Detection of Complex Genetic Effects in Genome-wide Association Studies

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

LI MA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: YANG DA

JUNE 2010

Acknowledgements

First, I would like to thank my adviser Dr. Yang Da for his kindness, patient guidance, and generous financial support. Dr. Yang Da has given me tremendous help for my PhD research in the last few years. I also want to thank my committee members, Dr. F. Abel Ponce de León, Dr. Leslie Hansen, Dr. Scott Fahrenkrug and Dr. Cavan Reilly for their valuable suggestions and time. I also want to thank Dr. H. Birali Runesha from Minnesota Supercomputing Institute. Dr. H. Birali Runesha gave me valuable suggestions and guidance on parallel computing programming. Thanks to John Garbe for his time and effort on part of the computer program implementation. Finally, I thank all those who gave me the possibility to finish this thesis. My thesis could not have been accomplished without all of you.

This research is supported in part by project MN-16-043 of the Agricultural Experiment Station at the University of Minnesota, by funding from the National Research Initiative Competitive Grants Program/United States Department of Agriculture (grant #2008-35205-18846) and by a research grant from the Minnesota Supercmputer Institute.

Abstract

The large number of single nucleotide polymorphisms (SNPs) available provides a powerful molecular resource for identifying complex genetic interactions associated with complex traits or diseases but also presents unprecedented data analysis challenges. In this work we developed new quantitative genetics methods and parallel computing tools to detect complex interactive SNP effects underlying complex traits or diseases using genome-wide association studies (GWAS). The new quantitative genetics methods allow detection of novel interactions between genes, sex and environment including second order and third order gene-gene, gene-sex, gene-environment interactions, where each gene may have additive, dominance or parent-of-origin effects. The parallel computing tools allow such complex analysis to be conducted in a timely manner for any large scale GWAS and can be scalable to meet growing data analysis challenges in the future. The analytical and computing methods were applied to the analysis of a Holstein cattle GWAS data set and the Framingham Heart Study (FHS) data. Significant epistasis and single-locus effects were detected affecting human cholestoral levels and dairy production, fertility and body traits. The analytical methods and computing tools will significantly facilitate the discovery of complex mechanisms underlying phenotypes using GWAS.

Table of Contents

Acknowledgements ………………………………………………………………… i
Abstract ………………………………………………………………………….. ii
List of Tables ……………………………………………………………………. v
List of Figures …………………………………………………………………… vii

Ch 1. Introduction ……………………………………………………….……… 1
Epigenetic effects: a broad range of complex genetic effects ……….……... 1
Unprecedented challenges for data analysis ……………………………….. 6
Specific aims ………………………………………………………………….. 9

Ch 2. Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies …...…………….. 10
Introduction ……..…………………………………………………….….…… 10
Background ……………………………………………...………….….…… 11
Methods …………………………………..…….….…………………….. 13
Results ……...…….………………………………………………….….….…… 16
Discussion ……..……………………………………………………….…….... 21
Conclusions ……………………………………………………………………… 23
Availability and requirements …………………………………………………… 23

Ch 3. Generalized least squares method to account for sib correlation for testing SNP single-locus and epistasis effects in genome-wide association analysis ……… 29
Introduction ……………………………………………………………..…..…… 30
Single-locus and epistasis effects of two bi-allelic loci ……….…………..….…… 32
Generalized least squares (GLS) tests of SNP effects ……………….….….… 34
Maximumlikelihood estimation of variance components ...……..………….… 37
Implementation for parallel and serial computing ………………………….. 37

Ch 4. Detection of epistasis effects of nine dairy traits in contemporary U.S. Holstein cows ……..……...……………………………………………………… 39
Introduction ……..……………………………………………………….….……. 39
Materials and methods………………………………………………..….…… 41
Results and discussion ………………….…………………………….…….. 42

Ch 5. Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham Heart Study Data ……..….… 66
Introduction ……..……………………………………………………….…….. 66
Background …………………………………………………….……….….…… 67
Results ………………….……………………………………………….……..… 69
Discussion ………..…………………………………………………….……….. 72
Conclusions …....………………………………………….……..…………..…… 76
Methods ………………………………………………………………………….. 76

Ch 6. A general approach for detecting imprinting, gene-gene, gene-sex, and gene-environment epigenetic effects of quantitative traits in genome-wide association studies ………………………………………………………………………………. 87

Introduction …………………………..……………………………….……….….. 88

Two-locus Epigenetic Model ……………..………………………….….…. 91

Statistical Testing of Epigenetic Effects Involving Gene, Sex, and Environment Factors ……...…………………………………………….….. 96

Discussion ……..………………………………………..…….….…… 101

Ch 7. Multi-locus Test Conditional on Confirmed Effects Leads to Increased Power in Genome-wide Association Analysis …………………………….…..…… 111

Introduction …………………………..…………………………….…. 111

Predicted Statistical Power of MLT and SLT …………………….….…… 112

Evaluation of Statistical Power of MLT and SLT Using Simulation and Real Data ……..…………………………………………..………….….. 115

Ch 8. References …..……………………………………………………….... 121

Appendix ..…………………..……………………………….…..……….……..... 140

List of Tables

Table 1.1 Estimated computing time (T) of parallel computing for epistasis testing of one phenotype 8

Table 2.1 Estimated single-processor computing time on the SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processor, and the total number of tests for two-locus and three-locus analysis. 25

Table 2.2 Example of distributing N SNPs to m(m+1)/2 processor cores (Pi, i = 1, m(m+1)/2) for the case where N/m is an integer, where m = N/n = number of subsets of SNPs with each subset having n SNPs (m and n are assumed integers). Each diagonal core receives one subset of n SNPs and computes [3n + 5n(n?1)/2] tests, and each off-diagonal core receives two subsets of total 2n SNPs and computes 5n2 tests. 25

Table 4.1 Distribution of top 50 most significant pairs for each of the 9 dairy traits by chromosome 46

Table 4.2 Top 10 epistasis effects for milk yield (MY) 47

Table 4.3 Top 10 epistasis effects for fat yield (FY) 47

Table 4.4 Top 10 epistasis effects for protein yield (PY) 48

Table 4.5 Top 10 epistasis effects for fat percent (FPC) 48

Table 4.6 Top 10 epistasis effects for protein percent (PPC) 49

Table 4.7 Top 10 epistasis effects for somatic cell score (SCS) 49

Table 4.8 Top 10 epistasis effects for daughter pregnancy rate (DPR) 50

Table 4.9 Top 10 epistasis effects for productive life (PL) 50

Table 4.10 Top 10 epistasis effects for net merit (NM) 51

Table 5.1 Single-locus SNP effects for TC with genome control (GC) adjusted $P < 7.2\times10^{-8}$ 80

Table 5.2 Single-locus SNP effects for HDL-C with genome control (GC) adjusted $P < 7.2\times10^{-8}$ 81

Table 5.3 Epistasis effects for TC and HDL-C with the significance of suggestive linkage 82

Table 5.4 Frequency and effect of gamete, allele-genotype or genotype-genotype combination in each epistasis effect with statistical significance of suggestive linkage 83

Table 6.1 Genotypes, genotypic frequencies (pjklm) and genotypic values (gjklm) of two quantitative trait loci. 103

Table 6.2 Second order epigenetic effects for one autosome SNP and one X chromosome SNP detectable in males. 103

Table 6.3 Observed computing time (hours) for pairwsie epistasis testing of four phenotypes on 6575 individuals from the Framingham Heart Study 104

Table 6.4 Sample size required to achieve 95% power with 5%type-I error. 104

Table 7.1 Power comparison between MLT (Power I) and SLT (Power II) with various conditional SNP effect sizes and constant effect size of 0.1 SD for the candidate SNP 119

Table 7.2 Comparison of P-values of MLT and SLT using the Framingham Heart Study data. 119

Table E1 Probability of each possible mating to produce a heterozygous offspring (A_iA_j) for a multi-allelic locus. 148

List of Figures

Figure 2.1 Observed and predicted run times of the EPISNPmpi program on Minnesota Supercomputing Institute's 2.6 GHz IBM BladeCenter Linux cluster (Blade) and the SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processor (Calhoun). 26

Figure 2.2 Examples of chromosome view of single-locus significance and sample size produced by the EPISNPPLOT program that draws chromosome views for all chromosomes by one command. 27

Figure 2.3 Examples of SNP epistasis network of a phenotype produced by the EPINET program that by default draws the 10 largest epistasis networks from the input test results. 28

Figure 4.1 Epistasis QTL map of net merit and its eight component traits detected in comtemporary Holstein cows 52-56

Figure 4.2 Significant SNP epistasis network of milk 57

Figure 4.3 Significant SNP epistasis network of fat 58

Figure 4.4 Significant SNP epistasis network of protein 59

Figure 4.5 Significant SNP epistasis network of fat percent 60

Figure 4.6 Significant SNP epistasis network of protein percent 61

Figure 4.7 Significant SNP epistasis network of sometic cell score 62

Figure 4.8 Significant SNP epistasis network of daughter pregnancy rate 63

Figure 4.9 Significant SNP epistasis network of productive life 64

Figure 4.10 Significant SNP epistasis network of net merit 65

Figure 5.1 Gene regions associated with total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C) 84

Figure 5.2 QQ plots for single-SNP whole genome association tests of total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C) 85

Figure 5.3 Distributions of total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C) in original scales and in log-transformed scales. 86

Figure 6.1 Imprinting effect of a quantitative trait is detected through the comparison of heterozygous individuals with different parental allele origins. 105

Figure 6.2 Observed (dotted lines) and predicted (solid lines) statistical power as a function of imprinting contrast heritability 106

Figure 6.3 Observed (dotted lines) and predicted (solid lines) statistical power as a function of imprinting contrast heritability 107

Figure 6.4 Observed (dotted lines) and predicted (solid lines) statistical power as a function of the population size 108

Figure 6.5 Observed (dotted lines) and predicted (solid lines) statistical power as a function of the population size 109

Figure 6.6 Mean square error (MSE) of the epistasis estimations as a function of the population size 110

Figure 7.1 Power improvement of the MLT method conditional on confirmed SNP effects over the SLT method. 120

Chapter 1

Introduction

Genome-wide association study (GWAS) is an unprecedented powerful tool for detecting complex epigenetic effects for three reasons; 1) The large number of SNPs available to GWAS increases the likelihood of discovering causative SNPs and increases the statistical power for detecting SNPs in linkage disequilibrium with causative mutations. 2) Sample size of GWAS is on the increase due to historical accumulations and availability of the same disease phenotypes from multiple centers. 3) The large sample sizes make possible the testing and estimating many levels of genetic and environment factors. The dense genome coverage of SNPs combined with other genomics advances has led to effective molecular and biological characterizations of significant SNP effects. To realize the full potential of GWAS for the discovery of genetic mechanisms underlying complex human diseases, methodology and computing tools capable of evaluating many genetic hypotheses are needed. However, methods are lacking for studying epigenetic factors underlying human diseases based on SNP-disease association and computing difficulty for complex epigenetic analysis is a severe bottleneck of large scale GWAS.

1.1 Epigenetic effects: a broad range of complex genetic effects

Epigenetics refers to changes in gene function that do not involve changes in the underlying DNA sequence of the organism (Van Speybroeck, 2002; Feinberg, 2007). By this definition, a broad range of complex genetic effects can be characterized as epigenetic effects, including imprinting, gene-gene, gene-sex and gene-environment interactions. A combination of these epigenetic interactions could be responsible for the elusive nature of genetic mechanisms underlying complex traits. Because of the large numbers of possible combinations among genetic and environment factors, epigenetics could be complex, a great challenge to understanding complex traits.

Epigenetic changes are crucial for the development and differentiation of the various cell types in an organism, as well as for normal cellular processes such as X-chromosome inactivation in female mammals, silencing of mating-type loci in yeast and both in human disease and normal development (Van Speybroeck, 2002; Feinberg, 2007; Eccleston, 2007; Johannes, 2007; Bird, 2007). Epigenetics is at the heart of phenotypic variation in health and disease so that understanding and manipulating the epigenome holds enormous promise for preventing and treating common human illness (Feinberg, 2007). The ability to discover each type of epigenetic effects and the interaction between various epigenetic effects will significantly advance the study of genetic mechanisms underlying human diseases.

The following reviews the significance of four categories of epigenetic effects for which our proposed work will provide methodology of detection.

1.1.1 Imprinting and epigenetics

Imprinting is a well-known epigenetic mechanism (Feinberg, 2007; Eccleston, 2007; Johannes, 2007; Bird, 2007; Esteller 2007). Imprinting corresponds to a parent-of-origin effect, meaning that the same alleles from the father and mother do not have the same gene activities owing to the silencing of an allele related to different levels of DNA methylation on the parental chromosome. Testing of interaction between imprinting and other genetic and environment factors including sex of the individual may provide new insights into the role of imprinting in disease expression, such as whether imprinting and other genetic and environmental factors modify or regulate each other. GWAS provides an opportunity to reveal the associations between SNP markers and a phenotype, such an association is suggested when two offsprings with heterozygous SNP genotypes, *Aa* and *aA*, where the allele on the left comes from the father and the allele on the right from the mother, do not have the same phenotypes (de Koning, 2002; London, 2004) either due to linkage disequilibrium with a causative imprinting factor or due to the SNP being a causative imprinting factor. Through this mechanism, the proposed works offers a broad range of tests for interactions between imprinting and other genetic effects, including additive and dominance effects, imprinting by sex effects, and imprinting by environmental factors.

1.1.2 Sex and epigenetics

Recent studies suggest that sex-specific genetic architecture also influences human phenotypes, including reproductive, physiological and disease traits, and genetic studies that ignore sex-specific effects in their design and interpretation could fail to identify a significant proportion of the genes that contribute to risk for complex diseases (Ober, 2008). Examples of reversal dominance (Osborn, 1916; Wentworth, 1916), an allele being dominant in one sex and recessive in the other, were reported in 1916, and more broad evidence of sex-gene interactions was discovered by recent studies (Ober, 2008). This proposal develops and implements broad tests for sex-gene interactions allowed by current parallel computing power, including interactions between sex and additive, dominance, imprinting, and second order gene-gene interactions.

1.1.3 Environment and epigenetics

Common diseases may involve phenotypic variants with both genetic variation and environmentally triggered epigenetic changes that modulate the effects of DNA sequence variation. Environmental factors (hormones, growth factors, toxins and dietary methyl donors) influence both the genome and epigenome and these epigenetic modifiers in turn are affected by variation in the genes that encode them (Feinberg, 2007; Bjornsson, 2004). Age-gene interaction as indicated by age-dependent penetrance is found for many human complex diseases (Farrer, 2006; Yarden, 2008; Crowe, 1977) including mental health (Rutte, 2005). Human genetic data often record a large number of environmental factors, such as age, smoking, drinking, diet, and medical and fitness treatments. While environment factors are known to affect disease expression, little is known about how environment affects various types of genetic effects at the level of genome-wide SNP-phenotype association. Methods to be developed in this proposal will allow the detection of a wide range of specific gene-environment interactions to define the nature of their interactions, including environment-imprinting, environment-allele, environment-genotype, or environment-epistasis interactions.

1.1.4 Epistasis and epigenetics

Epistasis refers to effects moderated by gene-gene interactions. Epistasis is an epigenetic effect because it results in variation in functional outcomes without changes in DNA sequences. The significance of epistasis in complex traits has been well recognized (Sanjuán, 2006; Carlborg, 2004; Moore, 2003). Observations of epistasis include deviations from Mendelian ratios, molecular interactions in gene regulation and biochemical and metabolic systems, non-constant positive effects of single polymorphisms, and gene interaction effects commonly found (Moore, 2003). Large epistasis networks found in yeast (Freudenberg-Hua, 2003; Tong, 2004) also point to the ubiquitous nature of epistasis. The global analysis of epistatic gene-interaction patterns bear a striking resemblance to what is now called systems biology (Phillips, 2008; Moore, 2005). Given recent work in this area, it is likely that for the next century the concept of epistasis will be even more central to biology than it has been over the past century (Phillips, 2008).

The quantitative genetics approach to study epistasis started with Fisher's work (1918) that defined epistasis as the residual genotypic value not explained by linear single gene effects. Cockerham (1954) and Kempthorne (1954) using different methods partitioned epistasis effects into four components of additive × additive, additive × dominance, dominance × additive, and dominance × dominance epistasis effect, with the genetic interpretation of allele × allele, allele × genotype, genotype × allele, and genotype × genotype interactions respectively. This partitioning can be used as a tool for identifying the exact mode of a gene interaction effect. Kempthorne's model uses the deviation of "genetic combination" from the sum of individual genetic factors. For example, an additive × additive effect (combination effect of two alleles each from a different locus) is a deviation of the gametic mean from the two single allelic means. This is clearly a measure for the gene combination effect and hence has a straightforward genetic interpretation of the epistasis effect. We have extended the Kempthorne epistasis approach to allow LD and Hardy-Weinberg disequilibrium (HWD) (Mao, 2006) and implemented this extended epistasis method in parallel and serial computing programs (Ma, 2008a; Ma, 2008b; Ma, 2008c), with the parallel version being capable of pairwise epistasis testing for any large GWAS currently in existence. Our cumulative work on quantitative genetics methods (London, 2004; Tong, 2004; Mao, 2005) and computing tools (Ma, 2008a; Ma, 2008b; Ma, 2008c) for epistasis detection will effectively facilitate achieving Aims in this proposal.

1.1.5 Epigenetics networks underlying phenotypic changes

Results from epigenetic analysis of SNP-phenotype association can be summarized as an epigenetic network showing gene-gene, gene-environment, and gene-sex interactions affecting a complex trait. This epigenetic network based on SNP-phenotype association is an important interaction network (Beyer, 2007) and can be combined with molecular functional (Bray, 2003) and functional (Gardner, 2003) networks towards the complete understanding of genetic mechanisms underlying complex diseases.

1.2 Unprecedented challenges for data analysis

GWAS provides unprecedented capabilities for identifying epigenetic factors underlying complex traits but also present unprecedented data analysis challenges. Analyzing two or more SNPs at a time for all possible SNP combinations is a demanding computational analysis due to the large number of SNP combinations. With current computing power, testing of second order epigenetic effects and third order gene-gene-sex, gene-gene-environment interaction are the only practical exhaustive epigenetic testing for all SNP combinations. For these types of epigenetic effects, current computing power should be able to analyze all SNPs in the human genome (10 million SNPs) based on the excellent scalability of our current parallel computing program (Ma, 2008b) for epistasis testing and the massive parallel computing power currently available.

For 10,000,000 SNPs, which would be all SNPs on the human genome, the estimated computing time using 200,000 intel 2.66 Ghz processor cores is about 24 hours for pairwise epistasis testing (**Table 1.1**). For the proposed additional types of interaction effects involving sex and environment factors, the estimated computing time should be multiplied by the number of sex and environment factors so that computing time could increase to days, which still would be practical. Using a large powerful parallel computing system could reduce the computing time. For example, using the fastest supercomputers such as the IBM BladeCenter Roadrunner QS22/LS21 Cluste2 could considerably reduce the computing time. A tempting solution would be to test epistasis effects for a subset of the SNPs with significant single-locus effects. However, it should be cautioned that requiring significant main effects for epistasis testing could miss many or even all significant epistasis effects (Ma, 2007).

For 500,000 ~ 1,000,000 SNPs, which are representative of current large scale GWAS, pairwise epistasis testing would require less than one day to complete. In our analysis of the 500,000 SNPs from the Framingham Heart Study, pairwise epistasis tests using our parallel computing program took about 14 hours using 800 processor cores.

For epigenetic testing using the methods to be developed in this proposal, computing time practically should be a few times as much as epistasis testing, which should be very manageable using today's supercomputer power. The capabilities of our proposed epigenetic testing are designed based on the current supercomputer power, i.e., the designed capabilities make maximum use of current computing and are yet doable.

Table 1.1. Estimated computing time (T) of parallel computing for epistasis testing of one phenotype

Number of SNPs (N)	Number of processor cores to be used	Two-locus analysis	Three-locus analysis
500,000	500	T ≈ 20 hours	T ≈ 400 years
	2000	T ≈ 6 hours	T ≈ 100 years
	200,000	T ≈ 4 minutes	T ≈ 1year
1,000,000	500	T ≈ 4 days	T ≈4,000 years
	2000	T ≈ 24 hours	T ≈ 500 years
10,000,000	100,000	T ≈ 2 days	T ≈ 2000 years
(All SNPs on human genome)	200,000	T ≈ 24 hours	T ≈ 1000 years

The estimated computing time was based on run times on the SGI Altix XE 1300 Linux cluster system (Calhoun) with 2.66 GHz Intel processor-cores (total 2048 cores) at the Minnesota Supercomputer Institute. The large numbers of tests for analyzing two or more SNPs is the reason of the computational bottleneck. The number of tests (M) for analyzing two or more SNPs jointly is, $M = N(N-1)/2$ for testing two SNPs at a time, and $M = N(N-1)(N-2)/6$ for testing three SNPs at a time.

1.3 Specific aims

The overall goal is to develop a new quantitative genetics methods and parallel computing tools to detect complex epigenetic SNP effects in genome-wide association studies (GWAS). Specific aims include: Aim 1, develop new quantitative genetics methods for detecting complex epigenetic effects affecting categorical traits; Aim 2, develop parallel computing tools to implement the novel quantitative genetics methods with the capability of analyzing large scale GWAS data in a timely manner; Aim 3, apply the new methods and computing tool to the analysis of the USDA/NRI funded dairy GWAS data and the Framingham Heart Study GWAS data for detecting epigenetic effects.

Successful realization of these aims will significantly facilitate epigenetics discovery in GWAS by providing the necessary analytical methods and computing tools to: 1) provide targets to identify molecular elements and pathways; 2) identify and quantify the effects of single genetic factors or interactions between genetic and environmental perturbation on the molecular elements underlying complex phenotypes; and 3) identify the role of non-genetic factors in complex SNP networks underlying phenotypes.

Chapter 2

Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies

2.1 Introduction

Genome-wide association studies (GWAS) using single nucleotide polymorphism (SNP) markers provide opportunities to detect epistatic SNPs associated with quantitative traits and to detect the exact mode of an epistasis effect. Computational difficulty is the main bottleneck for epistasis testing in large scale GWAS. The EPISNPmpi and EPISNP computer programs were developed for testing single-locus and epistatic SNP effects on quantitative traits in GWAS, including tests of three single-locus effects for each SNP (SNP genotypic effect, additive and dominance effects) and five epistasis effects for each pair of SNPs (two-locus interaction, additive × additive, additive × dominance, dominance × additive, and dominance × dominance) based on the extended Kempthorne model. EPISNPmpi is the parallel computing program for epistasis testing in large scale GWAS and achieved excellent scalability for large scale analysis and portability for various parallel computing platforms. EPISNP is the serial computing program based on the EPISNPmpi code for epistasis testing in small scale GWAS using commonly available operating systems and computer hardware. Three serial computing utility programs were developed for graphical viewing of test results and epistasis networks, and for estimating CPU time and disk space requirements. The EPISNPmpi parallel computing program provides an effective computing tool for epistasis testing in large scale GWAS, and the epiSNP serial computing programs are convenient tools for epistasis analysis in small scale GWAS using commonly available computer hardware.

2.2 Background

The large number of SNPs available provides opportunities for detecting DNA variations associated with complex traits through GWAS using SNP markers. This is because an increased number of SNPs increases the chance that some SNPs may be DNA variations affecting phenotypes (direct SNP effects) or results in increased linkage disequilibrium (LD) with DNA variations that have direct effects on the phenotypes (indirect SNP effects). With high throughput SNP genotyping technology, SNP genotyping of a large number of individuals is becoming increasingly practical. Such large scale SNP genotyping increases the effectiveness of SNP association studies and provides an unprecedented opportunity to study complex genetic effects such as epistasis. The significance of epistasis (gene interaction) in complex or quantitative traits has been well recognized (Balding, 2006; Carlborg, 2004; Li, 2000; Moore, 2003; Purcell, 2004). Large epistasis networks showing complex interactions among genes have been reported (Nishihara, 2007; Sambandan, 2006; Schadt, 2005). Fisher's partition (1918) of the nine genotypic values of two bi-allelic loci into single gene effects (additive and dominance effects) and an epistasis effect assuming Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) laid the foundation of a quantitative genetics approach to study epistasis. Also assuming HWE and LE, Cockerham (1954)] and Kempthorne (1954) partitioned Fisher's epistasis effect into four components using two different methods: additive × additive (A×A), additive × dominance (A×D), dominance × additive (D×A), and dominance × dominance (D×D) epistasis effects with the genetic interpretation of allele × allele, allele × genotype, genotype × allele, and genotype × genotype interactions respectively. This partitioning can be used as a tool for identifying the exact mode of a gene interaction effect. Kempthorne's partitioning of genotypic values has been extended to allow Hardy-Weinberg disequilibrium (HWD) and linkage disequilibrium (LD) so that Kempthorne's method could be used to test single-locus and epistasis effects in populations where HWD and LD may exist (Mao, 2006). With genome-wide detection of epistasis effects, epistasis networks affecting a quantitative trait could be established. Computational difficulty is the main bottleneck for epistasis testing in large scale GWAS due to the large number of SNP combinations. The number of SNP combinations (M) is $M = N(N-1)/2$ for testing two SNPs at a time, and is $M = N(N-1)(N-2)/6$ for testing three SNPs at a time, where N = number of SNPs. The computational difficulty of epistasis testing in large scale GWAS can be an open scale computing challenge that could exhaust the capabilities of any supercomputer in existence today. For example, pairwise epistasis testing of 1,000,000 SNPs would require 5 years using our EPISNP program and a single processor of the 2.66 GHz SGI Altix XE 1300 Linux cluster system at the Minnesota Supercomputer Institute, and this computing time could increase to 1.5 million years by adding just one SNP to the pairwise analysis (Table 2.1). With parallel computing, pairwise epistasis testing for any large scale GWAS currently in existence is possible. Large scale three-SNP epistasis testing may not be computationally feasible at this time. The parallel and serial computing software developed in this research is intended to provide computational tools for pairwise epistasis testing in GWAS on various parallel and serial computing platforms with the capability of pairwise epistasis testing for any large GWAS currently in existence.

2.3 Methods

The statistical methods implemented by the parallel and serial computing tools for detecting single-locus and epistasis effects include a general linear model for testing the marker effects of each SNP and each SNP pair, and include the extended Kempthorne model for testing additive and dominance effects of each SNP and for testing A×A, A×D, D×A, and D×D epistasis effects of each SNP pair. A two-step least squares analysis (Wolfinger, 2001) is used to implement the statistical tests. The first step corrects the phenotypic values for systematic effects such as gender and age. This step estimates fixed non-genetic effects and then removes the estimated fixed non-genetic effects from the original phenotypic observations to obtain the corrected phenotypic values (or residual values). The second step conducts epistasis and single-locus tests using the corrected phenotypic values as the phenotypic observations. This two-step analysis estimates and removes systematic effects only once and hence has considerable computational advantage when the number of SNPs is large. The single-locus analysis tests three genetic effects: the SNP genotypic effect, additive effect, and dominance effect. The statistical model for testing single-locus effects is $y = \mu + SNP + e$, where y = corrected phenotypic value, μ = common mean, SNP = the single-locus SNP genotypic effect, and e = random residual. The single-locus SNP genotypic effect was partitioned into additive and dominance effects. The single-locus genotypic effect answers the question whether the SNP had an effect on the phenotype whereas additive or dominance effect identifies the mode of the SNP effect. The statistical model for testing epistasis effects is y = μ + SNP1 + SNP2 + SNP1*SNP2 + e, where SNP1 and SNP2 are the two single-locus genotypic effects, and SNP1*SNP2 is the two-locus interaction effect (I-effect). The two-locus interaction effect was partitioned into four individual epistasis effects using the extended Kempthorne model that allows HWD and LD: A×A, A×D, D×A, and D×D epistasis effects. The two-locus interaction effect answers the question whether the two SNPs had an interaction effect whereas an individual epistasis effect (A×A, A×D, D×A, or D×D) identifies the mode of the interaction. The significance tests of the single-locus SNP effect and the two-locus interaction effect used an F-test. A t-test was used to test the significance of additive, dominance and epistasis effect using the following formula

$$\hat{T}_x = \frac{L_x}{\sqrt{\hat{\mathrm{var}}(L_x)}} = \frac{\mathbf{s}_i\hat{\mathbf{g}}}{\sqrt{\mathbf{s}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{s}'_i s^2}}$$

where L_x = contrast to estimate the genetic effect, $s^2 = (\mathbf{y} - \mathbf{X}\hat{\mathbf{g}})'(\mathbf{y} - \mathbf{X}\hat{\mathbf{g}})/(n-k)$ = estimated residual variance, $\hat{\mathbf{g}}$ = the least squares estimates of the SNP genotypic effects, and $\mathbf{s}_i$ = a function of marginal and conditional allelic and genotypic frequencies for estimating genetic effect i, which is either additive, dominance or an epistasis effect, and where n = number of observations and k = rank of **X** (Mao, 2006). For testing epistasis effects involving the X chromosome in mammals (or Z chromosome in birds), only females (or males in birds) can be included in the analysis. For epistasis analysis involving SNPs in pseudoautosomal regions, the analysis is the same as for autosomal SNPs. These epistasis testing methods were implemented in a parallel computing program intended for larges scale GWAS and in a serial computing program intended for small scale GWAS that could be analyzed on commonly available computer hardware.

Minimizing the processor memory required is critical to developing an efficient and successful parallel computing program because each individual processor has a limited amount of memory available. For example, each core of the quad-core processor on the SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processors (Calhoun) at the Minnesota Supercomputer Institute has a limit of 2 GB of memory. Therefore, the parallel code should use as little processor memory as possible to achieve scalability for large scale analysis that will otherwise require large processor memory. A two dimensional SNP data distribution scheme (Table 2.2) among processor cores was designed to minimize the memory requirement of each processor. To assign SNPs to each processor core, the N SNPs are evenly divided into m subsets with n SNPs in each subset such that the total number of processor cores (p) to be used is p = m(m+1)/2. For simplicity, n = N/m is assumed to be an integer. In the case N/m is not an integer, the leftover SNPs are assigned to an extra core. In Table 2.2, each diagonal core receives one subset of n SNPs and computes [3n + 5n(n−1)/2] tests, and each off-diagonal processor core receives two subsets of SNPs (2n SNPs) and computes $5n^2$ pairwise tests. Therefore, only (2n)/(mn) = 2/m of the N SNPs are stored in each off-diagonal processor, and only 1/m of the N SNPs are stored in each diagonal processor defined in Table 2.2. As the number of processor cores (p) increases, the number of SNP subsets (m) increases and the memory required for each processor core decreases. Therefore, the increased memory requirement per processor for large scale SNP analysis can be reduced by increasing the number of processor cores used. The parallel computing code was optimized to minimize inter-processor communications and was crafted for portability to various parallel and serial computing platforms. Testing results showed that the parallel computing code achieved excellent speedup and scalability and achieved excellent portability, as to be discussed below.

2.4 Results

A parallel computing program named EPISNPmpi and a serial computing program named EPISNP were developed for genome-wide pairwise epistasis testing. Three serial computing utility programs were developed to estimate computing time, to produce graphical chromosome view of significant single-locus results, and to produce graphical display of epistasis network.

2.4.1 The EPISNPmpi and EPISNP programs

The EPISNPmpi and EPISNP programs provide two sets of SNP tests: single-locus analysis and pairwise analysis. The single-locus analysis tests three effects of each SNP: SNP genotypic effect (M), additive (A) and dominance (D) effects. The pairwise analysis tests five effects of each pair of SNPs: The I-effect, A×A, A×D, D×A, and D×D. Three input files in text format are required, the phenotype file, the SNP genotype file, and the parameter file. The phenotype file contains observations of the quantitative trait(s), family ID, individual ID, individual gender, and non-genetic fixed effects such as smoking status and age of each individual. The SNP genotype file contains family ID, individual ID, individual gender, and SNP genotypes, and should be one file for each chromosome. The parameter file with file name parameter.dat provides various user-specified controls for the EPISNPmpi and EPISNP programs to have the flexibility to be generally applicable. These controls include the number of quantitative traits to be analyzed, user specified number of chromosomes, code for the sex chromosome, formats for SNP genotypes and missing values, and user specified number of fixed non-genetic factors to be included in the statistical model, where a fixed non-genetic factor can be an indicator variable or continuous variable (covariable). Both EPISNPmpi and EPISNP programs are applicable to populations with HWD and LD.

The speedup and scalability (Eager, 1989; Alabdulkareema, 2001) of the EPISNPmpi parallel program were evaluated for two supercomputer systems: a 2.6 GHz AMD Opteron IBM BladeCenter Linux cluster (Blade) and the Calhoun system. In parallel computing, speedup refers to how much a parallel algorithm is faster than a corresponding sequential algorithm and is defined as $S_k = T_1/T_k$, where k = number of processors, T_1 = the execution time of the sequential algorithm with one processor-core, and T_k = the execution time of the parallel algorithm with k processor-cores. Linear or ideal speedup is achieved when $S_k = k$. Scalability refers to the stability of average performance of a parallel program as the number of processors increases. Ideal scalability is achieved when the efficiency of k processors (E_k) is $E_k = S_k/k = 1$. Figure 2.1 shows the observed and predicted computing time using 15-528 processor cores, where each processor consists of four cores. The predicted computing time was calculated using the following formula assuming an ideal speedup or scalability

$$t_k = t_1/k \tag{1}$$

where k = number of processor cores, t_k = computing time using k processor cores, and t_1 = computing time using one processor core. In Figure 2.1, the computing times were normalized to the computing time on 15 processor-cores because the minimal number of cores used was 15. Results in Figure 2.1 showed that the observed computing time and the predicted computing time assuming ideal speedup and scalability matched very well, indicating that the EPISNPmpi coding achieved excellent speedup and scalability. Based on the observed run times of 0.20 and 19.3 hours for 50,000 and 500,000 SNPs respectively using 528 cores of the Calhoun system, the estimated computing time for pairwise epistasis tests is approximately an increasing quadratic function of the number of SNPs. Let N = the number of SNPs and N_0 = a smaller number of SNPs with a known computing time (t_0) for running EPISNPmpi such that N = N_0 (x). Then, the computing time required for analyzing N SNPs (t_N) is approximately

$$t_N = (t_0)(x^2) \tag{2}$$

The run time of 19.3 hours for 500,000 SNPs using 528 cores showed that pairwise epistasis testing for GWAS with about 500,000 SNPs could be completed in one day using about 25% of the 2048 cores of the Calhoun system. Based on this computing time and equations.(1-2), the predicted time for pairwise epistasis testing among 1,000,000 SNPs using all 2048 cores of the tCalhoun system would require about 20 hours to complete. This prediction indicates that EPISNPmpi is capable of completing pairwise epistasis analysis in one day for any large scale GWAS currently in existence, noting that the numbers of SNPs used in current large scale GWAS are in the range of 500,000 ~ 940,000, as represented by NIH's GAIN projects. Sample size, or the number of individuals, affects the computing time as well, but the increase in computing time due to increased sample size is minor. The EPISNPmpi code is highly portable to various computing platforms and has been ported to all supercomputer systems at the Minnesota Supercomputer Institute and to several popular serial computing platforms.

The EPISNP program is designed for epistasis analysis in small-scale GWAS on commonly available computer hardware. For example, an analysis of 5700 SNP markers took about 18 hours to complete on a PC with a single 3.8 GHz Pentium 4 processor.

The EPISNPmpi and EPISNP programs each produces two output files of the most significant results of single-locus tests and two output files of the most significant results of pairwise epistasis tests. The output file for significant epistasis results currently displays the names and chromosome locations of the two SNPs in each SNP pair with significant I-effect (interaction between the two loci), A×A, A×D, D×A, or D×D effect, significance level (p-value), and ordered estimates of individual effects that are useful for identifying the best and worst gene combinations affecting a phenotype. The second output file of single-locus tests is used as the input file of the EPISNPPLOT program and the second output file of pairwise epistasis tests is used as the input file of the EPINET program.

2.4.2 Three serial computing utility programs

The EPISNPPLOT program plots the chromosome view figures, where each figure shows the significance of each of the three single SNP effects and the sample size for all SNPs on each chromosome (Figure 2.2). The program produces chromosome view figures for all chromosomes by one command using an output file from EPISNPmpi or EPISNP as the input file. These chromosome views help identify chromosome regions with various degrees of significant effects and markers that did not have sufficient information to yield significant effects. By default, the EPISNPPLOT program draws chromosome view figures in the original marker order as in the input file. The user has the option to sort the input data by the marker significance, additive significance, dominance significance, or the number of observations. In Figure 2.2, the figure on the left is an example of a chromosome view based on the original marker order, and the figure on the right is an example of a chromosome view in ascending order of significant dominance effects. The EPINET program draws figures of epistasis networks of SNPs with significant epistasis effects at four user specified p-values. The program requires two input files: the parameter file to specify four significance levels (p values) for selecting loci in the epistasis networks, and the effect file that contains epistasis testing results from EPISNPmpi or EPISNP. The default input is to use 'effects.dat' as the input file and to print the 10 largest networks (Figure 2.3). Alternatively, the user can specify the file name on the command line. If the input file is specified, the number of networks to print can also be specified. The EPINET program uses four user specified node colors to represent the four significance levels defined by the corresponding p-values, and five program defined line colors to denote the five types of epistasis effects (Figure 2.3). The CPUHD estimates CPU time required to complete the data analysis using the EPISNP program and the total storage space required to store the output files. This is helpful for planning an epistasis analysis. For example, a potentially excessively long running time can be avoided by running CPUHD first. Detailed instructions for using EPISNPmpi, EPISNP and the three utility programs described below are available in two user manuals.

2.4.3 Commodity cluster-based processing of EPISNPmpi

EPISNPmpi has been developed and tested on many modern high-performance computers and supercomputer systems. Price-to-performance ratio of the computing system can be an important consideration in practice. To utilize commonly available computer hardware for high performance computing, EPISNPmpi has been implemented to run on commodity cluster or on an inexpensive network of workstations using MPICH message passing libraries. MPICH is a portable implementation of MPI, a standard for message-passing for distributed-memory applications, and is freely available at www.mcs.anl.gov/mpi/mpich1/download.html.

2.5 Discussion

Computational difficulty is the main bottleneck of epistasis testing in large scale GWAS. The computing tools we have developed help address the computational difficulty in epistasis analysis in large scale GWAS. The computing speed can be further improved if a more powerful computer system is used. However, serious computational challenges still exist in at least three areas: 1) Increased number of SNPs used in GWAS, 2) Integration of GWAS and a gene expression study, and 3) Joint epistasis testing for three or more SNPs at a time. The human genome has about 10 million SNPs. Although an exhaustive analysis of all human SNPs is not yet a reality, the number of SNPs used in GWAS is clearly rapidly increasing. Since the computing time for epistasis testing increases approximately as a quadratic function of the number of SNPs, computing difficulty will rapidly increase as the number of SNPs increases. Integration of large scale GWAS and a gene expression study using the same individuals poses another serious computational challenge. In this case, the computing time required is multiplied by the number of genes, where gene expression intensity of each gene is treated as one phenotype [19]. The joint epistasis testing for three or more SNPs could be the ultimate computing challenge. As shown in Table 2.1, adding just one SNP to the pairwise epistasis test for 1,000,000 SNPs could require 1/3 million times as much computing time. A tempting solution would be to test epistasis effects for a subset of SNPs with significant single-locus effects. However, this is not a good idea because requiring significant main effects for epistasis testing could miss many or even all significant epistasis effects with stringent p-values to declare significance. For example, the significant epistasis effects with $p < 10^{-7}$ for 5700 SNPs covering all 23 human chromosomes reported in Ma et al. [20] did not involve any SNPs with significant single-locus at $p < 10^{-4}$. Therefore, requiring significant single-locus effects at $p < 10^{-4}$ would have missed all the ten significant epistasis effects at $p < 10^{-7}$ among the 5700 SNPs. The EPISNPmpi and EPISNP programs provide capabilities for testing all possible pairwise epistasis effects. However, the use of these programs should be considered as only one step in GWAS analysis. Considerable work still may be required for digesting the test results.

2.6 Conclusions

The EPISNPmpi parallel computing program provides a computing tool capable of completing pairwise epistasis tests in large scale GWAS in a timely manner using a supercomputer system. The serial computing programs can be useful and convenient tools for epistasis analysis in small scale GWAS using commonly available computer hardware. EPISNPmpi is a portable program which not only exploits the capability of supercomputers but also runs on inexpensive loosely coupled cluster systems.

2.7 Availability and requirements

Project name: Parallel and serial computing for genome-wide SNP analysis

Project homepage: http://animalgene.umn.edu/

Operation systems:

1. EPISNPmpi [http://animalgene.umn.edu/episnpmpi/index.html]:

EPISNPmpi is the parallel computing program for testing single-locus and pairwise epistasis effects and is available for running on the majority of parallel computer systems. The following are the currently supported processors type, MPI libraries, compilers and corresponding binaries:

MPI library	*Compiler*	*Processor*	*Binary*
Voltaire MPI	Intel	Intel	EPISNPmpi_2.0_Voltaire_intel_intel.tar.gz
Voltaire MPI	Intel	AMD	EPISNPmpi_2.0_Voltaire_intel_AMD.tar.gz
Voltaire MPI	Intel	Intel (EM64T)	EPISNPmpi_2.0_Voltaire_suse_EM64T.tar.gz
PathMPI	Pathscale	AMD	EPISNPmpi_2.0_Pathscale_suse_AMD.tar.gz
IntelMPI	Intel	AMD	EPISNPmpi_2.0_intelMPI.suse_AMD.tar.gz
OpenMPI	Intel	Intel (EM64T)	EPISNPmpi_2.0_OpenMPI_suse_EM64T.tar.gz
IBM MPI	Intel	Power4	EPISNPmpi_2.0_IBM_AIX_pwr.tar.gz
MPT	Intel	Itanium	EPISNPmpi_2.0_SGI-Altix_SUSE_itanium.tar.gz

2. epiSNP [http://animalgene.umn.edu/episnp/index.html]:

The epiSNP package consists of four serial computing programs, EPISNP, CPUHD, EPISNPPLOT, and EPINET. EPISNP is the serial computing program for testing single-locus and pairwise epistasis effects. The following are the currently supported operation systems, processors types, and compilers used to generate binaries:

Operation system	*Compiler*	*Processor*	*Binary*
Widows	Intel	Intel/AMD	epiSNP_2.0_Widows.zip
Irix	SGI	MIPS	epiSNP_2.0_SGI_Irix_Mips.tar.gz
Linux (SUSE)	Intel	AMD	epiSNP_2.0_intel_suse_AMD.tar.gz
Linux (SUSE)	Intel	Intel (EM64T)	epiSNP_2.0_intel_suse_EM64T.tar.gz
Linux	Portland	Intel (32bit)	epiSNP_2.0_Linux_Portland_Intel.tar.gz
Linux (SUSE)	Pathscale	AMD	epiSNP_2.0_Pathscale_suse_AMD.tar.gz
Unix (AIX)	XLF	Power4	epiSNP_2.0_xlf_AIX_power.tar.gz

In the above binaries, epiSNP_2.0_Windows.zip contains all the four programs (EPINET, CPUHD, EPISNPPLOT, EPINET), while each of the other .gz file contains EPISNP and CPUHD only.

