MODELS AND METHODS FOR GENETIC LINKAGE AND ASSOCIATION ANALYSES

by

Jin Zheng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2010

Doctoral Committee:

Professor Gonçalo Abecasis, Chair
Professor Michael Lee Boehnke
Professor David T. Burke
Assistant Professor Zhaohui Qin

The memory of my dad.

Acknowledgements

The writing of a dissertation can be a lonely and isolating experience, yet it is obviously not possible without the personal and practical support of numerous people. This is a great opportunity to express my sincere gratitude to all of wonderful people who have helped me on this work.

First and foremost, I wish to thank my dissertation advisor, Professor Gonçalo R. Abecasis, for accepting me to be his student, for his continuous support for my pursuing the Ph.D., for his encouraging and advising me to complete the dissertation and all the research that lies behind it.

I would like to thank all my dissertation committee. A special thank to Professor Michael Boehnke for his detailed comments on my dissertation and supportive discussions with me. I am grateful to Professor David T. Burke for sharing his thoughts and sophisticated ideas with me. I am also pleased to thank Assistant Professor Zhaohui Qin for his great teaching and imparting the knowledge to me. In addition, other special thanks are due to Associate Professor Bhramar Mukherjee, Dr. Wei-min Chen, and Dr. Paul Scheet for their inspiring me to complete my dissertation.

Many friends and classmates have provided me with their warmest assistance and support. I greatly value their friendship and I deeply appreciate their belief in me.

Finally, I owe my greatest debts to my family. I thank my husband, Zhangsheng, my son, Zhichong, and my mom, Wenjuan for their constant and endless love!

Table of Contents

List of Figures

List of Tables

ABSTRACT

MODELS AND METHODS FOR GENETIC LINKAGE AND ASSOCIATION ANALYSES

by

Jin Zheng

Chair: Gonçalo Abecasis

Linkage and association analysis are both tools for mapping the locations of genes responsible for human traits. A common approach for quantitative trait linkage analysis in human pedigrees involves the use of variance component models. In the first part of this dissertation, I extended the variance-component method to allow for genetic and/or environmental variance components as functions of measured covariates. I show that our method can provide large gains in power when there is heterogeneity in heritability of the quantitative trait locus due to covariates, such as age and/or sex.

The recent availability of a high-density reference panel has allowed for the imputation of genotypes at single nucleotide polymorphism markers that were untyped in a cohort or case-control study but that have been characterized in the reference panel. In the second part of this dissertation, I compared the performance of three different strategies to take

account of the uncertainty of these imputed genotypes in the imputation-based association studies for quantitative traits. I found that for most realistic settings of genome-wide association studies (GWAS), the strategy of regressing the phenotype on the genetic dosages provided a good compromise between power and computational efficiency.

Although researchers have noticed the phenomenon of gene-environment interactions in disease etiology, it still remains uncertain how to trace the disease susceptibility loci by considering the role of environment and its potential to interact with genes, especially in GWAS. In the third part of this dissertation, I proposed a new likelihood-based method to identify genes involved in a gene-environment interaction, exploiting gene-environment independence at the population level. I compared its performance with the existing methods under different settings of parameters and by different criteria. The new likelihood-based approach shows merit in various settings, especially when the disease is not very rare. The simulation studies also showed that the empirical power of the new method was still great when the violation of the assumption was realistically modest.

Chapter 1

Introduction

1.1 Quantitative trait linkage analysis

There are many traits in human, such as blood pressure and serum lipid levels, which are best measured by continuous values. It has been well recognized that many of those quantitative traits are usually, though not necessarily always, inherited and determined by multiple genes, environmental and behavioral factors and interactions between them (Falconer and Mackay 1996, Lynch and Walsh 1998). One of genetic researchers' tasks is to understand the relationship between DNA sequence variation and variation in phenotypes for these quantitative traits, which would be important to predict disease risk and develop tailor therapeutic treatments in human populations. However, it is challenging to identify chromosomal locations and ultimately the genetic variants, or regulatory elements that affect the phenotypic expression of a trait, especially for a complex disease. There exist many quantitative trait loci with just small effects. In addition, phenomenon such as epistasis, pleiotropy and gene-environment interaction makes dissection of quantitative traits complicated.

A common tool for mapping the location of genes responsible for human quantitative

traits is linkage analysis, which aims to discover the cosegregation between the loci and genetic markers with known position. The basic methods currently used for QTL linkage mapping are based on Haseman-Elston regression methods (Haseman and Elston 1972) and variance components methods (Hopper and Mathews 1982; Amos 1994; Almasy and Blangero 2009). The fundamental idea behind these methods is that when a particular locus influences a trait, individuals that share more genetic material at that locus are likely to be more alike in their phenotypic values (Feingold 2001).

Haseman and Elston (1972) proposed the model-free linkage method by exploiting the inverse relationship between the squared trait difference and identity-by-decent (IBD) sharing between sib-pairs. Here, two alleles at a single locus are IBD if they are copies of the same allele in some earlier generation, *i.e.*, both are copies that arose by DNA replication from the same ancestral sequence without any intervening mutation. This method is relatively simple and robust (Allison *et al*. 2000), but the power is lower than variance components models (Fulker and Cherny 1996). Subsequent work extended and revised the Haseman-Elston regression to allow for multiple relative pairs from more kinds of pedigrees (Amos and Elson 1989, Sham *et al*. 2002), take account of information from all marker loci simultaneously (Fulker *et al*. 1995) and increase power (Xu *et al*. 2000).

When assumptions for the underlying quantitative trait distributions are valid, the variance components method has higher power than Haseman-Elston method (Forrest 2001, Tang and Siegmund 2001). In a typical variance components method widely used

today, variability among trait observations from individuals within pedigrees is expressed in terms of fixed effects from covariates and random effects due to an unobservable trait-affecting major locus, residual polygenic effects, and residual nongenetic variance.

In some cases, however, measured covariates can actually modify the size of genetic effects rather than directly affecting the trait mean. For example, Pilia *et al*. (2006) found that among 98 cardiovascular and personality traits, about half showed heterogeneity in variance components by age, by sex or both, and Weiss *et al*. (2006) also found substantial evidence for heterogeneity by sex in several human QTLs. To account for heterogeneity in genetic effects due to a measured covariate, in my first paper, I extended the variance components method to model random genetic and environmental variance components for each individual as a linear function of measured covariates. This model improves power in situations where genetic effects differ among individuals and these differences can be explained by a measured covariate.

1.2 Imputation-based association analysis

Association studies are another powerful tool for gene mapping. Association tests check for correlation between genetic variants and a trait of interest within a population. Linkage and association analysis are both based on the same principle that the genetic markers are close to the disease gene, we can identify a signal, even if the causal variant is not tested directly.

Association analysis is based on population data, including affected and unaffected individuals, which is generally easier to collect compared to families. The possibility of collecting large samples makes association studies attractive to detect those alleles with minor effects on a disease (Risch and Merikangas 1996; Christensen and Murray 2007). In addition, it has been proved that unrelated controls may be more powerful than individuals with the same pedigrees (Witte *et al*. 1999; Teng and Risch 1999). In an association analysis, late onset diseases can be studied (Teng and Risch 1999; Clark *et al*. 2005) and the actual disease allele is possible to be identified. Nevertheless, when the case group and the control group both are a mixture of subpopulations with different disease prevalence and allele frequency, even markers not associated with the disease will exhibit spurious association (Cardon and Palmer 2003). Linkage analysis is based on pedigree data, for which population stratification usually would not be a problem, because allele identity and the numbers of genetic variants at a locus are irrelevant (Rodriguez-Murillo and Greenberg 2008). Therefore, linkage analysis may be more appropriate for situations of rare variants.

Along with a revolution occurring in single nucleotide polymorphism (SNP) genotyping technology, it is now possible to genotype hundreds of thousands of alleles in parallel. This has made it possible to rapidly scan markers across the complete genomes of many people. The linkage and association between interesting traits and millions of markers could be tested. Recently, genome-wide linkage mapping and association studies have identified SNPs related to several complex diseases. Hundreds of thousands of SNPs in thousands of individuals have been assayed; hundreds of replicated associations have

now been reported for more than 80 diseases, traits and biological measurements as a result of genome-wide association (GWA) studies.
(http://www.genome.gov/gwastudies).

The completion of The International HapMap Project (HapMap) (International HapMap Consortium, 2007), has provided a possibility to impute missing genotypes that were not directly genotyped from a cohort or case-control study but were genotyped in the reference samples. Genotype imputation can increase the power of GWAS (Li *et al*. 2009). For example, in Willer *et al*. (2008), association of the low-density lipoprotein levels with variants in LDL receptor gene (*LDLR*) was detected only after imputation was performed, since the associated variant, rs6511720, was not selected for genotyping with the Affymetrix 500K array set, in which the best single marker tag has pair-wise r^2 of only 0.21. In addition, genotype imputation allows several different GWAS to be combined together. Those studies might be conducted at different times, with different sets of SNPs scanned.

There are several imputation programs available currently. For example, MERLIN (Abecasis *et al*. 2002, Abecasis and Wigginton 2005) and MENDEL (Lange *et al*. 1988, Lange *et al*. 2005) for genotype imputation in studies of related individuals; IMPUTE (Marchini *et al*. 2007), MACH (Li *et al*. 2006), fastPHASE/BIMBAM (Scheet and Stephens 2006, Servin and Stephens 2007), PLINK (Purcell *et al*. 2007), TUNA (Nicolae 2006), WHAP (Zaitlen *et al*. 2007), and BEAGLE (Browning 2006) for imputation in studies of unrelated individuals. For a recent review see Li *et al*. (2009).

Imputation provides probabilities of possible genotypes at the untyped positions. Then a natural question when implementing these procedures concerns how best to take account of uncertainty in imputed genotypes. In my second paper, I evaluate the relative performance of several different strategies for analyzing the distribution of imputed genotypes by simulated data with different effect sizes and sample sizes. These methods are: least-squares regression on the "best-guess" imputed genotype; regression on the expected genotype score or "dosage"; and mixture regression models that more fully incorporate posterior probabilities of genotypes at untyped SNPs. I found that for most realistic settings of genome-wide association studies (GWAS), such as modest genetic effects, large sample sizes and high average imputation accuracies, the strategy of regressing the phenotype on the genetic dosages provided a good compromise between power and computational efficiency.

1.3 Gene-environment-wide interaction studies

In contrast to Mendelian genetic diseases, complex diseases are ultimately determined by a number of genetic and environmental factors and their interactions (Schork 1997). From the beginning of the 20th century, researchers have already noted that the effects of genes could be modified by the environment. For example, Garrod (1902) suggested that the effect of individual's genotype in variation in response to drugs could be modified by diet.

The interaction between gene and environment means that how genetic and environmental factors influence the risk of a disease jointly. We usally describe the gene-environment multiplicative interaction as that the direction and magnitude of the genetic effect differs according to environmental exposure, or that genetic factors might modify the effect of an environmental exposure on disease risk.

Although researchers have noticed the phenomenon of gene-environment interactions, analysis of gene-environment interactions is included in only a small fraction of epidemiologic studies until now (Khoury and Wacholder 2008). Especially the role of environment and its potential to interact with genes generally has not been adequately addressed at a genome-wide level. During genome-wide association studies, marginal association between genetic information and disease status is usually tested in order to locate disease locus. It still remains uncertain how to trace the disease susceptibility loci by considering the gene-environment interactions. However, ignoring the gene-environment interaction could bring in biased estimation of the proportion of the disease that is explained by genes, by the environment, and/or by the interaction between them. (Hunter 2005).

Murcray *et al*. (2009) proposed a 2-step analysis of GWAS data to identify genes involved in a gene-environment interaction. Mukherjee and Chatterjee (2008) also conduct a novel empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency for testing gene-environment interaction. In my third paper, I describe a likelihood-based statistics for testing the interaction between gene and environment. At

the same time, I compare my new method, the empirical Bayes estimator method, the 2-step method, a case-only method and traditional logistic regression in terms of power and type I error by simulation studies with quite a range of different parameter settings under the GWAS framework. At some situation, our new method provides gains in power (compared to all alternative approaches), particularly when the trait being studied is common. The new likelihood-based approach provides a new way to screen the disease susceptibility loci after main genetic effects has been tested in GWAS.

Chapter 2

Variance Component Linkage Analysis Allowing for Heterogeneity in Effect Sizes due to Measured Covariates including Age and Sex

2.1 Introduction

A common approach for quantitative trait linkage analysis in human pedigrees involves the use of variance component models. Jinks and Fulker (1970) first used the variance components method to divide phenotypic variance into polygenic and environmental components by using data on relative pairs. By extending variance components to pedigree analysis involving likelihood theory, Lange *et al*. (1976) gave the theoretical foundation for linkage analysis of quantitative traits using variance components models. Goldgar (1990), Schork (1993) and Amos (1994) developed multipoint variance components methods by exploiting identity-by-descent (IBD) allele sharing between pairs of relatives. Fulker *et al*. (1995) and Almasy and Blangero (1998) further refined the approach by deriving practical strategies for the estimation of multipoint IBD matrices.

It is no secret that extensive genome-wide linkage analysis of numerous complex

phenotypes has only yielded modest results. It has even been argued that linkage analysis is not ideal for finding the genetic basis of complex phenotypes (e.g., Risch and Merikangus, 1996). However, in conventional variance component models, covariates are typically modeled as fixed effects which impact the trait mean, with the residual variance being decomposed into independent QTL, polygene and environmental effects. Covariate effects are assumed to be independent of the random effects. In some situations, measured covariates can actually modify the size of genetic effects rather than directly affecting the trait mean. This situation might be common for cardiovascular traits. Several papers report evidence for heterogeneity by age and sex in the architecture of several traits. For example, Weiss *et al*. (2006) also found substantial evidence for heterogeneity in variance components by sex for several human QTLs. Pilia *et al*. (2006) found that among total 98 quantitative traits, the 40 traits showing significant evidence for heterogeneity of variance components by sex included all five anthropometric traits and many of the blood test results (12 of 34), cardiovascular traits (8 of 20), and personality traits (15 of 35). They also found significant evidence for heterogeneity in variance components by age in 62 of the 98 traits examined, including a majority of traits in all categories.

Several methods have been developed to model genotype-covariate interaction. For example, Blangero *et al*. (1991) and Blangero (1993) applied multivariate segregation analyses to model the possibility of genotype-covariate interaction. Czerwinski *et al*. (2004) and Franceschini *et al*. (2006) extended the variance decomposition approach by including covariance-specific variance terms. Almasy *et al*. (2001) also presented several

extensions of the variance component methods for incorporating genotype x age interaction; however, in their evaluation, these extensions did not seem to increase power, which might be due to the conservative criterion for testing the heritability or the specific form of the models. Shi and Rao (2008) recently developed quantitative trait linkage analysis in the presence of temporal trends in genetics effects. They did so by modeling age effects and incorporating these effects into the covariance matrix directly. In that paper, a Gaussian function was used to model temporal trends.

To produce a computationally convenient approach for genetic linkage analysis that accounts for heterogeneity in genetic effects due to a measured covariate, we extended the variance-component method to model random genetic and environmental variance components for each individual as linear functions of measured covariates. Our model improves power in situations where genetic effects differ among individuals and these differences in effect can be explained by a measured covariate. We implemented the method in software, taking advantage of the MERLIN infrastructure (Abecasis *et al*. 2002, Abecasis and Wigginton 2005) for the analysis of human pedigrees.

SIMPLE TRAIT MODEL WITH HETEROGENEITY IN GENETIC EFFECTS

First, consider the standard additive model for a bi-allelic QTL with alleles ‘b’ and ‘B’. Let a be the additive genetic effect. In this model, the trait mean deviates for individuals with genotypes ‘b/b’, ‘B/b’ and ‘B/B’ are assumed to be $-a$, 0 (zero), and $+a$. We are interested in a situation where the additive genetic effect a is not the same for all

individuals but instead is influenced by a measured covariate. Specifically, we consider the case where the additive effect for individual i is $a_i = a + \alpha x_i$, where a is the baseline additive effect, x_i is a covariate that influences the additive genetic effect, and α quantifies the impact of the covariate on the additive genetic effect. In this model, the trait mean for individual i with genotypes 'b/b', 'B/b' or 'B/B' is assumed to be $-(a+\alpha x_i)$, 0 (zero), and $+(a+\alpha x_i)$. Depending on the sign and magnitude of the covariate effect α, the same genotype could increase phenotypic values for some individuals, but decrease phenotypic values for others.

2.2 Methods

Following to Amos (1994), the conventional variance components approach that we considered here models trait values as:

$$Y_i = \mu + MG_i(g_i) + PG_i + \sum_{j=1}^{s} \beta_j x_{ij} + e_i,$$

where Y_i represents the measured phenotype for the i^{th} individual, x_{ij} the j^{th} covariate value of the i^{th} individual, μ the overall mean, $\beta_j's$ the covariate effects, and MG_i, PG_i, and e_i the additive quantitative trait locus (QTL), polygene and environmental components of variance, respectively. For simplicity, we assume that $E(MG_i) = E(PG_i) = E(e_i) = 0$. The first two moments of the model are

$$E(Y_i) = \mu + \sum_{j=1}^{s} \beta_j x_{ij},$$

$$Var(Y_i) = \sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e,$$

$$Cov(Y_i, Y_j) = \pi_{ij}\sigma^2_{mg} + 2\Phi_{ij}\sigma^2_{pg}.$$

Here, Φ_{ij} is the kinship coefficient, defined as the probability that an allele drawn at random from an arbitrary locus in individual *i* is identical by descent (IBD) to an allele drawn at random from the same locus in individual *j* and π_{ij} is the estimated proportion of genes that are IBD at the locus of interest for individuals *i* and *j* based on the available marker data. In this model, the polygenic component is shared between individuals in proportion to their kinship coefficient; the major gene effect is also shared between individuals in proportion to the estimated IBD; and the environment component is unique to each individual. The locus specific heritability of the trait is

$$h^2_{mg} = \frac{\sigma^2_{mg}}{\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e}.$$

Under this model, the heritability, h^2_{mg}, is identical for every person in a pedigree. The hypothesis of testing this locus specific heritability is: $H_0 : \sigma^2_{mg} = 0$ versus $H_0 : \sigma^2_{mg} > 0$.

To allow for heterogeneity in genetic effects, we consider the extended model:

$$Y_i = \mu + MG_i(g_i, x_{i1}) + PG_i + \sum_{j=1}^{s} \beta_j x_{ij} + e_i.$$

In this model, $MG_i(g_i, x_{i1})$ is a function of the genotypes g_i but also of covariate x_{i1}, the value of the first covariate being considered in the model. Then the first two moments of the model are

$$E(Y_i) = \mu + \sum_{j=1}^{s} \beta_j x_{ij},$$

$$Var(Y_i) = \sigma^2_{mg}(1 + \beta_{mg} x_{i1})^2 + \sigma^2_{pg}(1 + \beta_{pg} x_{i1})^2 + \sigma^2_e, \qquad (2.1)$$

$$Cov(Y_i, Y_j) = \pi_{ij}\sigma^2_{mg}(1+\beta_{mg}x_{i1})(1+\beta_{mg}x_{j1}) + 2\Phi_{ij}\sigma^2_{pg}(1+\beta_{pg}x_{i1})(1+\beta_{pg}x_{j1}).$$

Notice that the major locus and polygenic contributions to the variance-covariance matrix now include a term related to the covariate. Here, we assume the components of the major gene and polygene have similar forms. See Appendix 2.1 for the explicit derivation of the parameters in above formulae.

Therefore, the genetic heritability of the quantitative trait due to the specific locus is

$$h^2_{mg}(x_{i1}) = \frac{\sigma^2_{mg}(1+\beta_{mg}x_{i1})^2}{\sigma^2_{mg}(1+\beta_{mg}x_{i1})^2 + \sigma^2_{pg}(1+\beta_{pg}x_{i1})^2 + \sigma^2_e},$$

which is a function of the covariate. When the value of the covariate x_{i1} increases, the value of the locus specific heritability $h^2_{mg}(x_{i1})$ might increase for some people or decrease for others, depending on the parameter β_{mg}. For example, in Pilia *et al*. 2006, the authors found that for the quantitative trait systolic blood pressure, the heritability is 8.2% among younger subjects (aged less than 42 years) and 29.8% among older subjects (aged 42 years or older). The hypothesis of testing this locus specific heritability becomes more complicated: $H_0 : \sigma^2_{mg} = 0, \beta_{mg} = 0$ versus $H_1 : \sigma^2_{mg} > 0$ and β_{mg} unconstrained.

Under the assumption that the trait is distributed as multivariate normal (Fisher 1918, Lange *et al*. 1976), maximum likelihood methods can be used to estimate parameters for the variance component model. Assume that there are N independent families with n_i individuals for each family. Then the likelihood is

$$L = \prod_{i=1}^{N} (2\pi^{-n_i/2} \left|\Omega_i\right|^{-1/2} e^{-1/2(y-\mu_i)^t \Omega_i^{-1}(y-\mu_i)}),$$

where μ_i and Ω_i are defined according to the expressions for $E(Y_i)$, $Var(Y_i)$ and $Cov(Y_i, Y_j)$ given above. To evaluate this likelihood, we first estimate the IBD coefficients for each pair of individuals within a family using Merlin (Abecasis *et al*., 2002) and, if necessary, allowing for marker-marker linkage disequilibrium (Abecasis and Wigginton 2005). We then maximize the log-likelihood function by the Nelder-Mead Simplex Method (1965) using multiple different starting values to guard against inadequate convergence. Finally, likelihood ratio tests allow us to assess the evidence for genetic linkage and to test for heterogeneity of QTL effects due to a measured covariate. We have implemented these three steps (estimation of IBD coefficients, estimation of parameter values, and likelihood ratio tests of linkage and heterogeneity) into a single package based on the MERLIN code, so that the entire process is seamless to users. The classical asymptotic distribution theory of the maximum likelihood estimates does not hold for the test statistic, since in expression (2.1), β_{mg} actually disappears under the null hypothesis. Therefore, we have not derived the exact number of degrees of freedom for these tests. Instead, we have examined their behavior through simulations.

