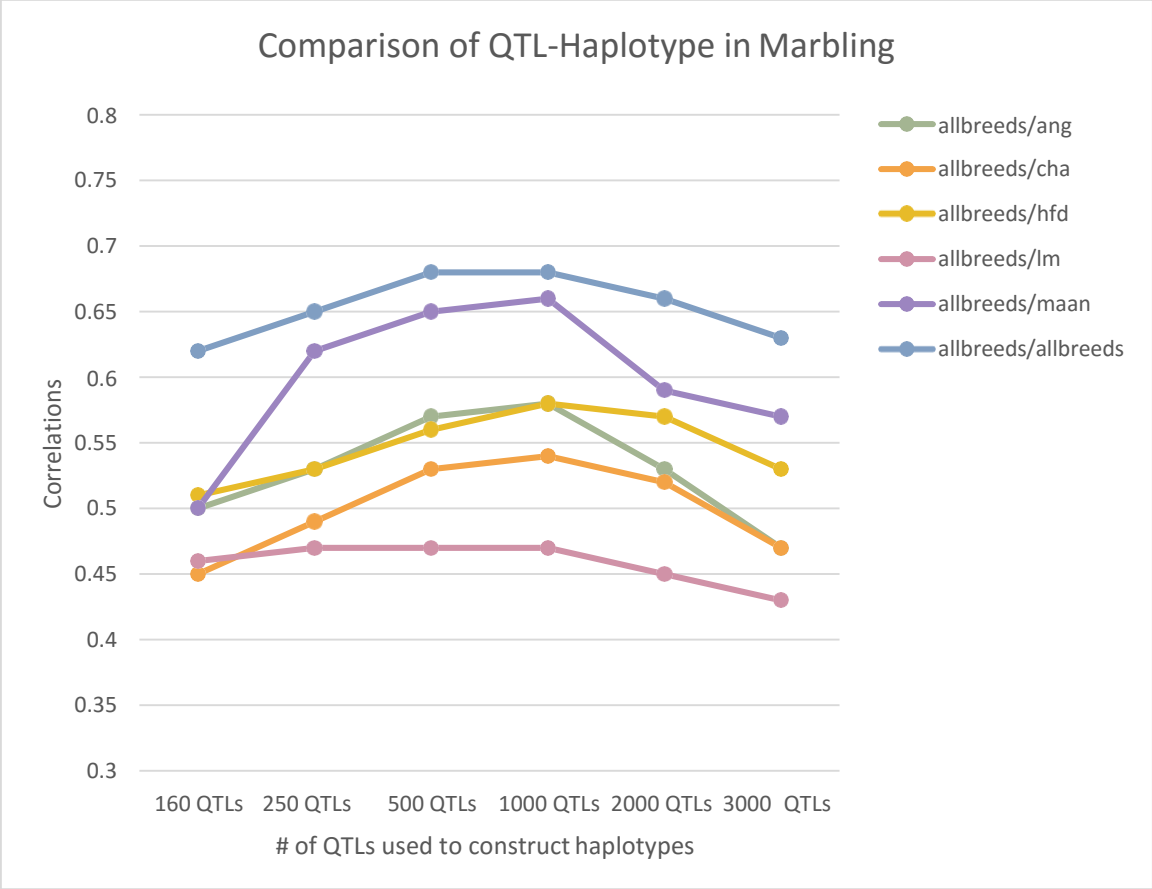

Figure 3.5 Comparison of WBSF correlations obtained from the different QTL-haplotypes predictions. The legend states the training breed/validation breed.

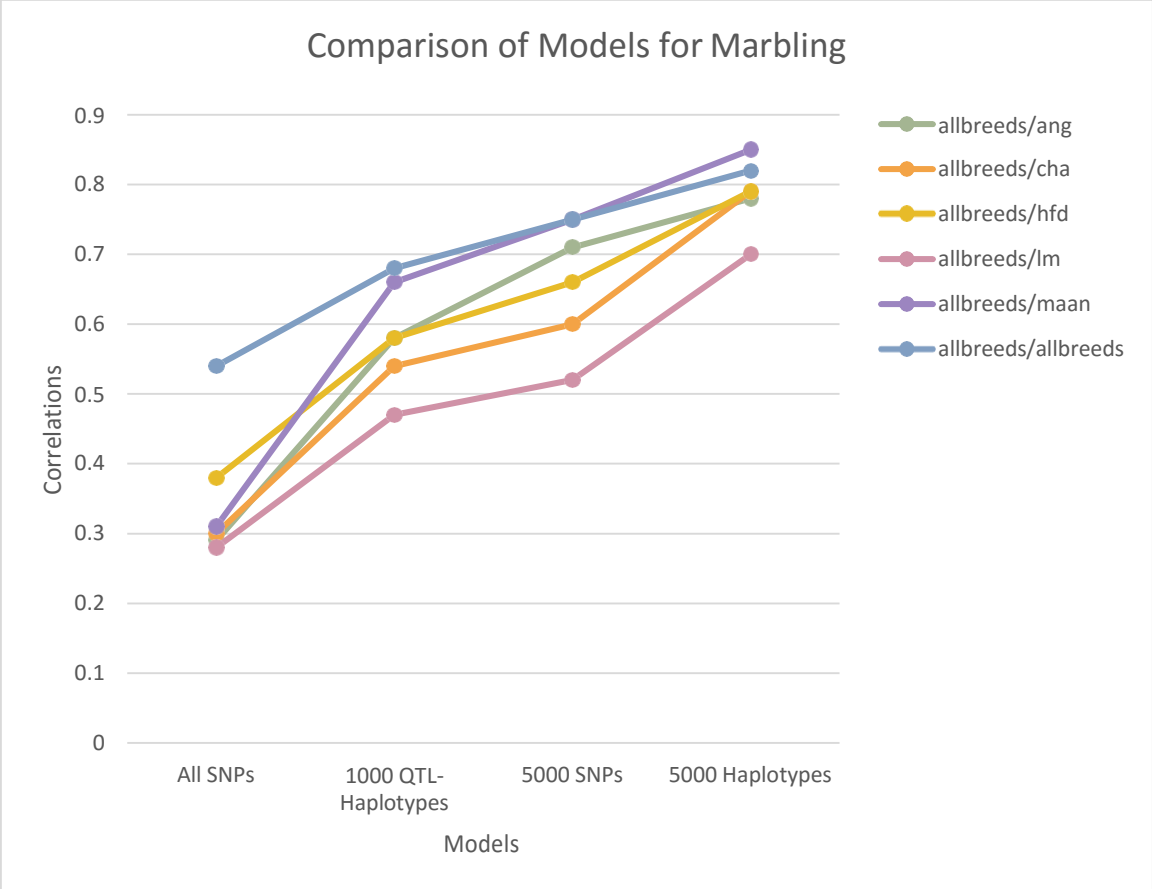

Figure 3.6 Comparison of MB correlations obtained from the different models. These models were the basic all SNP analyses, haplotype-based predictions, and predicting phenotypes using feature selection. The legend states the training breed/validation breed.

Figure 3.7 Comparison of HCW correlations obtained from the different models. These models were the basic all SNP analyses, haplotype-based predictions, and predicting phenotypes using feature selection. The legend states the training breed/validation breed.

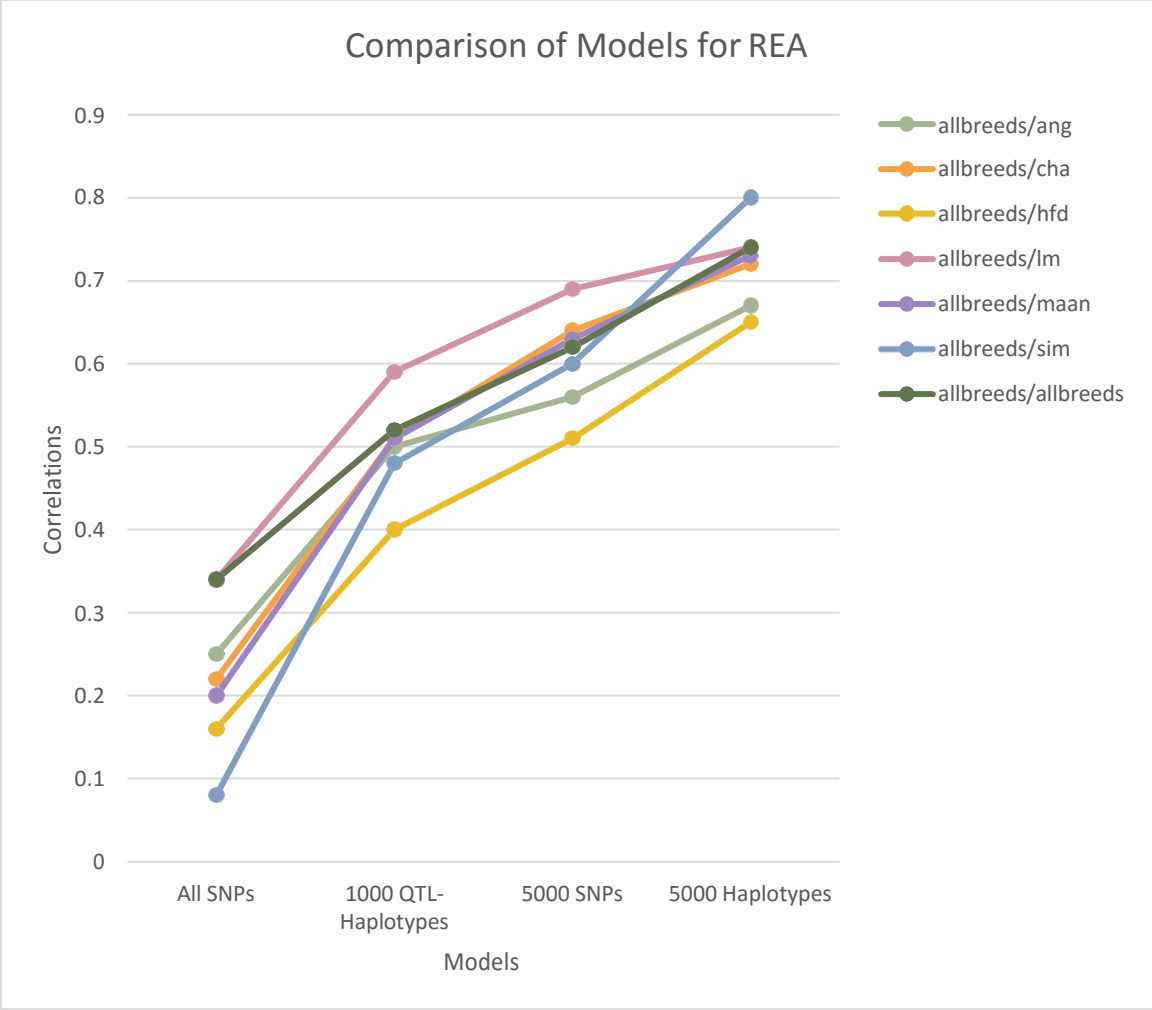

Figure 3.8 Comparison of REA correlations obtained from the different models. These models were the basic all SNP analyses, haplotype-based predictions, and predicting phenotypes using feature selection. The legend states the training breed/validation breed.

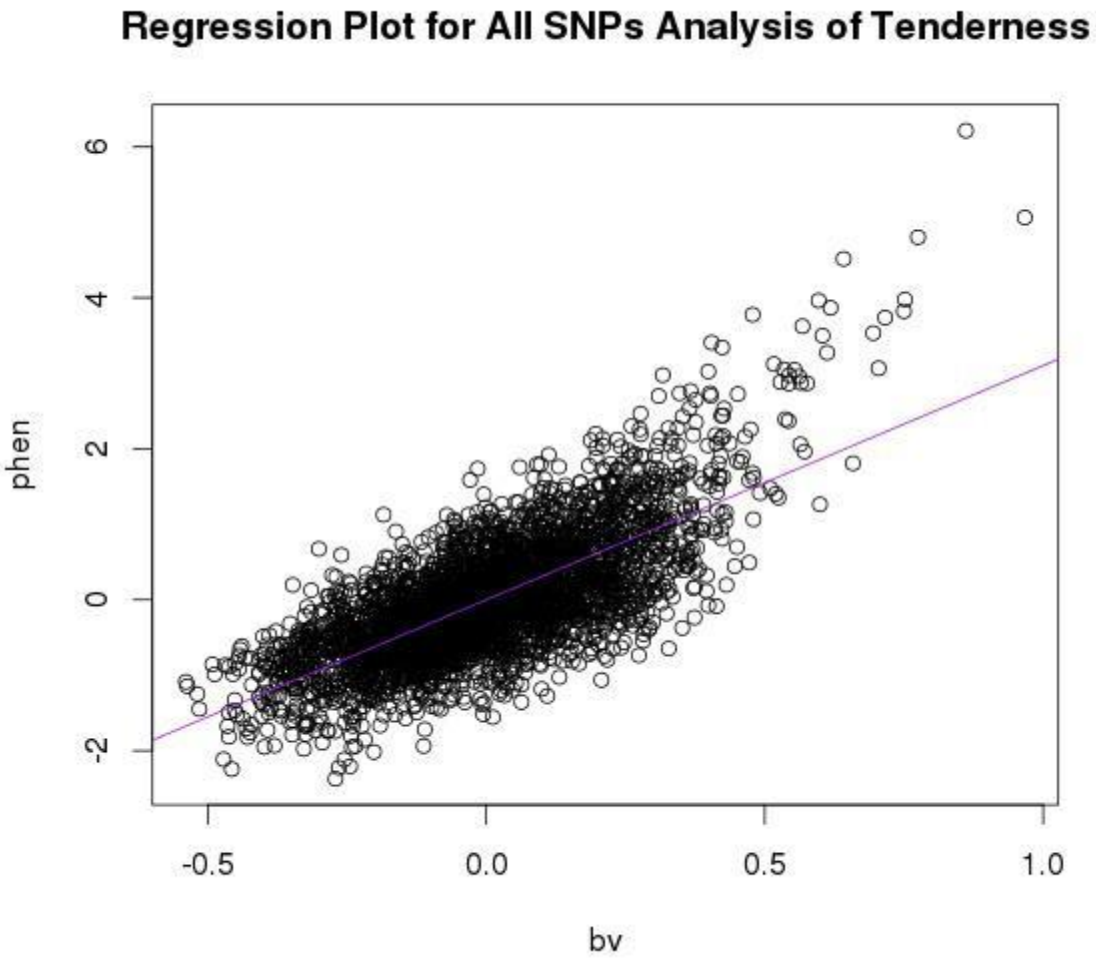

Figure 3.9 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting all of the SNPs for WBSF.

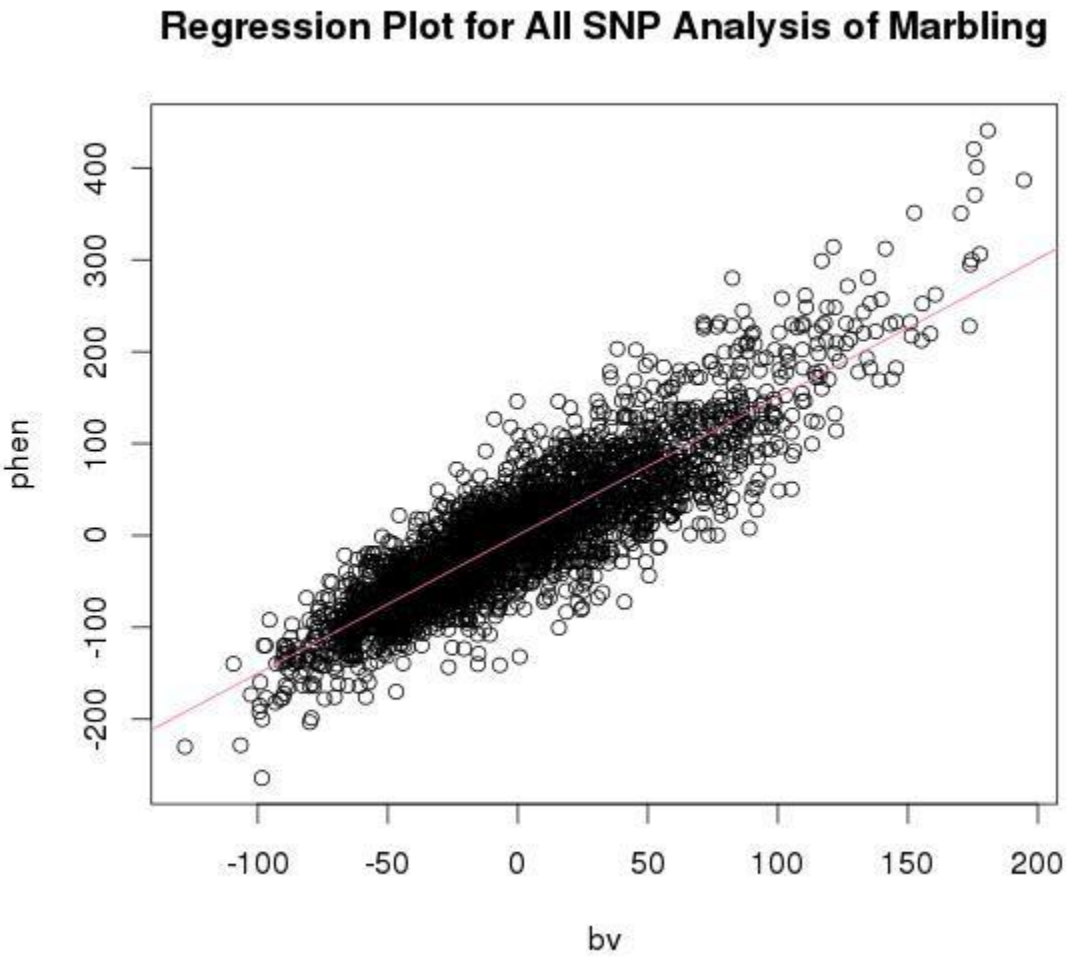

Figure 3.10 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting all of the SNPs for MB.

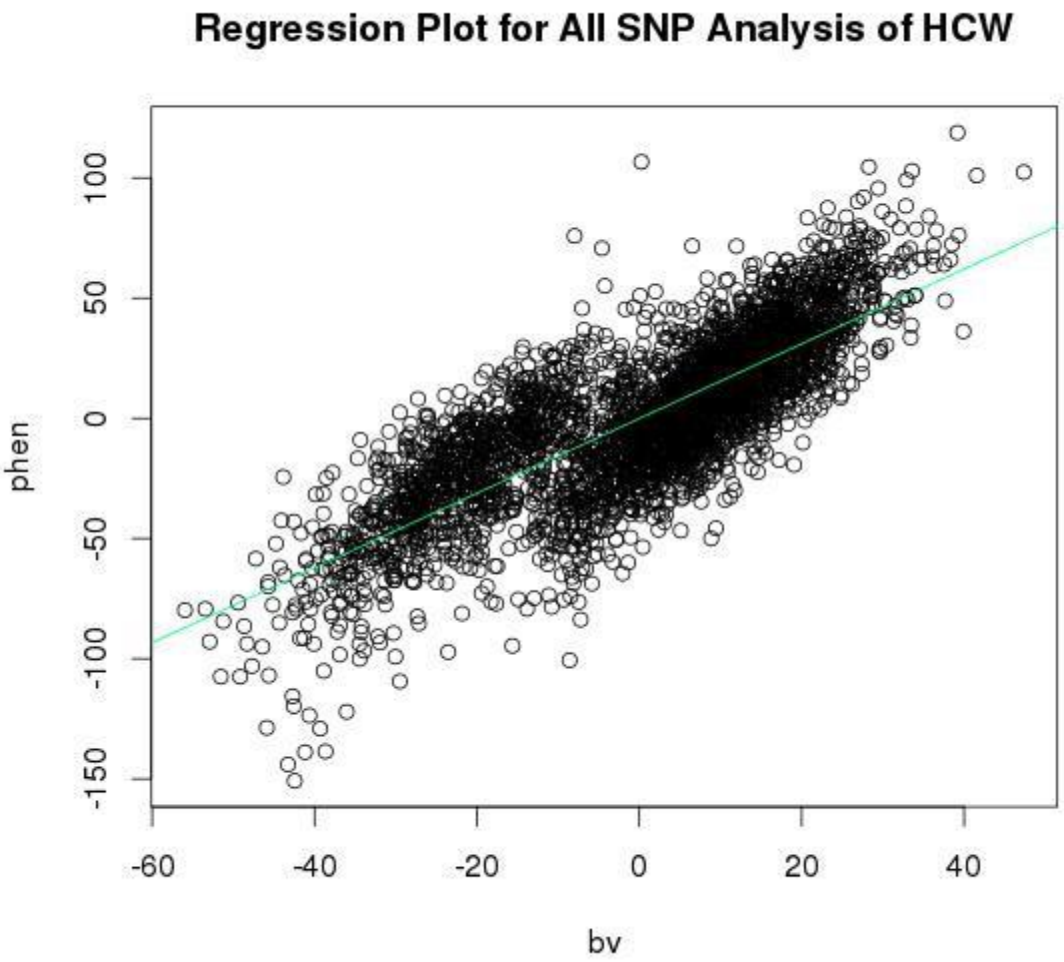

Figure 3.11 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting all of the SNPs for HCW.

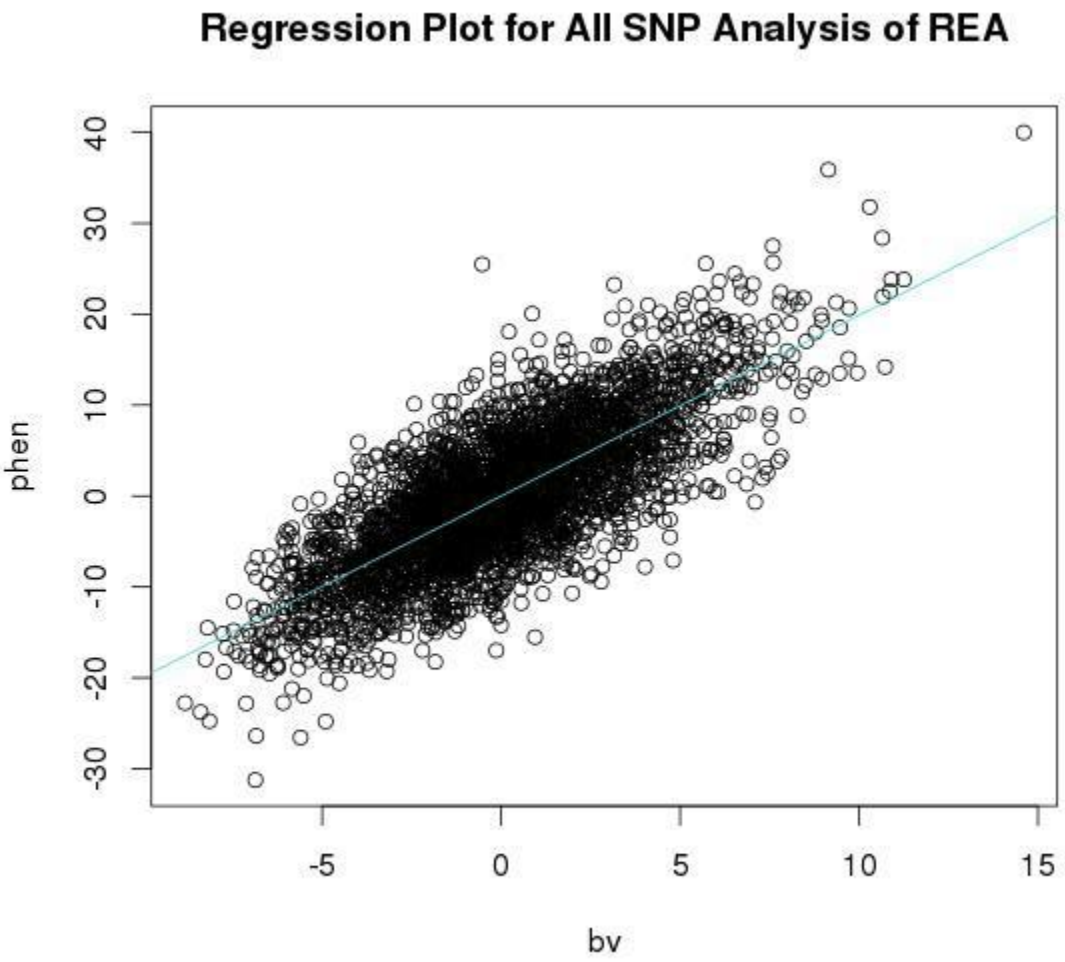

Figure 3.12 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting all of the SNPs for REA.

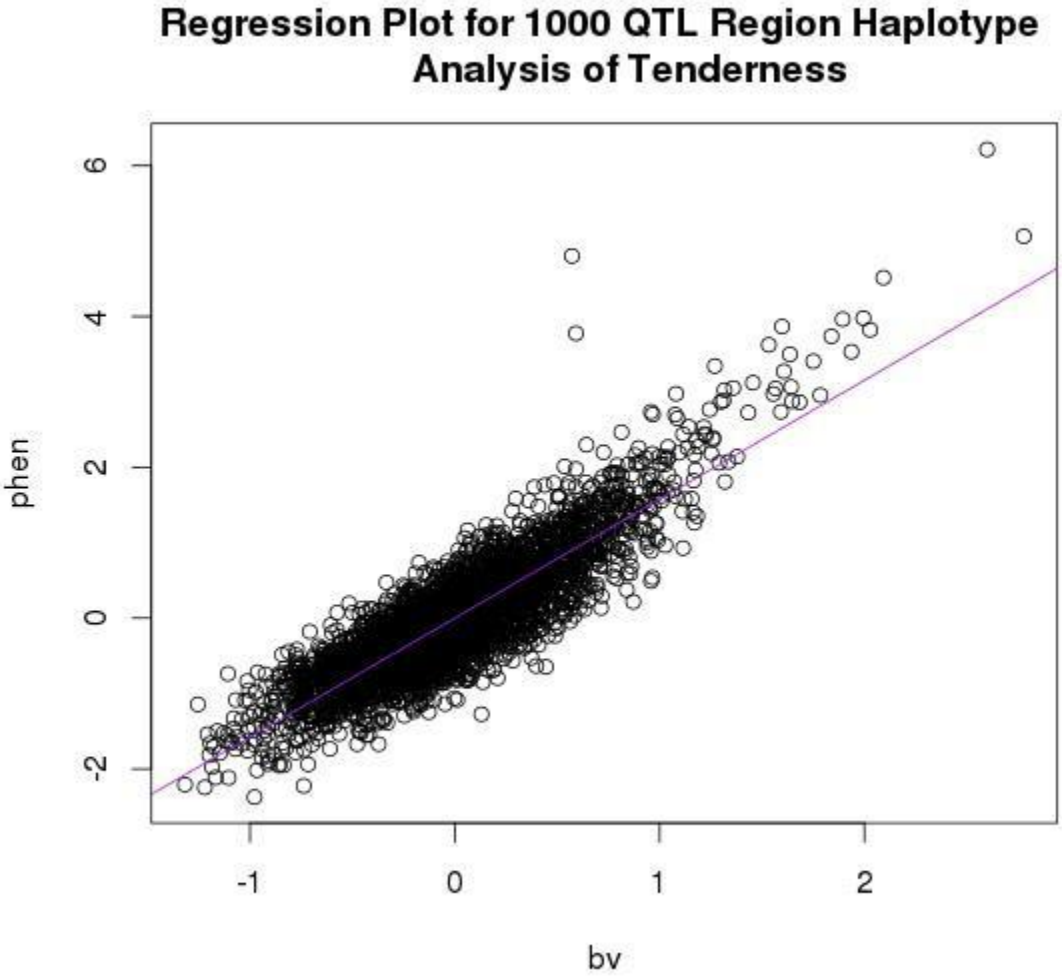

Figure 3.13 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting haplotypes selected from 1,000 QTL regions for WBSF.

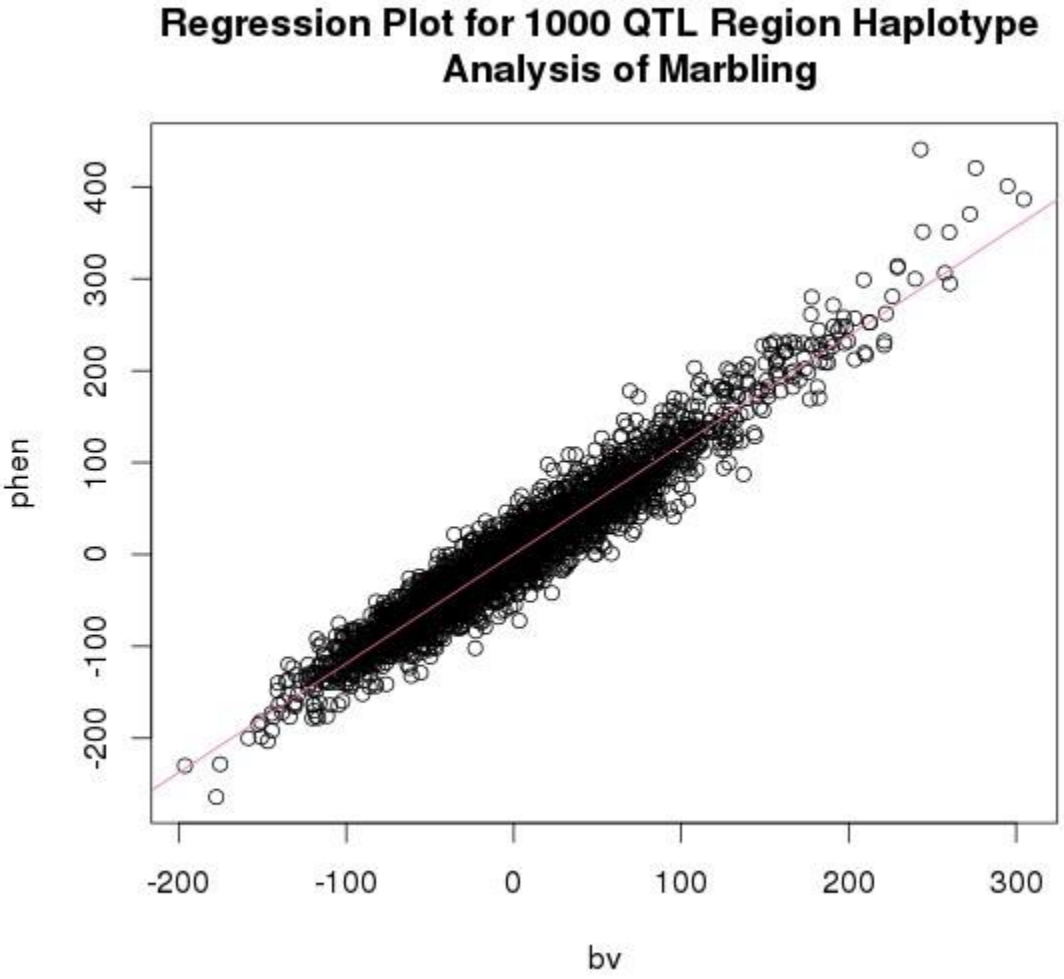

Figure 3.14 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting haplotypes selected from 1,000 QTL regions for MB.

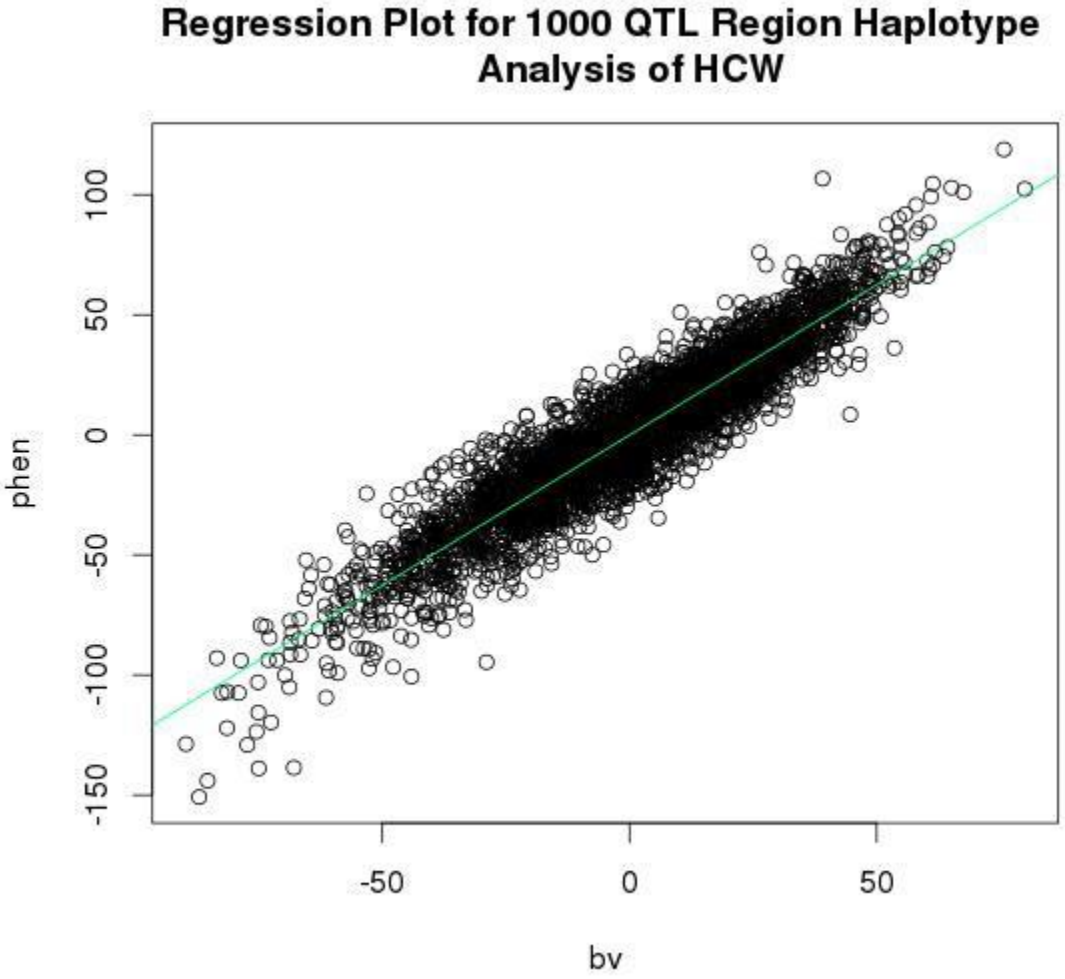

Figure 3.15 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting haplotypes selected from 1,000 QTL regions for HCW.

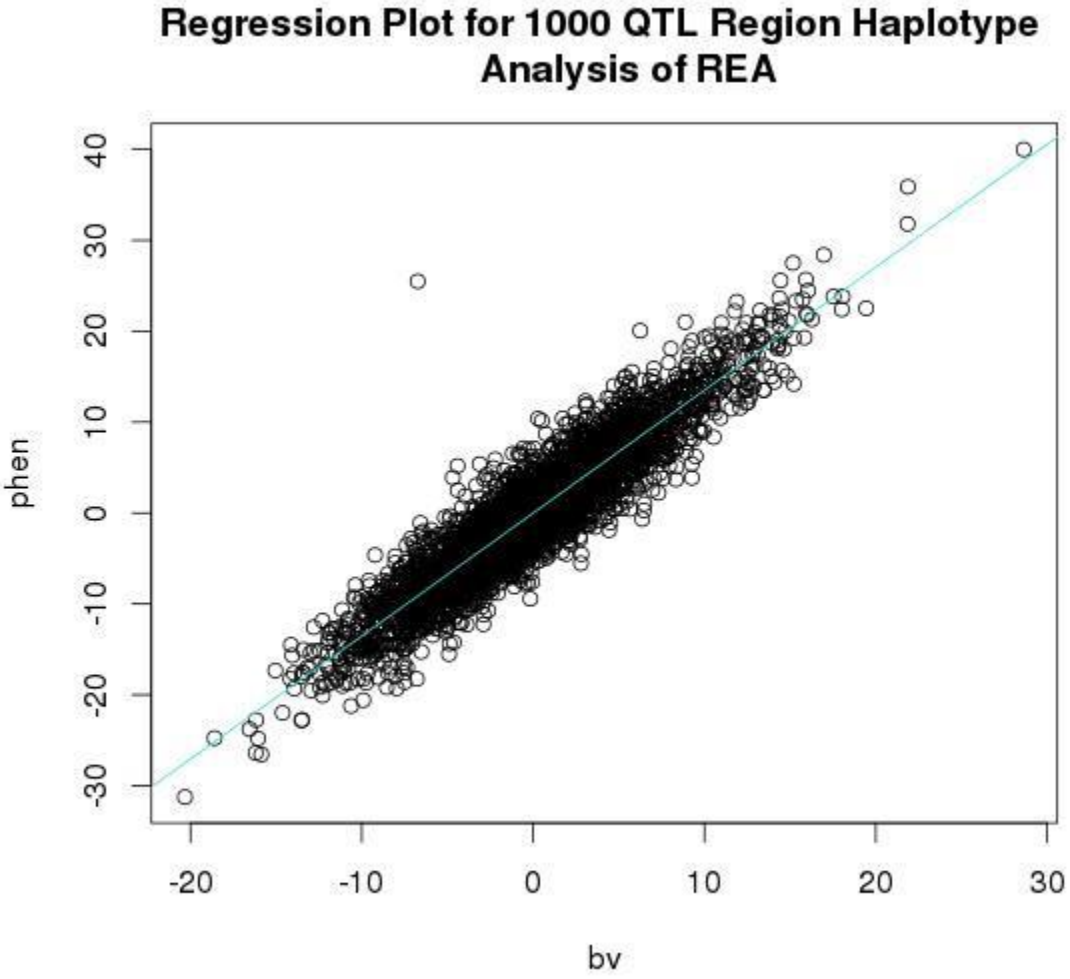

Figure 3.16 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting haplotypes selected from 1,000 QTL regions for REA.