Other requirements: None.

License: None.

Any restrictions to use by non-academics: None.

Table 2.1. Estimated single-processor computing time on the SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processor, and the total number of tests for two-locus and three-locus analysis.

Number of SNPs (N)		Two-locus analysis	Three-locus analysis
500,000	Computing time (T) Number of tests (M)	T ≈ 1.2 years $M = (1.25)\ 10^{11}$	T ≈ 200,000 years $M = (2.08)\ 10^{16}$
1,000,000	Computing time (T) Number of tests (M)	T ≈ 5 years $M = (5.0)\ 10^{11}$	T ≈ 1.5 million years $M = (1.67)\ 10^{17}$

Table 2.2. Example of distributing N SNPs to m(m+1)/2 processor cores (P_i, i = 1, m(m+1)/2) for the case where N/m is an integer, where m = N/n = number of subsets of SNPs with each subset having n SNPs (m and n are assumed integers). Each diagonal core receives one subset of n SNPs and computes [3n + 5n(n−1)/2] tests, and each off-diagonal core receives two subsets of total 2n SNPs and computes $5n^2$ tests.

Subset 1: SNP_1 … SNP_n	Subset 2: SNP_{n+1}… SNP_{2n}	… …	Subset m: $SNP_{n(m-1)+1}$ … SNP_N	
P_1	P_2	… …	P_m	Subset 1: SNP_1 … SNP_n
	P_{m+1}	… …	P_{2m-1}	Subset 2: SNP_{n+1}… SNP_{2n}
		… …	… …	… …
			$P_{m(m+1)/2}$	Subset m: $SNP_{n(m-1)+1}$ … SNP_N

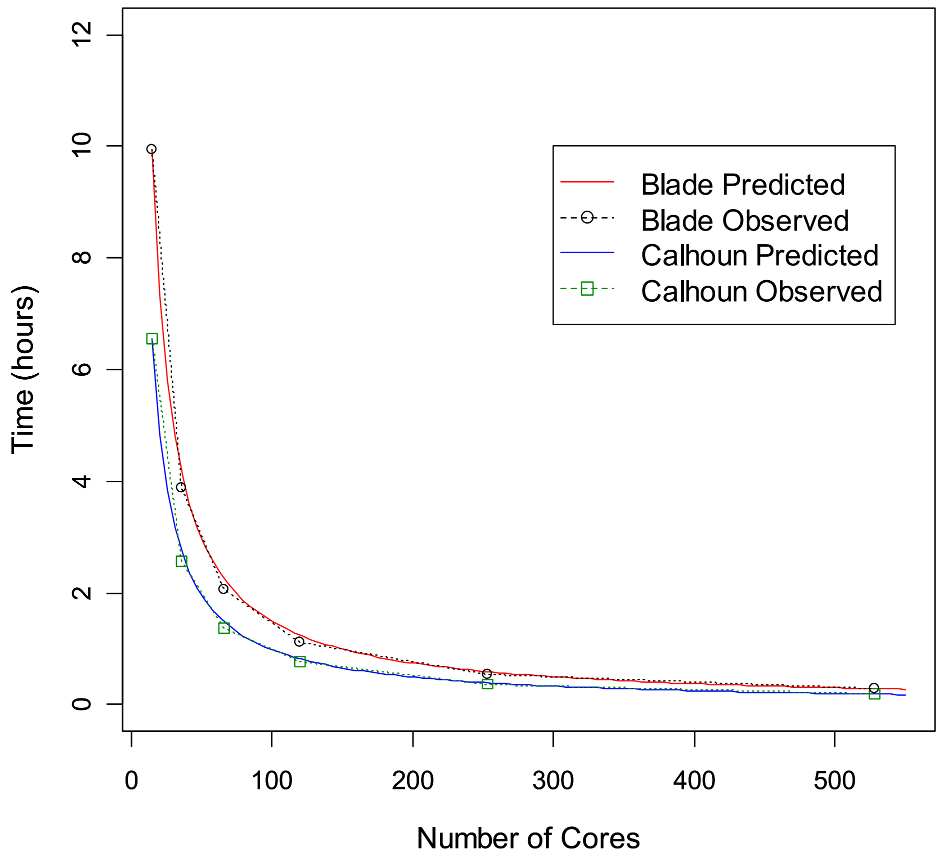


Figure 2.1

Observed and predicted run times of the EPISNPmpi program on Minnesota Supercomputing Institute's 2.6 GHz IBM BladeCenter Linux cluster (Blade) and the SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processor (Calhoun). The observed run times (circles representing Blade and squares representing Calhoun) matched well with the predicted run times under ideal speedup and scalability (solid line representing Blade and dotted line representing Calhoun). Analyses in this figure used a hypothetical GWAS data set with 50,000 SNPs and 2000 individuals.

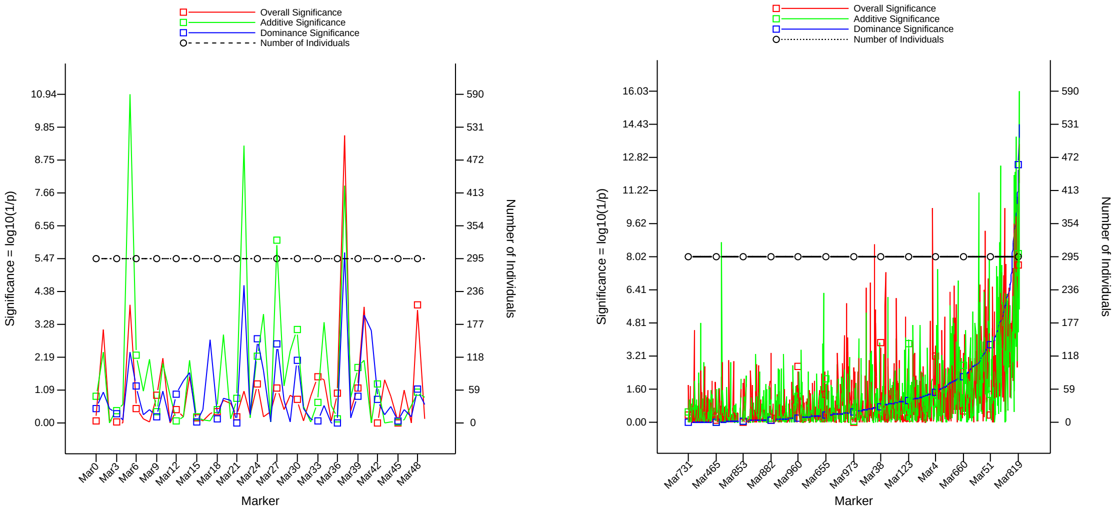


Figure 2.2
Examples of chromosome view of single-locus significance and sample size produced by the EPISNPPLOT program that draws chromosome views for all chromosomes by one command. The figure on the left is an example of chromosome view based on the original marker order, and the figure on the right is an example of chromosome view in ascending order of significant dominance effects.

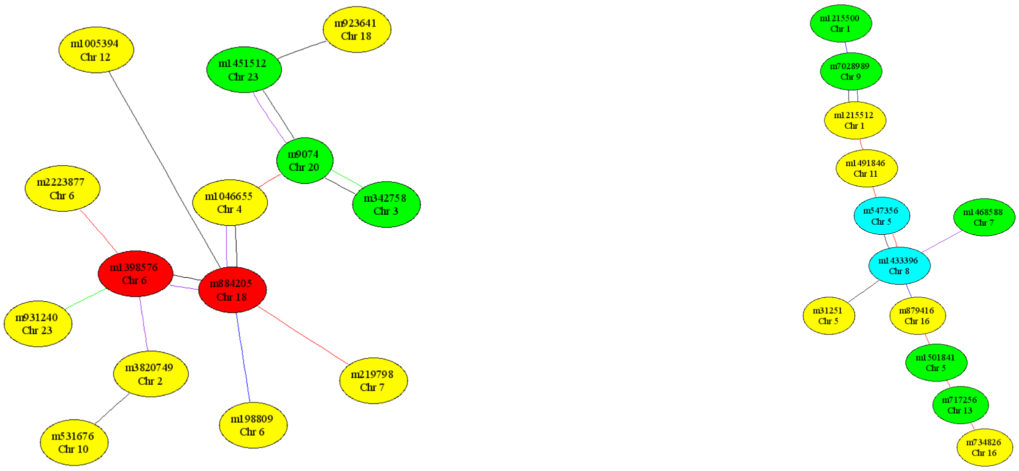


Figure 2.3
Examples of SNP epistasis network of a phenotype produced by the EPINET program that by default draws the 10 largest epistasis networks from the input test results. Line color: black = I-effect, red = A×A, purple = A×D, blue = D×A, green = D×D. Node color: red: $p<10^{-8}$, cyan: $p<10^{-7}$, green: $p<10^{-6}$, yellow: $p<10^{-5}$.

Chapter 3

Generalized least squares method to account for sib correlation for testing SNP single-locus and epistasis effects in genome-wide association analysis

The existence of a large number of single nucleotide polymorphisms (SNPs) provides opportunities to screen for DNA variations affecting complex traits using a genome-wide association (GWA) analysis. Family data are a commonly used data structure in genetic analysis. To account for correlations among individuals within the same family, a generalized least squares (GLS) method was developed for testing SNP single-locus and epistasis effects of a quantitative trait based on an extended Kempthorne model that allows Hardy-Weinberg disequilibrium and linkage disequilibrium. Simplified formulations were derived so that the most time consuming calculation in GLS analysis, the inverse of the phenotypic variance-covariance matrix, is no longer needed. Based on this GLS method, statistical tests were developed to test three single-locus effects for each SNP and five pairwise effects for each pair of SNPs. The GLS method has been implemented for supercomputer parallel computing for large scale pairwise epistasis testing.

3.1 Introduction

Genome-wide association (GWA) analysis of complex traits using a large number of SNP markers provides opportunities to study complex genetic effects such as epistasis effects. The work of Fisher (1918), Cockerham (1954) and Kempthorne (1954) laid the foundation for studying epistasis effects of quantitative traits using a quantitative genetics approach. Under the assumptions of Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE), Fisher partitioned the nine genotypic values of two loci into single gene and epistasis effects, and Cockerham and Kempthorne using different methods partitioned Fisher's epistasis effect into four components, additive × additive, additive × dominance, dominance × additive, and dominance × dominance epistasis effect with the genetic interpretation of allele × allele, allele × genotype, genotype × allele, and genotype × genotype interactions respectively. This partitioning can be used as a tool for identifying the exact mode of a gene interaction effect. Cockerham's partitioning uses orthogonal contrasts of genotypic values whereas Kempthorne's partitioning uses the deviation of "genetic combination" from the sum of individual genetic factors. For example, an additive × additive effect (combination effect of two alleles each from a different locus) is a deviation of the gametic mean from the two single allelic means. This is clearly a measure for the gene combination effect and hence has a more straightforward genetic interpretation of the epistasis effect. The Cockerham and Kempthorne epistasis models have been extended to allow more general assumptions such as linkage disequilibrium (LD) and Hardy-Weinberg disequilibrium (HWD) (Wang and Zeng, 2006; Mao et al., 2006). The method of extended Kempthorne model has been developed for detecting single-locus and pairwise epistasis effects in genome-wide association studies using SNP markers for random populations based on least squares (LS) analysis (Ma et al., 2008). In genetic studies, family data with two or more sibs may have sib correlations that are not accounted for by least squares analysis. Genetic variances contribute to sib correlations but can be difficult to have accurate estimation due to the need to invert matrices of various allelic and genotypic identity-by-decent probabilities (Kempthorne, 1954; Henderson, 1986, 1988). Environmental factors may also contribute to sib correlations but are sometimes confounded with genetic variances. For example, common environment effect and genetic effect are confounded in full sibs and cannot be separated using full sib data. Therefore, phenotypic correlation is a more complete measure for sib-correlation and is much easier to calculate than genetic variance components such as additive and dominance variance components. Since sib phenotypic values cannot be divided into dependent and independent variables, intra-class correlation is an appropriate measure of sib correlations.

The purpose of this study is develop a generalized least squares (GLS) approach to account for sib correlation measured by intra-class correlation based on the extended Kempthorne model to allow sib correlation for testing single-locus and epistasis effects in GWAS analysis. Both GLS and LS methods are unbiased linear estimators but GLS is also best linear estimator, where 'best' means minimal variance of the estimator. This is a desirable property for GWAS analysis because significant effects from GLS on average should be closer to the true effect locations than from LS effects.

3.2 Single-locus and epistasis effects of two bi-allelic loci

Eight single-locus and epistasis effects of two bi-allelic loci (SNP loci are assumed) in general populations where HWD and LD may exist (Mao et al., 2006) are defined from the extended Kempthorne model as follows:

$\alpha_1 = a_A - a_a = \mathrm{k}_2'\beta = \mathrm{s}_2\mathrm{g}$ = additive effect (gene substitution effect) of locus 1 (1.1)

$\alpha_2 = a_B - a_b = \mathrm{k}_3'\beta = \mathrm{s}_3\mathrm{g}$ = additive effect (gene substitution effect) of locus 2 (1.2)

$d_1 = d_{Aa} - (d_{AA} + d_{aa}) = \mathrm{k}_4'\beta = \mathrm{s}_4\mathrm{g}$ = dominance effect of locus 1 (1.3)

$d_2 = d_{Bb} - (d_{BB} + d_{bb}) = \mathrm{k}_5'\beta = \mathrm{s}_5\mathrm{g}$ = dominance effect of locus 2 (1.4)

$i_{aa} = (aa)_{AB} - (aa)_{Ab} - [(aa)_{aB} - (aa)_{ab}] = \mathrm{k}_6'\beta = \mathrm{s}_6\mathrm{g}$ = additive × additive effect (1.5)

$i_{ad} = \{(ad)_{ABb} - \frac{1}{2}[(ad)_{ABB} + (ad)_{Abb}]\} - \{(ad)_{aBb} - \frac{1}{2}[(ad)_{aBB} + (ad)_{abb}]\} = \mathrm{k}_7'\beta = \mathrm{s}_7\mathrm{g}$

= additive × dominance effect (1.6)

$i_{da} = \{(da)_{AaB} - \frac{1}{2}[(da)_{AAB} + (da)_{aaB}]\} - \{(da)_{Aab} - \frac{1}{2}[(da)_{AAb} + (da)_{aab}]\} = \mathrm{k}_8'\beta = \mathrm{s}_8\mathrm{g}$

= dominance × additive effect (1.7)

$i_{dd} = \{(dd)_{AaBb} - \frac{1}{2}[(dd)_{AaBB} + (dd)_{Aabb}]\} - \frac{1}{2}\{\{(dd)_{AABb} - \frac{1}{2}[(dd)_{AABB} + (dd)_{AAbb}]\}$
$+ \{(dd)_{aaBb} - \frac{1}{2}[(dd)_{aaBB} + (dd)_{aabb}]\}\} = \mathrm{k}_9'\beta = \mathrm{s}_9\mathrm{g}$ = dominance × dominance effect (1.8)

where k_j = contrast vector of β (j = 2, …9), β = column vector of the population mean and the 35 genetic effects, g = (column vector of the two-locus SNP genotypic values, and s_i = a function of marginal and conditional genotypic frequencies. Under the null hypothesis of no genetic effect, each of the four epistasis contrasts is expected to be zero. The above 8 genetic effects are orthogonal comparisons of the 35 individual genetic effects that have the following typical expressions:

$$a_i = \mu_i - \mu, \quad i = A, a;$$

$$a_k = \mu_k - \mu, \quad k = B, b;$$

$$d_{ij} = \mu_{ij} - \mu - a_i - a_j, \quad ij = AA, Aa, aa;$$

$$d_{kl} = \mu_{kl} - \mu - a_k - a_l, \quad kl = BB, Bb, bb\ ;$$

$$(aa)_{ik} = \mu_{ik} - \mu - a_i - a_k, \qquad k = AB, Ab, aB, ab;$$

$$(ad)_{ikl} = \mu i_{kl} - \mu - a_i - a_k - a_l - d_{kl} - (aa)_{ik} - (aa)_{il}, \qquad i = A, a;\ kl = BB, Bb, bb;$$

$$(da)_{ijk} = \mu_{ijk} - \mu - a_i - a_j - a_k - d_{ij} - (aa)_{ik} - (aa)_{jk}, \qquad ij = AA, Aa, aa;\ \ k = B, b\ ;$$

$$(dd)_{ijkl} = g_{ijkl} - \mu - a_i - a_j - a_k - a_l - d_{ij} - d_{kl} - (aa)_{ik} - (aa)_{il} - (aa)_{jk} - (aa)_{jl}$$

$$- (ad)_{ikl} - (ad)_{jkl} - (da)_{ijk} - (da)_{ijl}\ , \qquad ij = AA, Aa, aa;\ kl = BB, Bb, bb.$$

The 35 genetic effects represented by the above equations are the same as those under the assumptions of HWE and LE except that the calculations of the 27 population and marginal means to define the genetic effects are calculated under the assumptions of HWD and HD.

3.3 Generalized least squares (GLS) tests of SNP effects

The statistical test for each of the eight effects defined by Eqs. (1.1-1.8) requires estimating the nine genotypic values (g) by their estimates. To account for correlated individuals within families, we propose a generalized least squares method that uses intra-class correlation as a measure of sib correlations in testing the significance of each effect. The phenotypic values of a quantitative traits is assumed to be

$$y = X_1b + X_2g + Zf + e$$

where y = N x 1 vector of phenotypic values, X_1 = model matrix of fixed non-genetic effects, b = a column vector of fixed non-genetic effects such as gender, age, and smoking status, g = effects of the nine genotypes, X_2 = model matrix of g, f = random family effects with a common variance σ_f^2 for sibs in the same family that could include common genetic and environmental effects, Z = model matrix of f.

The variance-covariance matrix of the family effects is assumed to be $G = var(f) = I\sigma_f^2$. Then, the phenotypic variance-covariance matrix is $var(y) = V = ZGZ' + I\sigma_e^2$. The phenotypic values are assumed to follow a normal distribution with mean $X_1b + X_2g$ and variance-covariance matrix V, which can also be expressed as $\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \bigoplus_{i=1}^{m} \mathbf{V}_i$,

where V_i = variance-covariance matrix of phenotypic values of sibs in family i. All individuals are assumed to have the same variance (σ^2) and sibs within each family have the same covariance (c). The inverse of the V is the most computationally intensive operation. A simple formula of the V inverse is available so that the direct inversion is unnecessary. Let $\sigma^2 = \sigma_f^2 + \sigma_e^2$, $\rho = \sigma_f^2/\sigma^2$= intraclass correlation (Shrout and Fleiss, 1979; Kenneth and Wong, 1996) and n_i = the number of sibs in family i. Then, the n_i x n_i variance-covariance matrix of phenotypic values of sibs in family I and its inverse are:

$$\mathbf{V}_i^{-1} = (\mathbf{I} - r\mathbf{J})\frac{1}{\sigma_e^2}$$

with

$$r = \frac{\rho}{1 + (n_i - 1)\rho}$$

A simplied formula can be derived so that the V-inverse can be calculated without actually inverting V. Let V^{-1} = L'L, λ = r/(1-r), X_2^* = L'X_2 and y* = L'y, where L is an upper triangular matrix with

$$L_{ii} = \sqrt{\frac{(1+\lambda)\left[1-(i-1)\lambda\right]}{1-(i-2)\lambda}\frac{1-r}{\sigma_e^2}} \quad i = 1, \cdots, n$$

$$L_{ij} = -\lambda L_{ii} / [1-(i-1)\lambda], \quad i = 2,\cdots,n \quad j = i+1,\cdots,n-1$$

For simplicity of describing the test statistic of each genetic effect defined by Eqs.(1.1-1.8), fixed non-genetic effects are assumed to be estimated and removed from the phenotypic values. This assumption leads to a two-step regression analysis (Wolfinger et al., 2001; Ma et al., 2008), where the first step removes fixed non-genetic effects from the phenotypic values and the second step uses the residual values for subsequent analysis. Under this assumption, the generalized least squares estimator of the nine genotypic values is $\hat{\mathbf{g}} = \left(\mathbf{X}_2^{*\prime}\mathbf{X}_2^*\right)^{-1}\left(\mathbf{X}_2^{*\prime}\mathbf{y}^*\right) = \left(\mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_2\right)^{-1}\left(\mathbf{X}_2'\mathbf{V}^{-1}\mathbf{y}\right)$

where y* is the vector of corrected phenotypic values after removing fixed non-genetic effects. Let k = the rank of X. The test statistic for testing each genetic effect follows a Student-t distribution and is given by the following formula:

$$\hat{\mathrm{T}}_{\mathrm{X}} = \frac{L_{\mathrm{X}}}{\sqrt{\mathrm{V\hat{a}r}\left(L_{\mathrm{X}}\right)}} = \frac{\mathbf{s}_i\hat{\mathbf{g}}}{s\sqrt{\mathbf{s}_i\left(\mathbf{X}_2'\mathbf{V}^{-1}\mathbf{X}_2\right)^{-1}\mathbf{s}_i'}} = \frac{\mathbf{s}_i\hat{\mathbf{g}}}{s\sqrt{\mathbf{s}_i\left(\mathbf{X}_2^{*\prime}\mathbf{X}_2^*\right)^{-1}\mathbf{s}_i'}}$$

with

$$s^2 = \frac{\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \hat{\mathbf{g}}'\mathbf{X}_2'\mathbf{X}_2\hat{\mathbf{g}}}{N-k} = \frac{\mathbf{y}^{*\prime}\mathbf{y}^* - \hat{\mathbf{g}}'\mathbf{X}_2'\mathbf{X}_2\hat{\mathbf{g}}}{N-k}$$

3.4 Maximumlikelihood estimation of variance components

The GLS solution to the nine genotypic values (g) and the testing of each genetic effect requires values of the variance components in the V matrix. The variance components can be estimated using the following formulae of maximum likelihood method for estimating variance components:

$$\sigma_f^{2(j+1)} = \frac{\hat{\mathbf{f}}^{(j)\prime}\mathbf{G}^{(j)^{-1}}\left(\dfrac{\partial\mathbf{G}}{\partial\sigma_f^{2(j)}}\right)\mathbf{G}^{(j)^{-1}}\hat{\mathbf{f}}^{(j)}\sigma_e^{2(j)}\sigma_f^{2(j)}}{\mathbf{tr}\left\{\left(\dfrac{\partial\mathbf{G}}{\partial\sigma_f^{2(j)}}\right)\left[\mathbf{Z}'\mathbf{Z}-\mathbf{Z}'\mathbf{Z}\left(\mathbf{G}^{(j)^{-1}}\sigma_e^{2(j)}+\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Z}\right]\right\}}$$

$$\sigma_e^{2(j+1)} = \frac{\left(\mathbf{y}\text{ - }\mathbf{X}\widehat{\boldsymbol{\beta}}^{(j)}-\mathbf{Z}\hat{\mathbf{f}}^{(j)}\right)'\left(\mathbf{y}\text{ - }\mathbf{X}\widehat{\boldsymbol{\beta}}^{(j)}-\mathbf{Z}\hat{\mathbf{f}}^{(j)}\right)}{n-\mathrm{tr}\left[\mathbf{Z}'\mathbf{Z}\left(\mathbf{G}^{(j)^{-1}}\sigma_e^{2(j)}+\mathbf{Z}'\mathbf{Z}\right)^{-1}\right]}$$

The above estimation of variance components should be the first step of the GLS tests. Substituting the estimates of variance components in V matrix, the GLS testing of each genetic effect can be carried out.

3.5 Implementation for parallel and serial computing

The GLS method developed in this study has been implemented for parallel and serial computing and the resulting computer programs are available at the webpage of "Software Tools for Animal Gene Mapping" (http://animalgene.umn.edu). We have applied these computer programs to chicken data in collaboration with China Agricultural University and to the Framingham Heart Study data. Results showed that the GLS method was more accurate than the LS method for sib data, because the GLS method more accurately identified genes or chromosome locations that were confirmed in the literature or by on-going research. For example, In the Framingham Heart Study data, the two SNP effects on total cholestoral level near *CELSR2* and *PSRC1* or in *CELSR2* were ranked the most significant and second most significant effects by the GLS method but were ranked 6th and 15th by the LS method, noting that SNP effects near or in *CELSR2* and *PSRC1* had multiple confirmations in the literature.

Chapter 4

Detection of epistasis effects of nine dairy traits in contemporary U.S. Holstein cows

4.1 Introduction

Epistasis effects of net merit (NM) and its eight component traits were tested in 1654 contemporary U.S. Holstein cows using the BovineSNP50 chip (45,878 SNPs) and QTL maps of 450 epistasis effects involving 347 SNPs (top 50 effects per trait) were constructed. The top 50 epistasis effects of each trait explained 41-57% of the phenotypic variation of the trait. The 347 SNPs were distributed across 28 bovine autosomes and the X-chromosome. The X chromosome had the highest concentration of epistasis effects although it had the smallest number of SNP markers for analysis among all chromosomes: 25% of all 50 epistasis effects for milk yield (MY), 37% for fat yield (FY), 32% for protein yield (PY), 61% for fat percent (FPC), 54% for protein percent (PPC), 55% for somatic cell score (SCS), 63% for daughter pregnancy rate (DPR), 70% for productive life (PL), and 54% for NM. Epistasis effect between *PDE4B* (*phosphodiesterase 4B, cAMP-specific*) and *ROR1* (*receptor tyrosine kinase-like orphan receptor 1*) of chromosome 3 was the most significant epistasis effect for FY, PY and NM. X chromosome genes with or near most significant epistasis effects include *GRPR* (*gastrin-releasing peptide receptor*), *MAOB* (*monoamine oxidase B*), *FLNA* (*filamin A,* *alpha*), *LOC786985* (similar to *dystrophin*), *LOC520057* (similar to *type 1 protein phosphatase inhibitor*) and *SAT1*.

Net merit, component traits of net merit and calving traits are economically important traits of dairy cattle. Numerous studies of quantitative trait loci (QTL) associated with these traits (Ashwell, 2004; Ashwell, 2005; Ashwell, 1999; Ashwell, 2001; Baes, 2009; Boichard, 2003; Cole, 2009; Goertz, 2009) have been reported. Most of the previous dairy QTL reports used microsatellite markers. A genome scan of 1,536 single nucleotide polymorphism (SNP) markers selected from candidate genes using Canadian bulls (Kühn, 2003) studied six dairy functional traits, and a genome-wide analysis using the Illumina Bovine SNP50$^{\text{TM}}$ chip on U.S. Holstein bulls reported the size and distribution of QTL effects on all dairy traits with genetic evaluations (Cole, 2009). The availability of a large number of single nucleotide polymorphism (SNP) markers (Bovine HapMap Consortium, 2009; Wiggans, 2009) allows the construction of QTL maps with unprecedented high resolutions. Combined with bovine whole-genome sequence information, many SNP effects could be readily localized to specific genes or gene regions. Such high resolution QTL maps will provide valuable information for applying SNP markers in dairy breeding and selection practice and for understanding the genetic mechanism underlying dairy traits. Recently, results of single-locus analysis from genome-wide association studies (GWAS) of 31 dairy traits in contemporary U.S. Holstein cows using the Illumina BovineSNP50$^{\text{TM}}$ (Bovine HapMap Consortium, 2009) chip were reported (Cole et al., 2010a, 2010b; Wiggans et al., 2010; Ma et al., 2010). In this study, we analyzed epistasis effects of nine dairy traits in the aforementioned dairy GWAS (Cole et al., 2010a, 2010b; Wiggans et al., 2010; Ma et al., 2010).

4.2 Materials and Methods

The phenotypic data were predicted transmitting ability (PTA) of the nine dairy traits: milk yield (MY), fat yield (FY), protein yield (PY), fat percent (FPC), protein percent (PPC), somatic cell score (SCS), daughter pregnancy rate (DPR), productive life (PL), and lifetime net merit (NM). The study population included 1654 contemporary U.S. Holstein cows from Holstein Association USA (Brattleboro, VT), Genetic Visions (Middleton, WI), Genex Cooperative Inc. (Shawano, WI), Iowa State University (Ames, IA), Pennsylvania State University (University Park, PA), Virginia Polytechnic Institute and State University (Blacksburg, VA), University of Florida (Gainesville, FL), and University of Minnesota (St. Paul, MN). A total of 45,878 SNP markers from the Illumina BovineSNP50^{TM} chip (18) were selected based on two conditions: the minor allele frequency (MAF) was greater than 0 in the contemporary population, or the allele frequency difference was 2% or greater between the 1654 Holstein cows and a group of 301 Holstein cattle that remained unselected since 1964. Of the 45,878 SNP markers, 45,461 had known chromosome positions with average marker spacing of 58.45Kb.

DNA extraction and SNP genotyping were performed at the Bovine Functional Genomics Laboratory (Beltsville, MD), and marker genotypes were scored using Illumina's GenomeStudio software (v1.1.9). Statistical tests of QTL effects were implemented using the epiSNP computer package (Ma, 2008a) which implements the extended Kempthorne model that allows linkage disequilibrium between SNPs and Hardy-Weinberg disequilibrium of each SNP (Mao, 2006). For each SNP, two effects were tested, genotypic effect, and additive effects. Dominance effects also were investigated, but not expected to be significant because the traditional genetic evaluations used as input contain only additive effects. A genome-wide 5% type-I error with the Bonferroni correction was considered as the threshold p-value ($p < 4.75\times10^{-11}$) to declare genome-wide significance. SNP and gene locations were identified based on University of Maryland bovine genome assembly (UMD 3.0). SNP locations based on the Baylor College of Medicine bovine genome assembly Build 4.0 (Btau_4.0) from ENSEMBL and NCBI also were noted in the results. Figures of gene clusters were from ENSEMBL based on Btau_4.0 because such figures based on the UMD assembly were not yet available from ENSEMBL. The contribution of the top 100 SNP effects of each trait was measured by the coefficient of determination (R^2) and calculated using the linear regression procedure (PROC REG) of SAS.

4.3 Results and Discussion

Significant epistatic SNP effects: A large number of significant epistatic SNP effects reached genome-wide significance with the Bonferroni correction ($p < 4.75\times10^{-11}$). Due to the large number of epistasis effects exceeding the genome-wide significance, only top 50 effects of each trait are summarized here. SNPs with 50 epistasis effects of each trait accounted for 0.41-0.57 of the phenotypic variation of the trait (R^2 =0.41~0.57, Table

4.1). The 450 epistasis effects involved 347 SNPs on 28 autosomes and the X chromosome (Figure 4.1).

Heavy X-chromnsome involvement in epistasis effects: The X chromosome had the largest number of epistasis effects, account for about 50% of epistasis effects, ranging from 25% for MY to 70% for PL (Table 4.1), and had the most significant epistasis effects for 7 of the 9 traits except MY and PY, for which BTA12 was most significant for MY and BTA26 was most significant for PY (Table 4.2). BTA17 had the second largest number of epistasis effects, primarily involving yield traits (MY, FY, PY, FPC, PPC) and NM. BTA26 had the most significant epistasis effect for PY and had several single-locus effects on PY and PPC (reported separately). Three pairs of BTA14 SNPs with epistasis effects were near the region with single-locus effects on FPC, and one BTA7 SNP with epistasis effect was 300Kb downstream the most significant single-locus effect for SCS. QTL map of 381 SNPs with 450 epistatic SNP effects: The 450 SNP effects of 381 SNPs involved 28 Bos taurus (BTA) autosomes and the X chromosome (Figure 4.1) but X chromsome SNPs accounted for nearly 50% of the 450 SNP pairs, although the X chromosome had the smallest number of SNPs among all chromsomes. BTA17, BTA26, BTA1, BTA7 and BTA12 had localized concentrations of epistasis effects. BTA24 did not have significant epistasis effects. X chromosome and autosome genes: Of the 381 SNPs, 111 SNPs were in 107 genes (based on BTA_4.0). Of the 45,878 SNPs, 36% (16,516 SNPs) were located in 7434 coding genes. X chromosome genes with or near most significant epistasis effects that may involve a different chromosome include:

GRPR (*Gastrin-releasing peptide receptor)* and *SAT1 (Spermidine/spermine N1-acetyltransferase 1*) for DPR and PL; *LOC786985* (similar to *Dystrophin*) and *FLNA* (*Filamin A, alpha*) for SCS; *FLNA* for FY, FPC; *SLITRK4* (*SLIT and NTRK-like family, member 4*) and *TSPAN7* (Tetraspanin 7) for MY, *GRP50* (*G protein-coupled receptor 50*) for PY; *LOC520057* (similar to *Type 1 protein phosphatase inhibitor*) and *MAOB* (*Monoamine oxidase B*) for PPC, and *FGF16* (*Fibroblast growth factor 16*) and *NUDT11* (*Nudix (nucleoside diphosphate linked moiety X)-type motif 11*) for PL. Autosome genes with or near most significant epistasis effects include: *DUSP5* (*Dual specificity protein phosphatase 5*) of BTA 26 for PY (#1), *LOC614423* (similar to *phosphoglycerate mutase 1 (brain)*) and *ELN* (*Elastin*) for MY (#1 and #2), *LOC614109* (similar to *Solute carrier family 10*) and *AACS* (*Acetoacetyl-CoA synthetase*) of BTA17 for FY and PY (#2), *PCDH9* (*Protocadherin 9)* of BTA12 for SCS, *SORL1* (*Sortilin-related receptor, L(DLR class) A repeats-containing*) for PPC (#2), and *LOC785689* (similar to *Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4*) of BTA10 for FY (#2). A SNP in *ELAVL4* (*(embryonic lethal, abnormal vision, Drosophila)-like 4*) of BTA3 and a SNP near *ZFYVE9* of BTA3 were the 49th most significant epistasis effect for DPR. The full list of the 347 SNPs with 450 effects will be available to the public either as supplementary information to a journal article or through a website release.

The most striking result from the epistasis analysis was the heavy X chromosome involvement in epistasis effects, approximately accounting for 50% of the epistasis effects. This phenomenon was observed in various data analysis stages as more data became available, e.g., the analysis based on the University of Minnesota control and selection lines, the analysis based on about 1000 cows, and this final analysis based on all 1654 contemporary Holstein cows. The heavy involvement of the X chromosome in epistasis effects tend to point to regulatory roles of the X chromosome in dairy traits. Since one X chromosome is nearly all the genetic difference between a cow and a bull and all dairy traits in this study are female phenotypes, X chromosome regulatory roles in the female phenotypes should be likely.

Table 4.1. Distribution of top 50 most significant pairs for each of the 9 dairy traits by chromosome.

Chromosome	MY	FY	PY	FPC	PPC	NM	PL	SCS	DPR	SUM
1	1	1	7	0	5	6	2	1	5	**28**
2	4	2	1	0	1	0	0	0	0	8
3	4	8	8	0	**11**	8	1	1	9	**50**
4	2	2	0	0	0	0	0	1	0	5
5	1	0	0	1	0	4	2	3	0	**11**
6	2	0	0	3	0	1	0	9	1	**16**
7	0	0	3	1	0	4	4	1	2	**15**
8	7	0	0	0	0	1	0	1	0	9
9	4	3	3	0	0	0	0	3	1	**14**
10	0	8	4	0	0	2	2	0	6	**22**
11	8	0	2	0	0	0	0	2	0	**12**
12	3	2	2	2	6	4	2	**16**	4	**41**
13	2	0	1	0	0	1	0	0	0	4
14	7	1	2	2	0	2	2	0	0	**16**
15	1	2	0	9	9	0	0	1	0	**22**
16	2	0	0	0	0	0	0	0	0	2
17	9	**27**	**13**	5	1	9	1	0	0	**65**
18	0	1	0	3	5	0	0	0	2	**11**
19	1	0	0	0	0	0	0	0	0	1
20	0	0	1	4	0	1	5	0	2	**13**
21	2	0	0	0	0	0	0	0	1	3
22	2	0	0	0	0	0	0	1	1	4
23	2	0	6	3	1	1	7	0	0	**20**
24	2	0	0	0	1	0	0	2	0	5
25	2	0	3	0	0	0	0	0	0	5
26	0	5	**11**	6	2	0	0	0	0	**24**
27	0	1	1	0	2	2	0	1	0	7
28	0	0	0	0	0	0	0	0	0	0
29	3	0	0	0	2	0	0	0	0	5
X	**25**	**37**	**32**	**61**	**54**	**54**	**70**	**55**	**63**	**451**
U	4	0	0	0	0	0	2	2	3	**11**
P[a]	10^{-11}	10^{-21}	10^{-18}	10^{-17}	10^{-24}	10^{-29}	10^{-26}	10^{-20}	10^{-20}	
R^2	0.55	0.51	0.48	0.41	0.44	0.57	0.51	0.49	0.53	

[a]This is the rounded cut-off p-value for the top 50 significant epistasis effects

Table 4.2. Top 10 epistasis effects for milk yield (MY)

Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
9	25649959	HINT3	0.17	9	25793691	80Kb U HEY2	0.19	2.10×10^{-19}	AA	1.70×10^{-21}	650±68
14	3802542	1.8Kb D DENND3	0.18	16	64764674	9.9Kb U LOC100139054	0.07	1.50×10^{-14}	AA	1.00×10^{-16}	560±67
14	33999209	41.3Kb D PREX2	0.16	X	92552271	34.8Kb D PAGE4	0.06	1.80×10^{-14}	AA	9.40×10^{-17}	590±71
14	34639444	LOC615152	0.07	24	21262447	5.5Kb D MOCOS	0.19	8.80×10^{-15}	AA	1.10×10^{-16}	-560±66
17	991486	Blank	0.24	17	9336564	Blank	0.46	1.40×10^{-14}	AA	1.10×10^{-16}	-280±33
17	9336564	Blank	0.46	17	1041832	Blank	0.21	1.80×10^{-14}	AA	1.40×10^{-16}	-290±35
23	14580054	227Kb D LRFN2	0.34	23	14658449	Blank	0.27	5.20×10^{-15}	AA	1.50×10^{-17}	280±32
29	20199725	56Kb U MGC157332	0.32	Un	Un	LOC787351	0.09	2.90×10^{-15}	AA	1.50×10^{-17}	440±51
X	5764620	134Kb U LOC100140451	0.21	X	48657807	Blank	0.23	1.30×10^{-14}	AA	1.30×10^{-16}	330±39
X	92552271	34.8Kb D PAGE4	0.06	X	117435015	Blank	0.12	1.40×10^{-15}	AA	2.30×10^{-17}	630±73

Table 4.3. Top 10 epistasis effects for fat yield (FY)

Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
3	79333053	PDE4B	0.26	3	81745268	ROR1	0.18	1.70×10^{-33}	AA	2.80×10^{-36}	29±2.3
3	79333053	PDE4B	0.26	3	81842704	ROR1	0.18	5.10×10^{-33}	AA	8.30×10^{-36}	29±2.3
3	79378528	PDE4B	0.25	3	81745268	ROR1	0.18	6.30×10^{-34}	AA	2.30×10^{-36}	-29±2.3
3	79378528	PDE4B	0.25	3	81842704	ROR1	0.18	1.90×10^{-33}	AA	7.20×10^{-36}	-29±2.3
10	35312701	1.3Kb U THBS1	0.23	17	12258111	8Kb D LOC614109	0.14	8.50×10^{-32}	AA	7.00×10^{-34}	22±1.8
10	36946284	NDUFAF1	0.23	17	12258111	8Kb D LOC614109	0.14	2.00×10^{-30}	AA	8.10×10^{-33}	22±1.8
10	37022108	RTF1	0.23	17	12258111	8Kb D LOC614109	0.14	2.00×10^{-30}	AA	8.10×10^{-33}	22±1.8
17	12258111	8Kb D LOC614109	0.14	17	52960684	13.4Kb U AACS	0.2	2.40×10^{-31}	AA	2.10×10^{-33}	24±1.9
X	85589749	Blank	0.2	X	109172225	Blank	0.32	9.80×10^{-33}	AA	2.70×10^{-35}	-20±1.6
X	85589749	Blank	0.2	X	142828641	1Kb D CLCN4	0.38	5.90×10^{-30}	AA	2.20×10^{-32}	-18±1.5

Table 4.4. Top 10 epistasis effects for protein yield (PY)

Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
3	79333053	PDE4B	0.26	3	81745268	ROR1	0.18	7.80×10^{-33}	AA	2.30×10^{-35}	21±1.6
3	79333053	PDE4B	0.26	3	81842704	ROR1	0.18	1.50×10^{-32}	AA	4.40×10^{-35}	21±1.6
3	79378528	PDE4B	0.25	3	81745268	ROR1	0.18	1.10×10^{-32}	AA	8.90×10^{-35}	-21±1.6
3	79378528	PDE4B	0.25	3	81842704	ROR1	0.18	2.10×10^{-32}	AA	1.70×10^{-34}	-21±1.6
10	35048321	262Kb U THBS1	0.37	26	2915575	14.8Kb D LOC785617	0.09	1.70×10^{-27}	AA	8.40×10^{-29}	-17±1.5
17	12258111	8Kb D LOC614109	0.14	17	52960684	13.4Kb U AACS	0.2	5.30×10^{-27}	AA	1.70×10^{-28}	16±1.4
23	14170863	138Kb U LRFN2	0.44	23	14372603	21.5Kb D LRFN2	0.5	2.30×10^{-24}	AA	1.50×10^{-28}	15±1.3
23	14215024	93.5Kb U LRFN2	0.45	23	14372603	21.5Kb D LRFN2	0.5	3.70×10^{-26}	AA	1.40×10^{-30}	15±1.3
26	29972396	Blank	0.14	26	31267471	59.4Kb U DUSP5	0.19	1.20×10^{-26}	AA	1.20×10^{-28}	19±1.6
X	106241123	30.9Kb U LOC520057	0.13	X	106393188	121Kb U LOC520057	0.15	2.40×10^{-28}	AA	3.70×10^{-29}	19±1.6

