Abstract

LIU, YOUFANG. Analytical Tools for Population-Based Association Studies. (Under the direction of Dr. Jung-Ying Tzeng.)

Disease gene fine mapping is an important task in human genetic research. Association analysis is becoming a primary approach for localizing disease loci, especially when abundant SNPs are available due to the well improved genotyping technology during the last decades. Despite the rapid improvement of detection ability, there are many limitations of association strategy. In this dissertation, we focused on three different topics including haplotype similarity based test, association test incorporating genotyping error and simulation tool for large data set. 1) Previous haplotype similarity based tests don't have the ability to incorporate covariates in the test. In chapter 2, we proposed a new association method based on haplotype similarity that incorporates covariates and utilizes maximum amount of data information. We found that our method gives power improvement when neither LD nor allele frequency is too low and is comparable under other scenarios. 2) In chapter 3, we proposed a new strategy that incorporates the genotyping uncertainty to assess the association between traits and SNPs. Extensive simulation studies for case-control designs demonstrated that intensity information based association test can reduce the impact induced by genotyping error. 3) In chapter 4, we described simulation software, SimuGeno, which is used to simulate large scale genomic data for case-control association studies.

Analytical Tools for Population-Based Association Studies

by
Youfang Liu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2008

APPROVED BY:

Zhao-Bang Zeng
Co-Chair of Advisory Committee

Jung-Ying Tzeng
Co-Chair of Advisory Committee

Trudy F. C. Mackay

Daowen Zhang

Biography

Youfang Liu was born in Zhejiang province in China. She received her Bachelor of Science degree in biotechnology from Peking University in 1998. After working several years at the Institute of Microbiology of the Chinese Academy of Science, she went to Albert Einstein College of Medicine of Yeshiva University to pursue her Master of Science degree in molecular and cell biology. She completed her M.S. degree in 2003 and in the same year she started her doctoral work in the Ph.D. program in bioinformatics at North Carolina State University. During her study at NCSU, under the direction of Dr. Jung-Ying Tzeng, she focused on her research on method development for association mapping, particularly for population-based association analysis.

Acknowledgements

I would like to express my deepest gratitude to my advisor Dr. Jung-Ying Tzeng, for her influence both academically and personally. She is always patient and provides very helpful suggestions for the problems I encountered during my research. I also thank Dr. Zhao-Bang Zeng for his advice on my research and extremely generous support in the last several years. I also would like to thank my other committee members, Dr. Trudy Mackay and Dr. Daowen Zhang. Special thanks go to Dr. Mike Weale, who guided and helped me a lot to complete one project in my study.

I also appreciate the support from the staffs at the Bioinformatics Research Center. Juliebeth Briseno helped address many issues I had about the graduate school regulations. Tina Chen handled all my financial documents during the past years. Chris Smith helped solve many Unix/Linux system problems I encountered during my research. Many thanks to all the nice people I met here! Without their kind help, I would not be able to complete my doctoral study so smoothly.

Finally, I would like to express my special thanks to my husband, Xiaohua Gong. Without his sincerely love and generous support, I would never be who I am today.

Table of Contents

List of Tables

List of Figures

Chapter 1

Review

1.1 General introduction for association studies

Genetic association studies aim to detect association between the disease phenotypes and genetic polymorphisms. More and more attention has been focused on association analysis due to the rapid development of high-throughput genotyping technology, the availability of large amount of genetic marker and the completion of the initial wave of genetic maps (Neale 2004). During the past decade, association studies have been a promising tool to identify candidate genes or genomic regions that contribute to diseases. Causal genes for diseases, such as type I and type II diabetes, prostate cancer, breast cancer and inflammatory bowel diseases, have been found through genetic association studies (McCarthy 2008). These studies helped us to better understand the molecular mechanisms of diseases and will improve the medicine development in the future.

There are many different genetic markers that can be used to capture the genetic variation, such as restriction fragment length polymorphisms (RFLP's), microsatellites, single nucleotide polymorphisms (SNPs), and copy number variation (CNV). Among all the genetic markers, SNP is the most widely used one for human genetic disease mapping due to their high abundance across the human genome and the rapidly developed genotyping technology, although all the other genetic markers are still very important for genetic association research.