2.3 Simulations

First, we evaluated the type I error of our extended model under the null hypothesis for different significance thresholds. We simulated 50,000 datasets including 500 nuclear families each with 4 phenotyped siblings per family. No parental phenotype or genotype

data were simulated, although our analytical engine allows for arbitrary pedigree configurations (subject to the family size restrictions inherent to the MERLIN IBD calculation engine (Lange 1997)). Under the null, we simulated genotypes for 19 SNPs with equal-frequent alleles (spaced 1 cM apart). We simulated a trait where a random environmental effect accounted for 30% of the variance, polygenic effects accounted for 50% of the variance and a major locus accounted for 20% of variance. For each dataset, we first maximized the likelihood with the constraints $H_0 : \sigma^2_{mg} = 0, \beta_{mg} = 0$. Next, we maximized the likelihood with the constraints $H_1 : \sigma^2_{mg} > 0$ and β_{mg} unconstrained. We compared the empirical distribution of the resulting likelihood ratio test statistics and different mixtures of chi-squared distributions (Figure 2.1). Although the empirical significance thresholds for our method varied slightly according to the parameters used to generate simulated data, our results suggest that critical values derived using mixture of 50% of point mass at zero and a chi-squared distribution with 2 degrees of freedom provides a good approximation (Figure 2.1, top left panel). In this case, critical value should be 4.6 and 12.43 for type I error rates of 0.05 and 0.001, respectively. These thresholds can be used for an initial analysis and precise, simulation based, thresholds can be derived when potentially interesting signals are detected.

We next carried out simulations to evaluate the power of our approach, which models heterogeneity, compared to the conventional approach, which ignores it. We simulated 1,000 datasets, each with 500 families and 4 phenotyped siblings per family. We simulated a QTL locus (whose genotypes were masked) flanked by 19 SNPs (with consecutive SNPs separated by ~1 cM), with 2 alleles of equal frequency at each locus

and in Hardy-Weinberg equilibrium, which means $P(\mathrm{b}) = p = 0.5$ and $P(\mathrm{B}) = q = 0.5$, thus, $P(\mathrm{b/b}) = p^2 = 0.25$, $P(\mathrm{B/B}) = q^2 = 0.25$, and $P(\mathrm{b/B}) = 2pq = 0.5$. The effect of the QTL genotype was influenced by one normally distributed covariate. Our simulations show that, in the absence of heterogeneity, our method results in only a small power loss compared to the traditional variance component model. In contrast, when there is interaction between the measured covariate and the QTL, our method can provide large gains in power (Figure 2). In Figure 2.2, the parameter alpha quantifies the degree of heterogeneity in the genetic effect induced by the measured covariate. When alpha is zero, there is no heterogeneity and the major locus accounts for ~15% of variance in each individual. As alpha increases, this proportion of variance explained increases for some individuals and decreases for others. For example, when alpha is 0.20, the locus specific heritability attributable to the major gene varies between 8.0% and 22.3% for 95% of individuals – note that in this setting the conventional approach and our extended model still retain similar power. As α increases further, we gradually see a gain in power due to our extended model. For example, when $\alpha = 0.6$ and the heritability varies form 0.0% to 35.5% for 95% of individuals power is 86.9% for the conventional model but 93.2% for our extended model.

To compare powers of the two approaches under different settings of parameters, we performed more simulations by different proportions of variance components, different family sizes, and different type I error rates (Figure 2.3 and Figure 2.4). In general, and similar to our initial results summarized above, when the interaction effects between the covariate and the QTL are small, the power of the two approaches is similar and

furthermore, the power for both approaches decreases slightly as the size of the interaction effect increases. However, when the interaction between the covariate and genetic effects becomes large, the power of the extended model is markedly greater than the power for conventional analyses. All powers are based on the empirical distributions of the likelihood ratio test statistics, but not on our chi-square mixture approximation (Figure 2.3 and Figure 2.4).

2.4 Data Application

Hypertension is a common precursor of serious disorders including stroke, myocardial infarction, congestive heart failure, and renal failure in whites and to a greater extent in African Americans (Williams *et al*. 2000). The Hypertension Genetic Epidemiology Network (HyperGEN) is a constituent multi-center network participating in the National Heart, Lung and Blood Institute (NHLBI) Family Blood Pressure Program (FBPP) (Feinleib *et al*. 1979, Hunt *et al*. 1989, Pe′russe *et al*. 1989, Rice *et al*. 1989), a study designed to identify genetic contributions to hypertension. HyperGEN recruited two types of participants (hypertensive sibships and random samples of subjects) in African Americans and whites (Rao *et al*. 2003). These data are part of HyperGEN study. Phenotypes are the average of several systolic and diastolic sitting blood pressure measurements, abbreviated to SBP and DBP, respectively. Age, sex and race were recorded for each individual as well. Genotyping was performed by the Mammalian Genotyping Service (MGS) in Marshfield using a standard panel of 392 anonymous microsatellite markers approximately equally spaced every 9 cM throughout the genome.

We excluded 4 markers with unknown locations and markers on chromosome X. We also removed 7 subjects who contributed little information to pedigrees while making the pedigree too complex. All ungenotyped founders were included into the data for better understanding of the relationships among individuals. In the final dataset, there were 1135 subjects (530 males and 605 females; 664 whites and 471 African Americans) having full phenotype measurements and genotype information on 370 autosomal markers, and 1983 ungenotyped persons in the final data as well. Study subjects are arranged in 412 families in total. The average marker heterozygosity was 77.1%. The numbers of generations in each pedigree are 2 (1.0%), 3 (97.3%), and 4 (1.7%). Among genotyped participants with phenotype information, the mean and standard deviation were 115.57 mmHg and 15.00 mmHg for SBP, and 69.51 mmHg and 9.54 mmHg for DBP, respectively. Distributions of SBP and DBP among different age groups are plotted in Figure 2.5. All participants with full information were aged between 18 years and 65 years with a mean of 35.5 years and standard deviation of 8.7 years. Figure 2.6 includes distributions of age among different sex and race groups.

Our goal was to detect genetic regions related to the variability of blood pressure measurements from a genome-wide scan by both conventional and extended variance component models, adjusted for age, sex and race. In this data application, we focused on comparisons of manifestation of the two models when there was heterogeneity of heritability due to a specific locus.

In Figure 2.7(a) and 2.7(b), LOD scores from conventional and extended variance component model for DBP were plotted cross the genome. At the location of 120.9 cM on chromosome 4, we found the peak LOD scores as 3.30 and 3.70 for the conventional model and the extended model, respectively. The locus specific heritability was 42.9% from the conventional model (Table 2.1). From the extended model, the heritability due to the QTL would vary upon age, for example, 59.6% for people with 18 years and 14.3% for people with 65 years. Further, we calculated the heritability among subjects younger than 36 years (median age for all genotyped participants) and the rest, which deviation makes the two groups have similar sizes. The results showed heterogeneity in heritability with 43.1% for the younger people and 5.4% for the older people. Figure 2.7(c) and 2.7(d) are plots for the extended model among the two age groups; no strong signals could be found. The nearest markers are GATA62A12 (located at 114.04 cM on chromosome 4) and ATA26B08 (located at 129.92 cM on chromosome 4).

In Figure 2.8(a) and 2.8(b), LOD scores from conventional and extended variance component model for SBP were plotted across the genome. At the location of 56.2 cM on chromosome 9, there was little evidence of heterogeneity in heritability (30.7% for the younger people and 27.1% for the older people), and the peak LOD score from our extended model (1.98) was slightly less than that for the conventional model (2.28) (Table 2.2). In contrast, the peak LOD score from the extended model (2.03) was much greater than LOD score from the conventional model (0.76), where heritability was 0.0% for the younger people and 46.7% for the older people. The locus was at 40.87 cM on chromosome 13 and nearby marker was GATA6B07 (located at 38.96 cM on

chromosome 13).

In addition, we found the peak LOD score was even higher as 3.27 at 40.9 cM on chromosome 13 from the extended model among people aged no less than 36 years (Figure 2.9), which meant the linear relationship between the genetic effect and age was more suitable among people on that group. This might suggest a quadratic relationship between the major genetic effect due to this specific locus and the covariate age for the whole sample. The phenomenon of non-monotone relationship between the heritability and the value of covariate has been found in epidemiological studies. For example, Province and Rao (1985b) found that the genetic heritability of systolic blood pressure had a temporal trend which begins at 0.10 at birth and reaches a peak of 0.28 at age 36 years, then declines to a value of 0.1 at age 48 years.

Along with the simulation studies, our data application demonstrates that our extended variance component model can perform better than the conventional one when the heterogeneity in heritability is large.

2.5 Discussion and Conclusions

The use of variance component models for linkage analysis of quantitative traits has demonstrated to be an important and powerful tool for detecting and identifying QTLs (Hopper and Mathews 1982, Goldgar 1990, and Almasy and Blangero 1998). The basic idea of variance component models is to partition the total variability of the phenotypic

values into several independent parts. For example, the variance components could include a polygenic component which could be partitioned into additive and dominant components, and an environmental component which might have include shared sibling environment. After genotyping for each individual in pedigrees, the additive and/or dominant components due to a specific locus could also be considered into models. Further, the basic variance component model has been generalized in many directions to model more complex situations. For example, gene-by-gene interactions (epistasis) (Mitchell *et al*. 1997), gene-by-sex interactions (Towne *et al*. 1997), multivariate extension (Almasy, Dyer and Blangero 1997), imprinting (Shete and Amos 2002), and longitudinal data (de Andrade *et al*. 2002).

It is well known that several human traits, such as height and blood pressure, vary with age. Most investigators treat age as a nuisance parameter and attempt to "remove" its effect on a given phenotype by some sort of statistical adjustment. However, age is not merely a statistical nuisance parameter, it is a biological construct. As individuals grow from birth, many physiological and biological changes take place, including hormonal changes during puberty and menopause, and consequently increasing risk for many diseases as a result of accumulating and/or changing exposures to environmental triggers and variation in gene expression over time. Age represents a complex surrogate for a host of underlying phenomena. It is important to note that, while age effects on the mean and variance of a phenotype can be statistically adjusted for outside a model, age effects on the covariance between the phenotypes of two relatives requires explicit modeling. Age effects on the covariance can happen as genes turn on and off at various

stages in one's growth cycle, thus rendering heritability as a complex function of age. Indeed, rodent studies have demonstrated that heritability and genetic architecture vary with age, and that different genes may turn on and off at different ages (Cheverud *et al*., 1983, and Vaughn *et al*., 1999). Here we have incorporated age-dependent variation in genetic effects by treating age as a covariate in a variance components linkage analysis model.

Shi and Rao (2008) modified the genetic variance components directly by multiply a function of temporal trend for each individual in order to model the temporal trends in heterogeneity. They demonstrated the simulation results on choosing the Gaussian function of age as the temporal trend function. In this paper, we started from the original genetic effect due to the quantitative trait loci, modified it as a linear function of a covariate (such as age). Then we derived the corresponding locus specific variance component part, which could be more clearly interpreted.

In the data application, we calculated LOD scores from different variance component models Lander and Druglyak (1995) suggested a LOD score of 3.3 in linkage analysis using 1 degrees-of-freedom test and LOD score of 3.7 using a chi-square test with 2 degrees-of-freedom for "significant linkage". Because the sizes of sibships with all genotypes and phenotypes information available were not very large (ranging from 1 to 12, with distribution as 2 (36.4%), 1 (23.8%) and 3 (16.0%); the average sibship size is 1.44), the variance component estimates may not be highly accurate (DeWan *et al*. 2001).

At the simulation section, we compared the empirical distribution of the likelihood ratio test statistics under the null hypothesis from our extended model and different mixtures of chi-squared distributions. Our results suggest that critical values of the test statistics derived using mixture of 50% of point mass at zero and 50% of a chi-squared distribution with 2 degrees of freedom provides a good approximation. In this case, critical value should be 4.6 for type I error rates of 0.05. After analyzing the data, we did permutation study by MERLIN to find out the empirical critical value in a real dataset. MERLIN has the ability to perform gene dropping simulations which replace input data with simulated chromosomes conditional on family structure and actual marker spacings and allele frequencies, as well as missing data patterns (Sawcer *et al*. 1997, Kruglyak and Daly 1998). For the quantitative trait SBP, based on 14,868 randomly simulated markers, the empirical critical value is 4.48, which is very close to 4.6 from the mixture distributions.

Our approach could be helpful in the mapping of the many human quantitative trait loci whose effects vary according to covariates, including age, sex or any other type of covariate, for example, known (fixed) genetic effect at a specific locus. Our model can be conveniently extended further, for instance, incorporating the effects of binary or continuous covariates, the heterogeneity environmental effect varying with the covariates, different function forms of effects due to covariates, or even multiple covariates in one analysis.

SOFTWARE Source code and binaries implementing the methods described here will be posted online at www.sph.umich.edu/csg/abecasis/Merlin and will be available freely for

academic or commercial use.

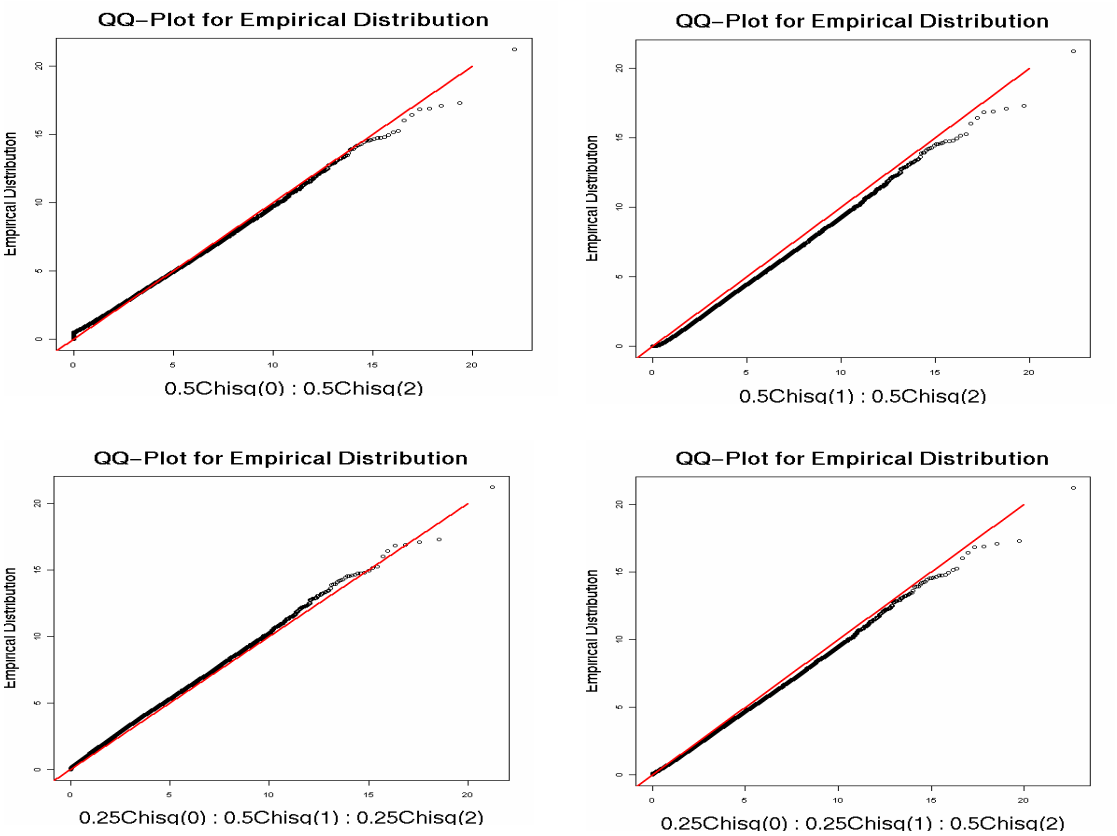


Figure 2.1: QQ-Plots comparing the distribution of empirical test statistics with different reference distributions.
To generate each reference distribution we simulated 50,000 chi-squared statistics by sampling from the appropriate mixture.

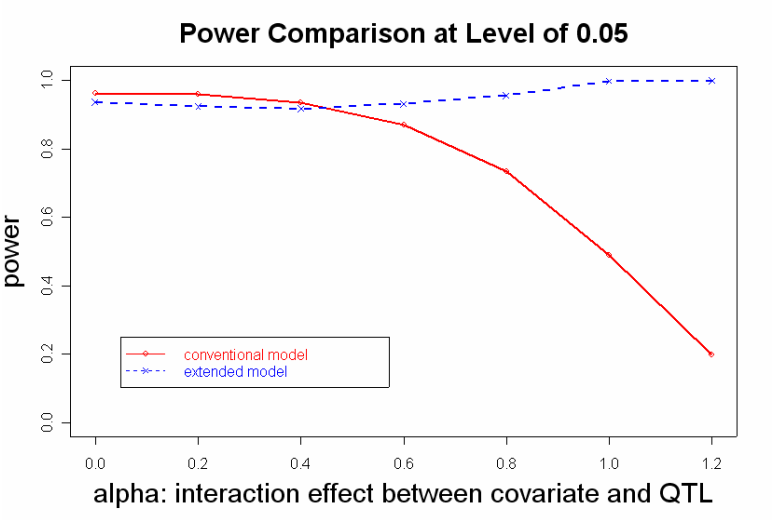


Figure 2.2: Power comparison between the conventional model and extended model. In our simulated data sets, 30% of the phenotypic variance is due to random environmental effect and 55% is due to the polygenic effect with the average heritability is set to 15%. When $\alpha = 0.0$, the QTL accounts for exactly 15% of the variance in all individuals. When $\alpha = 0.2$, the heritability varies from 11.4% to 26.8% for 95% of individuals. When $\alpha = 0.6$, the heritability varies form 0.0% to 35.5% for 95% of individuals. For larger values of α, there is extreme heterogeneity and the same genotype that raises phenotypic values in some individuals can decrease phenotypic values in others, depending on the covariate.

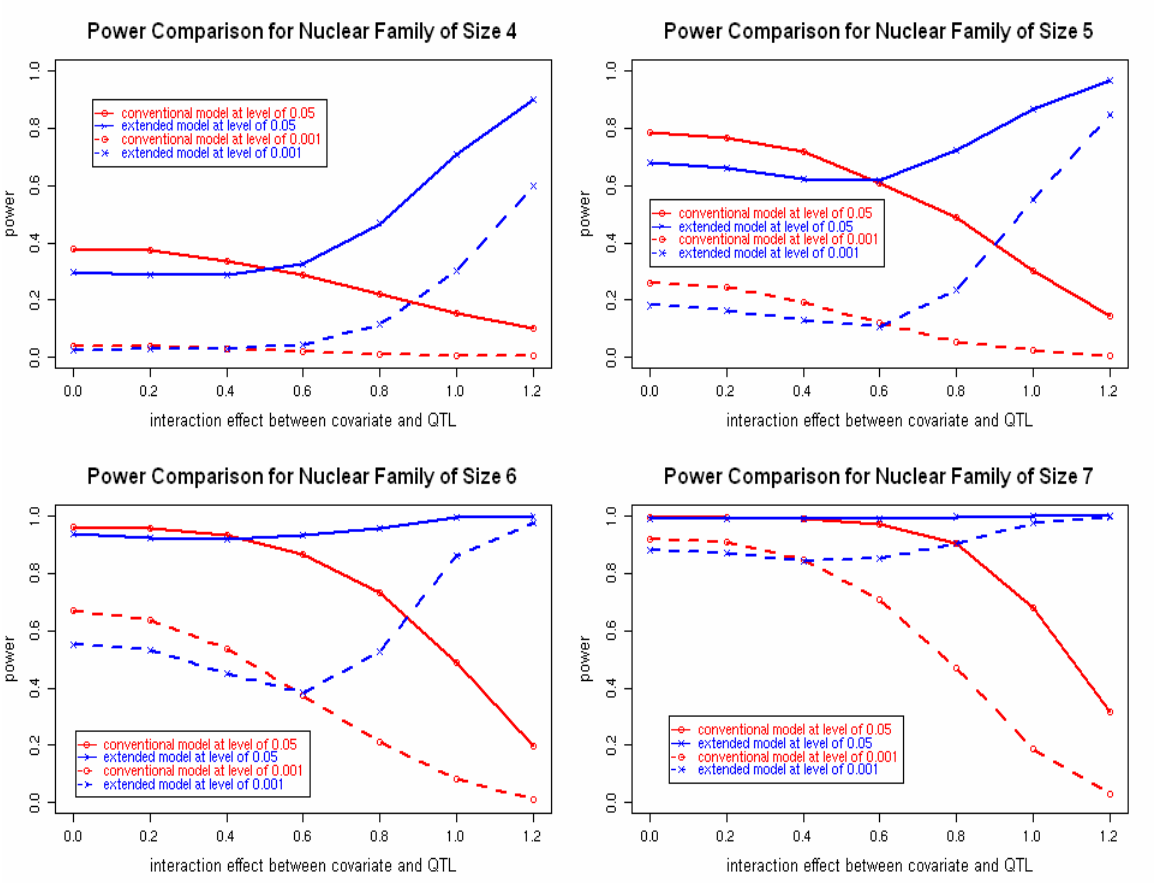


Figure 2.3: Power comparisons between the conventional models and extended models (1).

Compare for different family sizes, and two type I error rates. In these simulated data sets, 30% of the phenotypic variance is due to random environmental effect and 55% is due to the polygenic effect with the average heritability is set to 15%.

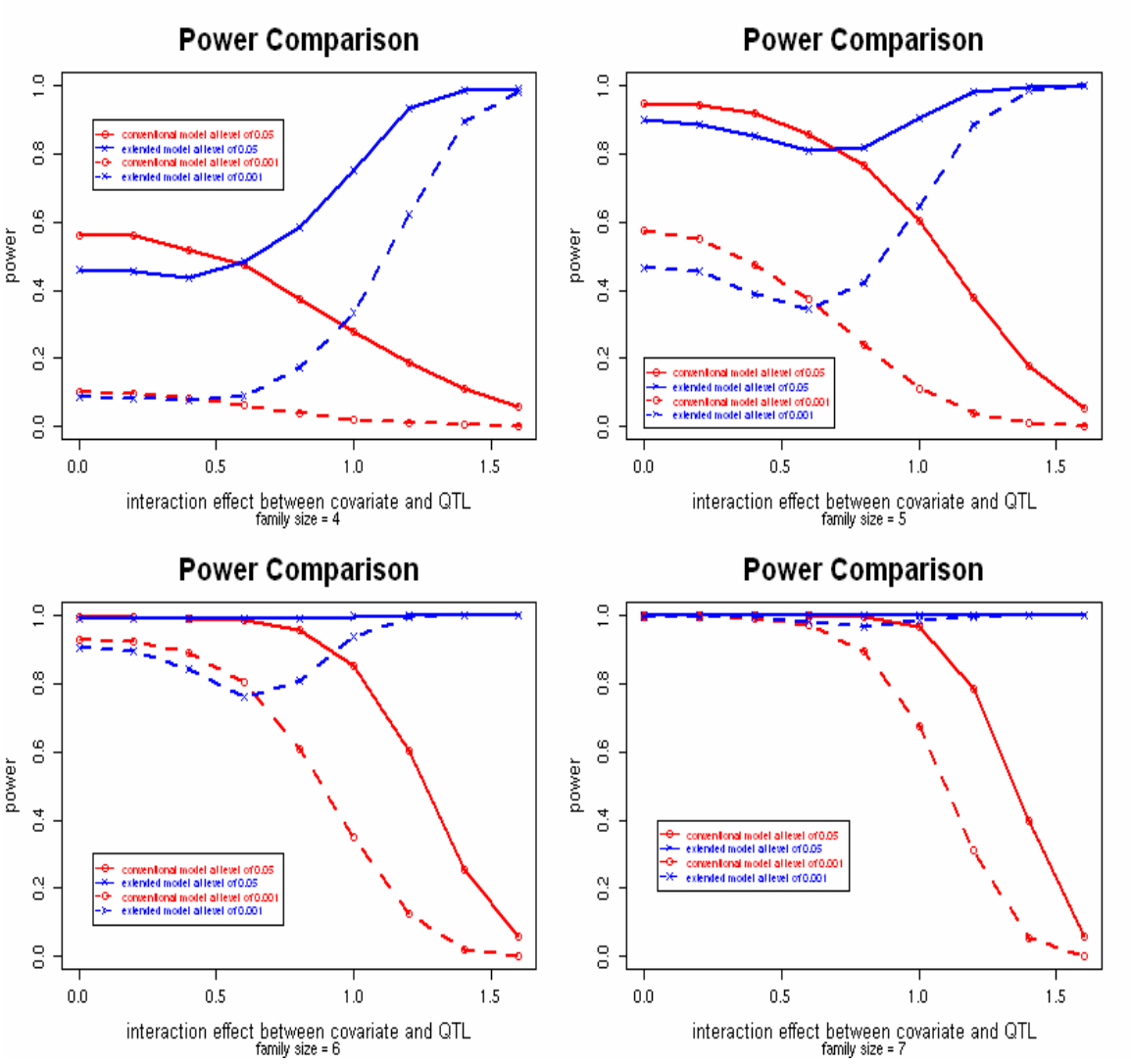


Figure 2.4: Power comparisons between the conventional models and extended models (2).