Table 4.5. Top 10 epistasis effects for fat percent (FPC)

Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
6	9062503	12.3Kb D LOC526064	0.21	X	106241123	30.9Kb U LOC520057	0.13	1.10×10^{-25}	AA	8.30×10^{-29}	0.062±0.0055
12	8201659	Blank	0.2	X	106241123	30.9Kb U LOC520057	0.13	5.20×10^{-26}	AA	3.00×10^{-28}	0.059±0.0053
15	32637662	SORL1	0.24	X	106241123	30.9Kb U LOC520057	0.13	1.10×10^{-24}	AA	5.10×10^{-28}	0.059±0.0052
15	34171356	31.6Kb D BSX	0.17	X	106241123	30.9Kb U LOC520057	0.13	3.40×10^{-26}	AA	1.40×10^{-29}	0.064±0.0056
23	32975635	ALDH5A1	0.17	X	106241123	30.9Kb U LOC520057	0.13	1.10×10^{-24}	AA	3.30×10^{-28}	-0.062±0.0056
26	29972396	Blank	0.14	X	106241123	30.9Kb U LOC520057	0.13	1.50×10^{-27}	AA	4.90×10^{-31}	0.074±0.0063
26	29972396	Blank	0.14	X	136196550	Blank	0.12	6.10×10^{-27}	AA	6.40×10^{-29}	-0.07±0.0062
26	29972396	Blank	0.14	X	136328915	Blank	0.13	3.20×10^{-26}	AA	2.40×10^{-28}	-0.069±0.0061
X	27387980	232Kb D SLITRK2	0.25	X	40319976	FLNA	0.17	3.70×10^{-28}	AA	4.50×10^{-29}	-0.054±0.0048
X	106241123	30.9Kb U LOC520057	0.13	X	106783368	Blank	0.34	1.80×10^{-25}	AA	4.10×10^{-28}	-0.056±0.0050

Table 4.6. Top 10 epistasis effects for protein percent (PPC)

Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
3	9579325	PEA15	0.17	X	106241123	30.9Kb U LOC520057	0.13	1.20×10^{-37}	AA	1.40×10^{-40}	-0.038±0.0028
12	8201659	Blank	0.2	X	106241123	30.9Kb U LOC520057	0.13	1.10×10^{-41}	AA	6.20×10^{-44}	0.036±0.0025
15	32713410	SORL1	0.23	X	106241123	30.9Kb U LOC520057	0.13	2.20×10^{-36}	AA	1.20×10^{-40}	0.035±0.0025
15	34171356	31.6Kb D BSX	0.17	X	106241123	30.9Kb U LOC520057	0.13	7.40×10^{-47}	AA	1.60×10^{-50}	0.041±0.0026
15	34171356	31.6Kb D BSX	0.17	X	106280810	8Kb D LOC520057	0.11	5.90×10^{-40}	AA	2.20×10^{-43}	-0.041±0.0029
15	34171356	31.6Kb D BSX	0.17	X	126985234	Blank	0.2	5.60×10^{-41}	AA	4.30×10^{-43}	-0.032±0.0023
15	34171356	31.6Kb D BSX	0.17	X	127636818	82.4Kb U LOC783117	0.2	1.70×10^{-41}	AA	6.30×10^{-44}	0.033±0.0023
17	33925677	Blank	0.15	X	105267785	MAOB	0.11	6.90×10^{-42}	AA	9.50×10^{-41}	-0.04±0.0029
X	27387980	232Kb D SLITRK2	0.25	X	40319976	FLNA	0.17	4.70×10^{-41}	AA	9.60×10^{-43}	-0.032±0.0022
X	105267785	MAOB	0.11	X	106241123	30.9Kb U LOC520057	0.13	1.10×10^{-38}	AA	1.60×10^{-41}	-0.044±0.0032

Table 4.7. Top 10 epistasis effects for somatic cell score (SCS)

Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
6	72382208	85.8Kb U SRD5A3	0.49	X	134602363	4.5Kb U GRPR	0.15	1.60×10^{-27}	AA	8.00×10^{-31}	0.14±0.012
7	93834124	Blank	0.15	X	134602363	4.5Kb U GRPR	0.15	6.70×10^{-29}	AA	4.80×10^{-30}	0.18±0.015
11	25544002	31.6Kb U LOC100139773	0.33	X	40319976	FLNA	0.17	1.50×10^{-28}	AA	3.80×10^{-30}	-0.14±0.012
12	40559886	PCDH9	0.28	X	115968607	LOC786985	0.16	5.30×10^{-34}	AA	5.30×10^{-35}	-0.17±0.013
12	45358430	Blank	0.18	X	115968607	LOC786985	0.16	9.20×10^{-34}	AA	7.40×10^{-34}	-0.18±0.015
12	46867626	DACH1	0.19	X	115968607	LOC786985	0.16	2.30×10^{-30}	AA	1.00×10^{-30}	0.17±0.014
12	49421215	Blank	0.15	X	115968607	LOC786985	0.16	5.50×10^{-32}	AA	4.20×10^{-32}	-0.19±0.016
12	52240216	111Kb U KCTD12	0.12	X	115968607	LOC786985	0.16	6.30×10^{-37}	AA	2.80×10^{-36}	0.21±0.016
12	84180733	83.5Kb U LOC100139606	0.16	X	115968607	LOC786985	0.16	6.40×10^{-28}	AA	9.70×10^{-31}	0.18±0.015
X	116335568	72.2Kb U LOC786944	0.16	X	134602363	4.5Kb U GRPR	0.15	5.40×10^{-28}	AA	1.80×10^{-31}	-0.18±0.016

Table 4.8. Top 10 epistasis effects for daughter pregnancy rate (DPR)

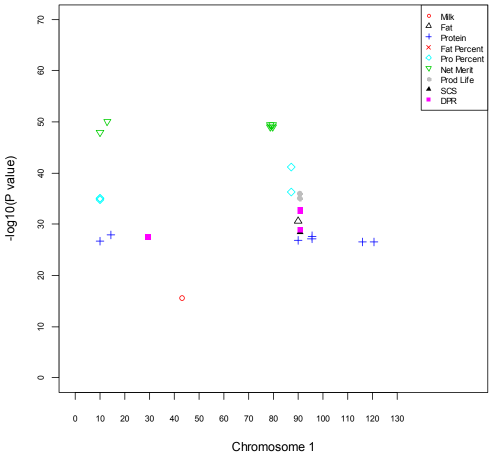
Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
1	91031267	Blank	0.27	1	89243673	111Kb U LOC784483	0.11	1.70×10^{-30}	AA	7.90×10^{-33}	1.6±0.13
1	91066621	Blank	0.26	1	89243673	111Kb U LOC784483	0.11	5.70×10^{-31}	AA	6.30×10^{-33}	1.6±0.13
12	45358430	Blank	0.18	X	59080285	Blank	0.19	2.00×10^{-29}	AA	6.10×10^{-33}	-1.3±0.11
21	68894872	LOC530121	0.24	X	48575295	Blank	0.3	6.70×10^{-32}	AA	5.40×10^{-33}	1±0.083
X	5471559	217Kb D C1GALT1C1	0.17	X	15464441	6.7Kb U LOC786185	0.12	2.90×10^{-37}	AA	1.00×10^{-39}	-1.7±0.13
X	8095146	Blank	0.21	X	15464441	6.7Kb U LOC786185	0.12	1.10×10^{-31}	AA	3.50×10^{-34}	1.4±0.11
X	9248137	81.9Kb U LOC785262	0.25	X	13523881	22.8Kb D OCRL	0.19	2.20×10^{-31}	AA	2.00×10^{-32}	1.4±0.11
X	40319976	FLNA	0.17	X	126985234	Blank	0.2	2.90×10^{-35}	AA	1.20×10^{-33}	1.4±0.11
X	40319976	FLNA	0.17	X	127636818	82.4Kb U LOC783117	0.2	2.40×10^{-36}	AA	1.80×10^{-34}	-1.4±0.11
X	116335568	72.2Kb U LOC786944	0.16	X	134602363	4.5Kb U GRPR	0.15	2.10×10^{-37}	AA	8.70×10^{-40}	1.7±0.12

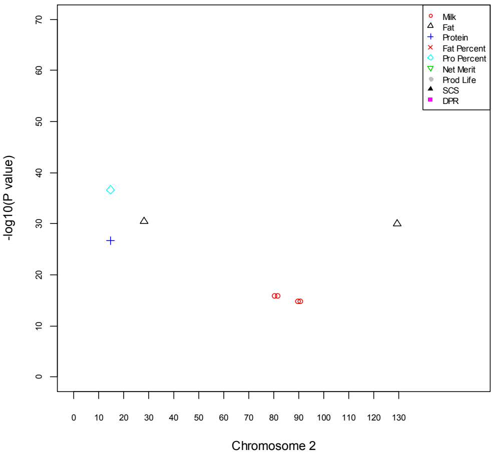
Table 4.9. Top 10 epistasis effects for productive life (PL)

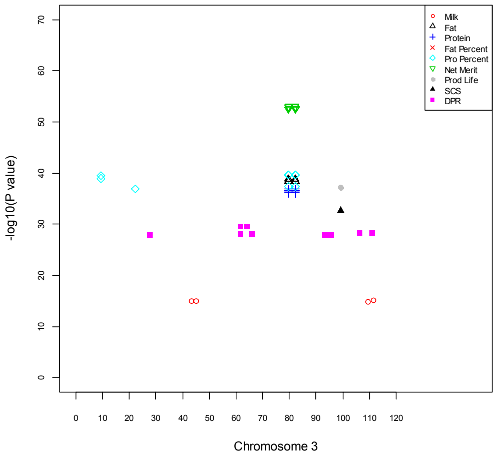
Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
7	93834124	Blank	0.15	23	32975635	ALDH5A1	0.17	1.30×10^{-36}	AA	1.60×10^{-37}	-2.2±0.17
10	36946284	NDUFAF1	0.23	X	2199470	Blank	0.21	1.90×10^{-38}	AA	2.00×10^{-37}	-1.8±0.14
10	37022108	RTF1	0.23	X	2199470	Blank	0.21	1.90×10^{-38}	AA	2.00×10^{-37}	-1.8±0.14
12	45358430	Blank	0.18	X	59080285	Blank	0.19	5.80×10^{-35}	AA	1.00×10^{-37}	-2.1±0.16
X	7042383	136Kb U GRIA3	0.17	X	60363808	Blank	0.2	2.90×10^{-34}	AA	9.30×10^{-36}	1.9±0.15
X	34049234	53.5Kb U VMA21	0.27	X	105132201	NDP	0.24	3.30×10^{-37}	AA	3.80×10^{-37}	-1.6±0.12
X	40319976	FLNA	0.17	X	127636818	82.4Kb U LOC783117	0.2	3.60×10^{-37}	AA	4.40×10^{-36}	-2.1±0.16
X	105132201	NDP	0.24	X	126472377	1.9Kb U SAT1	0.27	6.90×10^{-36}	AA	1.20×10^{-37}	-1.6±0.12
X	116335568	72.2Kb U LOC786944	0.16	X	134602363	4.5Kb U GRPR	0.15	9.80×10^{-36}	AA	2.10×10^{-38}	2.4±0.18
X	121076442	Blank	0.24	X	134602363	4.5Kb U GRPR	0.15	3.10×10^{-35}	AA	8.80×10^{-37}	2±0.15

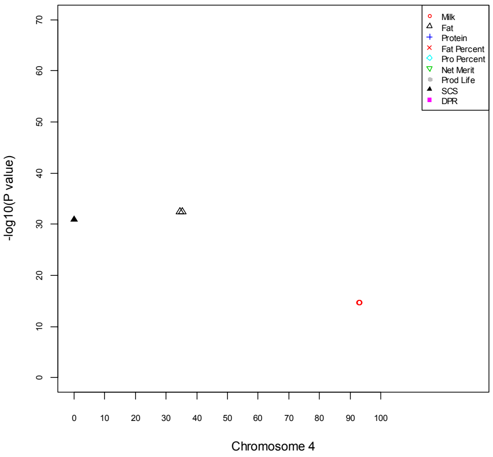
Table 4.10. Top 10 epistasis effects for net merit (NM)

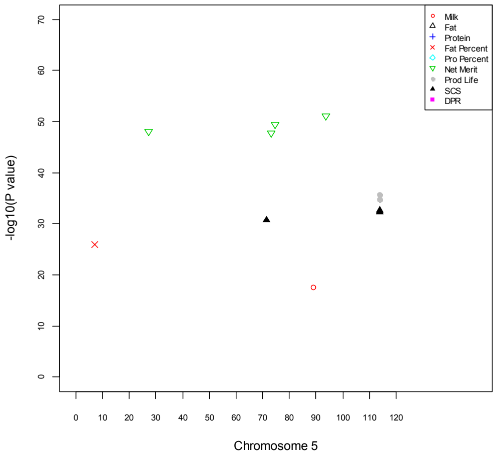
Chr1	Pos1	Gene1	MAF1	Chr2	Pos2	Gene2	MAF2	p-value			Effect±SD
								Genotype	Epistasis		
3	79333053	PDE4B	0.26	3	81745268	ROR1	0.18	3.10×10^{-45}	AA	1.20×10^{-48}	260±17
3	79333053	PDE4B	0.26	3	81842704	ROR1	0.18	7.70×10^{-45}	AA	3.00×10^{-48}	260±17
3	79378528	PDE4B	0.25	3	81745268	ROR1	0.18	4.80×10^{-46}	AA	3.40×10^{-49}	-260±17
3	79378528	PDE4B	0.25	3	81842704	ROR1	0.18	1.20×10^{-45}	AA	8.90×10^{-49}	-260±17
X	5471559	217Kb D C1GALT1C1	0.17	X	60363808	Blank	0.2	5.30×10^{-52}	AA	9.80×10^{-52}	-200±13
X	5471559	217Kb D C1GALT1C1	0.17	X	105267785	MAOB	0.11	4.70×10^{-50}	AA	6.90×10^{-49}	250±16
X	7042383	136Kb U GRIA3	0.17	X	60363808	Blank	0.2	6.40×10^{-48}	AA	4.40×10^{-49}	190±13
X	91958434	2.5Kb U LOC100141036	0.45	X	92853261	94Kb U CLCN5	0.31	1.00×10^{-53}	AA	1.50×10^{-53}	190±12
X	94321737	102Kb D NUDT11	0.11	X	103580973	260Kb U LOC100141229	0.2	2.90×10^{-48}	AA	1.40×10^{-48}	240±16
X	117783403	Blank	0.38	X	118022300	Blank	0.36	3.10×10^{-49}	AA	6.80×10^{-50}	160±10

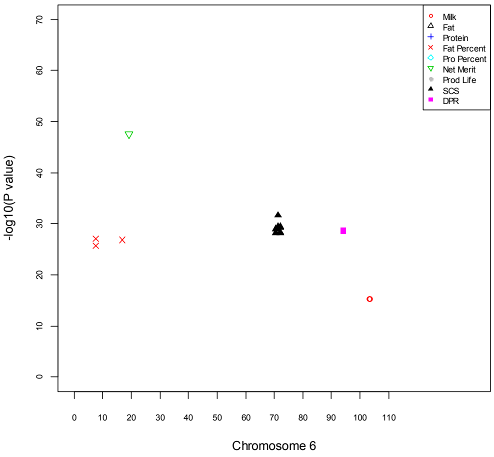

-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0 10 20 30 40 50 60 70 80 90 100 110 120 130
Chromosome 1

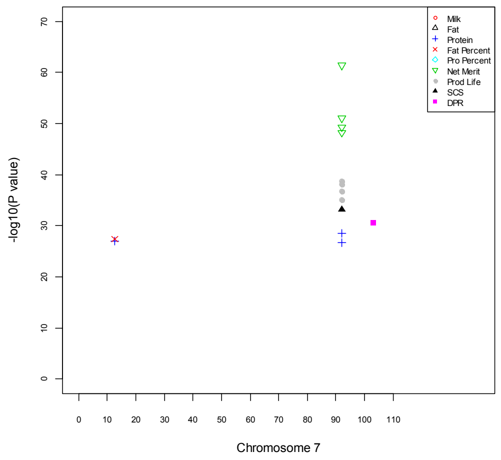

-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0 10 20 30 40 50 60 70 80 90 100 110 120 130
Chromosome 2


-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0 10 20 30 40 50 60 70 80 90 100 110 120
Chromosome 3

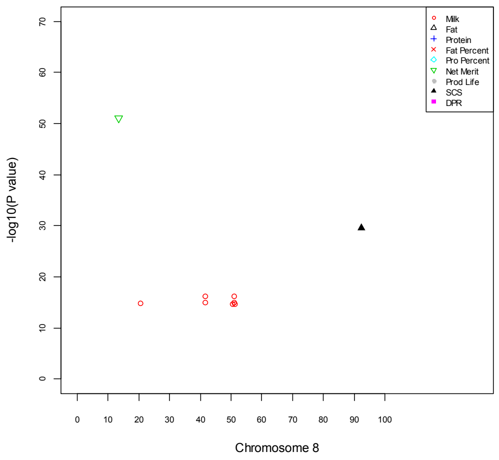

-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0 10 20 30 40 50 60 70 80 90 100
Chromosome 4


-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0 10 20 30 40 50 60 70 80 90 100 110 120
Chromosome 5

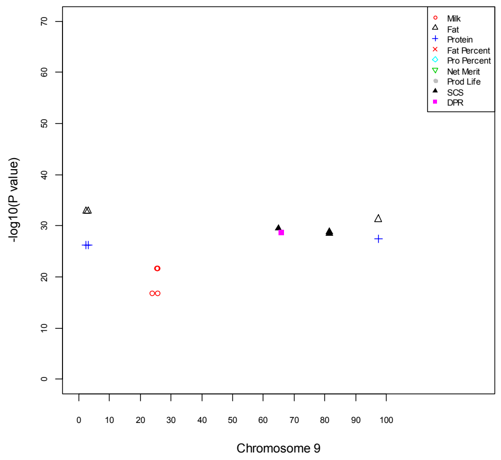

-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0 10 20 30 40 50 60 70 80 90 100 110
Chromosome 6



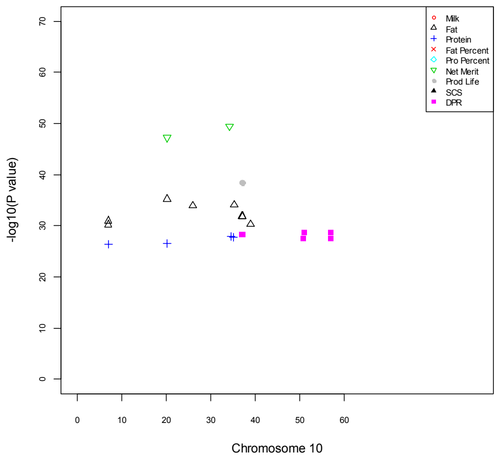
-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
70
80
90
100
110
Chromosome 7



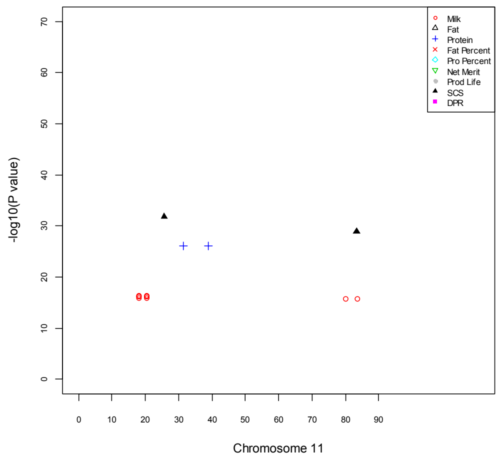
-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
70
80
90
100
Chromosome 8



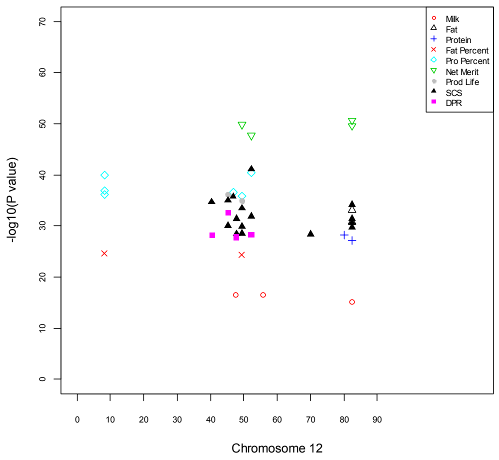
-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
70
80
90
100
Chromosome 9



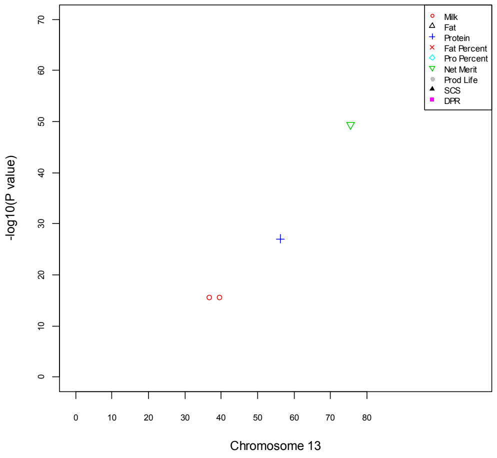
-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
Chromosome 10

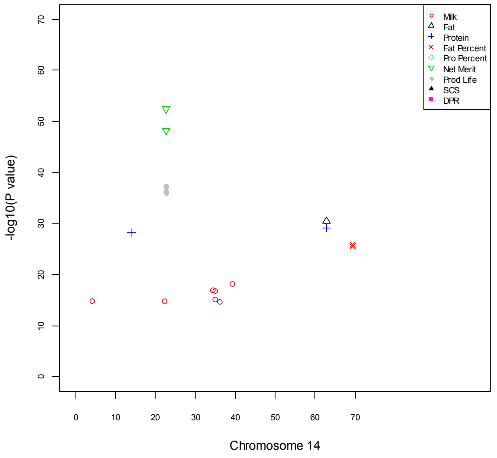


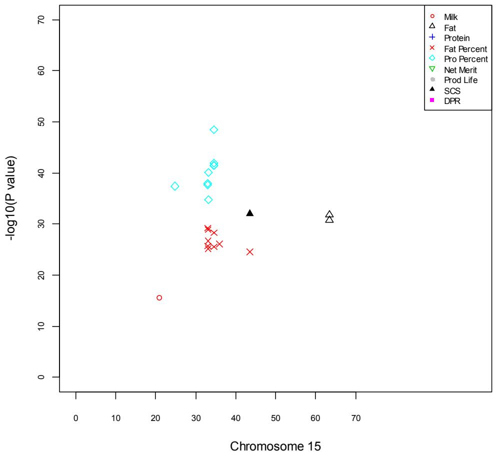
-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
70
80
90
Chromosome 11

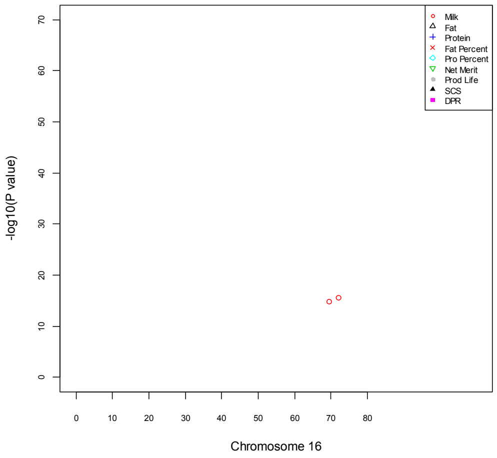


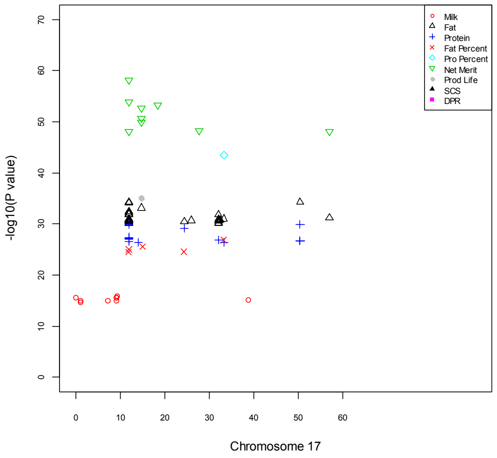
-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
70
80
90
Chromosome 12

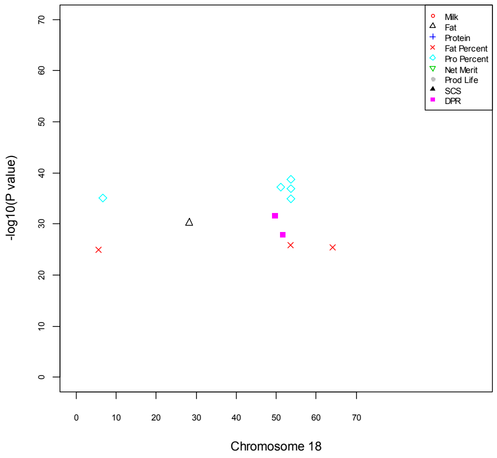

Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
-log10(P value)
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60 70 80
Chromosome 13

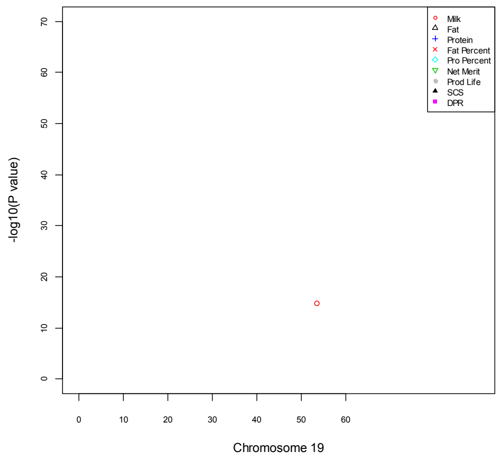

Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
-log10(P value)
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60 70
Chromosome 14

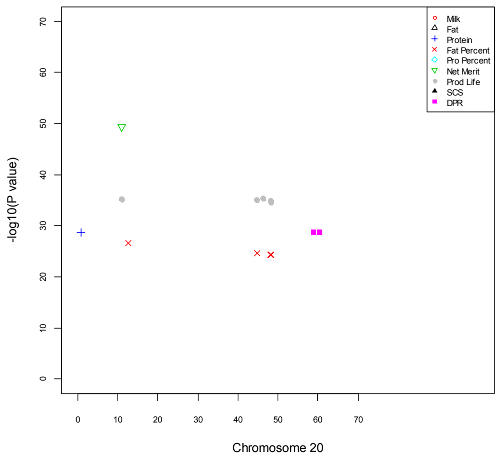

Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
-log10(P value)
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60 70
Chromosome 15

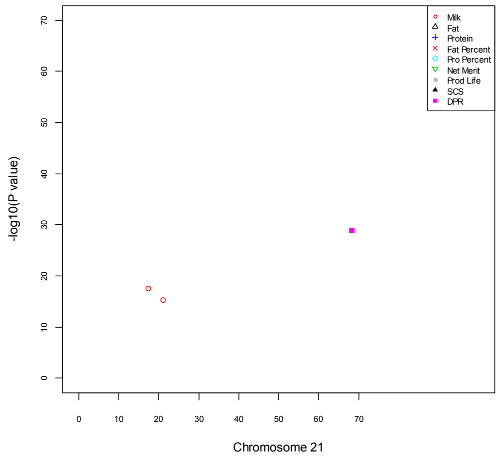

Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
-log10(P value)
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60 70 80
Chromosome 16

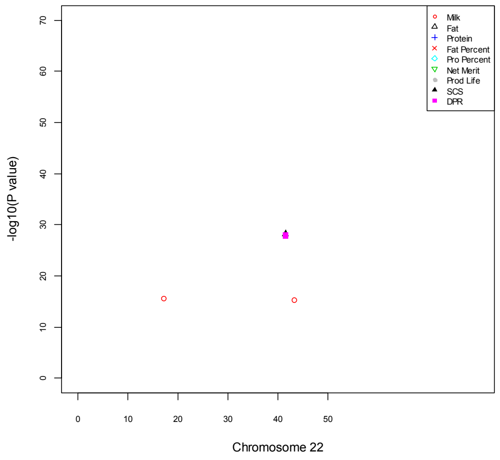

Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
-log10(P value)
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60
Chromosome 17

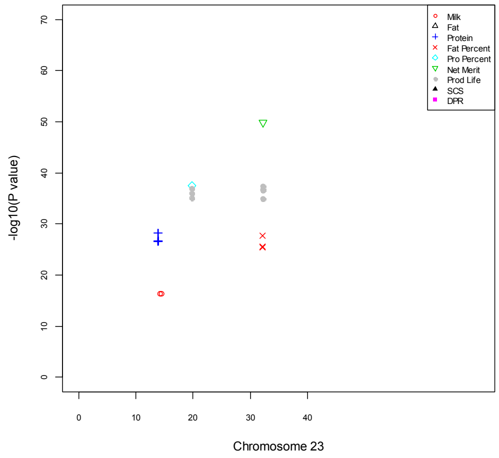

Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
-log10(P value)
0 10 20 30 40 50 60 70
0 10 20 30 40 50 60 70
Chromosome 18

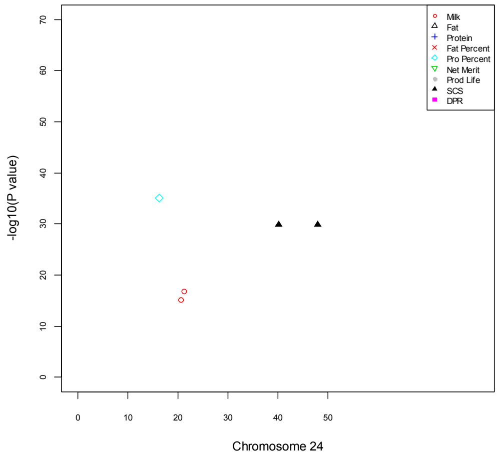

-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
Chromosome 19


-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
70
Chromosome 20


-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
60
70
Chromosome 21


-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
Chromosome 22


-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
Chromosome 23


-log10(P value)
0
10
20
30
40
50
60
70
Milk
Fat
Protein
Fat Percent
Pro Percent
Net Merit
Prod Life
SCS
DPR
0
10
20
30
40
50
Chromosome 24

\

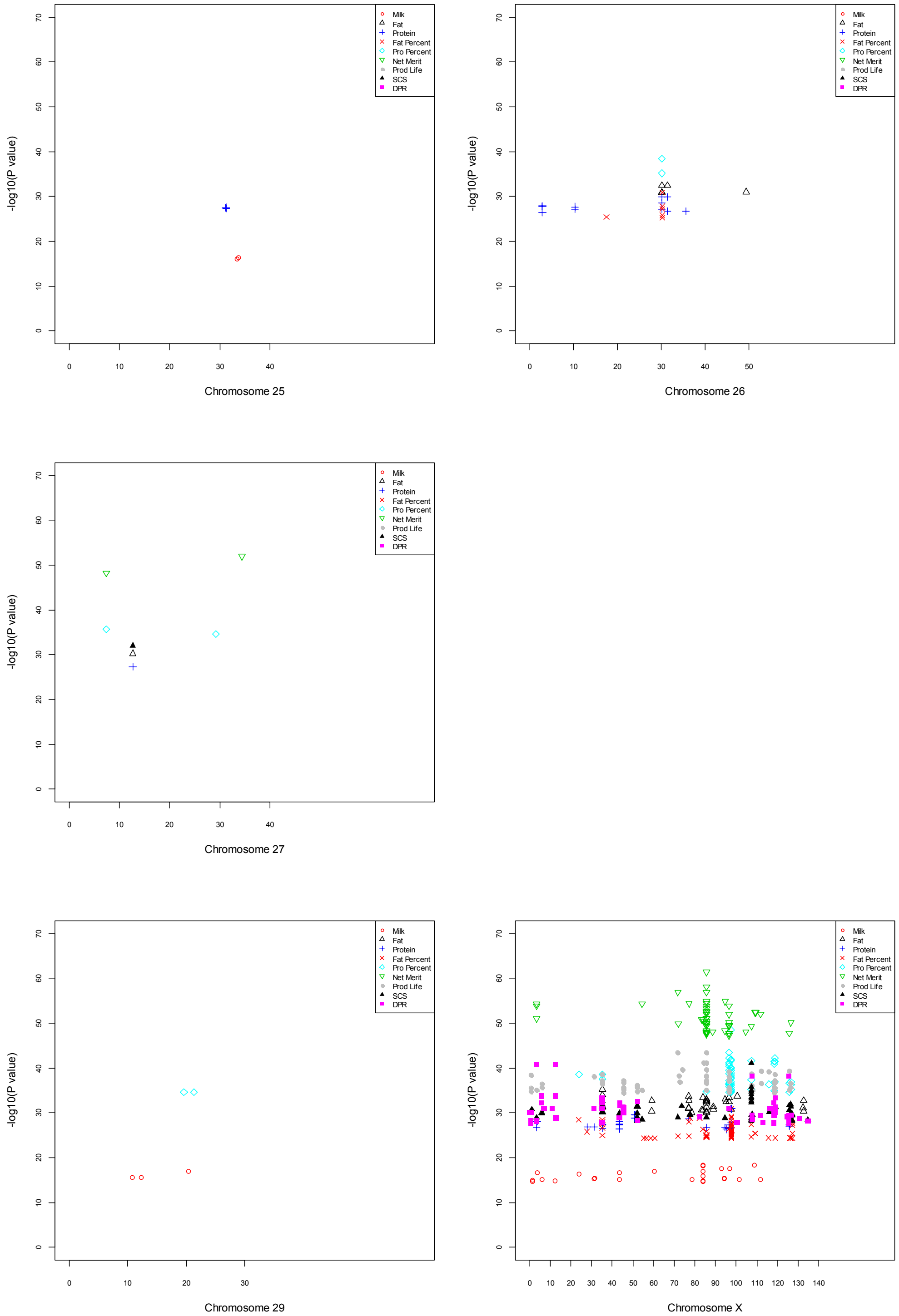


Figure 4.1 Epistasis QTL map of net merit and its eight component traits detected in comtemporary Holstein cows.

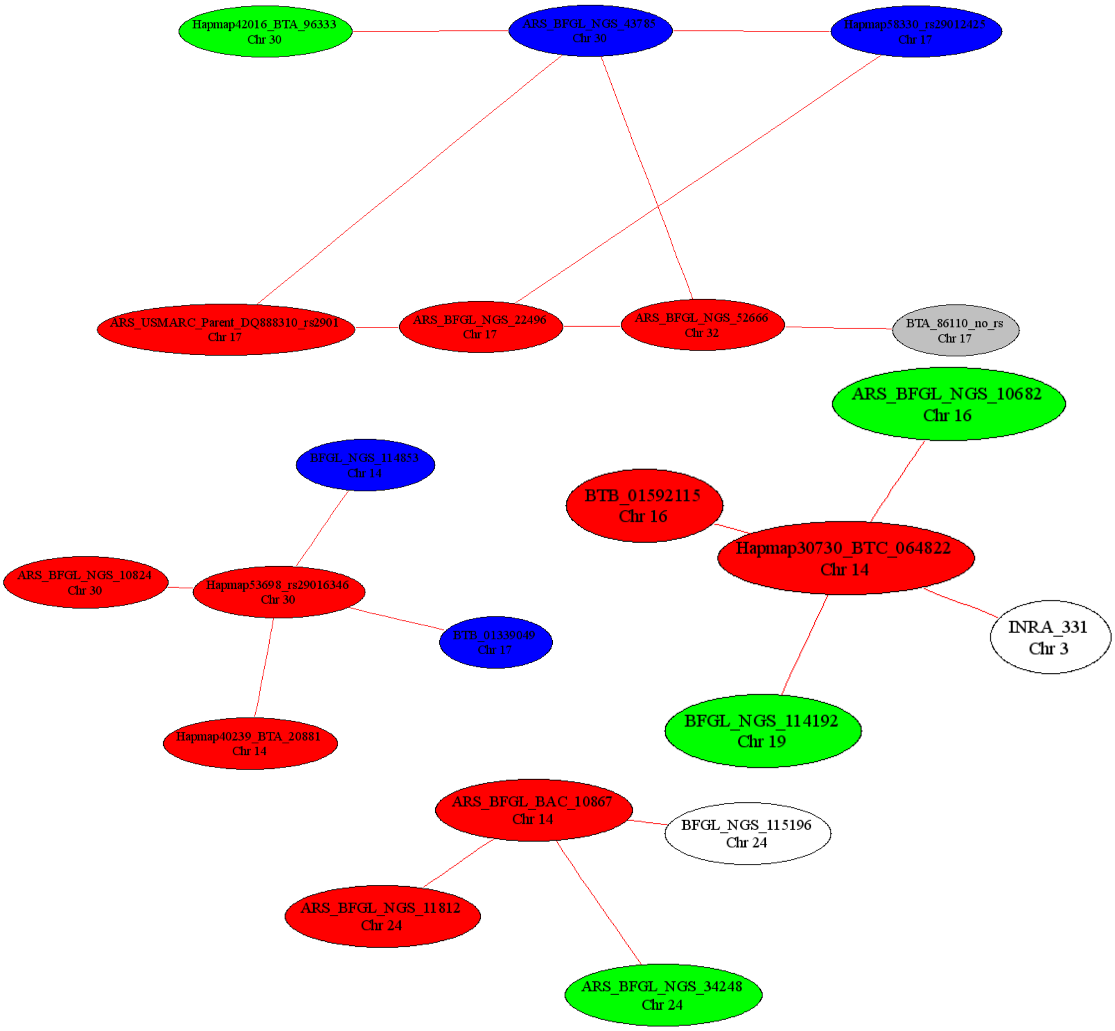


Figure 4.2 significant SNP epistasis network of Milk

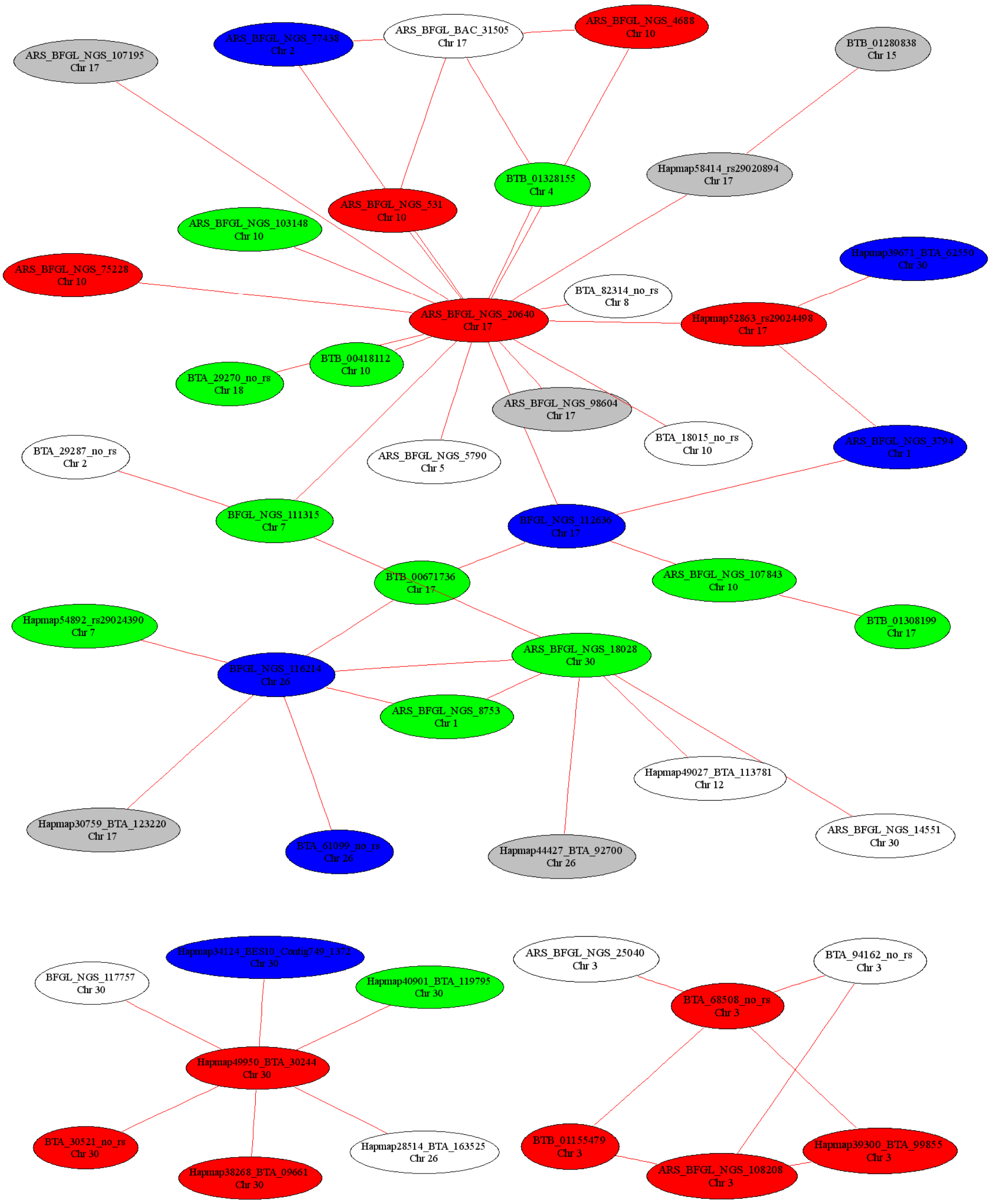

ARS_BFGL_NGS_107195 Chr 17
ARS_BFGL_NGS_77438 Chr 2
ARS_BFGL_BAC_31505 Chr 17
ARS_BFGL_NGS_4688 Chr 10
BTB_01280838 Chr 15
BTB_01328155 Chr 4
Hapmap58414_rs29020894 Chr 17
ARS_BFGL_NGS_531 Chr 10
ARS_BFGL_NGS_103148 Chr 10
Hapmap39671_BTA_62550 Chr 30
ARS_BFGL_NGS_75228 Chr 10
BTA_82314_no_rs Chr 8
ARS_BFGL_NGS_20640 Chr 17
Hapmap52863_rs29024498 Chr 17
BTB_00418112 Chr 10
BTA_29270_no_rs Chr 18
ARS_BFGL_NGS_98604 Chr 17
BTA_18015_no_rs Chr 10
ARS_BFGL_NGS_3794 Chr 1
BTA_29287_no_rs Chr 2
ARS_BFGL_NGS_5790 Chr 5
BFGL_NGS_111315 Chr 7
BFGL_NGS_112636 Chr 17
BTB_00671736 Chr 17
ARS_BFGL_NGS_107843 Chr 10
Hapmap54892_rs29024390 Chr 7
BTB_01308199 Chr 17
ARS_BFGL_NGS_18028 Chr 30
BFGL_NGS_116214 Chr 26
ARS_BFGL_NGS_8753 Chr 1
Hapmap49027_BTA_113781 Chr 12
ARS_BFGL_NGS_14551 Chr 30
Hapmap30759_BTA_123220 Chr 17
BTA_61099_no_rs Chr 26
Hapmap44427_BTA_92700 Chr 26
Hapmap34124_BES10_Contig749_1372 Chr 30
BFGL_NGS_117757 Chr 30
Hapmap40901_BTA_119795 Chr 30
ARS_BFGL_NGS_25040 Chr 3
BTA_94162_no_rs Chr 3
BTA_68508_no_rs Chr 3
Hapmap49950_BTA_30244 Chr 30
BTA_30521_no_rs Chr 30
Hapmap38268_BTA_09661 Chr 30
Hapmap28514_BTA_163525 Chr 26
BTB_01155479 Chr 3
ARS_BFGL_NGS_108208 Chr 3
Hapmap39300_BTA_99855 Chr 3

Figure 4.3 significant SNP epistasis network of fat yield.