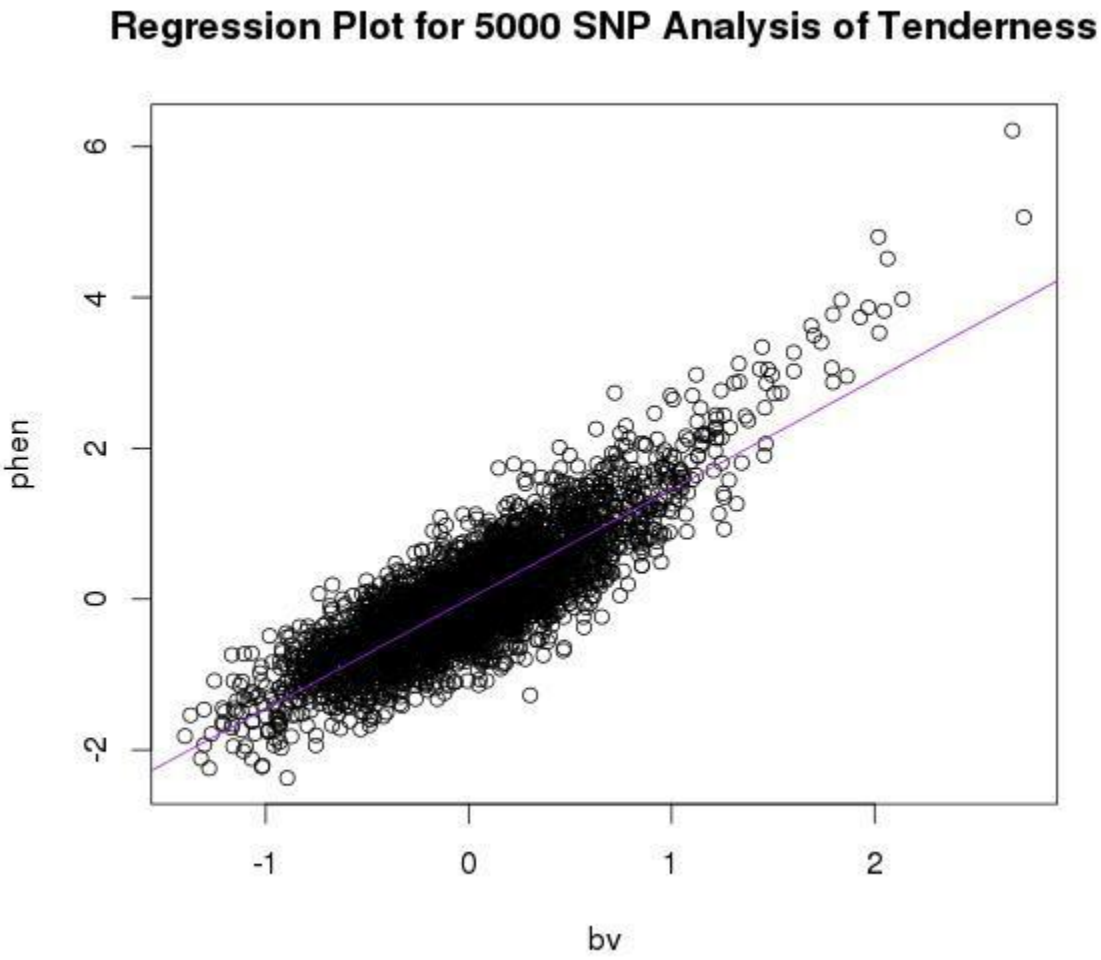

Figure 3.17 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the 5,000 SNPs that were used for the selection of haplotypes for WBSF.

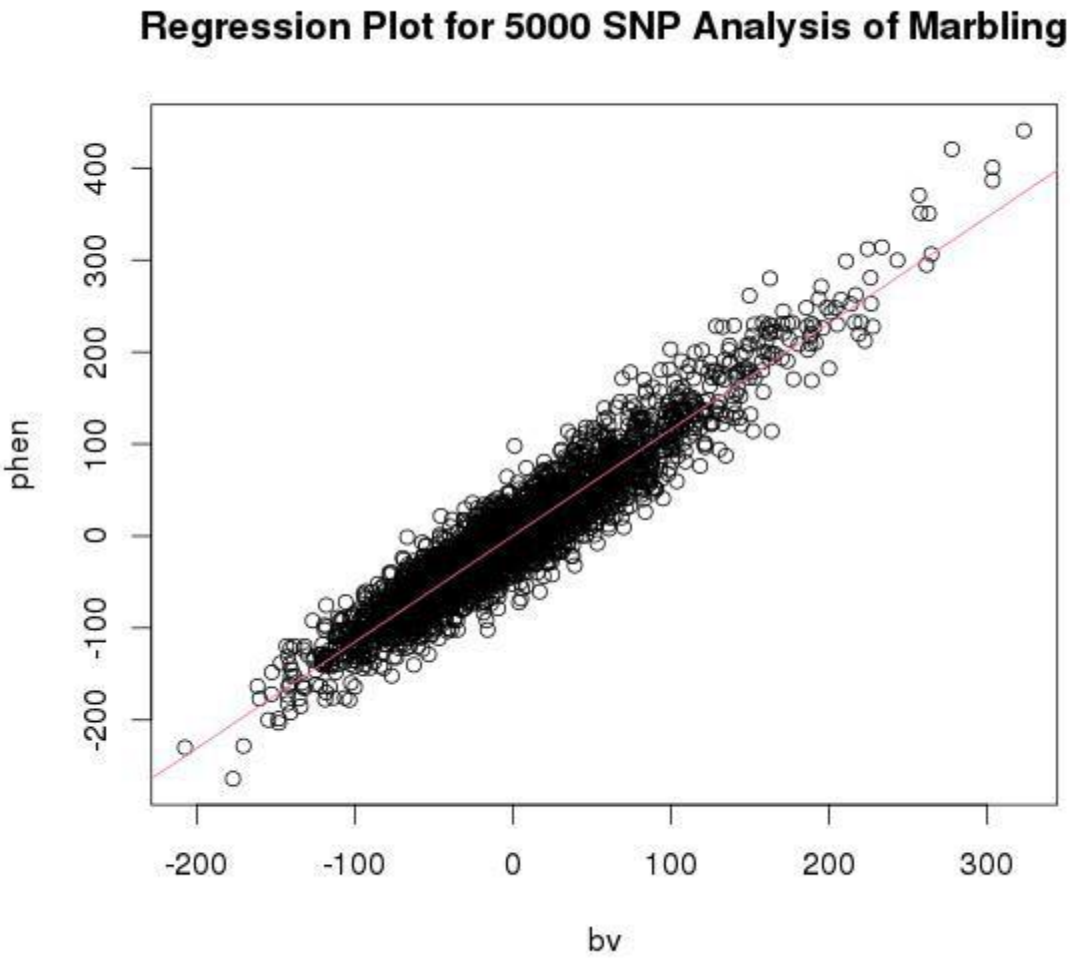

Figure 3.18 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the 5,000 SNPs that were used for the selection of haplotypes for MB.

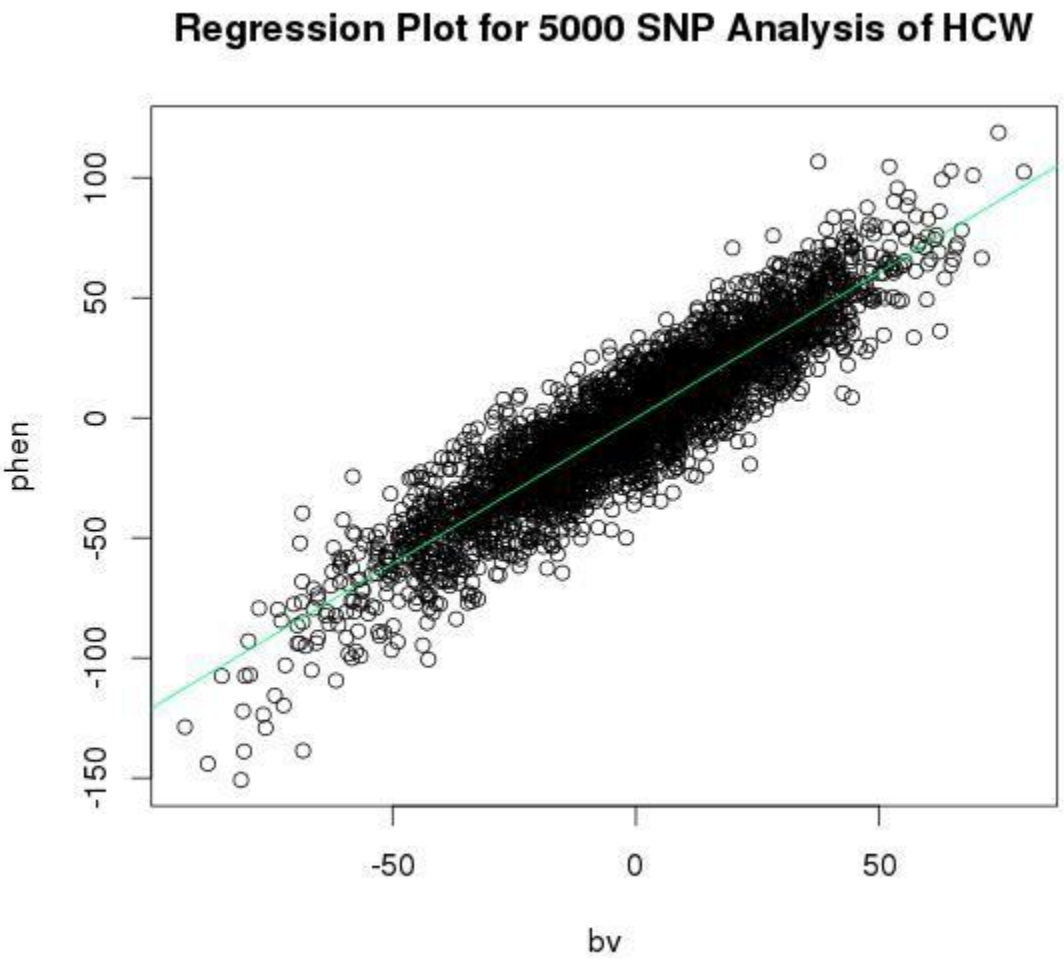

Figure 3.19 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the 5,000 SNPs that were used for the selection of haplotypes for HCW.

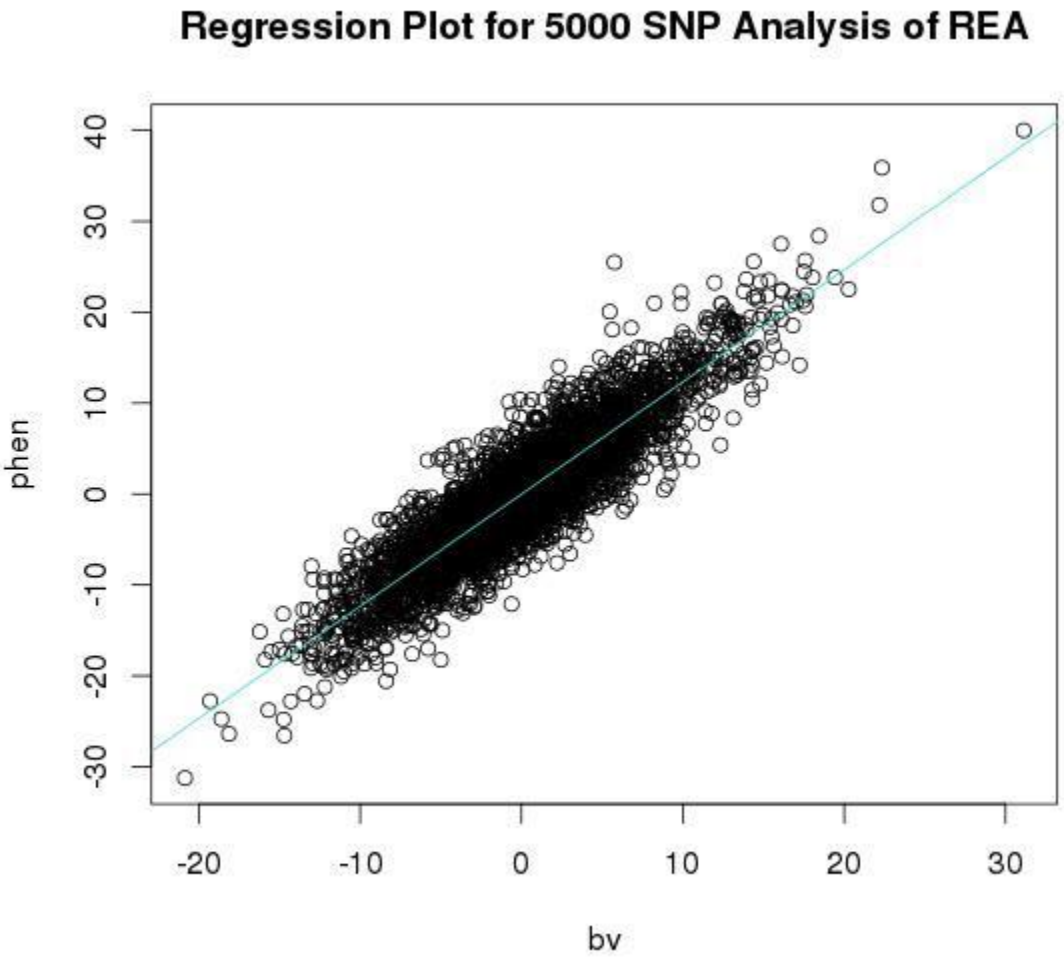

Figure 3.20 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the 5,000 SNPs that were used for the selection of haplotypes for REA.

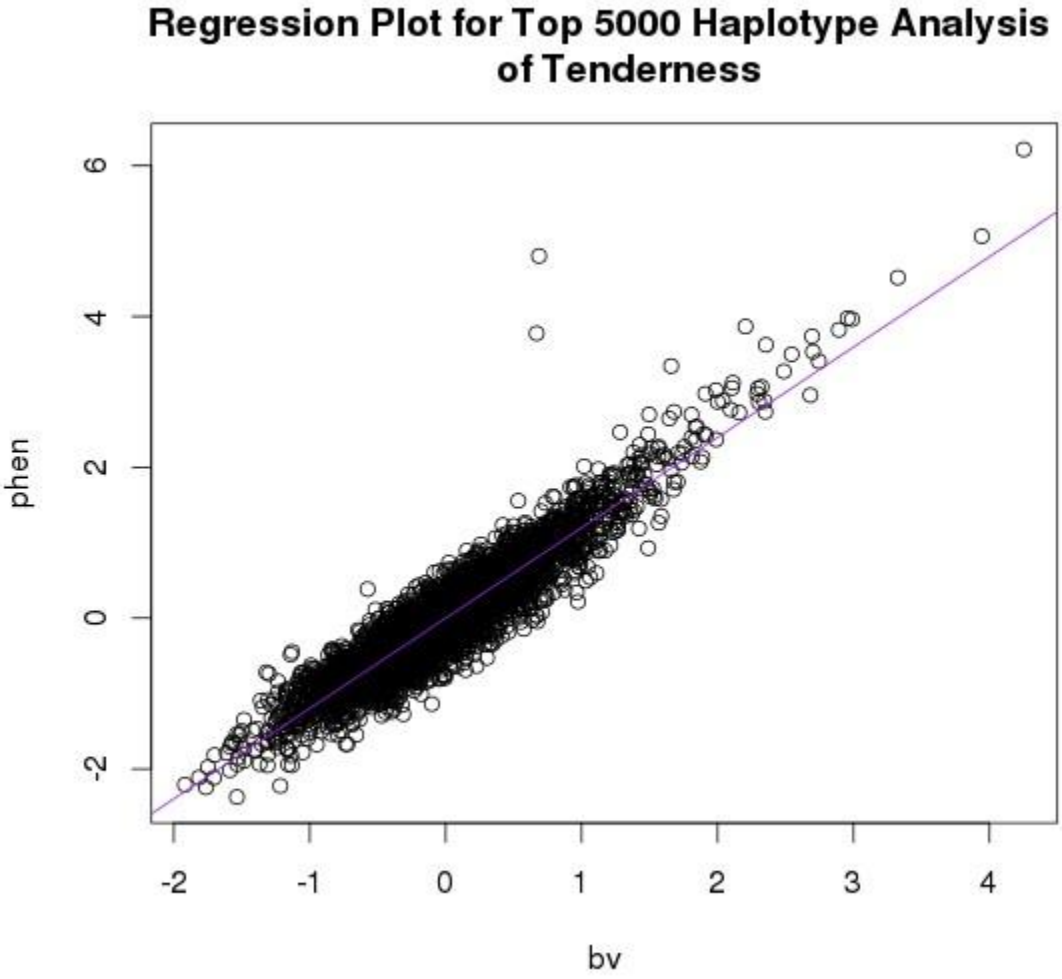

Figure 3.21 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the top 5,000 haplotypes with the largest effect from the 1,000 QTL region haplotype analysis for WBSF.

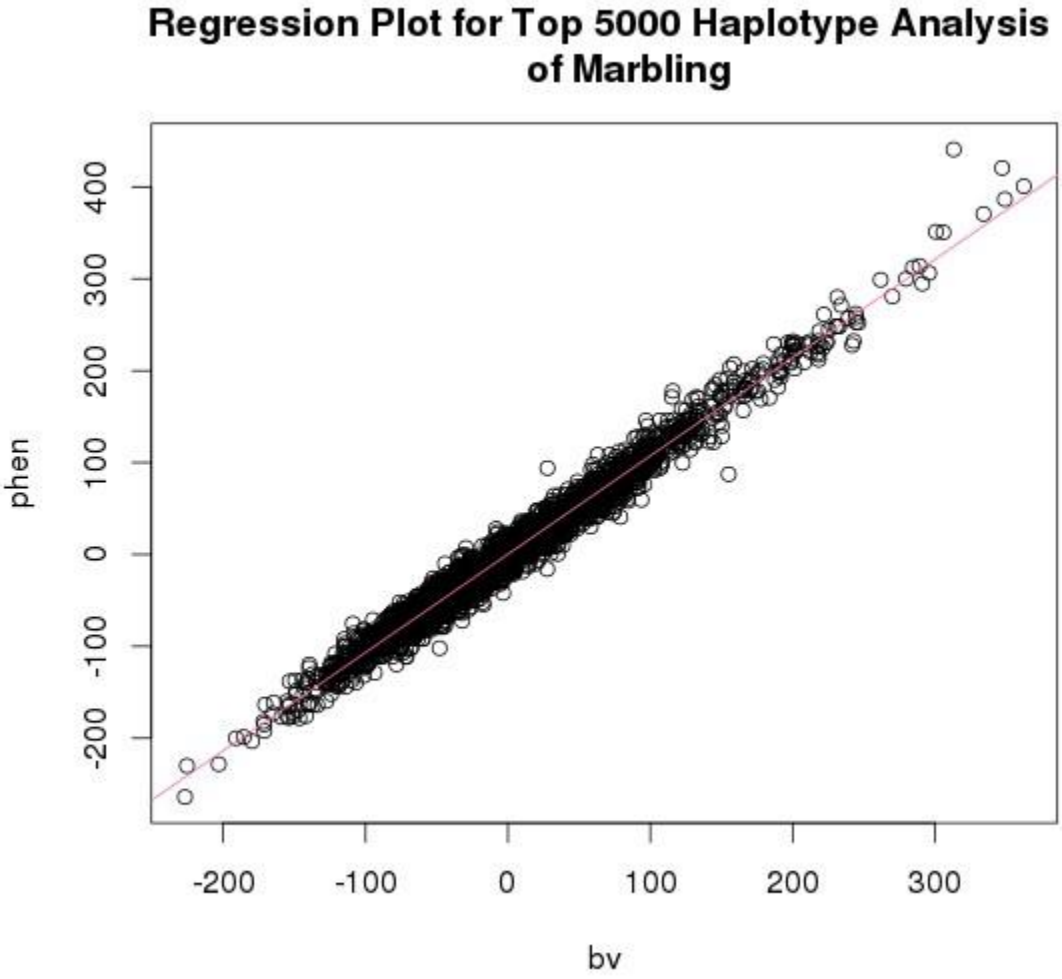

Figure 3.22 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the top 5,000 haplotypes with the largest effect from the 1,000 QTL region haplotype analysis for MB.

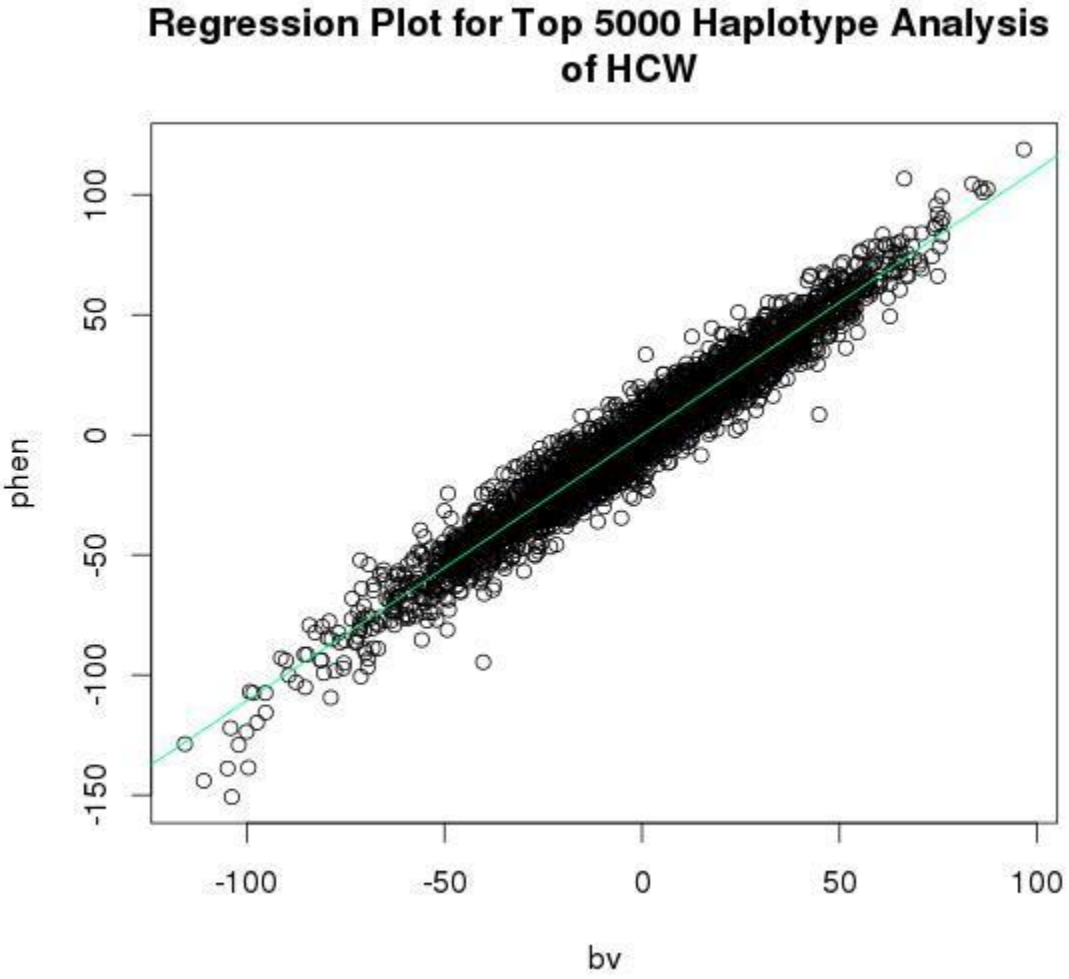

Figure 3.23 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the top 5,000 haplotypes with the largest effect from the 1,000 QTL region haplotype analysis for HCW.

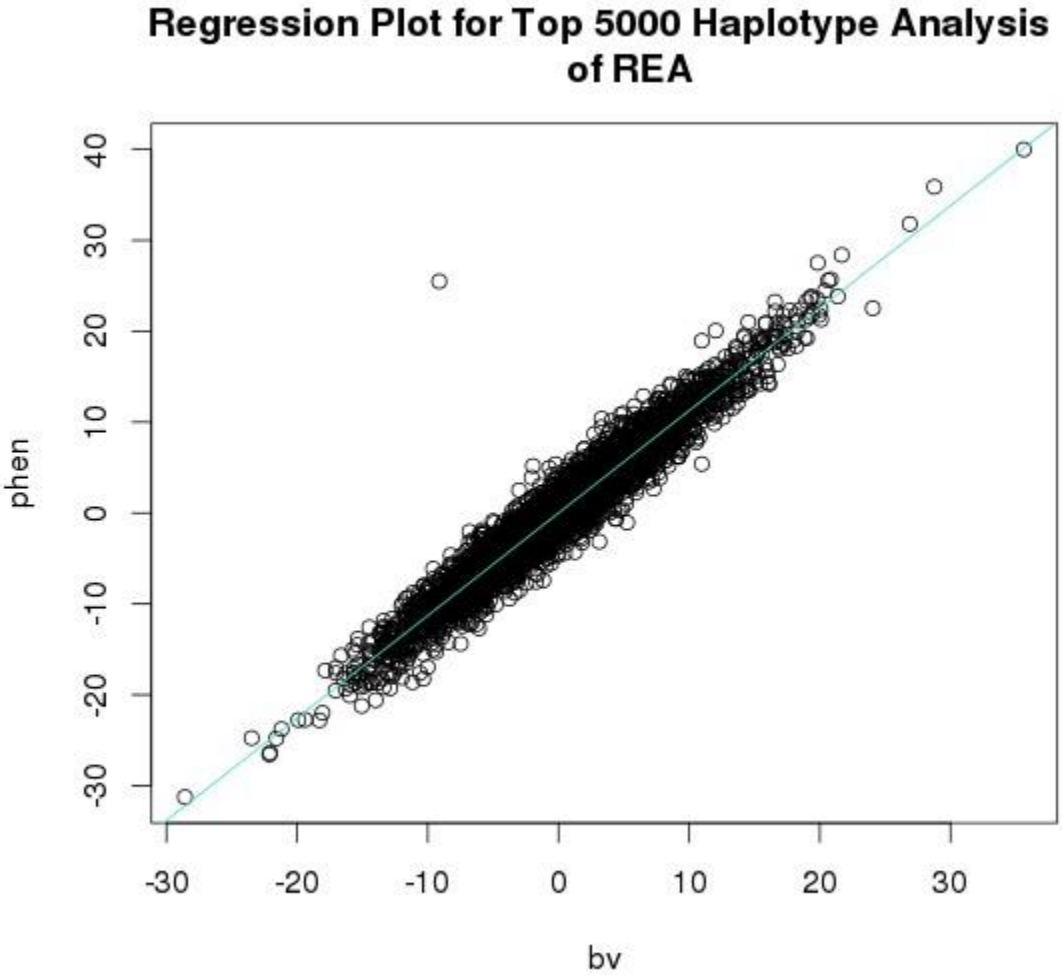

Figure 3.24 Regression Plot of adjusted phenotypes and the predicted breeding values from the analyses fitting the top 5,000 haplotypes with the largest effect from the 1,000 QTL region haplotype analysis for REA.

APPENDIX

GWAS Manhattan Plots

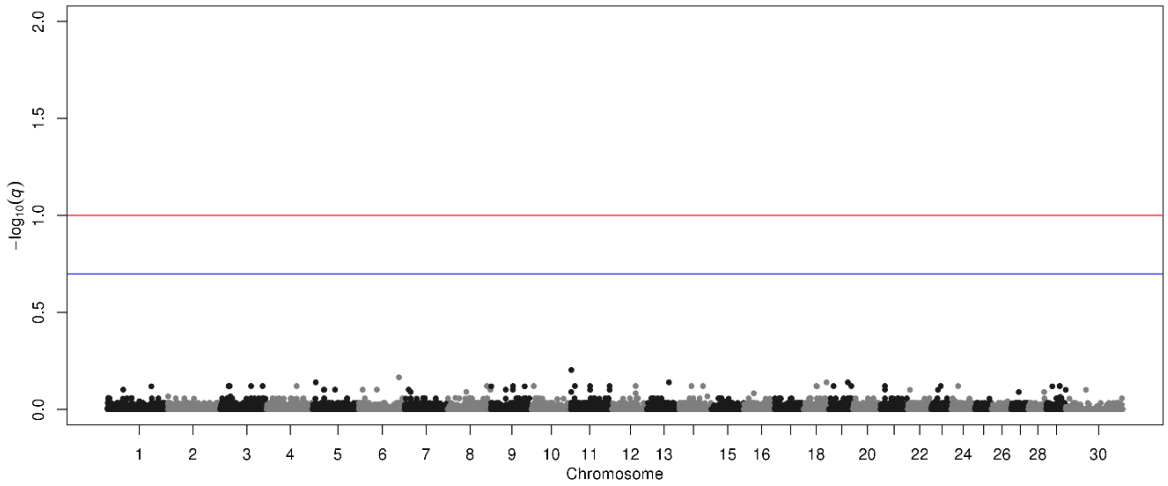

Figure a.1 Manhattan plot of SNP q-values estimated in the univariate analysis of MB in the multiple-breed population.

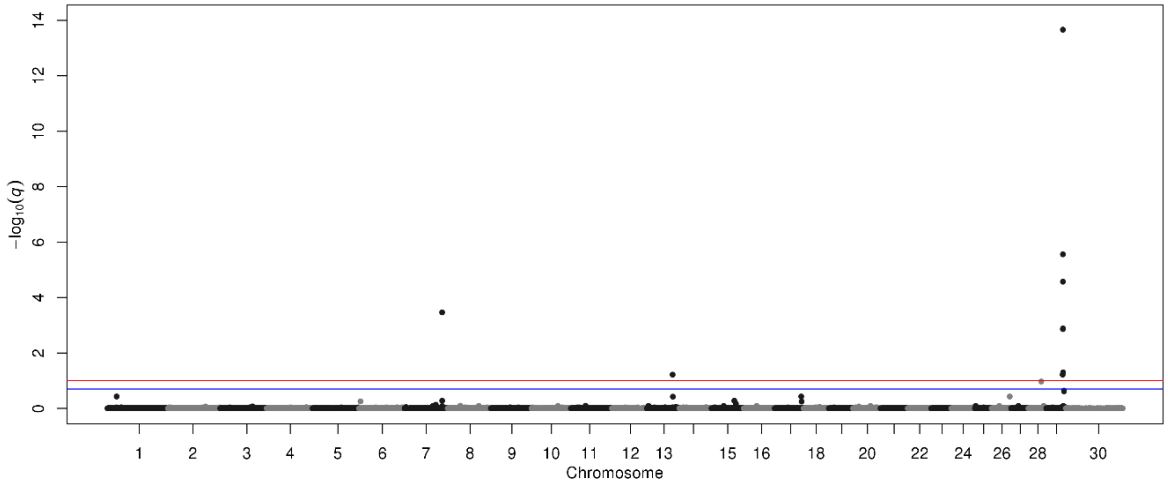

Figure a.2 Manhattan plot of SNP q-values estimated in the univariate analysis of WBSF in the multiple-breed population.

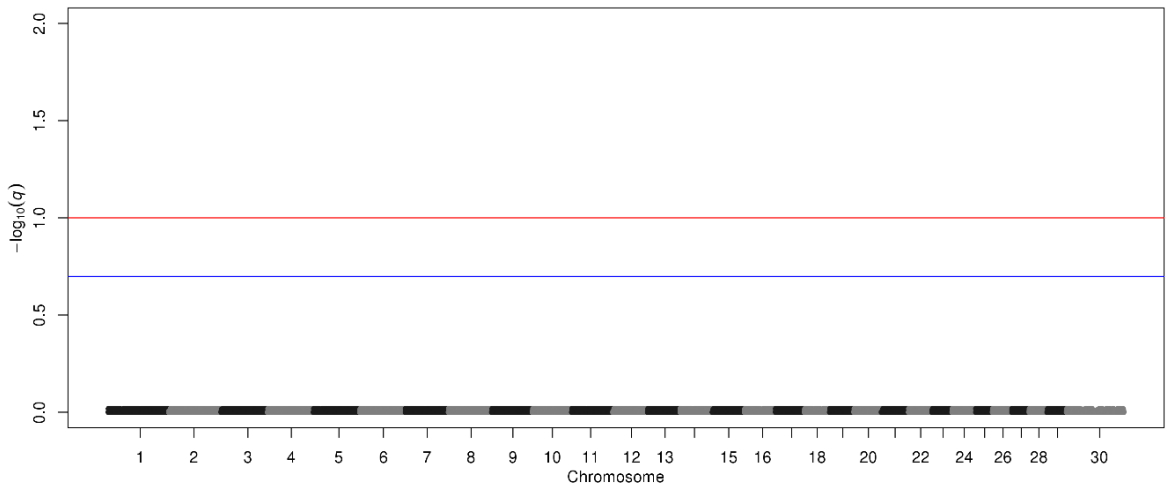

Figure a.3 Manhattan plot of SNP q-values estimated in the univariate analysis of CL in the multiple-breed population.

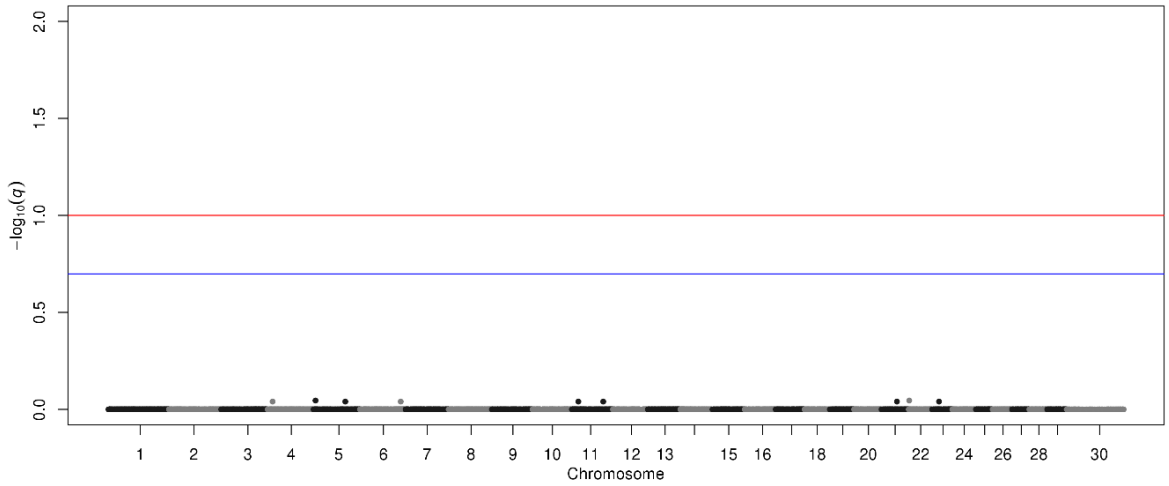

Figure a.4 Manhattan plot of SNP q-values estimated in the univariate analysis of KPH in the multiple-breed population.

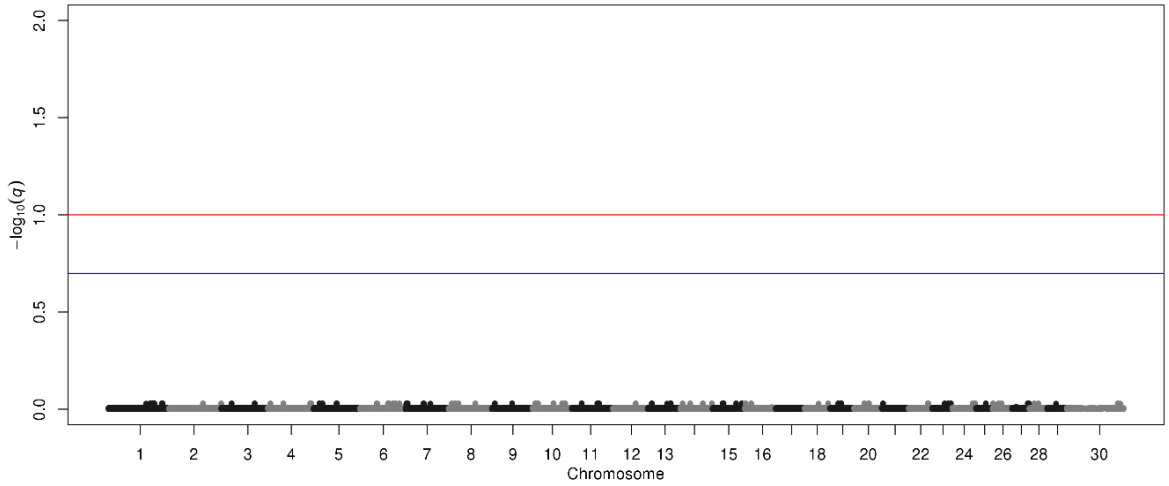

Figure a.5 Manhattan plot of SNP q-values estimated in the univariate analysis of IF in the multiple-breed population.

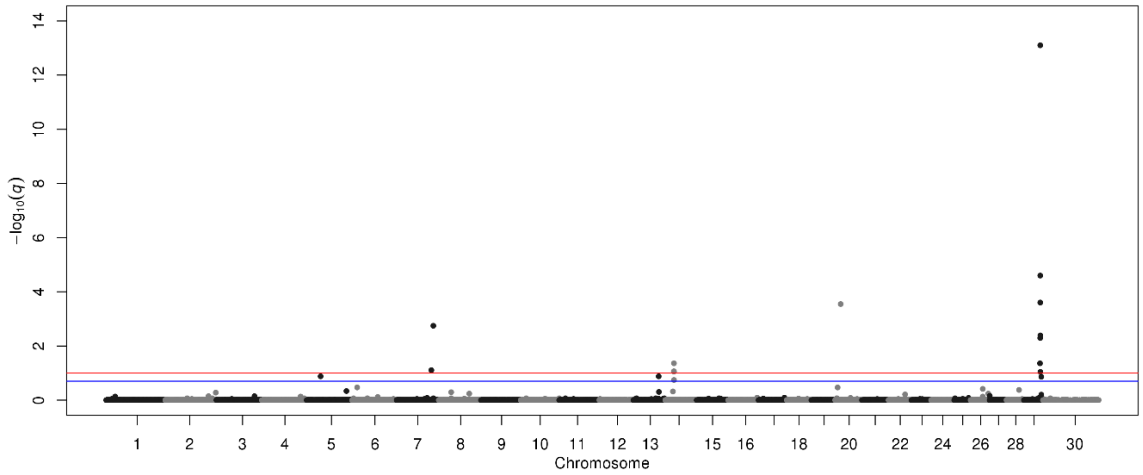

Figure a.6 Manhattan plot of SNP q-values estimated in the multivariate analysis of WBSF and HCW in the multiple-breed population.

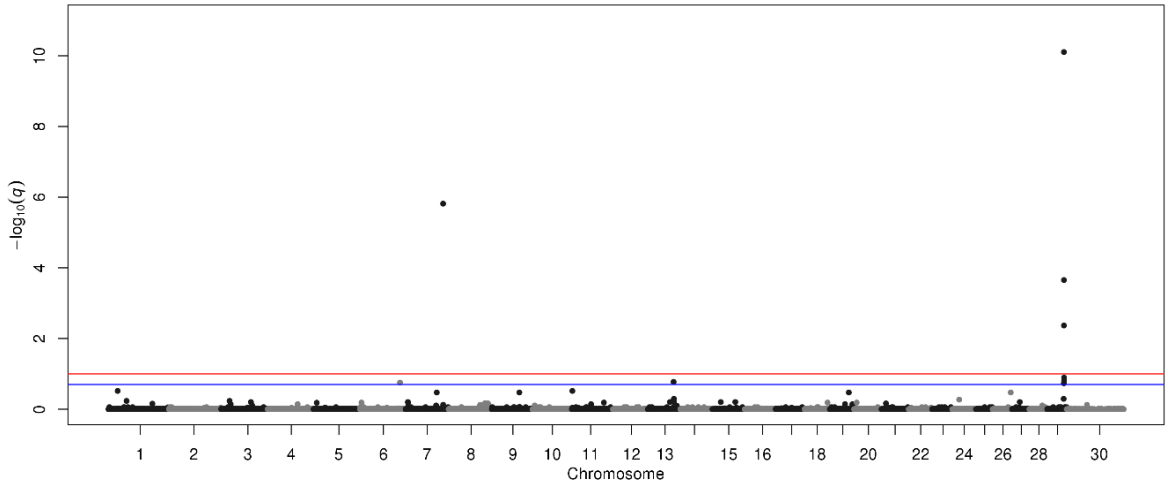

Figure a.7 Manhattan plot of SNP q-values estimated in the multivariate analysis of WBSF and MB in the multiple-breed population.