Compare for different family sizes, and two type I error rates. In these simulated data sets, 30% of the phenotypic variance is due to random environmental effect and 50% is due to the polygenic effect with the average heritability is set to 20%.

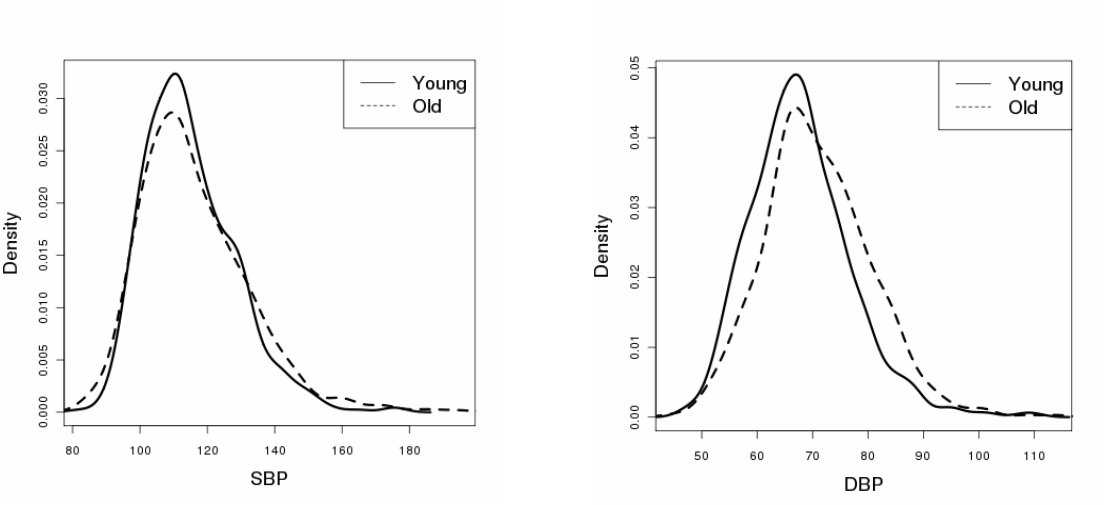


Figure 2.5: Distributions of systolic and diastolic blood pressures.
Kernel densities were plotted among younger people with age less than 36 years and older people with age equal to or greater than 36 years.

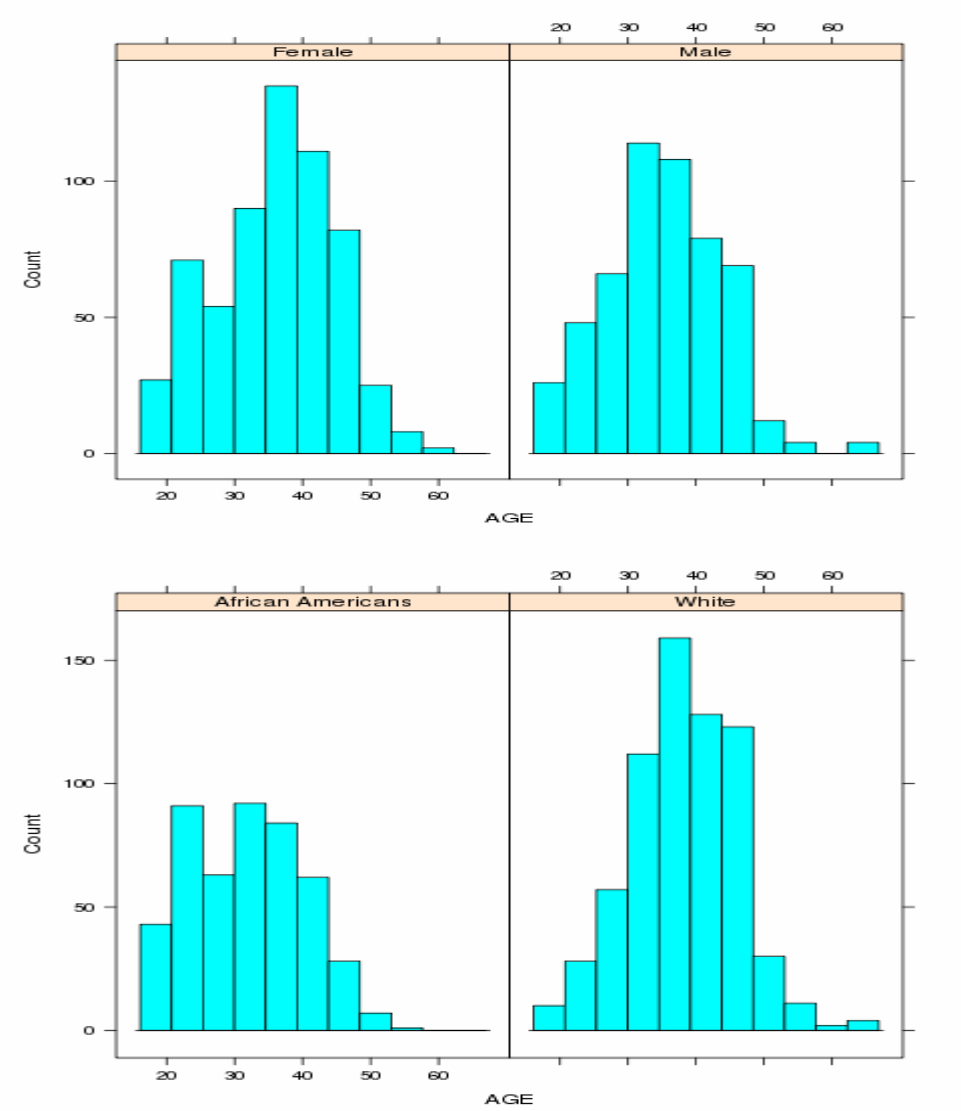


Figure 2.6: Distribution of age among different sex and race groups.

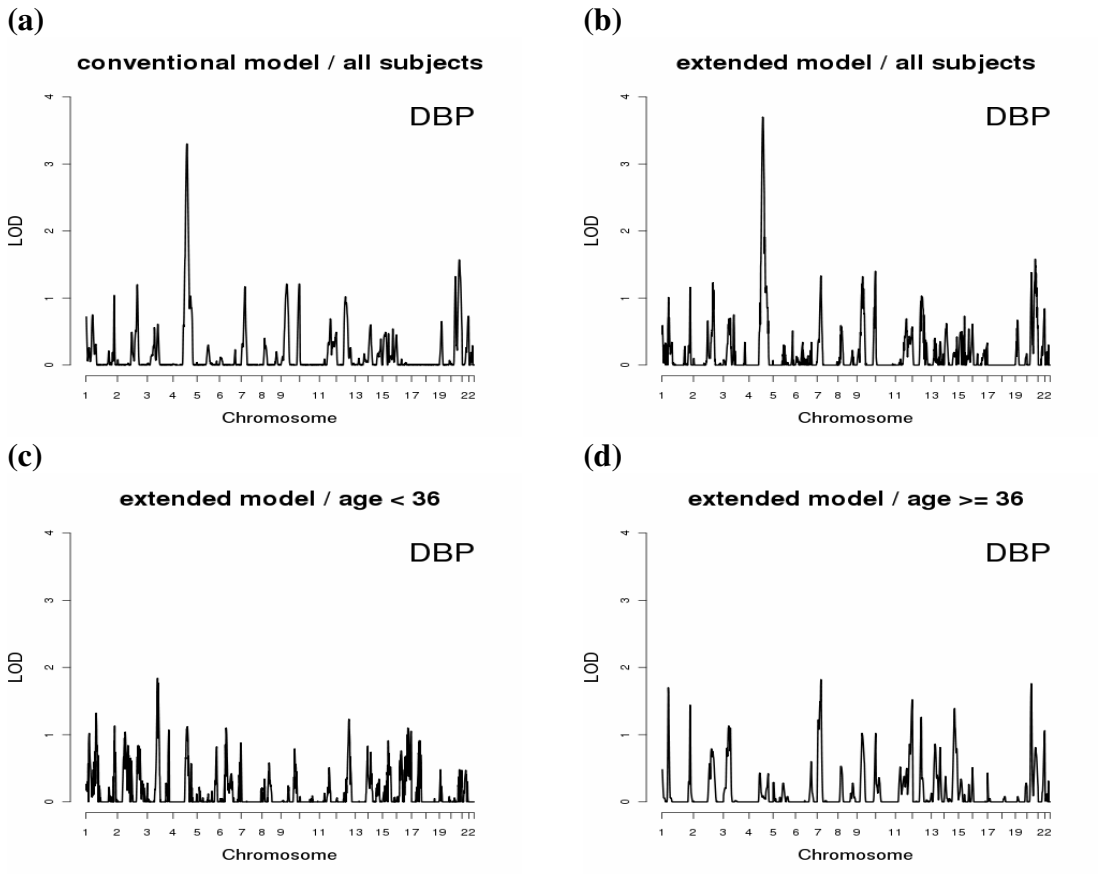


Figure 2.7: LOD scores from different models across the whole genome for DBP.

Position	Nearest Marker	Model	LOD	Locus-specific heritability
Chromosome 4 (120.93 cM)	GATA62A12 ATA26B08	Conventional	3.30	42.9% (43.1% for people aged 36 years or less; 5.4% for people older than 36 years)
		Extended	3.70	Varies with age. For example, estimated at 59.6% for people of age 18; estimated at 14.3% for people of age 65

Table 2.1: Local peak LOD scores for DBP.

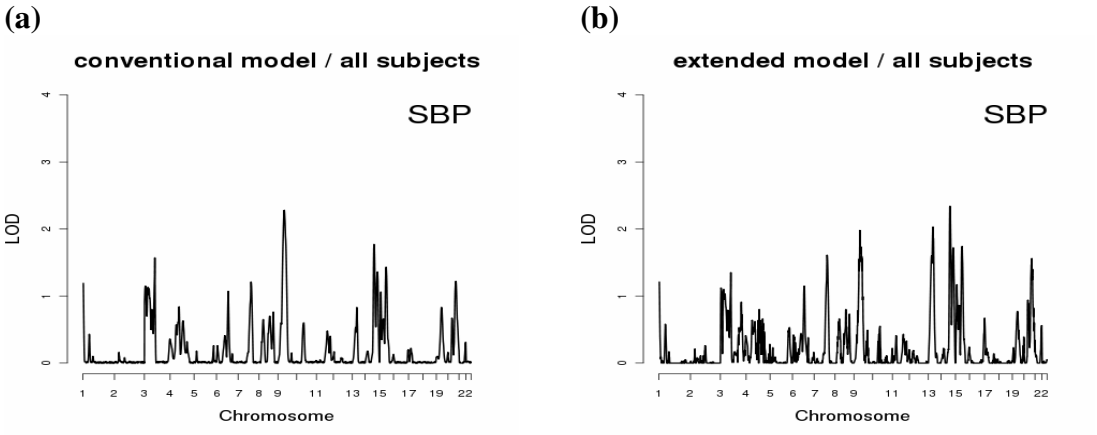


Figure 2.8: LOD scores from different models across the whole genome for SBP.

Position	Nearest Marker	Model	LOD	Locus-specific heritability
Chromosome 9 (56.23 cM)	GATA7D12	Conventional	2.28	38.3% (30.7% for people aged 36 years or less; 27.1% for people older than 36 years)
		Extended	1.98	Varies with age. For example, estimated at 31.2% for people of age 18; estimated at 27.5% for people of age 65
Chromosome 13 (40.87 cM)	GATA6B07	Conventional	0.76	18.6% (0.0% for people aged 36 years or less; 46.7% for people older than 36 years)
		Extended	2.03	Varies with age. For example, estimated at 4.00% for people of age 18; estimated at 59.1% for people of age 65

Table 2.2: Local peak LOD scores for SBP.

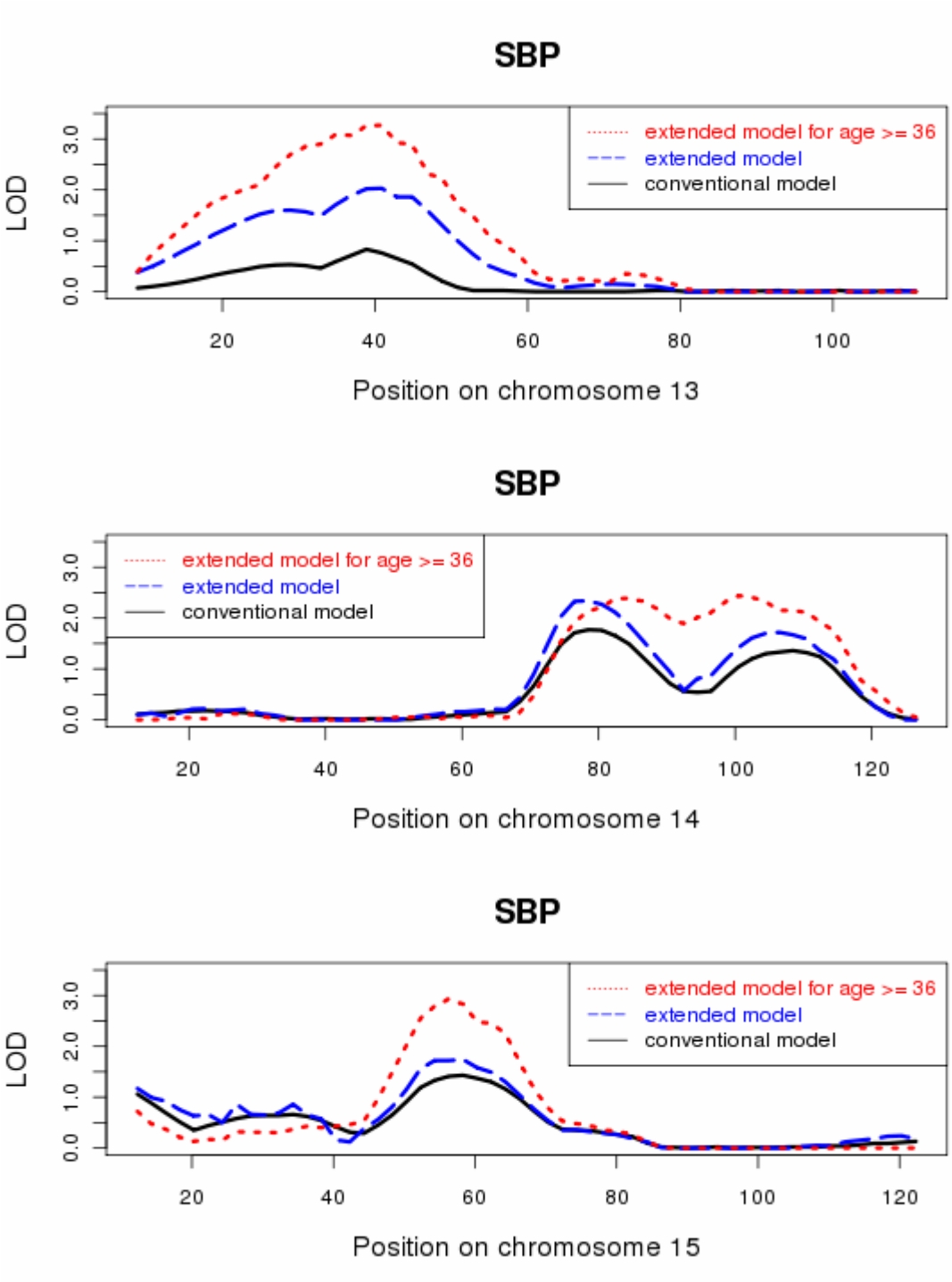

SBP
LOD
0.0
1.0
2.0
3.0
extended model for age >= 36
extended model
conventional model
20
40
60
80
100
Position on chromosome 13
SBP
LOD
0.0
1.0
2.0
3.0
extended model for age >= 36
extended model
conventional model
20
40
60
80
100
120
Position on chromosome 14
SBP
LOD
0.0
1.0
2.0
3.0
extended model for age >= 36
extended model
conventional model
20
40
60
80
100
120
Position on chromosome 15

Figure 2.9: Linkage analysis by different models and for different groups.

Appendix 2.1: Derivation of the parameters in the extended variance component model.

Both the conventional and extended variance component models considered in this paper include three variance components due to major gene (the QTL), polygene and environment.

Let us look at the variance of a quantitative trait explained by the major genetic effect first. As we described before, we consider the standard additive model for a bi-allelic QTL with alleles 'b' and 'B'. We first define the genotypic score for the individual *i*, Z_i, as follow,

$$Z_i = \begin{cases} -a, & \textit{if the genotype is } \text{b/b} \\ 0, & \textit{if the genotype is } \text{b/B} \\ a, & \textit{if the genotype is } \text{B/B} \end{cases},$$

where a is a constant. The allele frequency $P(\text{b}) = p$, and $P(\text{B}) = q$ with $p + q = 1$. We assume that the two alleles are equally likely and in Hardy-Weinberg equilibrium, which means $p = q = 0.5$, thus, $P(\text{b/b}) = p^2 = 0.25$, $P(\text{B/B}) = q^2 = 0.25$, and $P(\text{b/B}) = 2pq = 0.5$.

The variance due to the QTL effect for the individual *i* is

$$Var(Z_i) = E(Z_i^2) - (E(Z_i))^2$$

$$= \left[(-a)^2 \times p^2 + a^2 \times q^2 \right] - \left[-a \times p^2 + a \times q^2 \right]^2$$

$$= 2pqa^2$$

$$= 0.5a^2 .$$

This is actually the maximum variance from the major gene effect, since a marker is assumed to be fully informative with equally frequent alleles (Shi and Rao 2008).

The covariance due to the QTL effect between the sibling *i* and *j* is

$$Cov(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)E(Z_j)$$

$$= E(Z_i Z_j) \text{, when } \quad p = q = 0.5 .$$

As we mentioned before, π_{ij} is the proportion of genes that are identical by descent at the locus of interest for siblings *i* and *j*. We would find out $E(Z_i Z_j)$ based on the conditional probabilities given π_{ij}.

(I) If $\pi_{ij} = 0$, then the two persons will not share the same gene passed from the same parent. At this situation,

$$E(Z_i Z_j \mid \pi_{ij} = 0) = a^2 P(Z_i = -a, Z_j = -a \mid \pi_{ij} = 0) + a^2 P(Z_i = a, Z_j = a \mid \pi_{ij} = 0)$$

$$-a^2 P(Z_i = -a, Z_j = a \mid \pi_{ij} = 0) + a^2 P(Z_i = a, Z_j = -a \mid \pi_{ij} = 0)$$

$$= a^2 P(genotypes: \text{b/b \& b/b} \mid \pi_{ij} = 0)$$

$$+a^2 P(genotypes: \text{B/B \& B/B} \mid \pi_{ij} = 0)$$

$$-2 \times a^2 P(genotypes: \text{b/b \& B/B} \mid \pi_{ij} = 0) .$$

We denote b^n or B^n as the n^{th} gene with genotype b or B. Then if the parents of two

siblings have genotypes: $(b^1/b^2 \ \& \ b^3/b^4)$, then there are four possibilities for the genotypes of the two siblings: $(b^1/b^3 \ \& \ b^2/b^4)$, $(b^1/b^4 \ \& \ b^2/b^3)$, $(b^2/b^3 \ \& \ b^1/b^4)$, and $(b^2/b^4 \ \& \ b^1/b^3)$. So,

$$P(genotypes: \text{b/b \& b/b} \mid \pi_{ij} = 0)$$

$$= P(\text{b/b \& b/b} \mid Parents: \text{b/b \& b/b, and } \pi_{ij} = 0) \times P(Parents: \text{b/b \& b/b})$$

$$= 1 \times p^4 = p^4 .$$

With similarity,

$$P(genotypes: \text{B/B \& B/B} \mid \pi_{ij} = 0) = q^4 .$$

when $\pi_{ij} = 0$, two siblings with genotypes b/b and B/B would have both parents with genotype b/B. For this instance, with the parents' genotypes: $(b^1/B^2 \ \& \ b^3/B^4)$, the four possibilities for the two siblings' genotypes are: $(b^1/b^3 \ \& \ B^2/B^4)$, $(b^1/B^4 \ \& \ B^2/b^3)$, $(B^2/b^3 \ \& \ b^1/B^4)$, and $(B^2B^4/ \ \& \ b^1/b^3)$. So,

$$P(genotypes: \text{b/b \& B/B} \mid \pi_{ij} = 0)$$

$$= P(\text{b/b \& B/B} \mid Parents: \text{b/B \& b/B, and } \pi_{ij} = 0) \times P(Parents: \text{b/B \& b/B})$$

$$= (1/4) \times (2pq)^2$$

$$= p^2 q^2 .$$

Therefore,

$$E(Z_i Z_j \mid \pi_{ij} = 0) = a^2 p^4 + a^2 q^4 - 2 \times a^2 p^2 q^2$$

$$= 0 .$$

(II) If $\pi_{ij} = 0.5$, then the two persons will share only one gene passed from the same

parent. At this situation,

$$E(Z_iZ_j \mid \pi_{ij}=0.5) = a^2P(Z_i=-a, Z_j=-a \mid \pi_{ij}=0.5) + a^2P(Z_i=a, Z_j=a \mid \pi_{ij}=0.5)$$

$$-a^2P(Z_i=-a, Z_j=a \mid \pi_{ij}=0.5) + a^2P(Z_i=a, Z_j=-a \mid \pi_{ij}=0.5).$$

By the property of conditional probabilities, we could have,

$$P(Z_i=-a, Z_j=-a \mid \pi_{ij}=0.5)$$

$$= P(\text{b/b \& b/b} \mid Parents: \text{b/b \& b/B, and } \pi_{ij}=0.5) \times P(Parents: \text{b/b \& b/B})$$

$$+P(\text{b/b \& b/b} \mid Parents: \text{b/b \& b/b, and } \pi_{ij}=0.5) \times P(Parents: \text{b/b \& b/b})$$

We consider the following instance first: $\pi_{ij}=0.5$ and parents' genotypes $(b^1/b^2 \ \& \ b^3/B^4)$. Then the possible genotypes of the two siblings are: $(b^1/b^3 \ \& \ b^1/B^4)$, $(b^1/b^3 \ \& \ b^2/b^3)$, $(b^1/B^4 \ \& \ b^1/b^3)$, $(b^1/B^4 \ \& \ b^2/B^4)$, $(b^2/b^3 \ \& \ b^2/B^4)$, $(b^2/b^3 \ \& \ b^1/b^3)$, $(b^2/B^4 \ \& \ b^2/b^3)$, and $(b^2/B^4 \ \& \ b^1/B^4)$. Among the eight possible pairs of genotypes, there are two satisfy b/b and b/b. So,

$$P(\text{b/b \& b/b} \mid Parents: \text{b/b \& b/B, and } \pi_{ij}=0.5) \times P(Parents: \text{b/b \& b/B})$$

$$= (1/4) \times 2 \times p^2 \times 2pq$$

$$= p^3q.$$

By similar procedure, we have,

$$P(Z_i=-a, Z_j=-a \mid \pi_{ij}=0.5)$$

$$= 1 \times p^4 + p^3q$$

$$= p^3;$$

And,

$$P(Z_i=a, Z_j=a \mid \pi_{ij}=0.5)$$

$=1\times q^4+(1/4)\times 2\times q^2\times 2pq$

$=q^3$;

$P(Z_i=-a,Z_j=a\mid \pi_{ij}=0.5)=0$;

$P(Z_i=a,Z_j=-a\mid \pi_{ij}=0.5)=0$.