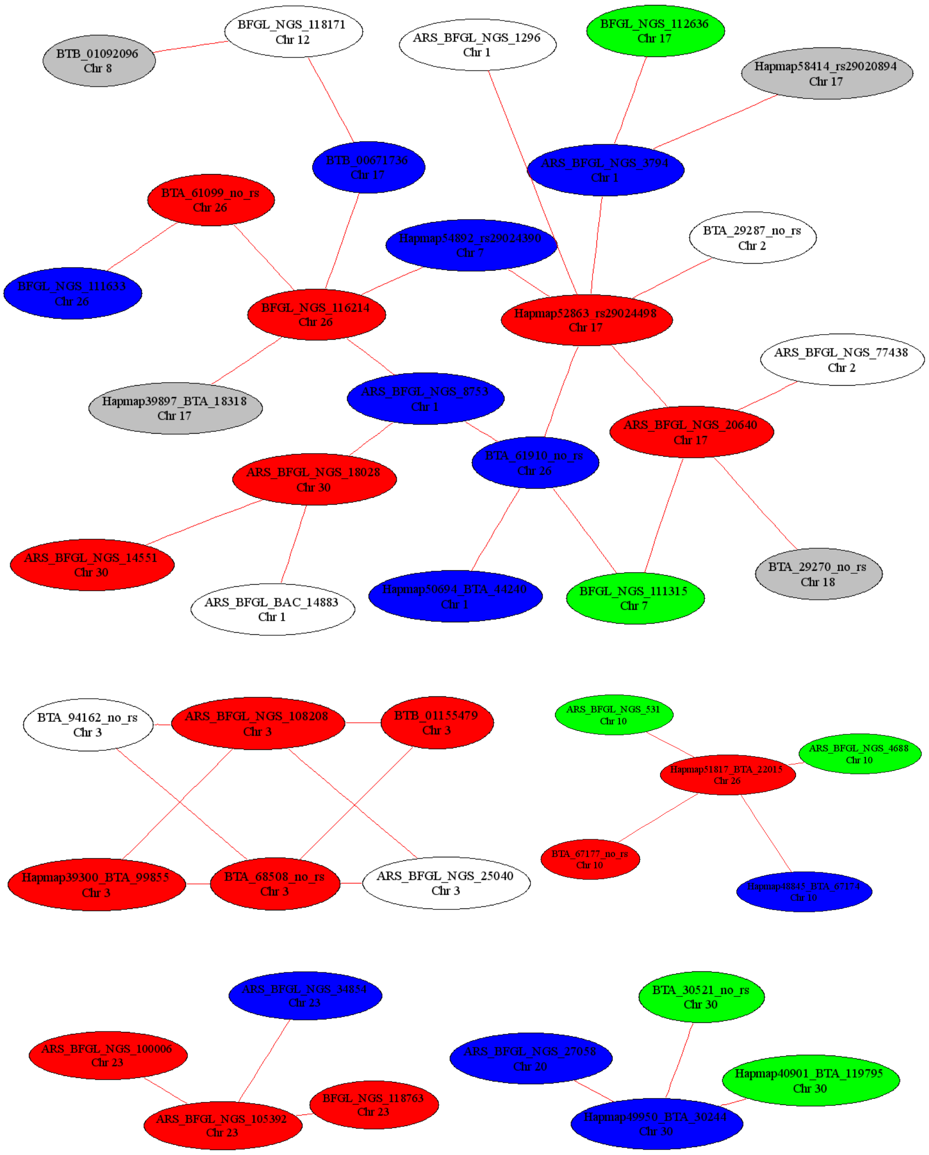


Figure 4.4 significant SNP epistasis network of protein yield.

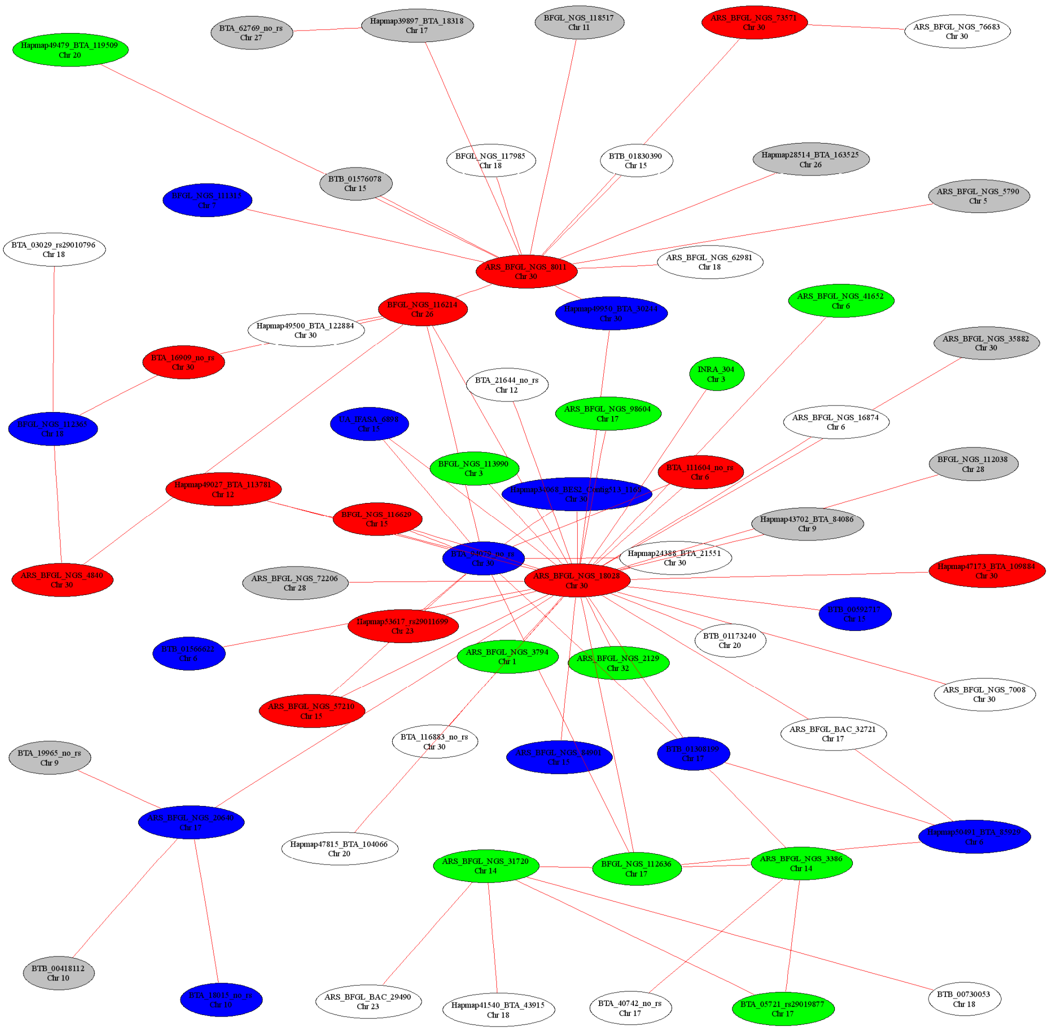


Figure 4.5 significant SNP epistasis network of fat percent.

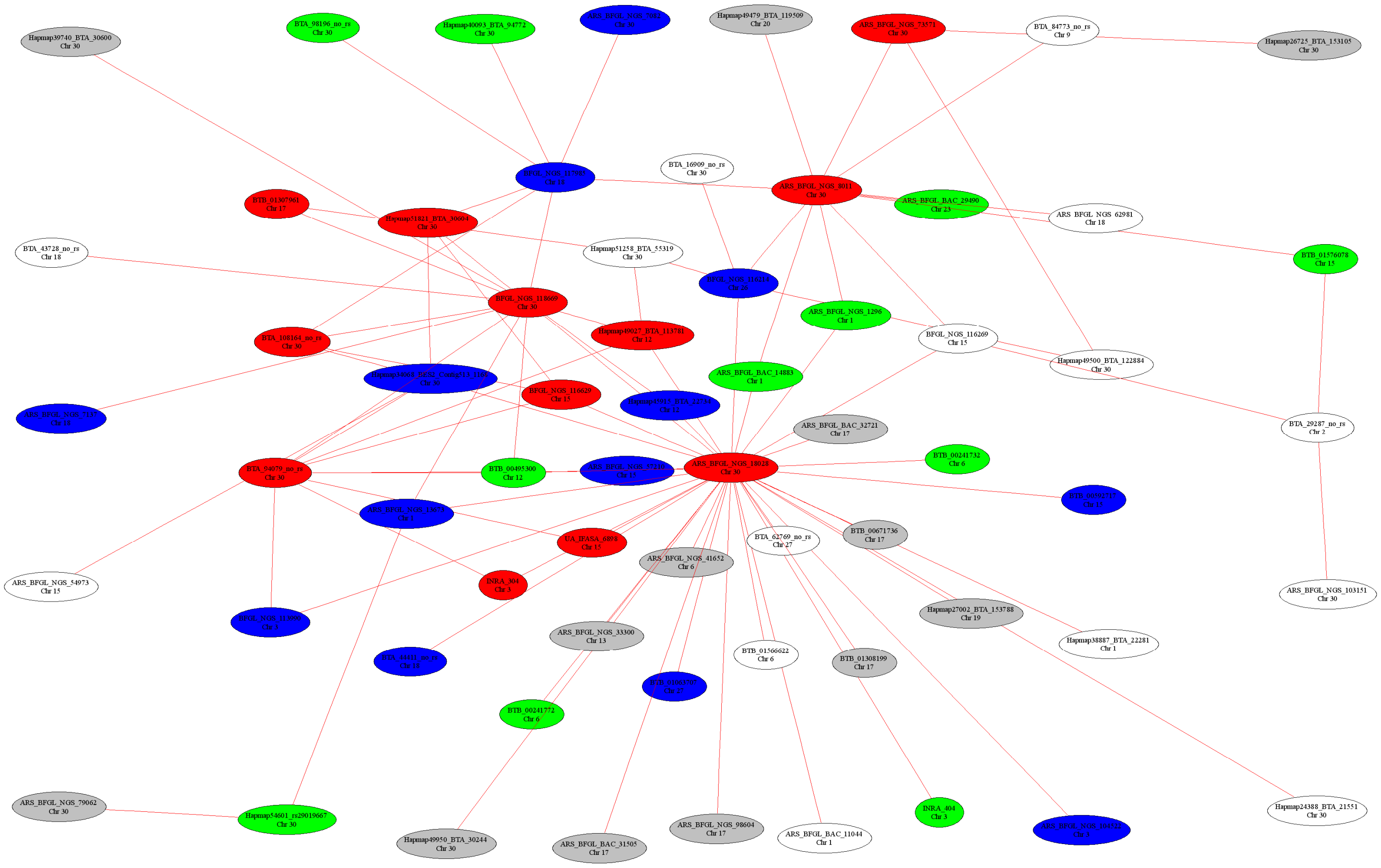

ARS_BFGL_NGS_7082 Chr 30
Hapmap49479_BTA_119509 Chr 20
ARS_BFGL_NGS_73571 Chr 30
BTA_98196_no_rs Chr 30
Hapmap40093_BTA_94772 Chr 30
BTA_84773_no_rs Chr 9
Hapmap39740_BTA_30600 Chr 30
Hapmap26725_BTA_153105 Chr 30
BTA_16909_no_rs Chr 30
BFGL_NGS_117985 Chr 18
ARS_BFGL_NGS_8011 Chr 30
ARS_BFGL_BAC_29490 Chr 23
BTB_01307961 Chr 17
Hapmap51821_BTA_30604 Chr 30
ARS_BFGL_NGS_62981 Chr 18
BTA_43728_no_rs Chr 18
Hapmap51258_BTA_55319 Chr 30
BTB_01576078 Chr 15
BFGL_NGS_116214 Chr 26
BFGL_NGS_118669 Chr 30
ARS_BFGL_NGS_1296 Chr 1
Hapmap49027_BTA_113781 Chr 12
BTA_108164_no_rs Chr 30
BFGL_NGS_116269 Chr 15
Hapmap49500_BTA_122884 Chr 30
Hapmap34068_BES2_Contig513_1168 Chr 30
ARS_BFGL_BAC_14883 Chr 1
BFGL_NGS_116629 Chr 15
Hapmap45915_BTA_22734 Chr 12
ARS_BFGL_NGS_7137 Chr 18
ARS_BFGL_BAC_32721 Chr 17
BTA_29287_no_rs Chr 2
BTB_00241732 Chr 6
BTA_94079_no_rs Chr 30
BTB_00495300 Chr 12
ARS_BFGL_NGS_57210 Chr 15
ARS_BFGL_NGS_18028 Chr 30
BTB_00592717 Chr 15
ARS_BFGL_NGS_13673 Chr 1
BTB_00671736 Chr 17
BTA_62769_no_rs Chr 27
UA_IFASA_6898 Chr 15
ARS_BFGL_NGS_41652 Chr 6
ARS_BFGL_NGS_54973 Chr 15
INRA_304 Chr 3
ARS_BFGL_NGS_103151 Chr 30
Hapmap27002_BTA_153788 Chr 19
BFGL_NGS_113990 Chr 3
Hapmap38887_BTA_22281 Chr 1
ARS_BFGL_NGS_33300 Chr 13
BTB_01566622 Chr 6
BTA_44411_no_rs Chr 18
BTB_01308199 Chr 17
BTB_01063707 Chr 27
BTB_00241772 Chr 6
ARS_BFGL_NGS_79062 Chr 30
INRA_404 Chr 3
Hapmap24388_BTA_21551 Chr 30
Hapmap54601_rs29019667 Chr 30
ARS_BFGL_NGS_98604 Chr 17
ARS_BFGL_NGS_104522 Chr 3
Hapmap49950_BTA_30244 Chr 30
ARS_BFGL_BAC_11044 Chr 1
ARS_BFGL_BAC_31505 Chr 17

Figure 4.6 significant SNP epistasis network of protein percent.

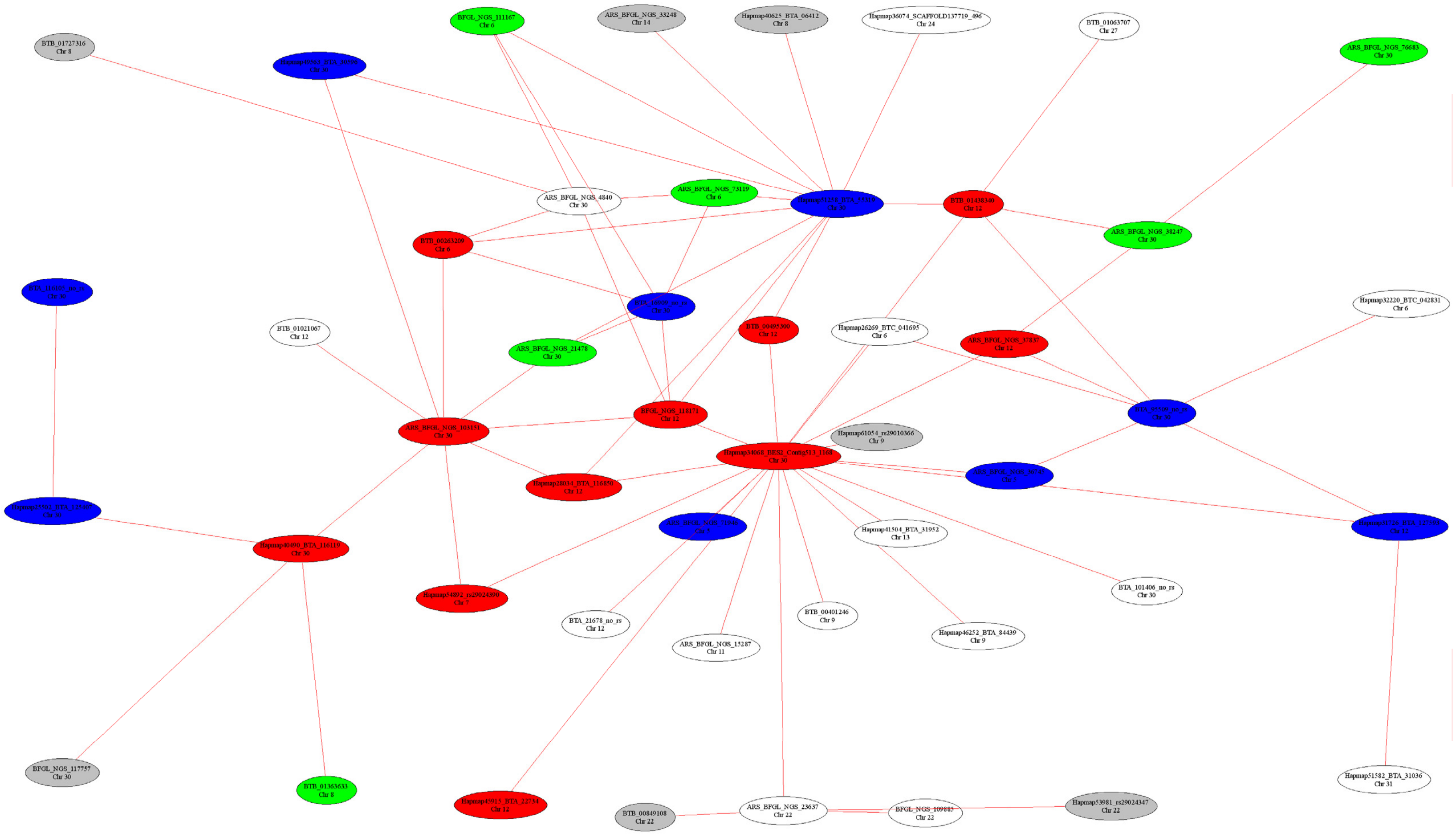

BFGL_NGS_111167 Chr 6
ARS_BFGL_NGS_33248 Chr 14
Hapmap40625_BTA_06412 Chr 8
Hapmap36074_SCAFFOLD137719_496 Chr 24
BTB_01063707 Chr 27
BTB_01727316 Chr 8
ARS_BFGL_NGS_76683 Chr 30
Hapmap49563_BTA_30596 Chr 30
ARS_BFGL_NGS_73119 Chr 6
ARS_BFGL_NGS_4840 Chr 30
Hapmap51258_BTA_55319 Chr 30
BTB_01438340 Chr 12
ARS_BFGL_NGS_38247 Chr 30
BTB_00263209 Chr 6
BTA_116105_no_rs Chr 30
BTA_16909_no_rs Chr 30
Hapmap32220_BTC_042831 Chr 6
BTB_01021067 Chr 12
BTB_00495300 Chr 12
Hapmap26269_BTC_041695 Chr 6
ARS_BFGL_NGS_21478 Chr 30
ARS_BFGL_NGS_37837 Chr 12
BTA_95509_no_rs Chr 30
BFGL_NGS_118171 Chr 12
ARS_BFGL_NGS_103151 Chr 30
Hapmap61054_rs29010366 Chr 9
Hapmap34068_BES2_Contig513_1168 Chr 30
Hapmap28034_BTA_116850 Chr 12
ARS_BFGL_NGS_36745 Chr 5
Hapmap25502_BTA_125407 Chr 30
ARS_BFGL_NGS_71946 Chr 5
Hapmap41504_BTA_31952 Chr 13
Hapmap31726_BTA_127593 Chr 12
Hapmap40490_BTA_116119 Chr 30
Hapmap54892_rs29024390 Chr 7
BTA_101406_no_rs Chr 30
BTA_21678_no_rs Chr 12
BTB_00401246 Chr 9
Hapmap46252_BTA_84439 Chr 9
ARS_BFGL_NGS_15287 Chr 11
BFGL_NGS_117757 Chr 30
Hapmap51582_BTA_31036 Chr 31
BTB_01363633 Chr 8
Hapmap45915_BTA_22734 Chr 12
BTB_00849108 Chr 22
ARS_BFGL_NGS_23637 Chr 22
BFGL_NGS_109885 Chr 22
Hapmap53981_rs29024347 Chr 22

Figure 4.7 significant SNP epistasis network of somatic cell score.

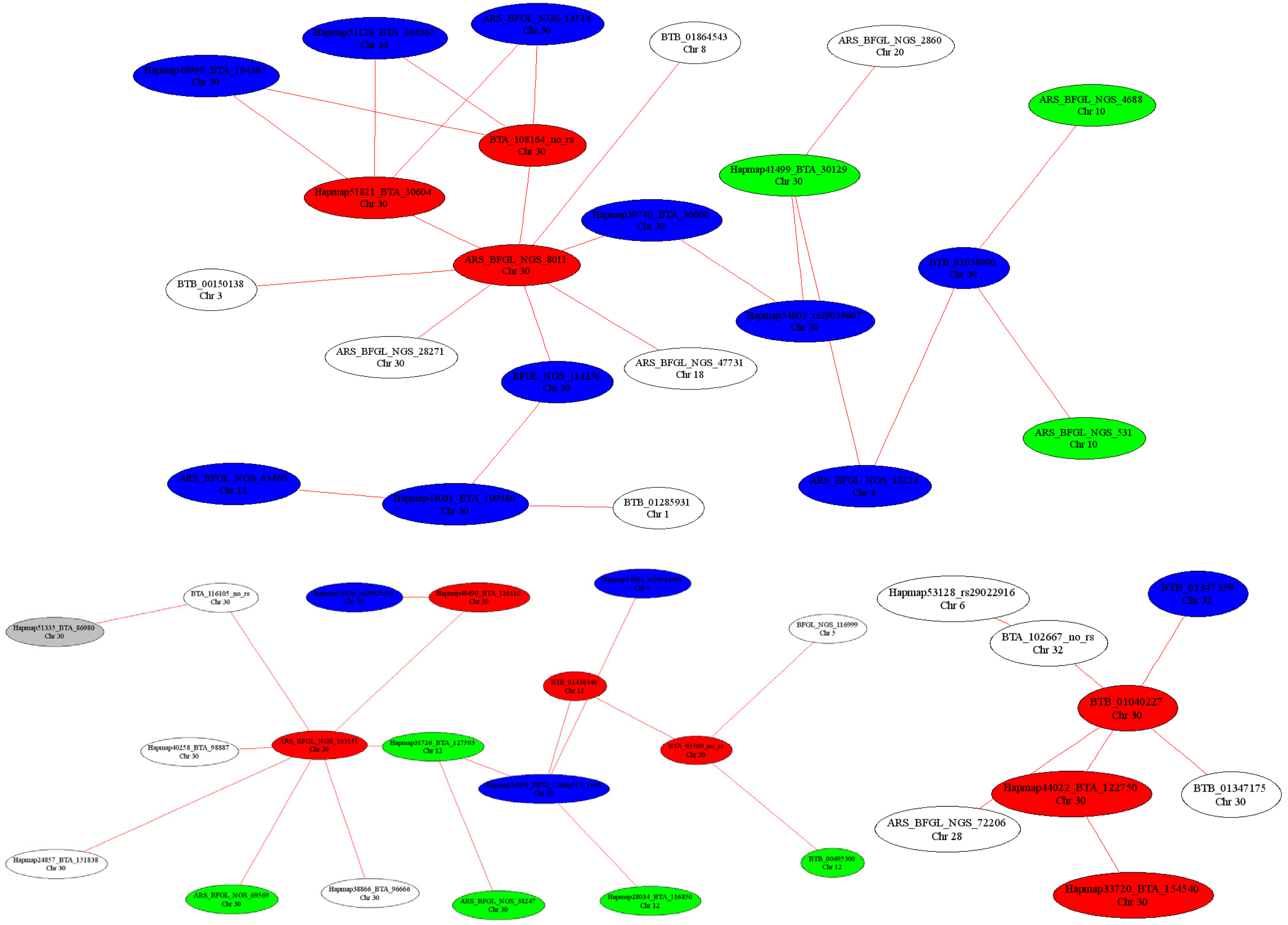


Figure 4.8 significant SNP epistasis network of daughter pregnancy rate.

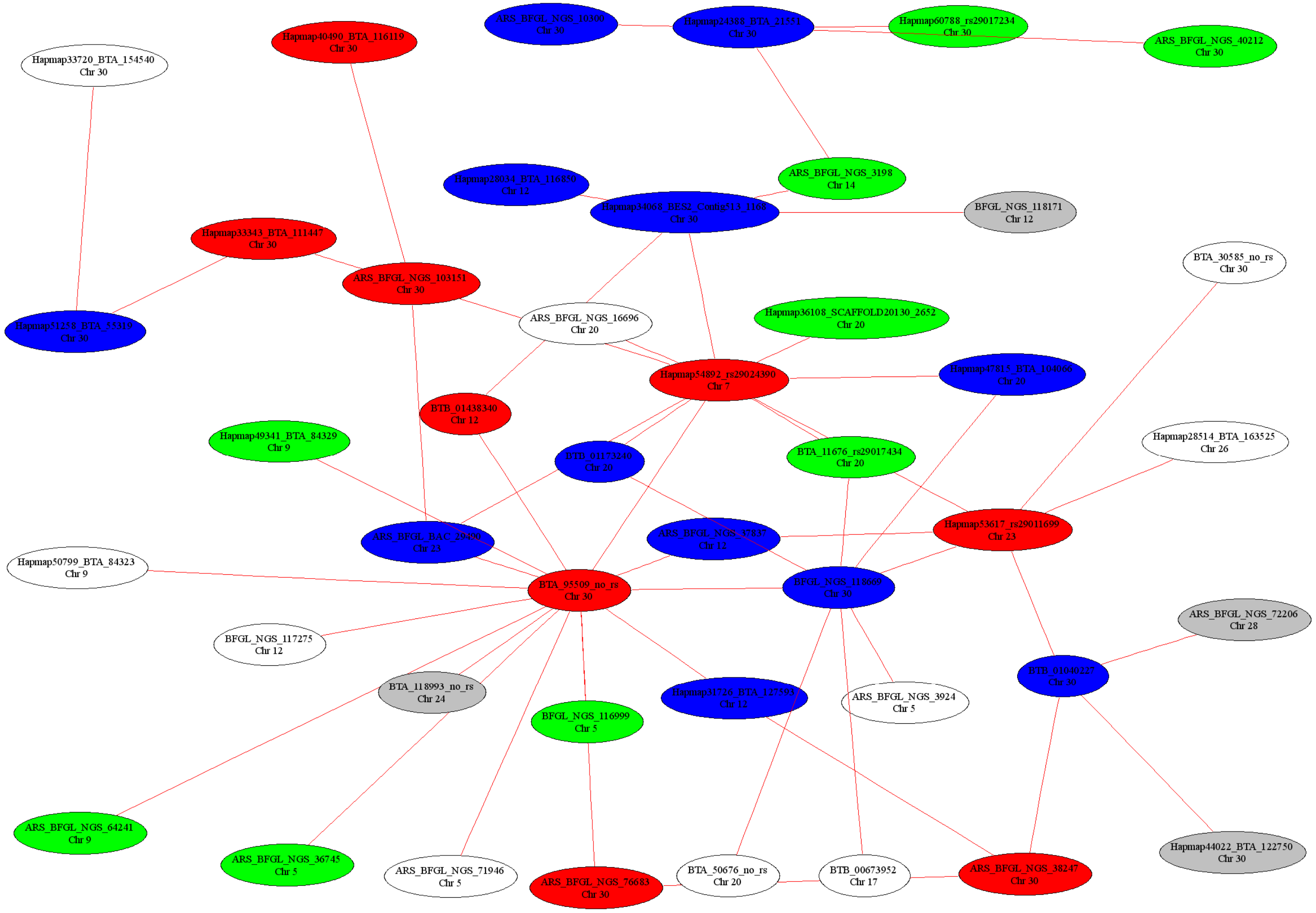


Figure 4.9 significant SNP epistasis network of productive life.

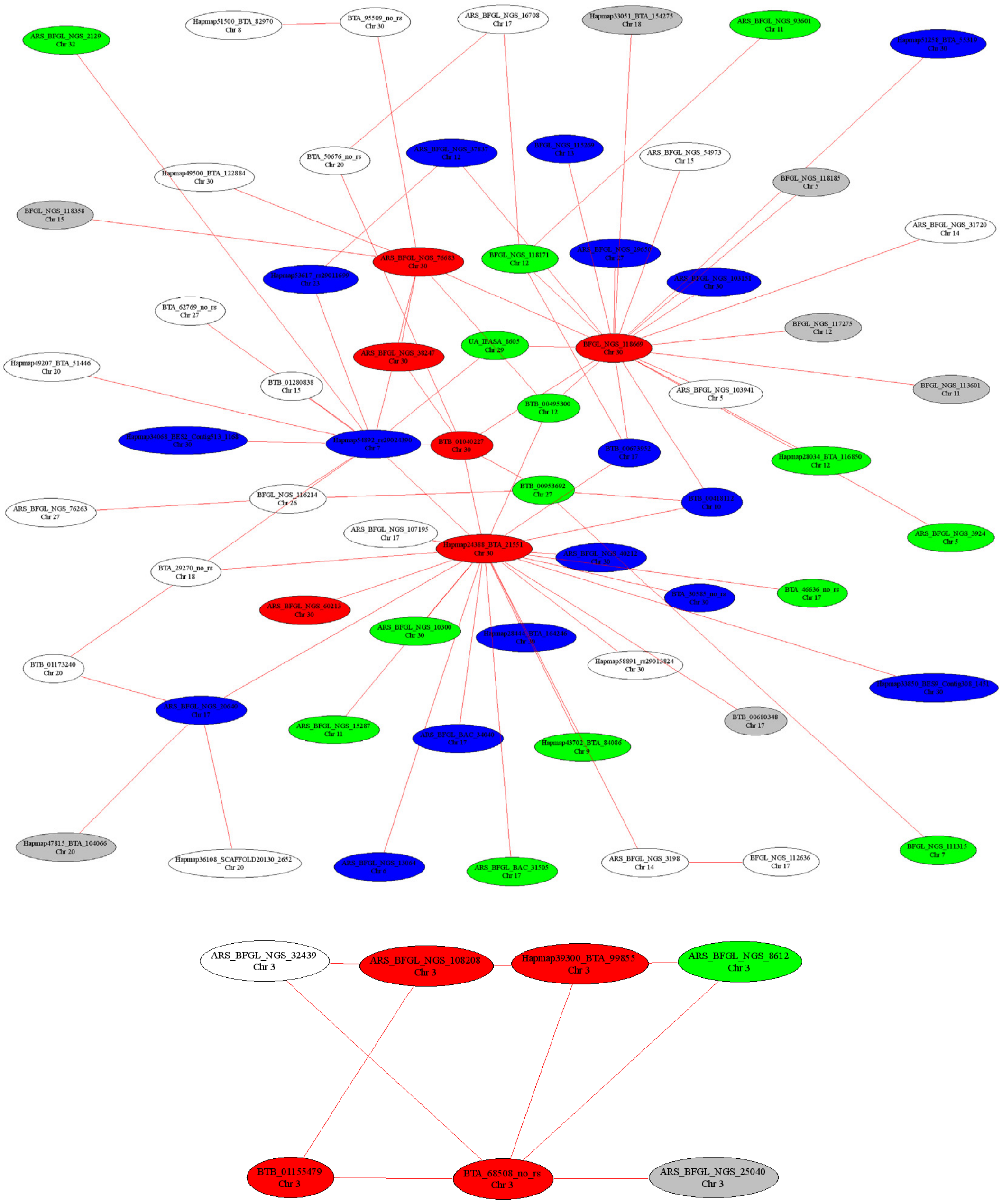

Hapmap51500_BTA_82970 Chr 8
BTA_95509_no_rs Chr 30
ARS_BFGL_NGS_16708 Chr 17
Hapmap33051_BTA_154275 Chr 18
ARS_BFGL_NGS_93601 Chr 11
ARS_BFGL_NGS_2129 Chr 32
Hapmap51258_BTA_55319 Chr 30
ARS_BFGL_NGS_37837 Chr 12
BFGL_NGS_115269 Chr 13
BTA_50676_no_rs Chr 20
ARS_BFGL_NGS_54973 Chr 15
Hapmap49500_BTA_122884 Chr 30
BFGL_NGS_118185 Chr 5
BFGL_NGS_118358 Chr 15
ARS_BFGL_NGS_31720 Chr 14
ARS_BFGL_NGS_29650 Chr 27
BFGL_NGS_118171 Chr 12
ARS_BFGL_NGS_76683 Chr 30
Hapmap53617_rs29011699 Chr 23
ARS_BFGL_NGS_103151 Chr 30
BTA_62769_no_rs Chr 27
BFGL_NGS_117275 Chr 12
UA_IFASA_8605 Chr 29
BFGL_NGS_118669 Chr 30
ARS_BFGL_NGS_38247 Chr 30
Hapmap49207_BTA_51446 Chr 20
BTB_01280838 Chr 15
BFGL_NGS_113601 Chr 11
ARS_BFGL_NGS_103941 Chr 5
BTB_00495300 Chr 12
Hapmap34068_BES2_Contig513_1168 Chr 30
Hapmap54892_rs29024390 Chr 7
BTB_01040227 Chr 30
BTB_00673952 Chr 17
Hapmap28034_BTA_116850 Chr 12
BTB_00953692 Chr 27
BFGL_NGS_116214 Chr 26
BTB_00418112 Chr 10
ARS_BFGL_NGS_76263 Chr 27
ARS_BFGL_NGS_107195 Chr 17
ARS_BFGL_NGS_3924 Chr 5
Hapmap24388_BTA_21551 Chr 30
ARS_BFGL_NGS_40212 Chr 30
BTA_29270_no_rs Chr 18
BTA_30585_no_rs Chr 30
BTA_46636_no_rs Chr 17
ARS_BFGL_NGS_60213 Chr 30
ARS_BFGL_NGS_10300 Chr 30
Hapmap28444_BTA_164246 Chr 30
Hapmap58891_rs29013824 Chr 30
BTB_01173240 Chr 20
Hapmap33850_BES9_Contig308_1451 Chr 30
ARS_BFGL_NGS_20640 Chr 17
BTB_00680348 Chr 17
ARS_BFGL_NGS_15287 Chr 11
ARS_BFGL_BAC_34040 Chr 17
Hapmap43702_BTA_84086 Chr 9
Hapmap47815_BTA_104066 Chr 20
Hapmap36108_SCAFFOLD20130_2652 Chr 20
ARS_BFGL_NGS_13064 Chr 6
ARS_BFGL_BAC_31505 Chr 17
ARS_BFGL_NGS_3198 Chr 14
BFGL_NGS_112636 Chr 17
BFGL_NGS_111315 Chr 7
ARS_BFGL_NGS_32439 Chr 3
ARS_BFGL_NGS_108208 Chr 3
Hapmap39300_BTA_99855 Chr 3
ARS_BFGL_NGS_8612 Chr 3
BTB_01155479 Chr 3
BTA_68508_no_rs Chr 3
ARS_BFGL_NGS_25040 Chr 3

Figure 4.10 significant SNP epistasis network of net merit.

Chapter 5

Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham Heart Study Data

5.1 Introduction

5.11 Background

Cholesterol concentrations in blood are related to cardiovascular diseases. Recent genome-wide association studies (GWAS) of cholesterol levels identified a number of single-locus effects on total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C) levels. Here, we report single-locus and epistasis SNP effects on TC and HDL-C using the Framingham Heart Study (FHS) data.

5.12 Results

Single-locus effects and pairwise epistasis effects of 432,096 SNP markers were tested for their significance on log-transformed TC and HDL-C levels. Twenty nine additive SNP effects reached single-locus genome-wide significance ($p < 7.2\times10^{-8}$) and no dominance effect reached genome-wide significance. Two new gene regions were detected, the *RAB3GAP1-R3HDM1-LCT-MCM6* region of chr02 for TC identified by six new SNPs, and the *OSBPL8-ZDHHC17* (chr12) region for HDL-C identified by one new SNP. The remaining 22 single-locus SNP effects confirmed previously reported genes or gene regions. For TC, three SNPs identified two gene regions that were tightly linked with previously reported genes associated with TC, including rs599839 that was 10 bases downstream of *PSRC1* and 3.498kb downstream of *CELSR2*, rs4970834 in *CELSR2*, and rs4245791 in *ABCG8* that slightly overlapped with *ABCG5*. For HDL-C, *LPL* was confirmed by 12 SNPs 8~45kb downstream, *CETP* by two SNPs 0.5~11kb upstream, and the *LIPG-ACAA2* region by five SNPs inside this region. Two epistasis effects on TC and thirteen epistasis effects on HDL-C reached the significance of "suggestive linkage". The most significant epistasis effect ($p = 5.72\times10^{-13}$) was close to reaching "significant linkage" and was a dominance × dominance effect of HDL-C between *LMBRD1* (chr06) and the *LRIG3* region (chr12), and this pair of gene regions had six other D×D effects with "suggestive linkage".

5.13 Conclusions

Genome-wide association analysis of the FHS data detected two new gene regions with genome-wide significance, detected epistatic SNP effects on TC and HDL-C with the significance of suggestive linkage in seven pairs of gene regions, and confirmed some previously reported gene regions associated with TC and HDL-C.

5.2 Background

Total cholesterol (TC) is related to coronary diseases and high-density lipoprotein (HDL-C) cholesterol is antiatherogenic. Genome-wide association studies (GWAS) and human genetic studies have identified a number of genes and gene regions affecting cholesterol phenotypes including TC and HDL-C (Aulchenko, 2008; Kathiresan, 2008a; Kathiresan, 2008b; Willer, 2008; Karvanen, 2009; Sandhu, 2008; Wallace, 2008; Samani, 2008; Nakayama, 2009; Schadt, 2008; Kathiresan, 2007). A meta-analysis of HDL-C levels that include the FHS data has previously been published (Kathiresan, 2008a). An early report on FHS (Murray, 2009) analyzed TC and HDL-C but used 100k SNPs and a sample size that was much smaller than the current FHS sample size. Epistasis analysis of TC and HDL-C was unavailable. Here, we apply a quantitative genetics approach offering simultaneous testing of eight genetic effects for each pair of loci based on an extended Kempthorne model that allows Hardy-Weinberger disequilibrium and linkage disequilibrium (Mao, 2006) for GWAS analysis of the FHS data using 432,096 SNP markers and over 6000 individuals to detect additive or dominance single-locus effects and epistasis effects on log-transformed TC and HDL-C. The epistasis effects we tested included additive × additive (A×A), additive × dominance (A×D) or dominance × additive (D×A), and dominance × dominance (D×D) effects, with genetic interpretations of allele × allele, allele × genotype or genotype × allele, and genotype × genotype interactions. The single-locus analysis was intended to detect new targets or confirm existing targets using a method of analysis different from those used in previous reports while the epistasis analysis of TC and HDL-C was the first such attempt using the FHS data and the 500k SNP panel.

5.3 Results

The single-locus tests detected nine SNPs with additive (or allelic) effects on TC and twenty SNPs with additive effects on HDL-C that reached genome-wide significance (**Tables 5.1-5.2**). No dominance effect reached genome-wide significance. Among the twenty nine SNP effects, twenty were new effects that were not reported in previous studies and nine were previously reported to be associated with various cholesterol phenotypes (Aulchenko, 2008; Kathiresan, 2008a; Kathiresan, 2008b; Willer, 2008; Karvanen, 2009; Sandhu, 2008; Wallace, 2008; Samani, 2008; Nakayama, 2009; Schadt, 2008; Kathiresan, 2007; Murray, 2009).. Seven SNPs identified two new gene regions while the remaining twenty two SNPs confirmed previously reported gene regions. Two epistasis effects on TC and thirteen epistasis effects on HDL-C representing seven pairs of gene regions reached the significance of "suggestive linkage".