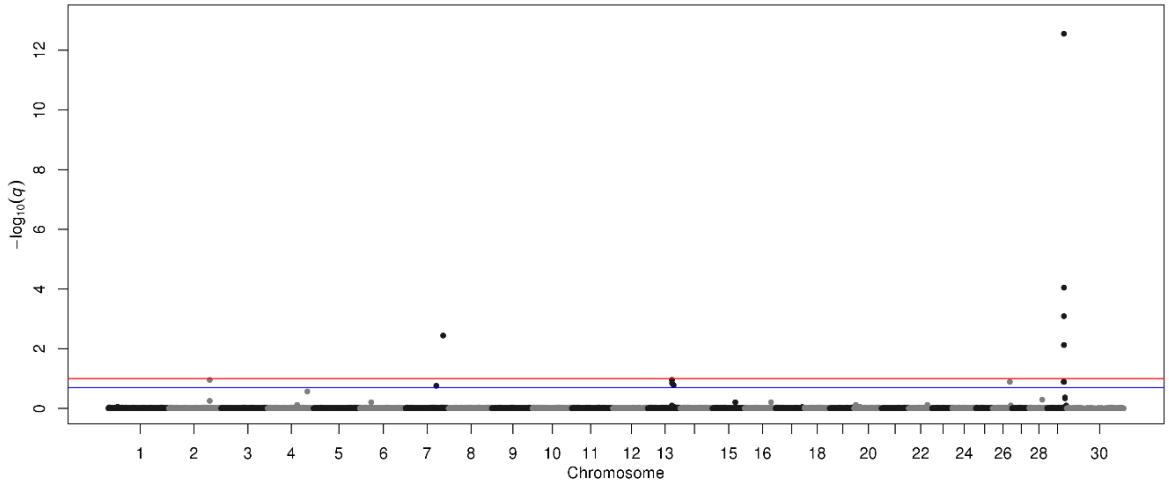

Figure a.8 Manhattan plot of SNP q-values estimated in the multivariate analysis of WBSF and FT in the multiple-breed population.

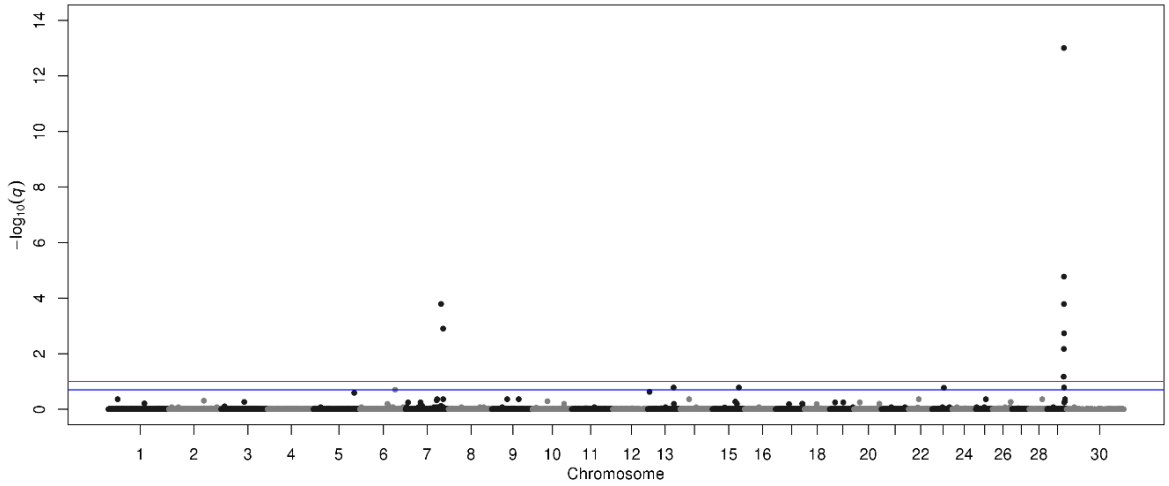

Figure a.9 Manhattan plot of SNP q-values estimated in the multivariate analysis of WBSF and REA in the multiple-breed population.

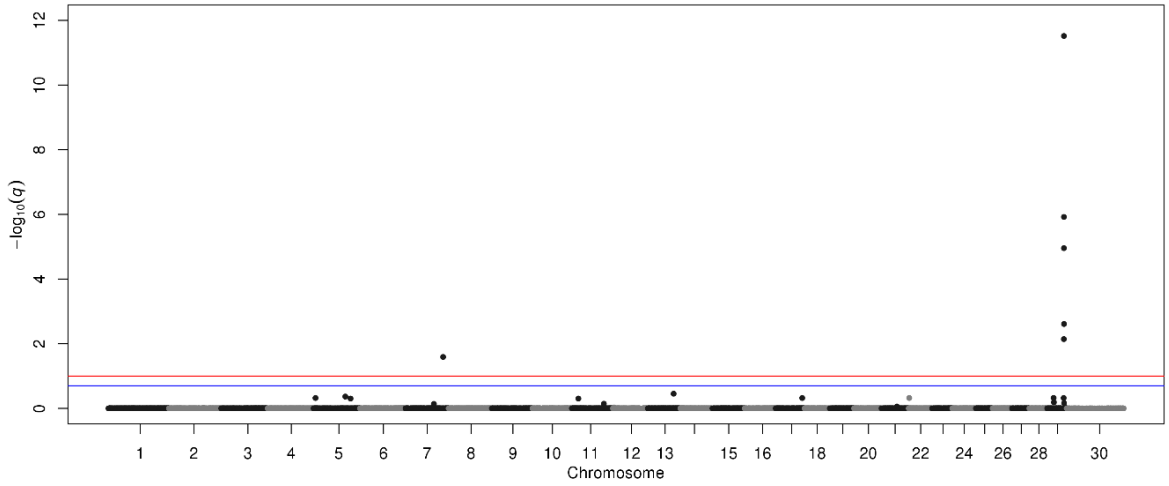

Figure a.10 Manhattan plot of SNP q-values estimated in the multivariate analysis of WBSF and KPH in the multiple-breed population.

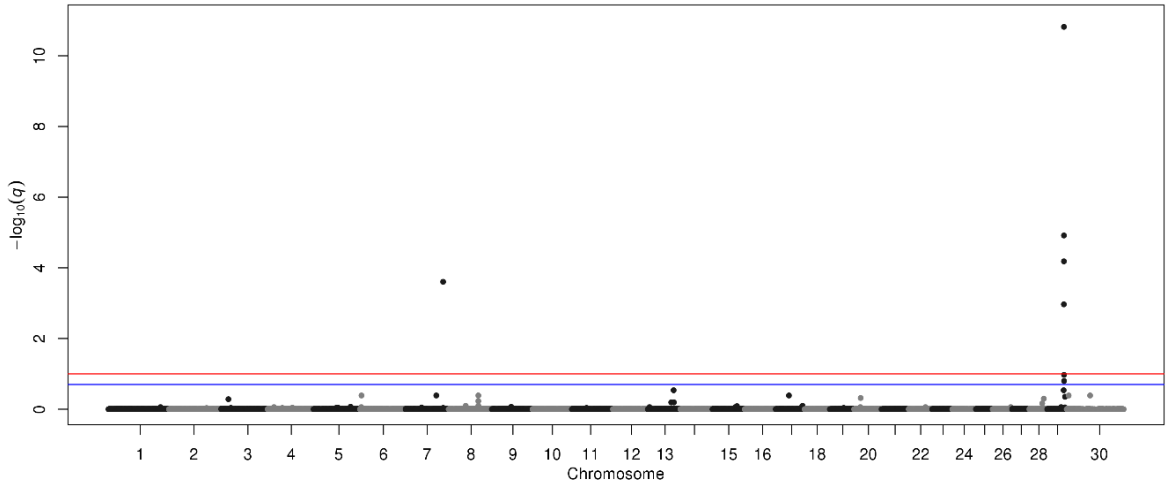

Figure a.11 Manhattan plot of SNP q-values estimated in the multivariate analysis of WBSF and CL in the multiple-breed population.

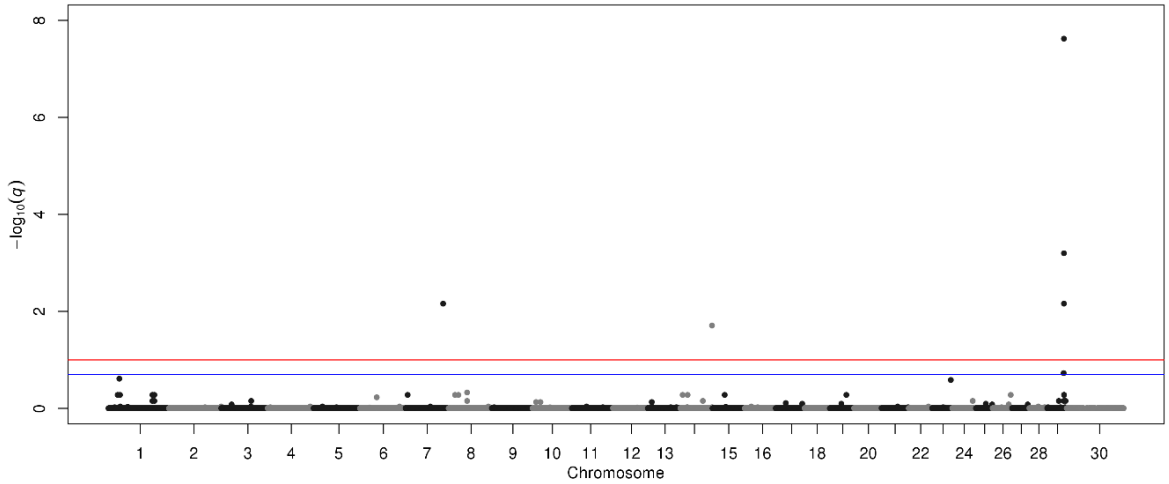

Figure a.12 Manhattan plot of SNP q-values estimated in the multivariate analysis of WBSF and IF in the multiple-breed population.

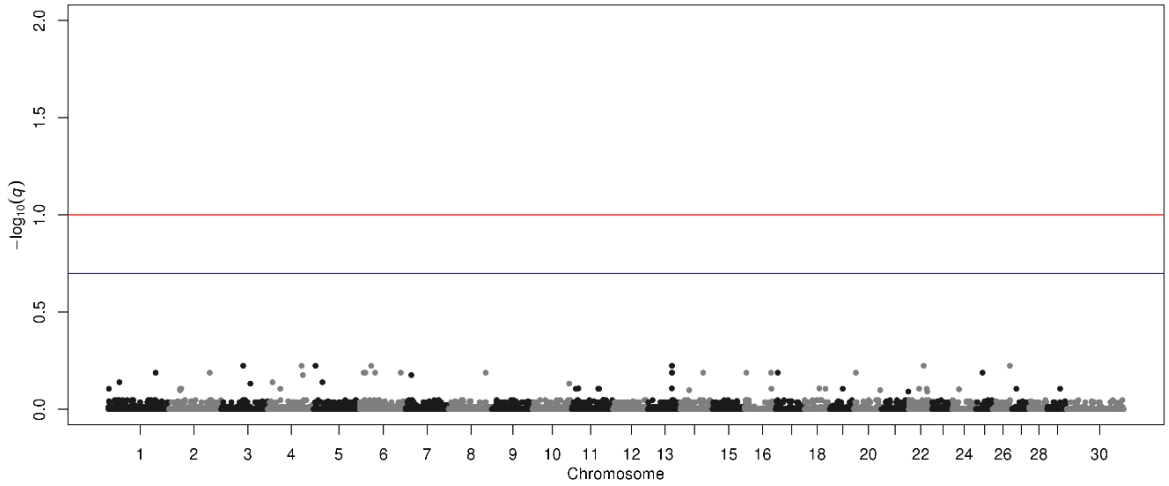

Figure a.13 Manhattan plot of SNP q-values estimated in the multivariate analysis of FT and KPH in the multiple-breed population.

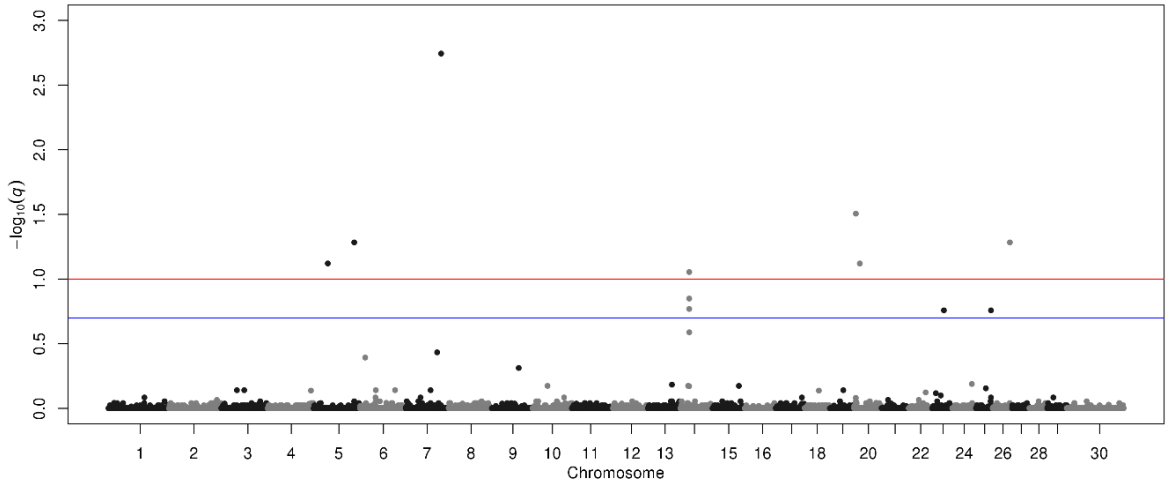

Figure a.14 Manhattan plot of SNP q-values estimated in the multivariate analysis of HCW and REA in the multiple-breed population.

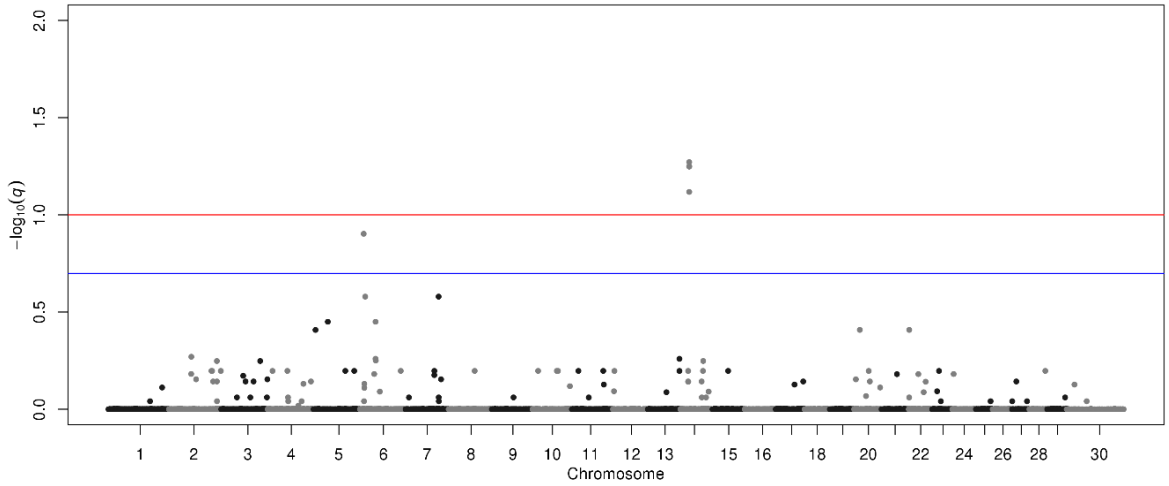

Figure a.15 Manhattan plot of SNP q-values estimated in the multivariate analysis of HCW and KPH in the multiple-breed population.

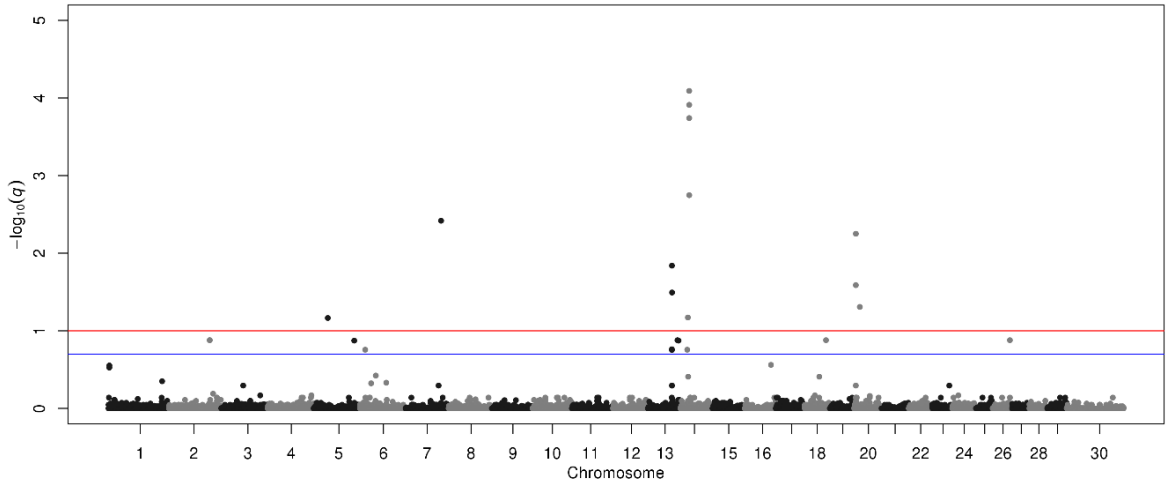

Figure a.16 Manhattan plot of SNP q-values estimated in the multivariate analysis of HCW and FT in the multiple-breed population.

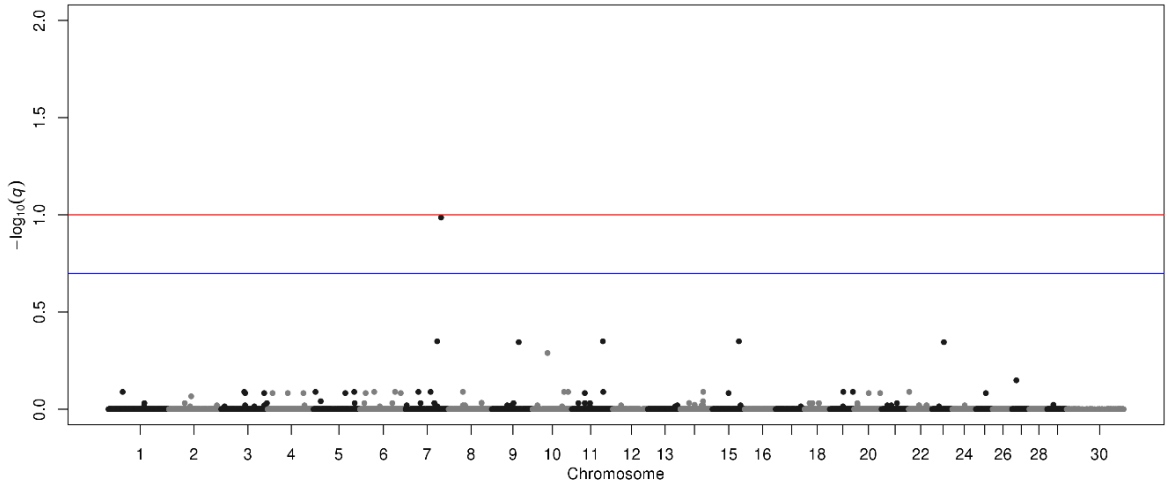

Figure a.17 Manhattan plot of SNP q-values estimated in the multivariate analysis of REA and KPH in the multiple-breed population.

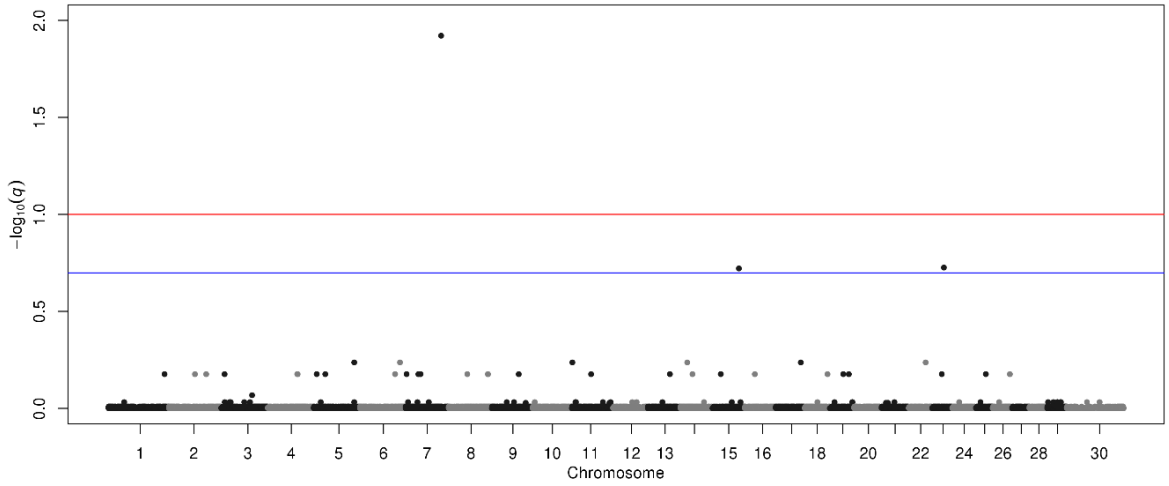

Figure a.18 Manhattan plot of SNP q-values estimated in the multivariate analysis of MB and REA in the multiple-breed population.

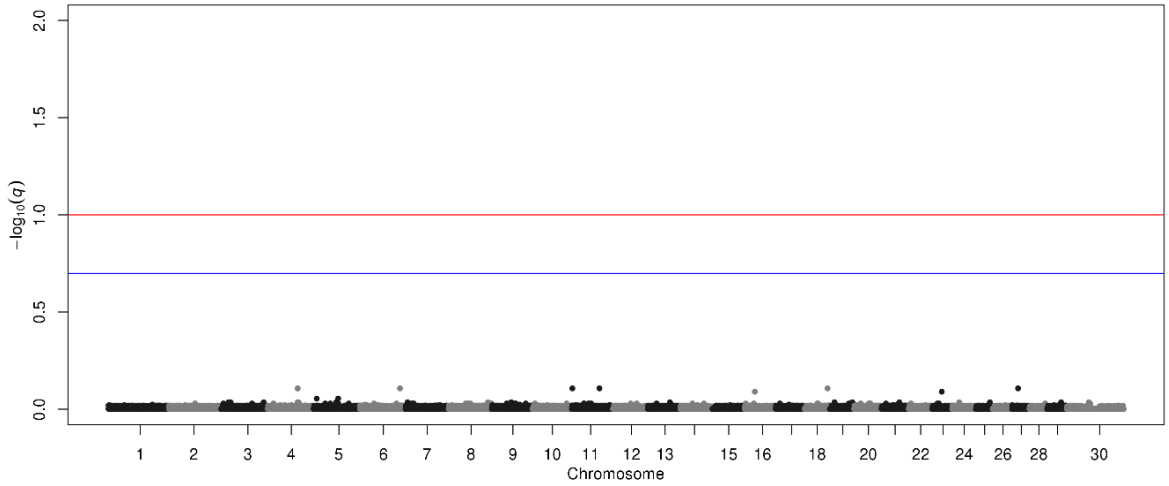

Figure a.19 Manhattan plot of SNP q-values estimated in the multivariate analysis of MB and CL in the multiple-breed population.

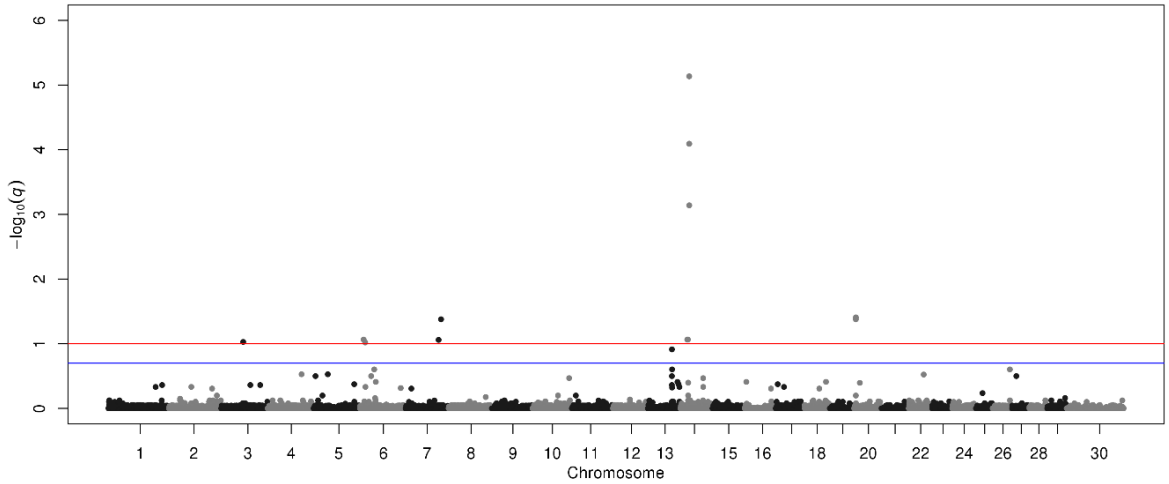

Figure a.20 Manhattan plot of SNP q-values estimated in the multivariate analysis of HCW, FT, and KPH in the multiple-breed population.

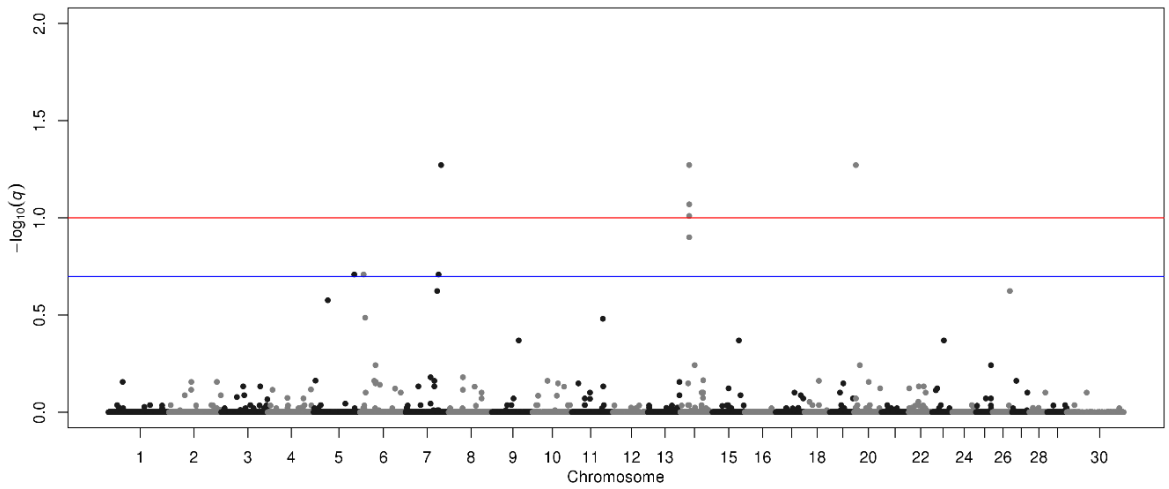

Figure a.21 Manhattan plot of SNP q-values estimated in the multivariate analysis of HCW, REA, and KPH in the multiple-breed population.

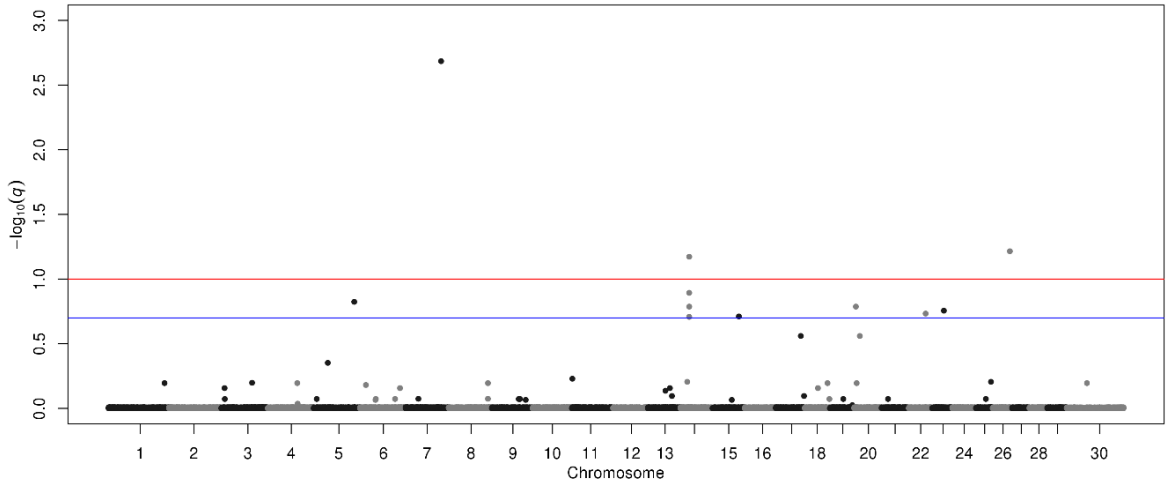

Figure a.22 Manhattan plot of SNP q-values estimated in the multivariate analysis of MB, HCW, and REA in the multiple-breed population.

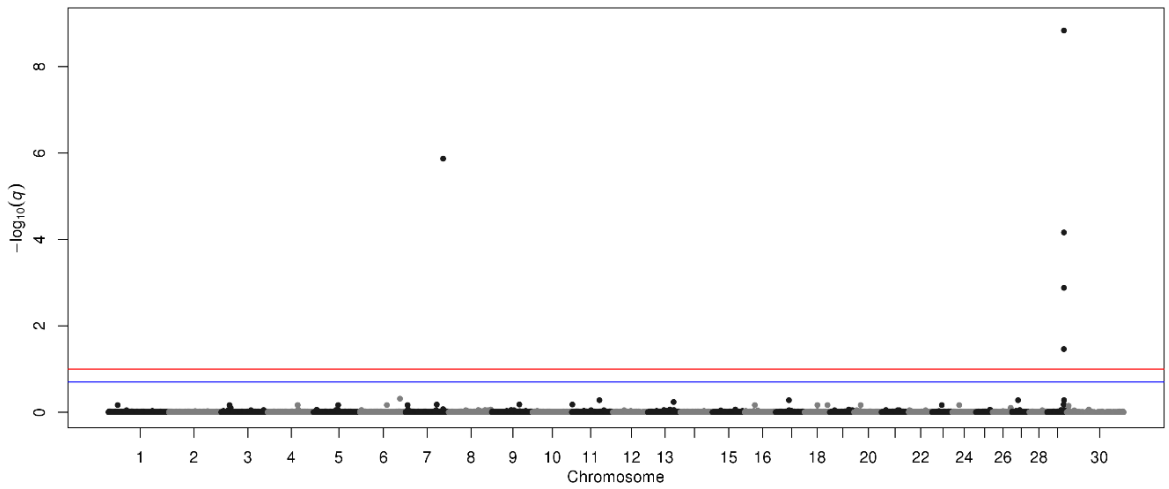

Figure a.23 Manhattan plot of SNP q-values estimated in the multivariate analysis of MB, WBSF, and CL in the multiple-breed population.

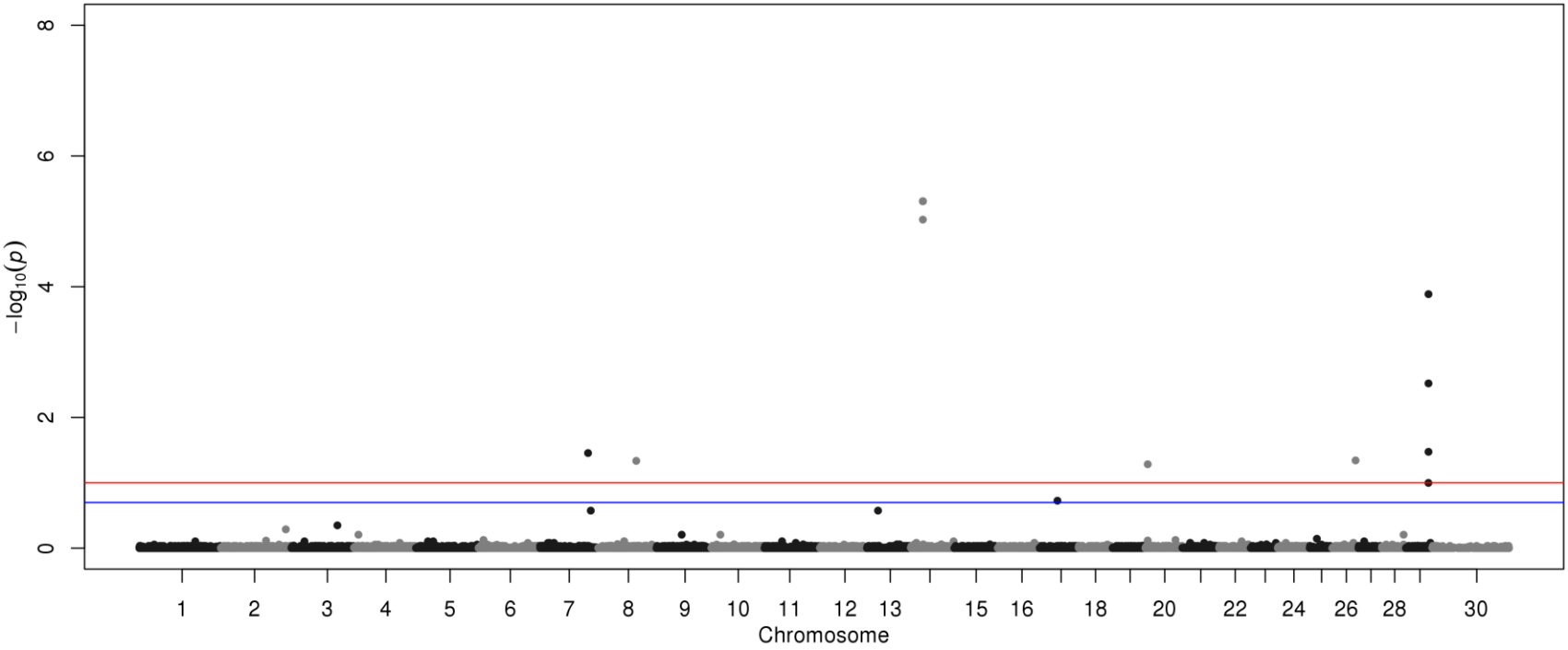

Figure a.24 Manhattan plot of SNP q-values estimated in the multivariate analysis of MB, WBSF, CL, HCW, FT, REA, IF, and KPH.

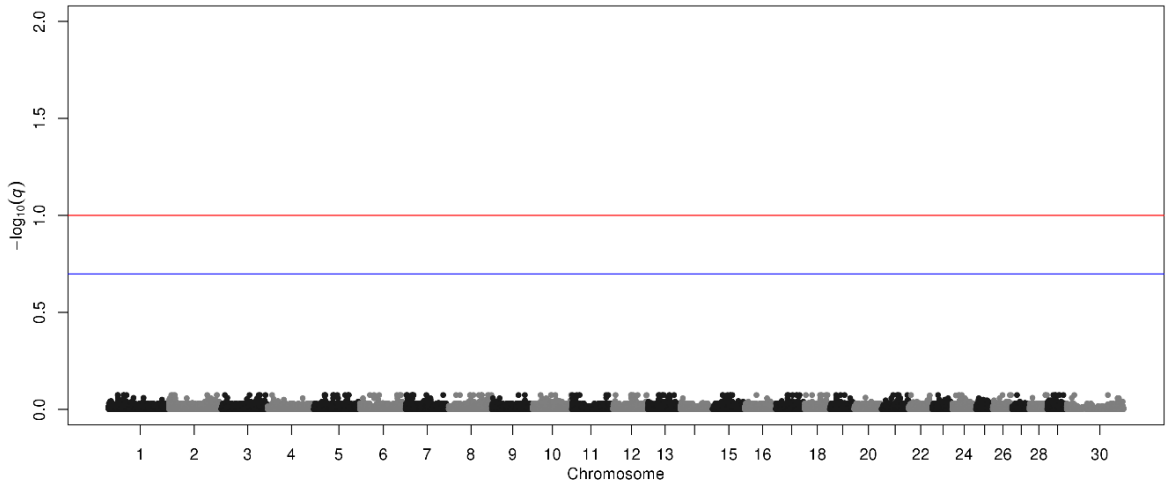

Figure a.25 Manhattan plot of SNP q-values estimated in the univariate analysis of MB in Angus.

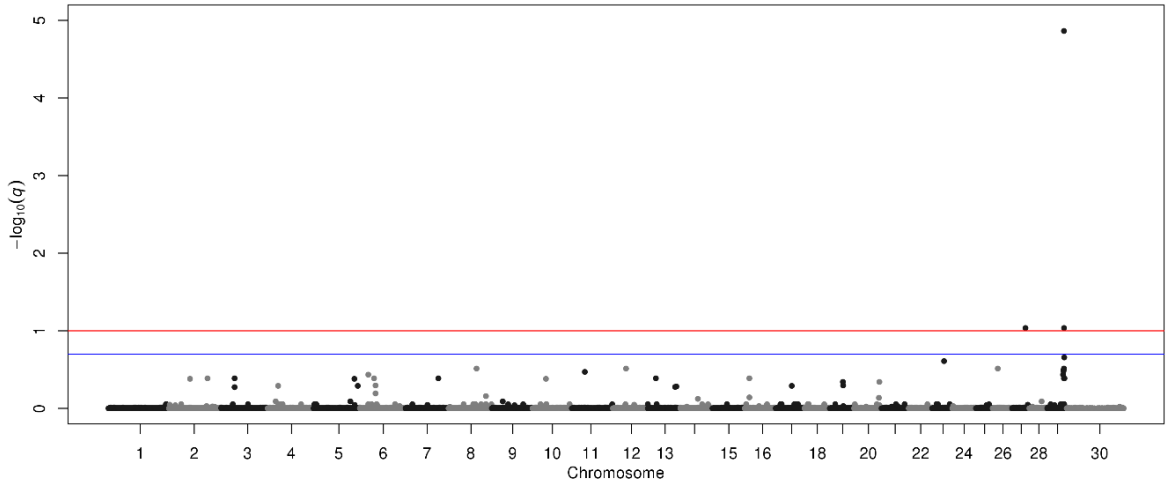

Figure a.26 Manhattan plot of SNP q-values estimated in the univariate analysis of WBSF in Angus.