Association between genetic variation and disease phenotypes can be generally grouped into three categories (Cordell 2005): 1) direct association, which means that the marker has a causal role; 2) indirect association, which means that the marker has no causal role but is associated with a nearby causal marker; and 3) confounded association, which is due to the population structure. Direct association is the easiest for association analysis and usually achieves the greatest power. Compared with the direct association, indirect association is much more difficult to detect and it is usually necessary to genotype more surrounding markers to pick up the causal marker. Population structure could result in false positive signals in association studies. To deal with confounded association, there are three strategies: matching by family, excluding population structure associated markers and using genomic control.

Based on how the samples are collected, association tests can be separated into two groups: population-based association test and family-based association test. Population-based analysis requires samples to be collected independently. The rational for population-based analysis is that the allele frequency distributions of the functional loci are different between cases and controls (Risch 2000). As we discussed in the previous paragraph, population structure is the main limitation of population-based test. The advantage of family-based test is that it is still valid even there was population structure. However, family data usually requires more resources in terms of money and time to collect data (Laird and Lange 2006). In the following discussion, we will only focus on population based association test.

Based on the way marker data being used, association test can be grouped into three categories: association test based on single SNP, multiple SNPs, or haplotype. Perhaps the most simple and natural association test is the single SNP based test. In the single SNP based test, the SNP was considered as the basic unit for testing. However, single SNP analyzing probably will neglect information due to joint effect of multiple SNPs. With increasing marker density, association is now often considered at the multiple marker level or haplotype level. Multiple SNPs based test can be used for testing the association between a gene and the phenotype given those SNPs are subject to an LD block within a gene. But the multiple SNP analysis can suffer from several problems: 1) too many parameters are needed to cover all the SNPs; 2) some of those SNPs are highly correlated. Another popular strategy suggested by the block-like structure of the human genome is to use haplotype to capture the correlation structure of SNPs in regions of little recombination. Haplotypes can capture the combined effects of tightly linked cis-acting causal variants.

Despite the rapid improvement of genetic association analysis to date, there are still many limitations in methodologies. Here, we will focus on three problems which still exist in association test: how to incorporate covariates in haplotype similarity based test, how to avoid the power reduction induced by genotyping error, and how to improve the efficiency of large scale data simulation.

1.2 Haplotype similarity based association test

Haplotype is the combination of closely linked SNPs located on the same chromosome. To perform the genetic association study, one can either use SNPs or the haplotypes. Generally, haplotype-based association tests have greater power when SNPs are in strong linkage disequilibrium with the disease locus (Akey 2001, Nielsen 2004, Zaitlen 2007) and are helpful in identifying rare causal variants (HapMap 2003, de Bakker 2005). However, the large dimensionality of haplotypes often leads to high degrees of freedom and the existence of rare haplotypes results in power loss in haplotype-based analyses (Seltman 2001, Molitor 2003(a), Thomas 2003, Zhang 2003, Durrant 2004, Sha 2005, Tzeng 2005, Yu 2005, Browning 2006).

To tackle the haplotype dimensionality problem, many methods have been proposed: (a) haplotype clustering (Seltman 2001, Molitor 2003(a), Durrant 2004, Tzeng 2005, Seltman 2003, Molitor 2003(b), Tzeng 2006), which clusters evolutionarily close haplotypes into groups, (b) haplotype smoothing (Molitor 2003(b), Thomas 2001, Schaid 2004), which smoothes haplotype effects by introducing a correlation structure on the effects of similar haplotypes, and (c) haplotype similarity (Houwen 1994, McPeek 1999, Su 2008), which looks for unusual sharing of chromosomal segments within homogeneous trait groups.

The general rationale behind the similarity method is that the haplotypes around a causative locus will be more similar among the cases descended from the common ancestors. Depending on how this concept is implemented, similarity methods can be divided into two categories: evolutionary based approaches and case-control based approaches.

However, neither of the existing haplotype-similarity approaches can incorporate covariate information into analysis. This makes them less attractive for studying complex traits where covariate adjustments can be crucial. In chapter 2, we proposed an approach that combines the two schools of similarity approaches and is easy to incorporate covariates.