5.31 Single-locus effects

For TC, nine SNPs with additive (or allelic) effects reached genome-wide significance with $p < 7.2\times10^{-8}$ (**Table 5.1**). Six SNPs inside or near four genes identified a new chr02 region containing *RAB3GAP1*, *R3HDM1*, *LCT* and *MCM6* to be associated with TC (**Figure 5.1a**). Of the six SNPs in the *RAB3GAP1-R3HDM1-LCT-MCM6* region, five SNPs were inside genes and one SNP was 4.2kb upstream *MCM6*. The most significant SNP in this region was rs2322660 in intron 12 of *LCT* (**Table 5.1**). The *RAB3GAP1-R3HDM1-LCT-MCM6* region contained two other genes (*ZRANB3* and *UBXD2*) that did not have significant SNPs. Eleven other SNPs spanning a 1.23 Mb region (**Figure 5.1a**) that includes *RAB3GAP1-R3HDM1-LCT-MCM6* had p-values between 1.27×10^{-5} and 7.13×10^{-7}, including one SNP upstream *ACMSD*, one SNP inside *ACMSD*, two SNPs inside *YSK4*, one SNP inside *R3HDM1*, one SNP inside *UBXD2* (also named *UBXN4* according to NCBI [13]), two SNPs inside *LCT*, and three SNPs downstream *DARS* (data not shown). These less significant results in the same neighborhood should add to the significance of the *RAB3GAP1-R3HDM1-LCT-MCM6* region to TC. Three SNPs identified two genes that were tightly linked with previously reported genes associated with TC (Aulchenko, 2008). These three SNPs were rs599839 that was 10 bases downstream *PSRC1* (chr01) and 3.498kb downstream *CELSR2*, rs4970834 in intron 28 of *CELSR2*, and rs4245791 in intron 3 of *ABCG8* (chr02) that slightly overlapped with *ABCG5*, where *CELSR2* and *ABCG5* regions were reported to be associated with TC in a recent GWAS report (Aulchenko, 2008). *PSRC1* and *ABCG8* also were reported to affect low-density lipoprotein cholesterol (LDL-C) (Kathiresan, 2008a; Kathiresan, 2008b; Willer, 2008; Sandhu, 2008; Wallace, 2008; Nakayama, 2009; Schadt, 2008). The SNP (rs599839) that was 10 bases downstream *PSRC1* had the most significant single-locus effect on TC (p = 3.7×10^{-16}), while the SNP inside *CELSR2* (rs4970834) had the second most significant single-locus effect on TC (p = 1.29×10^{-10}).

For HDL-C, twenty SNPs with additive effects reached genome-wide significance (**Table 5.2**). SNP rs17259942 identified a new gene region associated with HDL-C, the *OSBPL8- ZDHHC17* region (q21.2, **Figure 5.1b**), with rs17259942 being 117kb downstream *OSBPL8* and 85kb upstream *ZDHHC17* from ENSEMBL. According to NCBI, the *OSBPL8-ZDHHC17* region contained three pseudo-genes (*RPL7AP59*, *RPL21P98* and *RPL7P43*) and rs17259942 was 18kb downstream *RPL21P98* and 43kb upstream *RPL7P43*. The other nineteen SNPs confirmed previously reported gene regions, including twelve SNPs 8-45kb downstream *LPL*, two SNPs 0.5~11kb upstream *CETP*, and five SNPs in the *LIPG-ACAA2* region (39.812kb downstream *LIG* and 65.963kb upstream *ACAA2*) (Aulchenko, 2008; Kathiresan, 2008a; Kathiresan, 2008b; Willer, 2008; Wallace, 2008; Kathiresan, 2007; Murray, 2009. *LPL*, *CETP* and *LIPG* were reported to be associated with HDL-C in four recent GWAS reports (Aulchenko, 2008; Kathiresan, 2008a; Kathiresan, 2008b; Willer, 2008) while *ACAA2* was reported in (Kathiresan, 2008b). The SNP nearest to *CETP* (rs1800775) was the most significant effect ($p = 8.61\times10^{-34}$) in this study.

QQ plot for single SNP tests on TC and HDL-C showed that p-values of significant results all deviated from the expected p-values under the null hypothesis (**Figure 5.2**).

5.32 Epistasis effects

Two epistasis effects on TC and thirteen epistasis effects on HDL-C reached the significance of suggestive linkage defined in (Lander, 1995) (**Table 5.3**). The two epistasis effects on TC involved two different pairs of gene regions while the thirteen epistasis effects on HDL-C involved five different pairs of gene regions, so that the fifteen epistasis effects identified seven pairs of gene regions. Eight SNPs in introns 1, 5, 7, 9, and 14 of *LMBRD1* (chr06) (**Figure 5.1c**) interacted with a chr12 SNP about 53kb from LRIG3 (q14.1, **Figure 5.1d**) and all these eight pairs had D×D effects on HDL-C. One of the eight epistasis effects involving intron 14 of *LMBRD1* was the most significant epistasis effect that was close to reaching “significant linkage” defined in (Lander, 1995) or genome-wide significance with 5% Bonferroni corrected type-I error.

Among the seven different pairs of gene regions with epistasis effects, four pairs had A×A effects, one pair had A×D effect, and two pairs had D×D effects (**Table 5.3**). For the A×A effect on TC involving chr04 and chr10, the A-T gamete had the highest TC value while the A-G gamete had the lowest TC value (**Table 5.4**). This showed that the G and T alleles of rs705169 on chr10 had significantly different effects when combined with the A allele of rs4437278 on chr04, noting that rs705169 did not have significant single-locus effect. The same phenomenon was also observed for the other three A×A effects in **Table 5.4**. For the A×D effect, the A-GG allele-genotype combination had the highest TC value while the G-GG allele-genotype combination had the lowest TC value. The two D×D effects were on HDL-C. For the D×D effect of rs4706271 × rs6581219 representing the eight pairs of D×D effects involving the same gene regions, GT-GG had the highest HDL-C value while GG-GG had the lowest HDL-C value. For the remaining D×D effect of rs12596869 × rs6506699 representing the two D×D effects of the same gene regions, CC-AG had the highest HDL-C value while CC-AA had the lowest HDL-C value (**Table 5.4**).

5.4 Discussion

The single-locus results in this study had strong confirmations with existing studies. For TC, we confirmed *CELSR2* and *ABCG5* reported in (Aulchenko, 2008; Samani, 2008).. These confirmed TC results should be considered as strong confirmation because our study had no overlapping samples with studies of (Aulchenko, 2008; Samani, 2008). We detected seven effects on TC in the *RAB3GAP1-R3HDM1-LCT-MCM6* region with the SNP in *LCT* being the most significant. Six of these seven effects had p-values for LDL-C in the range of 0.007-0.056 from a meta-analysis, (**Table 5.1**). This could be an indication about the significance on TC from a meta-analysis because LDL-C is calculated from TC (Friedewald, 1972). A study in FINRISK cohorts with 14,140 individuals reported that *LCT* was associated with both TC and LDL-C with p-values in the range of 0.0005-0.005 (Silander, 2008). In *Silico* replication using 1231 Italian subjects from the InCHIANTI cohort (Tanaka, 2009) generally lacked confirmation for the TC results in **Table 5.1**. The first three markers had p-values in the range of 0.005-0.07 while the other effects had p-values greater than 0.14 from the InCHIANTI cohort. The biological function of *LCT* for digesting lactose could be a reason for agreements and disagreements in replicating *LCT* effects on cholesterol. *LCT* affects lactose digestion and long-term consumption of lactose in rats was found to affect aortic cholesterol levels (Wostmann, 1980). Therefore, dietary lactose levels that have not been considered by human GWAS could have affected the *LCT* results of different studies. *MCM6* contains two of the regulatory regions for *LCT* (Enattah, 2002) so that the significant effects in or near *MCM6* (**Table 5.1**) could be due to *MCM6*'s regulatory role to *LCT*. HDL-C had twenty significant SNP effects, but only one SNP identified a new gene region (*OSBPL8-ZDHHC17*) while all the other SNPs confirmed previously reported gene regions, although only seven of the twenty significant SNPs for HDL-C were reported previously

(**Table 5.2**). *OSBPL8* encodes a group of intracellular lipid receptors and suppresses *ABCA1* (Yan, 2007), and *ABCA1* was found to affect HDL-C level (Wang, 2000; Cohen, 2004). For HDL-C, the InCHIANTI cohort did not confirm the effects in the *LIPG-ACAA* region (p > 0.55) but confirmed the other effects. In light of different samples and different methods of data analysis between our study and those in previous reports, the confirmations of gene results we observed for TC and HDL-C should be considered strong confirmations. This study used log-transformed TC and HDL-C values while recent GWAS on TC (Aulchenko, 2008) and HDL-C (Aulchenko, 2008; Kathiresan, 2008a, Kathiresan, 2008b; Willer, 2008) used the original observations of TC and HDL-C that deviated from normal distribution. However, single-locus effects from our study and previous studies (Aulchenko, 2008; Kathiresan, 2008a, Kathiresan, 2008b; Willer, 2008) had remarkable mutual confirmation, indicating that single-locus analysis was somewhat robust to data distribution and possibly to methods of analysis.

Epistasis effects on TC and HDL-C were not reported in other GWAS so that a comparison between our epistasis results and those from others was unavailable. We detected eight SNP pairs indicating the interaction between gene *LMBRD1* and gene *LRIG3* with the significance of suggestive linkage. Both *LMBRD1* and *LRIG3* encode membrane proteins. *LMBRD1* gene is involved in the transportation and metabolism of vitamin B12 which is important for metabolism of branched chain amino acids and odd chain fatty acids (Rutsch, 2009). Replication using the InCHIANTI cohort did not confirm the epistasis results (p > 0.15). The statistical power of epistasis testing is less than that for testing a single-locus effect, particularly for epistasis effects involving dominance such as A×D and D×D effects, with D×D effect being the most difficult to detect. The reason for this difficulty was due to the fact that higher-order effects explain less phenotypic variation even if the effect sizes were the same as lower-order effects (Mao, 2006). The reduced power for epistasis testing could have contributed to the fact that the epistasis effects we detected only reached 'suggestive linkage' although the sample size was over 6000. The data analysis of this study showed that pairwise analysis was sensitive to outliers. This was due to the fact that artificially significant epistasis effects could occur when rare combinations of loci had extreme genotypic values by chance. This may happen when outliers exist due to the large number of pairwise effects arising from the large number of pairwise combinations. For example, over 466 billion pairwise effects (93,353,260,560 pairs × 5 effects per pair = 466,766,302,800 pairwise effects) were tested per trait in this study. A small fraction of random association between rare frequencies and outliers in opposite directions among a large number of pairs could yield a long list of artificially significant epistasis results. Therefore, dealing with outliers such as removing outliers and using data transformation is important in pairwise analysis. Pairwise analysis is computationally intensive but timely analysis is possible using parallel computing. Using 784 processor cores on the SGI Altix XE 1300 Linux cluster system with 2.66 GHz Intel Clovertown processor at the Minnesota Supercomputer Institute, the completion of pairwise epistasis analysis required about 15 hours per trait.

5.5 Conclusions

Genome-wide association analysis of the FHS data detected new single-locus and epistasis effects on TC and HDL-C and confirmed some previously reported effects. Additive effects were the primary single-locus effects of TC and HDL-C while epistasis effects involved allele × allele, allele × genotype (or genotype × allele), and genotype × genotype interactions.

5.6 Methods

5.6.1 Phenotype and SNP data

The FHS GWAS data (version 2) had 6575 individuals with SNP genotypes of the 500k SNP panel from dbGAP. Of the 6575 individuals, 6431 had observations on TC and 6078 individuals had observations on HDL-C. A total of 496,858 SNP markers had known chromosome locations and 432,096 of these SNP markers with minor allele frequencies greater than or equal to 0.01 were analyzed.

5.6.2 Statistical Analysis

Original TC and HDL-C observations deviated from normality and had outliers (**Figure 5.3a, 5.3d**). The Box-Cox transformation analysis (Box, 1964) implemented by the R statistical package (2008) showed that the log-transformation was approximately the best transformation to achieve normality for those two traits (**Figure 5.3b-5.3c, 5.3e-5.3f**). One TC outlier, the highest TC value, was removed from the data analysis while no

HDL-C outlier was removed. Log-transformed TC values were adjusted for blood sugar, body mass index, smoking status, and sex that had significant effects on log(TC). Age, age-squared, cholesterol treatment, and alcohol consumption were also tested for significant effects on log(TC) but were not included in the phenotypic model because they were insignificant. Log(HDL-C) was adjusted for age, age-squared, cholesterol treatment, blood sugar, body mass index, smoking status, number of cigars smoked, alcohol consumption and sex. Age was insignificant for HDL-C but was included because age-squared was nearly significant ($p < 0.0543$). Single-locus and epistasis effects for both traits were tested using the extended Kempthorne model that allows Hardy-Weinberg disequilibrium and linkage disequilibrium (Mao, 2006). For each SNP, three effects were tested, genotypic, additive (A) and dominance (D) effects. For each SNP pair, five effects were tested, two-locus genotypic effect, A×A, A×D, D×A, and D×D epistasis effects. The EPISNPmpi parallel computing program [28] with a modification to implement a generalized least squares (GLS) analysis to account for sib correlations (Ma, 2008d) was used to implement the statistical tests of single-locus and pairwise epistasis effects. For single-locus tests, $p = 7.2\times10^{-8}$ was used as the threshold p-value to declare genome-wide significance (Dudbridge, 2008). To assess genome-wide significance of pairwise epistasis results, we used 5% type-I error with the Bonferroni correction as the genome-wide significance. The 500k SNP data was estimated to have 276,666 independent SNPs (Moskvina, 2008). Each pairwise test was considered to have four independent tests although five effects were tested, because the two-locus marker genotypic effect was confounded with one of the four epistasis effects in reporting significant results. Therefore, the genome-wide 5% type-I errors with the Bonferroni correction was calculated as $p = 0.05[4(276{,}666)(276{,}665)/2]^{-1} = 3.266\times10^{-13}$. This 5% significance level is equivalent to “significant linkage” defined in (Lander, 1995). Since the Bonferroni correction is generally considered too severe, we also reported epistasis effects reaching “suggestive linkage” with statistical evidence that would be expected to occur one time at random in a genome-wide analysis (Lander, 1995). In addition to the GLS method to account for sib correlation, the genomic control (GC) method (Devlin, 1999) was used to account for potential sub-population structures in the three generation cohort of the FHS data set. For single-locus tests, all p-values were used to estimate inflation parameters for TC and HDL-C, yielding inflation parameter estimates of 1.14 and 1.11 respectively, and test statistics from the GLS tests were then adjusted by the estimates of inflation parameters and p-values were recalculated using the GC adjusted test statistics, which resulted in fewer significant effects. For the pairwise epistasis testing, we randomly selected 50,000 p-values and test statistics from over 466 billion pairwise tests for computational efficiency. Then we estimated the inflation parameters using two samples of 50,000 data points each for TC and HDL-C, yielding inflation parameter estimates of 1.01 and 1.05 respectively. All p-values were then adjusted using the inflation parameters and such adjustments also resulted in fewer significant epistasis results. Frequency of each subclass in an epistasis effect was calculated and each subclass was required to have a minimal number of five observations. After GC adjustment, QQ plots were made to show deviations of the observed p-values from the expected p-values under the null hypothesis for significant test results for single-locus tests only. QQ plot for epistasis effects were not made because the number of p-values for epistasis tests was too large. Gene locations of significant SNPs were identified according to ENSEMBL and NCBI based on Build 37.0 of the human genome.

Table 5.1: Single-locus SNP effects for TC with genome control (GC) adjusted $P < 7.2\times10^{-8}$.

SNP	Chr	Position	Gene Region	Reported SNP effect	MAF	Effect Type & P value	
						Genotype	Additive
rs4970834	1	109814880	CELSR2[a] (intron 28)	Non-HDL-C [5-7]	0.18	1.10E-08	1.75E-09
rs599839	1	109822166	10 bases downstream PSRC1[a]	TC [8] LDL-C [6,7,9,10] Non-HDL-C [5]	0.22	8.72E-14	2.46E-14
rs4245791	2	44074431	ABCG8 (intron 3)		0.32	1.82E-07	3.33E-08
rs6730157	2	135907088	RAB3GAP1 (intron 17)	LDL-C: P = 0.018 [2] [b]	0.45	8.51E-08	2.16E-08
rs12465802	2	136381348	R3HDM1 (intron 7)	LDL-C: P = 0.022 [2] [b]	0.44	2.63E-08	7.98E-09
rs4954280	2	136420690	R3HDM1 (intron18)	LDL-C: P = 0.007 [2] [b]	0.33	1.49E-07	5.87E-08
rs2322660	2	136557319	LCT (intron 12)	LDL-C: P = 0.055 [2] [b] TC: P = 0.003-0.005 [17] LDL-C: P = 0.002-0.0005 [17]	0.35	2.42E-08	7.08E-09
rs309180	2	136614255	MCM6 (intron 11)	LDL-C: P = 0.057 [2] [b]	0.36	2.43E-08	8.39E-09
rs632632	2	136638216	4.2kb upstream MCM6	LDL-C: P = 0.216 [2] [b]	0.36	2.50E-08	1.03E-08

[a] This gene was reported to be associated with TC [1].

[b] Available at www.sph.umich.edu/csg/abecasis/public/lipids2008/

Table 5.2: Single-locus SNP effects for HDL-C with genome control (GC) adjusted $P < 7.2\times10^{-8}$.

SNP	Chr	Position	Gene Region	Reported SNP effect	MAF	Effect Type & P value	
						Genotype	Additive
rs17482753	8	19832646	8kb downstream LPL[a]	Triglyceride [7]	0.10	1.27E-08	3.50E-09
rs10503669	8	19847690	23kb downstream LPL[a]	HDL-C [4,11]	0.09	3.75E-08	1.14E-08
rs17410962	8	19848080	23kb downstream LPL[a]		0.12	2.75E-08	4.27E-09
rs17489268	8	19852045	27kb downstream LPL[a]		0.27	2.28E-10	5.97E-11
rs17411031	8	19852310	28kb downstream LPL[a]	HDL-C [7]	0.27	3.06E-10	7.21E-11
rs17489282	8	19852518	28kb downstream LPL[a]		0.25	2.33E-09	6.54E-10
rs4922117	8	19852586	28kb downstream LPL[a]		0.25	2.28E-09	7.73E-10
rs17411126	8	19855272	31kb downstream LPL[a]		0.27	6.14E-10	1.23E-10
rs765547	8	19866274	42kb downstream LPL[a]		0.27	1.60E-10	3.07E-11
rs11986942	8	19867445	43kb downstream LPL[a]		0.33	2.22E-07	5.53E-08
rs1837842	8	19868290	44kb downstream LPL[a]		0.27	2.14E-10	4.09E-11
rs1919484	8	19869676	45kb downstream LPL[a]		0.27	4.12E-10	7.21E-11
rs17259942	12	77072077	OSBPL8-ZDHHC17		0.12	8.61E-08	1.81E-08
rs9989419	16	56985139	HERPUD1-CETP[a]	HDL-C [4,7]	0.40	4.57E-13	5.96E-14
rs1800775	16	56995236	0.5kb upstream CETP[a]	HDL-C [3,12]	0.45	1.54E-29	1.64E-30
rs7240405	18	47159090	LIPG[a]-ACAA2[b]		0.16	8.01E-08	1.58E-08
rs4939883	18	47167214	LIPG[a]-ACAA2[b]	HDL-C [1,2]	0.17	1.07E-07	1.85E-08
rs1943981	18	47169815	LIPG[a]-ACAA2[b]		0.17	1.49E-07	2.50E-08
rs2156552	18	47181668	LIPG[a]-ACAA2[b]	HDL-C [3,4]	0.16	5.87E-08	1.14E-08
rs6507945	18	47243912	LIPG[a]-ACAA2[b]		0.43	4.34E-08	6.91E-09

[a] This gene was reported in [1-4].

[b] This gene was reported in [3].

Table 5.3: Epistasis effects for TC and HDL-C with the significance of suggestive linkage.

SNP1	Chr1	Pos1	Gene1	MAF1	SNP2	Chr2	Pos2	Gene2	MAF2	P value		
										Genotype	Epistasis	
rs4437278	4	12488199	U6[a] (174kb)	0.15	rs705169	10	125285443	GRP26(140kb)	0.49	4.49E-10	AA	5.93E-12
rs4738150	8	72607907	U8[a] (120kb)	0.40	rs16918936	9	33009027	APTX, LOC646808	0.04	4.68E-13	AD	1.29E-12
rs10476539	5	91991628	AC026781.5[a] (62kb)	0.18	rs2392885	8	129003117	PVT1[a]	0.28	2.63E-10	AA	2.99E-12
rs4706271	6	70390132	LMBRD1(intron14)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	7.67E-11	DD	5.72E-13
rs7741758	6	70412380	LMBRD1(intron9)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	5.73E-10	DD	4.60E-12
rs9346333	6	70426479	LMBRD1(intron8)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	1.39E-10	DD	1.16E-12
rs9351772	6	70428200	LMBRD1(intron8)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	5.38E-10	DD	4.25E-12
rs7762400	6	70445634	LMBRD1(intron7)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	4.41E-10	DD	3.52E-12
rs9294851	6	70457629	LMBRD1(intron5)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	2.83E-10	DD	2.22E-12
rs9354890	6	70504296	LMBRD1(intron1)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	5.54E-10	DD	4.18E-12
rs9364063	6	70514750	LMBRD1(8kb)	0.41	rs6581219	12	59213144	LRIG3(53kb)	0.42	4.65E-10	DD	4.46E-12
rs2787520	6	106821428	ATG5(48kb)	0.35	rs7236739	18	20800715	CABLES1(intron4)	0.31	2.98E-10	AA	2.48E-12
rs2842169	10	128330713	AL583860.7[a] (85kb)	0.10	rs4756344	11	36765284	C11orf74(84kb)	0.26	3.56E-10	AA	3.73E-12
rs17623128	16	77630108	AC092724.2[a] (114kb)	0.33	rs6506699	18	9775566	RAB31	0.49	4.65E-10	DD	2.78E-12
rs12596869	16	77630380	AC092724.2[a] (114kb)	0.33	rs6506699	18	9775566	RAB31	0.49	3.81E-10	DD	2.30E-12

[a] This was an RNA gene.

Table 5.4: Frequency and effect of gamete, allele-genotype or genotype-genotype combination in each epistasis effect with statistical significance of suggestive linkage.

Trait	SNP1	Chr1	SNP2	Chr2										
TC	rs4437278	4	rs705169	10	Gamete	A-T	G-G	G-T	A-G					
					Frequency	0.0785	0.413	0.434	0.0742					
	rs4738150	8	rs16918936	9	Allele-genotype	A-GG	G-GT	A-TT	G-TT	A-GT	G-GG			
					Frequency	0.0015	0.0284	0.557	0.374	0.0387	0.001			
HDL-C	rs10476539	5	rs2392885	8	Gamete	A-C	G-T	G-C	A-T					
					Frequency	0.0538	0.593	0.226	0.127					
	rs4706271	6	rs6581219	12	Genotype-genotype	GT-GG	GG-AG	GT-AA	TT-AG	TT-AA	TT-GG	GT-AG	GG-AA	GG-GG
					Frequency	0.0817	0.0887	0.161	0.169	0.112	0.0619	0.24	0.055	0.0298
	rs2787520	6	rs7236739	18	Gamete	G-G	T-A	G-A	T-G					
					Frequency	0.106	0.448	0.241	0.206					
	rs2842169	10	rs4756344	11	Gamete	C-A	T-G	T-A	C-G					
					Frequency	0.0751	0.237	0.661	0.0272					
	rs12596869	16	rs6506699	18	Genotype-genotype	CC-AG	CT-AA	CT-GG	TT-AG	TT-GG	TT-AA	CT-AG	CC-GG	CC-AA
					Frequency	0.0501	0.103	0.115	0.208	0.131	0.115	0.215	0.0344	0.0279

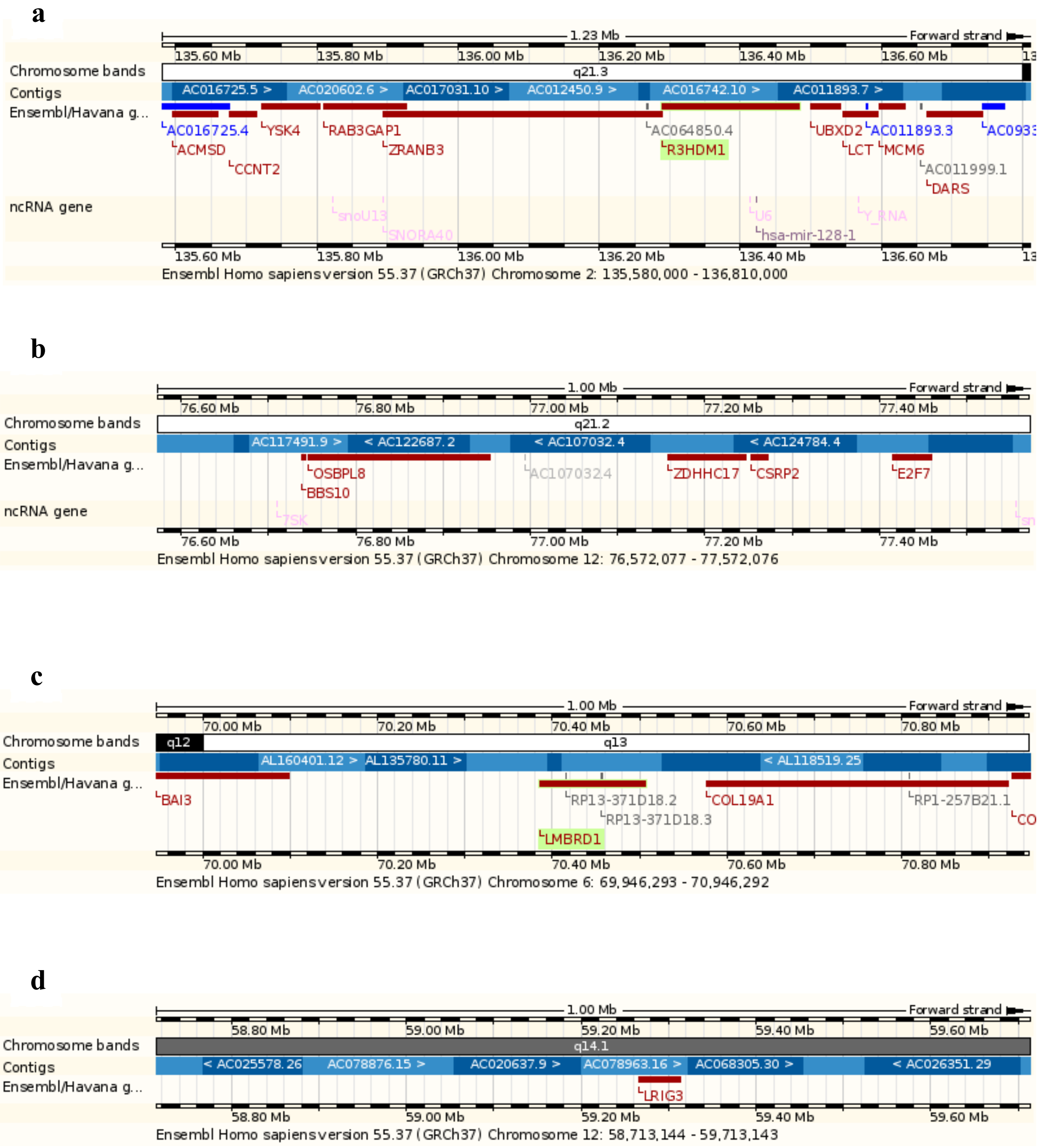


Figure 5.1

Gene regions associated with total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C). a, A 1.23 Mb region containing RAB3GAP1-R3HDM1-LCT-MCM6 with multiple SNP effects on TC. **b**, One Mb region containing OSBPL8-ZDHHC17 associated with HDL-C. **c**, One Mb region containing LMBRD1 that had multiple SNPs interacting with an SNP near LRIG3 for TC. **d**, One Mb region containing LRIG3 which was near an SNP interacting with multiple SNPs in LMBRD1 for TC.

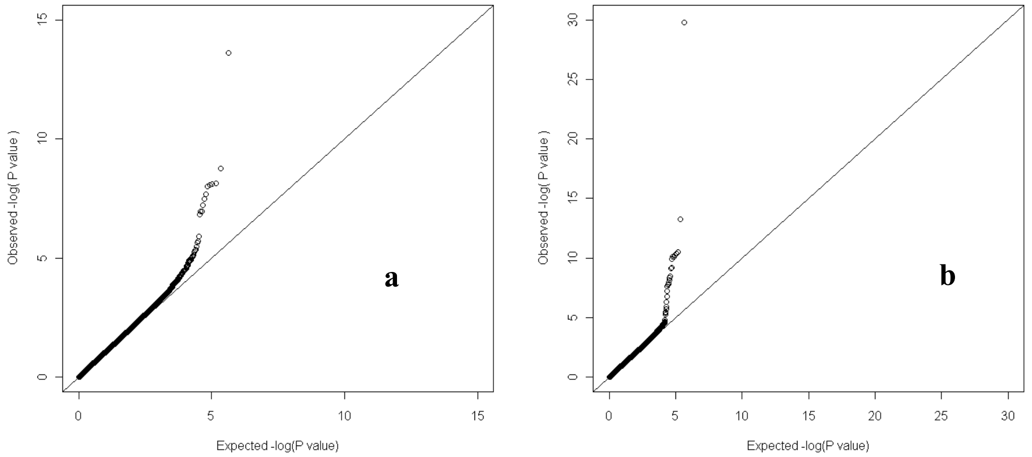


Figure 5.2
QQ plots for single-SNP whole genome association tests of total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C). a, TC. **b**, HDL-C.

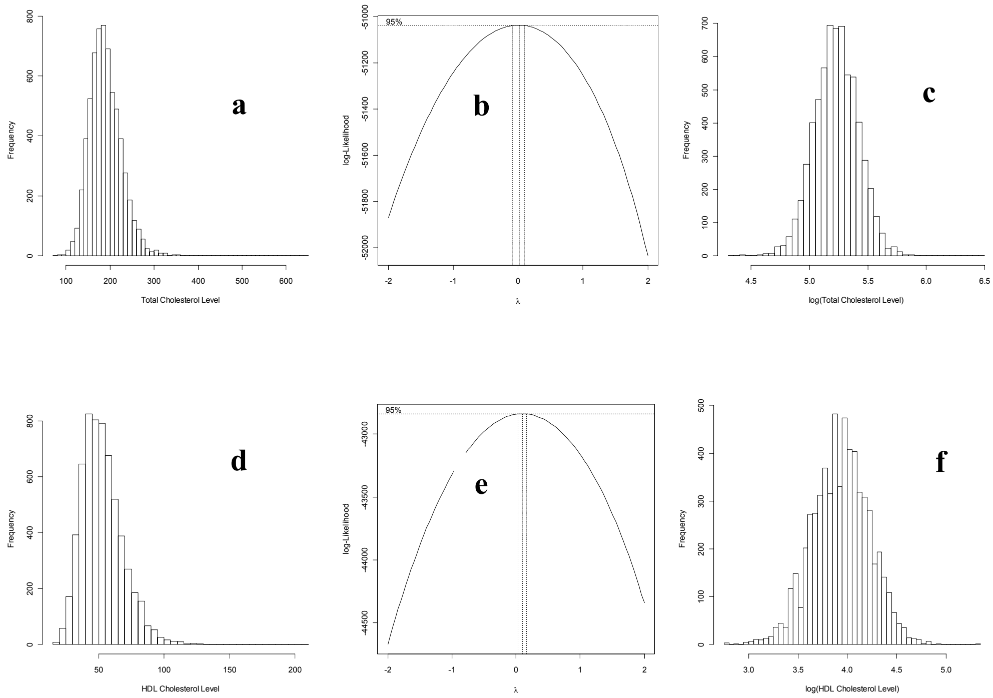


Figure 5.3

Distributions of total cholesterol (TC) and high-density lipoprotein cholesterol (HDL-C) in original scales and in log-transformed scales. a, Distribution of TC in original scale deviated from normality and had an outlier to the far right. **b**, The Box-Cox maximum likelihood analysis showed that log-transformation ($\lambda \approx 0$) was the best transformation to achieve normality for TC. **c**, Log-transformed TC values achieved normality. One outlier to the far right was removed from the data analysis. **d**, Distribution of HDL-C in original scale deviated from normality and had some outliers to the right. **e**, The Box-Cox maximum likelihood analysis showed that log-transformation ($\lambda \approx 0$) was the best transformation to achieve normality for HDL-C. **f**, Log-transformed HDL-C values achieved normality without serious outliers.

Chapter 6

A general approach for detecting imprinting, gene-gene, gene-sex, and gene-environment epigenetic effects of quantitative traits in genome-wide association studies

Epigenetics refers to changes in gene action that do not involve changes in the underlying DNA sequence of the organism. By this definition, imprinting, gene-gene, gene-sex, and gene-environment interactions belong to epigenetic effects. In this study, we develop a new general approach for detecting interactions between additive, dominance, imprinting, sex, and environment effects for each pair of bi-allelic loci in general populations where Hardy-Weinberg disequilibrium (HWD) and linkage disequilibrium (LD) may exist. This approach starts with tests of fifteen main effects and gene-gene interaction effects for two loci, including seven new effects involving imprinting and eight previously defined effects. The tests of single-locus imprinting effects were based on contrasts of dominance definitions of heterozygous genotypes with opposite parental allele origins. Gene-gene interactions involving imprinting effects, including additive × imprinting, imprinting × additive, dominance × imprinting, imprinting × dominance, and imprinting × imprinting, were defined as contrasts of additive × dominance, dominance × additive, and dominance × dominance effects. The genotypic values of two loci are then expressed as the population mean and fifteen genetic and epigenetic effects, with functions of marginal and conditional probabilities being the model matrices. Tests of a gene-sex or gene-environment interaction is based on the Hadamard product of the model matrix for a gene effect and that for sex or environment effect, yielding tests of fifteen gene-sex effects and fifteen gene-environment effects. We estimate that this new approach could analyze any large scale genome-wide association study contemporary using parallel computing.

6.1 Introduction

Epigenetic changes are crucial for the development and differentiation of the various cell types in an organism, as well as for normal cellular processes such as X-chromosome inactivation in female mammals and silencing of mating-type loci in yeast, and are involved in human disease as well as normal development (Bird, 2007; Eccleston et al., 2007; Feinberg, 2007; Johannes et al., 2007; Van Speybroeck, 2002). Epigenetics is at the heart of phenotypic variation in health and disease so that understanding and manipulating the epigenome holds enormous promise for preventing and treating common human illness (Eccleston et al., 2007).

Epigenetics refers to changes in gene action that do not involve changes in the underlying DNA sequence of the organism (Feinberg, 2007; Van Speybroeck, 2002). By this definition, a broad range of complex genetic effects can be characterized as epigenetic effects, including imprinting, gene-gene, gene-sex, and gene-environment interactions. One or a combination of these epigenetic factors could be responsible for the elusive nature of genetic mechanisms underlying complex phenotypes. Imprinting is widely considered as a major mechanism of epigenetics (Bird, 2007; Eccleston et al., 2007; Esteller, 2007; Feinberg, 2007; Johannes et al., 2007). Imprinting effect is the parent-of-origin effect, meaning that the same alleles from the father and mother do not have the same gene expression levels owing to the silencing of the allele related to differential levels of DNA methylation. Imprinting of the X-chromosome is a mechanism of X-chromosome inactivation (Okamoto et al., 2004). Testing of interaction between imprinting and other genetic and environment factors including sex of the individual may provide new insights into the role of imprinting in phenotypic expression. Recent studies suggest that sex-specific genetic architecture also influences human phenotypes, including reproductive, physiological and disease traits, and that genetic studies that ignore sex-specific effects in their design and interpretation could fail to identify a significant proportion of the genes that contribute to risk for complex diseases (Ober et al., 2008). Examples of reversal dominance (an allele being dominant in one sex and recessive in the other) were reported in 1916 (Osborn , 1916; Wentworth, 1916) and more broad evidence of sex-gene interactions was discovered by recent studies (Ober et al., 2008). Common diseases may involve phenotypic variants with both genetic variation and environmentally triggered epigenetic change that modulates the effects of DNA sequence variation. Environmental factors (hormones, growth factors, toxins and dietary methyl donors) influence both the genome and epigenome and these epigenetic modifiers in turn are affected by variation in the genes that encode them (Bjornsson et al., 2004; Feinberg, 2007). Age-gene interaction as indicated by age-dependent penetrance affects many human complex diseases including mental health (Crowe and Smouse, 1977; Farrer, 2006; Rutte, 2005; Yarden et al., 2008). Human genetic data often record a large number of environmental factors, such as age, smoking, drinking, diet, and medical and fitness treatments. While environment factors are known to affect disease expression, little is known how environment affects various types of genetic effects at the level of genome-wide SNP-phenotype association. Epistasis refers to gene-gene interaction effects (Fisher, 1918; Cockerham, 1954; Kempthorne, 1954). Epistasis belongs to epigenetic effects because epistasis effects involves changes in gene-gene combinations that do not involve DNA sequence changes of the loci. The significance of epistasis in complex traits has been well recognized (Carlborg and Haley,2004; Moore, 2003; Sanjuán and Elena, 2006). The global analysis of gene-interaction patterns bears a striking resemblance to what is now called systems biology (Moore and Williams, 2005; Phillips 2008) Given recent work in this area, it is likely that for the next century the concept of epistasis will be even more central to biology than it was over the past century (Phillips, 2008).

Genome-wide association study (GWAS) is an unprecedented powerful tool for detecting complex epigenetic effects. The large number of SNPs available to GWAS increases the likelihood to discover causative SNPs and increases the statistical power for detecting SNPs in linkage disequilibrium with causative mutations. Sample sizes of GWAS are on the increase. The large sample sizes make testing and estimating many levels of genetic and environment factors possible. Analytical and computing tools are needed to fully realize the potential of GWAS in understanding the genetic mechanisms including epigenetics underlying complex phenotypes.

In this study, we develop a general quantitative genetics approach for detecting imprinting, gene-gene, gene-sex, and gene-environment epigenetic effects in GWAS. This approach is expected to computationally feasible for analyzing any large-scale GWAS.