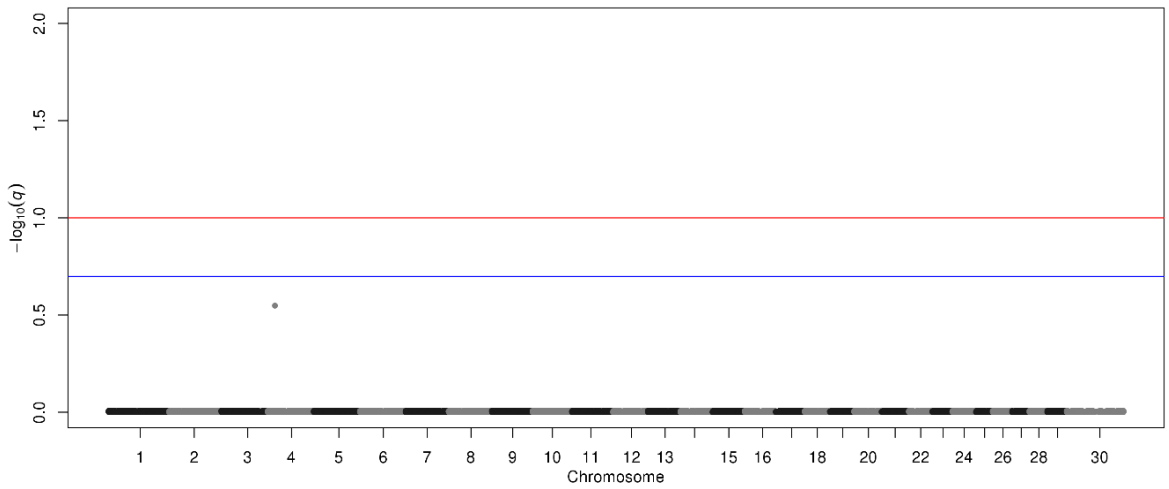

Figure a.27 Manhattan plot of SNP q-values estimated in the univariate analysis of CL in Angus.

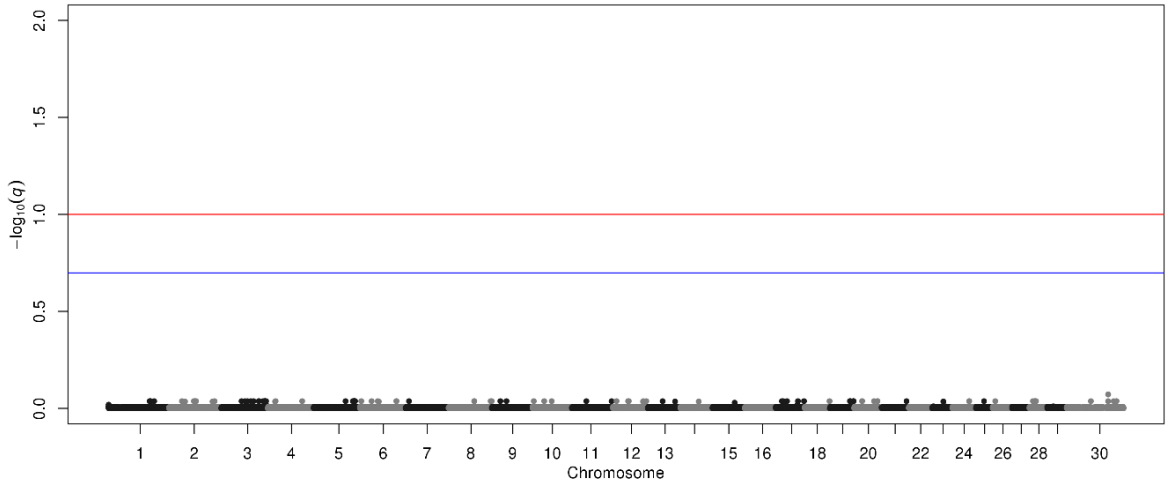

Figure a.28 Manhattan plot of SNP q-values estimated in the univariate analysis of HCW in Angus.

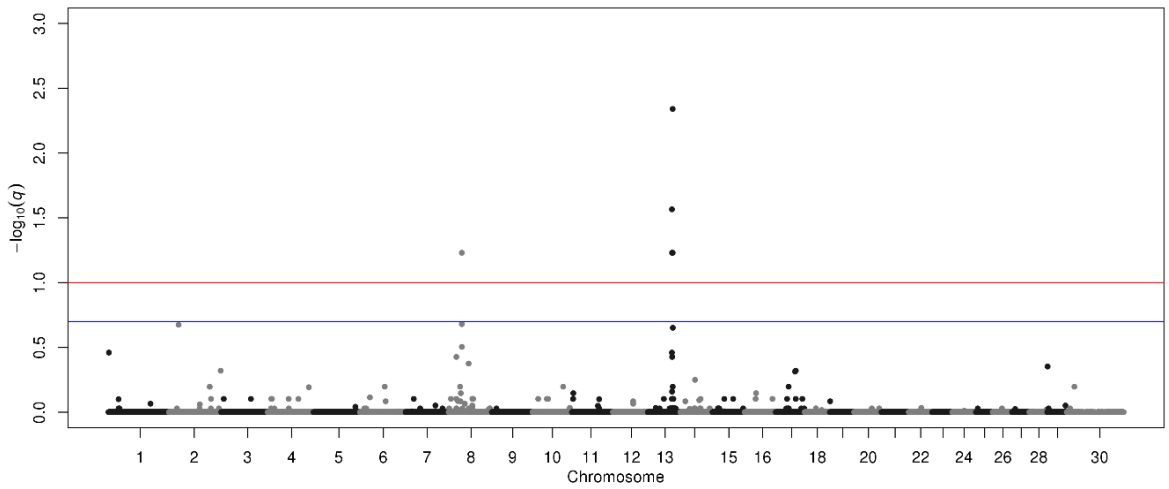

Figure a.29 Manhattan plot of SNP q-values estimated in the univariate analysis of FT in Angus.

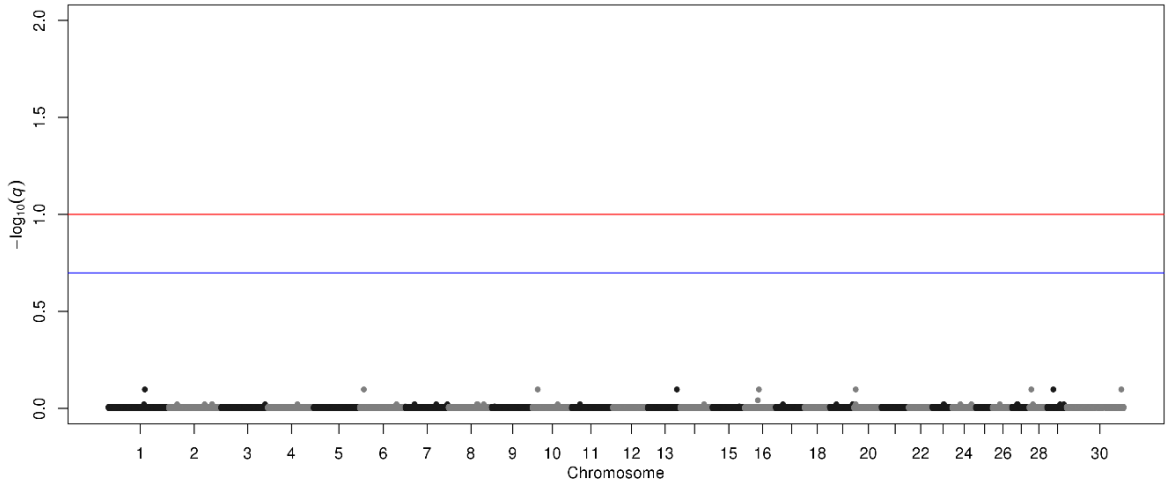

Figure a.30 Manhattan plot of SNP q-values estimated in the univariate analysis of REA in Angus.

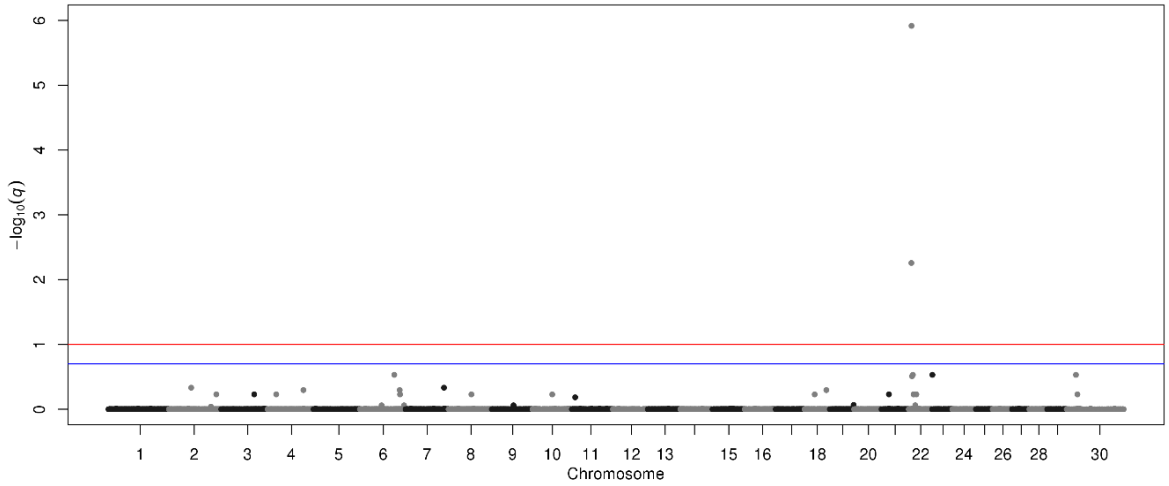

Figure a.31 Manhattan plot of SNP q-values estimated in the univariate analysis of KPH in Angus.

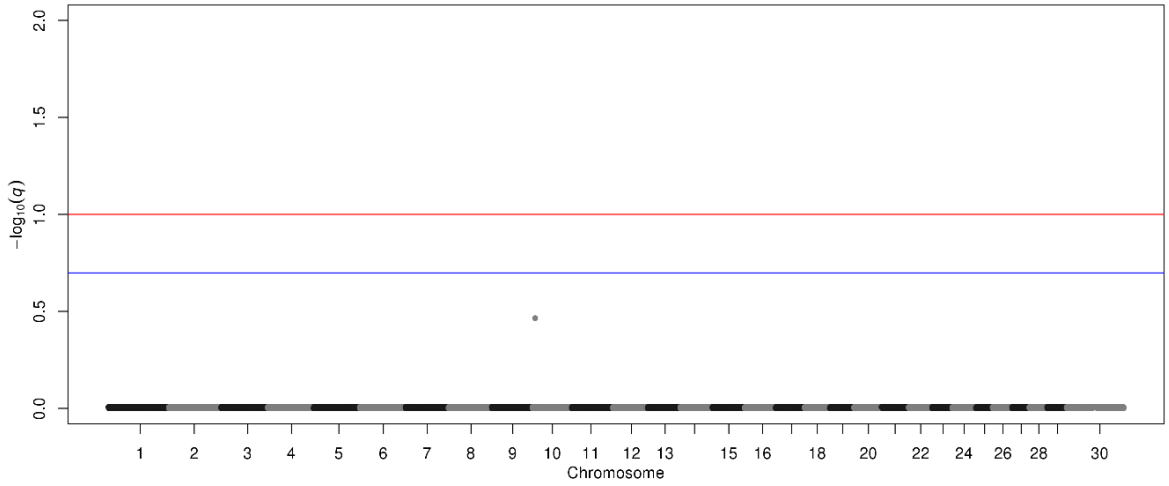

Figure a.32 Manhattan plot of SNP q-values estimated in the univariate analysis of IF in Angus.

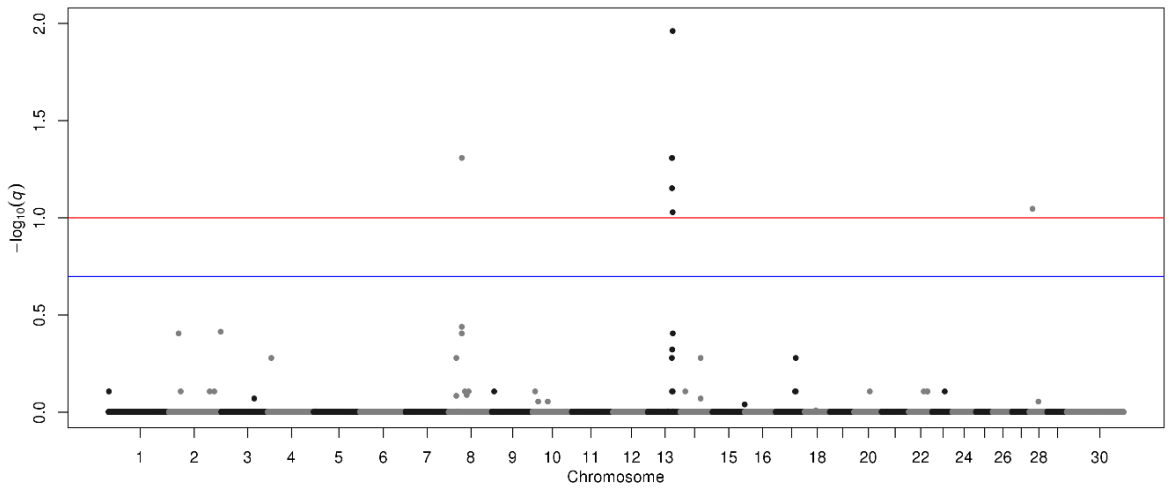

Figure a.33 Manhattan plot of SNP q-values estimated in the multivariate analysis of FT and KPH in Angus.

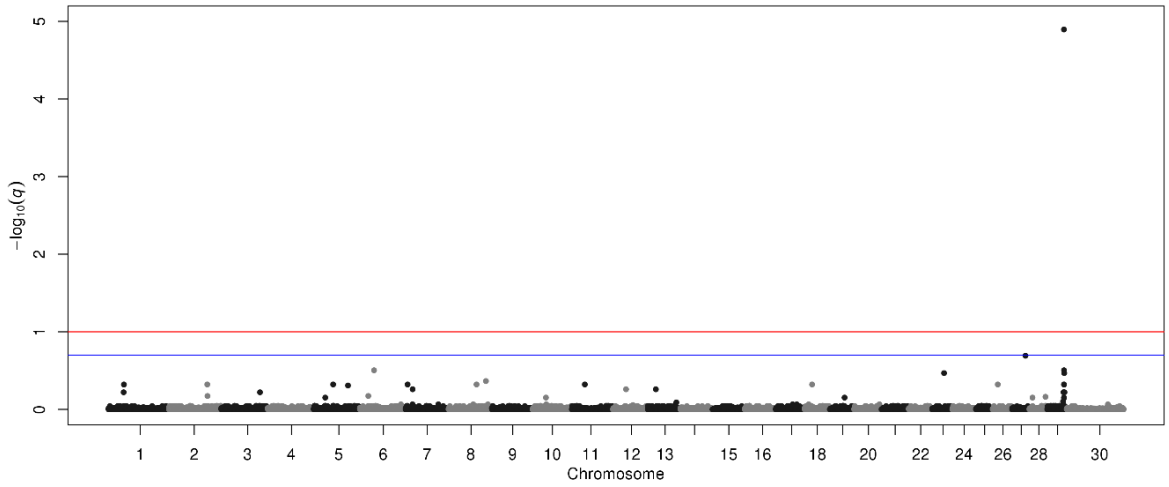

Figure a.34 Manhattan plot of SNP q-values estimated in the multivariate analysis of MB and WBSF in Angus.

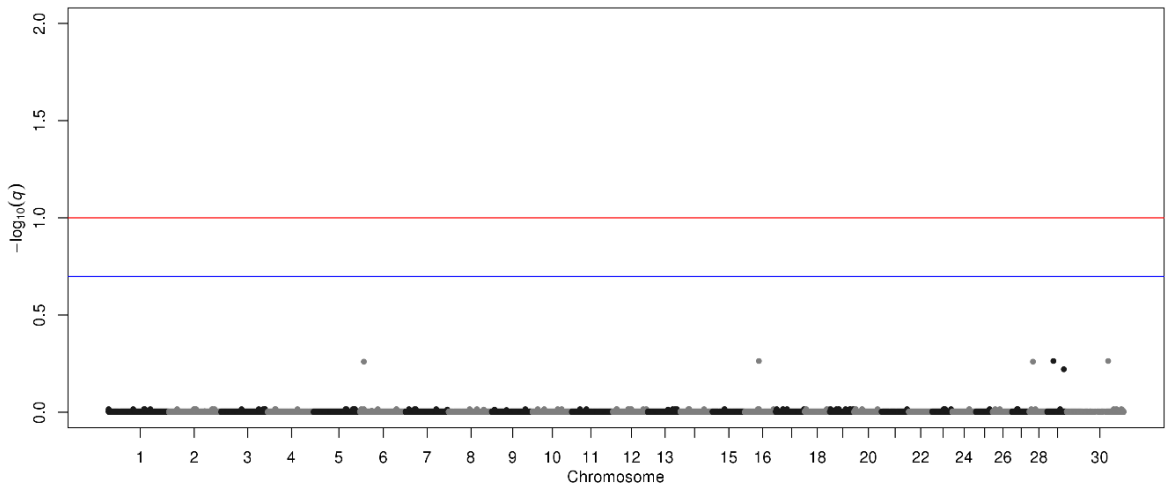

Figure a.35 Manhattan plot of SNP q-values estimated in the multivariate analysis of HCW and REA in Angus.

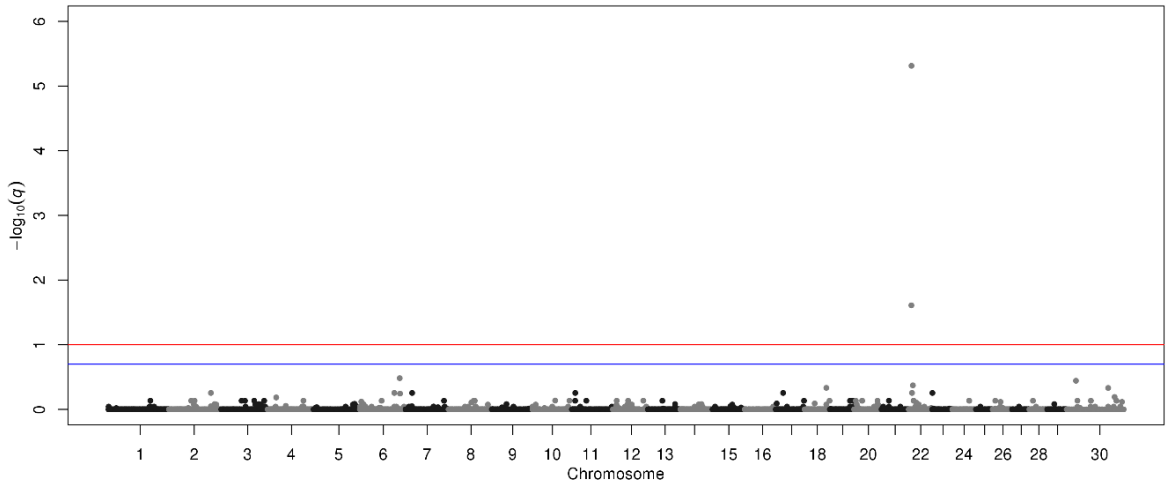

Figure a.36 Manhattan plot of SNP q-values estimated in the multivariate analysis of HCW and KPH in Angus.

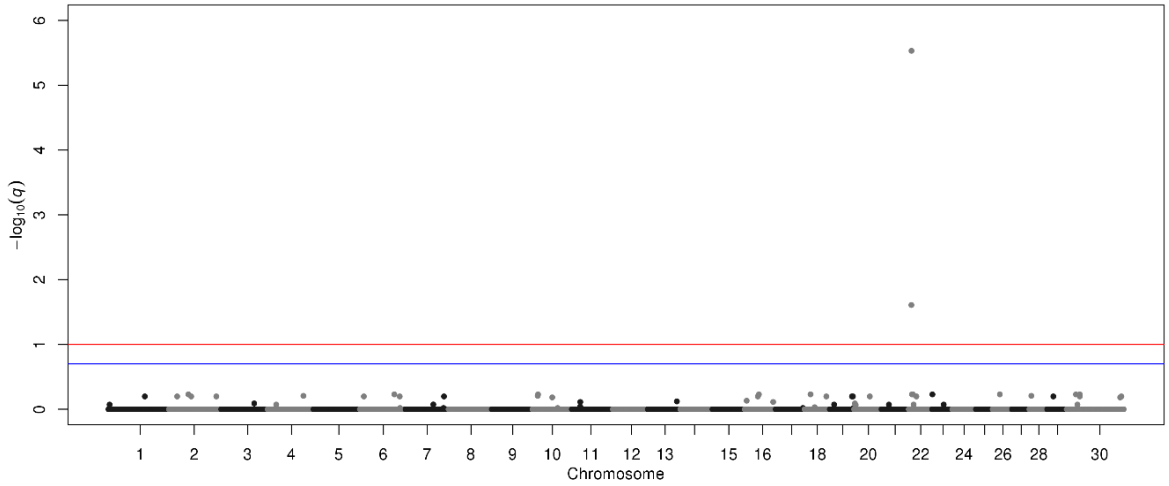

Figure a.37 Manhattan plot of SNP q-values estimated in the multivariate analysis of REA and KPH in Angus.

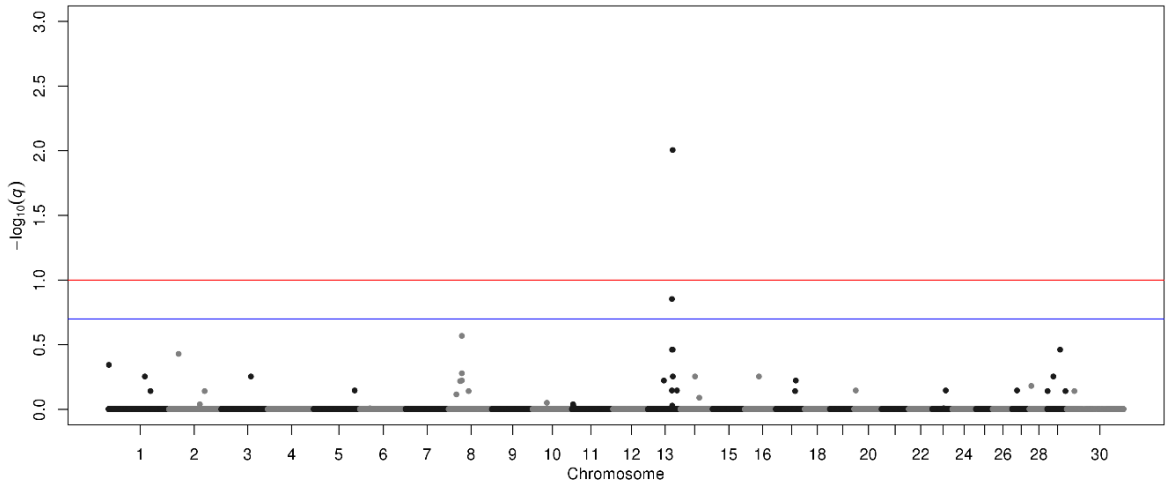

Figure a.38 Manhattan plot of SNP q-values estimated in the multivariate analysis of FT and REA in Angus.

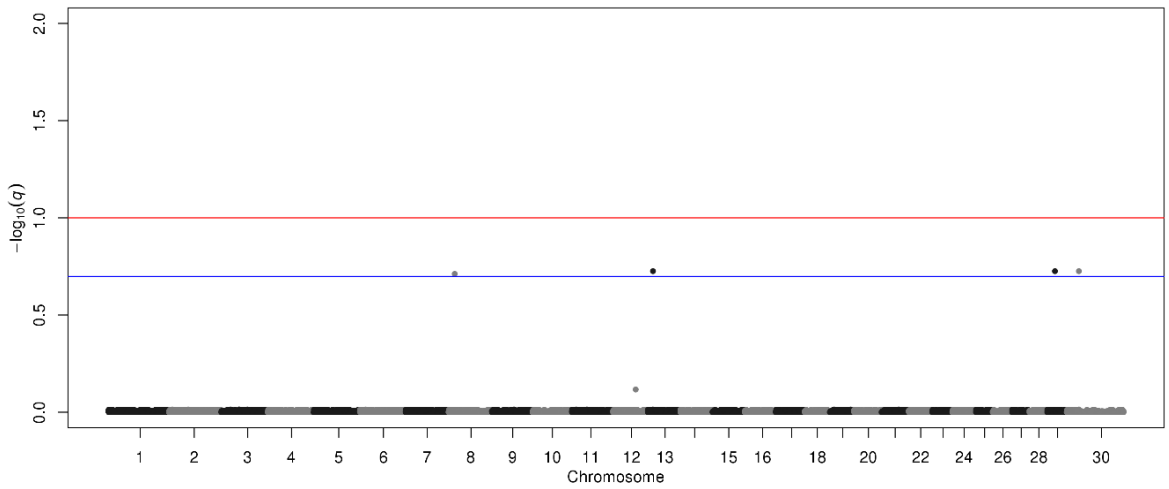

Figure a.39 Manhattan plot of SNP q-values estimated in the univariate analysis of MB in Charolais.

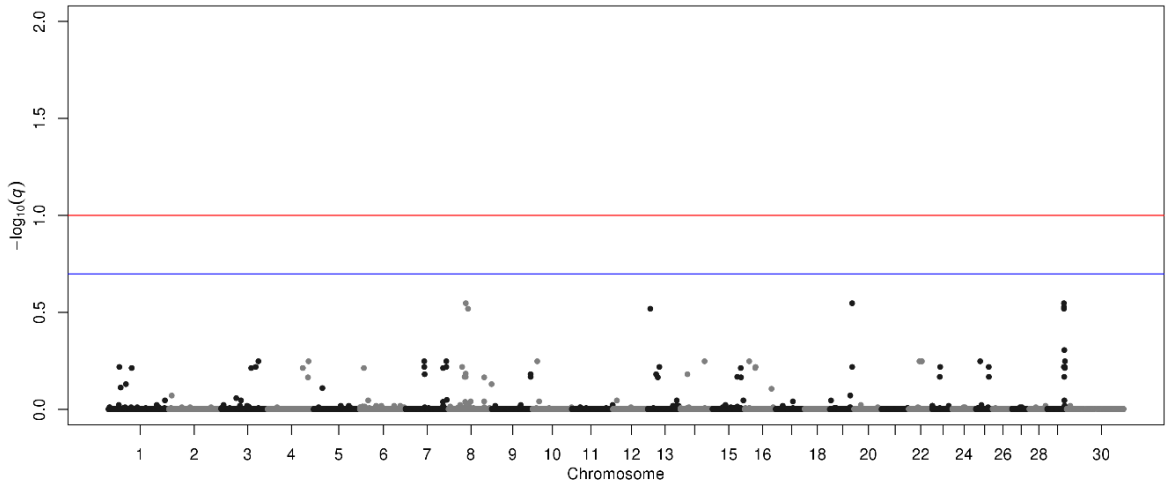

Figure a.40 Manhattan plot of SNP q-values estimated in the univariate analysis of WBSF in Charolais.

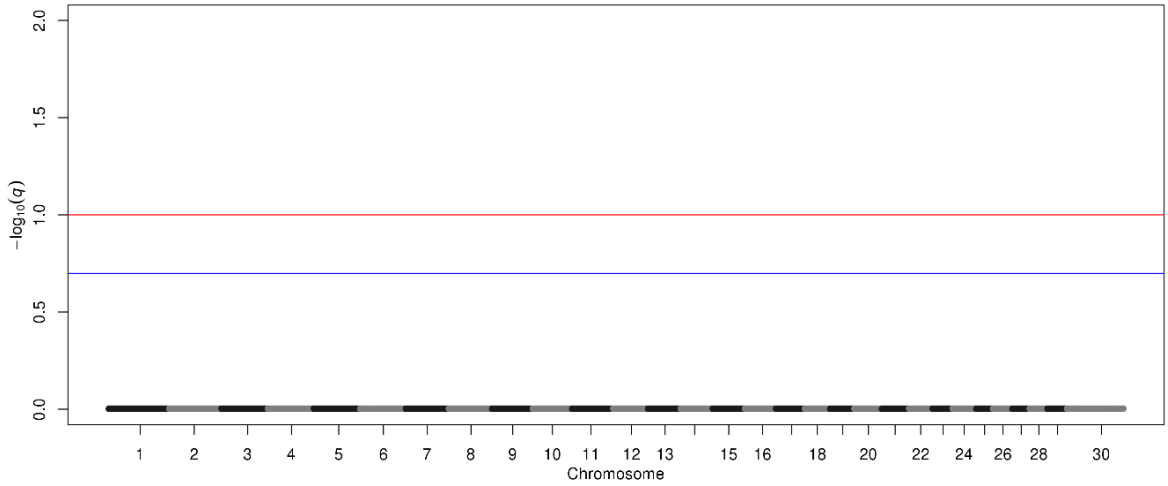

Figure a.41 Manhattan plot of SNP *q*-values estimated in the univariate analysis of CL in Charolais.

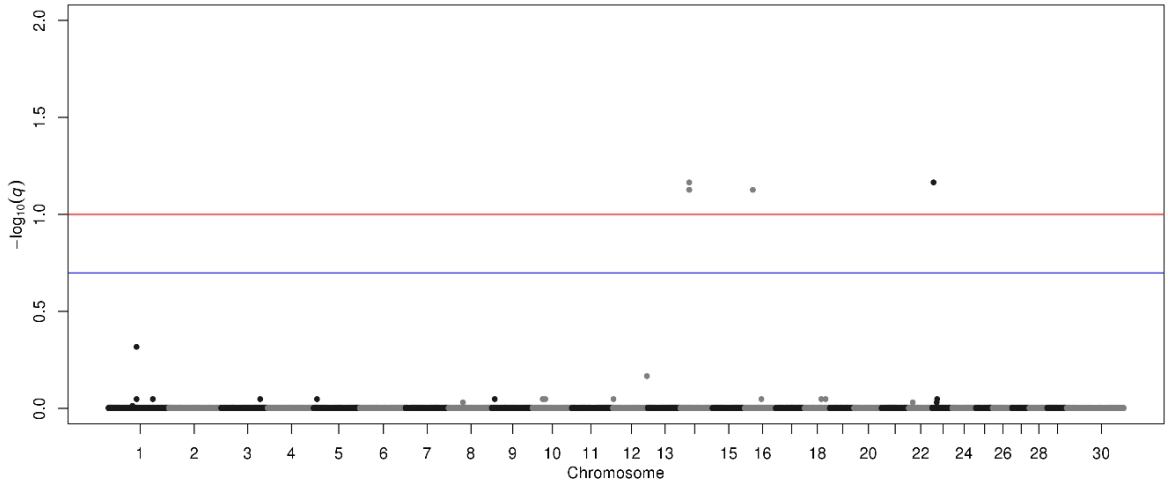

Figure a.42 Manhattan plot of SNP *q*-values estimated in the univariate analysis of HCW in Charolais.

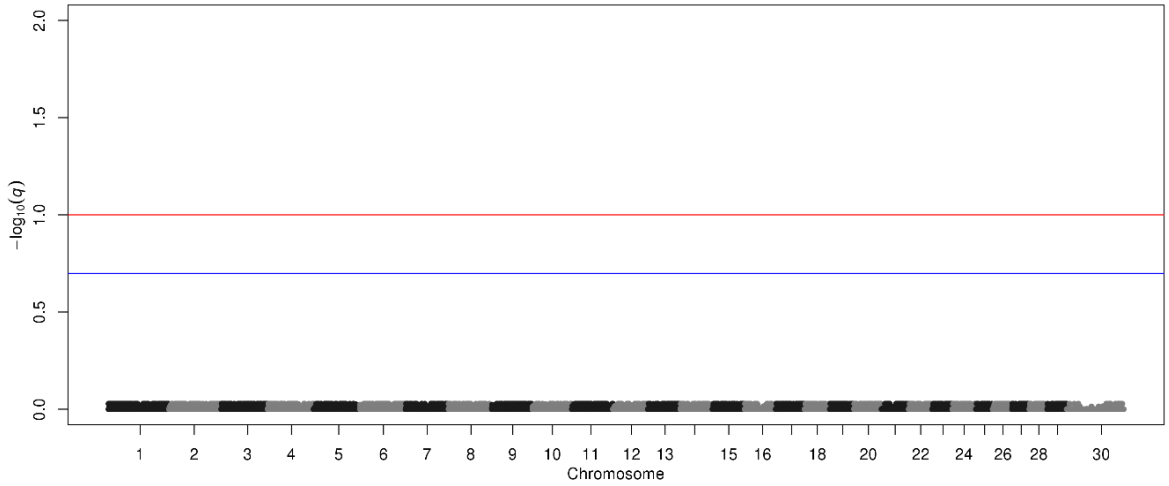

Figure a.43 Manhattan plot of SNP *q*-values estimated in the univariate analysis of FT in Charolais.

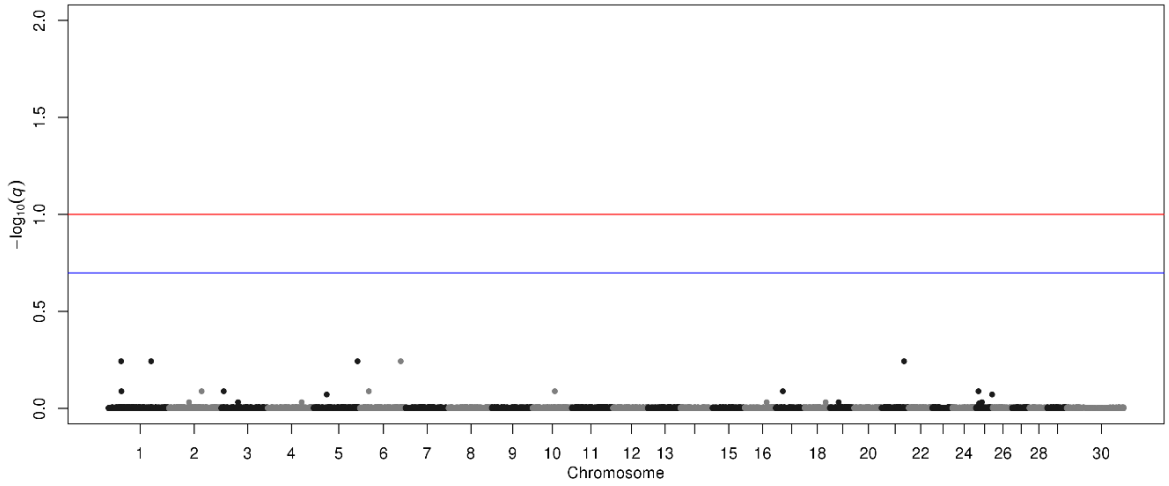

Figure a.44 Manhattan plot of SNP *q*-values estimated in the univariate analysis of KPH in Charolais.

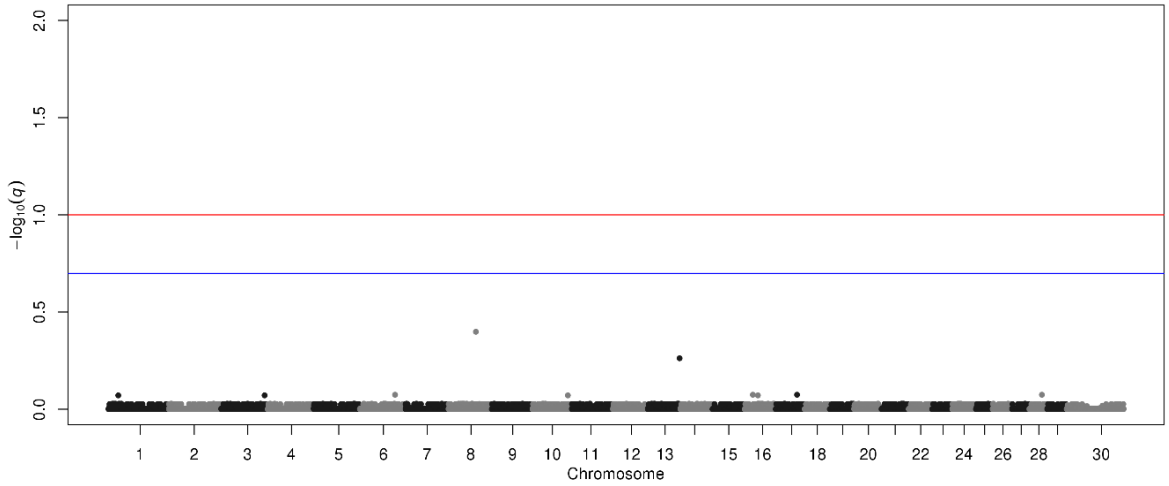

Figure a.45 Manhattan plot of SNP *q*-values estimated in the univariate analysis of REA in Charolais.

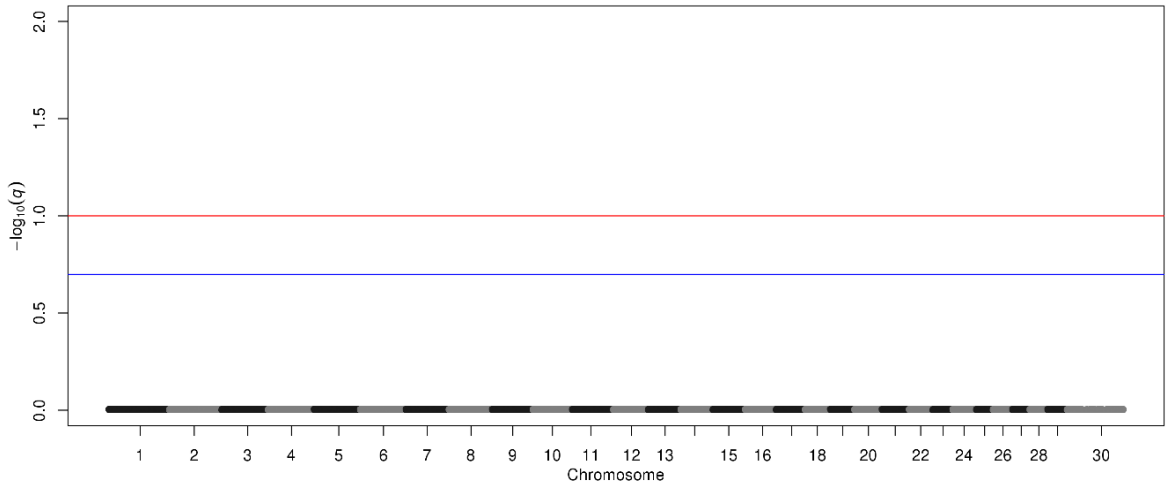

Figure a.46 Manhattan plot of SNP *q*-values estimated in the univariate analysis of IF in Charolais.

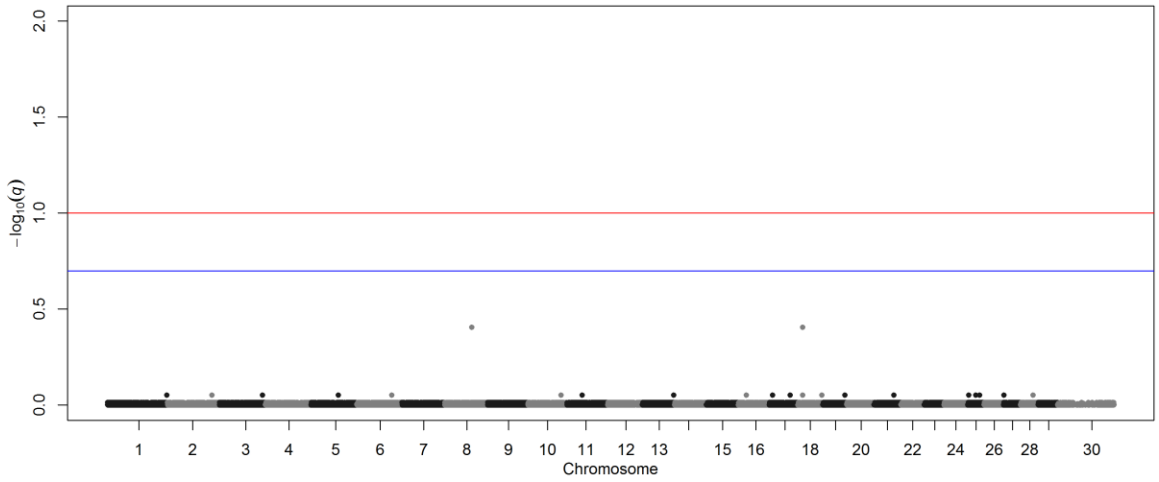

Figure a.47 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of FT and REA in Charolais.