1.3 Genotyping error

High-throughput genome-wide SNP genotyping assay across many thousands of samples is required for association mapping study. It is important to notice that the performance of the genetic association methods depends on the novel high-fidelity genotyping technology and the accurate genotype determination. Many whole genome scan SNP chips have been developed recently, such as Illumina BeadArray, Affymatrix, Perlegen, and Tagman. Although genotyping technology has been considerably improved recently, further improvements are still necessary.

Many genotyping scoring algorithms have been published. Those genotyping scoring algorithms can be divided into two groups: (a) classification based method, and (b) distribution based method. The clustering of individuals for genotyping was widely used in genotyping scoring since 2002, such as K-means algorithm (Oliver 2002), RLMM (Rabbee 2006) and BRLMM (Affymetrix 2007). Distribution and likelihood based methods are developed recently. The original genotyping data can be easily fitted into normal mixture distribution, *t* mixture distribution or gamma distribution and the genotype can be determined by the probability of likelihood (Moorhead 2006, Xiao 2007, Teo 2007, WTCCC 2007).

Genotyping error is defined as the proportion of mistyping in all called genotypes. Genotyping error includes the technological error and the scoring error (Kang 2004). Those technological problems have been improved during the recent years due to the technological development of genotyping whereas the scoring error is still a considerable problem. Genotyping error could result in (1) incorrect estimates of allele frequency, linkage disequilibrium, genetic distance and (2) less power of association studies and linkage analysis (Goldstein 1997, Abecasis 2001, Akey 2001, Gordon 2002, Kang 2004, Hao 2004, Ahn 2006).

It is difficult to avoid genotyping scoring error under traditional association strategy. Recently, a couple of papers have been published that tried to incorporate genotyping uncertainty in association tests (Kang 2004, Zhu 2006). They used genotype probabilistic scoring instead of genotype as input to assess the association analysis. Simulation studies show that their methods can reduce the impact induced by genotyping errors because genotype probabilistic data provides more quantitative information. These two papers discussed above focused on haplotype inference and haplotype association test. In chapter 3, to avoid the genotyping scoring error, we proposed a new score test that incorporates the genotyping uncertainty to assess the single SNP association analysis.

1.4 Software for simulation

One key issue for developing novel association test is how to evaluate the power of each method under realistic settings. Simulation is an efficient way to evaluate the ability of novel methods to detect the disease markers. There are three main approaches for simulation (Liu 2008): 1) “backwards”, which starts with the samples that will form your simulated dataset, then works backwards in time to construct the genealogical information; 2) “forwards”, which starts with the entire population of individuals and then follows how all the genetic data are passed on from one generation to the next; 3) “Sidewards”, which starts with a collection of real genetic data, and uses these as a template for generating new simulated data with similar properties.

With the steady increase in public-available genomewide SNP data, such as the HapMap project, the potential advantage of the “sidewards” simulation approach has been realized recently. HapMap project, as a natural extension of the Human Genome Project, accelerates the pace of biomedical research. By providing abundant human genomic information, such as population information, LD block estimation, and accurate haplotype determination, HapMap project became a popular public data source for statistical genetic research. HapMap data based simulations also have been already widely used in association study (Bakker 2005, Pe’er 2006). In chapter 4, we described the new software, SimuGeno, which is a HapMap data based simulation tool. It offers several different ways to generate genotype data and provides causal region simulation to accelerate the simulation.

Chapter 2

A regression-based association test using inferred ancestral haplotype similarity

Youfang Liu, Yi-Ju Li , Glen A. Satten, Andrew S. Allen and Jung-Ying Tzeng

2.1 Abstract

Objective: We propose a new association method based on haplotype similarity that incorporates covariates and utilizes maximum amount data information.

Methods: We first estimate the ancestral haplotypes of case individual and then, for each individual, an ancestral haplotype based similarity score is computed by comparing that individual's observed genotype with the estimated ancestral haplotypes. Trait values are then regressed onto the similarity scores. Covariates can easily be incorporated under the regression framework. To minimize the bias of raw p-values due to variation in ancestral haplotype estimation, a permutation procedure is adopted to obtain empirical p-values.