6.2 Two-locus Epigenetic Model

The general epigenetic approach starts with the two-locus epigenetic model that allows Hardy-Weinberg disequilibrium (HWD) and linkage disequilibrium (LD). Two biallelic autosome loci, locus 1 with *A* and *a* alleles and locus 2 with *B* and *b*, are assumed to affect a complex trait of quantitative nature. Application to X chromosome and pseudoautosomal region s will be based on results under the autosome assumption. Parental origin of each allele is assumed to be known so that two bi-allelic loci have 16 possible genotypes. Let g_{jklm} and p_{jklm} denote the genotypic value and genotypic probability of individuals possessing genotype jk at locus 1 (jk = *AA*, *Aa*, *aA*, *aa*) and lm at locus 2 (lm = *BB*, *Bb*, *bB*, *bb*) (Table 6.1). The partition of genotypic values uses the extended Kempthorne model (Mao et al., 2006). Under the Kempthorne model, each genetic effect is defined as a deviation of genetic combination effect from the sum of individual genetic factors in the genetic combination and has a straightforward biological interpretation. Imprinting effect of each locus is defined based on the comparison of two alternative heterozygous genotypes (de Koning et al., 2002; London, 2004) , e.g., *Aa* and *aA*, where the allele on the left is assumed to originate from the father, and the allele on the right is assumed to be from the mother. The rationale of this definition of imprinting effect is illustrated in Figure 6.1. In this study, imprinting effect of each heterozygous genotype is defined as the deviation of the dominance deviation of this genotype from the average of dominance deviations of the two alternative heterozygous genotypes. This type of definition for imprinting effect as is then extended to pairwise interaction effects involving imprinting. The two-locus epigenetic model requires calculations of a series marginal and conditional probabilities as well as a series marginal and conditional means (online **Supplementary Material**). Let μ = the mean genotypic value in the population, μ_j = the marginal mean of genotypic values for individuals with allele j (j = A, a), μ_l = the marginal mean of genotypic values for individuals with allele l (l = B, b), μ_{jk} = the marginal mean of genotypic values for individuals with genotype jk at locus 1 with $\mu_{jk} \neq \mu_{kj}$ if j ≠ k and imprinting is present, μ_{lm} = the marginal mean of genotypic values for individuals with genotype lm at locus 2 with $\mu_{lm} \neq \mu_{ml}$ if l ≠ m and imprinting is present, μ_{jl} = the marginal mean of genotypic values for individuals with allele j at locus 1 and allele l at locus 2 (j = A, a; l = B, b), μ_{jlm} = the marginal mean of genotypic values for individuals with allele j at locus 1 and genotype lm at locus 2 with $\mu_{jlm} \neq \mu_{jml}$ if l ≠ m and imprinting is present, and μ_{jkl} = the marginal mean of genotypic values for individuals with genotype jk at locus 1 and allele l at locus 2 with $\mu_{jkl} \neq \mu_{kjl}$ if j ≠ k and imprinting is present. Then, the two-locus epigenetic model and can be expressed as follows,

$$\begin{aligned} g_{jklm} &= \mu + a_j + a_k + a_l + a_m + d_{jk} + d_{lm} + i_{jk} + i_{lm} + (aa)_{jl} + (aa)_{jm} + (aa)_{kl} + (aa)_{km} \\ &\quad + (ad)_{jlm} + (ad)_{klm} + (ai)_{jlm} + (ai)_{klm} + (da)_{jkl} + (da)_{jkm} + (ia)_{jkl} + (ia)_{jkm} \\ &\quad + (id)_{jklm} + (di)_{jklm} + (ii)_{jklm} + (dd)_{jklm} \qquad [1] \end{aligned}$$

In Eq[1], terms involving additive effects only have the same definitions as in the extended Kempthorne model (Mao et al., 2006), i.e., $a_j = \mu_j - \mu$ = additive effect of allele j of locus 1 (j = *A*, *a*), $a_l = \mu_l - \mu$ = additive effect of allele l of locus 2 (l = *B*, *b*), $(aa)_{jl} = \mu_{jl} - \mu - a_j - a_l$ = additive × additive epistasis effect of genotypes with alleles j and l. Other effects in Eq[1] are defined based on the following effects that are essentially the same as in the extended Kempthorne model except imprinting implications of terms involving dominance deviations and calculation details (online **Supplementary Material**). These effects are: $\delta_{jk} = \mu_{jk} - \mu - a_j - a_k$ = dominance effect of locus 1 with $\delta_{jk} \neq \delta_{kj}$ if $j \neq k$ and imprinting is present; $\delta_{lm} = \mu_{lm} - \mu - a_l - a_m$ = dominance effect of locus 2 with $\delta_{lm} \neq \delta_{ml}$ if $l \neq m$ and imprinting is present; $(a\delta)_{jlm} = \mu_{jlm} - \mu - a_j - a_l - a_m - \delta_{lm} - (aa)_{jl} - (aa)_{jm}$ = additive ×dominance epistasis effect with $\delta_{lm} \neq \delta_{ml}$ if $l \neq m$ and imprinting is present; $(\delta a)_{jkl} = \mu_{jkl} - \mu - a_j - a_k - a_l - \delta_{jk}$ain $-(aa)_{jl} - (aa)_{kl}$ = dominance × additive epistasis effect with $\delta_{jk} \neq \delta_{kj}$ if $j \neq k$ and imprinting is present; and $(\delta\delta)_{jklm} = g_{jklm} - \mu - a_j - a_k - a_l - a_m - \delta_{jk} - \delta_{lm} - (aa)_{jl} - (aa)_{jm} - (aa)_{kl} - (aa)_{km} - (a\delta)_{jlm} - (a\delta)_{klm} - (\delta a)_{jkl} - (\delta a)_{jkm}$ = dominance × dominance epistasis effect, with $(\delta\delta)_{jklm} \neq (\delta\delta)_{jkml}$ for $l \neq m$, $(\delta\delta)_{jklm} \neq (\delta\delta)_{kjlm}$ for $j \neq k$, and $(\delta\delta)_{jklm} \neq (\delta\delta)_{kjml}$ for $j \neq k$ and $l \neq m$, if imprinting is present. With these definitions, each genetic effect in Eq[1] involving dominance or imprinting is defined as average deviations or a deviation from an average of deviations:

$d_{jk} = \frac{1}{2}(\delta_{jk} + \delta_{kj})$ = dominance effect of jk genotype of locus 1 [2.1]

$d_{lm} = \frac{1}{2}(\delta_{lm} + \delta_{ml})$ = dominance effect of lm genotype of locus 2 [2.2]

$i_{jk} = \delta_{jk} - \frac{1}{2}(\delta_{jk} + \delta_{kj}) = \frac{1}{2}(\delta_{jk} - \delta_{kj})$ = imprinting effect of jk genotype of locus 1 [2.3]

$i_{lm} = \delta_{lm} - \frac{1}{2}(\delta_{lm} + \delta_{ml}) = \frac{1}{2}(\delta_{lm} - \delta_{ml})$ = imprinting effect of lm genotype of locus 2 [2.4]

$(ad)_{jlm} = (ad)_{jml} = ½[(a\delta)_{jlm} + (a\delta)_{jml}]$ [2.5]

= additive ×dominance epistasis effect

$(ai)_{jlm} = (a\delta)_{jlm} - ½[(a\delta)_{jlm} + (a\delta)_{jml}] = ½[(a\delta)_{jlm} - (a\delta)_{jml}]$ [2.6]

= additive ×imprinting epigenetic effect

$(da)_{jkl} = (da)_{kjl} = ½[(\delta a)_{jkl} + (\delta a)_{kjl}]$ [2.7]

= dominance × additive epistasis effect

$(ia)_{jkl} = (\delta a)_{jkl} - ½[(\delta a)_{jkl} + (\delta a)_{kjl}] = ½[(\delta a)_{jkl} - (\delta a)_{kjl}]$ [2.8]

= imprinting × additive epigenetic effect

$(dd)_{hlm} = ½[(\delta\delta)_{jklm} + (\delta\delta)_{kjlm}]$ with subscript h denoting $j \neq k$ [2.9]

= dominance ×dominance epistasis effect when locus 1 is heterozygous

= $(\delta\delta)_{jjlm}$ or $(\delta\delta)_{kklm}$ if $j = k$

$(dd)_{jkh} = ½[(\delta\delta)_{jklm} + (\delta\delta)_{jkml}]$ with subscript h denoting $l \neq m$ [2.10]

= dominance ×dominance epistasis effect when locus 2 is heterozygous

= $(\delta\delta)_{jkll}$ or $(\delta\delta)_{jkmm}$ if $l = m$

$(dd)_{jklm} = ¼[(\delta\delta)_{jklm} + (\delta\delta)_{jkml} + (\delta\delta)_{kjlm} + (\delta\delta)_{kjml}]$ [2.11]

= dominance ×dominance epistasis effect of ijkl genotype

$(di)_{jklm} = (dd)_{jklm} - (dd)_{jkh} = ¼[(\delta\delta)_{kjlm} + (\delta\delta)_{kjml} -(\delta\delta)_{jklm} - (\delta\delta)_{jkml}]$ [2.12]

= dominance ×imprinting epigenetic effect

$(id)_{jklm} = (dd)_{jklm} - (dd)_{hlm} = ¼[-(\delta\delta)_{jklm} + (\delta\delta)_{jkml} - (\delta\delta)_{kjlm} + (\delta\delta)_{kjml}]$ [2.13]

= imprinting × dominance epigenetic effect

$(ii)_{jklm} = (\delta\delta)_{jklm} - (id)_{jklm} - (di)_{jklm} - (dd)_{jklm} = ¼[(\delta\delta)_{jklm} - (\delta\delta)_{jkml} - (\delta\delta)_{kjlm} + (\delta\delta)_{kjml}]$

= imprinting ×imprinting epigenetic effect. [2.14]

Note that $i_{jk} = (ia)_{jkl} = (di)_{jklm} = 0$ if $j = k$, $i_{lm} = (ai)_{jlm} = (id)_{jklm} = 0$ if $l = m$, and $(ii)_{jklm} = 0$ if $j = k$ or $l = m$.

Based on the genetic and epigenetic effects defined in Eq[1] and by Eq[2.1-2.14], fifteen contrasts can be defined for testing fifteen genetic and epigenetic effects of the two-locus epigenetic model given by Eq[1], i.e., b_{a1}, b_{a2}, = additive effects of locus 1 and locus 2, b_{d1}, b_{d2} = dominance effects of locus 1 and locus 2, b_{i1}, b_{i2} = imprinting effects of locus 1 and locus 2, b_{aa} additive × additive epistasis effect, b_{ad} = additive × dominance epistasis effect, b_{da} = dominance × additive epistasis effect, b_{ai} = additive × imprinting effect, b_{ia} = dominance × additive epistasis effect, b_{dd} dominance ×dominance epistasis effect, b_{di} = dominance × imprinting effect, b_{id} = imprinting × dominance effect, and b_{ii} = imprinting × imprinting epigenetic effect (online **Supplementary Material**). Each of these contrasts is expected to be null if the true effect is absent. Let

$\mathbf{g} = (g_{AABB}, g_{AABb}, g_{AAbB}, g_{AAbb}, g_{AaBB}, g_{AaBb}, g_{AabB}, g_{Aabb}, g_{aABB}, g_{aABb}, g_{aAbB}, g_{aAbb}, g_{aaBB}, g_{aaBb}, g_{aabB}, g_{aabb})'$ [3]

$\mathbf{b} = (\mu, b_{a1}, b_{a2}, b_{d1}, b_{d2}, b_{i1}, b_{i2}, b_{aa}, b_{ad}, b_{da}, b_{ai}, b_{ia}, b_{dd}, b_{di}, b_{id}, b_{ii})'$. [4]

Then, **b** can be expressed as linear functions of genotypic values of Eq.[3], i.e.,

$\mathbf{b} = \mathbf{Sg}$ [5]

where $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6, \mathbf{s}_7, \mathbf{s}_8, \mathbf{s}_9, \mathbf{s}_{10}, \mathbf{s}_{11}, \mathbf{s}_{12}, \mathbf{s}_{13}, \mathbf{s}_{14}, \mathbf{s}_{15}, \mathbf{s}_{16})$ = function of marginal and conditional probabilities (online **Supplementary Material**). From Eq[4-5], the unique solutions of the genotypic values in terms of population mean and the fifteen genetic and epigenetic effects are:

$\mathbf{g} = \mathbf{S}^{-1}\mathbf{b} = \mathbf{Tb} = \mathbf{t}_1 b_1 + \mathbf{t}_2 b_2 + ... + \mathbf{t}_{16} b_{16}$ [6]

Each of the fifteen genetic and epigenetics effects in Eq[4] is a linear function of the sixteen genotypic values. Renaming the fifteen genetic and epigenetic effects in Eq[4] as $b_2, ..., b_{16}$. Then,

$$b_k = \mathbf{s}_k\mathbf{g}, \qquad k = 2, ..., 16 \qquad [7]$$

The significance of b_k can be tested using the estimated value of Eq[6] from phenotypic observations.

6.3 Statistical Testing of Epigenetic Effects Involving Gene, Sex, and Environment Factors

The statistical test of interactions between two effects is through the product of the coefficients of the two effects. For example, if the main effect model of two effects is $y = \mu + x_1b_1 + x_2b_2 + e$, then the interaction between effects 1 and 2 (b_{12}) is tested by the model $y = \mu + x_1b_1 + x_2b_2 + (x_1 * x_2)b_{12} + e$. Combining this statistical result with the two-locus epigenetic model of Eq[1-7], the statistical test of epigenetic effects can be developed. For convenience, only one environmental factor is assumed because extension to two or more environmental factors is straightforward. From Eq[6], the deviation of SNP genotypic values from the population mean of genotypic values is:

$$\mathbf{g}_c = \mathbf{g} - \mathbf{t}_1b_1 = \mathbf{t}_2b_2 + ... + \mathbf{t}_{16}b_{16} = (\mathbf{t}_2, ..., \mathbf{t}_{16})\mathbf{b}_g \qquad [8]$$

where

$$\mathbf{b}_g = (\, b_{a1}, b_{a2}, b_{d1}, b_{d2}, b_{i1}, b_{i2}, b_{aa}, b_{ad}, b_{da}, b_{ai}, b_{ia}, b_{dd}, b_{di}, b_{id}, b_{ii}\,)' = (b_2, ..., b_{16})'. \qquad [9]$$

Eq[8] is the total genetic effect as a summation of all genetic and epigenetic effects. The phenotypic model with epigenetic effects involving gene, sex and environment factors is: y = (intercept) + (sex effect) + (environment effect) + (other non-genetic effects) + (15 SNP effects) + sex×(15 SNP effects) + environment×(15 SNP effects) + (random residual). Using matrix notation, this model can be described as:

$$\mathbf{y} = \mathbf{X}_0 b_0 + \mathbf{X}_s s + \mathbf{X}_e \mathbf{e} + \mathbf{X}_f \mathbf{f} + \mathbf{X}_g(t_2, ..., \mathbf{t}_{16})\mathbf{b}_g + \mathbf{X}_s\#[\mathbf{X}_g(\mathbf{t}_2, ...t_{16})]\mathbf{b}_{gs} + \mathbf{X}_e\#[\mathbf{X}_g(\mathbf{t}_2, ..., t_{16})]\mathbf{b}_{ge} + \boldsymbol{\varepsilon} \quad [10]$$

where $\mathbf{y}$ = column vector of phenotypic observations, b_0 = the intercept, $\mathbf{X}_0$ = model matrix for b_0, s = sex effect, $\mathbf{X}_s$ = model matrix (column vector) of sex effect with '1' for male and '-1' for female, $\mathbf{X}_g$ = model matrix of SNP genotypic effect ($\mathbf{g}_c$), $\mathbf{e}$ = column vector of the environment effect(s) that may interact with genetic effects in $\mathbf{b}_g$, $\mathbf{X}_e$ = model matrix of $\mathbf{e}$, $\mathbf{f}$ = an arbitrary number of non-genetic factors that are not tested for interaction with genetic factors, $\mathbf{X}_f$ = model matrix of $\mathbf{f}$, # denotes the Hadamard product of two matrices (Searle, 1982), $[\mathbf{X}_s\#(\mathbf{X}_g\mathbf{t}_2), ..., \mathbf{X}_s\#(\mathbf{X}_g\mathbf{t}_{16})]$ = model matrices of gene-sex and gene-gene-sex interactions, $[\mathbf{X}_e\#(\mathbf{X}_g\mathbf{t}_2), ..., \mathbf{X}_e\#(\mathbf{X}_g\mathbf{t}_{16})]$ = model matrices of gene-environment and gene-gene-environment interactions, $\boldsymbol{\varepsilon}$ = random residual values, and

$$\mathbf{b}_{gs} = (b_{a1s}, b_{a2s}, b_{d1s}, b_{d2s}, b_{i1s}, b_{i2s}, b_{aas}, b_{ads}, b_{das}, b_{ais}, b_{ias}, b_{dds}, b_{dis}, b_{ids}, b_{iis})' = (b_{17}, ..., b_{31})' \quad [11]$$

$$\mathbf{b}_{ge} = (b_{a1e}, b_{a2e}, b_{d1e}, b_{d2e}, b_{i1e}, b_{i2e}, b_{aae}, b_{ade}, b_{dae}, b_{aie}, b_{iae}, b_{dde}, b_{die}, b_{ide}, b_{iie})' = (b_{32}, ..., b_{46})'. \quad [12]$$

Interpretations of the 'b' effects in Eq[11-12] are given by the subscripts, e.g., $b_{ids} = b_{30}$ = epigenetic effect due to the 3rd order interaction between imprinting effect of locus 1, dominance effect of locus 2, and sex effect; and $b_{aie} = b_{42}$ = epigenetic effect due to the 3rd order interaction between additive effect of locus 1, imprinting effect of locus 2, and the environment effect. The expression of environment effects, $\mathbf{X}_e\mathbf{e}$, has the flexibility to accommodate multiple levels of environment effect. For computational convenience, only one level in the **e** vector is assumed so that the gene-environment interactions can be defined by Eq[12] . For binary environment effects such as smoking status and treatment status (yes or no), $\mathbf{X}_e$ can be defined as a column vector with '1' for one level and '-1' for the other level. Continuous environment factors such as age and weight have only one column for $\mathbf{X}_e$, and environment factors with more than two levels can also be treated as a continuous variable. For computational efficiency with large numbers of tests, a two-step regression analysis can be used (Wolfinger et al., 2002; Mao et al., 2006). Let $\mathbf{y}_c = \mathbf{y} - (\mathbf{X}_s s + \mathbf{X}_e e + \mathbf{X}_f\mathbf{f})$ = corrected phenotypic values as the first step of the two-step regression analysis. From Eq[10], the phenotypic model for the second step of the two-step regression analysis is:

$$\begin{aligned}\mathbf{y}_c &= \mathbf{X}_0 b_0 + (\mathbf{Z}_2 \ldots \mathbf{Z}_{16})\mathbf{b}_g + (\mathbf{Z}_{17} \ldots \mathbf{Z}_{31})\mathbf{b}_{gs} + (\mathbf{Z}_{32} \ldots \mathbf{Z}_{46})\mathbf{b}_{ge} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}_0 b_0 + (\mathbf{Z}_g, \mathbf{Z}_{gs}, \mathbf{Z}_{ge})\mathbf{b}_G + \boldsymbol{\varepsilon} \qquad [13]\end{aligned}$$

where $\mathbf{Z}_g = (\mathbf{Z}_2 \ldots \mathbf{Z}_{16}) = \mathbf{X}_g(t_2, \ldots, \mathbf{t}_{16})$, $\mathbf{Z}_{gs} = (\mathbf{Z}_{17} \ldots \mathbf{Z}_{31}) = [\mathbf{X}_s\#(\mathbf{X}_g\mathbf{t}_2), \ldots, \mathbf{X}_s\#(\mathbf{X}_g\mathbf{t}_{16})]$, $\mathbf{Z}_{ge} = (\mathbf{Z}_{32} \ldots \mathbf{Z}_{46}) = [\mathbf{X}_e\#(\mathbf{X}_g\mathbf{t}_2), \ldots, \mathbf{X}_e\#(\mathbf{X}_g\mathbf{t}_{16})]$, and $\mathbf{b}_G = (\mathbf{b}_g', \mathbf{b}_{gs}', \mathbf{b}_{ge}')' = (b_2, \ldots b_{46})'$ (Eq[9] and Eq[11-12]). To account for potential correlation among observations, the variance-covariance matrix of $\mathbf{y}_c$ (or $\mathbf{y}$) is assumed as: $\text{var}(\mathbf{y}_c) = \mathbf{V}\sigma^2$, where $\mathbf{V} = \mathbf{I}$ = identity matrix for uncorrelated observations, and σ^2 = variance of a random residual. Then, a genetic or epigenetic effect in Eq.[13] can be tested as:

$$\hat{T}_k = \frac{\left|\hat{b}_k\right|}{\sqrt{s^2 v^{kk}}}, \text{ k=2,}\ldots\text{, 46} \qquad [14]$$

where $s^2 = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{g}}_c\right)' \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{g}}_c\right)/\left(n - r\right)$ = estimate of σ^2, v^{kk} = diagonal element at position (k, k) of $(\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}$, and where $\mathbf{W}$ = ($\mathbf{X}_0$, $\mathbf{Z}_g$, $\mathbf{Z}_{gs}$, $\mathbf{Z}_{ge}$), n = number of observations, r = rank of **W**. The test statistic follows a Student t-distribution with n−r degrees of freedom. Note that defining r = rank of **W** is based on Eq[13], which is the phenotypic model for the second step of the two-step regression analysis. To consider the degrees of freedom consumed in the first step of the two-step regression analysis, r should be based on the fill model given by Eq[10], as we did in the implementation of the two-step regression analysis for pairwise epistasis testing (Ma et al., 2008).

6.3.1 Application to X Chromosome and Pseudo-autosomal Regions

When either or both of the two SNPs are on the X chromosome, the interactions involving imprinting effects may provide insights about X chromosome inactivation because imprinting of a gene on the X chromosome could be part of the X chromosome inactivation mechanism[7]. For SNPs on the X chromosome, females can be used for detecting all but sex related epigenetic effects, whether one or both SNPs are X chromosome SNPs, but males can be used for detecting a much smaller number of effects. For two X chromosome SNPs, males can be used for detecting additive (A) and A×A effects so that using all females and males can detect A×S and A×A×S epigenetic effects, where ‘S’ denotes sex effect. For one autosome SNP and one X chromosome SNP, males can also be used for detecting A×D and A×i effects (Table 6.2) so that A×D×S and A×i×S can be tested when all females and males are used. The epigenetic modeling and testing leading to Eq[14] apply to pseudo-autosomal regions of the Y and X chromosomes.

6.3.2 Statistical Power, Sample Size, and Accuracy of Effect Estimates

Statistical power and sample size as functions of several statistical and genetic parameters were evaluated using analytical formulae (online **Supplementary Material**) and simulations. The simulated phenotypic value of each individual is obtained as the summation of the individual QTL genotypic value and a random residual following N(0,1) distribution. Each simulation generated 10,000 repeats. Statistical powers observed from the simulated data agreed well with the predicted powers. In general, imprinting was more difficult to detect than additive and dominance effects, interaction involving imprinting was more difficult to detect than the same order interaction effects that do not involving imprinting, and higher order interactions were difficult to detect than lower order interactions (Figures 6.2-6.5). The same trend was also true for the accuracy of effect estimates measured by mean square errors (Figure 6.6) and for sample size requirement (Table 6.4). Statistical power and sample size were sensitive to the effect size measured as the effect variance relative to the total phenotypic variance and improved rapidly as the effect size increases (Figures 6.2-6.3, Table 6.4). Statistical power increases and as sample size increases for a given effect size (Figures 6.4-6.5). The results show that detecting an epigenetic effect that accounts for 3% or more of the phenotypic variation is practical in GWAS with sample size above 2000. This estimate becomes more optimistic for larger sample sizes such as that the Framingham Heart Study that currently has 6575 individuals with SNP genotypes.

6.4 Discussion

Parental allele transmission for detecting imprinting related effects: The testing of imprinting related effect requires knowing parental allele transmission to the heterozygous offspring (Figure 6.1), but parental allele transmission cannot be determined unequivocally when parents and the child all have the same heterozygous SNP genotypes. Under the assumption of HWE, the probability that the parental alleles in a heterozygous offspring SNP genotype can be determined is $w_2 = 1 - (2pq)^2/Q$ when both parents have known SNP genotypes, and is $w_1 = 1 - 2pq/Q$ when one parent has known SNP genotype, where p and q are allele frequencies ($p + q = 1$), and $Q = 1 - p^4 - q^4$ (online **Supplementary Material**). For equal allele frequencies, $w_2 = 5/7 = 71\%$, and $w_1 = 3/7 = 43\%$, so that 29~57% of the heterozygous offspring could not be used for testing imprinting related effects if unequivocal determination of parental allele transmission is required. An ideal solution to the issue of parental allele transmission is to impute SNP genotype (Marchini et al., 2007; Scheet and Stephens, 2006) because all individuals can be used using imputed genotypic data. For data sets without parental genotypes, imputing is the only way to implement tests of imprinting related effects. In this case, a significant imprinting effect cannot be characterized as male or female imprinting.

Computational considerations: The genetic and phenotypic modeling (Eq[1-13]) leading to the general statistical testing of epigenetic effects could be extended for testing higher order interaction effects. Such extension to include more environment factors is computationally feasible using parallel computing because the number of environment factors that could be considered for epigenetic effects is limited. However, such extension to include interactions involving three SNPs or more is not yet feasible due to the large number of SNP combinations generated by adding the third SNP (Ma et al., 2008). Table 6.3 shows the observed computing times for pairwise analysis of the association between the 500k SNPs and four phenotypes on 6575 individuals from the Framingham Heart Study data. Based on those observed computing time, the completion of the epigenetic analysis in 24 hours would require about 5000 processor cores for 500k SNPs and 20,000 processor cores for 1000k SNPs. Therefore, timely epigenetic analysis for any large scale GWAS is possible using a massive parallel computing system such as the IBM Blue Gene system with over one million processor cores.

Extreme Allele Frequencies: Extreme allele frequencies cause highly unbalanced allelic and genotypic distributions and could increase the chance of false significant effects because of the increased chance of random association between the rare alleles or genotypes and extreme phenotypic values. Extreme allele frequencies could cause even worse unbalanced haplotype and genotypic distributions for pairwise analysis. In experimental and agricultural species, a designed population such as the F2 and backcross populations for mapping quantitative trait loci (QTL) should be helpful for minimizing the potential problems of extreme frequencies in GWAS analysis.

Table 6.1. Genotypes, genotypic frequencies (p_{jklm}) and genotypic values (g_{jklm}) of two quantitative trait loci.

AABB	*AABb*	*AAbB*	*AAbb*	*AaBB*	*aABB*	*AaBb*	*AabB*
p_{AABB}	p_{AABb}	p_{AAbB}	p_{AAbb}	p_{AaBB}	p_{aABB}	p_{AaBb}	p_{AabB}
g_{AABB}	g_{AABb}	g_{AAbB}	g_{AAbb}	g_{AaBB}	g_{aABB}	g_{AaBb}	g_{AabB}
aABb	*aAbB*	*Aabb*	*aAbb*	*aaBB*	*aaBb*	*aabB*	*aabb*
p_{aABb}	p_{aAbB}	p_{Aabb}	p_{aAbb}	p_{aaBB}	p_{aaBb}	p_{aAbB}	p_{aabb}
g_{aABb}	g_{aAbB}	g_{Aabb}	g_{aAbb}	g_{aaBB}	g_{aaBb}	g_{aAbB}	g_{aabb}

Table 6.2. Second order epigenetic effects for one autosome SNP and one X chromosome SNP detectable in males.

		Autosome SNP		
	Gene Effect	additive (A)	dominance (D)	imprinting (i)
X chromosome SNP	additive (A)	A×A	A×D	A×i

Table 6.3. Observed computing time (hours) for pairwsie epistasis testing of four phenotypes on 6575 individuals from the Framingham Heart Study.

Minor allele frequency required	Number of SNPs	Number of cores	Systolic blood pressure	Diastolic blood pressure	HDL cholesterol	Total cholesterol
0.05	385,182	632	18.5	18.31		
0.01	432,096	672	18.01	18.38	18.96	19

The total number of SNPs in the Framingham 500k data is 496,858 without minor allele frequency requirement. The computing time was based on run times on the SGI Altix XE 1300 Linux cluster system (Calhoun) with 2.66 GHz Intel processor-cores (total 2048 cores) at the Minnesota Supercomputer Institute.

Table 6.4. Sample size required to achieve 95% power with 5%type-I error.

Contrast heritability [a]				i	i × a	i × d	i × i
H_i^2	H_{ia}^2	H_{id}^2	H_{ii}^2	n_i	n_{ia}	n_{id}	n_{ii}
0.004	0.0030	0.0009	0.0007	2958	4006	13584	17975
0.008	0.0059	0.0017	0.0013	1330	1801	6107	8081
0.012	0.0089	0.0026	0.0020	787	1066	3615	4784
0.016	0.0119	0.0035	0.0026	516	699	2369	3135
0.020	0.0148	0.0044	0.0033	353	478	1621	2145
0.024	0.0178	0.0052	0.0039	245	331	1123	1486
0.028	0.0208	0.0061	0.0046	167	226	767	1015
0.032	0.0237	0.0070	0.0053	109	147	500	661
0.036	0.0267	0.0078	0.0059	64	86	292	387
0.040	0.0297	0.0087	0.0066	27	37	126	167

[a] '$b_{a1} = b_{a2} = b_{d1} = b_{d2} = b_{i1} = b_{i2} = b_{aa} = b_{ad} = b_{da} = b_{ai} = b_{ia} = b_{dd} = b_{di} = b_{id} = b_{ii}$' indicates that all the fifteen effects are assumed to be of the same size in defining each heritability. The H_i^2 values of 0.004~0.04 correspond to 0.1787 ~ 0.5650 of b_i/σ_y , i.e., the size of b_i is about 0.1787 ~ 0.5650 phenotypic standard deviations. The SNP genotypic frequencies for each pair of SNPs assumed the following values that allowed Hardy-Weinberg disequilibrium and linkage disequilibrium: **p** = (0.13, 0.05, 0.04, 0.02, 0.05, 0.08, 0.08, 0.04, 0.05, 0.08, 0.09, 0.03, 0.03, 0.03, 0.03, 0.17)'.

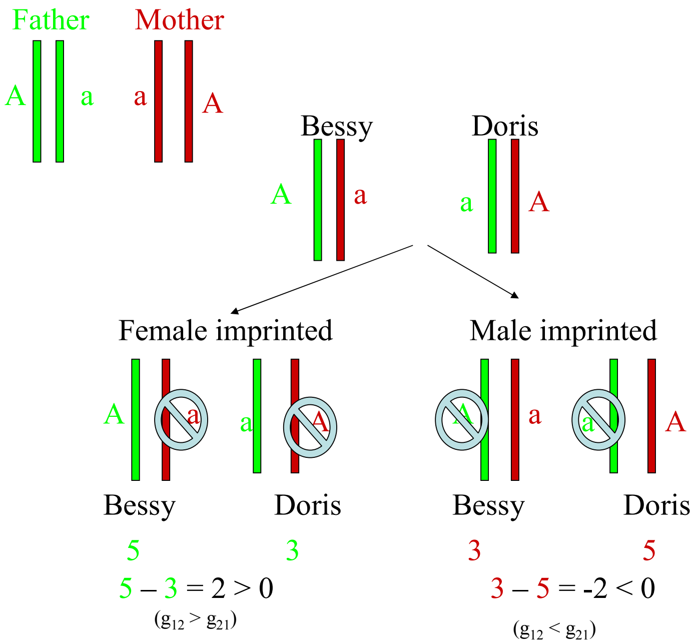


Figure 6.1. Imprinting effect of a quantitative trait is detected through the comparison of heterozygous individuals with different parental allele origins. The contribution of alleles to genotypic value is assumed to be: A=5, a=3. Bessy has genotype *Aa* with genotypic value of g_{12}, and Doris has genotype *aA* with genotypic value of g_{21}, where the allele on the left is from the father and the allele on the right from the mother. If the locus is female imprinted (silenced in the chromosome coming from the mother), Bessy's genotypic value is larger than Doris', leading to a positive estimate of imprinting effect $i = ½(g_{12}–g_{21}) > 0$. If the locus is male imprinted, Doris has a larger genotypic value than Bessy, giving a negative estimate of imprinting effect $i = ½(g_{12}–g_{21}) < 0$. In the absence of imprinting, $g_{12} = g_{21}$ and $i = 0$.
(Source: N.R. London, 2004. Ph.D. Thesis, Department of Animal Science, University of Minnesota)

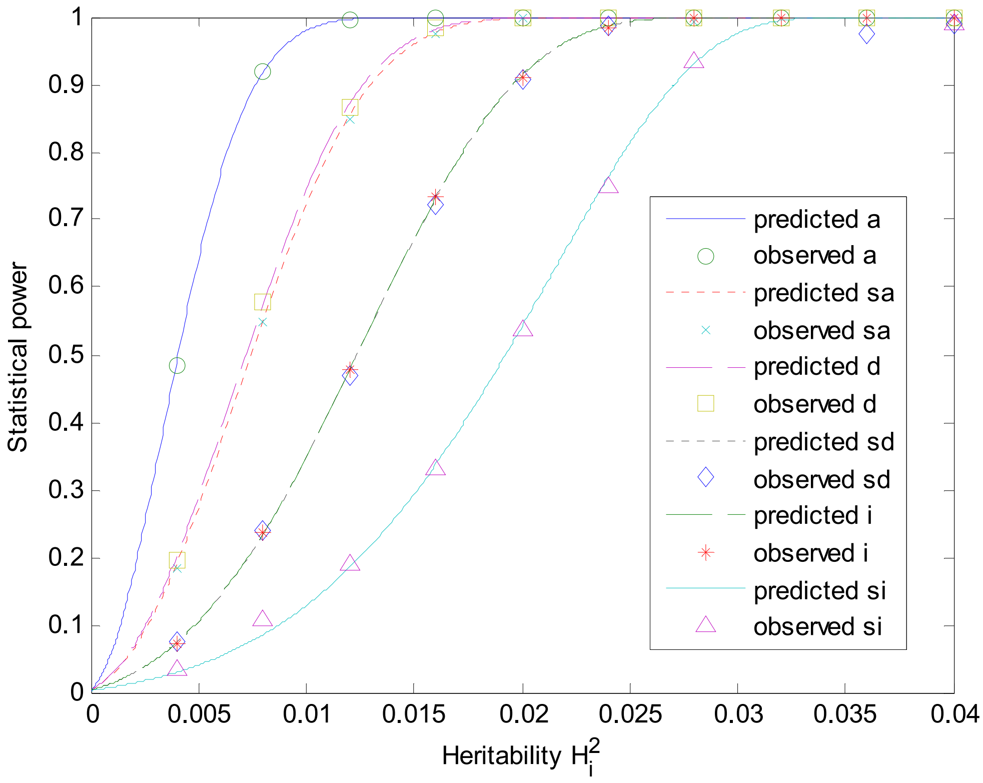


Figure 6.2. Observed (dotted lines) and predicted (solid lines) statistical power as a function of imprinting contrast heritability (H_i^2) using the new method for detecting single-locus effects and second order gene-sex epigenetic effects allowing Hardy-Weinberg and linkage disequilibria. ($\alpha = 0.0034$, $n = 500$, where $\alpha = 0.0034$ is the threshold significance level for 'suggestive linkage', Lander and Kruglyak, 1995). Gene-environment effects have the same statistical power as gene-gene-sex effects for binary environment factors.

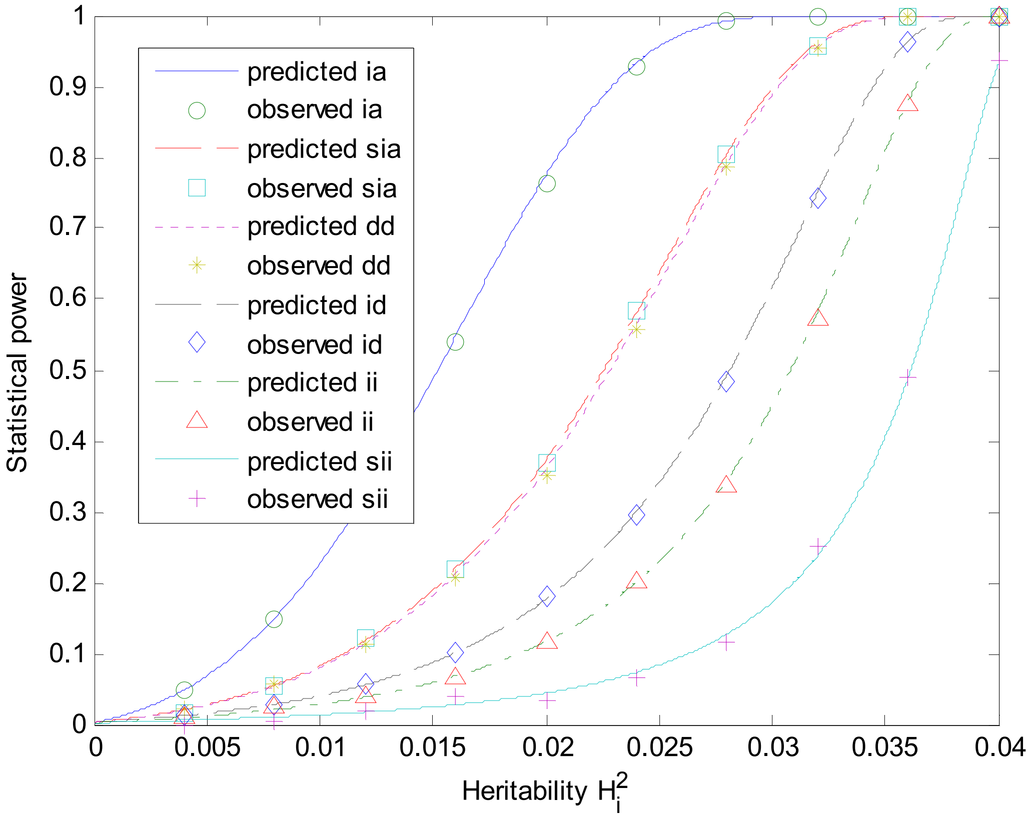


Figure 6.3. Observed (dotted lines) and predicted (solid lines) statistical power as a function of imprinting contrast heritability (H_i^2) using the new general method for detecting second order gene-gene and third order gene-gene-sex epigenetic effects allowing Hardy-Weinberg and linkage disequilibria. ($\alpha = 0.0034$, $n = 500$). Gene-gene-environment effects have the same statistical power as gene-gene-sex effects for binary environment factors.

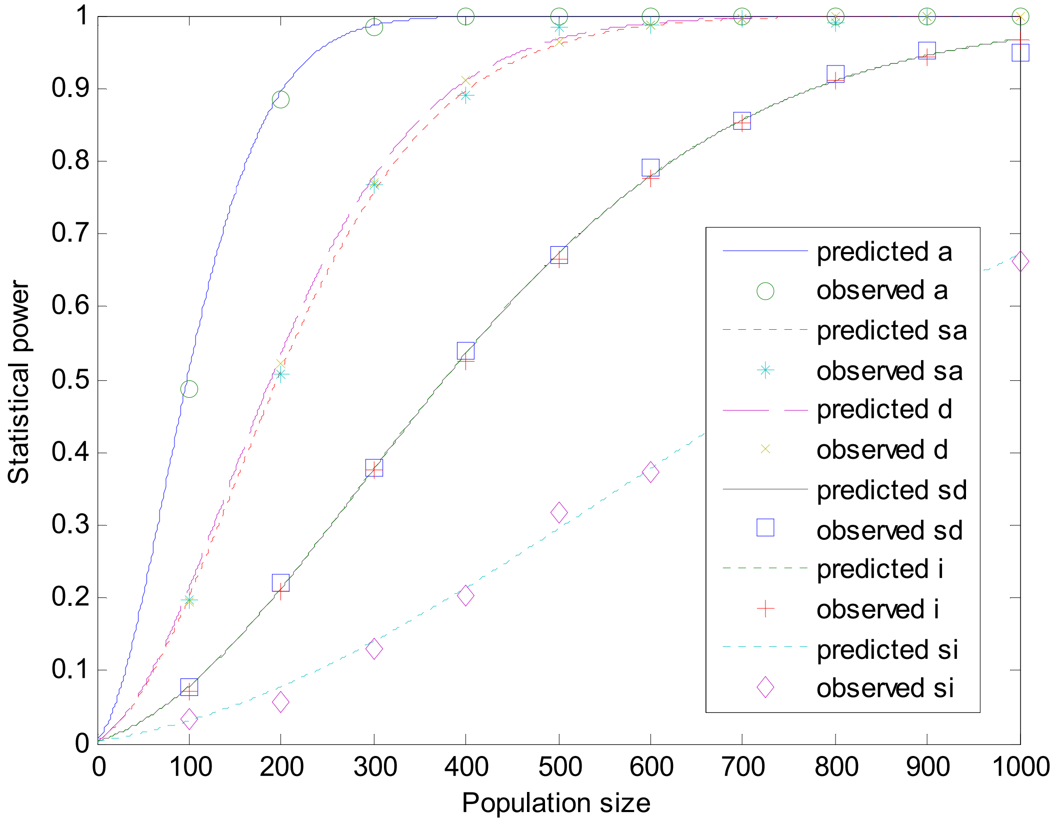


Figure 6.4. Observed (dotted lines) and predicted (solid lines) statistical power as a function of the population size using the new method for detecting epistasis effects allowing Hardy-Weinberg and linkage disequilibria. ($\alpha = 0.0034$, $H_i^2 = 0.015$).

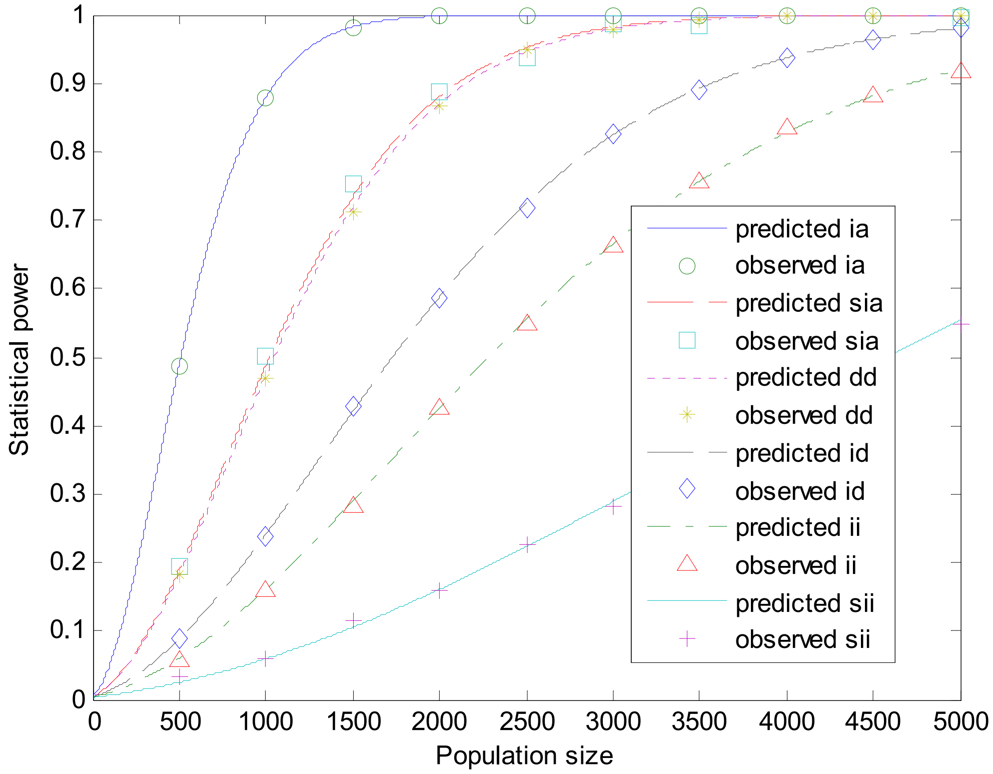


Figure 6.5. Observed (dotted lines) and predicted (solid lines) statistical power as a function of the population size using the new method for detecting epistasis effects allowing Hardy-Weinberg and linkage disequilibria. ($\alpha = 0.0034$, $H_i^2 = 0.015$).

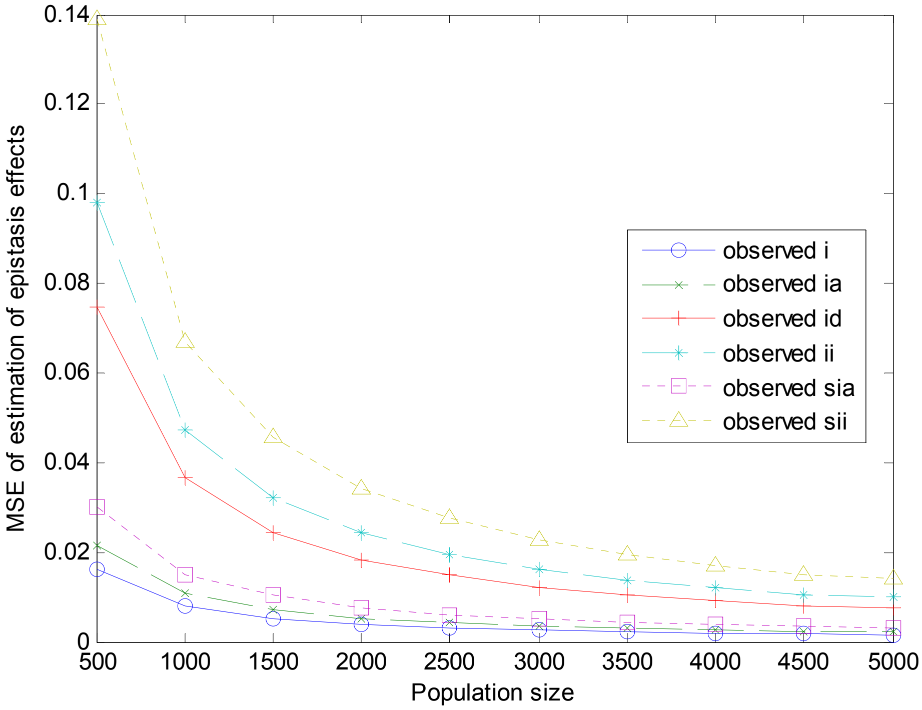


Figure 6.6. Mean square error (MSE) of the epistasis estimations as a function of the population size for the same data in Figures 1 and 2. ($\alpha = 0.0034$, $H_i^2 = 0.02$).