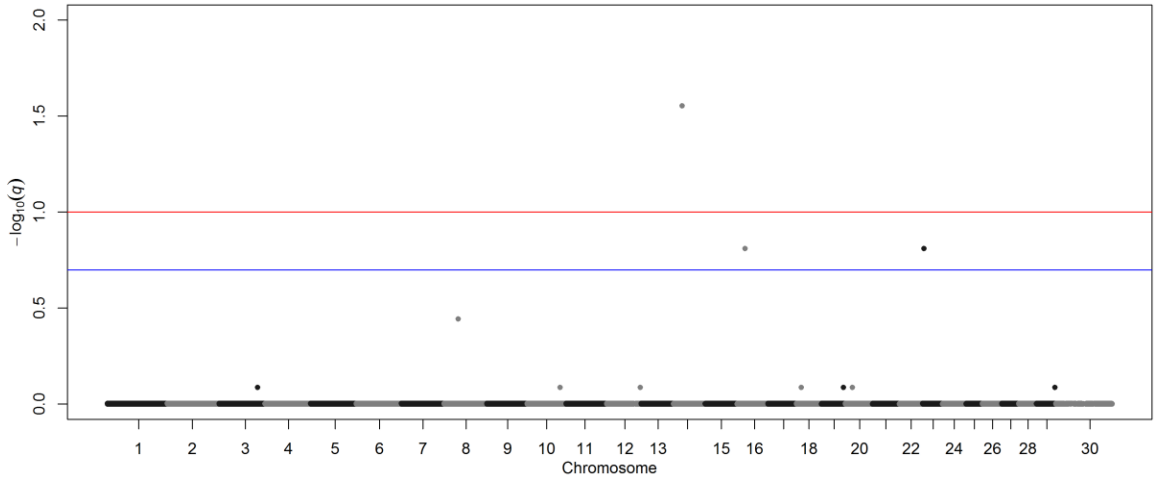

Figure a.48 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW and FT in Charolais.

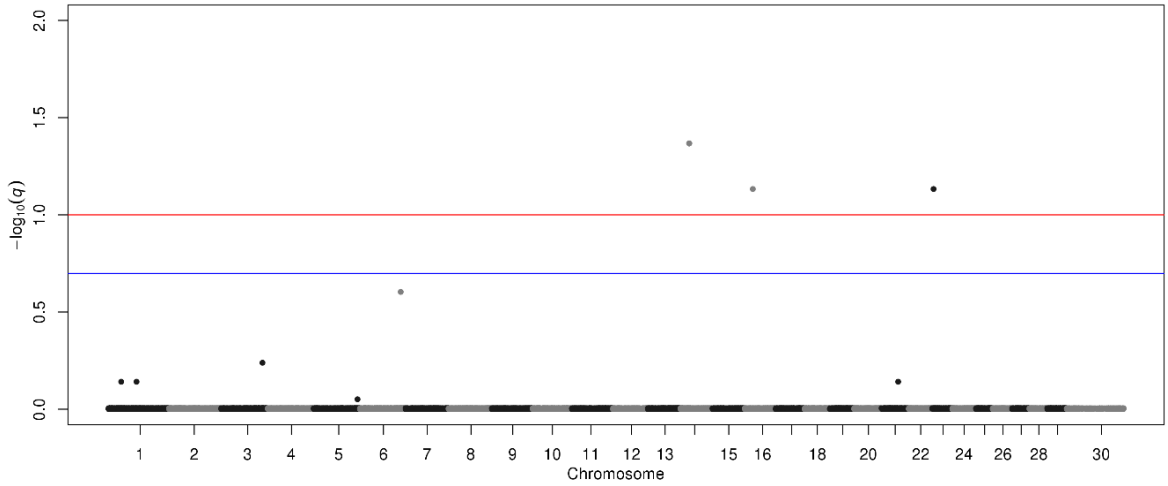

Figure a.49 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW and KPH in Charolais.

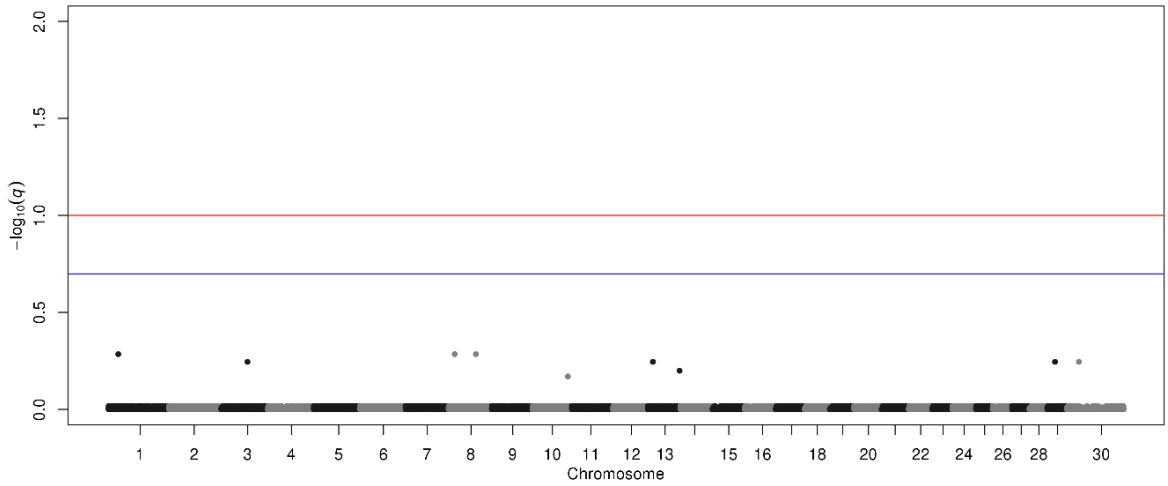

Figure a.50 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and REA in Charolais.

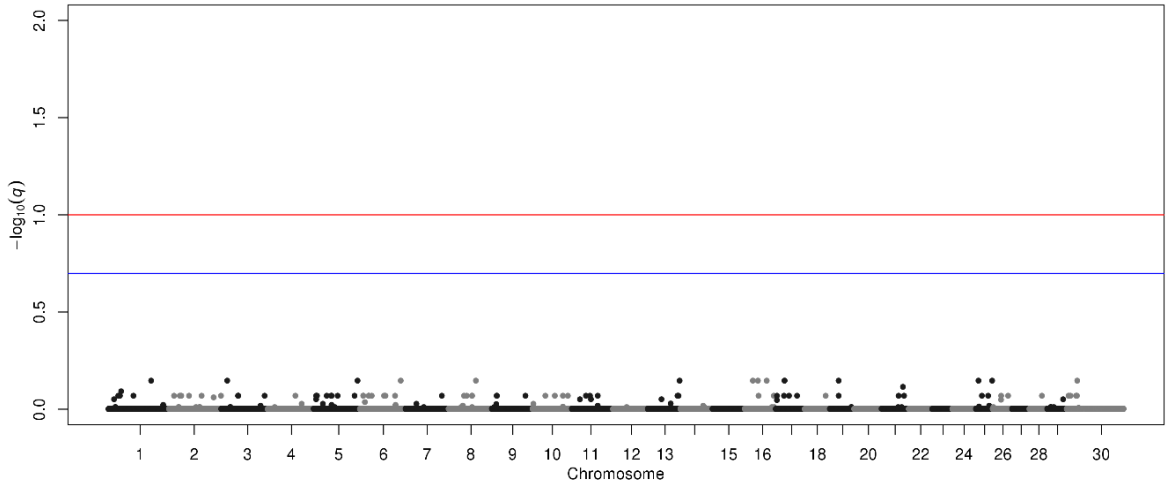

Figure a.51 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of REA and KPH in Charolais.

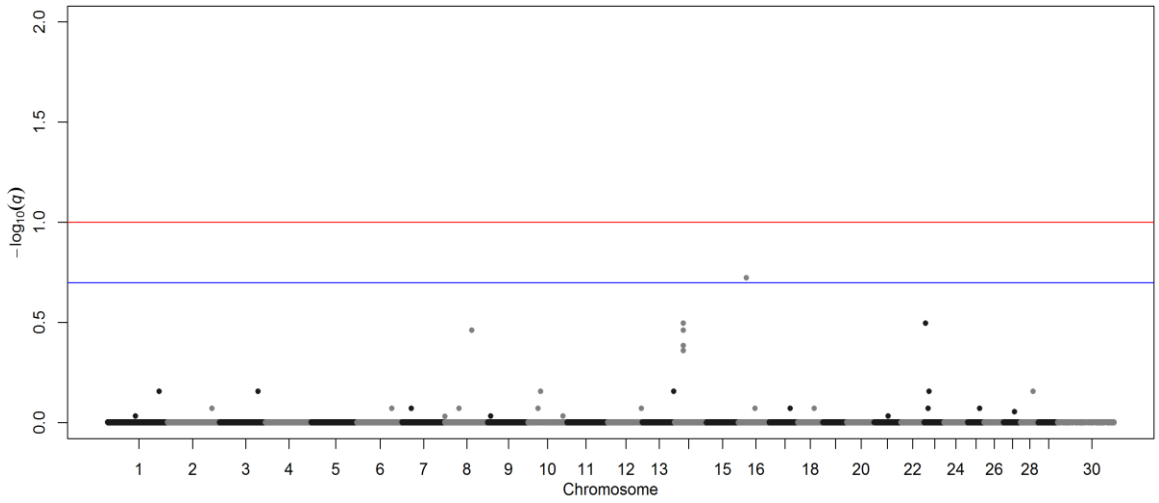

Figure a.52 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW and REA in Charolais.

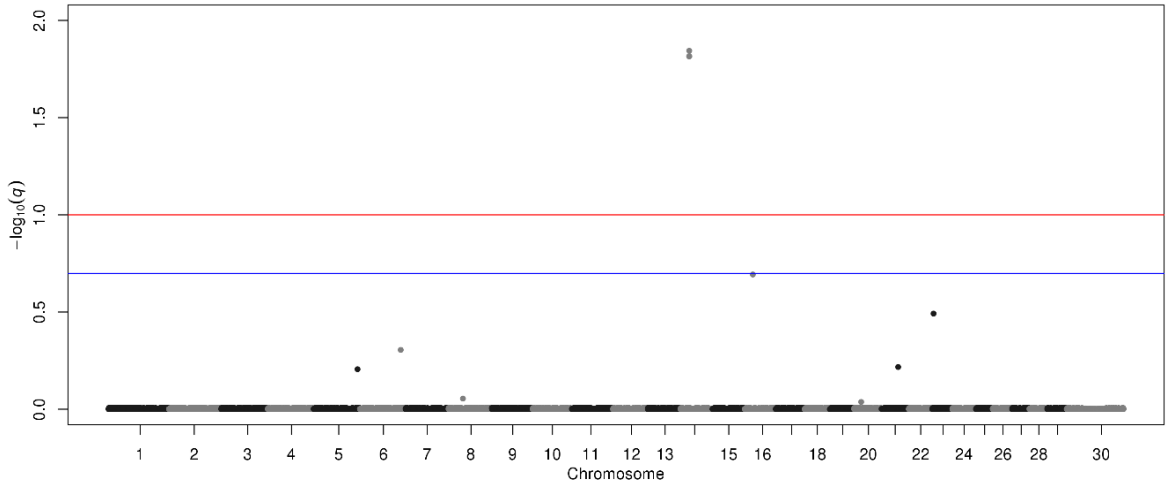

Figure a.53 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT and KPH in Charolais.

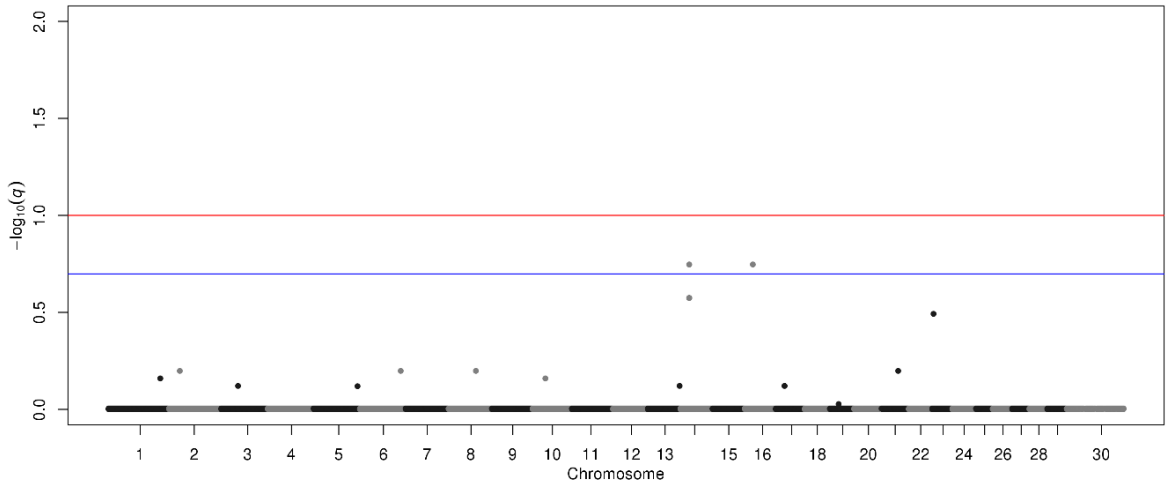

Figure a.54 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, REA and KPH in Charolais.

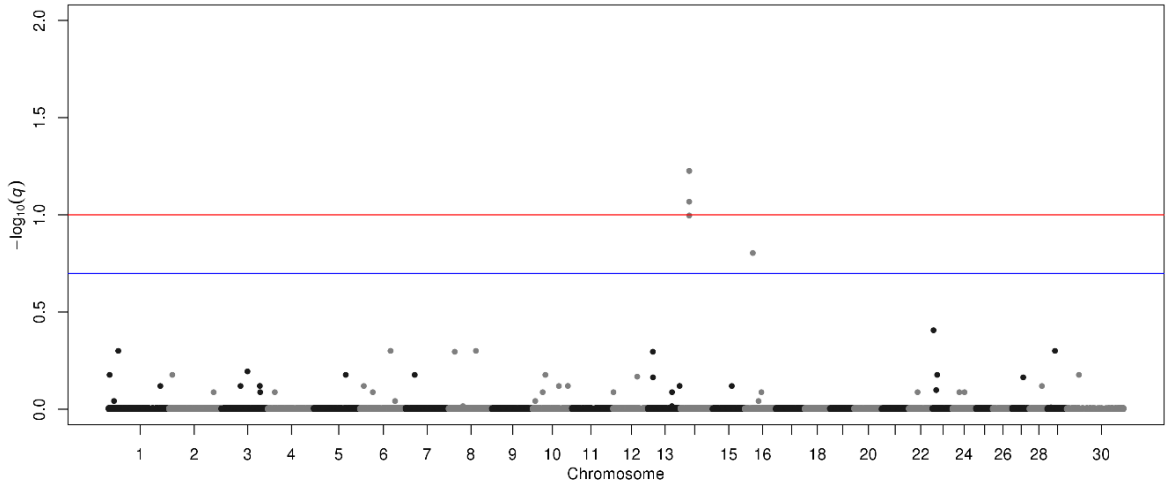

Figure a.55 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB, HCW, and REA in Charolais.

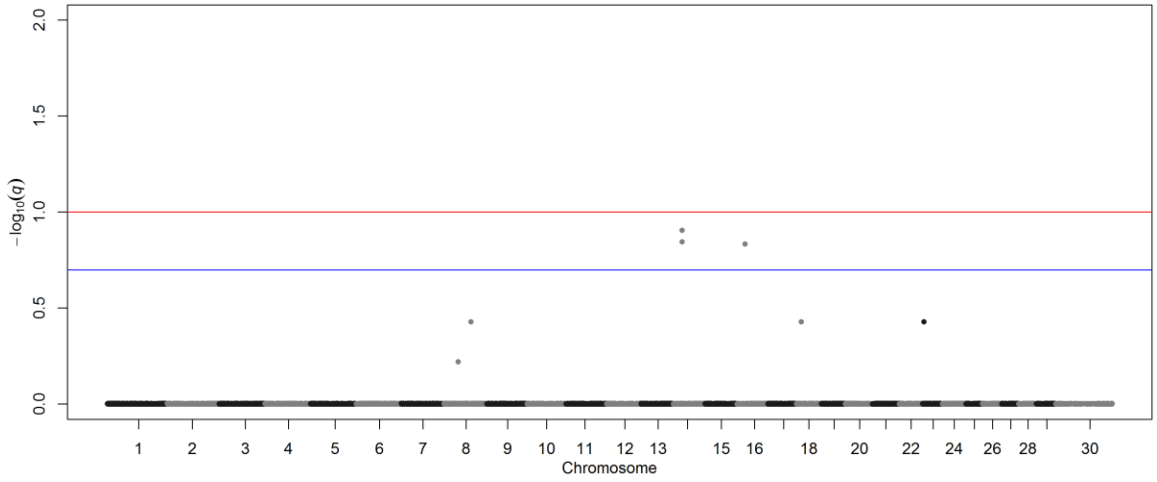

Figure a.56 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT and REA in Charolais.

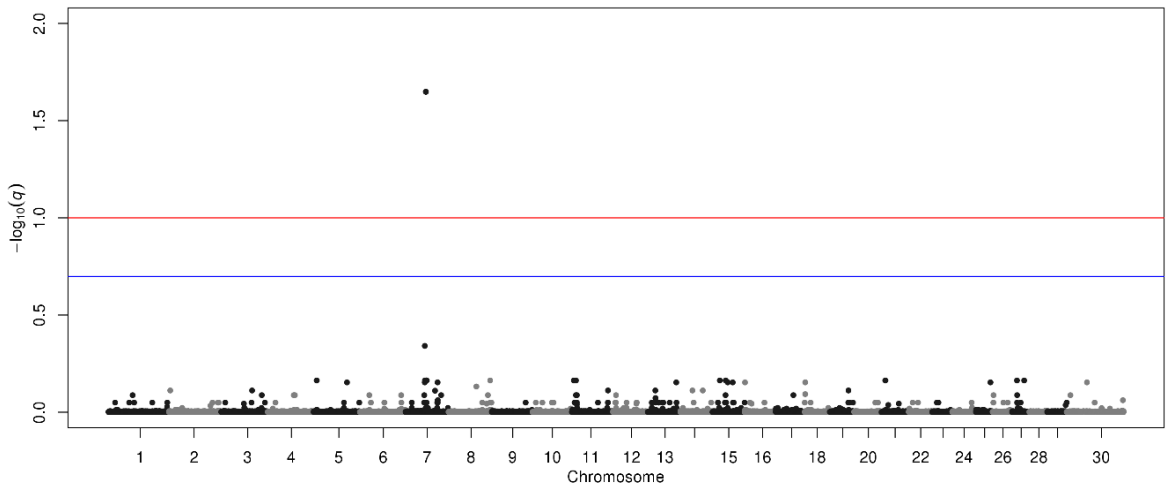

Figure a.57 Manhattan plot of SNP *q*-values estimated in the univariate analysis of MB in Hereford.

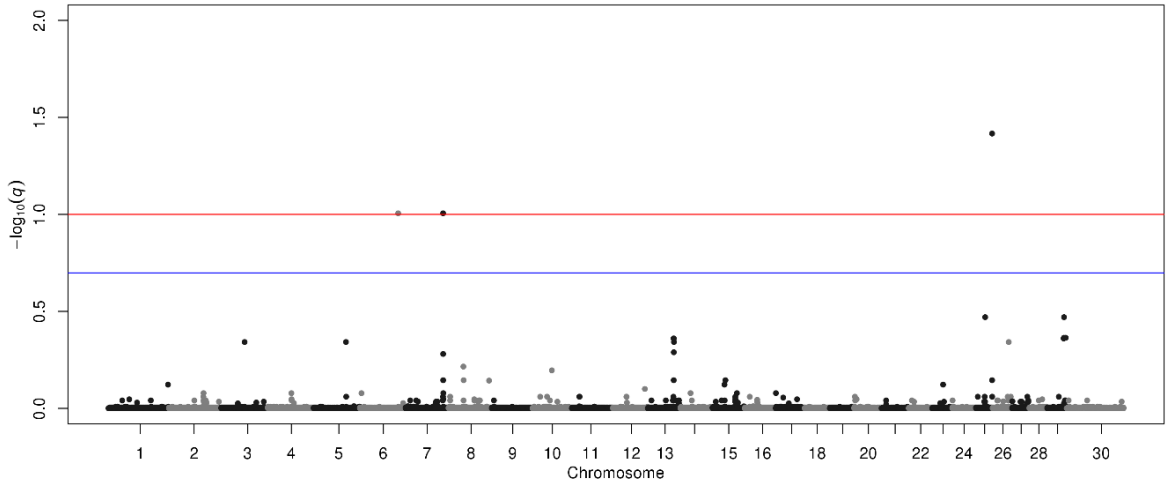

Figure a.58 Manhattan plot of SNP *q*-values estimated in the univariate analysis of WBSF in Hereford.

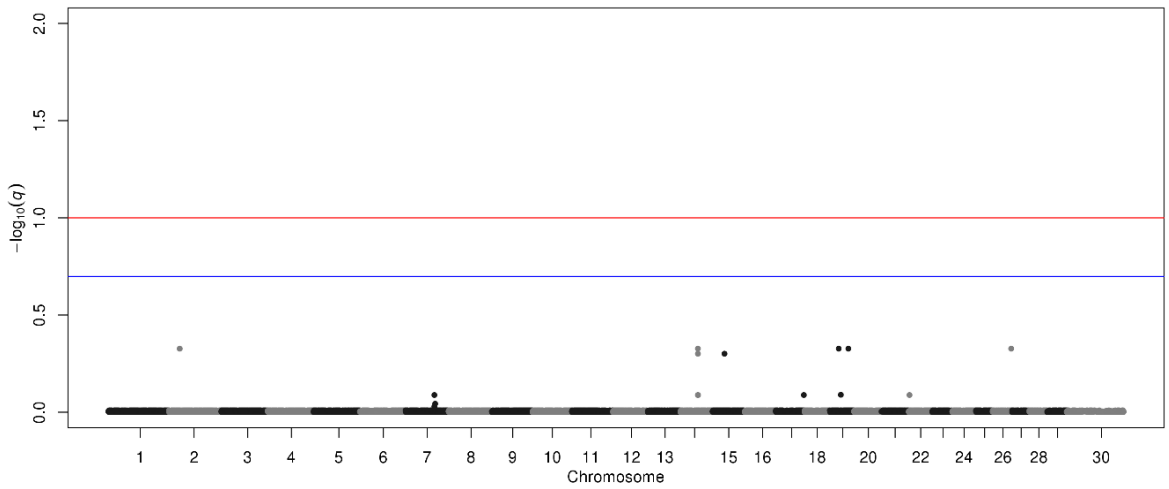

Figure a.59 Manhattan plot of SNP *q*-values estimated in the univariate analysis of CL in Hereford.

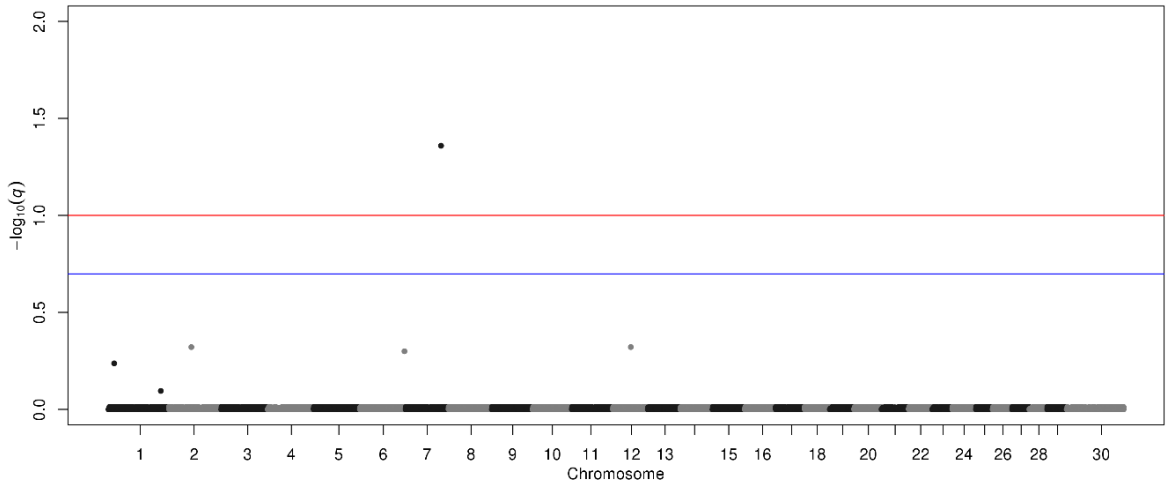

Figure a.60 Manhattan plot of SNP *q*-values estimated in the univariate analysis of HCW in Hereford.

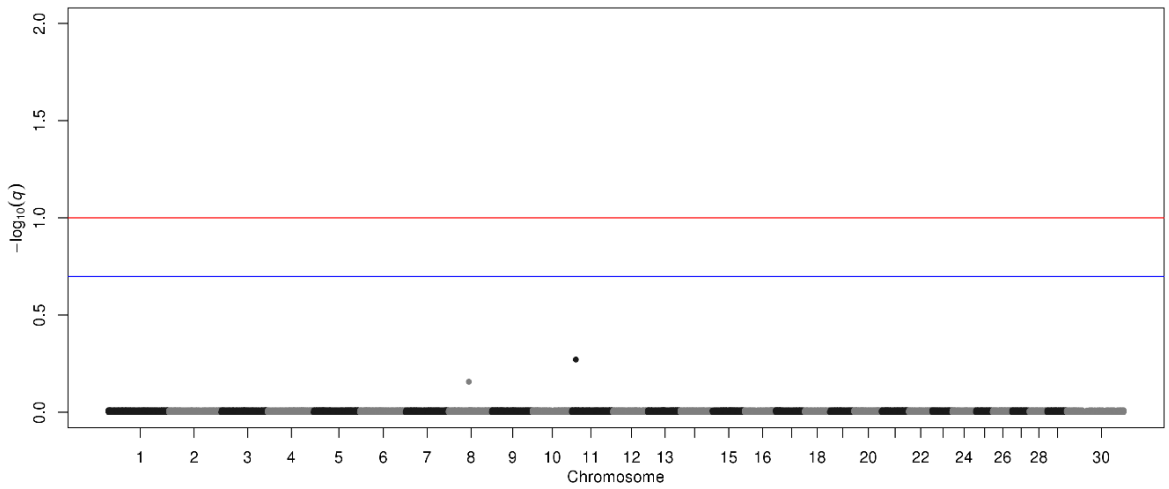

Figure a.61 Manhattan plot of SNP *q*-values estimated in the univariate analysis of FT in Hereford.

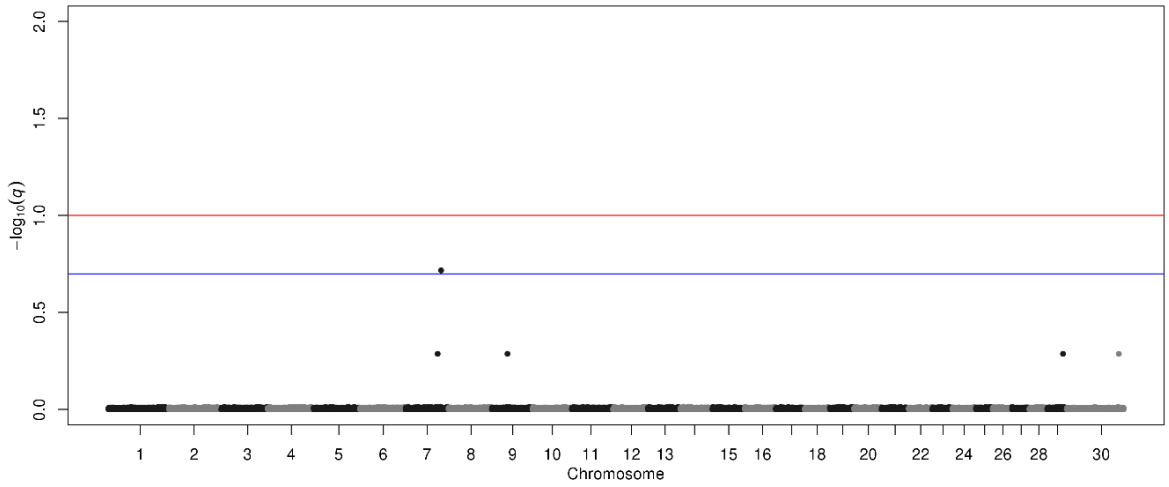

Figure a.62 Manhattan plot of SNP *q*-values estimated in the univariate analysis of REA in Hereford.

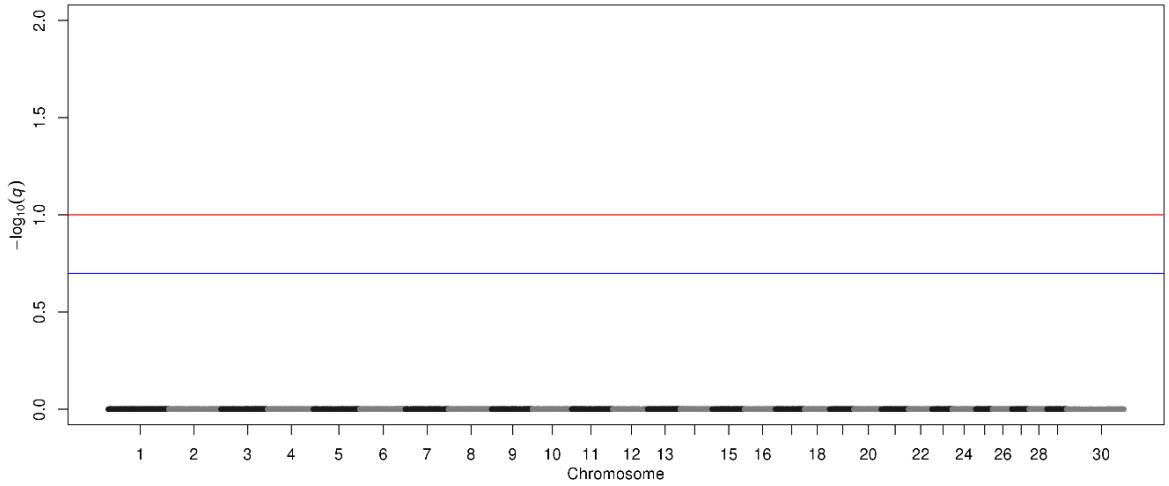

Figure a.63 Manhattan plot of SNP *q*-values estimated in the univariate analysis of KPH in Hereford.

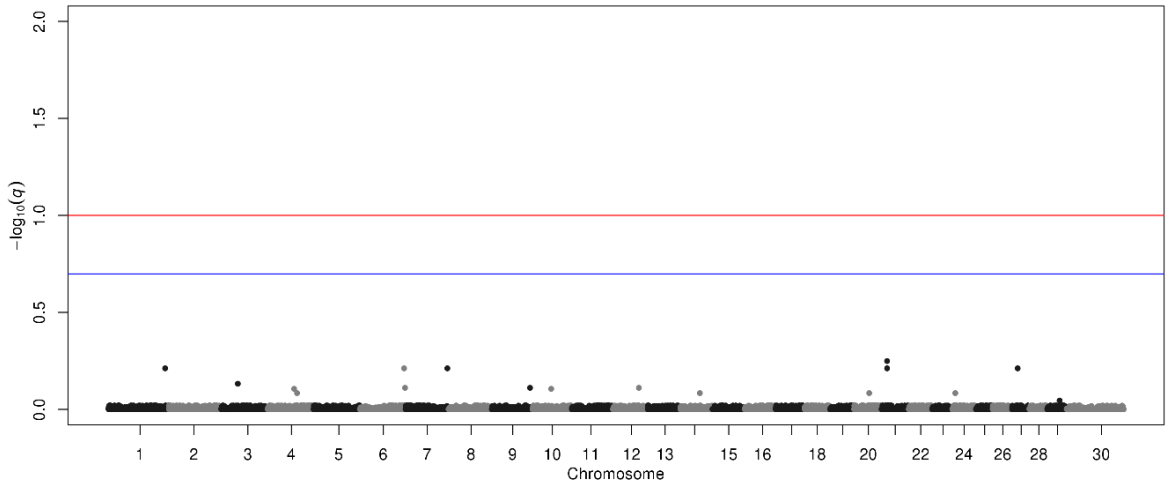

Figure a.64 Manhattan plot of SNP *q*-values estimated in the univariate analysis of IF in Hereford.

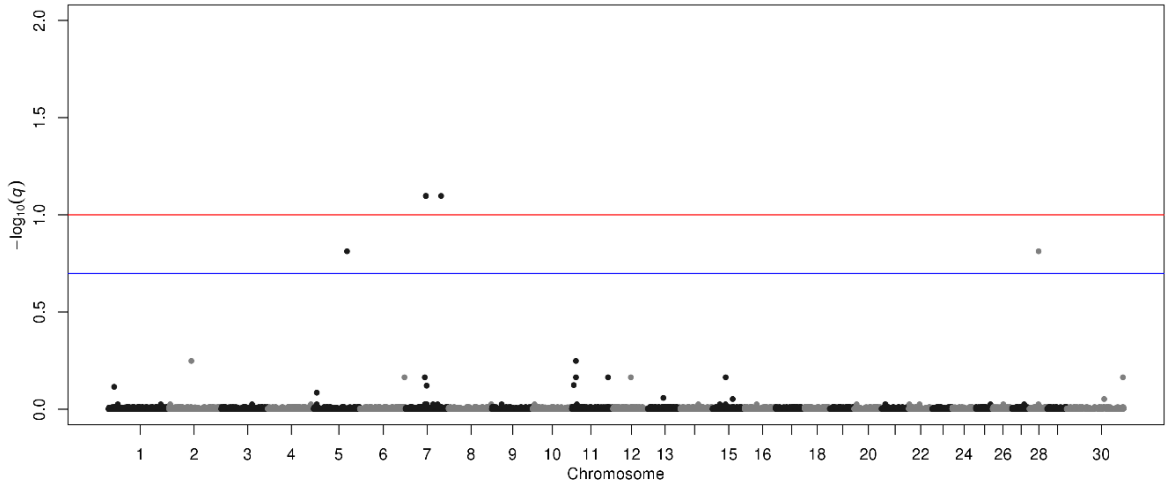

Figure a.65 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and HCW in Hereford.

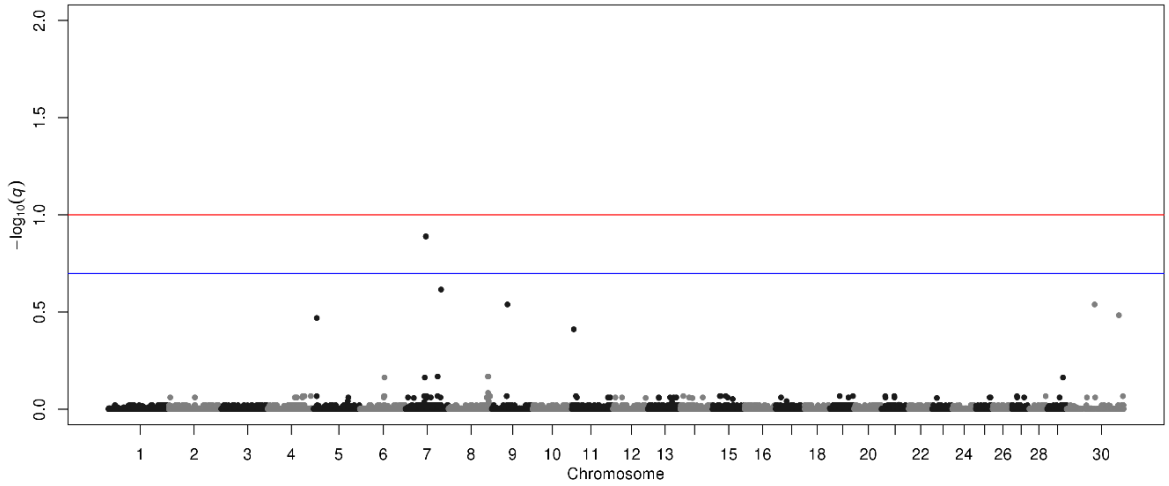

Figure a.66 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and REA in Hereford.

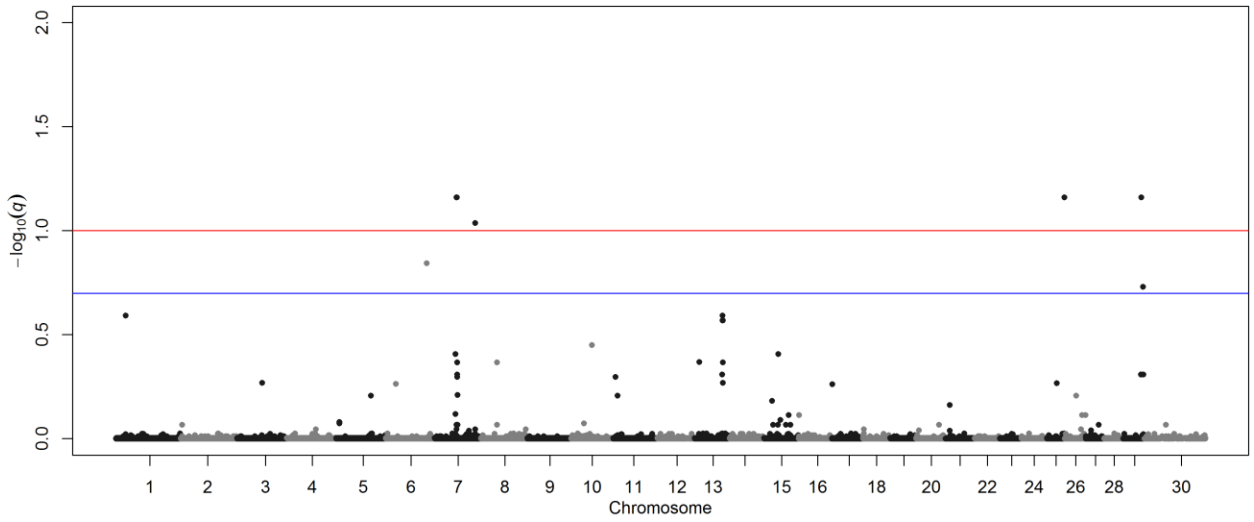

Figure a.67 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and WBSF in Hereford.