Therefore,

$$E(Z_iZ_j\mid \pi_{ij}=0.5)=a^2p^3+a^2q^3$$

$$=0.5a^2\pi_{ij}.$$

In general,

$$Cov(Z_i,Z_j)=E(Z_iZ_j)=0.5a^2\pi_{ij}.$$

(III) If $\pi_{ij}=1$, then the two persons will have exactly the same two genes passed from their parents. At this situation,

$$E(Z_iZ_j\mid \pi_{ij}=1)=a^2P(Z_i=-a,Z_j=-a\mid \pi_{ij}=1)+a^2P(Z_i=a,Z_j=a\mid \pi_{ij}=1)$$

$$-a^2P(Z_i=-a,Z_j=a\mid \pi_{ij}=1)+a^2P(Z_i=a,Z_j=-a\mid \pi_{ij}=1),$$

in which,

$$P(Z_i=-a,Z_j=-a\mid \pi_{ij}=1)$$

$$=P(\text{b/b \& b/b}\mid Parents:\text{b/b \& b/b, and }\pi_{ij}=1)\times P(Parents:\text{b/b \& b/b})$$

$$+P(\text{b/b \& b/b}\mid Parents:\text{b/b \& b/B, and }\pi_{ij}=1)\times P(Parents:\text{b/b \& b/B})$$

$$+P(\text{b/b \& b/b}\mid Parents:\text{b/B \& b/B, and }\pi_{ij}=1)\times P(Parents:\text{b/B \& b/B})$$

$$=1\times p^4+(1/2)\times 2p^2(2pq)+(1/4)\times(2pq)^2$$

$= p^2$;

And,

$$P(Z_i = a, Z_j = a \mid \pi_{ij} = 1)$$

$$= 1 \times q^4 + (1/2) \times 2q^2(2pq) + (1/4) \times (2pq)^2$$

$$= q^2;$$

$$P(Z_i = -a, Z_j = a \mid \pi_{ij} = 1) = 0;$$

$$P(Z_i = a, Z_j = -a \mid \pi_{ij} = 1) = 0.$$

Thus,

$$E(Z_i Z_j \mid \pi_{ij} = 1) = a^2 p^2 + a^2 q^2$$

$$= 0.5a^2 \pi_{ij}.$$

Now, we define the genotypic score for the individual *i*, Z_i, as follow,

$$Z_i = \begin{cases} -(a + \alpha x_i), & \textit{if the genotype is } \text{b/b} \\ 0, & \textit{if the genotype is } \text{b/B} \\ a + \alpha x_i, & \textit{if the genotype is } \text{B/B} \end{cases},$$

where the additive genetic effect is not the same for all individuals but instead is influenced by a measured covariate through a linear function. Assuming the covariate *X* and *Z* are independent, with the similar derivation as above, we could have,

$$Var(Z_i) = 0.5(a + \alpha x_i)^2 \pi_{ii},$$

where $\pi_{ii} = 1$; and,

$$Cov(Z_i, Z_j) = 0.5(a + \alpha x_i)(a + \alpha x_j)\pi_{ij}.$$

Without loss of generality, we define the variance and covariance due to the QTL as:

$$Var(Z_i) = \sigma_{mg}^2 (1 + \beta_{mg} x_{i1})^2 ,$$

and,

$$Cov(Z_i, Z_j) = \pi_{ij} \sigma_{mg}^2 (1 + \beta_{mg} x_i)(1 + \beta_{mg} x_j) .$$

We also consider the same structure of the variance and covariance due to the polygenic effect in our extend variance component model accommodating the heterogeneity of genetic effect.

Chapter 3

A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes

3.1 Introduction

The shared ancestry of chromosomes in a population results in haplotype stretches in covered by different individuals that are very similar to each other. Making use of these haplotype stretches, and thereby accounting for the correlation of alleles at nearby markers (linkage disequilibrium; LD), statistical algorithms can make inferences about unobserved alleles. To estimate a missing allele at a specific single-nucleotide polymorphism (SNP) on a haplotype, these algorithms compare flanking markers with those from other haplotypes in the sample to find appropriate "template" or *reference* haplotypes from which to make a guess about the missing allele.

Recently there has been considerable interest in the imputation of missing genotype data for the analysis of genome-wide association (GWA) studies. The availability of panels of extensively-genotyped reference samples, such as those from The International HapMap Project (HapMap) (International HapMap Consortium, 2007), has allowed for the indirect measurement of SNP genotypes that were not directly typed from a cohort or

case–control study but only typed in the reference samples. This strategy has aided the discovery of multiple loci associated with diseases (e.g. Barrett *et al*. 2008, Scott *et al*. 2007, The Wellcome Trust Case Control Consortium, 2007) or quantitative trait (Lettre *et al*. 2008, Loos *et al*. 2008, Willer *et al*. 2008). For example, in Willer *et al*. (2008), the *LDLR* (cholesterol receptor) signal was detected only after imputation was performed, since the associated variant (rs6511720) was poorly tagged in samples genotypes with the Afymetrix 500K array set (maximum $R^2 \approx 0.21$).

This imputation-based mapping protocol is a 2-step process. First, unmeasured genotypes are imputed in the GWA data. Then, imputed genotypes are tested for association with phenotypes. Multiple methods exist for imputing genotypes from population genetic data (Browning and Browning, 2007; Greenspan and Geiger 2004, Li *et al*. 2009, Scheet and Stephens 2006, Stephens and Scheet 2005); for a recent review see Browning (2008). Here we focus on the second step, testing the imputed genotypes for association with a trait of interest.

Specifically, we aim to evaluate the relative performance of several strategies for analyzing the distribution of imputed genotypes in downstream analyses. One summary of these probabilities comes from imputing a "best-guess" genotype for each individual, which corresponds to the marginal mode of the posterior distribution of the unmeasured genotype. This approach ignores the uncertainty in the imputed genotype. When imputation is accurate, the correspondence between the true and imputed genotype is strong and an analysis of the imputed genotypes might result in little loss in power

compared with the true genotypes. However, if imputation accuracy is low there may be a weak correlation between the true genotypes and the guesses, which will mask any real association between genotype and phenotype.

We also consider two approaches that attempt to account for this uncertainty. The first of these uses the mean of the distribution of imputed genotypes, which corresponds to an expected allelic or genotypic count, or "dosage", for each individual. This approach may do well, relative to using the "best guess" genotype, when there is some uncertainty about the true genotype, since it retains more of the available information, differentiating genotypes that were imputed very confidently from those that are more uncertain.

A final approach uses mixture regression models to take full advantage of the individual genotype posterior probabilities. This approach should be superior when there is uncertainty in the imputed genotypes, and information about the relationship between genotype and phenotype is not related by an average. For example, this may occur when the posterior probabilities are high for the two homozygote genotypes, and an average dosage would indicate the unmeasured genotype was a heterozygote.

We find that for most realistic settings of GWAS, such as modest genetic effects, large sample sizes and high average imputation accuracies, the strategy of regressing the phenotype on the genetic dosages provides adequate performance. In fact, for these settings, small gains from using the full mixture models are rendered negligible by the increased model complexity and associated "cost" of estimating additional parameters.

3.2 Methods

Overview

To simulate data from realistic cohort-based association studies, we first generated dense genotype data from a coalescent model. Then, conditional on these genotypes, we simulated quantitative trait data for all individuals in each cohort. To mimic the marker density from a GWA study, we masked a fraction of the SNPs and imputed these genotypes, conditional on a set of simulated reference haplotypes and the remaining observed SNPs. Finally, we performed analyses to test for association between imputed genotypes and phenotypes.

Simulations

Genotype data

For each 100 one-megabase (1 Mb) region, we simulated 10,000 chromosomes from a coalescent model that mimics LD in real data, accounts for variations in local recombination rates, and models population history consistent with the HapMap CEU and YRI analysis panels (Schaffner *et al*. 2005).

For each 1-Mb region, we then took a random subset of 120 simulated chromosomes to generate a region-specific "pseudo-HapMap". We randomly paired (assuming Hardy-Weinberg equilibrium) a random subset of 2,000 chromosomes of the remaining 9,880 chromosomes to create 1,000 diploid individuals.

For the simulated HapMap data, polymorphic sites were ascertained and thinned to match the corresponding (CEU or YRI) Phase II HapMap International HapMap Consortium (2007) marker density, allele frequency spectrum and LD patterns, resulting in $\approx$ 1,000 SNPs for each region for the panel of 120 HapMap chromosomes. Based on the thinned HapMap panel, we selected a set of 100 tagSNPs for each region that included the 90 tagSNPs with the largest number of proxies and 10 additional SNPs picked at random among the remaining tags (Carlson *et al*. 2004). The tagSNP selection approach taken above resulted in tagSNP sets that captured $\approx$ 78% of the common variants (MAF > 5%) in the simulated CEU HapMap, similar to the observed performance of the Illumina HumanHap300 Beadchip SNP genotyping platform. The genotypes at these 100 tagSNPs constituted the observed data for each simulated sample.

Quantitative trait

We generated phenotype values on each of the n individuals for a large and small sample (n = 1000, 50), conditional on their simulated genotypes. We simulated trait values separately for four genetic models, with varying degrees of dominance, and also for a null model where genotypes and phenotypes were independent.

At each SNP, the genotype label (0, 1, 2) is represented by the count of an arbitrarily chosen allele. Table 3.1 contains a summary of notation for the frequencies and genetic effect sizes ("phenotypic deviations") of each genotype. Since allele frequency affects the power to detect phenotype association, we adjust the phenotypic deviations separately for

each SNP, so that we may tabulate results over all SNPs. To accomplish this, we maintain constant genetic variance of 0.0293 (respectively 1.4874) for $n = 1000$ ($n = 50$), calculated so as to achieve approximately 90% power at type-I error of 5×10^{-5} when analyzing the simulated genotypes under an additive genetic model with equal allele frequencies of one-half. We used the following formula for genetic variance V_G (from Equation [8.8] of Falconer (1989), p. 129):

$$V_G = 2pq[a + d(q - p)]^2 + [apqd]^2 \qquad (1)$$

where p and $q = 1 - p$ are allele frequencies, and a and d are additive and dominance effects (Table 3.1).

We performed the above trait simulations for 83,327 SNPs in turn for the following genetic models: additive ($d = 0$); partially-dominant ($d = \frac{1}{2}a$); dominant ($d = a$); and over-dominant ($d = \frac{6}{5}a$), corresponding to a value for the heterozygote that is 10% greater than the difference between the two homozygotes.

To simulate trait data y_i for individual i (1, … , n) at a single SNP, we used the following model:

$$y_i^* = \mu^* + (-a)I_{\{g_i=0\}} + (d)I_{\{g_i=1\}} + (a)I_{\{g_i=2\}} + \varepsilon_i \qquad (2)$$

where g_i is the true genotype for individual i, a and d are chosen according to (1), the indicator variable $I_{\{A\}}$ is one if A is true and zero otherwise, and $\varepsilon_i \sim N(0,1)$.

Genotype imputation

To obtain posterior probabilities and imputed genotypes, we used the software package fastPHASE (Scheet and Stephens, 2006). For each simulated region, we fit the LD model to the reference chromosomes only, and then applied this fitted model to the pseudo individuals in the simulated cohort. (For convenience we set the number of haplotype clusters K to be 20.) We assess imputation accuracy with the square of the Pearson correlation coefficient between the true and best-guess genotypes (R^2), which is more informative about power at different allele frequencies than a simple genotype imputation error rate measure. For our simulations, the median R^2 for these data was 0.90 and the mean was 0.75.

Regression analysis

We used regression analysis to test the effectiveness of multiple summaries of the imputed genotypes. Let p_{ki} denote the conditional ("posterior") probabilities for the imputed genotypes of individual $i(1, \ldots, n)$, where k (0, 1, 2) indexes the genotype by its label. We evaluated the performance of the following three summaries of the genotype probabilities conditional on the observed data:

1. *best guess* — "maximum *a posteriori*";
2. *dosage* — estimated allelic or genotypic counts; and
3. *posterior probabilities* — probabilities of the 3 possible genotypes obtained from imputation.

For comparison, we also analyzed the true (simulated) genotypes.

First we give the models used for ordinary least squares (OLS) regression. Then we explain the use of mixture models for regression. For each method, we consider both *additive* (1-parameter) and *non-additive* (2-parameter) regression models for analysis. In what follows, let y_i denote the quantitative trait value for individual *i* at a SNP.

Ordinary regression on genotype imputation features

Additive. Let x_i represent a particular feature of the imputation procedure or the true genotype (g_i) at a SNP under consideration, *i.e.*

$$x_i = \begin{cases} \arg\max_{k\in\{0,1,2\}} \{p_{ki}\}, & best-guess \quad genotype \\ p_{1i} + 2p_{2i} \quad , & allelic \quad dosage \\ g_i \quad , & true \quad genotype \end{cases}$$

The additive model is written as

$$y_i = \mu + \beta x_i + \varepsilon_i, \tag{3}$$

where $\varepsilon_i \sim N(0, \sigma^2)$, independently for all *i*. We use ordinary least squares (OLS) regression to test the null hypothesis $H_0 : \beta = 0$ vs. $H_0 : \beta \neq 0$. We compute an F-statistic.

Non-additive. Under a non-additive model, we expand x_i to be comprised of two components $\left(x_i^{(1)}, x_i^{(2)}\right)$ as follows:

$$\left(x_i^{(1)}, x_i^{(2)}\right) = \begin{cases} (I_{\{x_i=1\}}, I_{\{x_i=2\}}), & best-guess \quad genotype \\ (p_{1i}, p_{2i}) \quad , & allelic \quad dosage \\ (I_{\{g_i=1\}}, I_{\{g_i=2\}}), & true \quad genotype \end{cases}.$$

We write the dominance model as

$$y_i = \mu + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \varepsilon_i, \quad (4)$$

where $\varepsilon_i \sim N(0, \sigma^2)$, as above. Again we evaluate the null hypothesis that there is no effect for any genotype, *i.e.* $H_0 : \beta_1 = 0, \beta_2 = 0$ vs. $H_0 : \beta_1 \neq 0 or \beta_2 \neq 0$. We apply OLS regression to compute an F-statistic.

Mixture of regression models

To investigate the approach of multiple-imputation, we fit a mixture of regression models to the phenotype data and posterior genotype probabilities. The composite regression model may be written as

$$y_i = \sum_{k=0}^{2} p_{ki} f_i(\mu, \beta, \varepsilon_i), \quad (5)$$

where the regression function $f_k(\cdot)$ is a function of the assumed genetic model, *i.e.* additive or non-additive (see below).

For each assumed model, we construct a likelihood ratio statistics to test for a genetic effect. To estimate the parameters (μ, β), we maximize the log-likelihood function using the Nelder-Mead Simplex Method (Nelder and Mead, 1965), implemented in the R package *optim*.

Additive. Under an assumption of additivity of the allelic effects, the regression function $f_k(\cdot)$ is

$$f_k(\mu,\beta,\varepsilon_i)=\begin{cases}\mu+\varepsilon_i & , \ k=0\\ \mu+\beta+\varepsilon_i & , \ k=1\\ \mu+2\beta+\varepsilon_i & , \ k=2\end{cases}, \qquad (6)$$

where $\varepsilon_i \sim N(0,\sigma^2)$.

To test the hypothesis $H_0:\beta=0$ vs. $H_0:\beta\neq 0$, we construct a likelihood ratio test.

Non-additive. Relaxing the assumption of additivity (allowing for dominance) of the allelic effects, we expand β to be (β_1,β_2), and the regression function $f_k(\cdot)$ is

$$f_k(\mu,\beta_1,\beta_2,\varepsilon_i)=\begin{cases}\mu+\varepsilon_i & , \ k=0\\ \mu+\beta_1+\varepsilon_i & , \ k=1\\ \mu+\beta_1+\beta_2+\varepsilon_i & , \ k=2\end{cases}, \qquad (7)$$

where $\varepsilon_i \sim N(0,\sigma^2)$. To test the hypothesis $H_0:\beta_1=0,\beta_2=0$ vs. $H_0:\beta_1\neq 0 or \beta_2\neq 0$, we construct a likelihood ratio test.

3.3 Results

Large sample size with small effects

We computed power empirically, based on the analysis of $\approx$ 1 million null data sets. Results from analysis based on our various imputation strategies and regression models, for the large sample of 1,000 individuals in the simulated studies, are reported in Table 3.2. In general, there was a consistent gain in performance achieved from using the dosage summaries or mixture models in comparison to using the best guess genotypes. This improvement was larger for the 2-parameter regression models, regardless of the

underlying genetic model, with absolute gains in power of $\approx$ 14%. For additive or 1-parameter models, the average gain was more modest (3-4%). All differences between the dosage and mixture model strategies were small (< 2%).

We also examined the effect of imputation accuracy and allele frequencies on the power to detect association (Figure 3.2). We summarized accuracy at each SNP with the square of the Pearson correlation coefficient between the imputed and true genotypes (coded as 0, 1, or 2), which we refer to as R^2.

When the accuracy is high ($R^2 > 0.9$), using the best-guess genotype from the imputation procedure results in little loss of power. The gain from using a dosage or mixture model is greatest at intermediate accuracies, since posterior probabilities are informative about the underlying genetic variation, even if they do not allow accurate best-guess imputation of genotypes. For all three strategies, at low imputation accuracies, the lines of the additive regression models converge together; so do the lines of the dominant regression models. An important factor in overall power summaries, such as those in Tables 3.2 and 3.3, is the allele frequency distribution of SNPs present in the reference panel, at which genotypes are being imputed in the study samples, since the tables are constructed with averages over all SNPs. In Figures 3.2(c) and 3.3(c), where phenotypes were simulated from an additive genetic model, powers of all regression models increase substantially when minor allele frequencies are relatively low. This may reflect the relative difficulty of accurate imputation at SNPs with a lower MAF. (Under the correct additive model, power for the true genotypes is unaffected, since, we attempted to make power

independent of allele frequency for the purposes of making general comparisons among analysis strategies; see Methods.) For data simulated under a dominant genetic model, the powers of different regression models are much greater for SNPs with low to moderate allele frequencies of the dominant allele. Methods that assume the correct dominant model for analysis are superior at a greater range of allele frequencies.

Small sample size with large effects

For SNPs with modest genetic effects, as above, there is little gain from the increased computational demands of applying mixture models for the analyses. To examine a scenario where the mixture models might offer an advantage, we repeated the above simulations with larger genetic effects (and thus smaller sample sizes so that power was below 100%). This situation might be found at expression quantitative trait loci (eQTL) mappings. These results are in Table 3.3. Here, the advantage of applying mixture models is apparent, with average power gains of 10-12%. The contrast is greater at lower imputation accuracies (top row of Figure) and is maintained even when we applied the incorrect additive regression model to data simulated with a strong genetic effect (Figure 3.3(b)).

3.4 Discussion

Several software packages have been developed to impute and test SNPs that were not typed directly, such as BIMBAM (Servin and Stephens, 2007), IMPUTE (Marchini *et al*. 2007), Mach (Li *et al*. 2006) and Beagle (Browning and Browning 2009). Two of these

methods (BIMBAM and IMPUTE) assess association between genotype and phenotype with a Bayes Factor. Here we do not consider the Bayesian approach, which is discussed by Guan and Stephens (2008).

Multiple factors will impact power of imputation-based strategies for the analysis of GWAS, including differences in the patterns of LD and allele frequencies between the study and reference populations. However, for the single-marker analyses examined in our study, the impact of these factors can be measured via their effect on imputation accuracy, since the missing (unmeasured) genotypes are the quantities of interest for analysis.

Here, we have made no attempt to model the correlation of genotypes among SNPs during analysis. To detect interactions among genotypes at nearby SNPs, it may be beneficial to model this dependence during imputation and analysis. The imputation procedures mentioned above may obtain correlated genotypes by sampling entire chromosomes of untyped SNPs, instead of the data at each SNP, marginally.

It may be possible to do better in such a setting by using genuine "multiple imputation" methods. However, in our setting, by applying a mixture of regression models, we hope to capture a range of possible phenotype–genotype relationships, and the gain from multiple imputation over the mixture model should not be large. Therefore, we felt that the mixture model provided a close approximation to an optimal analysis procedure.