Chapter 7

Multi-locus Test Conditional on Confirmed Effects Leads to Increased Power in Genome-wide Association Analysis

7.1 Introduction

Complex diseases or phenotypes may involve multiple genetic variants and interactions between genetic, environmental and other factors. Current genome-wide association studies (GWAS) mostly used single-locus analysis and had identified genetic effects with multiple confirmations. Such confirmed SNP effects were likely to be true genetic effects and ignoring this information in testing new effects of the same phenotype results in decreased statistical power due to increased residual variance that has a component of the omitted effects. In this study, a multi-locus test (MLT) conditional on SNPs with confirmed effects was proposed for GWAS analysis to improve statistical power. Analytical formulae for statistical power were derived and were verified by simulation for MLT accounting for confirmed SNPs and for single-locus test (SLT) without accounting for confirmed SNPs. Statistical power of the two methods was compared using case studies and the Framingham Heart Study (FHS) GWAS data. In the GWAS case studies on four cholesterol phenotypes and serum metabolites, the MLT method improved statistical power by 5% to 38% over SLT depending on the number and effect sizes of the conditional SNPs. For the analysis of HDL cholesterol (HDL-C) and total cholesterol (TC) of the FHS data, the MLT method conditional on confirmed SNPs from GWAS catalog and NCBI improved statistical significance over SLT.

Genome-wide association studies (GWAS) have identified genetic variants associated with a number of complex diseases or traits [Hindorff et al., 2009; Illig et al., 2010; Sandhu et al., 2008] and some of these variants had confirmations from several studies [Hindorff et al., 2009]. Published GWAS studies typically used a single-locus test (SLT), in which each variant was tested individually for association with a specific phenotype. Single-locus analysis may not be the best approach in the presence of confirmed SNP effects, because confirmed effects become a component of the random residuals and decrease statistical power for detecting new effects if those true effects are omitted in the analysis. In this study, we propose a multi-locus test (MLT) conditional on confirmed effects for GWAS analysis to increase the statistical power for detecting new SNP effects, and we demonstrated the MLT method had increased statistical power relative to SLT using analytical formulae derived in this study and using simulation, case studies, and the Framingham Heart Study (FHS) data.

7.2 Predicted Statistical Power of MLT and SLT

The multiple linear regression model for the MLT method can be expressed as:

$$Y_i = \mu + \boldsymbol{Z}_i \boldsymbol{\beta}_Z + \boldsymbol{G}_i \boldsymbol{\beta}_G + G_{i,s}\beta_s + e_i \qquad (1)$$

where μ = the population mean of the phenotypic values, $\boldsymbol{Z}_i$ = 1 × p vector of the p covariates for subject i (i = 1, …, N) to account for environment, population stratification, and other factors; $\boldsymbol{\beta}_Z$ = p × 1 vector of the partial regression coefficients of the covariates

$\boldsymbol{Z}_i$; $\boldsymbol{G}_i = (G_{i,1}, G_{i,2}, \ldots, G_{i,s\text{-}1}) = 1 \times (S-1)$ vector of the $S-1$ SNP genotypes for subject i that were confirmed to be associated with the phenotype, with $G_{i,j}$ taking values of 0, 1 or 2 according to the number of copies of the minor allele for subject i at SNP j; $\boldsymbol{\beta}_G = (S-1) \times 1$ vector of partial regression coefficients of $S-1$ conditional SNPs; $G_{i,s}$ = the genotype value of the candidate SNP; βs = the partial regression coefficient of the candidate SNP, and e_i = random residual that is assumed follow $N(0, \sigma^2)$ normal distribution. The residual variance of Equation 1 is

$$\sigma_{e_i}^2 = Var\left(y_i \middle| G_{i,s}, \boldsymbol{G}_i, \boldsymbol{Z}_i\right) = Var\left(e_i\right) = \sigma^2 \tag{2}$$

A standard t-test can be used for testing the significance of the candidate SNP based on testing the following hypotheses, H_0: $\beta_s = 0$, where β_s is the regression coefficient of the candidate SNP and is the Mth element of $\boldsymbol{\beta} = (\mu, \boldsymbol{\beta}_Z, \boldsymbol{\beta}_G, \beta_s)$ and $M = 1 + p + S$. Statistical power for one-sided t-test was derived in this study for simplicity (statistical power for two-sided t-test can be obtained similarly but is not considered here). Using a multiple linear regression framework, the power of the one-sided t-test can be formulated as:

$$P_I = \int_{t_\alpha}^{+\infty} f\left(t, \lambda, N-M\right) dt \tag{3}$$

where t_α denotes the ordinate of the t-distribution with $N-M$ degrees of freedom corresponding to α and f denotes a non-central t-distribution with $N-M$ degrees of freedom and non-centrality parameter:

$$\lambda = \frac{\hat{\beta}_s}{\sqrt{\operatorname{var}\left(\hat{\beta}_s\right)}} = \frac{\hat{\beta}_s}{\sqrt{\hat{\boldsymbol{\Sigma}}_{MM}}} \tag{4}$$

where $\hat{\boldsymbol{\Sigma}}_{MM}$ = the element at the *M*th row and *M*th column of variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{\Sigma}}$ is a $M \times M$ variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and can be estimated as:

$$\hat{\boldsymbol{\Sigma}} = Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X'X}\right)^{-1}\sigma^2 \tag{5}$$

where $\boldsymbol{X}$ is the design matrix in Equation 1.

For SLT, the statistical model is the same as Equation 1 except that the residual term is now a summation of confirmed effects and random residuals. The residual variance for the SLT model is no longer σ^2 and has the following mathematical expression:

$$\sigma_{\varepsilon_i}{}^2 = Var\left(y_i \middle| \boldsymbol{Z}_i, G_{i,s}\right) = Var\left(\boldsymbol{G}_i\boldsymbol{\beta_G} + e_i\right) = \boldsymbol{\beta}_{\boldsymbol{G}}^T Var\left(\boldsymbol{G}_i\right)\boldsymbol{\beta_G} + \sigma^2 \tag{6}$$

where $Var\left(\boldsymbol{G}_i\right)$ is calculated based on the $\boldsymbol{G}_i$ values defined in Equation 1. Equation 6 shows that the residual variance of the SLT model is inflated over the MLT model of Equation 1 due to the fact that confirmed SNP effects are now in the residual term of Equation 6. Using this inflated estimated residual variance, the variance of $\hat{\beta}_s$ can be estimated by:

$$Var\left(\hat{\beta}_s\right) = c\sigma_{\varepsilon_i}^2 \tag{7}$$

where *c* is the element at the (*p* + 2)th row and (*p* + 2)th column of matrix $\left(\boldsymbol{\bar{X}'\bar{X}}\right)^{-1}$ and $\boldsymbol{\bar{X}}$ is the design matrix for the regression model of SLT. Therefore, the t-test statistic for SLT does not have a t-distribution but a t-distribution divided by a constant, $\tau = \sqrt{\dfrac{c\sigma_{\varepsilon_i}^2}{\hat{\Sigma}_{MM}}}$.

Similar to Equations 3, the power of the one-sided t-test can be formulated as:

$$P_{II} = \int_{\tau t_\alpha}^{+\infty} f\left(t, \lambda, N - p - 2\right) dt \tag{8}$$

where t_α denotes the ordinate of the t-distribution with $N-p-2$ degrees of freedom corresponding to α and f denotes a non-central t-distribution with $N-p-2$ degrees of freedom and the same non-centrality parameter λ as that in Equation 4.

7.3 Evaluation of Statistical Power of MLT and SLT Using Simulation and Real Data

The analytical formulae for statistical power for MLT accounting for confirmed effects and for SLT without accounting for confirmed effects (Equation 3 and Equation 8) were validated by simulated data of 2000 subjects for various effect sizes of the candidate SNP and confirmed SNPs with 10,000 repeats. The phenotypic values were simulated by the summation of a population mean, three additive SNP effects and a random error which followed a standard normal distribution. The three SNPs were simulated under Hardy-Weinberg equilibrium and linkage equilibrium with allele frequencies, 0.3, 0.4 and 0.2. The first two SNPs were assumed to have confirmed effects and the last SNP was assumed to be the candidate SNP. The candidate SNP was tested by the MLT and SLT methods in each simulation. Empirical power was calculated as the proportion of significant results from all 10,000 simulation results. We fixed the effects of the two confirmed SNPs as 0.3 and 0.2 standard deviation of residuals (SD) and varied the effect of the candidate SNP from 0.04 to 0.2 SD. Simulated statistical power were nearly identical to the predicted power for MLT and SLT (results not shown). With this knowledge of the power formulae being correct, predicted statistical power for MLT and SLT were calculated for various effect sizes of the confirmed SNPs (Table 7.1), showing that MLT results in higher power over SLT as the effect sizes of confirmed SNPs increase.

We further evaluated predicted statistical power using reported effect sizes for some confirmed SNP effects. We collected all reported SNP effects for HDL cholesterol (HDL-C), LDL cholesterol (LDL-C), triglycerides (TG), total cholesterol (TC), and serum metabolites (SM) from the GWAS catalog [Hindorff et al., 2009]. The effect sizes and risk allele frequencies of those SNPs were extracted and utilized for the power calculations. After filtering out SNPs in high linkage disequilibrium (LD) by only keeping one SNP with the largest effect size in each high LD region, the final selection of confirmed SNP markers included 22 relatively independent SNPs with effect sizes of 0.07-0.24 SD for HDL-C, 24 SNPs with effect sizes of 0.07-0.35 SD for LDL-C, 13 SNPs with effect sizes of 0.06-0.42 SD for TG, and 19 SNPs with effect sizes of 0.06-0.24 SD for TC. For SM, we extracted five SNPs which explained 5.6 to 36.3 percent of the total phenotypic variation [Hindorff et al., 2009; Illig et al., 2010]. Conditional on those known SNP effects, statistical power of MLT increased over that of SLT by about 4-5% for HDL-C, LDL-C, TG and TC. The patterns of the heatmaps for statistical power improvement were similar for these four traits and only the heatmap for HDL-C was shown in Figure 7.1A. Largest improvements appeared in those regions where the candidate SNP had small effect size and large allele frequency or the candidate SNP had medium effect size and small allele frequency (0.1-0.2). The increase in statistical power of MLT over SLT was much larger for SM, varying from 10% to 30%, because of the larger effect sizes of the conditional SNPs (Figure 7.1B).

For GWAS analysis using real data, true statistical power is not observable but MLT is expected to have more significant results than SLT. To compare observed statistical significance of MLT and SLT, we used the FHS GWAS data (version 2) that had 6575 individuals with SNP genotypes of the 500k SNP panel from dbGAP [Mailman et al., 2007]. Of the 6575 individuals, 6078 individuals had observations on HDL-C and 6431 individuals had observations on TC. From the 500k SNP panel, 432,096 SNP markers with known locations and minor allele frequencies 0.01 or greater were selected and tested. The original cholesterol measures deviated from normality and had outliers. The Box-Cox transformation analysis [Box and Cox, 1964] implemented by the R package [R Development Core Team, 2008] showed that the log-transformation was approximately the best transformation to achieve normality for HDL-C and TC. Age, age-squared, cholesterol treatment, blood sugar, body mass index, smoking status, number of cigars smoked, alcohol consumption and sex were adjusted for transformed HDL-C. Blood sugar, body mass index, smoking status, and sex were adjusted for transformed TC. The testing of SNP effects used the generalized least squares version [Ma et al., 2008a] of epiSNP [Ma et al., 2008b]. From the GWAS catalog [Hindorff et al., 2009] and NCBI (http://www.ncbi.nlm.nih.gov), we selected six SNP markers (Table 7.2) with multiple confirmations. These six SNP markers were independent of each other because pairwise correlations measured by R-squared among these SNPs were nearly zero. Results showed that MLT had more significant results than SLT for both HDL-C and TC (Table 7.2). The first two markers in Table 7.2 had the largest improvement in observed significance (reduced *P*-value), while the remaining markers only had minor improvement. In our analysis, we did not impute genotypic data to fill in missing genotypes so that the MLT test had smaller sample size than SLT. The observed significance should have been larger than observed in Table 7.2 if the missing genotypes were filled by imputing genotypes using software such as MACH [Li and Abecasis, 2006] and BEAGLE [Browning and Browning, 2007]. Although improvement in statistical significance is small in some cases, such improvement could be easily achieved without additional cost using GWAS analysis software such as PLINK [Purcell et al., 2007] and epiSNP [Ma et al., 2008b] that provide the option to incorporate covariates.

Table 7.1. Power comparison between MLT (Power I) and SLT (Power II) with various conditional SNP effect sizes and constant effect size of 0.1 SD for the candidate SNP.

Effect Sizes of Confirmed SNPs	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Power I	0.713	0.713	0.713	0.713	0.713	0.713	0.713	0.713	0.713	0.713
Power II	0.710	0.701	0.686	0.665	0.638	0.606	0.568	0.525	0.479	0.429
Improvement	0.003	0.012	0.027	0.048	0.075	0.108	0.146	0.188	0.235	0.284

Table 7.2. Comparison of *P*-values of MLT and SLT using the Framingham Heart Study data.

SNP	*P*-value		Phenotype	Conditional SNPs
	SLT	MLT		
rs1800775	7.06×10^{-34}	4.85×10^{-35}	HDL-C	rs765547, rs6507945, rs17259942
rs765547	2.19×10^{-12}	9.53×10^{-13}	HDL-C	rs1800775, rs6507945, rs17259942
rs6507945	9.77×10^{-10}	9.51×10^{-10}	HDL-C	rs1800775, rs765547, rs17259942
rs17259942	2.77×10^{-09}	1.52×10^{-09}	HDL-C	rs1800775, rs765547, rs6507945
rs599839	1.23×10^{-15}	9.77×10^{-16}	TC	rs4245791
rs4245791	3.32×10^{-09}	3.21×10^{-09}	TC	rs599839

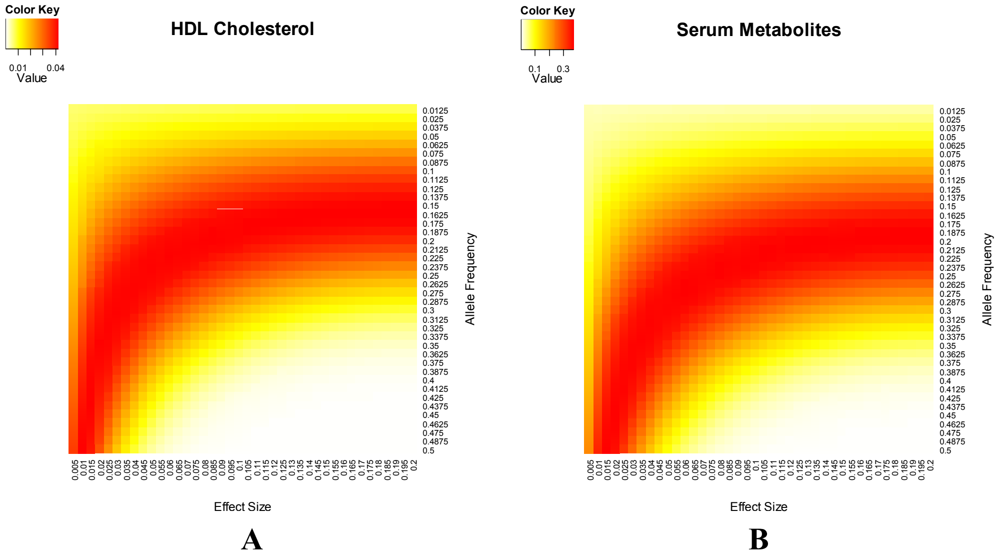


Figure 7.1. Power improvement of the MLT method conditional on confirmed SNP effects over the SLT method. **a**, HDL cholesterol. **b**, Serum metabolities. (The intensity of red color denotes the power difference between MLT and CLT)

Chapter 8

References

1. Alabdulkareema M, Lakshmivarahan S, Dhallb SK (2001) Scalability analysis of large codes using factorial designs. Parallel Computing, 27: 1145-1171
2. Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, Torsten Hothorn. (2010) mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9.
3. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, Penninx BWJH, Janssens ACJW, Wilson JF, Spector T, Martin NG, Pedersen NL, Kyvik KO, Kaprio J, Hofman A, Freimer NB, Jarvelin MR, Gyllensten U, Campbell H, Rudan I, Johansson A, Marroni F, Hayward C, Vitart V, Jonasson I, Pattaro C, Wright A, Hastie N, Pichler I, Hicks AA, Falchi M, Willemsen G, Hottenga JJ, De Geus EJC, Montgomery GW, Whitfield J, Magnusson P, Saharinen J, Perola M, Silander K, Isaacs A, Sijbrands EJG, Uitterlinden AG, Witteman JCM, Oostra BA, Elliott P, Ruokonen A, Sabatti C, Gieger C, Meitinger T, Kronenberg F, Döring A, Wichmann HE, Smit JH, McCarthy MI, Duijn CM, Leena (2008) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet, 41:47 - 55.
4. Ashwell MS, and Van Tassell CP. (1999) Detection of putative loci affecting milk, health, and type traits in a US Holstein population using 70 microsatellite markers in a genome scan. J Dairy Sci 82:2497-2502.

5. Ashwell MS, Van Tassell CP, and Sonstegard TS. (2001) A genome scan to identify quantitative trait loci affecting economically important traits in a US Holstein population, J Dairy Sci 84:2535-2542.

6. Ashwell MS, Heyen DW, Sonstegard TS, Van Tassell CP, Da Y, VanRaden PM, Ron M, Weller JI and Lewin HA. (2004) Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. J Dairy Sci 87:468-475.

7. Ashwell MS, Heyen DW, Weller JI, Ron M, Sonstegard TS, Van Tassell CP, and Lewin HA. (2005) Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle. J Dairy Sci 88:4111-4119.

8. Baes C, Brand B, Mayer M, Kühn C, Liu Z, Reinhardt F, and Reinsch N. Refined (2009) positioning of a quantitative trait locus affecting somatic cell score on chromosome 18 in the German Holstein using linkage disequilibrium. J. Dairy Sci. 92: 4046-4054.

9. Balding, DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet, 7: 781-791.

10. Beyer A, Bandyopadhyay S & Ideker T. (2007) Integrating physical and genetic maps: from genomes to interaction networks. Nature Reviews Genetics 8, 699-710.

11. Bird A. (2007). Perceptions of epigenetics. Nature 447, 396-398.

12. Boichard D, Grohs C, Bourgeois F, Cerqueira F, Faugeras R, Neau A, Rupp R, Amigues Y, Boscher MY, and Leveziel H. (2003) Detection of genes influencing economic traits in three French dairy cattle breeds. Genet Sel Evol 35:77_101.

13. Bovine HapMap Consortium, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM, Lien S, Matukumalli LK, McEwan JC, Nazareth LV, Schnabel RD, Weinstock GM, Wheeler DA, Ajmone-Marsan P, Boettcher PJ, Caetano AR, Garcia JF, Hanotte O, Mariani P, Skow LC, Sonstegard TS, Williams JL, Diallo B, Hailemariam L, Martinez ML, Morris CA, Silva LO, Spelman RJ, Mulatu W, Zhao K, Abbey CA, Agaba M, Araujo FR, Bunch RJ, Burton J, Gorni C,Olivier H, Harrison BE, Luff B, Machado MA, Mwakaya J, Plastow G, Sim W, Smith T, Thomas MB, Valentini A, Williams P, Womack J, Woolliams JA, Liu Y, Qin X, Worley KC, Gao C, Jiang H, Moore SS, Ren Y, Song XZ, Bustamante CD, Hernandez RD, Muzny DM, Patil S, San Lucas A, Fu Q, Kent MP, Vega R, Matukumalli A, McWilliam S, Sclep G, Bryc K, Choi J, Gao H, Grefenstette JJ, Murdoch B, Stella A, Villa-Angulo R, Wright M, Aerts J, Jann O, Negrini R, Goddard ME, Hayes BJ, Bradley DG, Barbosa da Silva M, Lau LP, Liu GE, Lynn DJ, Panzitta F, and Dodds KG. (2009) Genome wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324(5926):528-532.

14. Box, GEP, Cox, DR (1964) An analysis of transformations (with discussion). J. Roy. Stat. Soc B, 26, 211–252.

15. Bjornsson HT, Fallin MD & Feinberg AP. (2004) An integrated epigenetic and genetic approach to common human disease. Trends Genet. 20, 350–358.

16. Bray, D. (2003) Molecular Networks: The Top-Down View. Science 301, 1864-1865.

17. Browning BL (2008) PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation and one and two-stage genetic associations studies. BMC Bioinformatics 9: 309

18. Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097.

19. Carlborg O & Haley CS. (2004) Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. 5: 618-625.

20. Cheverud JM (2001) A simple correction for multiple comparisions in interval mapping genome scans. Heredity 87:52-58.

21. Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R (2004) Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood 103(7):2771-8

22. Chiaretti S, Li X, Gentleman R, Vitale A, Wang KS, Mandelli F, Foà R, Ritz J (2005) Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. Clin Cancer Res 11(20):7209-19

23. Cockerham CC. (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 859-882, 1954.

24. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, and Hobbs HH: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 2004, 305:869–872.

25. Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD, Taylor JF, and Wiggans GR. (2009) Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92: 2931-2946.

26. Cole, J.B., G.R. Wiggans, L. Ma, T.S. Sonstegard, B.A. Crooker, C. P. Van Tassell, J. Yang, L. K. Matukumalli, and Y. Da. High resolution QTL maps of 31 traits in contemporary U.S. Holstein cows. (2010a) Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, Aug. 1–6. (Accepted)

27. Cole, J.B., T.S. Sonstegard, L. Ma, G. R. Wiggans, B.A. Crooker, C. P. Van Tassell, J. Yang, L. K. Matukumalli, Y. Da. (2010b) High resolution QTL map of net merit component traits and calving traits from genome-wide association analysis in contemporary U.S. Holstein cows. Abstract and poster #P565 for Plant and Animal Genome XVIII, January 9-13. San Diego.

28. Crowe RR & Smouse PE. (1977). The genetic implications of age-dependent penetrance in manic-depressive illness. Journal of Psychiatric Research 13, 273-285.

29. Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999, 55, 997–1004.

30. Dudbridge F, Gusnanto A: Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 2008, 32: 227–234.

31. Eager DL, Zahorjan J, Lazowska ED: Speedup versus efficiency in parallel systems. Trans. On Competers. 1989, C-38: 408-423.

32. Eccleston A, DeWitt N, Gunter C, Marte1 B, Nath D. (2007) Epigenetics. Nature 447, 395.

33. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Leena Peltonen L, Järvelä I: Identification of a variant associated with adult-type hypolactasia Nat Genet 2002, 30, 233 – 237.

34. Esteller M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. Nature Reviews Genetics 8, 286-298.

35. Farrer MJ. (2006) Genetics of Parkinson disease: paradigm shifts and future prospects. Nature Reviews Genetics 7, 306-318.

36. Feinberg AP (2007) Phenotypic plasticity and the epigenetics of human disease. Nature 447, 433-440.

37. Fisher RA. (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinburgh 52: 399-433.

38. Ronald A. Fisher (1925). Statistical Methods for Research Workers. Oliver and Boyd.

39. Freudenberg-Hua Y, Freudenberg J, Kluck N, Cichon S, Propping P & Nothen MM. (2003) Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. Genome Res. 13: 2271-2276.

40. Friedewald WT, Levy RI, Fredrickson DS: Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 1972, 18:499-502.

41. Gardner TS., Bernardo DD, Lorenz D & Collins JJ. (2003) Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling Science 301, 102-105.

42. Genetic Association Information Network (GAIN) [http://www.fnih.org/GAIN2/platforms.shtml]

43. Goertz I, Baes C, Weimann CC, Reinsch N, and Erhardt G. (2009) Association between single nucleotide polymorphisms in the CXCR1 gene and somatic cell score in Holstein dairy cattle. J. Dairy Sci. 92: 4018-4022.

44. Henderson, C. R. 1986. Recent developments in variance and covariance estimation. J. Anim. Sci. 63:208.

45. Henderson, C. R. 1988. Theoretical basis and computational methods for a number of different animal models. J. Dairy Sci. 71(Suppl. 2):216.

46. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-

wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106: 9362-9367.

47. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmüller G, Kato BS, Mewes HW, Meitinger T, de Angelis MH, Kronenberg F, Soranzo N, Wichmann HE, Spector TD, Adamski J, Suhre K. 2010. A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42(2):137-41.

48. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet. 2001, 17: 388–391

49. Johannes F, Colot V & Jansen RC. (2007) Opinion: Epigenome dynamics: a quantitative genetics perspective Nature Reviews Genetics 9, 883-890.

50. Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, Gianniny L, Burtt NP, Melander O, Orho-Melander M, Arnett DK, Peloso GM, Ordovas JM, Cupples LA: A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Med Genet 2007, 19:8 Suppl 1:S17.

51. Karvanen J, Silander K, Kee F, Tiret L, Salomaa V, Kuulasmaa K, Wiklund PG, Virtamo J, Saarela O, Perret C, Perola M, Peltonen L, Cambien F, Erdmann J, Samani NJ, Schunkert H and Evans A: The impact of newly identified loci on coronary heart disease, stroke and total mortality in the MORGAM prospective cohorts. Genet Epidemiol 2009, 33: 237-246.

52. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T, Voight BF, Bonnycastle LL, Jackson AU, Crawford G, Surti A, Guiducci C, Burtt NP, Parish S, Clarke R, Zelenika D, Kubalanza KA, Morken MA, Scott LJ, Stringham HM, Galan P, Swift AJ, Kuusisto J, Bergman RN, Sundvall J, Laakso M, Ferrucci L, Scheet P, Sanna S, Uda M, Yang Q, Lunetta KL, Dupuis J, De Bakker PIW, O'Donnell CJ, Chambers JC, Kooner JS, Hercberg S, Meneton P, Lakatta EG, Scuteri A, Schlessinger D, Tuomilehto J, Collins FS, Groop L, Altshuler D, Collins R, Lathrop GM, Melander O, Salomaa V, Peltonen L, Orho-Melander M, Ordovas JM, Boehnke M, Abecasis GR, Mohlke KL, Cupples LA: Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 2008a, 41:56- 65.

53. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, Wahlstrand B, Hedner T, Corella D, Tai ES, Ordovas JM, Berglund G, Vartiainen E, Jousilahti P, Hedblad B, Taskinen M-R, Newton-Cheh C, Salomaa V, Peltonen L, Groop L, Altshuler DM, Orho-Melander M: Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 2008b, 40:189 – 197.

54. Kempthorne O. (1954) The correlation between relatives in a random mating population. Proc. R. Soc. Lond B Biol. Sci. 143: 102-113.

55. Kenneth O. McGraw & S. P. Wong (1996). "Forming inferences about some intraclass correlation coefficients". Psychological Methods 1: 30–46.

56. de Koning, DJ, Bovenhuis, H, van Arendonk, JAM. (2002) On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. Genetics 161, 931-938.

57. Kühn C, Bennewitz J, Reinsch N, Xu N, Thomsen H, Looft C, Brockmann GA, Schwerin M, Weimann C, Hiendleder S, Erhardt G, Medjugorac I, F?rster M, Brenig B, Reinhardt F, Reents R, Russ I, Averdunk G, Bl¨¹mel J, and Kalm E. (2003) Quantitative trait loci mapping of functional traits in the German Holstein cattle population. J Dairy Sci 86:360-368.

58. Eccleston A, DeWitt N, Gunter C, Marte1 B, Nath D. (2007) Epigenetics. Nature 447, 395.

59. Lander ES, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 1995, 11: 241–247.

60. Li W, Reich J: A complete enumeration and classification of two-locus disease models. Hum. Hered. 2000, 50: 334-349

61. London NR. (2004) Statistical theory and methods for mapping gender-affected genes and quantitative trait loci. Ph.D. Thesis. Department of Animal Science and Program in Molecular Veterinary Bioscience, University of Minnesota.

62. Ma L, Dvorkin D, Garbe JR & Da Y. (2007) Genome-wide analysis of single-locus and epistasis SNP effects on anti-cyclic citrullinated peptide as a measure of rheumatoid arthritis. BMC Proceedings: 1(Suppl 1), S127.

63. Ma L, Runesha HB, Dvorkin D, Garbe JR & Da Y. (2008a) Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative

traits in genome-wide association studies. BMC Bioinformatics 9:315. Article URL: http://www.biomedcentral.com/1471-2105/9/315.

64. Ma L, Runesha HB & Da Y. (2008b) EPISNPmpi: A supercomputer parallel computing program for epistasis testing in genome-wide association studies. User manual version 2.0. Department of Animal Science and Supercomputing Institute, University of Minnesota.http://animalgene.umn.edu/episnpmpi/EPiSNPmpi_manual_2.0.pdf .

65. Ma L, Runesha HB, Dvorkin D, Garbe JR & Da Y. (2008c) epiSNP: A computer package of serial computing programs for epistasis testing in genome-wide association studies. User manual version 2.0. Department of Animal Science and Supercomputing Institute, University of Minnesota. http://animalgene.umn.edu/episnp/epiSNP_manual_2.0.pdf .

66. Ma L, Amos CI, Yang Da Y. (2008d) Accounting for correlations among individuals for testing SNP single-locus and epistasis effects in Genome-wide association analysis (Abstract). Plant & Animal Genomes XVI Conference, January 12-16, 2008, San Diego, CA. http://www.intl-pag.org/16/abstracts/PAG16 P11 903.html (Last accessed 6/22/09)

67. Ma, L., T. S. Sonstegard, J.B. Cole, G. R. Wiggans, B.A. Crooker, C.P. Van Tassell, J. Yang, L.K. Matukumalli, Y. Da. (2010) X chromosome SNPs were heavily involved in epistasis effects of net merit component traits in contemporary U.S. Holstein cows. Abstract and poster #P542 for Plant and Animal Genome XVIII, January 9-13. San Diego.

68. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1007-1181.

69. Mao Y. London NR, Ma L, Dvorkin D & Da Y. (2006) Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model. Physiol. Genomics 28: 46-52.

70. Mao Y & Da Y. (2005) Statistical power for detecting epistasis QTL effects under the F-2 Design. Gene. Sel. Evol. 37: 129-150.

71. Marchini J, Howie B, Myers S, McVean G, & Donnelly P. (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39 : 906-913

72. Moore JH. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum. Hered. 56: 73-82.

73. Moore JH & Williams SM. (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 27, 637–646.

74. Moskvina V, Schmidt K: On multiple-testing correction in genome-wide association studies. Genet Epidemiol 2008, 32:567–573.

75. Murray A, Cluett C, Bandinelli S, Corsi AM, Ferrucci L, Guralnik J, Singleton A, Frayling T, Melzer D: Common lipid-altering gene variants are associated with

therapeutic intervention thresholds of lipid levels in older people. Eur Heart J 2009, 30:1711-1719.

76. Nakayama K, Bayasgalan T, Yamanaka K, Kumada M, Gotoh T, Utsumi N, Yanagisawa Y, Okayama M, Kajii E, Ishibashi S, Iwamoto S, and The Jichi Community Genetics Team (JCOG): Large scale replication analysis of loci associated with lipid concentrations in a Japanese population. J Med Genet 2009, 46: 370-374.

77. National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov (Last accessed 6/22/09).

78. Nishihara E, Tsaih SW, Tsukahara C, Langley S, Sheehan S, DiPetrillo K, Kunita S, Yagami K, Churchill GA, Paigenn B, Sugiyama F: Quantitative trait loci associated with blood pressure of metabolic syndrome in the progeny of NZO/HILtJ × C3H/HeJ intercrosses. Mammalian Genome 2007, 18: 573-583

79. Ober C, Loisel DA, Gilad Y. (2008) Sex-specific genetic architecture of human disease. Nature Rev. Genet. 9, 911-922.

80. Okamoto I, Otte A, Allis C, Reinberg D, Heard E (2004). Epigenetic dynamics of imprinted X inactivation during early mouse development. Science 303, 644–649.

81. Osborn D. (1916) Inheritance of baldness. Various patterns due to heredity and sometimes present at birth—a sex-limited character-dominant in man—women not bald unless they inherit tendency from both parents. J. Hered. 7, 347-355.

82. Phillips PC. (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9, 855-867.

83. Purcell S, Sham PC: Epistasis in quantitative trait locus linkage analysis: interaction or main effect? Behav. Genet. 2004, 34: 143-152

84. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559-575.

85. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, 2008.URL http://www.R-project.org (Last accessed 6/22/09)

86. Rutsch F, Gailus S, Miousse IR, Suormala T, Sagné C, Toliat MR, Nurnberg G, Wittkampf T, Buers I, Sharifi A, Stucki M, Becker C, Baumgartner M, Robenek H, Marquardt T, Hohne W, Gasnier B, Rosenblatt DS, Fowler B, Nurnberg P: Identification of a putative lysosomal cobalamin exporter altered in the cblF defect of vitamin B(12) metabolism. Nat Genet 2009, 41, 234-239.

87. Rutte M. (2005) How the environment affects mental health. The British Journal of Psychiatry 186, 4-6.

88. Sambandan S, Yamamoto A, Fanara JJ, Mackay TFC, Anholt RRH: Dynamic genetic interactions determine odor-guided behavior in drosophila melanogaster. Genetics. 2006, 74: 1349–1363.

89. Samani NJ, Braund PS, Erdmann J, Gotz A, Tomaszewski M, Linsel-Nitschke P, Hajat C, Mangino M, Hengstenberg C, Stark K, Ziegler A, Caulfield M, Burton PR, Schunkert H and Tobin MD: The novel genetic variant predisposing to

coronary artery disease in the region of the PSRC1 and CELSR2 genes on chromosome 1 associates with serum cholesterol. J Mol Med 2008, 86: 1233-1241.

90. Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, Song K, Yuan X, Johnson T, Ashford S, Inouye M, Luben R, Sims M, Hadley D, McArdle W, Barter P, Kesäniemi YA, Mahley RW, McPherson R, Grundy SM; Wellcome Trust Case Control Consortium, Bingham SA, Khaw KT, Loos RJ, Waeber G, Barroso I, Strachan DP, Deloukas P, Vollenweider P, Wareham NJ, Mooser V: LDL-cholesterol concentrations: a genome-wide association study. Lancet 2008, 371: 483-491.
91. Sanjuán R & Elena SF. (2006) Epistasis correlates to genomic complexity. Proc Natl Acad Sci U S A. 103, 14402–14405.
92. Searle SR. (1982) Matrix algebra useful for statistics. John Wiley & Sons, New York.
93. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Thakurta DG, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ: An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. 2005 37: 710 – 717.
94. Scheet P & Stephens MA. (2006) Fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78, 629–644.

95. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH and Ulrich R: Mapping the genetic architecture of gene expression in human liver. PLoS Biol 2008, 6: e107.

96. Shrout, P.E. & Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, 2, 420-428.

97. Silander K, Alanne M, Kristiansson K, Saarela O, Ripatti S, Auro K, Karvanen J, Kulathinal S, Niemelä M, Ellonen P, Vartiainen E, Jousilahti P, Saarela J, Kuulasmaa K, Evans A, Perola M, Salomaa V, Peltonen L: Gender Differences in Genetic Risk Profiles for Cardiovascular Disease. PLoS ONE 2008, 3, e3615.

98. Sonstegard, T. S., L. Ma, C.P. Van Tassell, E-S. Kim, J.B. Cole, G.R. Wiggans, B.A. Crooker, B.D. Mariani , L.K. Matukumalli, J.R. Garbe, S.C. Fahrenkrug, G. Liu, and Y. Da. (2010) Forty years of artificial selection in U.S. Holstein cattle had genome-wide signatures. Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, Aug. 1–6. (Accepted)

99. Tanaka T, Shen J, Abecasis GR, Kisialiou A, Ordovas JM, Guralnik JM, Andrew Singleton A, Bandinelli S, Cherubini A, Arnett D, Tsai MY, Ferrucci L: Genome-Wide Association Study of Plasma Polyunsaturated Fatty Acids in the InCHIANTI Study. PLoS Genet. 2009, 5, e1000338.

100. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H & Boone C. (2004) Global mapping of the yeast genetic interaction network. Science 303: 808-813.

101. Van Speybroeck, L. (2002) From epigenesis to epigenetics: the case of C. H. Waddington. Ann. NY Acad. Sci. 981, 61–81.

102. Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, Ahmadi K, Dobson RJ, Marcano AC, Hajat C, Burton P, Deloukas P, Brown M, Connell JM, Dominiczak A, Lathrop GM, Webster J, Farrall M, Spector T, Samani NJ, Caulfield MJ, Munroe PB: Genome-wide association study identifies genes for biomarkers of cardiovascular disease: Serum urate and dyslipidemia. Am J Hum Genet 2008, 82:139-149.

103. Wang J, Burnett JR, Near S, Young K, Zinman B, Hanley AJ, Connelly PW, Harris SB, Hegele RA: Common and rare ABCA1 variants affecting plasma HDL cholesterol. Arterioscler Thromb Vasc Biol. 2000; 20:1983–1989.

104. Weir BS & Cockerham CC. (1989) Complete characterization of disequilibrium at two loci. Pg. 86-110 in Mathematical Evolutionary Theory. M. E. Feldman (Editor). Princeton University Press, Princeton.

105. Wentworth EN (1916) A sex-limited color in Ayshire cattle J. Agric. Res. 6, 141-147.

106. Wiggans GR, Sonstegard TS, VanRaden PM, Matukumalli MK, Schnabel RD, Taylor JF, F. Schenkel S, and Van Tassell CP. (2009) Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92: 3431-3436.

107. Wiggans, G.R. , L. Ma , T. S. Sonstegard, J.B. Cole, B.A. Crooker, C.P. Van Tassell, J. Yang , L. K. Matukumalli, Y. Da. (2010) High resolution QTL map of body conformation traits from genome-wide association analysis in contemporary U.S. Holstein cows. Abstract and poster #P547 for Plant and Animal Genome XVIII, January 9-13. San Diego.

108. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR: Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 2008, 40: 161-169.

109. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. 2001, 8: 625-37

110. Wostmann BS, Bruckner-Kardoss E: The effect of long-term feeding of 10% dietary lactose on serum, liver and aortic cholesterol of the rat and the gerbil. J Nutr 1980, 110, 82-89.

111. Yan D, Mäyränpää MI, Wong J, Perttilä J, Lehto M, Jauhiainen M, Kovanen PT, Ehnholm C, Brown AJ, Olkkonen VM: OSBP-related protein 8 (ORP8) suppresses ABCA1 expression and cholesterol efflux from macrophages. J Biol Chem 2007, 283:332–340.

112. Yarden RI, Friedman E, Metsuyanim S, Olender T, Ben-Asher E, Papa MZ. (2008) J. Breast Cancer Research and Treatment 111:497-504.