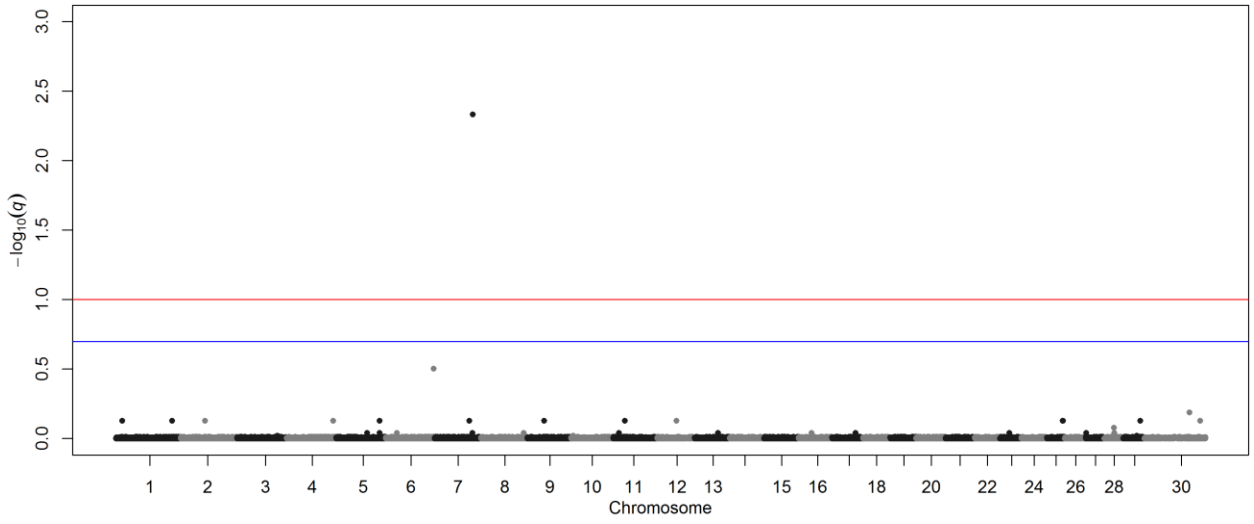

Figure a.68 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW and REA in Hereford.

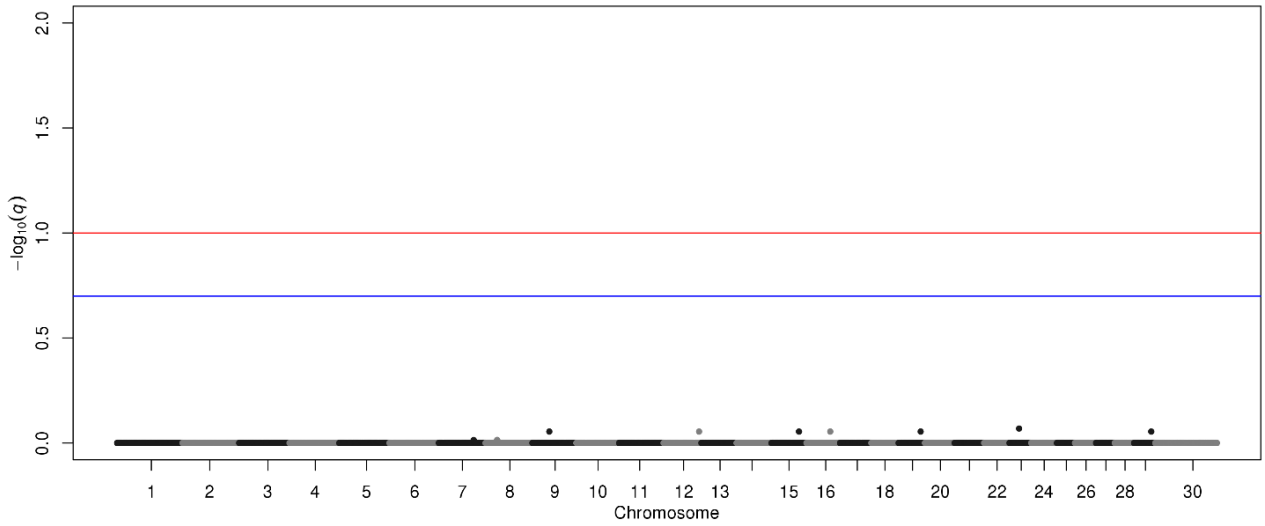

Figure a.69 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of REA and KPH in Hereford.

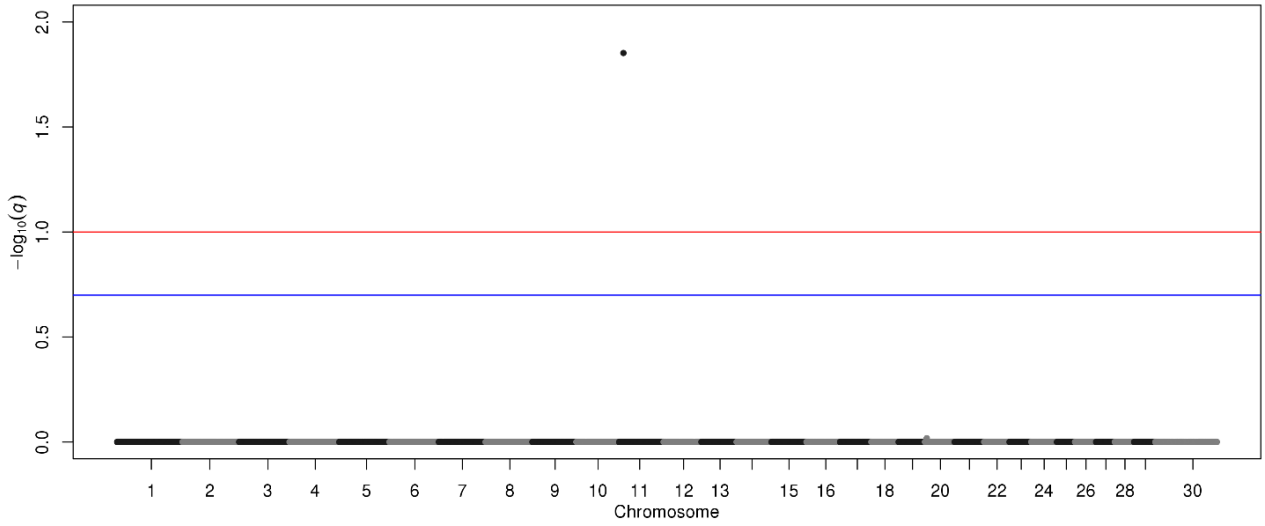

Figure a.70 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of FT and KPH in Hereford.

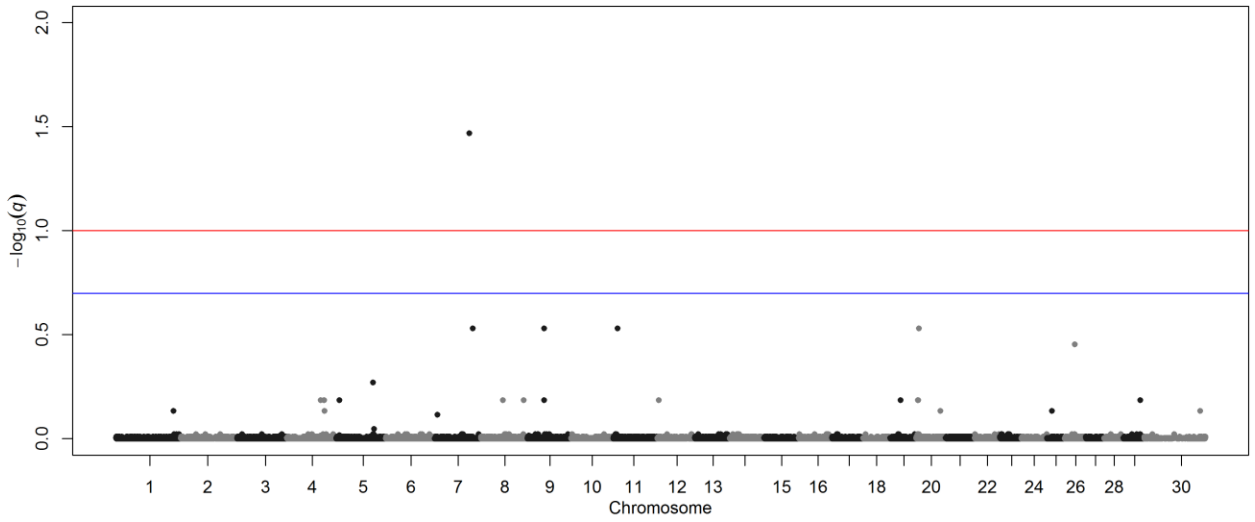

Figure a.71 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of FT and REA in Hereford.

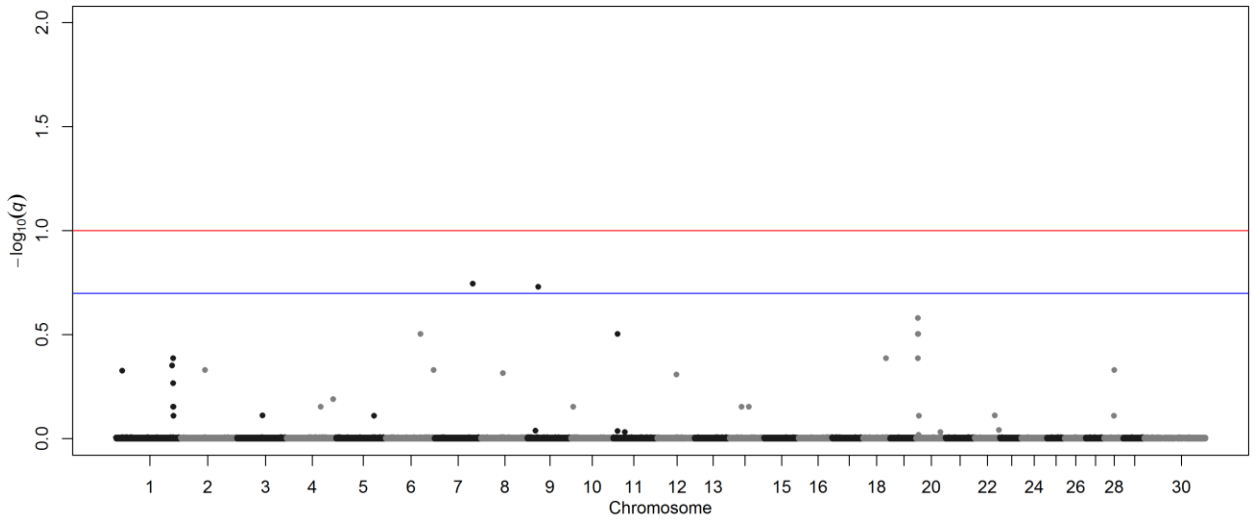

Figure a.72 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW and FT in Hereford.

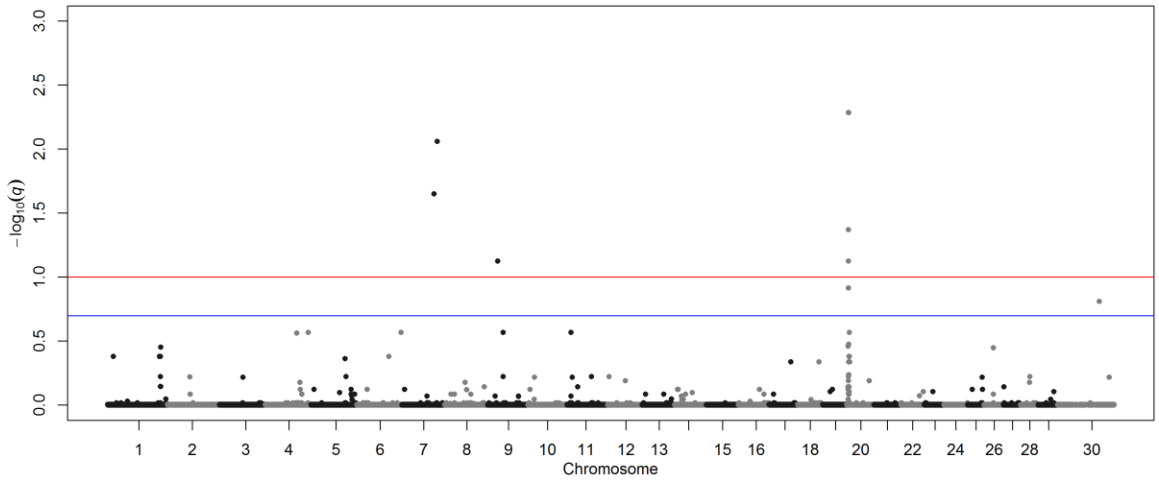

Figure a.73 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT and REA in Hereford.

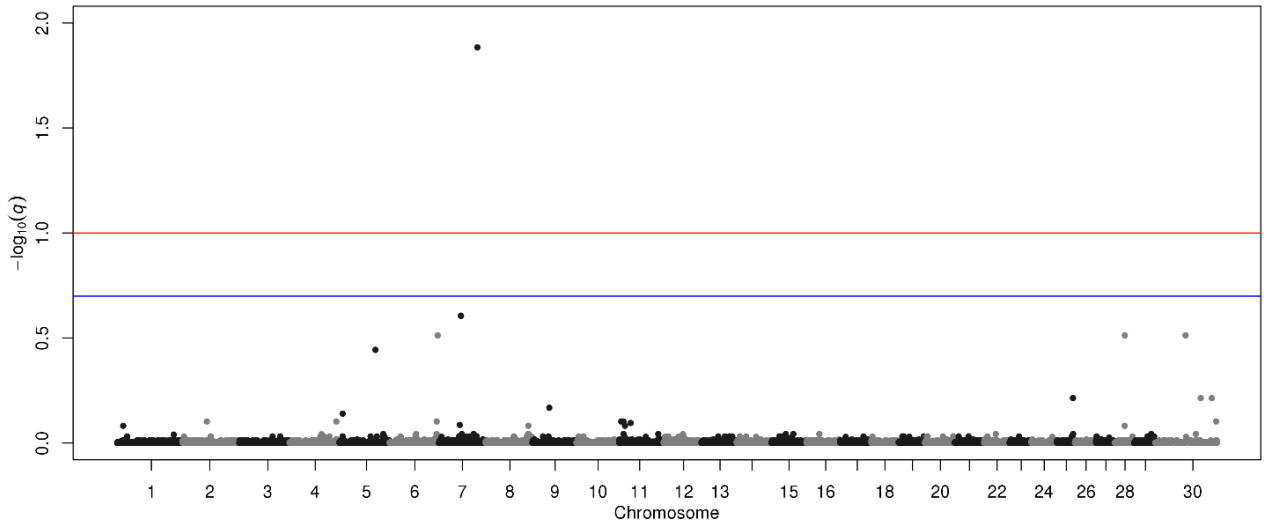

Figure a.74 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB, HCW and REA in Hereford.

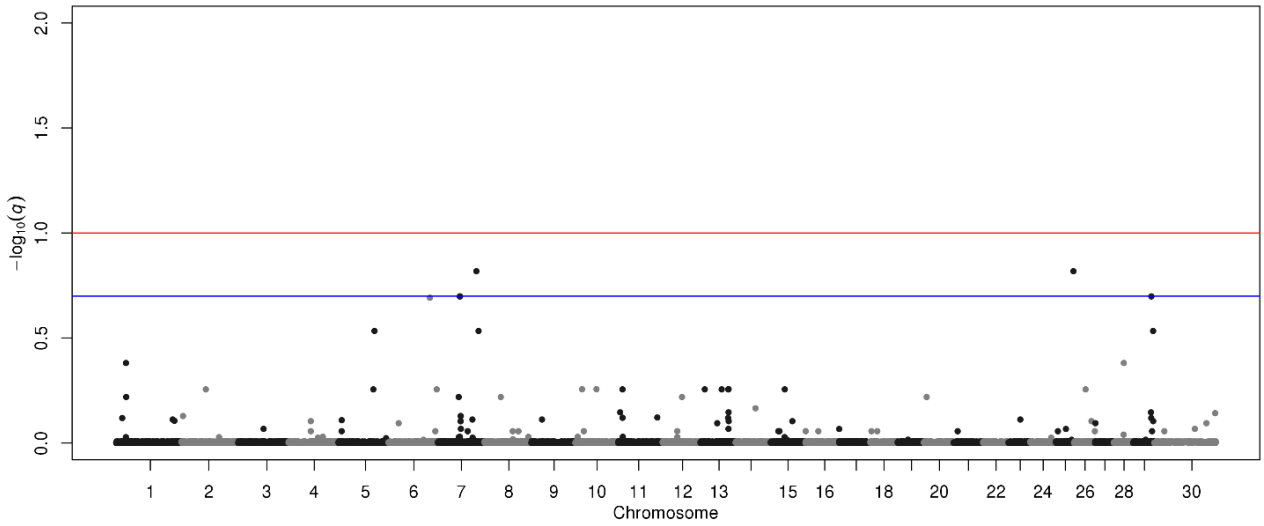

Figure a.75 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB, WBSF and HCW in Hereford.

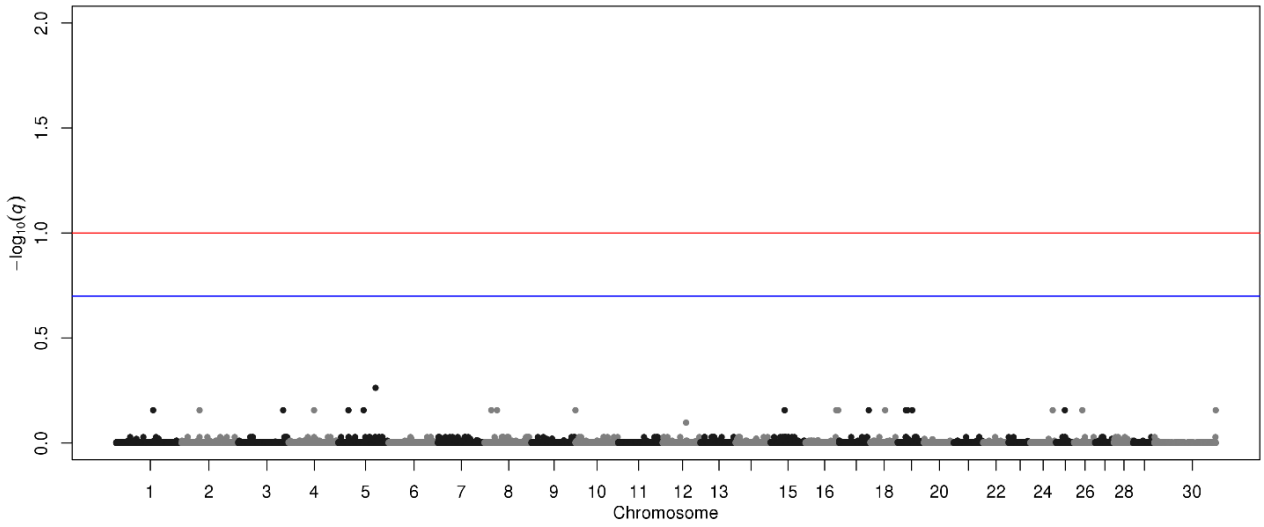

Figure a.76 Manhattan plot of SNP *q*-values estimated in the univariate analysis of WBSF in Limousin.

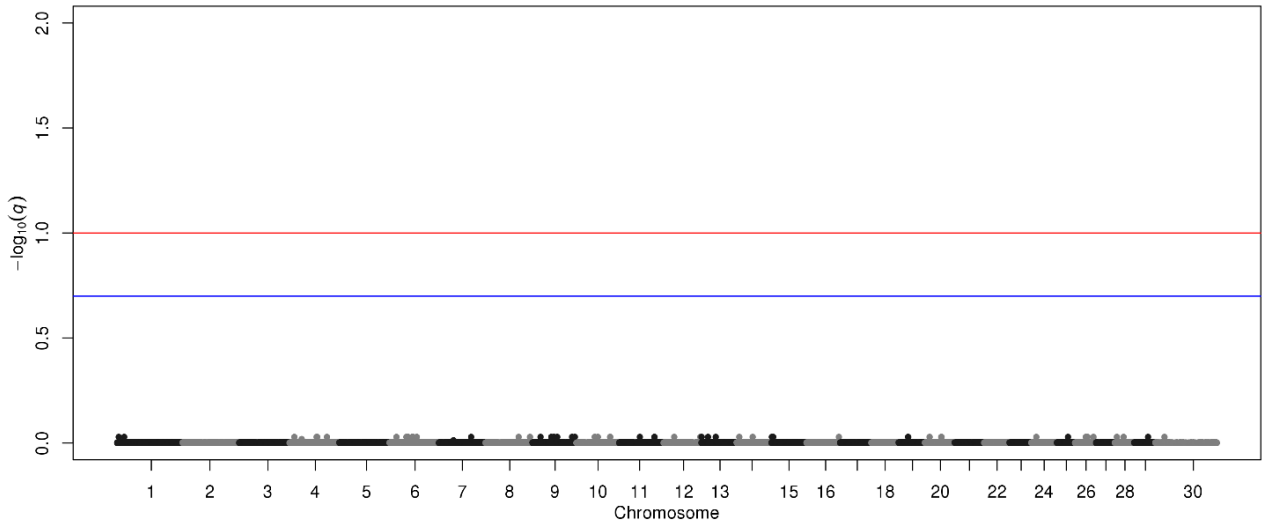

Figure a.77 Manhattan plot of SNP *q*-values estimated in the univariate analysis of CL in Limousin.

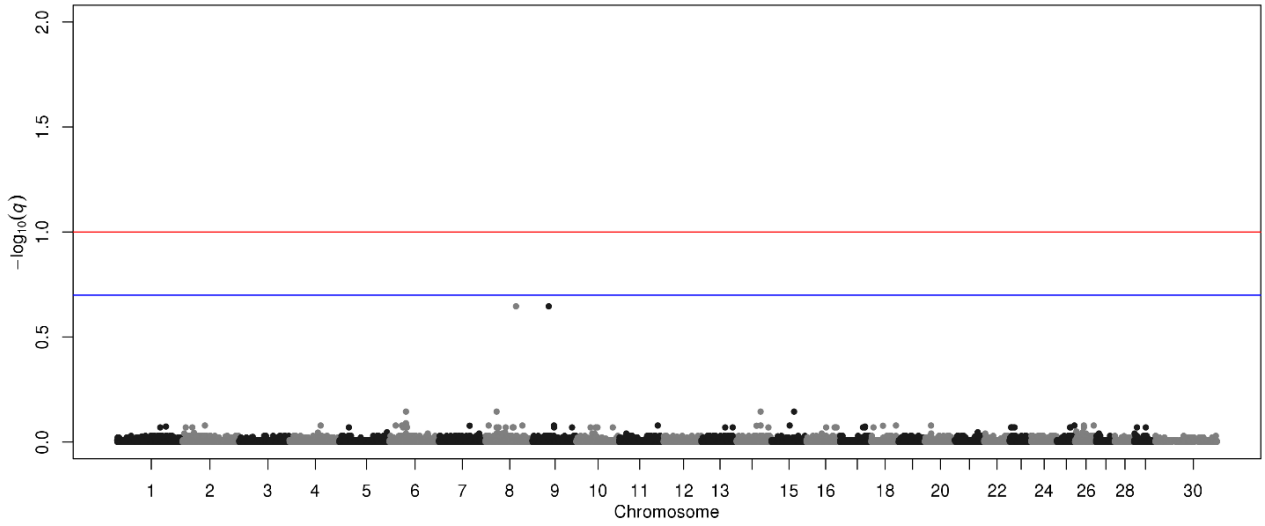

Figure a.78 Manhattan plot of SNP *q*-values estimated in the univariate analysis of HCW in Limousin.

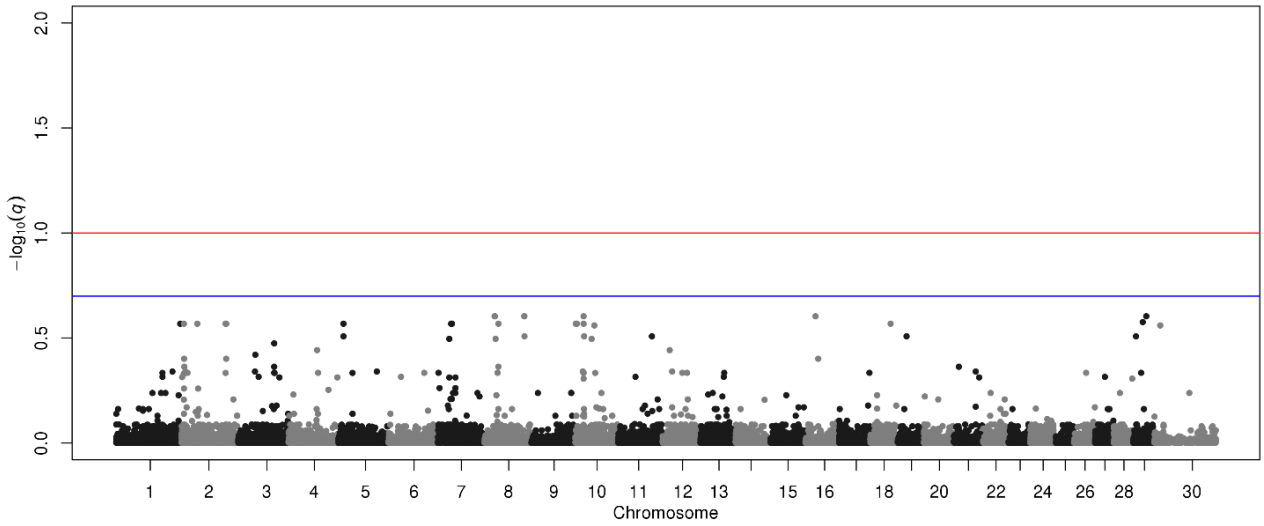

Figure a.79 Manhattan plot of SNP *q*-values estimated in the univariate analysis of FT in Limousin.

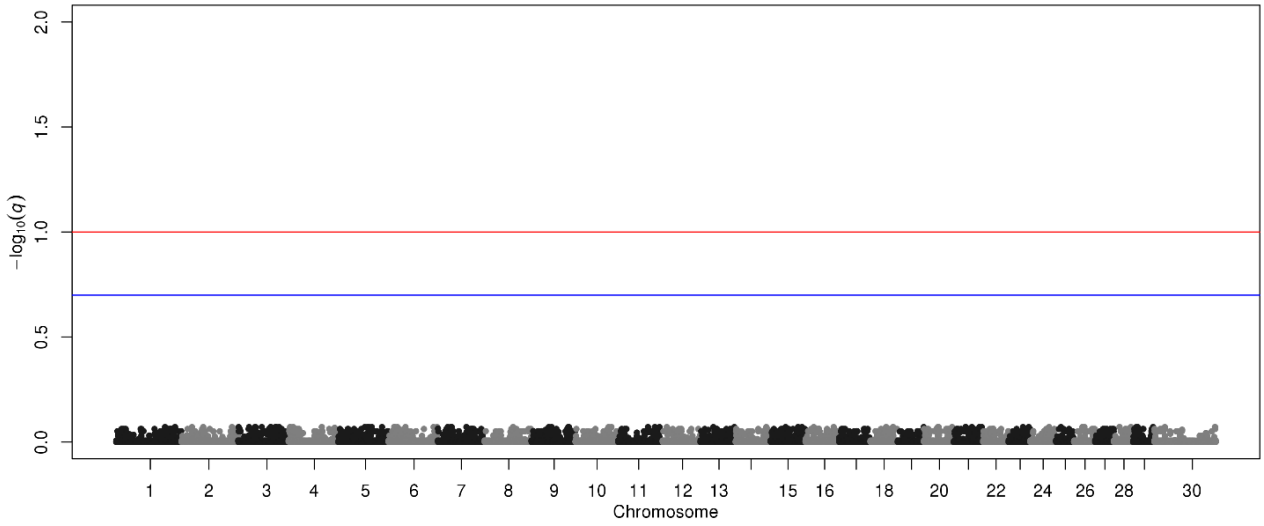

Figure a.80 Manhattan plot of SNP *q*-values estimated in the univariate analysis of KPH in Limousin.

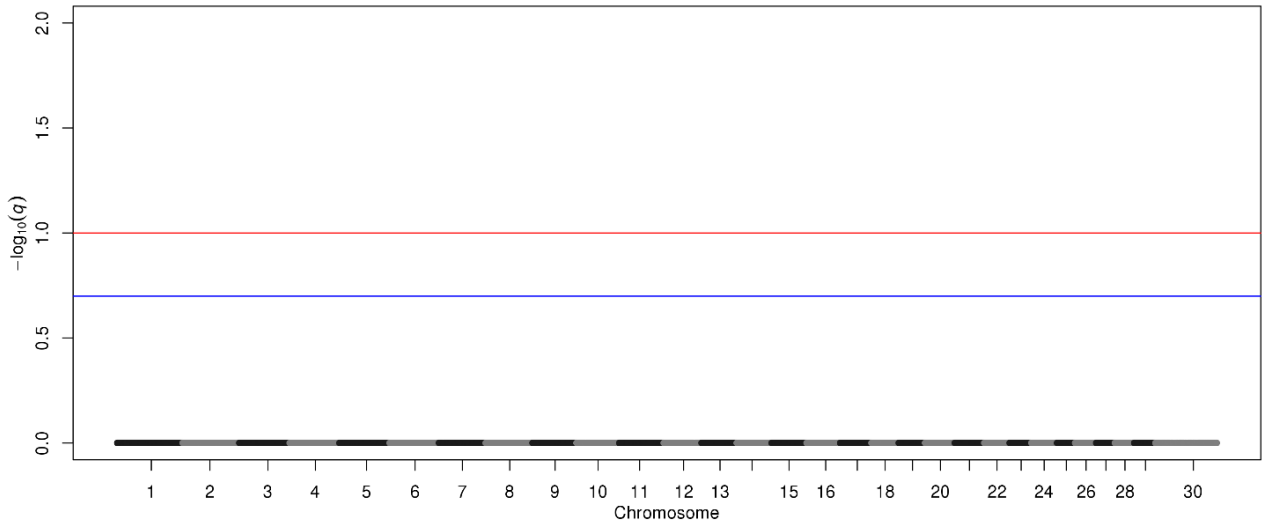

Figure a.81 Manhattan plot of SNP *q*-values estimated in the univariate analysis of FT in Limousin.

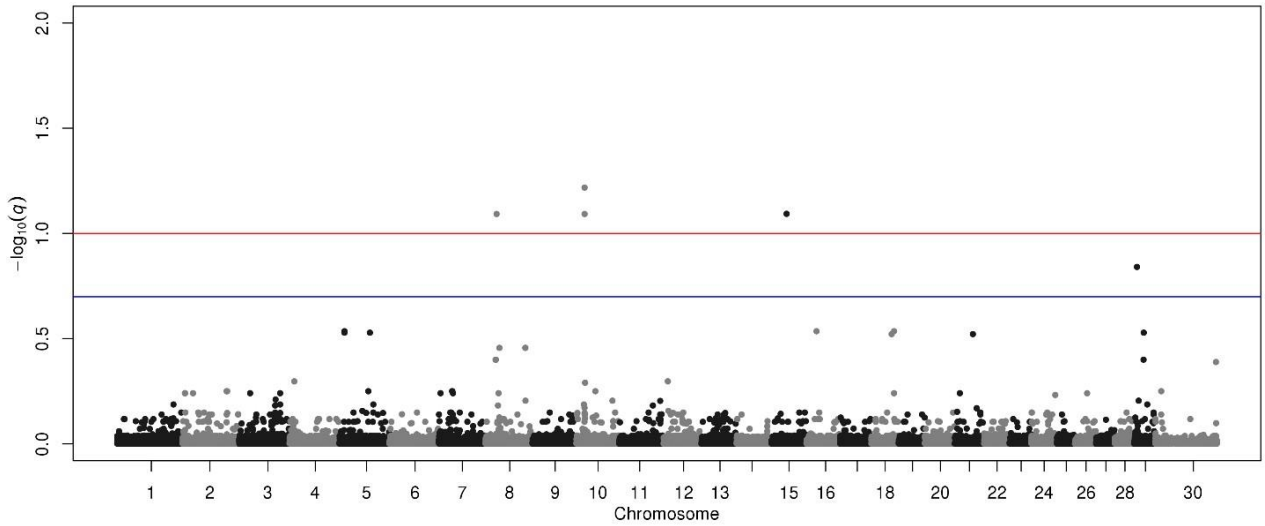

Figure a.82 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of FT and KPH in Limousin.

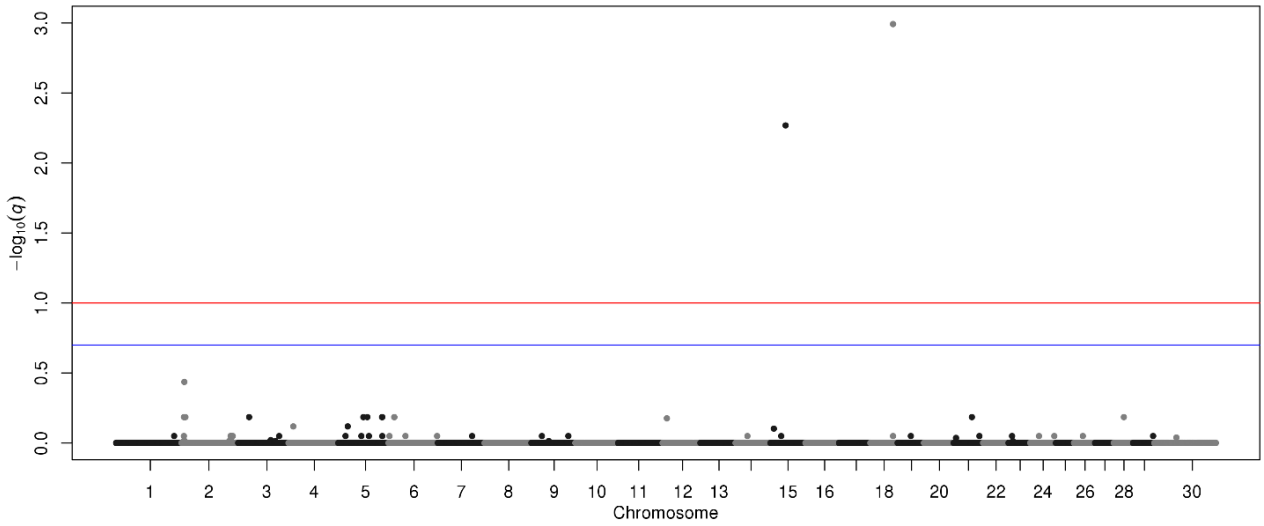

Figure a.83 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of REA and KPH in Limousin.

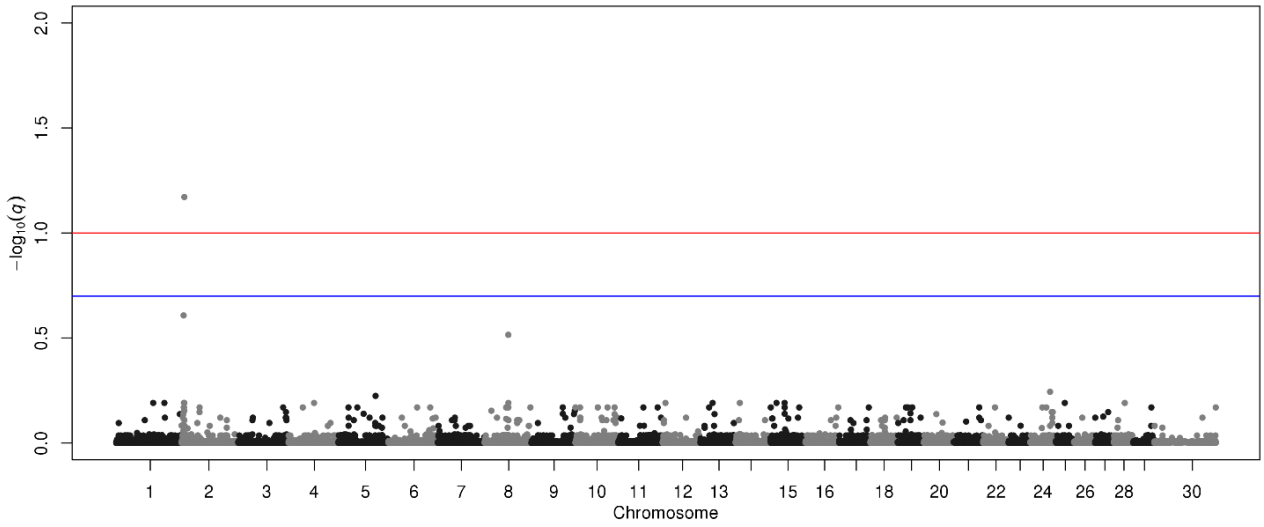

Figure a.84 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and WBSF in Limousin.

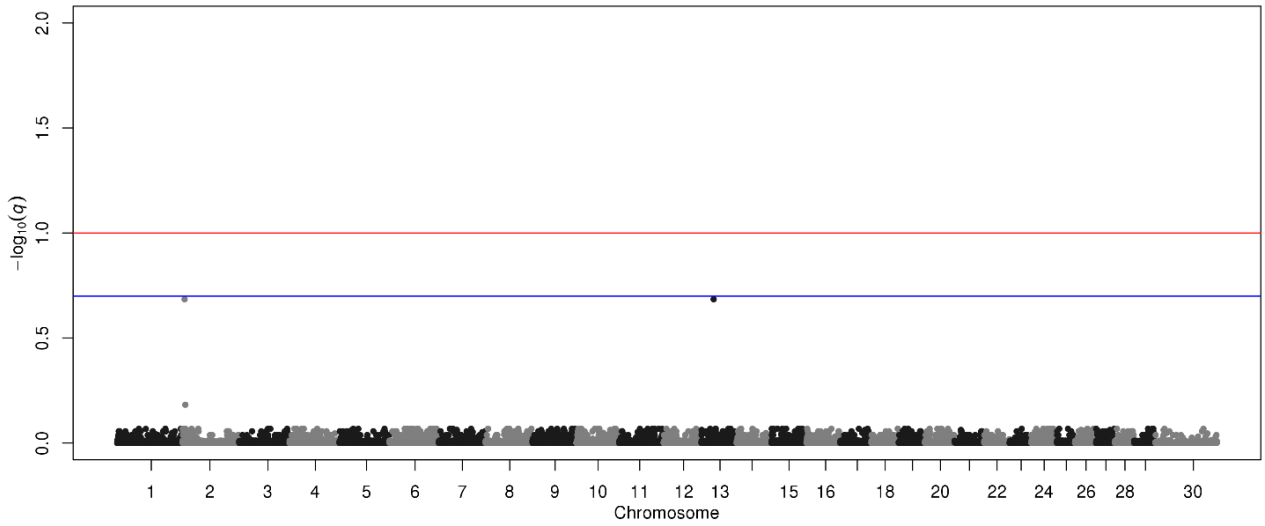

Figure a.85 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and CL in Limousin.