In our most relevant comparisons with modest effects and large sample sizes, use of the dosage summaries was as powerful as using the mixture model methods, at a fraction of computational cost. The exception to this result is apparent only at SNPs with very large genetic effects. In such situations of large effects, most methods will be effective at detecting an association. This difference is most pronounced at poorly imputed SNPs. Overall, use of the dosage quantities appear to be effective and efficient to account for the uncertainty in the imputed genotypes.

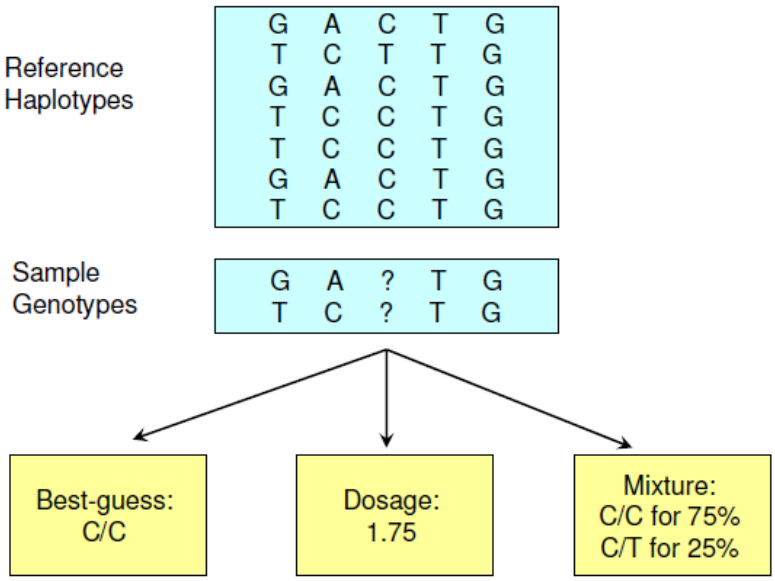


Figure 10.1: A didactic figure illustrating the three strategies.

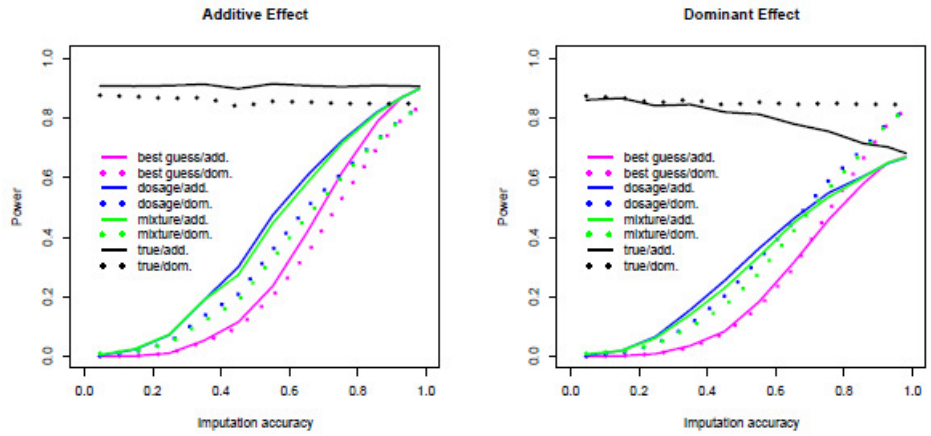
	Genotype		
	A/A	A/a	a/a
Labels	0	1	2
Frequencies	q^2	$2pq$	p^2
Phenotypic deviation	$-a$	d	a

Table 3.1: Genotype and phenotype values. Genotype labels are the counts of an arbitrarily chosen allele.

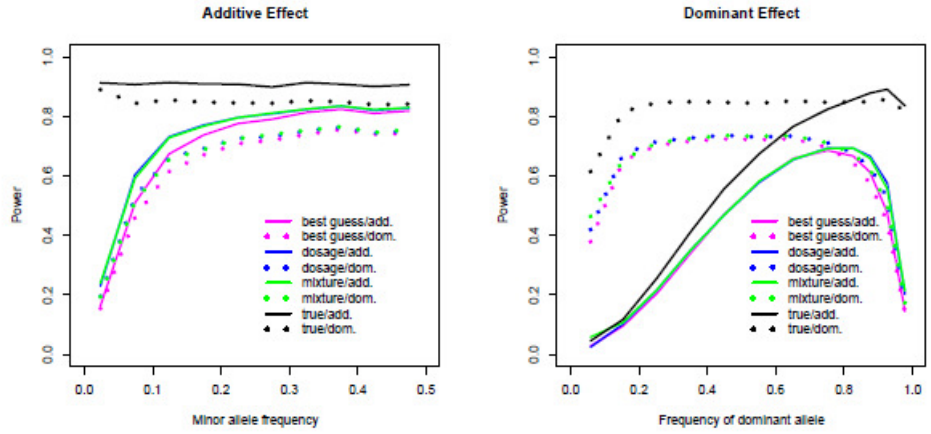
The phenotypic deviations are the deviations from the mean μ^* in expression (2) used in the simulation, and vary by SNP. (This table is adapted from Table 7.3 of Falconer, 1989, p. 121.)

Analysis Strategy	Simulated Trait			
Imputation summary/ Regression model	Additive ($d = 0$)	Partially dominant ($d = \frac{1}{2}a$)	Dominant ($d = a$)	Over-dominant ($d = \frac{6}{5}a$)
best-guess/additive	.635	.599	.478	.435
best-guess/dominance	.466	.463	.448	.449
dosage/additive	.660	.620	.489	.447
dosage/dominance	.603	.598	.588	.588
mixture/additive	.668	.628	.499	.456
mixture/dominance	.604	.600	.587	.588
True/1 df	.897	.865	.730	.683
True/2 df	.708	.711	.706	.709

Table 4.2: Power results for small effects and large sample size.
The "Analysis Strategy" specifies the combination of imputation quantity/summary (e.g. best guess, dosage, or mixture model) and whether the regression model allows for deviations from a strict additive model. Results are based on a cohort of 1,000 individuals. Power was computed at a fixed type-I error rate (α) of 5×10^{-5}, based on empirical quantiles from analysis of 916,597 "null" data sets, with a trait simulated independent of genotype. Quantitative traits were simulated to have constant genetic variance of 0.0293.



(a) Power vs. R^2 with an additive effect (b) Power vs. R^2 under complete dominance



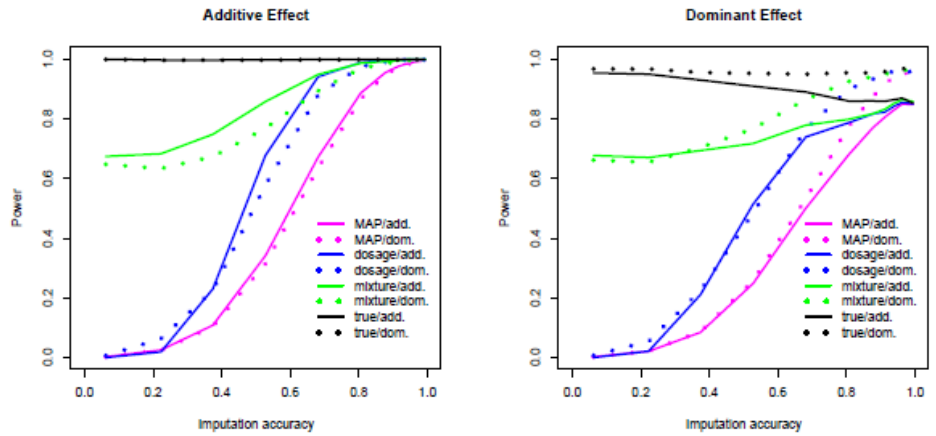
(c) Power vs. frequency of minor allele with an additive effect (d) Power vs. frequency of dominant allele under complete dominance

Figure 5.2: Power vs. accuracy and allele frequency for large sample size and small effects.

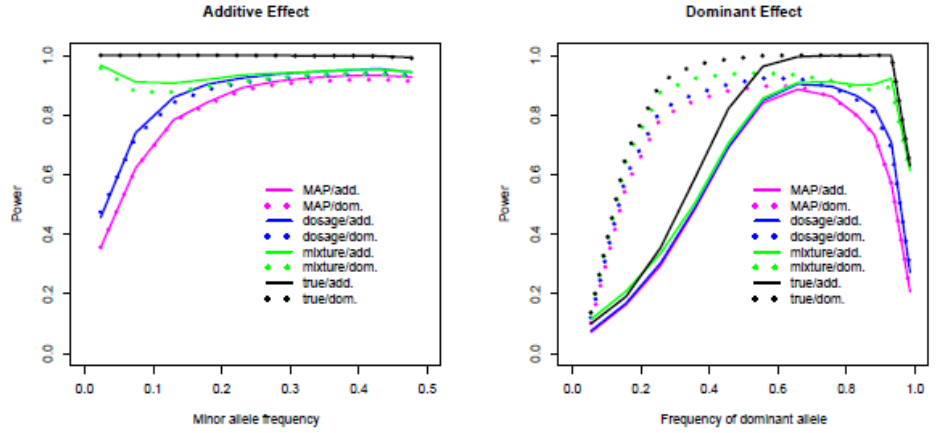
For each summary and the true genotypes, both an additive (solid line) and dominant (dotted line) model were analyzed. Figures A and C are based on data simulated with an additive effect; Figures B and D are based on data simulated under a model of complete dominance. Power was computed at a fixed type-I error rate (α) of 5×10^{-5}. The sample size was 1000. TOP: Power is plotted against R^2, a measure of imputation accuracy. BOTTOM: Power is plotted against allele frequency.

Analysis Strategy	Simulated Trait			
Imputation summary/ Regression model	Additive ($d = 0$)	Partially dominant ($d = \frac{1}{2}a$)	Dominant ($d = a$)	Over-dominant ($d = \frac{6}{5}a$)
best-guess/additive	.701	.688	.582	.546
best-guess/dominance	.682	.670	.629	.636
dosage/additive	.755	.743	.629	.590
dosage/dominance	.745	.736	.702	.707
mixture/additive	**.850**	**.837**	.721	.686
mixture/dominance	.829	.828	**.805**	**.810**
Truth/1 df	.916	.911	.802	.767
Truth/2 df	.913	.910	.873	.890

Table 6.3: Power results for large effects and small sample size.
The "Analysis Strategy" specifies the combination of imputation quantity/summary (e.g. best guess, dosage, or mixture model) and whether the regression model allows for deviations from a strict additive model. Results are based on a cohort of 50 individuals. Power was computed at a fixed type-I error rate (α) of 5×10^{-5}, based on empirical quantiles from analysis of $\approx$ 1 million "null" data sets, with a trait simulated independent of genotype. Quantitative traits were simulated to have constant genetic variance (see Methods), given the genetic model and allele frequencies at each SNP.



(a) Power vs. R^2 with an additive effect (b) Power vs. R^2 under complete dominance



(c) Power vs. frequency of minor allele with an additive effect (d) Power vs. frequency of dominant allele under complete dominance

Figure 7.3: Power vs. accuracy and allele frequency for small sample size and large effects.

Power was computed at a fixed type-I error rate (α) of 5×10^{-5}. The sample size was 50. For each summary and the true genotypes, both an additive (solid line) and dominant (dotted line) model were analyzed. Figures A and C are based on data simulated with an additive effect; Figures B and D are based on data simulated under a model of complete dominance. TOP: Power is plotted against R^2, a measure of imputation accuracy. BOTTOM: Power is plotted against allele frequency.

Chapter 4

Locate Complex Disease Susceptibility Loci by Investigating Gene and Environment Interaction for Genome-Wide Association Studies

4.1 Introduction

Susceptibility to most of complex diseases is influenced by a combination of genetic factors, environmental factors, and interactions between them. Understanding the interaction of genes and environment will lead us to new methods of disease detection and prevention. If we know how genetic variation would cause people to respond differently to a drug, then the drug treatment could be made safer and more effective. For example, current treatment guidelines for coronary artery disease prevention require risk stratification of the patient. Quantification of the patient's coronary artery disease risk guides the intensity of evidence-based drug treatment of modifiable risk factors (Lanktree and Hegele 2009). Studying the gene-environment interaction could also strengthen the associations between environmental factors and diseases by examining these factors in genetically susceptible individuals (Hunter 2005). Failure to analyze genetic and environmental factors together would weaken the observed associations between a true

risk factor and disease occurrence, when susceptible and non-susceptible persons are mixed (Khoury *et al*. 1988, Khoury and Wacholder 2009).

In general, interactions are differences in the strength of association between a gene and phenotype based upon the presence of or quantitative variations in another factor. The additional factor could be an environmental factor, behavior quantity, or another genetic variant (for example, genotype at another locus). The multiplicative interaction between genetic and environmental factors is often investigated for detecting the disease susceptibility loci. Usually, the multiplicative interaction is described as that an association of gene and environment in diseased subjects (cases) is different than that in healthy subjects (controls). Mechanistically, an interaction could be that the direction and magnitude of the genetic effect differs according to environmental exposure. A classical example is phenylketonuria (PKU), a human genetic condition caused by mutations to a gene coding for the liver enzyme phenylalanine hydroxylase. If newborns are put on a special, phenylalanine-free diet right away and stay on it, they avoid the severe mental retardation that typically results from PKU (Baker 2004). An alternative mechanistic explanation is that genetic factors might modify the effect of an environmental exposure on disease risk (Kraft *et al*. 2007). An example is the interaction between sunlight exposure and skin color: sunlight exposure has a much stronger influence on skin cancer risk in fair-skinned humans than among individuals with an inherited tendency to darker skin (Green and Trichopoulos 2002). Statistically, the two mechanistic routes are indistinguishable.

During the current decade, increasing efficiency and decreasing cost of genotyping and improved statistical methods have made genome-wide associate (GWA) studies the method of choice for localizing common susceptibility variants. Risch and Merikangas, in 1996, showed that GWAS can have high power to identify alleles with modest effects. In GWAS, the marginal genetic effect is usually tested in order to detect disease genes. However, we still do not know the best way to locate complex disease susceptibility loci by exploiting the gene-environment interaction, especially in the GWAS framework, in which hundreds of thousands markers are scanned. In the context of GWAS, multiple comparisons would be considered and even the hypotheses might possibly be different than those in traditional contexts. For example, we might be more interested in whether there is at least one locus showing significant evidence of the gene-environment interaction, rather than which specific locus shows significant evidence of the gene-environment interaction.

The standard model for testing the gene-environment interaction is to evaluate a logistic regression model, with genotypic status, exposure status, and an interaction term between them as covariates. Another approach to detect gene-environment interaction is to use in case-only data and test whether genotypic status and exposure status are correlated among cases; this approach assumes that the two are independent at the population level and the disease is rare. Recently, Mukherjee and Chatterjee (2008) proposed an empirical Bayes-type shrinkage estimator to test the relationship between gene and environmental factors. The estimator is a weighted average of the case-only and case-control estimators of the logarithm of the interaction. This estimator balances the bias of case-only estimator

and the inefficiency of case-only estimator. Murcray *et al*. (2009), developed a 2-step method. They examined the marginal correlation of gene and environment first and then used the traditional logistic model as the second step at markers selected by the first step. This approach is generally more powerful than the traditional logistic regression-based approach.

In this paper, we describe a new powerful method for identifying gene-environment interactions based on likelihood ratio tests, which models the interaction of genetic and environmental factors among cases and controls under the assumption that gene and environment are independent at the population level. We also compare its performance to above existing approaches in the setting of large scale association studies by different definitions of powers and type I errors, as well as ranked p-values. We show that our approach provides great gains in power (compared to all alternative approaches) when the trait being studied is common and performs similarly to the case-only approach when the trait is rare. When the departure to the assumption of gene-environment independence is modest, which closes to the reality, our new method still performs best in terms of the empirical power.

4.2 Methods

The traditional logistic regression model is:

$$logit\left[P(D=1\mid g,e)\right]=\beta_0+\beta_g G+\beta_e E+\beta_{ge}GE, \qquad (1)$$

where D is the disease status for each individual, coded as 1 for affected and 0 for

unaffected, E is the exposure status with 1 for exposed and 0 for unexposed. For simplicity and demonstration purposes, we assume a binary coding of the genotype such that G = 1 means carriers of at least one risk allele and G = 0 means non-carriers. $\beta_g = log(OR_{g|E=0})$ is the natural logarithm of the odds ratio between disease status and genotypes among individuals with E = 0, $\beta_e = log(OR_{e|G=0})$ is the natural logarithm of the odds ratio between disease status and exposure category among individuals with G = 0, and $\beta_{ge} = log(OR_{g|E=1} / OR_{g|E=0})$ is the natural logarithm of the ratio of the genetic odds ratios comparing exposed to unexposed subjects. If $\beta_{ge} = 0$, the odds ratio between disease status and genotypes for exposed people is the same as that for unexposed people, and there is no multiplicative gene-environment interaction at the tested marker. In the GWAS, the null hypothesis $H_0 : \beta_{ge} = 0$ would be tested for each marker with a one-degree-of-freedom likelihood ratio test. The power would be corrected for multiple comparisons by Bonferroni criterion.

Under the assumption that genetic effect and environment factor are independent at the population level and the disease is rare, the case-only method estimates the interaction with better precision than the traditional logistic regression (Piegorsch *et al*. 1994; Khoury and Flanders, 1996), which means the variance of the estimator is smaller than that from the standard logistic model in case-control studies. The case-only method uses a simpler logistic regression model on diseased individuals only:

$$logit\left[P(E = 1 \mid g)\right] = \gamma_0 + \gamma_g G . \qquad (2)$$

In a GWA study, the one-degree-of-freedom likelihood ratio test for testing $H_0 : \gamma_g = 0$

could be performed at each marker within the case group, and again must be adjusted for multiple comparisons.

Murcray *et al*., in 2009, developed a 2-step test. At Step 1, they performed a likelihood-ratio test for the association between gene and environement at all *M* SNPs, based on the logistic regression model (2). The subset of *m* SNPs that exceed a given significance threshold (*i.e.* with p-value < α_1) for the test of $H_0 : \gamma_g = 0$ would be analyzed at Step 2. Similar to the case-only method, this step also assumes that gene and environment are independent at population level and the disease is rare. At Step 2, the *m* SNPs that passed Step 1 are tested by the traditional test of gene-environment interaction based on model (1). These two steps are proved to be independent, therefore, this 2-step method maintains correct type I error rates and is robust to the assumption of gene-environment independence. They demonstrated that this method generally show higher powers than the logistic model in various situations

Mukherjee and Chatterjee (2008) proposed an empirical Bayes-type shrinkage estimator to analyze case-control data. The estimator is a weighted average of the case-only and case-control estimators of the logarithm of the interaction. When the difference between the case-only and case-control estimators is bigger, the weight of the case-control estimator is larger. When the variance of case-only estimator is bigger, the weight of case-only estimator is larger. This empirical Bayes estimator methodis given as:

$$\hat{\beta}_{EB} = \frac{\hat{\sigma}_{CC}^2}{(\hat{\tau}^2 + \hat{\sigma}_{CC}^2)} \hat{\beta}_{CO} + \frac{\hat{\tau}^2}{(\hat{\tau}^2 + \hat{\sigma}_{CC}^2)} \hat{\beta}_{CC} .$$

Here, $\hat{\beta}_{CO}$ and $\hat{\beta}_{CC}$ are the case-only and case-control estimator, respectively, $\hat{\tau}^2$ is square of the difference between these two estimators. $\hat{\sigma}^2_{CC}$ is the variance of case-control estimator. The authors use Wald test statistics for testing the gene-environment interaction.

Our new likelihood-based approach tests the interaction between gene and environment by likelihood ratio test statisteics, exploiting the gene-environment independence at underlying population. We assume gene-environment independent at population level, but we do not require rare disease assumption. Table 4.1 gives the data structure, which is a 2x2x2 contingency table. The core of the log likelihood function of the table is

$$l = \sum_i n_i \times \log(p_i) + \sum_j m_j \times \log(q_j).$$

Let D, G and E denote variables of disease status, genotype and environmental factor, respectively. The disease prevalence is denoted as $f = P(D)$. At the population level, we have

$$P(G,E) = P(G,E \mid D) \times P(D) + P(G,E \mid \bar{D}) \times P(\bar{D}) = p_1 f + q_1(1-f),$$

$$P(G,\bar{E}) = P(G,\bar{E} \mid D) \times P(D) + P(G,\bar{E} \mid \bar{D}) \times P(\bar{D}) = p_2 f + q_2(1-f),$$

$$P(\bar{G},E) = P(\bar{G},E \mid D) \times P(D) + P(\bar{G},E \mid \bar{D}) \times P(\bar{D}) = p_3 f + q_3(1-f), \text{ and}$$

$$P(\bar{G},\bar{E}) = P(\bar{G},\bar{E} \mid D) \times P(D) + P(\bar{G},\bar{E} \mid \bar{D}) \times P(\bar{D}) = p_4 f + q_4(1-f).$$

Under the assumption of independence of gene and environment,

$$P(G,E) \times P(\bar{G},\bar{E}) = P(G,\bar{E}) \times P(\bar{G},E).$$

The null hypothesis of no gene-environment multiplicative interaction in our method is

equivalent to $H_0 : OR_{ge|d} = OR_{ge|\bar{d}}$, which is exactly the same as $\frac{p_1 p_4}{p_2 p_3} = \frac{q_1 q_4}{q_2 q_3}$, corresponding to $H_0 : \beta_{ge} = 0$ for model (1).

Thus, under the null, we have

$$\begin{cases} p_1 + p_2 + p_3 + p_4 = 1 \\ q_1 + q_2 + q_3 + q_4 = 1 \\ P(G,E) \times P(\bar{G},\bar{E}) = P(G,\bar{E}) \times P(\bar{G},E), \\ \frac{p_1 p_4}{p_2 p_3} = \frac{q_1 q_4}{q_2 q_3} \end{cases}$$

with four unknown parameters (the rests are nuisance parameters). Under the alternative,

$$\begin{cases} p_1 + p_2 + p_3 + p_4 = 1 \\ q_1 + q_2 + q_3 + q_4 = 1 \\ P(G,E) \times P(\bar{G},\bar{E}) = P(G,\bar{E}) \times P(\bar{G},E) \end{cases},$$

with five unknown parameters in total. Thus, we can carry out a one-degree-of-freedom likelihood ratio test for the gene-environment interaction at each SNP, corrected by Bonferroni criterion.