Appendix

A. MARGINAL AND CONDITIONAL PROBABILITIES

Marginal probabilities

Genotype × allele probabilities:
$p_{AAB} = p_{AABB} + \frac{1}{2}p_{AABb} + \frac{1}{2}p_{AAbB}$, $p_{AaB} = p_{AaBB} + \frac{1}{2}p_{AaBb} + \frac{1}{2}p_{AabB}$,
$p_{aAB} = p_{aABB} + \frac{1}{2}p_{aABb} + \frac{1}{2}p_{aAbB}$, $p_{aaB} = p_{aaBB} + \frac{1}{2}p_{aaBb} + \frac{1}{2}p_{aabB}$,
$p_{AAb} = p_{AAbb} + \frac{1}{2}p_{AABb} + \frac{1}{2}p_{AAbB}$, $p_{Aab} = p_{Aabb} + \frac{1}{2}p_{AaBb} + \frac{1}{2}p_{AabB}$,
$p_{aAb} = p_{aAbb} + \frac{1}{2}p_{aABb} + \frac{1}{2}p_{aAbB}$, $p_{aab} = p_{aabb} + \frac{1}{2}p_{aaBb} + \frac{1}{2}p_{aabB}$,
$p_{ABB} = p_{AABB} + \frac{1}{2}p_{AaBB} + \frac{1}{2}p_{aABB}$, $p_{ABb} = p_{AABb} + \frac{1}{2}p_{AaBb} + \frac{1}{2}p_{aABb}$,
$p_{AbB} = p_{AAbB} + \frac{1}{2}p_{AabB} + \frac{1}{2}p_{aAbB}$, $p_{Abb} = p_{AAbb} + \frac{1}{2}p_{Aabb} + \frac{1}{2}p_{aAbb}$,
$p_{aBB} = p_{aaBB} + \frac{1}{2}p_{AaBB} + \frac{1}{2}p_{aABB}$, $p_{aBb} = p_{aaBb} + \frac{1}{2}p_{AaBb} + \frac{1}{2}p_{aABb}$,
$p_{abB} = p_{aabB} + \frac{1}{2}p_{AabB} + \frac{1}{2}p_{aAbB}$, $p_{abb} = p_{aabb} + \frac{1}{2}p_{Aabb} + \frac{1}{2}p_{aAbb}$

Genotype probabilities for each locus:
$p_{AA} = p_{AABB} + p_{AABb} + p_{AAbB} + p_{AAbb}$, $p_{Aa} = p_{AaBB} + p_{AaBb} + p_{AabB} + p_{Aabb}$,
$p_{aA} = p_{aABB} + p_{aABb} + p_{aAbB} + p_{aAbb}$, $p_{aa} = p_{aaBB} + p_{aaBb} + p_{aabB} + p_{aabb}$,
$p_{BB} = p_{AABB} + p_{AaBB} + p_{aABB} + p_{aaBB}$, $p_{Bb} = p_{AABb} + p_{AaBb} + p_{aABb} + p_{aaBb}$,
$p_{bB} = p_{AAbB} + p_{AabB} + p_{aAbB} + p_{aabB}$, $p_{bb} = p_{AAbb} + p_{Aabb} + p_{aAbb} + p_{aabb}$

Allelic probabilities:
$p_A = p_{AA} + \frac{1}{2}p_{Aa} + \frac{1}{2}p_{aA}$, $p_a = p_{aa} + \frac{1}{2}p_{Aa} + \frac{1}{2}p_{aA}$, $p_B = p_{BB} + \frac{1}{2}p_{Bb} + \frac{1}{2}p_{bB}$, $p_b = p_{bb} + \frac{1}{2}p_{Bb} + \frac{1}{2}p_{bB}$

Haplotype probabilities:

Haplotype probabilities p_{AB}, p_{Ab}, p_{aB}, and p_{ab} are obtained based on the formula in Weir (1996) with slight modifications,

$$p_{Ab}^{(i)} = p_A - p_{AB}^{(i)}, \quad p_{aB}^{(i)} = p_B - p_{AB}^{(i)}, \quad p_{ab}^{(i)} = 1 - p_A - p_B + p_{AB}^{(i)}$$

$$p_{AB}^{(i+1)} = p_{AABB} + \frac{1}{2}\left[p_{AABb} + p_{AAbB} + p_{AaBB} + p_{aABB} + \frac{p_{AB}^{(i)}p_{ab}^{(i)}}{p_{AB}^{(i)}p_{ab}^{(i)} + p_{Ab}^{(i)}p_{aB}^{(i)}}\left(p_{AaBb} + p_{AabB} + p_{aABb} + p_{aAbB} \right) \right]$$

Conditional probabilities

Genotype × genotype probabilities conditional on each allele:
$p_{AABB|A} = p_{AABB}/p_A$, $p_{AABb|A} = p_{AABb}/p_A$, $p_{AAbB|A} = p_{AAbB}/p_A$, $p_{AAbb|A} = p_{AAbb}/p_A$,
$p_{AaBB|A} = p_{AaBB}/(2p_A)$,$p_{aABB|A} = p_{aABB}/(2p_A)$, $p_{AaBb|A} = p_{AaBb}/(2p_A)$, $p_{aABb|A} = p_{aABb}/(2p_A)$
$p_{AabB|A} = p_{AabB}/(2p_A)$, $p_{aAbB|A} = p_{aAbB}/(2p_A)$, $p_{Aabb|A} = p_{Aabb}/(2p_A)$, $p_{aAbb|A} = p_{aAbb}/(2p_A)$
$p_{aaBB|a} = p_{aaBB}/p_a$, $p_{aaBb|a} = p_{aaBb}/p_a$, $p_{aabB|a} = p_{aabB}/p_a$, $p_{aabb|a} = p_{aabb}/p_a$,
$p_{AaBB|a} = p_{AaBB}/(2p_a)$,$p_{aABB|a} = p_{aABB}/(2p_a)$, $p_{AaBb|a} = p_{AaBb}/(2p_a)$, $p_{aABb|a} = p_{aABb}/(2p_a)$

$p_{AabB|a} = p_{AabB}/(2p_a)$, $p_{aAbB|a} = p_{aAbB}/(2p_a)$, $p_{Aabb|a} = p_{Aabb}/(2p_a)$, $p_{aAbb|a} = p_{aAbb}/(2p_a)$
$p_{AABB|B} = p_{AABB}/p_B$, $p_{AaBB|B} = p_{AaBB}/p_B$, $p_{aABB|B} = p_{aABB}/p_B$, $p_{aaBB|B} = p_{aaBB}/p_B$,
$p_{AABb|B} = p_{AABb}/(2p_B)$,$p_{aABb|B} = p_{aABb}/(2p_B)$, $p_{AaBb|B} = p_{AaBb}/(2p_B)$, $p_{aaBb|B} = p_{aaBb}/(2p_B)$
$p_{AAbB|B} = p_{AAbB}/(2p_B)$, $p_{AabB|B} = p_{AabB}/(2p_B)$, $p_{aAbB|B} = p_{aAbB}/(2p_B)$, $p_{aabB|B} = p_{aabB}/(2p_B)$
$p_{AAbb|b} = p_{AAbb}/p_b$, $p_{Aabb|b} = p_{Aabb}/p_b$, $p_{aAbb|b} = p_{aAbb}/p_b$, $p_{aabb|b} = p_{aabb}/p_b$,
$p_{AABb|b} = p_{AABb}/(2p_b)$,$p_{aABb|b} = p_{aABb}/(2p_b)$, $p_{AaBb|b} = p_{AaBb}/(2p_b)$, $p_{aaBb|b} = p_{aaBb}/(2p_b)$
$p_{AAbB|b} = p_{AAbB}/(2p_b)$, $p_{AabB|b} = p_{AabB}/(2p_b)$, $p_{aAbB|b} = p_{aAbB}/(2p_b)$, $p_{aabB|b} = p_{aabB}/(2p_b)$

Genotype × genotype probabilities conditional on each genotype of one locus:
$p_{AABB|AA} = p_{AABB}/p_{AA}$, $p_{AABb|AA} = p_{AABb}/p_{AA}$, $p_{AAbB|AA} = p_{AAbB}/p_{AA}$, $p_{AAbb|AA} = p_{AAbb}/p_{AA}$,
$p_{AaBB|Aa} = p_{AaBB}/p_{Aa}$, $p_{AaBb|Aa} = p_{AaBb}/p_{Aa}$, $p_{AabB|Aa} = p_{AabB}/p_{Aa}$, $p_{Aabb|Aa} = p_{Aabb}/p_{Aa}$,
$p_{aABB|aA} = p_{aABB}/p_{aA}$, $p_{aABb|aA} = p_{aABb}/p_{aA}$, $p_{aAbB|aA} = p_{aAbB}/p_{aA}$, $p_{aAbb|aA} = p_{aAbb}/p_{aA}$,
$p_{aaBB|aa} = p_{aaBB}/p_{aa}$, $p_{aaBb|aa} = p_{aaBb}/p_{aa}$, $p_{aabB|aa} = p_{aabB}/p_{aa}$, $p_{aabb|aa} = p_{aabb}/p_{aa}$,
$p_{AABB|BB} = p_{AABB}/p_{BB}$, $p_{AaBB|BB} = p_{AaBB}/p_{BB}$, $p_{aABB|BB} = p_{aABB}/p_{BB}$, $p_{aaBB|BB} = p_{aaBB}/p_{BB}$,
$p_{AABb|Bb} = p_{AABb}/p_{Bb}$, $p_{AaBb|Bb} = p_{AaBb}/p_{Bb}$, $p_{aABb|Bb} = p_{aABb}/p_{Bb}$, $p_{aaBb|Bb} = p_{aaBb}/p_{Bb}$,
$p_{AAbB|bB} = p_{AAbB}/p_{bB}$, $p_{AabB|bB} = p_{AabB}/p_{bB}$, $p_{aAbB|bB} = p_{aAbB}/p_{bB}$, $p_{aabB|bB} = p_{aabB}/p_{bB}$,
$p_{AAbb|bb} = p_{AAbb}/p_{bb}$, $p_{Aabb|bb} = p_{Aabb}/p_{bb}$, $p_{aAbb|bb} = p_{aAbb}/p_{bb}$, $p_{aabb|bb} = p_{aabb}/p_{bb}$

Genotype × genotype probabilities conditional on each pairing haplotypes:
$p_{AABB|AB} = p_{AABB}/p_{AB}$, $p_{AABb|AB} = p_{AABb}/(2p_{AB})$, $p_{AAbB|AB} = p_{AAbB}/(2p_{AB})$,
$p_{AaBB|AB} = p_{AaBB}/(2p_{AB})$, $p_{aABB|AB} = p_{aABB}/(2p_{AB})$, $p_{AaBb|AB} = qp_{AaBb}/(2p_{AB})$,
$p_{AabB|AB} = qp_{AabB}/(2p_{AB})$, $p_{aABb|AB} = qp_{aABb}/(2p_{AB})$, $p_{aAbB|AB} = qp_{aAbB}/(2p_{AB})$,
$p_{AAbb|Ab} = p_{AAbb}/p_{Ab}$, $p_{AABb|Ab} = p_{AABb}/(2p_{Ab})$, $p_{AAbB|Ab} = p_{AAbB}/(2p_{Ab})$,
$p_{Aabb|Ab} = p_{Aabb}/(2p_{Ab})$, $p_{aAbb|Ab} = p_{aAbb}/(2p_{Ab})$, $p_{AaBb|Ab} = (1-q)p_{AaBb}/(2p_{Ab})$,
$p_{AabB|Ab} = (1-q)p_{AabB}/(2p_{Ab})$, $p_{aABb|Ab} = (1-q)p_{aABb}/(2p_{Ab})$, $p_{aAbB|Ab} = (1-q)p_{aAbB}/(2p_{Ab})$,
$p_{aaBB|aB} = p_{aaBB}/p_{aB}$, $p_{aaBb|aB} = p_{aaBb}/(2p_{aB})$, $p_{aabB|aB} = p_{aabB}/(2p_{aB})$,
$p_{AaBB|aB} = p_{AaBB}/(2p_{aB})$, $p_{aABB|aB} = p_{aABB}/(2p_{aB})$, $p_{AaBb|aB} = (1-q)p_{AaBb}/(2p_{aB})$,
$p_{AabB|aB} = (1-q)p_{AabB}/(2p_{aB})$, $p_{aABb|aB} = (1-q)p_{aABb}/(2p_{aB})$, $p_{aAbB|aB} = (1-q)p_{aAbB}/(2p_{aB})$,
$p_{aabb|ab} = p_{aabb}/p_{ab}$, $p_{aaBb|ab} = p_{aaBb}/(2p_{ab})$, $p_{aabB|ab} = p_{aabB}/(2p_{ab})$,
$p_{Aabb|ab} = p_{Aabb}/(2p_{ab})$, $p_{aAbb|ab} = p_{aAbb}/(2p_{ab})$, $p_{AaBb|ab} = qp_{AaBb}/(2p_{ab})$,
$p_{AabB|ab} = qp_{AabB}/(2p_{ab})$, $p_{aABb|ab} = qp_{aABb}/(2p_{ab})$, $p_{aAbB|ab} = qp_{aAbB}/(2p_{ab})$

where $q = \dfrac{p_{AB}p_{ab}}{p_{AB}p_{ab} + p_{Ab}p_{aB}}$

Genotype × genotype probabilities conditional on each genotype allele combination:
$p_{AABB|ABB} = p_{AABB}/p_{ABB}$, $p_{AaBB|ABB} = p_{AaBB}/(2p_{ABB})$, $p_{aABB|ABB} = p_{aABB}/(2p_{ABB})$,
$p_{AABb|ABb} = p_{AABb}/p_{ABb}$, $p_{AaBb|ABb} = p_{AaBb}/(2p_{ABb})$, $p_{aABb|ABb} = p_{aABb}/(2p_{ABb})$,
$p_{AAbB|AbB} = p_{AAbB}/p_{AbB}$, $p_{AabB|AbB} = p_{AabB}/(2p_{AbB})$, $p_{aAbB|AbB} = p_{aAbB}/(2p_{AbB})$,
$p_{AAbb|Abb} = p_{AAbb}/p_{Abb}$, $p_{Aabb|Abb} = p_{Aabb}/(2p_{Abb})$, $p_{aAbb|Abb} = p_{aAbb}/(2p_{Abb})$,

$p_{aaBB|aBB} = p_{aaBB}/p_{aBB}$, $p_{AaBB|aBB} = p_{AaBB}/(2p_{aBB})$, $p_{aABB|aBB} = p_{aABB}/(2p_{aBB})$,
$p_{aaBb|aBb} = p_{aaBb}/p_{aBb}$, $p_{AaBb|aBb} = p_{AaBb}/(2p_{aBb})$, $p_{aABb|aBb} = p_{aABb}/(2p_{aBb})$,

$p_{aabB|abB} = p_{aabB}/p_{abB}$, $p_{AabB|abB} = p_{AabB}/(2p_{abB})$, $p_{aAbB|abB} = p_{aAbB}/(2p_{abB})$,
$p_{aabb|abb} = p_{aabb}/p_{abb}$, $p_{Aabb|abb} = p_{Aabb}/(2p_{abb})$, $p_{aAbb|abb} = p_{aAbb}/(2p_{abb})$,

$p_{AABB|AAB} = p_{AABB}/p_{AAB}$, $p_{AABb|AAB} = p_{AABb}/(2p_{AAB})$, $p_{AAbB|AAB} = p_{AAbB}/(2p_{AAB})$,
$p_{AaBB|AaB} = p_{AaBB}/p_{AaB}$, $p_{AaBb|AaB} = p_{AaBb}/(2p_{AaB})$, $p_{AabB|AaB} = p_{AabB}/(2p_{AaB})$,
$p_{aABB|aAB} = p_{aABB}/p_{aAB}$, $p_{aABb|aAB} = p_{aABb}/(2p_{aAB})$, $p_{aAbB|aAB} = p_{aAbB}/(2p_{aAB})$,
$p_{aaBB|aaB} = p_{aaBB}/p_{aaB}$, $p_{aaBb|aaB} = p_{aaBb}/(2p_{aaB})$, $p_{aabB|aaB} = p_{aabB}/(2p_{aaB})$,

$p_{AABB|AAB} = p_{AABB}/p_{AAB}$, $p_{AABb|AAB} = p_{AABb}/(2p_{AAB})$, $p_{AAbB|AAB} = p_{AAbB}/(2p_{AAB})$,
$p_{AaBB|AaB} = p_{AaBB}/p_{AaB}$, $p_{AaBb|AaB} = p_{AaBb}/(2p_{AaB})$, $p_{AabB|AaB} = p_{AabB}/(2p_{AaB})$,
$p_{aABB|aAB} = p_{aABB}/p_{aAB}$, $p_{aABb|aAB} = p_{aABb}/(2p_{aAB})$, $p_{aAbB|aAB} = p_{aAbB}/(2p_{aAB})$,
$p_{aaBB|aaB} = p_{aaBB}/p_{aaB}$, $p_{aaBb|aaB} = p_{aaBb}/(2p_{aaB})$, $p_{aabB|aaB} = p_{aabB}/(2p_{aaB})$

B. CALCULATION OF MEANS

Population mean

$\mu = (p_{AABB}, p_{AABb}, p_{AAbB}, p_{AAbb}, p_{AaBB}, p_{AaBb}, p_{AabB}, p_{Aabb}, p_{aABB}, p_{aABb}, p_{aAbB}, p_{aAbb}, p_{aaBB}, p_{aaBb}, p_{aabB}, p_{aabb})\mathbf{g} = \mathbf{s_1}'\mathbf{g}$ [B.1]

Allelic means:

$\mu_A = (p_{AABB|A}, p_{AABb|A}, p_{AAbB|A}, p_{AAbb|A}, p_{AaBB|A}, p_{AaBb|A}, p_{AabB|A}, p_{Aabb|A}, p_{aABB|A}, p_{aABb|A}, p_{aAbB|A}, p_{aAbb|A}, 0, 0, 0, 0)\mathbf{g}$
$\mu_a = (0, 0, 0, 0, p_{AaBB|a}, p_{AaBb|a}, p_{AabB|a}, p_{Aabb|a}, p_{aABB|a}, p_{aABb|a}, p_{aAbB|a}, p_{aAbb|a}, p_{aaBB|a}, p_{aaBb|a}, p_{aabB|a}, p_{aabb|a})\mathbf{g}$
$\mu_B = (p_{AABB|B}, p_{AABb|B}, p_{AAbB|B}, 0, p_{AaBB|B}, p_{AaBb|B}, p_{AabB|B}, 0, p_{aABB|B}, p_{aABb|B}, p_{aAbB|B}, 0, p_{aaBB|B}, p_{aaBb|B}, p_{aabB|B}, 0)\mathbf{g}$
$\mu_b = (0, p_{AABb|b}, p_{AAbB|b}, p_{AAbb|b}, 0, p_{AaBb|b}, p_{AabB|b}, p_{Aabb|b}, 0, p_{aABb|b}, p_{aAbB|b}, p_{aAbb|b}, 0, p_{aaBb|b}, p_{aabB|b}, p_{aabb|b})\mathbf{g}$

Single locus genotypic means

$\mu_{AA} = (p_{AABB|AA}, p_{AABb|AA}, p_{AAbB|AA}, p_{AAbb|AA}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{Aa} = (0, 0, 0, 0, p_{AaBB|Aa}, p_{AaBb|Aa}, p_{AabB|Aa}, p_{Aabb|Aa}, 0, 0, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aA} = (0, 0, 0, 0, 0, 0, 0, 0, p_{aABB|aA}, p_{aABb|aA}, p_{aAbB|aA}, p_{aAbb|aA}, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aa} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, p_{aaBB|aa}, p_{aaBb|aa}, p_{aabB|aa}, p_{aabb|aa})\mathbf{g}$
$\mu_{BB} = (p_{AABB|BB}, 0, 0, 0, p_{AaBB|BB}, 0, 0, 0, p_{aABB|BB}, 0, 0, 0, p_{aaBB|BB}, 0, 0, 0)\mathbf{g}$
$\mu_{Bb} = (0, p_{AABb|Bb}, 0, 0, 0, p_{AaBb|Bb}, 0, 0, 0, p_{aABb|Bb}, 0, 0, 0, p_{aaBb|Bb}, 0, 0)\mathbf{g}$
$\mu_{bB} = (0, 0, p_{AAbB|bB}, 0, 0, 0, p_{AabB|bB}, 0, 0, 0, p_{aAbB|bB}, 0, 0, 0, p_{aabB|bB}, 0)\mathbf{g}$
$\mu_{bb} = (0, 0, 0, p_{AAbb|bb}, 0, 0, 0, p_{Aabb|bb}, 0, 0, 0, p_{aAbb|bb}, 0, 0, 0, p_{aabb|bb})\mathbf{g}$

Gametic means

$\mu_{AB} = (p_{AABB|AB}, p_{AABb|AB}, p_{AAbB|AB}, 0, p_{AaBB|AB}, p_{AaBb|AB}, p_{AabB|AB}, 0, p_{aABB|AB}, p_{aABb|AB}, p_{aAbB|AB}, 0, 0, 0, 0, 0)\mathbf{g}$

$\mu_{Ab} = (0, p_{AABb|Ab}, p_{AAbB|Ab}, p_{AAbb|Ab}, 0, p_{AaBb|Ab}, p_{AabB|Ab}, p_{Aabb|Ab}, 0, p_{aABb|Ab}, p_{aAbB|Ab}, p_{aAbb|Ab}, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aB} = (0, 0, 0, 0, p_{AaBB|aB}, p_{AaBb|aB}, p_{AabB|aB}, 0, p_{aABB|aB}, p_{aABb|aB}, p_{aAbB|aB}, 0, p_{aaBB|aB}, p_{aaBb|aB}, p_{aabB|aB}, 0)\mathbf{g}$
$\mu_{ab} = (0, 0, 0, 0, 0, p_{AaBb|ab}, p_{AabB|ab}, p_{Aabb|ab}, 0, p_{aABb|ab}, p_{aAbB|aB}, p_{aAbb|ab}, 0, p_{aaBb|ab}, p_{aabB|ab}, p_{aabb|ab})\mathbf{g}$

<u>Means of genotype-allele combinations</u>

$\mu_{ABB} = (p_{AABB|ABB}, 0, 0, 0, p_{AaBB|ABB}, 0, 0, 0, p_{aABB|ABB}, 0, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{ABb} = (0, p_{AABb|ABb}, 0, 0, 0, p_{AaBb|ABb}, 0, 0, 0, p_{aABb|ABb}, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{AbB} = (0, 0, p_{AAbB|AbB}, 0, 0, 0, p_{AabB|AbB}, 0, 0, 0, p_{aAbB|AbB}, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{Abb} = (0, 0, 0, p_{AAbb|Abb}, 0, 0, 0, p_{Aabb|Abb}, 0, 0, 0, p_{aAbb|Abb}, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aBB} = (0, 0, 0, 0, p_{AaBB|aBB}, 0, 0, 0, p_{aABB|aBB}, 0, 0, 0, p_{aaBB|aBB}, 0, 0, 0)\mathbf{g}$
$\mu_{aBb} = (0, 0, 0, 0, 0, p_{AaBb|aBb}, 0, 0, 0, p_{aABb|aBb}, 0, 0, 0, p_{aaBb|aBb}, 0, 0)\mathbf{g}$
$\mu_{abB} = (0, 0, 0, 0, 0, 0, p_{AabB|abB}, 0, 0, 0, p_{aAbB|abB}, 0, 0, 0, p_{aabB|abB}, 0)\mathbf{g}$
$\mu_{abb} = (0, 0, 0, 0, 0, 0, 0, p_{Aabb|abb}, 0, 0, 0, p_{aAbb|abb}, 0, 0, 0, p_{aabb|abb})\mathbf{g}$

$\mu_{AAB} = (p_{AABB|AAB}, p_{AABb|AAB}, p_{AAbB|AAB}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{AaB} = (0, 0, 0, 0, p_{AaBB|AaB}, p_{AaBb|AaB}, p_{AabB|AaB}, 0, 0, 0, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aAB} = (0, 0, 0, 0, 0, 0, 0, 0, p_{aABB|aAB}, p_{aABb|aAB}, p_{aAbB|aAB}, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aaB} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, p_{aaBB|aaB}, p_{aaBb|aaB}, p_{aabB|aaB}, 0)\mathbf{g}$
$\mu_{AAb} = (0, p_{AABb|AAb}, p_{AAbB|AAb}, p_{AAbb|AAb}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{Aab} = (0, 0, 0, 0, 0, p_{AaBb|Aab}, p_{AabB|Aab}, p_{Aabb|Aab}, 0, 0, 0, 0, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aAb} = (0, 0, 0, 0, 0, 0, 0, 0, 0, p_{aABb|aAb}, p_{aAbB|aAb}, p_{aAbb|aAb}, 0, 0, 0, 0)\mathbf{g}$
$\mu_{aab} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, p_{aaBb|aab}, p_{aabB|aab}, p_{aabb|aab})\mathbf{g}$

C. CONTRASTS OF GENE-GENE GENETIC AND EPIGENETIC EFFECTS OF EQUATION [9]

The epigenetic model defined by Eq[1] has 80 nonzero effects, noting that $i_{jk} = (ia)_{jkl} = (di)_{jklm} = 0$ if $j = k$, $i_{lm} = (ai)_{jlm} = (id)_{jklm} = 0$ if $l = m$, and $(ii)_{jklm} = 0$ if $j = k$ or $l = m$. With the sixteen genotypic values defined in **g** (Eq[3], fifteen unique contrasts of the 80 nonzero effects can be defined such that each contrast defines and tests a unique genetic or epigenetic effect. The general expression of these fifteen contrasts is:

$$b_j = \mathbf{k}_j'\boldsymbol{\beta} = \mathbf{s}_j\mathbf{g} \qquad [C.1]$$

where $\boldsymbol{\beta}$ = 81 × 1column vector of the population mean and the 80 nonzero genetic and epigenetic effects in Eq[1], and $\mathbf{k}_j$ = 81 × 1column vector of contrast coefficients of $\boldsymbol{\beta}$ ($j = 2, \ldots 16$). Then,

$$b_2 = b_{a1} = a_A - a_a = \mu_A - \mu_a = \mathbf{k}_2'\boldsymbol{\beta} = \mathbf{s}_2\mathbf{g} \qquad [C.2]$$

$b_3 = b_{a2} = a_B - a_b = \mu_B - \mu_b = \mathbf{k}_3'\boldsymbol{\beta} = \mathbf{s}_3\mathbf{g}$ [C.3]

$b_4 = b_{d1} = d_{Aa} - \frac{1}{2}\,(d_{AA} + d_{aa}) = \frac{1}{2}(\delta_{Aa} + \delta_{aA} - \delta_{AA} - \delta_{aa}) = \frac{1}{2}(\mu_{Aa} + \mu_{aA} - \mu_{AA} - \mu_{aa})$
$= \mathbf{k}_4'\boldsymbol{\beta} = \mathbf{s}_4\mathbf{g}$ [C.4]

$b_5 = b_{d2} = d_{Bb} - \frac{1}{2}\,(d_{BB} + d_{bb}) = \frac{1}{2}(\delta_{Bb} + \delta_{bB} - \delta_{BB} - \delta_{bb}) = \frac{1}{2}(\mu_{Bb} + \mu_{bB} - \mu_{BB} - \mu_{bb})$
$= \mathbf{k}_5'\boldsymbol{\beta} = \mathbf{s}_5\mathbf{g}$ [C.5]

$b_6 = b_{i1} = i_{Aa} - i_{aA} = \delta_{Aa} - \delta_{aA} = \mu_{Aa} - \mu_{aA} = \mathbf{k}_6'\boldsymbol{\beta} = \mathbf{s}_6\mathbf{g}$ [C.6]

$b_7 = b_{i2} = i_{Bb} - i_{bB} = \delta_{Bb} - \delta_{bB} = \mu_{Bb} - \mu_{bB} = \mathbf{k}_7'\boldsymbol{\beta} = \mathbf{s}_7\mathbf{g}$ [C.7]

$b_8 = b_{aa} = [(aa)_{AB} - (aa)_{Ab}] - [(aa)_{aB} - (aa)_{ab}] = \mu_{AB} - \mu_{Ab} - \mu_{aB} + \mu_{ab} = \mathbf{k}_8'\boldsymbol{\beta} = \mathbf{s}_8\mathbf{g}$ [C.8]

$b_9 = b_{ad} = \{(ad)_{ABb} - \frac{1}{2}\,[(ad)_{ABB} + (ad)_{Abb}]\} - \{(ad)_{aBb} - \frac{1}{2}\,[(ad)_{aBB} + (ad)_{abb}]\}$
$= \frac{1}{2}[(a\delta)_{ABb} + (a\delta)_{AbB} - (a\delta)_{ABB} - (a\delta)_{Abb} - (a\delta)_{aBb} - (a\delta)_{abB} + (a\delta)_{aBB} + (a\delta)_{abb}]$
$= \frac{1}{2}(\mu_{ABb} + \mu_{AbB} - \mu_{ABB} - \mu_{Abb} - \mu_{aBb} - \mu_{abB} + \mu_{aBB} + \mu_{abb})$
$= \mathbf{k}_9'\boldsymbol{\beta} = \mathbf{s}_9\mathbf{g}$ [C.9]

$b_{10} = b_{da} = \{(da)_{AaB} - \frac{1}{2}\,[(da)_{AAB} + (da)_{aaB}]\} - \{(da)_{Aab} - \frac{1}{2}\,[(da)_{AAb} + (da)_{aab}]\}$
$= \frac{1}{2}[(\delta a)_{AaB} + (\delta a)_{aAB} - (\delta a)_{AAB} - (\delta a)_{aaB} - (\delta a)_{Aab} - (\delta a)_{aAb} + (\delta a)_{AAb} + (\delta a)_{aab}]$
$= \frac{1}{2}(\mu_{AaB} + \mu_{aAB} - \mu_{AAB} - \mu_{aaB} - \mu_{Aab} - \mu_{aAb} + \mu_{AAb} + \mu_{aab})$
$= \mathbf{k}_{10}'\boldsymbol{\beta} = \mathbf{s}_{10}\mathbf{g}$ [C.10]

$b_{11} = b_{ai} = [(ai)_{ABb} - (ai)_{AbB}] - [(ai)_{aBb} - (ai)_{abB}] = [(a\delta)_{ABb} - (a\delta)_{AbB}] - [(a\delta)_{aBb} - (a\delta)_{abB}]$
$= \mu_{ABb} - \mu_{AbB} - \mu_{aBb} + \mu_{abB}$
$= \mathbf{k}_{11}'\boldsymbol{\beta} = \mathbf{s}_{11}\mathbf{g}$ [C.11]

$b_{12} = b_{ia} = [(ia)_{AaB} - (ia)_{aAB}] - [(ia)_{Aab} - (ia)_{aAb}] = [(\delta a)_{AaB} - (\delta a)_{aAB}] - [(\delta a)_{Aab} - (\delta a)_{aAb}]$
$= \mu_{AaB} - \mu_{aAB} - \mu_{Aab} + \mu_{aAb}$
$= \mathbf{k}_{12}'\boldsymbol{\beta} = \mathbf{s}_{12}\mathbf{g}$ [C.12]

$b_{13} = b_{dd} = \{(dd)_{AaBb} - \frac{1}{2}[(dd)_{AaBB} + (dd)_{AaBB}]\} - \frac{1}{2}\{\{(dd)_{AABb} - \frac{1}{2}[(dd)_{AABB}+(dd)_{AAbb}]\}$
$+ \{\{(dd)_{aaBb} - \frac{1}{2}[(dd)_{aaBB}+(dd)_{aabb}]\}\}$
$= \frac{1}{4}[(\delta\delta)_{AaBb} + (\delta\delta)_{AabB} + (\delta\delta)_{aABb} + (\delta\delta)_{aAbB} - (\delta\delta)_{AaBB} - (\delta\delta)_{aABB} - (\delta\delta)_{Aabb} -$
$(\delta\delta)_{aAbb} -$
$(\delta\delta)_{AABb} - (\delta\delta)_{AAbB} + (\delta\delta)_{AABB} + (\delta\delta)_{AAbb} - (\delta\delta)_{aaBb} - (\delta\delta)_{aabB} + (\delta\delta)_{aaBB} + (\delta\delta)_{aabb}]$
$= \frac{1}{4}(g_{AaBb} + g_{AabB} + g_{aABb} + g_{aAbB} - g_{AaBB} - g_{aABB} - g_{Aabb} - g_{aAbb} - g_{AABb} - g_{AAbB} +$
$g_{AABB} +$
$g_{AAbb} - g_{aaBb} - g_{aabB} + g_{aaBB} + g_{aabb})$
$= \mathbf{k}_{13}'\boldsymbol{\beta} = \mathbf{s}_{13}\mathbf{g}$ [C.13]

$b_{14} = b_{di} = [(di)_{AaBb} - (di)_{AabB}] - \frac{1}{2}\,\{[(di)_{AABb} - (di)_{AAbB}] + [(di)_{aaBb} - (di)_{aabB}]\}$

$= \frac{1}{2}[(\delta\delta)_{AaBb}+(\delta\delta)_{aABb}-(\delta\delta)_{AabB}-(\delta\delta)_{aAbB}-(\delta\delta)_{AABb}+(\delta\delta)_{AAbB}-(\delta\delta)_{aaBb}+(\delta\delta)_{aabB}]$
$= \frac{1}{2}(g_{AaBb} + g_{aABb} - g_{AabB} - g_{aAbB} - g_{AABb} + g_{AAbB} - g_{aaBb} + g_{aabB})$
$= \mathbf{k}_{14}'\boldsymbol{\beta} = \mathbf{s}_{14}\mathbf{g}$ [C.14]

$b_{15} = b_{id} = [(id)_{AaBb} - (id)_{aABb}] - \frac{1}{2}\{[(id)_{AaBB} - (id)_{aABB}] + [(id)_{Aabb} - (id)_{aAbb}]\}$
$= \frac{1}{2}[(\delta\delta)_{AaBb}+(\delta\delta)_{AabB}-(\delta\delta)_{aABb}-(\delta\delta)_{aAbB}-(\delta\delta)_{AaBB}+(\delta\delta)_{aABB}-(\delta\delta)_{Aabb}+(\delta\delta)_{aAbb}]$
$= \frac{1}{2}(g_{AaBb} + g_{AabB} - g_{aABb} - g_{aAbB} - g_{AaBB} + g_{aABB} - g_{Aabb} + g_{aAbb})$
$= \mathbf{k}_{15}'\boldsymbol{\beta} = \mathbf{s}_{15}\mathbf{g}$ [C.15]

$b_{16} = b_{ii} = [(ii)_{AaBb}-(ii)_{AabB}] - [(ii)_{aABb}-(ii)_{aAbB}] = (\delta\delta)_{AaBb}-(\delta\delta)_{AabB}-(\delta\delta)_{aABb}+(\delta\delta)_{aAbB}$
$= g_{AaBb} - g_{AabB} - g_{aABb} + g_{aAbB}$
$= \mathbf{k}_{16}'\boldsymbol{\beta} = \mathbf{s}_{16}\mathbf{g}$ [C.16]

where $\mathbf{s}_i$ = row vector of marginal and conditional frequencies derived from those for calculating means in Section B that in turn were used for calculating effects in Eq[1], b_{a1}, b_{a2}, = additive effects of locus 1 and locus 2, b_{d1}, b_{d2} = dominance effects of locus 1 and locus 2, b_{i1}, b_{i2} = imprinting effects of locus 1 and locus 2, b_{aa} additive × additive epistasis effect, b_{ad} = additive × dominance epistasis effect, b_{da} = dominance × additive epistasis effect, b_{ai} = additive × imprinting effect, b_{ia} = dominance × additive epistasis effect, b_{dd} dominance ×dominance epistasis effect, b_{di} = dominance × imprinting effect, b_{id} = imprinting × dominance effect, and b_{ii} = imprinting × imprinting epigenetic effect. Each of these contrasts is expected to be null if the true effect is absent. It can be shown that $\mathbf{s}_i$ is a contrast vector for i=1,…,16.

D. STATISTICAL POWER AND SAMPLE SIZE

In Eq[6], the first column of **T** is **1**, a column vector of 1's. Using Eq[3], the total genetic variance can be partitioned into components attributable to the genetic contrasts defined by Eq[C.2-C.16], i.e.,

$$\sigma_g^2 = \mathbf{g}'\left[\text{diag}(\mathbf{p})\text{-}\mathbf{pp}'\right]\mathbf{g}=\mathbf{b}'\left\{\mathbf{T}'\left[\text{diag}(\mathbf{p})\text{-}\mathbf{pp}'\right]\mathbf{T}\right\}\mathbf{b}=\mathbf{b}_g{}'\mathbf{C}\mathbf{b}_g$$
$$=\sum_{i=1}^{15} c_{ii}b_{i+1}^2+2\sum_{j=i+1}^{15} c_{ij}b_{i+1}b_{j+1}=\sum_{i=2}^{16}\sigma_i^2+2\sum_{i=2}^{15}\sum_{j=i+1}^{16}\sigma_{ij} \quad \text{[D.1]}$$

Where

$\mathbf{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16})'$
$= (p_{AABB}, p_{AABb}, p_{AAbB}, p_{AAbb}, p_{AaBB}, p_{AaBb}, p_{AabB}, p_{Aabb},$
$p_{aABB}, p_{aABb}, p_{aAbB}, p_{aAbb}, p_{aaBB}, p_{aaBb}, p_{aabB}, p_{aabb})'$ [D.2]

diag(**p**) = diagonal matrix with p_i as the (i,i)th element, $\mathbf{b}_g = (b_2, ..., b_{16})'$ (Eq[9]), $\sigma^2_{i+1}=c_{ii}b^2_{i+1}$, $\sigma_{i+1,j+1}=c_{ii}b_{i+1}b_{j+1}$, and

$$\mathbf{C} = \mathbf{T}_g' [\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}']\mathbf{T}_g \qquad [D.3]$$

with $\mathbf{T}_g = (\mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4, \mathbf{t}_5, \mathbf{t}_6, \mathbf{t}_7, \mathbf{t}_8, \mathbf{t}_9, \mathbf{t}_{10}, \mathbf{t}_{11}, \mathbf{t}_{12}, \mathbf{t}_{13}, \mathbf{t}_{14}, \mathbf{t}_{15}, \mathbf{t}_{16})$, obtained from matrix **T** by removing the first column of **T**.

Statistical power (π) is the probability that an effect is detected when the effect is present, commonly denoted by π = 1 – β, where β is the type II error, i.e., the probability of false 'negatives'. A standardized normal distribution denoted by N(0,1) is assumed for deriving the statistical power. The normal distribution is chosen because the calculation of the exact residual degrees of freedom is unnecessary, providing analytical simplicity. Since the residual degrees of freedom are sufficiently large for the sample sizes discussed in this article ($N \geq 100$), the normal distribution practically yields identical results as the t-distribution that is often used in QTL analysis. The general expression for π is:

$$\pi = 1 - \beta = 1 - \Pr(Z < z_x) = 1 - \Phi(z_x) \qquad [D.4]$$

where Z is a N(0,1) random variable, z_x is the ordinate of the standardized normal curve corresponding to the type II error of β, and Φ is the cumulative distribution function of standard normal random variable. Let b_x = contrast for testing an imprinting related effect, x = i, ia, id, ii. Using $E(L_x)$ in place of b_x for x = i, ia, id, ii as defined by Eq[C.1], the z_x value in Eq. (C5) can be expressed as

$$z_x = z_{\alpha/2} - \frac{E(b_x)}{\sqrt{\mathrm{var}(b_x)}} = z_{\alpha/2} - \frac{\sqrt{n_x}E(b_x)}{\sqrt{V_x}} \qquad [D.5]$$

where n_x is the sample size and $V_x = n_x\mathrm{var}(L_x)$, for x = i, ia, id, ii. For convenience, V_x will be referred to as the 'kernel' of the contrast variance, meaning that V_x differs from $\mathrm{var}(b_x)$ only by a constant of n_x. The expressions of V_x in terms of genetic parameters are given as follows,

$$V_i = \sigma_e^2 \sum_{i=1}^{16} p_i^{-1} s_{6i}^2 \qquad [D.6]$$

$$V_{ia} = \sigma_e^2 \sum_{i=1}^{16} p_i^{-1} s_{12i}^2 \qquad [D.7]$$

$$V_{id} = \sigma_e^2 \sum_{i=1}^{16} p_i^{-1} s_{15i}^2 \qquad [D.8]$$

$$V_{ii} = \sigma_e^2 \sum_{i=1}^{16} p_i^{-1} s_{16i}^2 \qquad [D.9]$$

where **p** = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16})' = (p_{AABB}, p_{AABb}, p_{AAbB}, p_{AAbb}, p_{AaBB}, p_{AaBb}, p_{AabB}, p_{Aabb}, p_{aABB}, p_{aABb}, p_{aAbB}, p_{aAbb}, p_{aaBB}, p_{aaBb}, p_{aabB}, p_{aabb})'. The proof of Eq[D.6-D.9] is similar to that in Mao et al. (2006). Letting $\lambda_x = E(b_x)/\sqrt{V_x}$, z_x of Eq[D.5] can be expressed in terms of genetic parameters as

$$z_x = z_{a/2} - \sqrt{n_x}\lambda_x \qquad [D.10]$$

where

$$\lambda_i = \frac{H_i}{\sqrt{\left(1-H_G^2\right)c_{5,5}\sum_{i=1}^{16} p_i^{-1}s_{6i}^2}} \qquad [D.11]$$

$$\lambda_{ia} = \frac{H_i}{\sqrt{\left(1-H_G^2\right)c_{11,11}\sum_{i=1}^{16} p_i^{-1}s_{12i}^2}} \qquad [D.12]$$

$$\lambda_{id} = \frac{H_i}{\sqrt{\left(1-H_G^2\right)c_{14,14}\sum_{i=1}^{16} p_i^{-1}s_{15i}^2}} \qquad [D.13]$$

$$\lambda_{ii} = \frac{H_i}{\sqrt{\left(1-H_G^2\right)c_{15,15}\sum_{i=1}^{16} p_i^{-1}s_{16i}^2}} \qquad [D.14]$$

Eqs.(C11-C15), $H_i^2 = \sigma_6^2/\sigma_y^2$, $H_{ia}^2 = \sigma_{12}^2/\sigma_y^2$, $H_{id}^2 = \sigma_{15}^2/\sigma_y^2$, $H_{ii}^2 = \sigma_{16}^2/\sigma_y^2$ are contrast heritabilities with σ_i^2 (i= 6, 12, 15, 16) given in Eq.(C3-C4), and $H_G^2 = \sigma_G^2/\sigma_y^2$.

Using the above results, the minimum sample size required for given levels of type I and type II errors can be expressed as:

$$n_x = \frac{V_x\left(z_{\alpha/2}+z_\beta\right)^2}{E\left(L_x\right)^2} = \frac{\left(z_{\alpha/2}+z_\beta\right)^2}{\lambda_x^2} \qquad [D.15]$$

where $z_{\alpha/2}$ and z_β are the ordinate of the standardized normal curve corresponding to the probabilities of α/2 and β. The sample size given by Eq[D.15] is a decreasing function of type-I and type-II errors, and size of the epistasis effect (or contrast heritability).

E. FREQUENCY OF HETEROZYGOUS SNP GENOTYPE WITH KNOWN PARENTAL ALLELE ORIGIN

For imprinting related effects, distinguish between *Aa* and *aA* genotypes requires knowing parental origin of the two alleles. Based on Table E1, the probability that parental allele origins of a heterozygous genotype for a multi-allelic locus can be derived as

$$w_2 = P_{a2}/Q_a = 1 - P_2^2/Q_a \qquad [E.1]$$
$$w_1 = P_{a1}/Q_a = 1 - [P_2(1 - P_6)]/Q_a \qquad [E.2]$$

where w_m = probability that a heterozygous genotype has known parental allele origins when m known grandparents (m =1, 2), $Q_a = 1 - P_1^2 - P_3^2 - P_4^2 - P_5^2 - P_6^2 - 2P_1P_3 - 2P_4P_5 - 2P_6(1-P_6)$ = probability of all matings that can produce the A_iA_j offspring, $P_{a2} = Q_a - P_2^2$ = probability of all matings that can produce the A_iA_j offspring with known parental allele origins when both parents have known genotypes, and $P_{a1} = Q_a - P_2(1 - P_6)$ = probability of all matings that can produce the A_iA_j offspring with known parental allele origins when one parent has known genotypes. For a bi-allelic locus with *A* and *a* aleles such as an SNP, Eq[E1-E2] reduce to:

$$w_2 = 1 - P_2^2/(1 - P_1^2 - P_4^2) = 1 - (2pq)^2/(1 - p^4 - q^4) \qquad [E.3]$$
$$w_1 = 1 - P_2/(1 - P_1^2 - P_4^2) = 1 - 2pq/ (1 - p^4 - q^4) \qquad [E.4]$$

where p = frequency of *A* allele and *q* = frequency of a allele.

Table E1. Probability of each possible mating to produce a heterozygous offspring (A_iA_j) for a multi-allelic locus.

		A_iA_i	A_iA_j	A_iA_k	A_jA_j	A_jA_k	A_kA_l
		P_1	P_2	P_3	P_4	P_5	P_6
A_iA_i	$P_1 = p_i^2$	0	P_1P_2	0	P_1P_4	P_1P_5	0
A_iA_j	$P_2 = 2p_ip_j$	P_1P_2	P_2^2	P_2P_3	P_2P_4	P_2P_5	0
A_iA_k	$P_3 = 2p_i(1- p_i - p_j)$	0	P_2P_3	0	P_3P_4	P_3P_5	0
A_jA_j	$P_4 = p_j^2$	P_1P_4	P_2P_4	P_3P_4	0	0	0
A_jA_k	$P_5 = 2p_j(1- p_i - p_j)$	P_1P_5	P_2P_5	P_3P_5	0	0	0
A_kA_l	$P_6 = (1- p_i - p_j)^2$	0	0	0	0	0	0