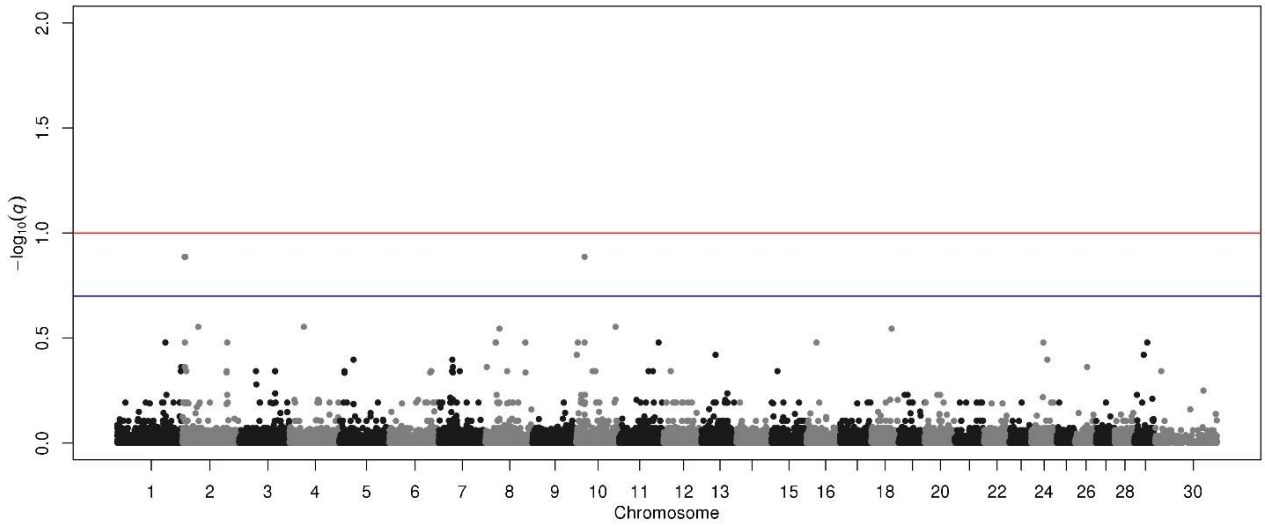

Figure a.86 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and FT in Limousin.

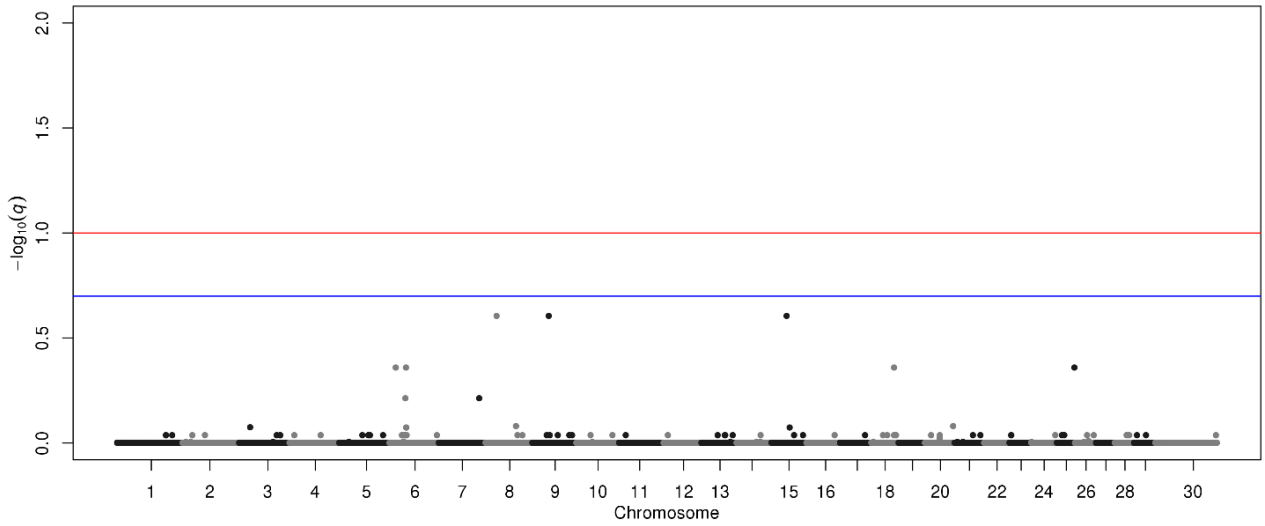

Figure a.87 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW and KPH in Limousin.

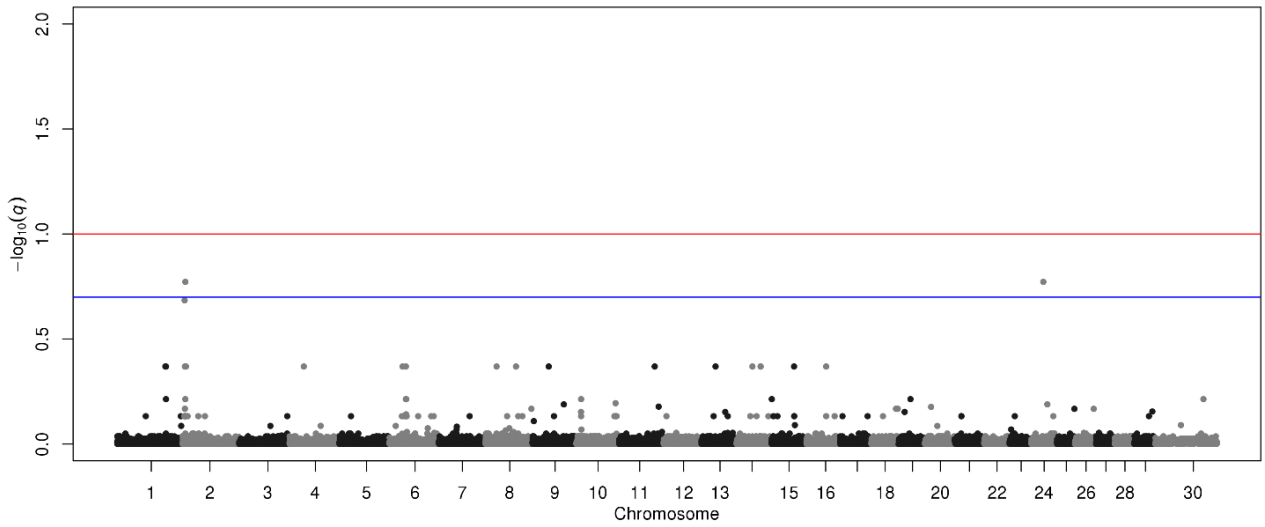

Figure a.88 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB and HCW in Limousin.

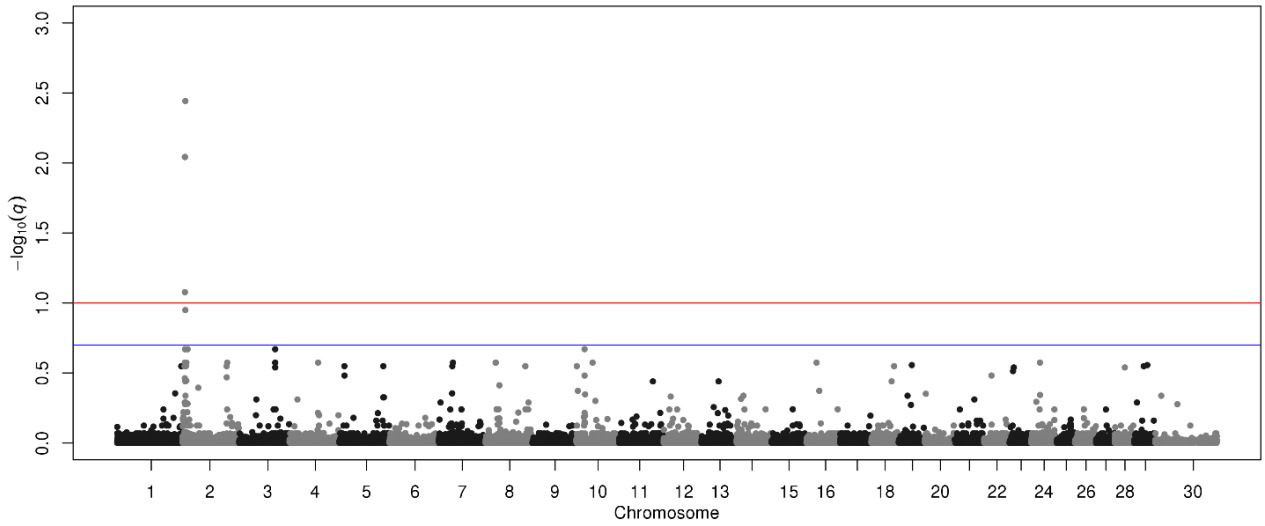

Figure a.89 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of FT and REA in Limousin.

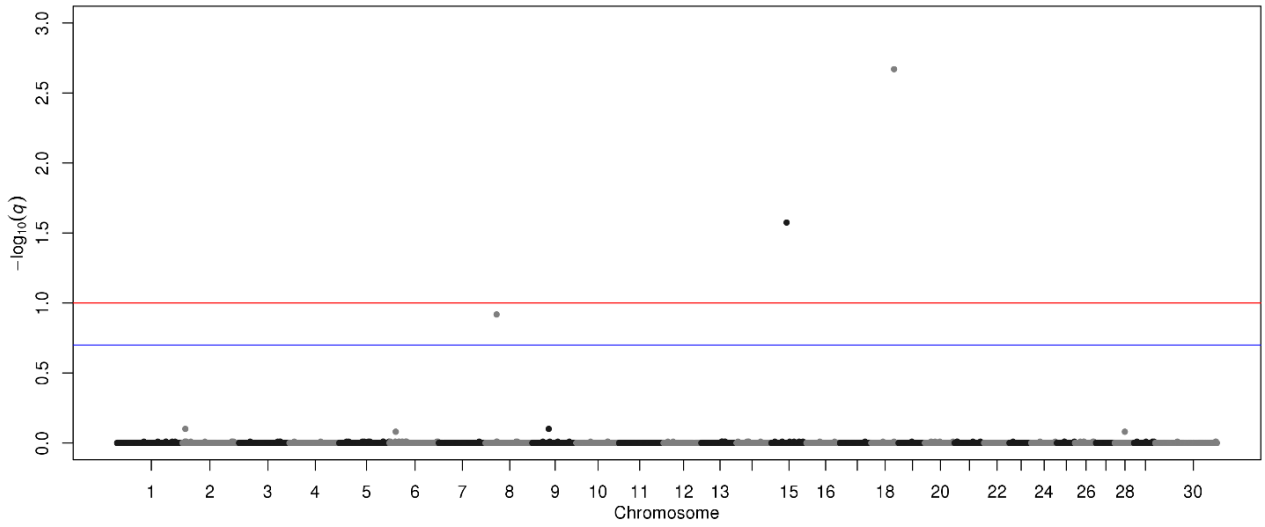

Figure a.90 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, REA, and KPH in Limousin.

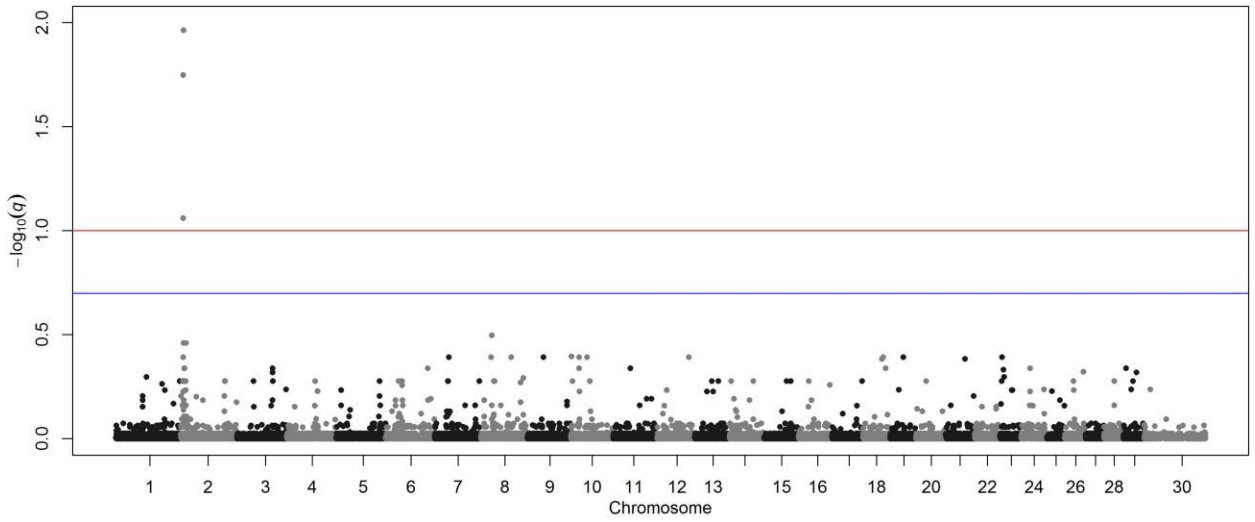

Figure a.91 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT, and REA in Limousin.

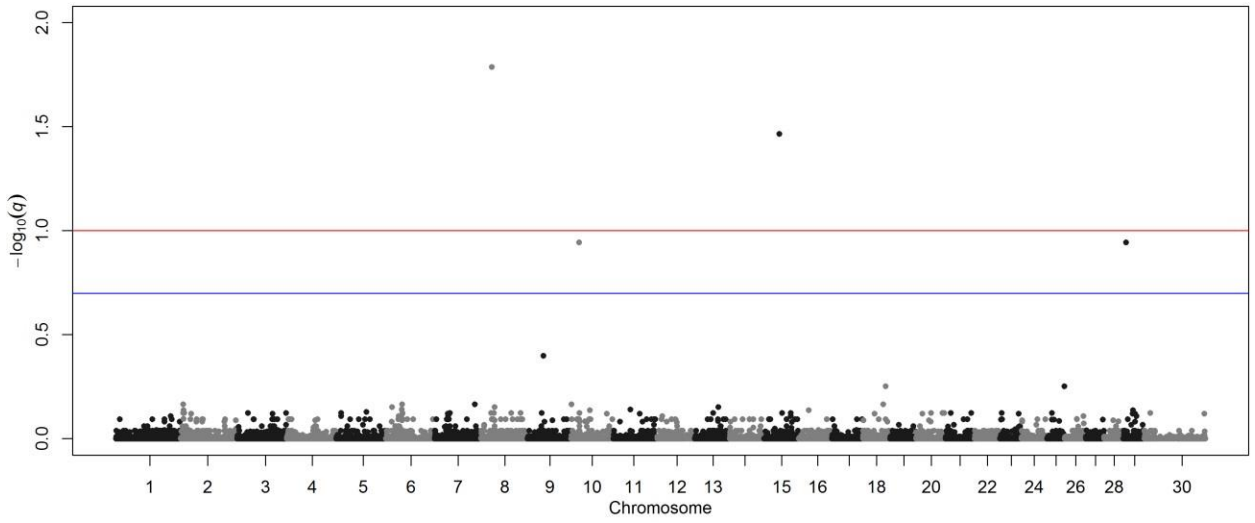

Figure a.92 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT, and KPH in Limousin.

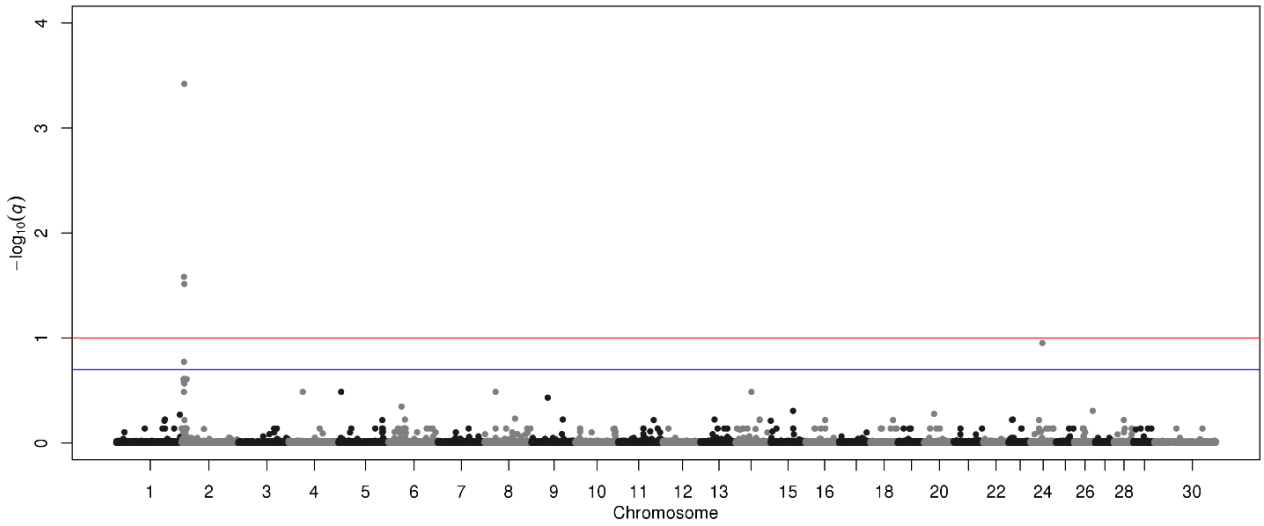

Figure a.93 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB, HCW, and REA in Limousin.

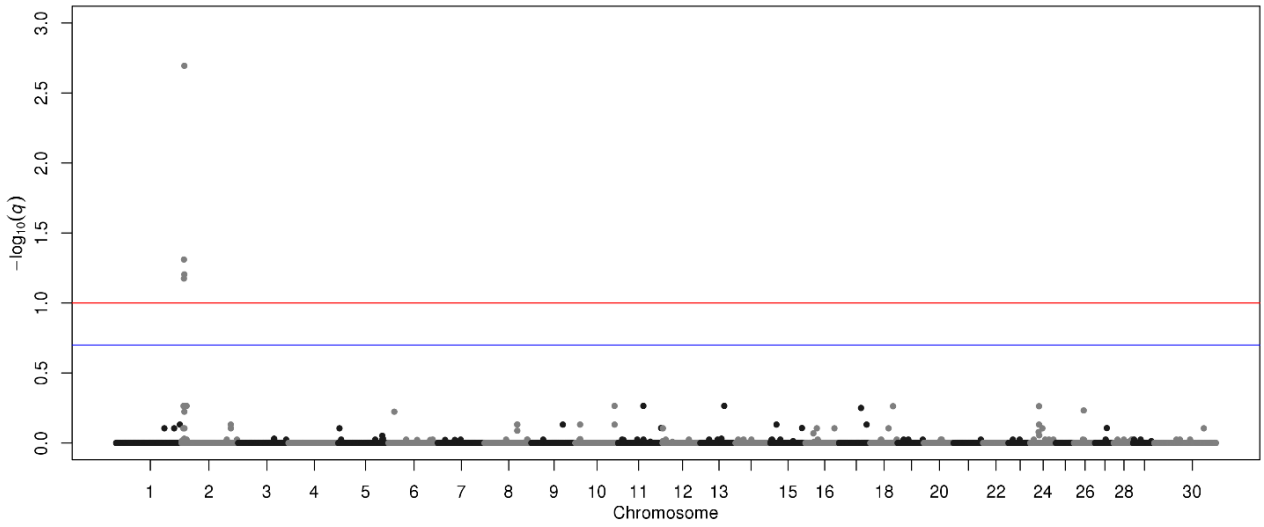

Figure a.94 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB, REA, and IF in Limousin.

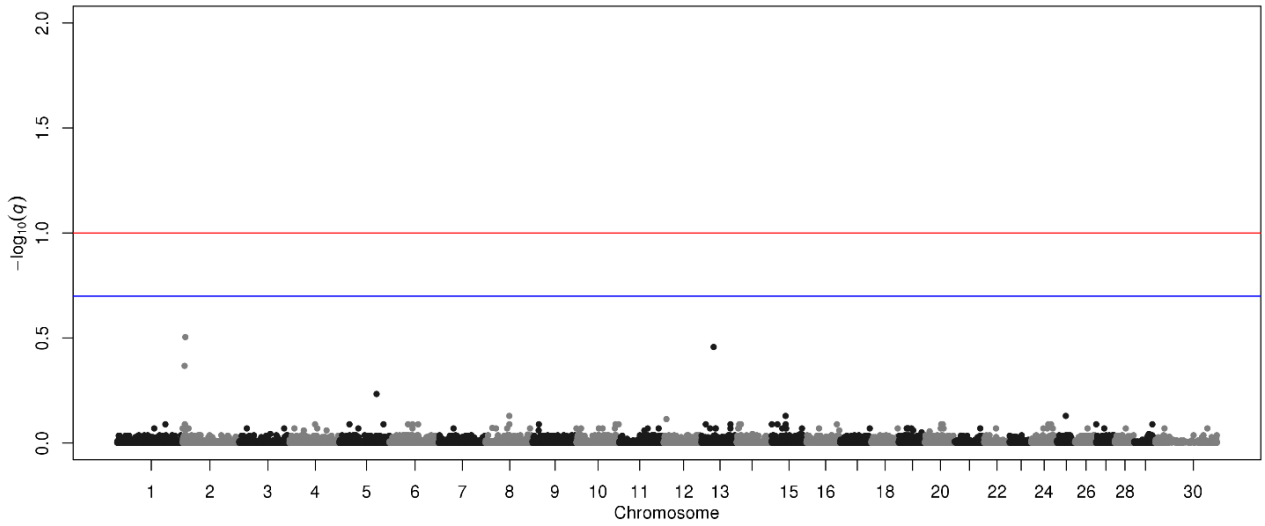

Figure a.95 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of MB, WBSF, and CL in Limousin.

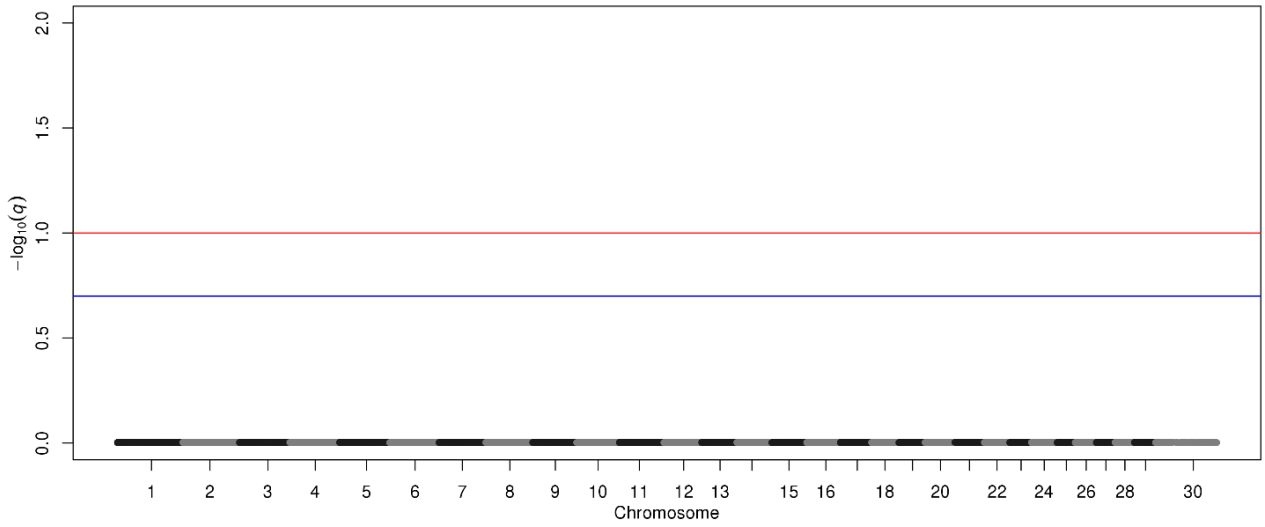

Figure a.96 Manhattan plot of SNP *q*-values estimated in the univariate analysis of MB in Maine-Anjou.

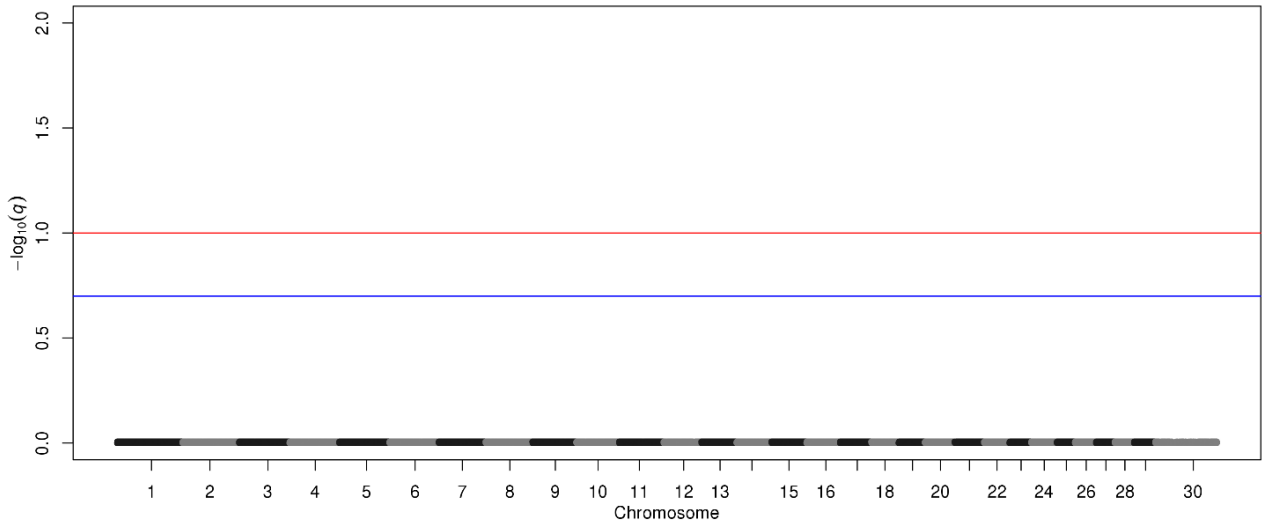

Figure a.97 Manhattan plot of SNP *q*-values estimated in the univariate analysis of WBSF in Maine-Anjou.

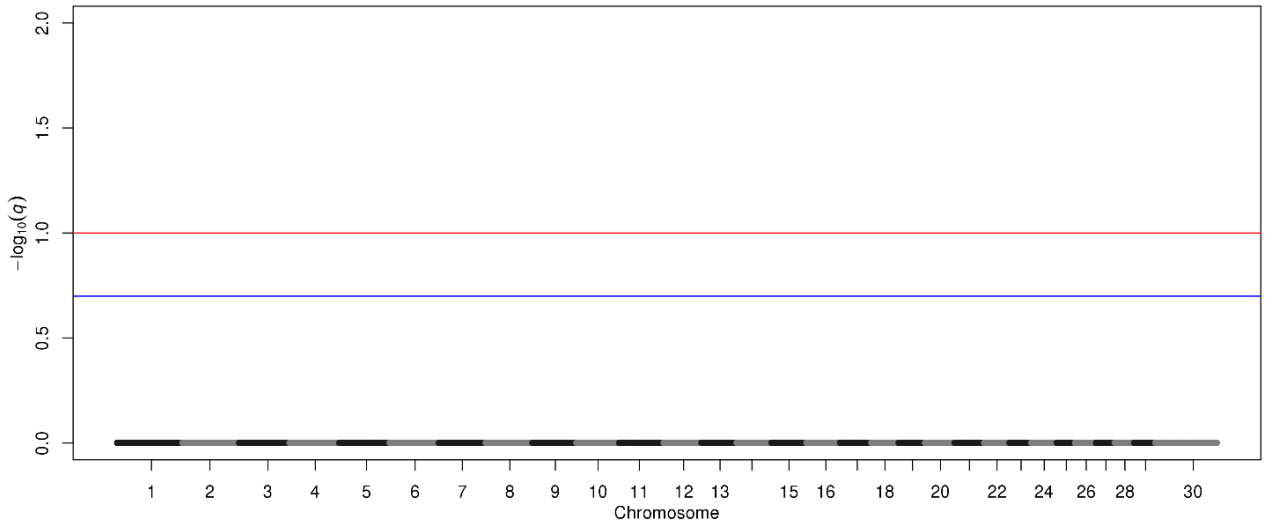

Figure a.98 Manhattan plot of SNP *q*-values estimated in the univariate analysis of CL in Maine-Anjou.

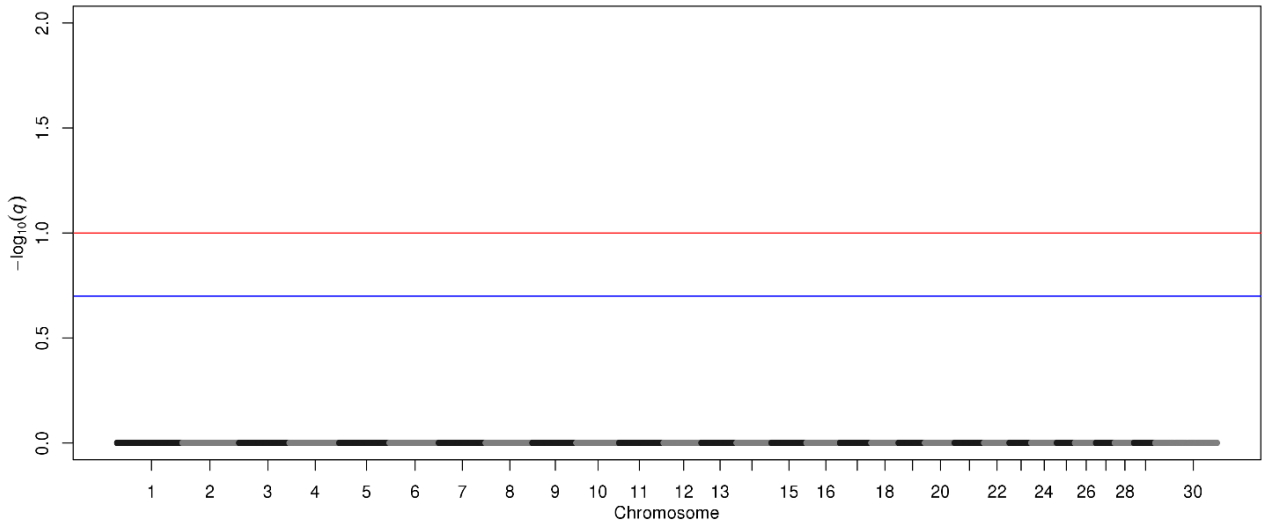

Figure a.99 Manhattan plot of SNP *q*-values estimated in the univariate analysis of HCW in Maine-Anjou.

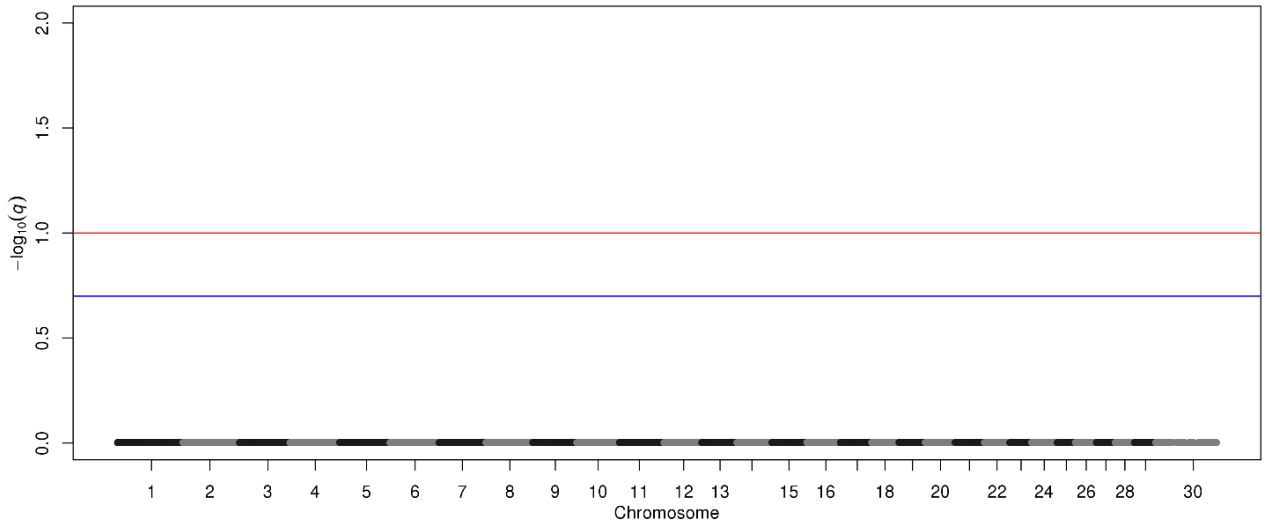

Figure a.100 Manhattan plot of SNP *q*-values estimated in the univariate analysis of FT in Maine-Anjou.

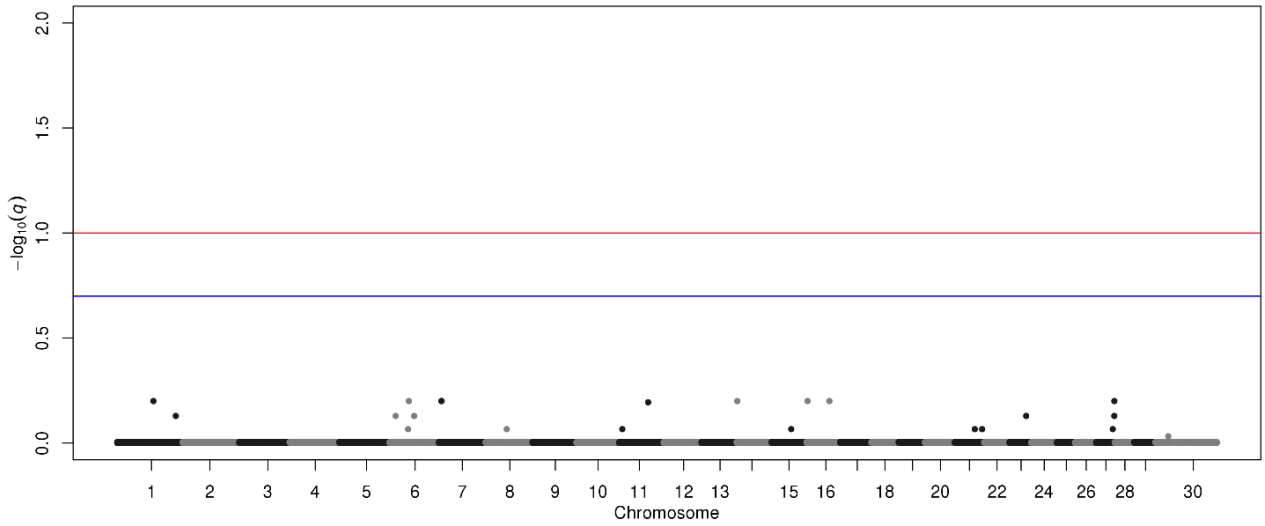

Figure a.101 Manhattan plot of SNP *q*-values estimated in the univariate analysis of REA in Maine-Anjou.

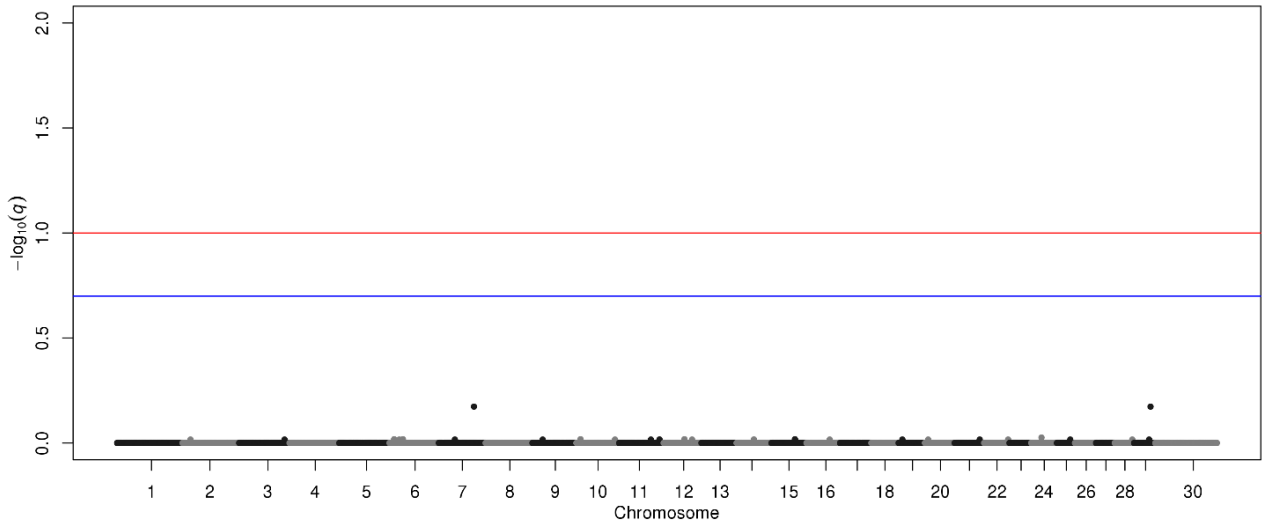

Figure a.102 Manhattan plot of SNP *q*-values estimated in the univariate analysis of KPH in Maine-Anjou.

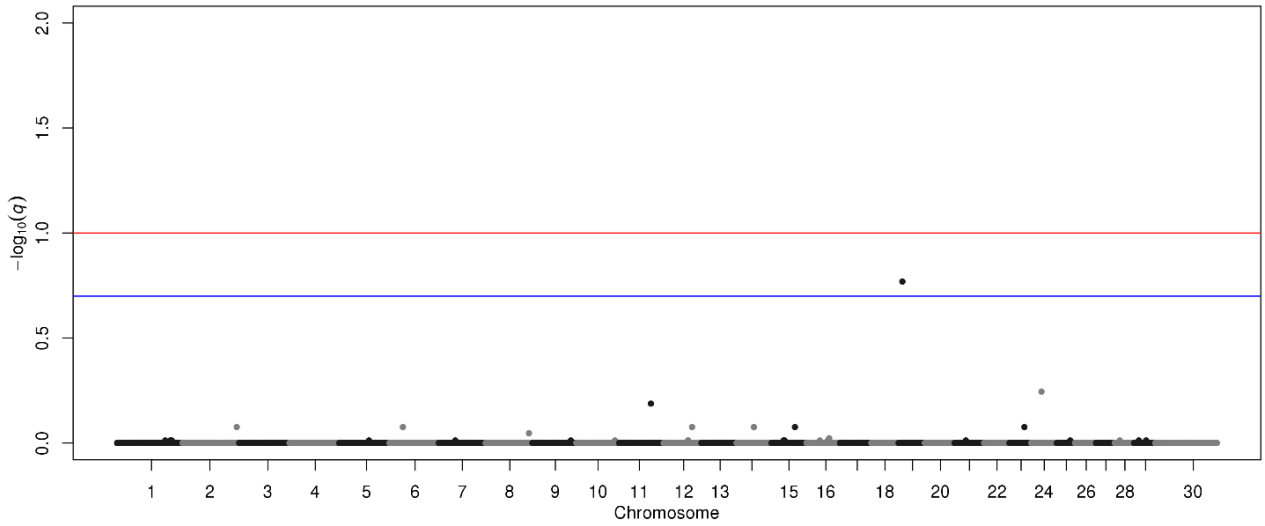

Figure a.103 Manhattan plot of SNP *q*-values estimated in the univariate analysis of IF in Maine-Anjou.