We performed simulations to compare the above five methods: traditional logistic regression model, the case-only method, 2-step method by Murcray *et al*., empirical-Bayes estimator and our new likelihood-based approach. In the simulations, we adopted the parameter settings similar to those studied by Murcray *et al*. (2009), based on the model (1). For each of 500 replicate data sets, we simulated 500 cases and 500 controls, M = 10,000, 25,000, or 50,000 independent markers for each individual, including one true disease susceptibility locus and the rest independent of disease status. Minor allele

frequency at the disease locus was set to $q_A = 0.1, 0.2, \text{or } 0.25$; distributions of minor allele frequency at null markers was set to $q \sim Uniform\ (0.05, 0.5)$; disease prevalence $p_d = 0.05, 0.1, 0.3 \text{ or } 0.5$; and environmental exposure frequency $p_E = 0.1, 0.25, \text{or } 0.5$. We considered different combinations of main effects and interaction in the model (1): $R_g = exp(\beta_g)$, $R_e = exp(\beta_e)$, and $R_{ge} = exp(\beta_{ge})$.

To examine the sensitivity of all methods when the gene-environment independence at population level assumption is violated, we also simulated the situations in which genetic and environmental factors are correlated at a small portion of markers. We define $p_{ge} = 0.01 \text{ and } 0.05$, as the probability of gene-environment association at a given null marker in the population. $p_{ge} = 0$ indicates that gene and environment are independent at all simulated markers. If the marker is not independent to exposure status, the population marker-exposure odds ratio ($exp(\theta_{ge})$) would be simulated as 1.1, 1.2, 1.5 or 2.0.

In this paper, we compared different methods by different definitions of measurements as follow:

1. *Experiment-wise power and type I error*. Power is calculated as the proportion of the total 500 replicates in which the disease susceptibility locus was detected at $p < 0.05/M$, where M is the total number of SNPs tested. For the 2-step method, the criterion is $p < 0.05/m$ at Step 2, where m is the number SNPs selected by the first step. The type I error rate is estimated as the proportion of the total 500 replicates in which at least one of the

null markers is found significant after Bonferroni correction for multiple comparisons. Both criteria are introduced by Murcray *et al*. 2009.

2. *Top selections*. The proportion of the total 500 replicates in which the disease susceptibility locus is among the top 10 or top 25 most significant SNPs is calculated based on the ranked p-values.

3. *Empirical power*. Among all null markers ($M \times 500$ in total), the empirical cutoff is defined as the value that makes exactly 5% of total 500 replicates (*i.e*. 25 replicates here) in which at least one null marker is detected significant after adjusting for multiple comparisons. The empirical power is then estimated based on this empirical cutoff. The purpose of the empirical power is to control the corresponding type I error at exactly 0.05.

4. *Integrated type I error and power*. Mukherjee *et al*. (2008) introduced these criteria. They evaluated average power and type I error rate for different tests for interaction under some distributions for the genotype-exposure odds-ratio parameters that were likely to hold in large-scale association studies, instead of assuming a fixed value of that parameter

5. *Tabulated type I error*. This type I error is the proportion of significant markers with $p < 0.05$ among all simulated null markers across all replicates ($(M-1)$ SNPs in each of 500 replicates totally).

4.3 Results

Figure 4.1 shows power comparison for the traditional logistic model, the case-only approach, the 2-step approach, the empirical Bayes estimator method and our new likelihood-based approach across different interaction effect sizes R_{ge}. The powers of all methods increase as the interaction effect size increases. The power is at nominal level where there is no interaction, while powers approach 1 when the interaction is very large (say, $log(5)$). At intermediate interaction effect sizes, the likelihood-based approach and the case-only method both have highest powers. The power of the logistic model is consistently low. The empirical Bayes estimator method performs similarly as the 2-step method at this framwork.

In comparison to the basic setting of parameters: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$, various alternatives are considered at Table 4.2, Table 4.3 and Table 4.4, with experiment-wise powers, experiment-wise type I errors, and empirical powers, respectively. These tables need to be assessed simultaneously. For example, under the base model setting, the experiment-wise power is 0.330 for the logistic model, while the type I error rate is 0.048. Based on the empirical threshold controlling the type I error as exactly 0.05, the empirical power becomes slightly higher, 0.334. Overall, all methods seem maintain correct type I error rates with slight variation in Table 4.3. The power of likelihood-based approach is always highest under these settings and the traditional logistic models have the lowest powers for most

situations. In many situations, the power of the likelihood-based approach is more than two times greater than the power of the logistic model. All methods have relatively high powers for common exposures and alleles. We notice that in Table 4.3, the experiment-wise type I error rate of the empirical Bayes estimator method is always very small. For example, under the basic setting of parameters, the experiment-wise type I error rate is as low as 0.008, while the type I error rates range between 0.036 and 0.050 for the other four methods. We then plotted empirical power comparisons in Figure 4.2 with the same settings as Figure 4.1. By comparing both figures, we find that the empirical power of the empirical Bayes method is generally little higher than that of the 2-step method.

When the disease prevalence $p_d = 0.05$, the case-only method always performs well. However, as disease prevalence increases, the power of the case-only method drops very quickly (Table 4.2, Table 4.3), since one of the assumptions of case-only method (rare disease) is violated. At the extreme case of disease prevalence 0.5, the power of the likelihood-based approach becomes similar as the logistic model, whereas the powers of 2-step method and case-only method are very low. To further exam the performance for a common disease, we plotted experiment-wise powers and empirical powers of different models upon varying interaction effect sizes for disease prevalence 0.2 (Figure 4.3 and Figure 4.4, respectively). The power line of the case-only method overlaps that of the 2-step method and both are intermediate to the other two methods. The empirical power line of the empirical Bayes estimator method is very close to that of the logistic regression model.

Figure 4.5 displays the percentage of replicates in which the disease susceptibility locus would be picked up among the top 10 most significant markers upon different interaction effect sizes; findings mirror those in Figure 4.1 for power comparisons. When the disease prevalence is as low as 0.05, the case-only method performs as well as the likelihood-based approach, while the five methods perform more similarly in Figure 4.5 compared with performance in power comparison in Figure 4.1. However, when the disease prevalence is 0.2, the case-only method is worse than the new method, and the 2-step method is even worse than the logistic model in terms of power to select the disease susceptibility locus by the top-10-ranked p-values (Figure 4.4). These results are confirmed in Table 4.5. The likelihood-based approach is the most robust over all settings of parameters. For example, when the exposure prevalence p_E = 0.1, there are still 73.2% of replicates where the disease susceptibility locus is among the top 10 selection, although the power of the likelihood-based approach dropped down to 0.262 (Table 4.2). Although the power line of the empirical Bayes estimator method is lowest in Figure 4.3 and Figure 4.4, the ability to pick up the disease susceptibility locus of the empirical Bayes estimator method is better than that of logistic regression model and the 2-step method in Figure 4.6.

Since the two methods with generally higher powers both exploit the assumption of gene and environment independence at the population level, we did simulations to exam the performance of the methods when this assumption is violated. We denote p_{ge} as the probability that one specific marker is correlated with an environmental factor. We also

define θ_{ge} as the natural logarithm of odds ratio between gene and environment in the population, if the marker is associated with the environmental factor. As expected, the experiment-wise type I errors of case-only method and the likelihood-based approach are inflated at all situations (Table 4.6). The stronger the gene–environment association, the greater the inflation of type I error rates for both methods. For example, when $exp(\theta_{ge})$=1.5, the type I errors of the new method are as high as 0.370 and 0.884 when p_{ge}=0.01 and p_{ge}=0.05, respectively. The experiment-wise powers are meaningless in these circumstances. However, we could investigate the empirical powers at Table 4.7 instead. By controlling the adjusted type I error at exactly 0.05, the empirical powers of the likelihood-based approach are 0.640 and 0.414 for the case of p_{ge}=0.01 and p_{ge}=0.05, respectively, which are still relatively higher.

In the simulations above, we fixed the value of the key parameter θ_{ge}, which describes the degree to which the gene is associated with environment at the population level. Mukherjee *et al*. (2008) introduced an idea that the value of θ_{ge} is distributed as a normal distribution, instead of a fixed value. This scenario is more like realistic situation in GWAS, since there is possibility that at a small portion of markers, not all markers over the genome, the gene and environment might be associated to various degrees. For example, when p_{ge}=0.1 and $\theta_{ge} \sim N(0, log(1.2)/2)$, there are 90% of markers are independent of the environment, and for 95% of the rest 10%, the GE odds ratio is between $-log(1.2)$ and $log(1.2)$. Table 4.8 lists integrated power and type I error for different parameter settings, as well as the corresponding empirical power. Again, when

the departure from the gene-environment independence assumption is modest, our new likelihood-based approach still has relatively high empirical powers. For example, when 85% of markers are independent of the environment, and for 95% of the rest 15% markers, the gene-environment odds ratio is between $-log(1.3)$ and $log(1.3)$, the empirical power of the likelihood-based approach is 0.768, which is much higher than that of the 2-step method and the empirical Bayes estimator method. When the gene-environment independence assumption does not hold, the 2-step method and the empirical Bayes estimator method perform similarly. When the departure to the assumption is small, the empirical Bayes estimator method has a little higher power, whereas the 2-step is a little better when the departure is large.

Finally, we compared the power to select the disease susceptibility locus for all methods when the gene-environment independence assumption was not satisfied (Table 4.9). The likelihood-based approach has very robust performance at all situations, despite highly inflated experiment-wise type I errors at some cases.

4.4 Discussion

In the context of GWAS, gene-environment interaction is still a new frontier. In this paper, we proposed a new likelihood-based approach to test for the gene-environment interaction, and compared its performance by simulation studies to the traditional logistic regression model, a case-only method (Piegorsch *et al*. 1994; Khoury and Flanders, 1996), the 2-step method (Murcray *et al*. 2009) and the empirical Bayes-type shrinkage

estimator method (Mukherjee and Chatterjee 2008), through various criteria and under a range of settings of parameters. The likelihood-based approach has increased power compared to the other methods when the gene-environment independence assumption is satisfied. When the assumption is slightly violated, the likelihood-based approach would have inflated type I error rate. However, the empirical power of the likelihood-based approach is still relatively high under modest departure from independence assumption. For example, the probability of a marker associated with environment is less than 0.05 and the gene-environment odds ratio at population level for the associated marker is less than 1.5. These are very realistic situations, suggested by Liu *et al*. (2004), that violation of gene-environment independence would likely to be modest in most situations when it occurs.

In our simulation studies, the traditional logistic model had poor performance as expected. It maintains theoretically correct type I error rates under all circumstances considered, whereas the power is influenced greatly by the numbers of samples within each cell in the data structure (Table 4.1).

The case-only method assumes gene-environment independence at the population level and that the disease is rare. The rare disease assumption for all levels of both genetic and environmental exposures might not be always valid. Schmidt and Schaid (1999) describe a situation where the marginal probability of the disease may be small in the population but high for certain subgroups. Our simulations show that when the disease is comparatively common (prevalence greater than 0.1), both power and the ability to select

the disease susceptibility locus decreases substantially, although the type I error rates remain at their nominal levels.

The 2-step method proposed by Murcray *et al*. (2009) conducts a screening test at the first step, which reduces the number of multiple comparisons at the independent second step reduced. Therefore, it has higher power than the traditional logistic regression model. Similar to the case-only method, the first step of the 2-step method also assumes gene-environment independence and rare disease, which is the reason why the 2-step method performs much worse at common disease situations in our simulation studies. Since the two steps are proved to be un-correlated, the 2-step method maintains the correct type I error rates.

Mukherjee and Bhatterjee (2008) described a novel empirical Bayes-type shrinkage estimator to detect gene-environment interaction. This strategy balanced the bias and efficiency, since the case-control estimator is always unbiased and the case-only estimator would be much more efficient when the gene-environment independence and rare disease assumption holds. When genetic and environmental factors are independent at the underlying population, this method would exploit this assumption to get higher power; when the assumption does not hold, this method would ensure lower false positives.

Our new likelihood-based method exploits the entire information of the independence between genetic and environmental factors in general population but does not require the

rare disease assumption, which makes its power higher than other methods. This assumption has been exploited in many methods, for example, Self *et al*. 1991, Hwang *et al*. 1994, Piegorsch *et al*. 1994, Umbach and Weidinger 1997, Modan *et al*. 2001, Chatterjee and Carroll 2005, Kraft *et al*. 2007. The gene-environment independence assumption is reasonable for "randomized exposures" (such as treatments assigned in a randomized trial), and for external environment agents such as carcinogens from a nearby chemical factory (Chatterjee and Carroll 2005). In some situations, genotype and exposure may co-vary according to other factors (such as ethnicity), however, this assumption might not be valid (Umbach and Weidinger 1997).

In the paper, we also evaluated the sensitivity to the violation of the gene-environment independence assumption for all methods under several parameter settings, either fixing the value of gene-environment odds ratio or considering it as normal distributed partially. The later scenario is similar to the prior distribution of a key parameter in Bayesian analyses. Apparently, the "prior" of mixed normal distribution is more close to real situation than the "prior" as a constant, especially in a GWA study. Then the corresponding power and type I error could be treated as a weighted average of those values with weights obtained from the specified mixture distribution (Mukherjee *et al*. 2008).

When the experiment-wise type I errors could not be restricted as 0.05, for example, type I errors are inflated at some situations, we are more interested in the empirical powers rather than the experiment-wise power. The empirical power is calculated by controlling

the adjusted type I error at exactly 0.05, *i.e.*, the false positives are controlled. On the basis of empirical powers, the performance of different models could be compared fairly. In a real situation, there might not be enough null markers to calculate the empirical threshold. We could get the empirical cutoffs through the null distribution developed by permutation or bootstrap resampling.

When the departure to the gene-environment interaction assumption is large, we could try to find definable homogeneous strata in which the assumption would hold or perhaps use recently developed approaches (for example, genomic control based methods (Devlin *et al*. 2001) or principal component of ancestry covariates (Alexander *et al*. 2009)) to control for differences in genetic background among samples.

In this paper, we conducted many different definitions of powers and type I errors to perform the model comparisons. We would like to investigate two definitions of type I errors more here. The corresponding hypotheses of the experiment-wise type I error and the tabulated type I error (Table 4.10) are different. Under the GWA study framework similar to our simulations, the null hypothesis of the experiment-wise type I error is that there is no any marker showing gene-environment interaction for the whole set of markers; the alternative hypothesis is that there is at lease one null marker found significant after the multiple comparisons correction. These hypotheses are reasonable for a GWA study. In contrast, the null hypothesis of the tabulated type I error is that for a specific marker, there is no evidence of gene-environment interaction, while the alternative is that this specific marker shows significant evidence of gene-environment

interaction. This is the reason why we do not need to adjust the cutoff when estimating the tabulated type I error. The integrated type I error described by Mukherjee *et al*. (2008) is corresponding to the tabulated type I error here. For example, when p_{ge} = 0.2 and $\theta_{ge} \sim N(log(2)/2)$, the tabulated type I error in Table 4.10 is 0.0740 at level of 0.05; while the corresponding integrated type I error at Mukherjee *et al*. (2008) is 0.070 for case-only method.

We could easily extend our method to accommodate multi-genotypes other than dichotomous genotypes by extending the contingency table 4.1 to bigger dimensions. The environmental factor E here could be an external factor, a behavior quantity, or another genetic variant. The extension to the continuous phenotypes is also possible. Our method provides a new way to screen the disease susceptibility loci after main genetic effects has been tested in GWAS.

	Case (D = 1)		Control (D = 0)	
	E = 1	E = 0	E = 1	E = 0
G = 1	p_1 (n_1)	p_2 (n_2)	q_1 (m_1)	q_2 (m_2)
G = 0	p_3 (n_3)	p_4 (n_4)	q_3 (m_3)	q_4 (m_4)

Table 11.1: Data structure for case-control studies.
D is the disease status for each individual, valued as 1 for affected and 0 for unaffected, *E* is the exposure status with 1 for exposed person and 0 for unexposed. *G* = 1 means carriers of at least one risk allele and *G* = 0 means non-carriers. $p_i's$ and $q_i's$ are the probability of each cell, $n_i's$ and $m_i's$ are the counts for each sub-group.

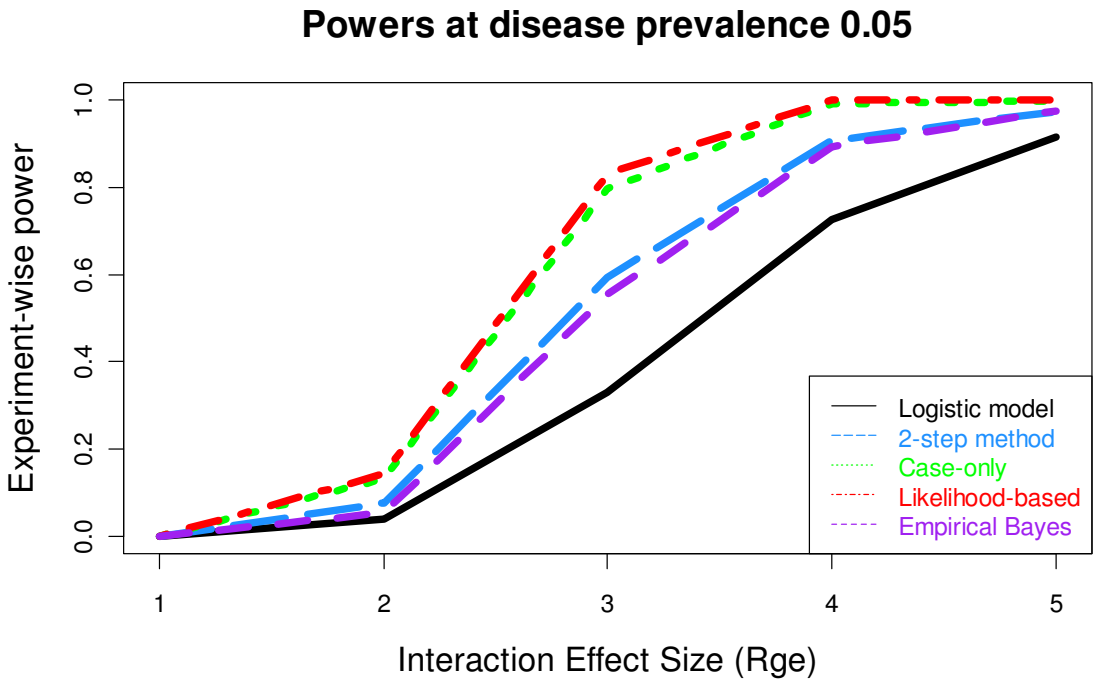


Figure 12.1: Experiment-wise power comparison at diseases prevalence 0.05. The base setting of parameters is $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$ and $p_{ge} = 0$. The interaction effect size R_{ge} is varying.

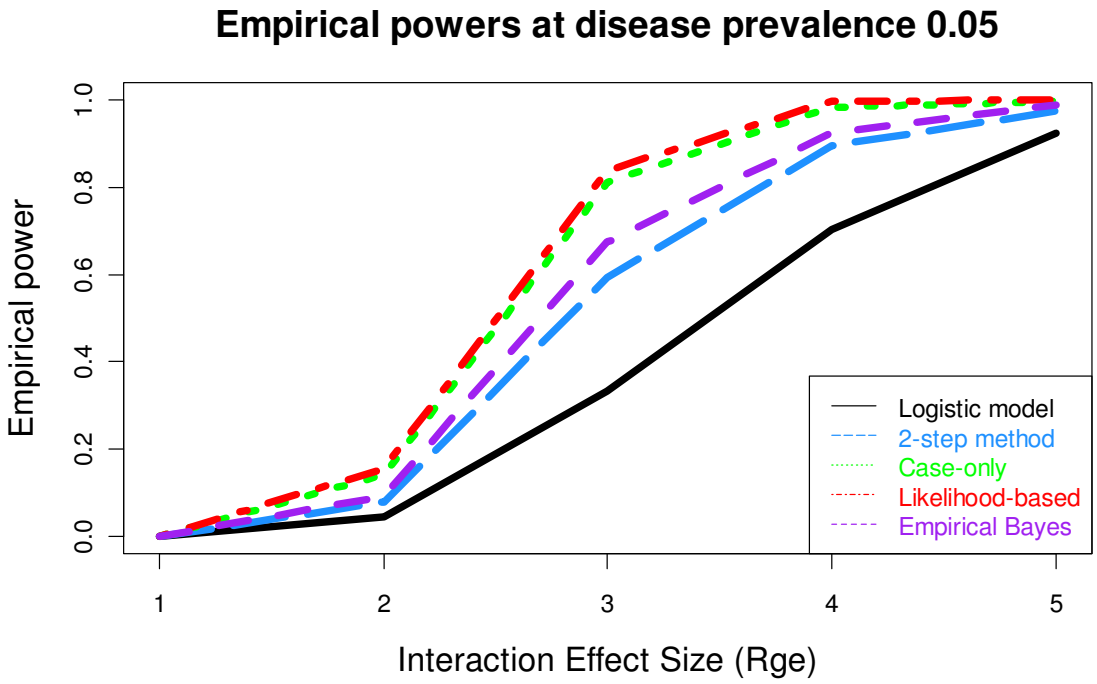


Figure 13.2: Empirical power comparison at diseases prevalence 0.05.
The base setting of parameters is $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$ and $p_{ge} = 0$. The interaction effect size R_{ge} is varying.

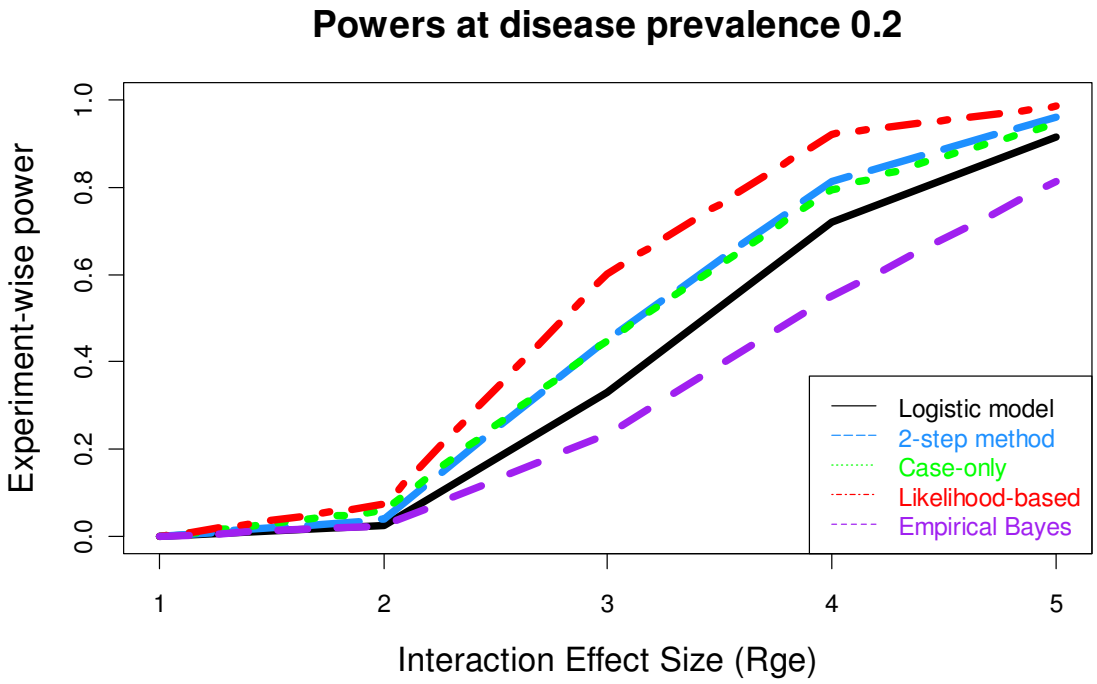


Figure 14.3: Experiment-wise power comparison at diseases prevalence 0.2.
The base setting of parameters is $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$ and $p_{ge} = 0$. The interaction effect size R_{ge} is varying.

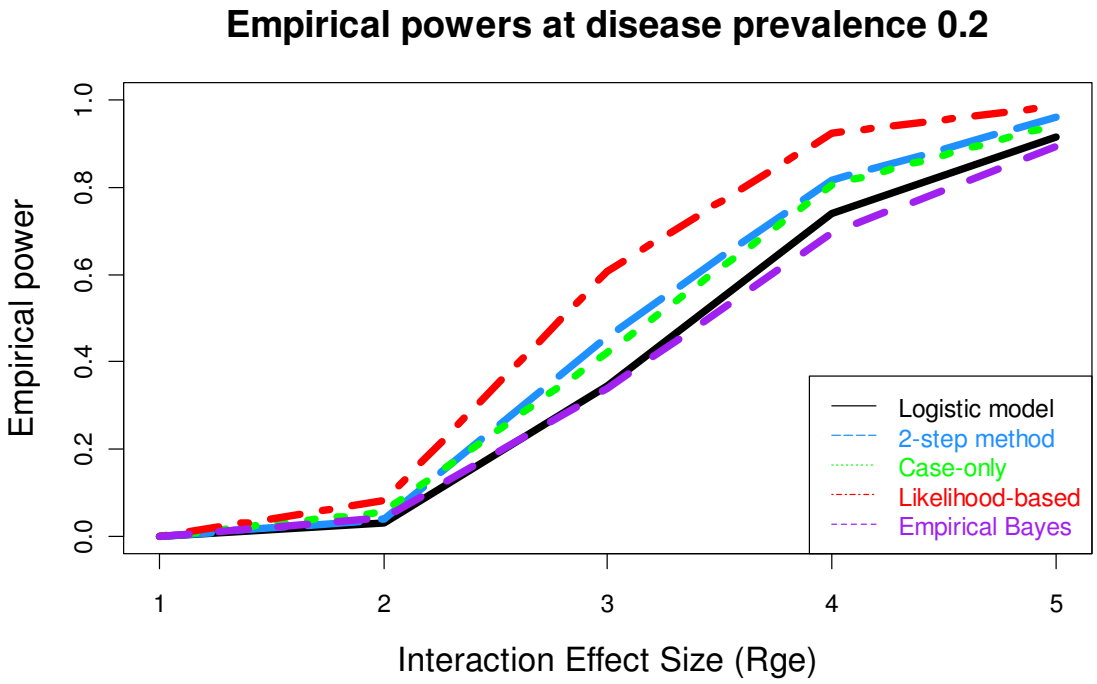


Figure 15.4: Empirical power comparison at diseases prevalence 0.2.
The base setting of parameters is $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$ and $p_{ge} = 0$. The interaction effect size R_{ge} is varying.

Model	Logistic	2-step	Case-only	Empirical Bayes	Likelihood -based
BASE	0.330	0.592	0.796	0.554	0.830
Disease susceptibility locus allele frequency (q_A)					
0.1	0.172	0.346	0.644	0.320	0.656
0.25	0.336	0.598	0.806	0.610	0.848
Exposure prevalence (p_E)					
0.1	0.022	0.090	0.246	0.098	0.262
0.25	0.214	0.476	0.778	0.468	0.800
Effect sizes (R_g , R_e , R_{ge})					
123	0.228	0.486	0.566	0.394	0.648
213	0.308	0.568	0.700	0.568	0.762
223	0.200	0.446	0.380	0.260	0.542
No. of markers (*M*)					
25,000	0.292	0.490	0.762	0.512	0.792
50,000	0.262	0.478	0.712	0.572	0.762
Disease prevalence (p_d)					
0.1	0.318	0.532	0.674	0.420	0.726
0.2	0.330	0.452	0.450	0.234	0.602
0.3	0.328	0.236	0.210	0.156	0.434
0.5	0.308	0.028	0.020	0.148	0.308

Table 16.2: Experiment-wise power comparison.

The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$.

Model	Logistic	2-step	Case-only	Empirical Bayes	Likelihood -based
BASE	0.048	0.050	0.036	0.008	0.042
Disease susceptibility locus allele frequency (q_A)					
0.1	0.074	0.066	0.056	0.012	0.052
0.25	0.050	0.046	0.054	0.012	0.056
Exposure prevalence (p_E)					
0.1	0.050	0.048	0.042	0.006	0.042
0.25	0.048	0.058	0.052	0.016	0.062
Effect sizes (R_g , R_e , R_{ge})					
123	0.050	0.054	0.046	0.008	0.046
213	0.046	0.046	0.038	0.014	0.038
223	0.050	0.064	0.044	0.006	0.040
No. Of markers (*M*)					
25,000	0.038	0.052	0.040	0.014	0.040
50,000	0.074	0.056	0.060	0.012	0.048
Disease prevalence (p_d)					
0.1	0.056	0.040	0.052	0.008	0.048
0.2	0.038	0.046	0.054	0.012	0.042
0.3	0.040	0.060	0.052	0.014	0.048
0.5	0.050	0.058	0.060	0.012	0.050

Table 17.3: Experiment-wise type I error comparison.

The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$.

	Logistic	2-step	Case-only	Empirical Bayes	Likelihood -based
BASE	0.334	0.592	0.810	0.674	0.838
Disease susceptibility locus allele frequency (q_A)					
0.1	0.144	0.326	0.636	0.400	0.652
0.25	0.336	0.610	0.802	0.690	0.844
Exposure prevalence (p_E)					
0.1	0.022	0.090	0.254	0.164	0.268
0.25	0.216	0.454	0.772	0.554	0.798
Effect sizes (R_g , R_e , R_{ge})					
123	0.228	0.482	0.572	0.502	0.668
213	0.308	0.576	0.730	0.568	0.788
223	0.200	0.434	0.392	0.366	0.562
No. of markers (*M*)					
25,000	0.310	0.472	0.770	0.658	0.808
50,000	0.262	0.478	0.712	0.572	0.762
Disease prevalence (p_d)					
0.1	0.310	0.546	0.672	0.532	0.734
0.2	0.344	0.458	0.422	0.340	0.608
0.3	0.344	0.222	0.206	0.220	0.438
0.5	0.308	0.028	0.018	0.212	0.308

Table 18.4: Empirical power comparison.

The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$.

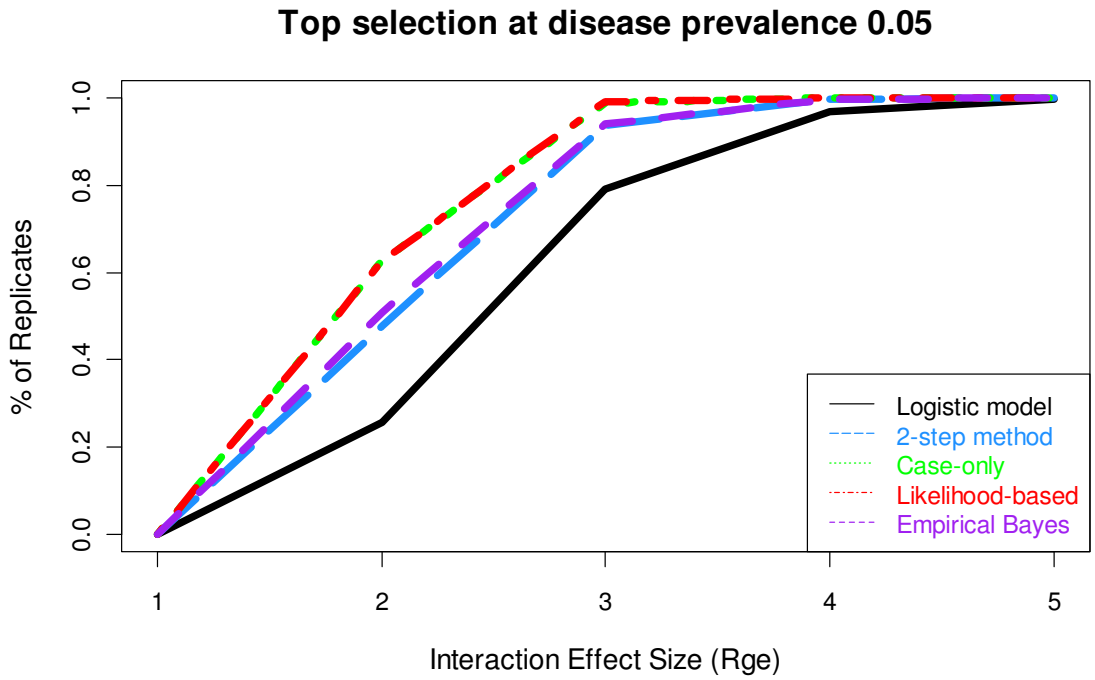


Figure 19.5: Comparison of top selection at diseases prevalence 0.05.
Percentages of replicates for which the p-value for disease susceptibility locus is ranked in the top 10 marker p-values are plotted. The base setting of parameters is $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$ and $p_{ge} = 0$. The interaction effect size R_{ge} is varying.

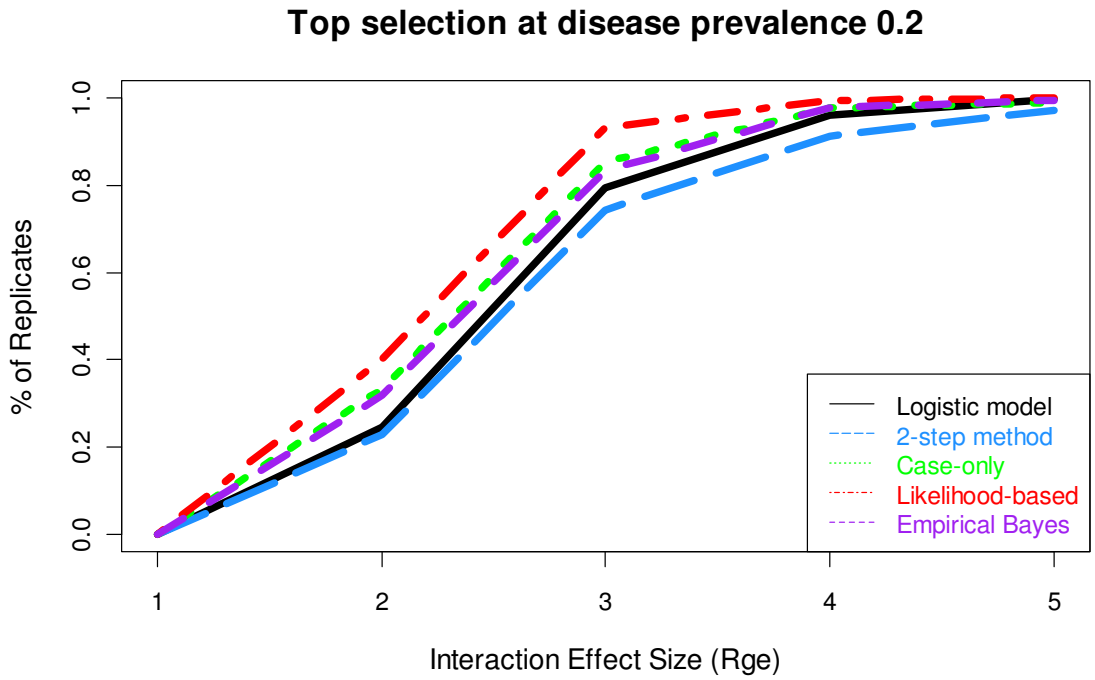


Figure 20.6: Comparison of top selection at diseases prevalence 0.2.
Percentages of replicates for which the p-value for disease susceptibility locus is ranked in the top 10 marker p-values are plotted. The base setting of parameters is $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$ and $p_{ge} = 0$. The interaction effect size R_{ge} is varying.

Model	Logistic		2-step		Case-only		Empirical Bayes		Likelihood-based	
	Top 10	Top 25	Top 10	Top 25	Top 10	Top 10	Top 25	Top 25	Top 10	Top 25
BASE	.790	.868	.938	.972	.990	.992	.942	.962	.992	.996
Disease susceptibility locus allele frequency (q_A)										
0.1	.582	.660	.814	.880	.938	.946	.852	.894	.946	.974
0.25	.806	.874	.944	.976	.978	.984	.950	.968	.984	.998
Exposure prevalence (p_E)										
0.1	.276	.372	.530	.620	.722	.732	.576	.684	.732	.824
0.25	.718	.796	.908	.954	.984	.986	.916	.944	.986	.990
Effect sizes (R_g , R_e , R_{ge})										
123	.730	.810	.914	.942	.928	.956	.890	.918	.956	.980
213	.766	.836	.942	.968	.952	.962	.920	.954	.962	.976
223	.678	.762	.926	.966	.840	.936	.846	.898	.936	.958
No. Of markers (M)										
25,000	.722	.796	.912	.954	.98	.982	.924	.948	.982	.986
50,000	.670	.746	.874	.928	.960	.982	.910	.938	.970	.988
Disease prevalence (p_d)										
0.1	.740	.844	.914	.940	.954	.972	.944	.972	.972	.990
0.2	.794	.858	.744	.758	.854	.932	.834	.894	.932	.958
0.3	.774	.862	.388	.394	.642	.740	.740	.818	.866	.924
0.5	.798	.856	.048	.052	.204	.800	.638	.766	.800	.856

Table 8.5: Comparison of top selection.
Percentages of replicates for which the p-value for disease susceptibility locus is ranked in the top 10 and top 25 marker p-values are listed. The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$.

Model	Logistic	2-step	Case-only	Empirical Bayes	Likelihood -based
BASE	0.048	0.05	0.036	0.008	0.042
Population gene-environment association (p_{ge}=0.01)					
$exp(\theta_{ge})$=1.1	0.040	0.042	0.058	0.014	0.058
$exp(\theta_{ge})$=1.2	0.040	0.042	0.068	0.014	0.068
$exp(\theta_{ge})$=1.5	0.040	0.040	0.474	0.040	0.370
$exp(\theta_{ge})$=2	0.040	0.036	1.000	0.028	1.000
Population gene-environment association (p_{ge}=0.5)					
$exp(\theta_{ge})$=1.1	0.040	0.046	0.070	0.020	0.072
$exp(\theta_{ge})$=1.2	0.044	0.046	0.114	0.024	0.108
$exp(\theta_{ge})$=1.5	0.038	0.062	0.952	0.118	0.884
$exp(\theta_{ge})$=2	0.040	0.034	1.000	0.080	1.000

Table 9.6: Experiment-wise type I error when gene and environment are correlated. The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$.

Model	Logistic	2-step	Case-only	Empirical Bayes	Likelihood -based
BASE	0.334	0.592	0.81	0.674	0.838
Population gene-environment association (p_{ge}=0.01)					
$exp(\theta_{ge})$=1.1	0.334	0.608	0.780	0.654	0.820
$exp(\theta_{ge})$=1.2	0.334	0.608	0.774	0.650	0.800
$exp(\theta_{ge})$=1.5	0.334	0.602	0.550	0.588	0.640
$exp(\theta_{ge})$=2	0.334	0.608	0.078	0.614	0.144
Population gene-environment association (p_{ge}=0.5)					
$exp(\theta_{ge})$=1.1	0.330	0.606	0.764	0.648	0.796
$exp(\theta_{ge})$=1.2	0.322	0.592	0.740	0.622	0.766
$exp(\theta_{ge})$=1.5	0.338	0.536	0.362	0.466	0.414
$exp(\theta_{ge})$=2	0.334	0.572	0.032	0.498	0.054

Table 10.7: Empirical power when gene and environment are correlated.
The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$.

	Logistic	2-step	Case -only	Empirical Bayes	Likelihood -based
BASE					
Power	0.330	0.592	0.796	0.554	0.830
Type I error	0.048	0.050	0.036	0.008	0.042
Empirical power	0.334	0.592	0.810	0.674	0.838
p_{ge}=0.1; $\theta_{ge} \sim N(0, log(1.2)/2)$					
Power	0.338	0.572	0.824	0.608	0.844
Type I error	0.042	0.030	0.064	0.014	0.056
Empirical power	0.352	0.602	0.818	0.664	0.840
p_{ge}=0.15; $\theta_{ge} \sim N(0, log(1.3)/2)$					
Power	0.356	0.574	0.824	0.564	0.840
Type I error	0.064	0.048	0.262	0.040	0.226
Empirical power	0.346	0.586	0.708	0.576	0.768
p_{ge}=0.2; $\theta_{ge} \sim N(0, log(1.5)/2)$					
Power	0.360	0.576	0.850	0.616	0.874
Type I error	0.044	0.050	0.948	0.126	0.878
Empirical power	0.370	0.578	0.294	0.510	0.344

Table 11.8: Integrated power and type I error comparison.
The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$. Power and type I error above refer to integrated power and integrated type I error.

		Logistic		2-step		Case-only		Empirical Bayes		Likelihood-based	
p_{ge}	$exp(\theta_{ge})$	Top 10	Top 25	Top 10	Top 25	Top 10	Top 25	Top 10	Top 25	Top 10	Top 25
0	-	.790	.868	.938	.972	.990	.996	.942	.962	.992	.996
.01	1.1	.808	.870	.932	.964	.982	.994	.942	.962	.986	.994
	1.2	.808	.870	.932	.964	.980	.994	.940	.962	.986	.994
	1.5	.808	.870	.932	.960	.970	.992	.938	.962	.978	.992
	2	.808	.870	.932	.960	.738	.942	.938	.962	.828	.942
.05	1.1	.808	.870	.930	.962	.982	.994	.944	.962	.986	.994
	1.2	.808	.870	.928	.962	.980	.994	.940	.962	.986	.994
	1.5	.810	.870	.916	.952	.908	.974	.928	.962	.930	.974
	2	.810	.866	.910	.948	.464	.724	.934	.962	.564	.724
.1	$\theta_{ge} \sim N(0, log(1.2)/2)$	.786	.862	.952	.978	.994	.996	.950	.968	.996	.996
.15	$\theta_{ge} \sim N(0, log(1.5)/2)$	.796	.858	.934	.972	.978	.996	.936	.958	.984	.996
.2	$\theta_{ge} \sim N(0, log(2)/2)$	.816	.872	.946	.958	.930	.978	.930	.950	.954	.978

Table 12.9: Comparison of top selection when gene and environment are correlated.
Percentages of replicates for which the p-value for disease susceptibility locus is ranked in the top 10 and top 25 marker p-values are listed. The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$.
The first row of $p_{ge} = 0$ indicates the base model.

		Logistic		2-step		Case-only		Empirical Bayes		Likelihood-based	
p_{ge}	$exp(\theta_{ge})$	.05	.001	.05	.001	.05	.001	.05	.001	.05	.001
0	-	.0505	.00102	.0491	.00102	.0504	.00103	.0392	.00048	.0504	.00103
.01	1.1	.0507	.00103	.0494	.00104	.0506	.00103	.0395	.00049	.0506	.00103
	1.2	.0507	.00103	.0498	.00103	.0513	.00108	.0397	.00056	.0513	.00108
	1.5	.0507	.00103	.0499	.00104	.0546	.00188	.0396	.00052	.0546	.00188
	2	.0507	.00103	.0500	.00103	.0587	.00579	.0392	.00048	.0587	.00579
.05	1.1	.0507	.00103	.0499	.00103	.0515	.00109	.0399	.00051	.0515	.00109
	1.2	.,507	.00103	.0499	.00104	.0548	.00136	.0409	.00058	.0548	.00136
	1.5	.0507	.00103	.0500	.00103	.0717	.00531	.0420	.00083	.0717	.00531
	2	.0507	.00103	.0500	.00101	.0921	.02498	.0407	.00063	.0921	.02498
.1	$\theta_{ge} \sim N(0, log(1.2)$	.0508	.00103	.0499	.00102	.0527	.00117	.0403	.00054	.0527	.00117
.15	$\theta_{ge} \sim N(0, log(1.5)$	.0506	.00103	.0500	.00102	.0571	.00168	.0415	.00065	.0571	.00168
.2	$\theta_{ge} \sim N(0, log(2)$	.0506	.00103	.0500	.00105	.0717	.00435	.0441	.00090	.0717	.00435

Table 13.10: Tabulated type I error when gene and environment are correlated.

The parameters of the base model are: $M = 10{,}000$, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_d = 0.05$, and $p_{ge} = 0$. The first row of $p_{ge} = 0$ indicates the base model. The tabulated type I errors are estimated at level of 0.05 and 0.001, respectively.

Chapter 5

Conclusions and Discussions

One of the greatest challenges for genetic researches is the identification of genes that are responsible for complex traits. Unlike classical Mendelian disorders, complex diseases do not show Mendelian patterns of inheritance and include a multiplicity of genetic and environmental factors. The contribution of each factor might be small and different factors might be interactive (Cardon and Bell 2001). Further, confounding factors such as heterogeneity, phenocopies, genetic imprinting and reduced penetrance make thorough genetic dissection difficult, if not impossible.

Although assigning genes to chromosomal locations is ultimately a physical exercise, much can be done with statistical analysis (Weir 2000). With the advent of more cost-efficient high-throughput genotyping technology, appropriate statistical methods are needed to best exploit best information from different types of data, and sophisticated models with higher power are required. In this dissertation, I proposed and analyzed statistical methods for genetic linkage and association analysis.

In the second chapter, I extended the model for linkage analysis of quantitative trait loci. Usually, methods for humans QTL linkage analysis rely on a partitioning of the total variability of trait values. Through variance component models, I estimated and tested the

proportion of phenotypic variance explained by the major genetic, polygenic and environmental factors. Conventionally, the heritability due to the specific locus, polygene or environment is assumed independent of other factors, *i.e.*, identical for different individuals. However, more and more researches have discovered heterogeneity in heritability by age, sex or other covariates. I proposed an extended model to accommodate this type of heterogeneity based on the common variance component model. In the extended model, the genetic effect is a linear function of a covariate, which leads to distinct variances and covariances for different individuals within a pedigree upon the covariate. Simulation studies considering different proportions of variance components, different family sizes, and different significance levels showed that allowing for the heterogeneity lead to an increase in power to detect linkage, especially when the heterogeneity in heritability is large. I also applied the new method to data from HyperGEN network. At the quantitative trait loci where the heritability are very different among different age groups, the new model considering the heterogeneity due to age gave us stronger linkage signals.