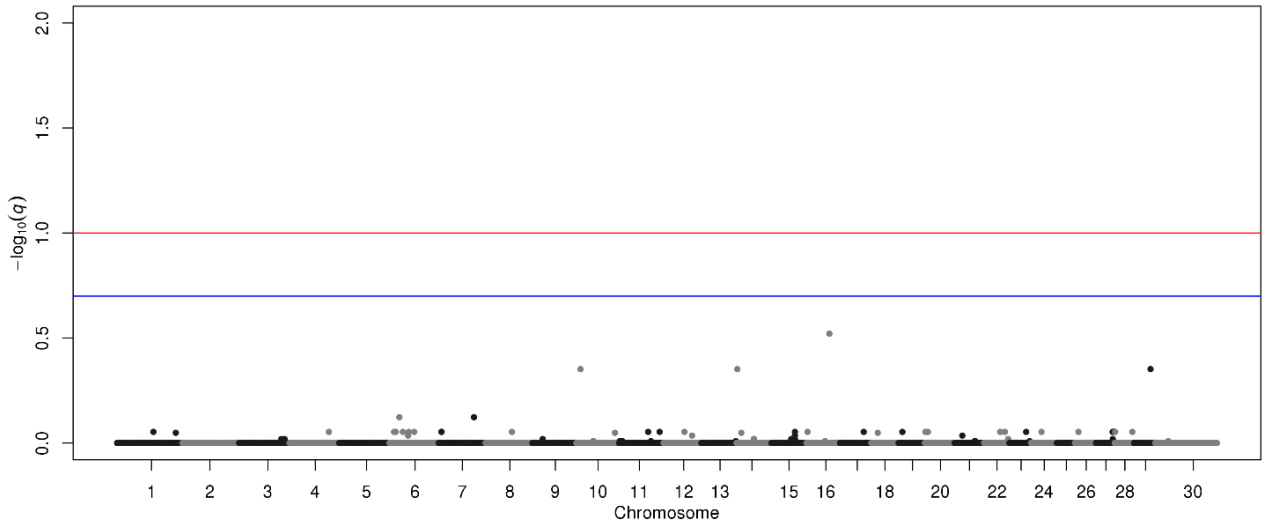

Figure a.104 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of REA and KPH in Maine-Anjou.

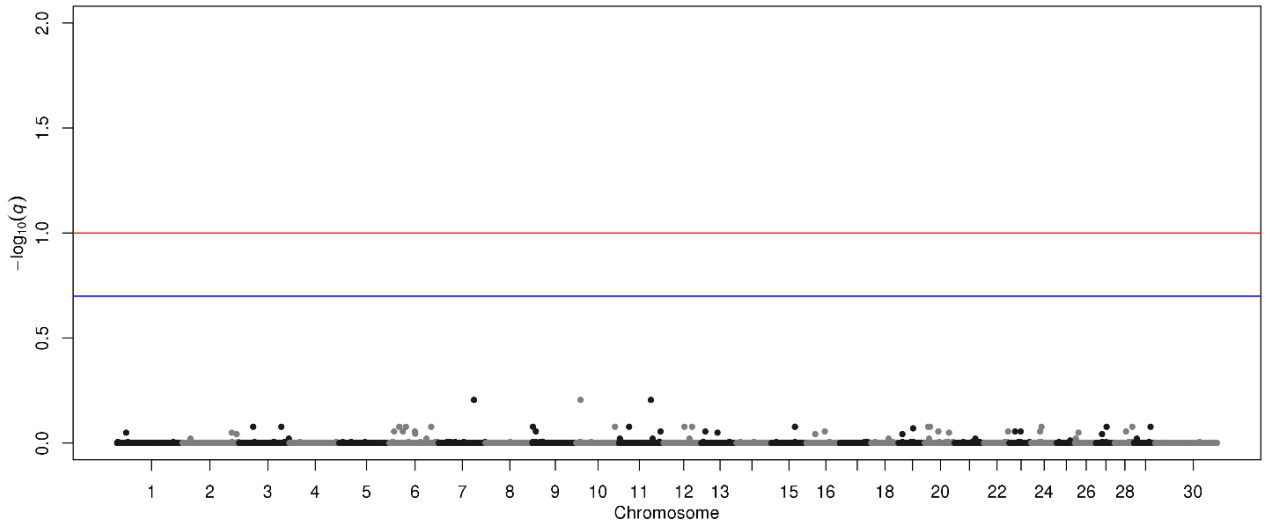

Figure a.105 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of FT and KPH in Maine-Anjou.

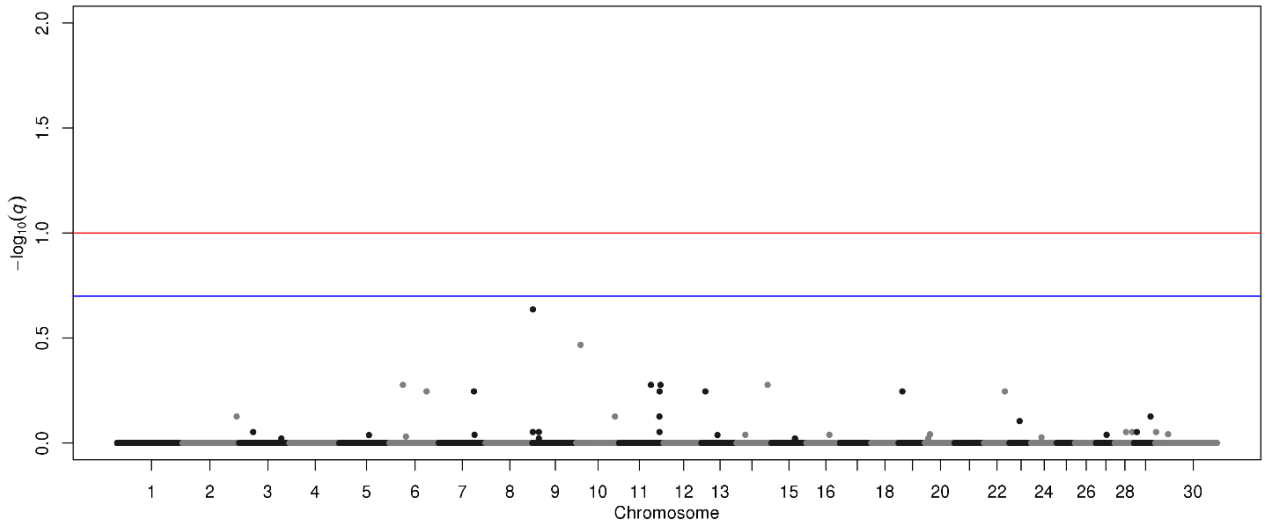

Figure a.106 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT, and KPH in Maine-Anjou.

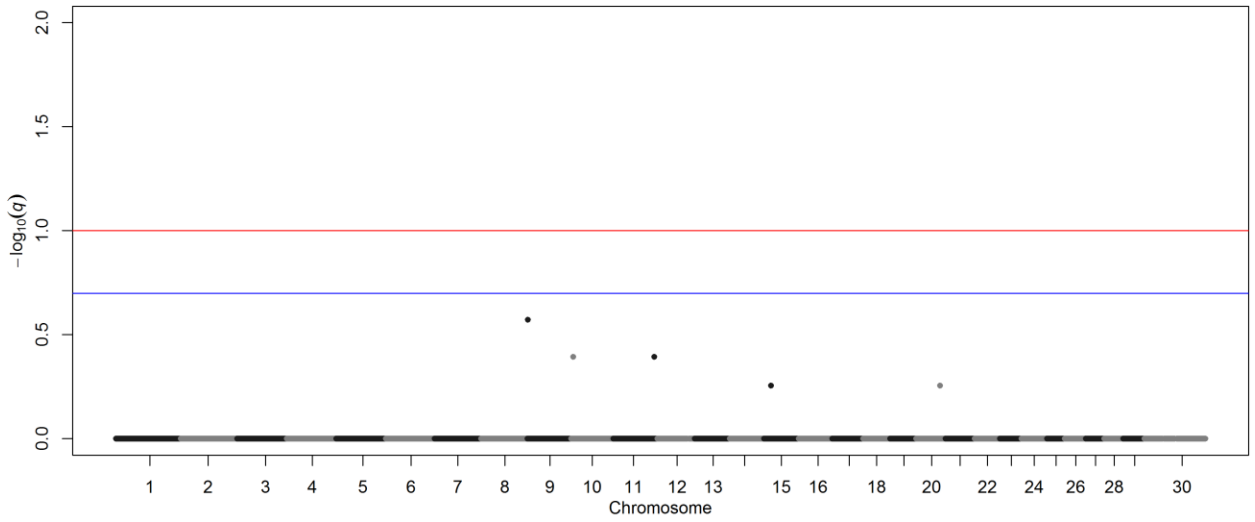

Figure a.107 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT, and REA in Maine-Anjou.

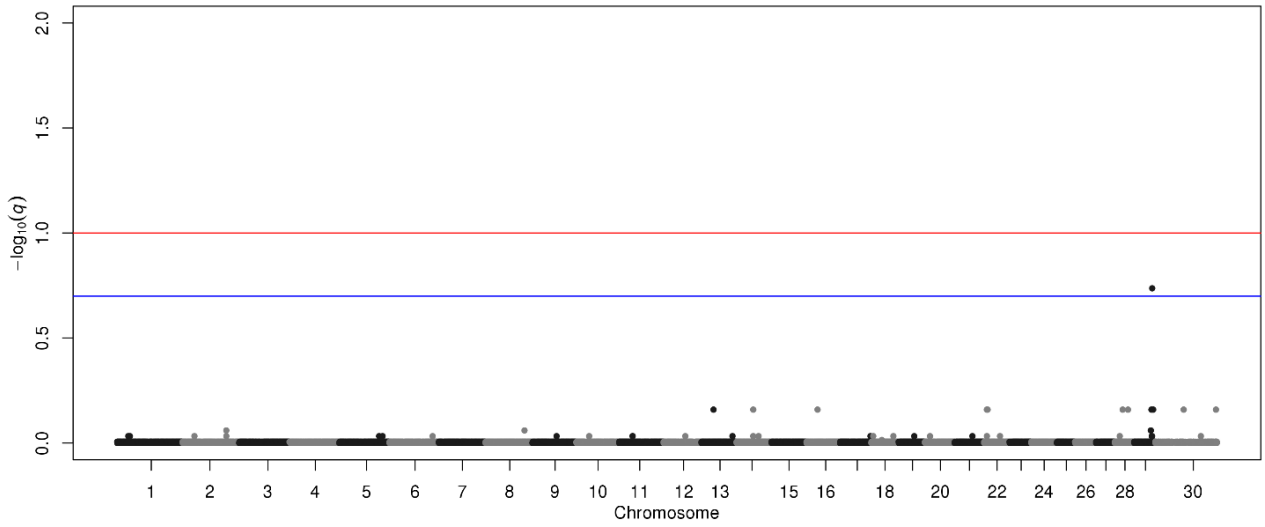

Figure a.108 Manhattan plot of SNP *q*-values estimated in the univariate analysis of WBSF in Simmental.

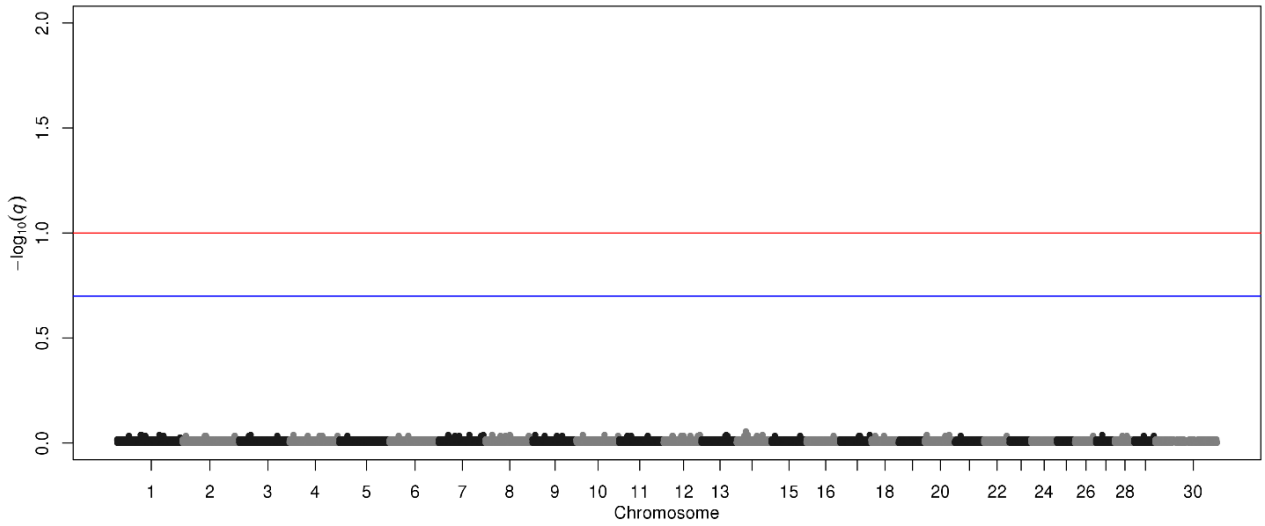

Figure a.109 Manhattan plot of SNP *q*-values estimated in the univariate analysis of CL in Simmental.

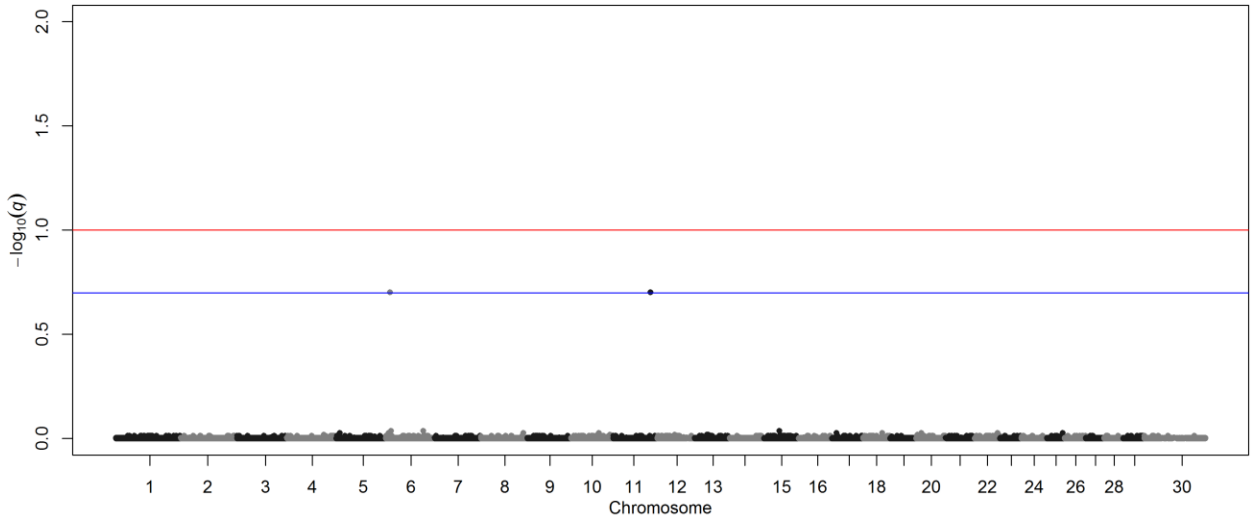

Figure a.110 Manhattan plot of SNP *q*-values estimated in the univariate analysis of HCW in Simmental.

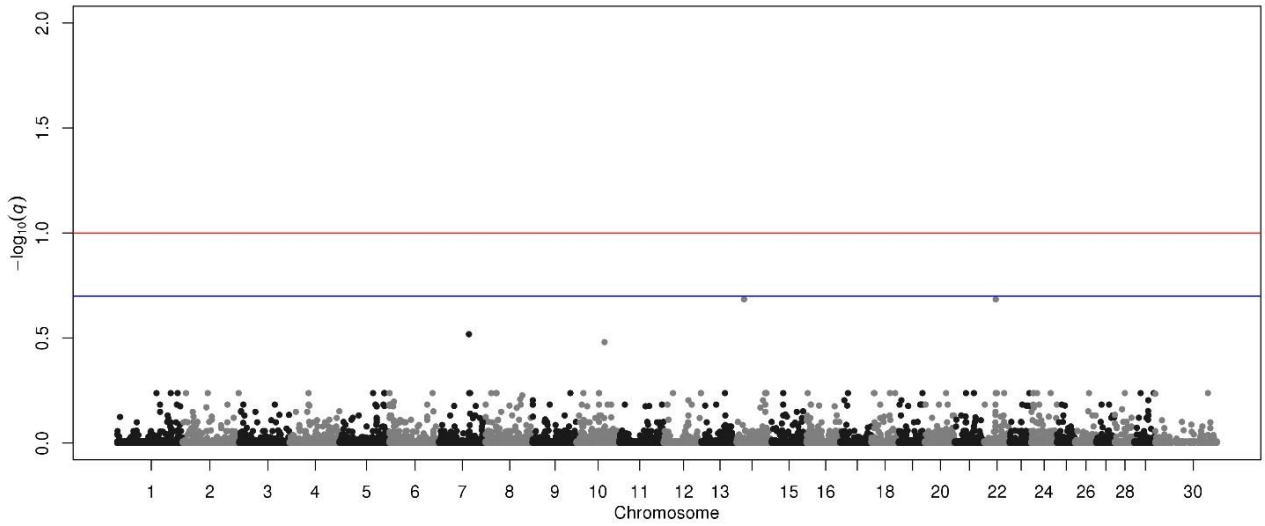

Figure a.111 Manhattan plot of SNP *q*-values estimated in the univariate analysis of FT in Simmental.

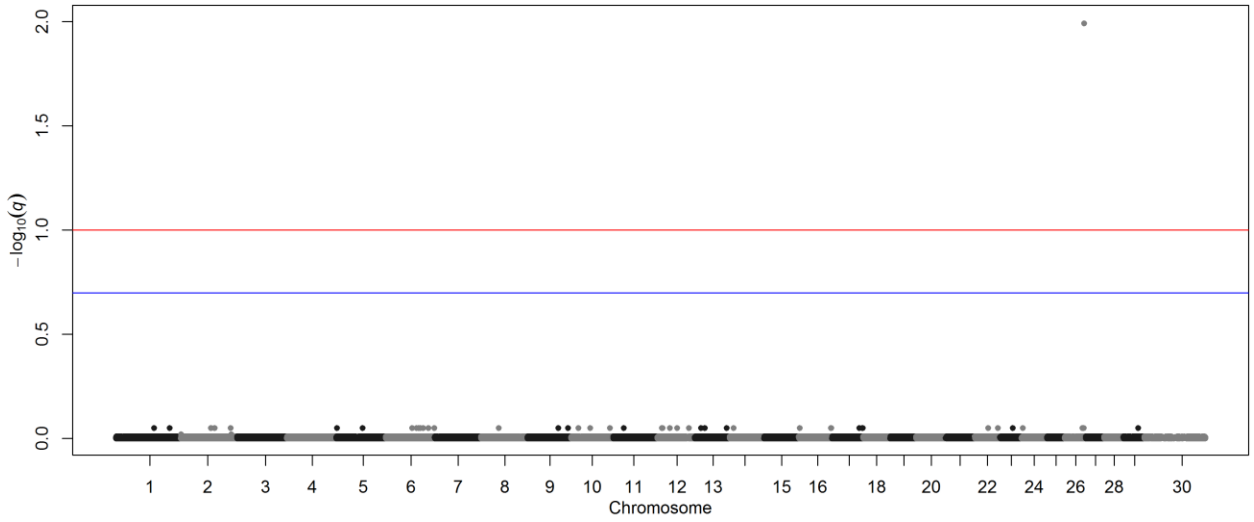

Figure a.112 Manhattan plot of SNP *q*-values estimated in the univariate analysis of REA in Simmental.

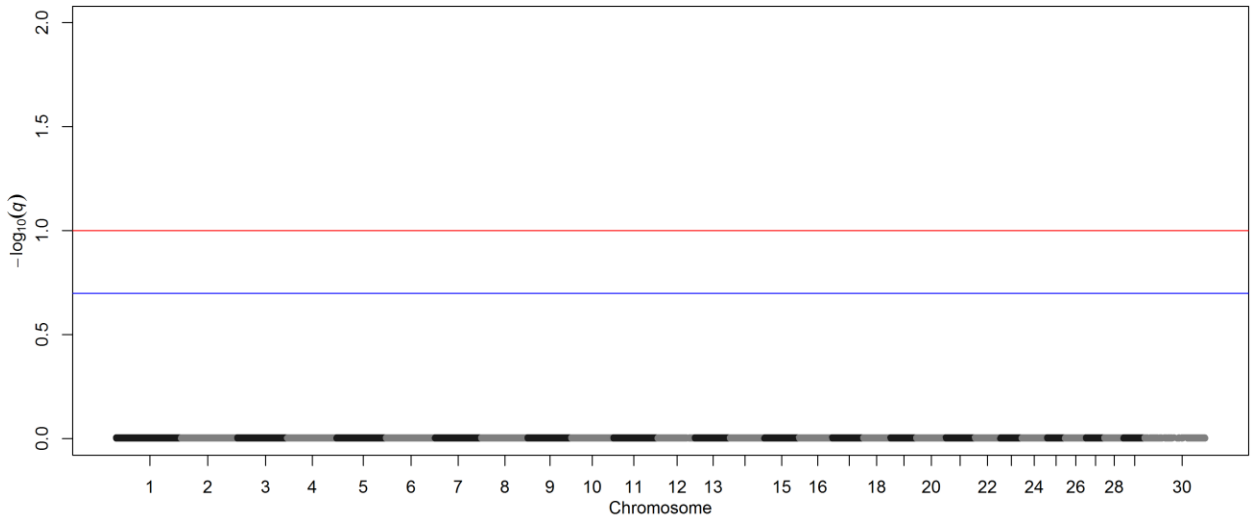

Figure a.113 Manhattan plot of SNP *q*-values estimated in the univariate analysis of KPH in Simmental.

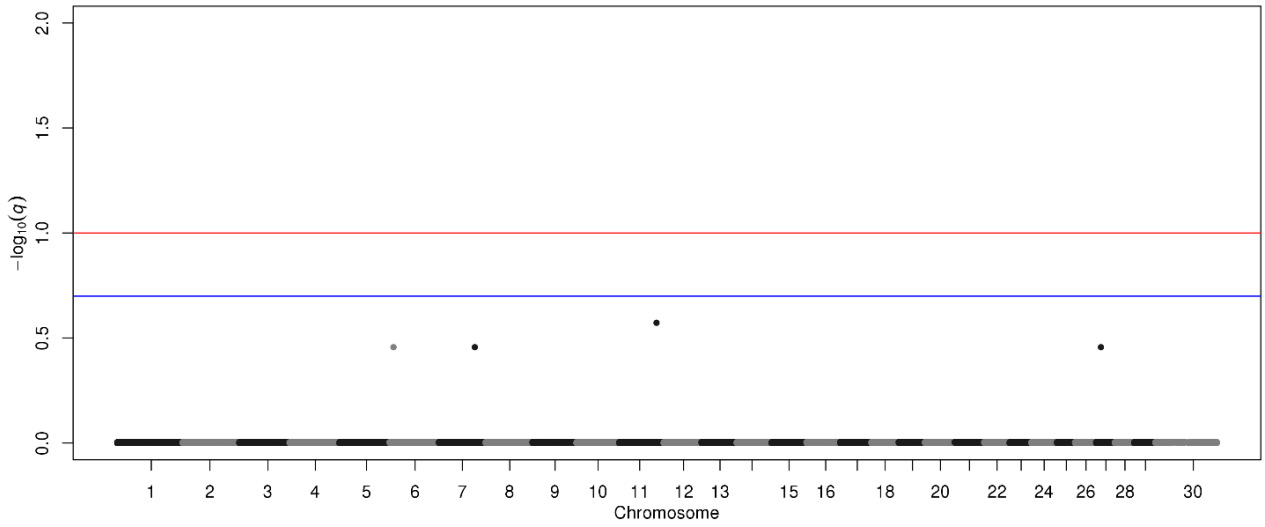

Figure a.114 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW and KPH in Simmental.

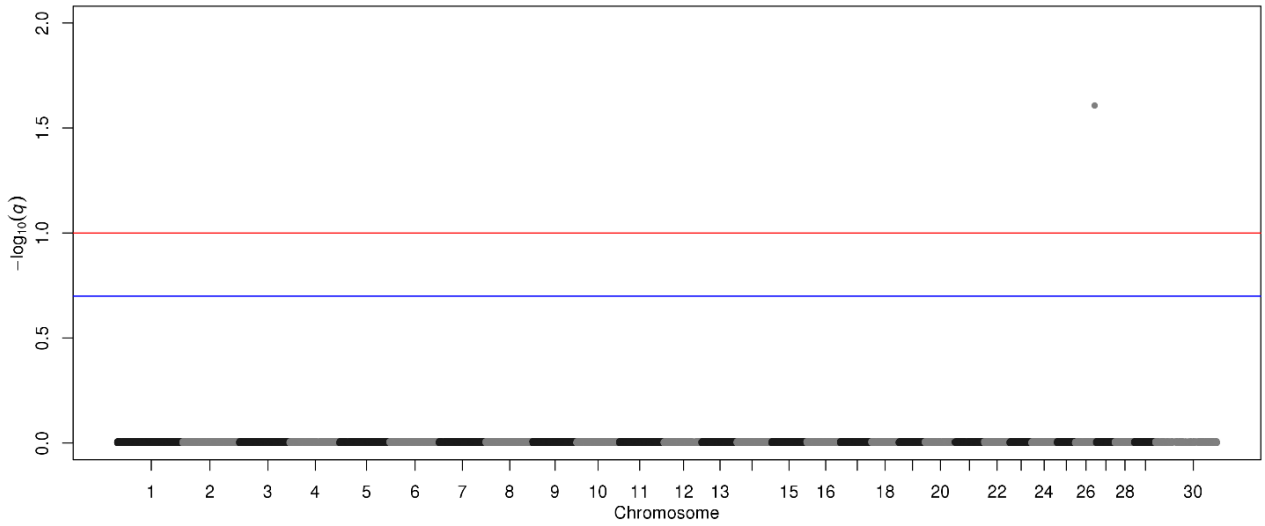

Figure a.115 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of REA and KPH in Simmental.

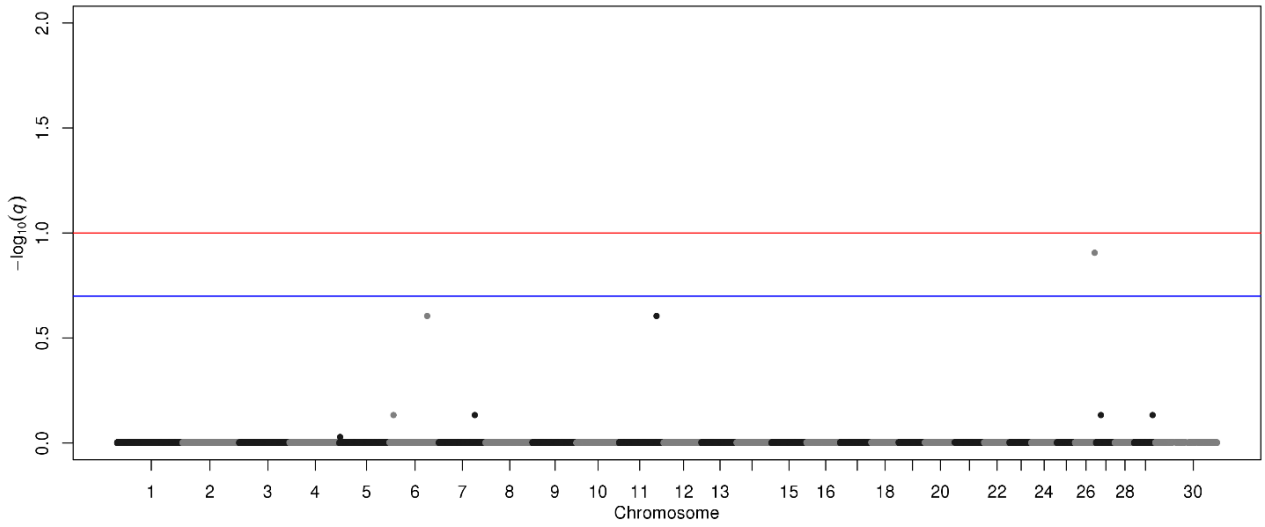

Figure a.116 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, REA, and KPH in Simmental.

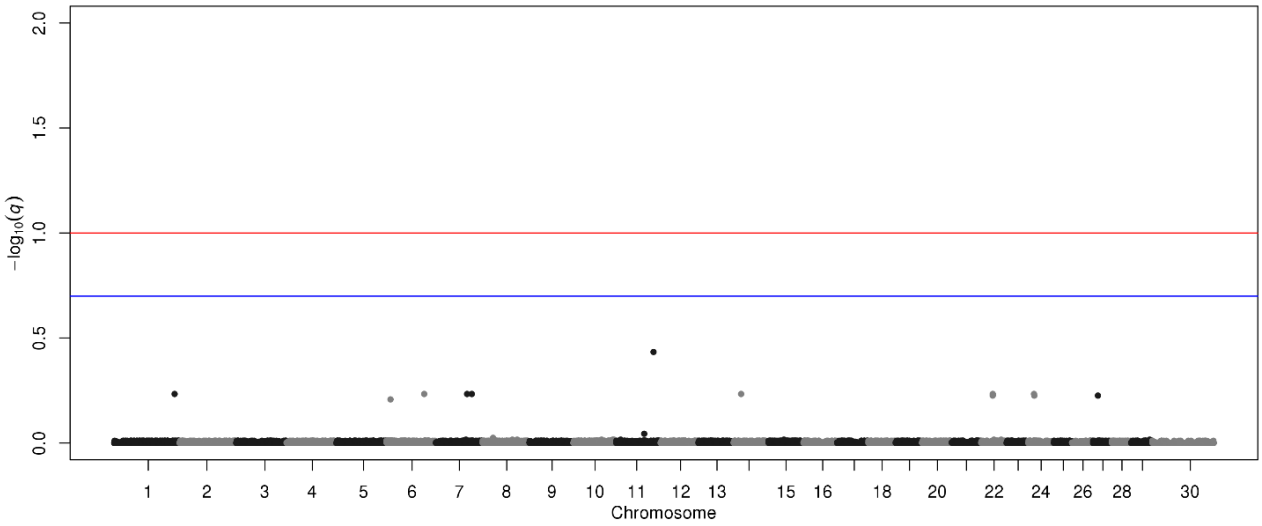

Figure a.117 Manhattan plot of SNP *q*-values estimated in the multivariate analysis of HCW, FT, and KPH in Simmental.

LITERATURE CITED

Alexander, L.J. et al., 2009. A Limousin specific myostatin allele affects longissimus muscle area and fatty acid profiles in a Wagyu-Limousin F2 population. *Journal of animal science*, 87(5), pp.1576–1581.

Anon, bok%3A978-1-62703-447-0.

Aulchenko, Y.S. et al., 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10), pp.1294–1296.

Beermann, D.H., 2002. Beta-Adrenergic receptor agonist modulation of skeletal muscle growth. *Distribution*, 80(E-Suppl_1), pp.E18–23. Available at: http://jas.fass.org/cgi/content/abstract/80/E-Suppl_1/E18.

Bensaad, K. et al., 2006. TIGAR, a p53-Inducible Regulator of Glycolysis and Apoptosis. *Cell*, 126(1), pp.107–120.

Berger, S. et al., 2015. Effectiveness of Shrinkage and Variable Selection Methods for the Prediction of Complex Human Traits using Data from Distantly Related Individuals. *Annals of Human Genetics*, 79(2), pp.122–135.

Boichard, D. et al., 2012. Genomic selection in French dairy cattle. *Animal Production Science*.

Bolormaa, S. et al., 2014. A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. *PLoS Genetics*, 10(3).

Breuza, L. et al., 2004. Proteomics of endoplasmic reticulum-golgi intermediate compartment (ERGIC) membranes from brefeldin A-treated HepG2 cells identifies ERGIC-32, a new cycling protein that interacts with human Erv46. *Journal of Biological Chemistry*, 279(45), pp.47242–47253.

Browning, S.R. & Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, 81(5), pp.1084–97.

Button, K.S. et al., 2013. Power failure: why small sample size undermines the

reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5), pp.365–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23571845.

Charan, J. & Kantharia, N., 2013. How to calculate sample size in animal studies? *Journal of Pharmacology and Pharmacotherapeutics*, 4(4), p.303.

Cuyabano, B.C., Su, G. & Lund, M.S., 2015. Ge n e t i c s Selection Evolution Selection of haplotype variables from a high-density marker map for genomic prediction. *Genetics Selection Evolution*, 47.

Cuyabano, B.C., Su, G. & Lund, M.S., 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics*, 15. Available at: http://www.biomedcentral.com/1471-2164/15/1171.

Decker, J.E., 2015. Agricultural Genomics: Commercial Applications Bring Increased Basic Research Power G. Gibson, ed. *PLoS Genetics*, 11(11), p.3.

Devon, S., 1998. Carcass Merit Project: Dna Marker Validation. , pp.1–21.

Dikeman, M.E. et al., 2005. Phenotypic ranges and relationships among carcass and meat palatability traits for fourteen cattle breeds, and heritabilities and expected progeny differences for Warner-Bratzler shear force in three beef cattle breeds. *Journal of Animal Science*, 83(10), pp.2461–2467.

Fortes, M.R.S. et al., 2013. Evidence for pleiotropism and recent selection in the PLAG1 region in Australian Beef cattle. *Animal Genetics*, 44(6), pp.636–647.

Gianola, D. et al., 2009. Additive genetic variability and the Bayesian alphabet. *Genetics*.

Habier, D. et al., Extension of the bayesian alphabet for genomic selection.

Habier, D. et al., 2011. Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1), p.186.

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

Hayes, B.J. et al., 2010. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein

cattle as contrasting model traits. *PLoS genetics*, 6(9), p.e1001139.

Hulsman Hanna, L.L. et al., 2014. Comparison of breeding value prediction for two traits in a Nellore-Angus crossbred population using different bayesian modeling methodologies. *Genetics and Molecular Biology*.

Kachman, S.D., 2008. AN INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS Stephen D . Kachman. *Statistics*, 24, pp.59–73. Available at: http://armyconference.org/ACAS2003CD/ACAS2003/McCullochCharles/mcculloch.pdf.

Kachman, S.D. et al., 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genetics, Selection, Evolution : GSE*, 45(1), p.30.

Kang, H.M. et al., 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4), pp.348–354. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=mnh&AN=20208533&lang=fr&site=ehost-live.

Karim, L. et al., 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nature genetics*, 43(5), pp.405–413.

Korte, A. et al., 2012. technical reports A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Publishing Group*, 44(9), pp.1066–1071. Available at: http://dx.doi.org/10.1038/ng.2376.

Legarra, a, Aguilar, I. & Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. *Journal of dairy science*, 92(9), pp.4656–4663. Available at: http://dx.doi.org/10.3168/jds.2009-2061.

de Los Campos, G., Vazquez, A.I., et al., 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS genetics*, 9(7), p.e1003608.

de Los Campos, G., Hickey, J.M., et al., 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2), pp.327–45.

de los Campos, G., Sorensen, D. & Gianola, D., 2015. Genomic heritability: what

is it? *PLoS genetics*, 11(5), p.e1005048. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=25942577&retmode=ref&cmd=prlinks.

Lynch, M., Walsh, B. & others, 1998. *Genetics and analysis of quantitative traits*, Sinauer Sunderland, MA.

Makowsky, R. et al., 2011. Beyond missing heritability: Prediction of complex traits. *PLoS Genetics*, 7(4).

McClure, M.C. et al., 2012. Genome-wide association analysis for quantitative trait loci influencing Warner-Bratzler shear force in five taurine cattle breeds. *Animal Genetics*, 43(6), pp.662–673.

Mersmann, H.J., 1998. Overview of the Effects of ??-Adrenergic Receptor Agonists on Animal Growth Including Mechanisms of Action. *Journal of Animal Science*, 76(1), pp.160–172.

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

Miar, Y. et al., 2013. Estimation of Genetic and Phenotypic Parameters for Ultrasound and Carcass Merit Traits in Crossbred Beef Cattle. *Canadian Journal of Animal Science*, 94, pp.1–8. Available at: http://pubs.aic.ca/doi/abs/10.4141/CJAS2013-115.

Misztal, I., Aggrey, S.E. & Muir, W.M., 2013. Experiences with a single-step genome evaluation. *Poultry science*, 92(9), pp.2530–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23960138.

Moghaddar, N., Swan, A.A. & Hj Van Der Werf, J., 2014. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. , 46, pp.1–10.

Nishimura, S. et al., 2012. Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. *BMC genetics*, 13(1), p.40. Available at: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-13-40.

Patwari, P. et al., 2011. The arrestin domain-containing 3 protein regulates body mass and energy expenditure. *Cell Metabolism*, 14(5), pp.671–683.

Patwari, P. & Lee, R.T., 2012. An expanded family of arrestins regulate

metabolism. *Trends in Endocrinology and Metabolism*, 23(5), pp.216–222.

Powell, J.E. & Zietsch, B.P., 2011. Predicting Sensation Seeking From Dopamine Genes: Use and Misuse of Genetic Prediction. *Psychological Science*, 22(3), pp.413–415. Available at: http://pss.sagepub.com/lookup/doi/10.1177/0956797610397669.

Ramayo-Caldas, Y. et al., 2014. A marker-derived gene network reveals the regulatory role of PPARGC1A, HNF4G, and FOXP3 in intramuscular fat deposition of beef cattle. In *Journal of Animal Science*. pp. 2832–2845.

Rolf, M.M. et al., 2015. Comparison of Bayesian models to estimate direct genomic values in multi-breed commercial beef cattle. *Genetics Selection Evolution*, 47(1), pp.1–14. Available at: http://www.gsejournal.org/content/47/1/23.

Saatchi, M. et al., 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics, selection, evolution : GSE*, 43(1), p.40.

Saatchi, M. et al., 2014. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC genomics*, 15(1), p.442. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4102727&tool=pmcentrez&rendertype=abstract.

Segura, V. et al., 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44(7), pp.825–830. Available at: http://dx.doi.org/10.1038/ng.2314.

Sellick, G.S. et al., 2007. Effect of myostatin F94L on carcass yield in cattle. *Animal Genetics*, 38(5), pp.440–446.

Snelling, W.M. et al., 2015. A survey of polymorphisms detected from sequences of popular beef breeds. *Journal of animal science*, 93(11), pp.5128–43. Available at: https://www.animalsciencepublications.org/publications/jas/articles/93/11/5128#comments.

Stephens, M., 2013. A Unified Framework for Association Analysis with Multiple Related Phenotypes. , 8(7).

Taylor, J.F., 2014. Implementation and accuracy of genomic selection.

Aquaculture, 420-421.

Turner, S.D., 2014. *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots*,

Villumsen, T.M., Janss, L. & Lund, M.S., 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. *Journal of Animal Breeding and Genetics*.

Weber, K.L. et al., 2012. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes. *Journal of Animal Science*, 90(12), pp.4177–4190.

Wray, N.R. et al., 2013. Pitfalls of predicting complex traits from SNPs.

Yang, J. et al., 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7), pp.565–9.

Yang, J. et al., 2011. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1), pp.76–82.

Zhang, H. et al., 2012. Progress of genome wide association study in domestic animals. *Journal of Animal Science and Biotechnology*, 3(1), p.26.

Zhang, Z. et al., 2010. association studies. , 42(4), pp.355–360.

Zhou, X., Carbonetto, P. & Stephens, M., 2013. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics*.

Zhou, X. & Stephens, M., 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4), pp.407–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24531419.