In this paper, I test the hypothesis $H_0 : \sigma^2_{mg} = 0, \beta_{mg} = 0$ *vs.* $H_1 : \sigma^2_{mg} > 0$ and β_{mg} unconstrained. The classical asymptotic distribution theory of the maximum likelihood estimates does not hold for the test statistic, since at the expression (2.1), β_{mg} actually disappears under the null hypothesis. I simulated the distribution of the likelihood ratio test statistic under the null hypothesis, and calculate powers based on the empirical cutoffs. By adopting the ideas of admixture test for linkage heterogeneity (Chiu, Liang and Beaty 2002), I would like to establish a new approach next that could

eliminate the nuisance parameters in the test statistic, thereby the theoretical distributions and properties could be explicitly assessed. Based on the theoretical approach, I would also discover more appropriate LOD score of evidence for significant linkage in data application. The relationship between locus specific genetic effect and covariates (linear relationship in this paper) would be further investigated.

In the third chapter, I evaluated and compared several imputation-based association methods to account for uncertainty of imputed genotypes for quantitative traits analyses. During high-throughput genotyping across the whole genome, there are always missing genotype data at some SNP sites due to assay failures and/or by design. Imputation-based association methods provide a powerful framework for testing untyped variants for association with phenotypes and for combining results from multiple studies that use different genotyping platforms (Guan and Stephens 2008). I assessed the powers of three methods to summarize the outputs of genotype imputation in testing the association between the genotypes and the trait of interest by simulation studies. The three strategies are least-squares regression on the "best-guess" imputed genotype, regression on the expected genotype score or "dosage", and mixture regression models that more fully incorporate posterior probabilities of genotypes at untyped SNPs. For most realistic settings of GWAS, such as modest genetic effects, large sample sizes, and high average imputation accuracies, dosage-based analysis provides adequate performance.

In this paper, I focused on testing quantitative phenotypes for association with genotypes. I would be interested in comparison of the three strategies in case-control association

studies (*i.e*. binary phenotypes), especially establishing a sophisticated model to take full advantage of the individual posterior probabilities, which is similar to the mixture regression model described in the paper. I would also like to determine whether the performance of the three strategies is robust to different imputation methods while I used fastPHASE (Scheet and Stephens, 2006) here. In addition to the likelihood ratio test (the F test statistics used in the first two methods is asymptotically equivalent to the likelihood ratio test statistics) I used in this paper, Guan and Stephen (2008) proposed a Bayesian approach. They defined the Bayes factor as the strength of the evidence for alternative hypothesis versus null hypothesis, and demonstrated some advantages of the Bayes factor method than the standard likelihood ratio test. I would compare this approach to others under the same settings.

In the fourth chapter, I proposed a new likelihood-based approach to identify the susceptibility loci by detecting the interaction between gene and environment in the GWAS framework. The gene-environment interaction is a common and important factor for complex diseases. However, current GWAS are designed to test the direct association of a SNP or cluster of SNPs with disease (Browning and Browning 2007, Zhao *et al*. 2006). Investigators may, therefore, miss important genetic variants that are specific to subgroups of the population defined by some environmental exposure (Engelman *et al*. 2009). In comparison to the standard logistic regression model, case-only method, the 2-step method (Murcray *et al*. 2009) and the empirical Bayes estimator method (Mukherjee and Chatterjee 2008), my new likelihood ratio test for the multiplicative gene-environment interaction exploiting the assumption of gene-environment

independence at population level is more powerful than any of others in many circumstances, especially when the disease being studied is common. When the departure to the gene-environment independence is modest across the whole genome, the empirical power (by controlling the adjusted type I error at level of 0.05) of the new method still shows dominant.

In the future, I would like to borrow the idea of conducting a weighted average estimator to combine the case-control estimator and my proposed one. At the same time, I would give the close form of the test statistic for the new likelihood ratio test and further assess the theoretical properties. The 2-step method in Murcray *et al*. (2009) is very robust to the relationship between genotype and environment in a population. I would also replace the original test by the new likelihood ratio test in the first step, since they claimed that when the power of the first step test is high, the chance that a true positive will be carried to the second step is also high.

In summary, my dissertation focused on models and methods for genetic linkage and association studies. I extended the variance component model to allow the heterogeneity in heritability; I evaluated different imputation-based association method in GWAS; I proposed a new test for identifying susceptibility loci by detecting the gene-environment interaction in GWAS. My new models show advantages over existing ones and my analyses provide reference and justification for further studies.

References

Abecasis GR and Wigginton GR. Handling marker-marker linkage disequilibrium: Pedigree analysis with clustered markers. *Am J Hum Genet* **77**:754-767 (2005).

Abecasis GR, Cherny SS, Cookson WO and Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**:97-101 (2002).

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655-1664 (2009).

Allison DB, Fernandez JR, Heo M, and Beasley TM. Testing the robustness of the new Haseman – Elston quantitativetrait loci-mapping procedure. *Am J Hum Genet* **67**: 249-252 (2000).

Almasy L and Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**:1198-1211 (1998).

Almasy L and Blangero J. Human QTL linkage mapping. *Genetica* **136**:333-340 (2009).

Almasy L, Dyer TD and Blangero J. Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkage. *Genet Epidemiol* **14**:953–958 (1997).

Almasy L, Towne B, Peterson C and Blangero J. Detecting genotype × age interaction. *Genet Epid* **21**(Suppl 1):S819-S824 (2001).

Amos CI and Elston RC. Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet Epidemiol* **6**:349-360 (1989).

Arnett DK, Miller MB, Coon H, Ellison RC, North KE, Province M, *et al*. Genome-wide linkage analysis replicates susceptibility locus for fasting plasma triglycerides: NHLBI Family Heart Study. *Hum Genet* **115**:468-474 (2004).

Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* **54**:535-543 (1994).

Austin MA, Edwards KL, Monks SA, Koprowicz KM, Brunzell JD, Motulsky AG, *et al*. Genome-wide scan for quantitative trait loci influencing LDL size and plasma triglyceride in familial hypertriglyceridemia. *J Lipid Res* **44**:2161-2168 (2003).

Baker C. Behavioral genetics: an introduction to how genes and environments interact through development to shape differences in mood, personality, and intelligence. Chapter three: How do environments impinge upon genes? AAAS (American Association for the Advancement of Science) (2004).

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, *et al*. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**:955–962 (2008).

Blangero J and Konigsberg LW. Multivariate segregation analysis using the mixed model. *Genet Epid* **8**:299-316 (1991).

Blangero J. Statistical genetic approaches to human adaptability. *Hum Biol* **65**(6):941-966 (1993).

Browning SR. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* **78**:903-13 (2006).

Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet* **124**:439-450 (2008).

Browning SR and Browning BL. 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 (2007).

Browning SR and Browning BL. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am J Hum Genet* **84**:210-223 (2009).

Cardon LR and Palmer LJ. Population stratification and spurious allelic association. *Lancet* **361**:598–604 (2003).

Cardon LR and Bell JI. Association study designs for complex diseases. *Nat Rev Genet* **2**: 91-99 (2001).

Carlson C, Eberle M, Rieder M, Yi Q, Kruglyak L and Nickerson D. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**:106-120 (2004).

Chatterjee N and Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92**:399-418 (2005).

Chatterjee N and Wacholder S. Invited commentary: Efficient testing of gene-environment interaction. *Am J Epidemiol* **169**:231-233 (2009).

Cheverud JM, Rutledge JJ and Atchley WR. Quantitative genetics of development: Genetic correlations among age-specific trait values and the evolution of

ontogeny. *Evolution* **37**:895-905 (1983).

Chiu YF, Liang KY and Beaty TH. Multipoint linkage detection I the presence of heterogeneity. *Biostatistics* **3**:195-211 (2002).

Christensen K and Murray JC. What genome-wide association studies can do for medicine. *New Engl J Med* **356**:1094-1097 (2007).

Clark AG, Boerwinkle E, Hixson J and Sing CF. Determinants of the success of whole-genome association testing. *Genome Res* **15**:1463-1467 (2005).

Clark AG, Boerwinkle E, Hixson J and Sing CF. Determinants of the success of whole-genome association (2005).

Coon H, Leppert MF, Eckfeldt JH, Oberman A, Myers RH, Peacock JM, *et al*. Genome-wide linkage analysis of lipids in the Hypertension Genetic Epidemiology Network (HyperGEN) blood pressure study. *Arterioscler Thromb Vasc Biol* **21**:1969-1976 (2001).

Covault J, Gelernter J, Hesselbrock V, Nellissery M and Kranzler HR. Allelic and haplotypic association of GABRA2 with alcohol dependence. *Am J Med Genet B Neuropsychiatr Genet* **129**:104-109 (2004).

Czerwinski SA, Mahaney MC, Rainwater DL, Vandeberg JL, Maccluer JW, Stern MP and Blangero J. Gene by smoking interaction: Evidence for effects on low-density lipoprotein size and plasma levels of triglyceride and high-density lipoprotein cholesterol. *Hum Biol* **76**:863-876 (2004).

De Andrade M, Gueguen R, Visvikis S, Sass C, Siest G and Amos CI. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. *Genet Epidemiol* **22**(suppl): 221–232 (2002).

Devlin B, Roeder K and Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* **60**:155-66 (2001).

DeWan AT, Arnett DK, Atwood LD, Province MA, Lewis CE, Hunt SC and Eckfeldt J. A genomen scan for renal function among hypertensives: the HyperGEN study. *Am J Hum Genet* **68**:136-144 (2001).

Duggirala R, Blangero J, Almasy L, Dyer TD, Williams KL, Leach RJ, *et al*. A major susceptibility locus influencing plasma triglyceride concentrations is located on chromosome 15q in Mexican Americans. *Am J Hum Genet* **66**:1237-1245 (2000).

Edenberg HJ, Dick DM, Xuei X, Tian H, Almasy L, Bauer LO, *et al*. Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations. *Am J Hum Genet* **74**:705-714

(2004).

Engelman CD, Baurley JW, Chiu YF, Joubert BR, Lewinger JP, Maenner MJ, Murcray CE, Shi G, and Gauderman WJ. Detecting gene-environment interactions in genome-wide association data. *Genet Epidemiol* **33** (*Suppl* 1): S68–S732 (2009).

Falconer DS. Introduction to Quantitative Genetics (third ed.) (1989).

Falconer D S and Mackay T F C. Introduction to Quantitative Genetics (Addison Wesley Longman, Harlow) (1996).

Feingold E. Methods for Linkage Analysis of Quantitative Trait Loci in Humans. *Theor Popul Biol* **60**:167-180 (2001).

Feinleib M and Garrison RJ. The contribution of family studies to the partitioning of population variation of blood pressure. In: Sing CF, Skolnick M, eds. The Genetic Analysis of Common Diseases. New York: Alan R Liss; 653–73 (1979).

Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *T Roy Soc Edin* **52**:399-433 (1918).

Forrest, W. Weighting improves the "new Haseman – Elston" method. *Hum Hered* **52**: 47-54 (2001).

Franceschini N, MacCluer JW, Cöring HHH, Cole SA, Rose KM, Almasy L, Diego V, Laston S, Lee ET, Howard BW, Best LG, Fabsitz RR, Roman MJ and North KE. A quantitative trait loci-specific gene-by-sex interaction on systolic blood pressure among American Indians. *Hypertension* **48**:266-270 (2006).

Fulker DW, Cherny SS. An improved multipoint sib-pair analysis of quantitative traits. *Behavior Genet* **26**:527-532 (1996).

Fulker DW, Cherny SS, Cardon LR. Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am J Hum Genet* **56**:1224-1233 (1995).

Garrod A. The incidence of alkatonuria: a study in chemical individuality. *Lancet* **160**:1616-1620 (1902).

Goldgar DE. Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* **47**:957-967 (1990).

Green A and Trichopoulos D. Skin cancer. In Textbook of Cancer Epidemiology (eds Adami, H., Hunter, D. & Trichopoulos, D.) 281-300 (2002).

Greenspan G and Geiger D. Model-based Inference of Haplotype Block Variation. *J*

Comput Biol **11**:493-504 (2004).

Guan Y and Stephens M. Practical Issues in Imputation-Based Association Mapping. *PLoS Genetics* **4**e1000279 (2008).

Han J, Hankinson SE, Colditz GA and Hunter DJ. Genetic variation in XRCC1, sun exposure, and risk of skin cancer. *Br J Cancer* **91**: 1604-1609 (2004).

Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**:3-19 (1972).

Hopper JL and Mathews JD. Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* **46**373–383 (1982).

Hwang SJ, Beaty TH, Panny SR, Street NA, Joseph JM, Gordon S, McIntosh I and Francomano CA. Association study of transforming growth factor alpha (TGFa)¹aq1 polymorphism and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects, *Am J Epidemiol* **141:**629-636 (1995).

Hunt SC, Hasstedt SJ, Kuida H, Stults BM, Hopkins PH and Williams RR. Genetic heritability and common environmental components of resting and stressed blood pressures, lipids, and body mass index in Utah pedigrees and twins. *Am J Epidemiol* **129**:625-638 (1989).

Hunter DJ. Gene-enviornment interactions in human diseases. *Nature Review* **6**:287-298 (2005).

International HapMap Consortium. A second generation human haplotypemap of over 3.1 million SNPs. *Nature* **449**:851-861 (2007).

Jinks JL and Fulker DW. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychol Bull* **73**: 311-349 (1970).

Khoury MJ and Flanders WD. ÔNontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controlsÕ. *Am J Epidemiol* **144**: 207-213 (1996).

Khoury MJ, Adams MJ and Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet* **42**:89-95 (1988).

Khoury MJ and Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies – challenges and opportunities. *Am J Epidemiol* **169**:227-230 (2009)

Kraft P, Yen YC, Stram DO, Morrison J and Gauderman WJ. Exploiting

gene-environment interaction to detect genetic associations. *Hum Hered* **63**:111-119 (2007).

Kruglyak L and Daly MJ. Linkage thresholds for two-stage genome scans. *Am J Hum Genet* **62**:994-7 (1998).

Lander E and Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**:241–247 (1995).

Lange K. Mathematical and statistical methods for genetic analysis (1997).

Lange K, Sinsheimer JS, Sobel E. Association testing with Mendel. *Genet. Epidemio*. **29**:36-50 (2005).

Lange K, Weeks D, Boehnke M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet. Epidemiol* **5**:471-72 (1988).

Lange K, Westlake J and Spence MA. Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann Hum Genet* **39**: 485-491 (1976).

Lanktree MB and Hegele RA. Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease. *Genome Medicine* **1**:28 (2009)

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IM, Jacobs KB, Lyssenko V, Uda M, Boehnke M, Chanock SJ, Groop LC, Hu FB, Isomaa B, Kraft P, Peltonen L, Salomaa V, Schlessinger D, *et al*. Identification of ten loci associated with height highlights newbiological pathways in human growth. *Nat Genet* **40**:584–591 (2008).

Li Y, Ding J, Abecasis GR. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet* **79**:S2290 (2006).

Li Y, Willer CJ, Sanna S, and Abecasis GR. Genotype imputation. *Annu Rev Genomics Hum Genet* (2009).

Lin DY, Hu Y and Huang BE. Simple and efficient analysis of disease association with missing genotype data. *Am J Hum Genet* **82**:444-452 (2008).

Liu X, Fallin MD, Kao WH. Genetic dissection methods: designs used for tests of gene-environment interaction. *Curr Opin Genet Dev* **14**:241-245 (2004).

Long JC,Knowler WC,Hanson RL,Robin RW,Urbane kM,Moore E *et al*. Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 froman autosome-wide scan in anAmericanIndian population. *Am J Med Genet*

81(3):216–221 (1998).

Loos RJF, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, Cooper C, Smith GD, Day I, Dina C, De S, Dermitzakis ET, Doney ASF, *et al*. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**:768–775 (2008).

Lynch M and Walsh B. Genetics and analysis of quantitative traits (Sinauer Associates, Sunderland, Massachusetts) (1998).

Mackay TFC, Stone EA, Ayroles1 JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* **10:** 565-577 (2009).

Marchini J, Howie B, Myers S, McVean G and Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**:906–913 (2007).

McConnell R, Berhane K, Yao L, Jerrett M, Lurmann F, Gilliland F, Kunzli N, GaudermanJ, Avol E, Thomas D and Peters J. Traffic, susceptibility, and childhood asthma. *Environ Health Perspect* **114**: 766-772 (2006).

Mitchell BD, Ghosh S, Schneider JL, Birznieks G and Blangero J. Power of variance component linkage analysis to detect epistasis. *Genet Epidemiol* **14**:1017-1022 (1997).

Modan MD, Hartge P, *et al*. Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New Engl J Med* **345:** 235-40 (2001).

Montana G. Briefings in bioinformatics. *Statistical Methods in Genetics* **7**: 297-308 (2006)

Mukherjee B and Chatterjee C. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**:685-694 (2008).

Mukherjee B, AhnJ,Gruber SB, *et al*. Tests for gene-environment interaction from case-control data: a novel study of type I error, power, and designs. *Genet Epidemiol* **32**:615–626 (2008).

Murcray CE, Lewinger JP and Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* **169**:219-226 (2009).

Nelder JA and Mead R. A simplex algorithm for function minimization. *Computer*

Journal 7:308–313 (1965).

Nicolae DL. Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet. Epidemiol* **30**:718-27 (2006).

Pe´russe L, Rice T, Bouchard C, Vogler GP and Rao DC. Cardiovascular risk factors in a French Canadian population: Resolution of genetic and familial environmental effects on blood pressure by using extensive information on environmental correlates. *Am J Hum Genet* **45**:240-251 (1989).

Piegorsch W, Weinberg C and Taylor J: Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* **13**: 153-162 (1994).

Pilia G, Chen WM, Scuteri A, Orrú M, Albai G, Dei M, Lai S, Usala G, Lai M, Loi P, Mameli C, Vacca L, Deiana M, Olla N, Masala M, Cao A, Najjar SS, Terracciano A, Nedrezov T, Sharov A, Zonderman AB, Abecasis GR, Costa P, Lakatta E and Schlessinger D. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**:e132 (2006).

Province MA and Rao DC. A new model for the resolution of cultural and biological inheritance in the presence of temporal trends: application to systolic blood pressure. *Genet Epidemiol* **2**:363-374 (1985b).

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, *et al*. PLINK: a toolset for whole genome association and population-based linkage analyses. *Am J Hum Genet* **81**:559-75 (2007).

Ramos RG and Olden K. Gene-environment interactions in the development of complex disease phenotypes. *Int J Environ Res Public health* **5**:4-11 (2008).

Rao DC, Province MA, Leppert MF, Oberman A, Heiss G, Ellison RC, Arnett DK, Eckfeldt JH, Schwander K, Mockrin SC, and Hunt SC. A genome-wide affected sibpair linkage analysis of hypertension: the hyperGEN network. *Am J Hypertens* **16**:148-150 (2003)

Rice T, Vogler GP, Pe´russe L, Bouchard C and Rao DC. Cardiovascular risk factors in a French Canadian population: resolution of genetic and familial environmental effects on blood pressure using twins, adoptees and extensive information on environmental correlates. *Genet Epidemiol* **6**:571-588 (1989).

Risch N and Merikangas K. The future of genetic studies of complex human diseases. *Science* **273**: 1616-1617 (1996).

Rodriguez-Murillo L and Greenberg DA. Genetic association analysis: a primer on how it works, its strengths and its weaknesses. *Int J Androl* **31**:546-556 (2008).

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN and Clayton D. Empirical genomewide significance levels established by whole genome simulations. *Genet Epidemiol* **14**:223-9 (1997).

Schaffner S, Foo C, Gabriel S, Reich D, Daly M and Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15**:1576-1583 (2005).

Scheet P and Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**:629-44 (2006).

Schmidt S and Schaid DJ. Potential misinterpretation of the case-only study to assess geneenvironment interaction. *Am J Epidemiol* **150**: 878-85 (1999).

Schork N. Genetics of complex disease. *Am J Respir Crit Care Med* **156**:S103-S109 (1997).

Schork NJ. Extended multipoint identity-by-descent analysis of human quantitative traits: Efficiency, power, and modeling considerations. *Am J Hum Genet* **53**:1306-1319 (1993).

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, *et al*. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science* **316**:1341-1345 (2007).

Self SG, Longton G, Kopecky KJ and Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* **47**:53-61 (1991).

Servin B and Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**:e114 (2007).

Sham PC, Purcell S, Cherny SS and Abecasis GR. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet* **71**:238-253 (2002).

Shete S and Amos CI. Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *Am J Hum Genet* **70**:751-757 (2002).

Shi G and Rao DC. Ignoring temporal trends in genetic effects substantially reduces power of quantitative trait linkage analysis. *Genet Epidemiol* **32**:61-72 (2008).

Stephens M and Scheet P. Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. *Am J Hum Genet* **76**:449-462 (2005).

Tang HK and Siegmund D. Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2**, 147-162 (2001).

Teng J and Risch N. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res* **9**, 234-241 (1999).

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**:661-678 (2007).

Towne B, Siervogel RM and Blangero J. Effects of genotype-by-sex interaction on quantitative trait linkage analysis. *Genet Epidemiol* **14**:1053-1058 (1997).

Umbach DM and Weinberg CR. Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Stat Med* **16**:1731-1743 (1995).

Vaughn TT, Pletscher LS, Peripato A, King-Ellison K, Adams E, Erikson C, and Cheverud JM. Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. *Genet Res Camb* **74**:313-322 (1999).

Weir BS. Challenges facing statistical genetics. *J Am Stat Assoc* **95**:449 (2000).

Weiss LA, Pan L, Abney M and Ober C. The sex-specific genetic architecture of quantitative traits in humans. *Nat Genet* **38**:218-222h (2006).

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, *et al*. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**:161-169 (2008).

Williams RR, Rao DC, Ellison RC, Arnett DK, Heiss G, Oberman A, Eckfeldt JH, Leppert MF, Province MA, Mockrin SC and Hunt SC. NHLBI family blood pressure program: methodology and recruitment in the HyperGEN Network. *Ann Epidemiol* **10**:389-400 (2000).

Witte JS, Gauderman WJ and Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* **149**:693-705 (1999).

Xu X, Weiss S, Xu X and Wei LJ. A unified Haseman–Elston method for testing linkage with quantitative traits. *Am J Hum Genet* **67**: 1025-1028 (2000).

Zaitlen N, Kang HM, Eskin E and Halperin E. Leveraging the HapMap correlation structure in association studies. *Am J Hum Genet* **80**:683-91 (2007).

Zhao J, Jin L and Xiong M. Nonlinear tests for genomewide association studies. *Genetics* **174**:1529-1538 (2006).