Results and Conclusion: To evaluate the power and type I error of our method, we conducted simulations for various scenarios of LD and allele frequency and compared our method with the standard haplotype score test. We found that our method gives power improvement when neither LD nor allele frequency is too low and is comparable under other scenarios. We also applied our method to the GAW15 simulated SNP data for Rheumatoid Arthritis (RA). In an 8cM causal region, our method successfully pinpoints a stretch of SNPs that covers the fine-scale region where the two causal locus of RA, the HLA DR locus and D locus, is located.

2.2 Introduction

Association analysis is becoming a primary approach for localizing disease loci, especially for detecting genes with modest effects on a disease. To access the association between genetic variants and disease, one can either consider individual SNPs or the haplotypes of closely linked SNPs. Although studies of their relative efficiency revealed varying conclusions, it is generally appreciated that haplotype-based analyses have greater power when SNPs are in strong multilocus linkage disequilibrium with the disease locus (Akey 2001, Neilsen 2004, Zaitlen 2007), and are helpful in identifying rare causal variants (HapMap 2003, de Bakker 2005). However, practical potential of haplotype-based analysis may not be fully realized due to the difficulties balancing the dimensionality of the haplotypes and the amount of information (Seltman 2001, Molitor 2003(a), Thomas 2003, Zhang 2003, Durrant 2004, Sha 2005, Tzeng 2005, Yu 2005, Browning 2006). The large dimensionality often leads to high degrees of freedom and the existence of rare haplotypes results in power loss in haplotype-based analyses.

Many thoughts have been proposed to tackle the dimensionality problem in haplotype methods. These approaches include: (a) haplotype clustering (Seltman 2001, Molitor 2003a, Durrant 2004, Tzeng 2005, Seltman 2003, Molitor 2003b, Tzeng 2006), which clusters evolutionarily close haplotypes into groups, (b) haplotype smoothing (Molitor 2003b, Thomas 2001, Schaid 2004), which smoothes haplotype effects by introducing a correlation structure on the effects of similar haplotypes, and (c) haplotype similarity (Houwen 1994, McPeek 1999, Su 2008), which looks for unusual sharing of chromosomal segments within homogeneous trait groups. In this study, we focus on the haplotype-similarity approach, and introduce a method that aims to combine the merits of current similarity methods and incorporate covariate information.

Haplotype similarity methods have been constructed for association testing or for LD mapping. The general rationale behind the similarity method is that haplotypes around a causative locus will be more similar among cases descended from the common ancestors. Depending on how this concept is implemented, similarity methods can be divided into two categories: evolutionary based approaches and two-sample based approaches. Evolutionary based approaches tend to apply to cases only and the excess similarity among cases is identified by comparing to the similarity level expected from the genealogical process (Durham 1997, Service 1999). It takes direct advantage of the decay process of haplotype sharing, which is the underlying driving mechanism for haplotype similarity (McPeek 1999, Su 2008, Morris 2002, Morris 2003, Morris 2005). However, it becomes more challenging to model the genealogical process for complex diseases because the causal variants have a relatively modest impact on total disease risk (Zöllner 2005).

The two-sample based methods hence use the sharing level among control haplotypes as baseline for comparisons (Van der Meulen 1997, Tzeng 2003a, Tzeng 2003b). Although the use of case-control samples bypasses the need to modeling the evolutionary process, these methods tend to be applicable only to binary traits, limit similarity calculations to the concordant samples (i.e., case-case similarity and control-control similarity). Further, many of these methods do not use information obtained from case-control similarity. Sha et al. (2007) showed that by accounting for the information from discordant pairs, the power of using haplotype similarity to detect association is significantly improved (Sha 2007). Finally, the existing haplotype-similarity methods do not incorporate covariate information. This makes them less attractive for studying complex traits where covariate adjustments can be crucial.

In this article, we propose an approach that combines the two schools of similarity approaches to addresses current concerns. Our method follows the framework of two-sample approaches, while the level of similarity in each sample is quantified following the spirit of evolutionary-based approaches. Specifically we estimate the ancestral haplotypes of cases using the multilocus decay-of-haplotype-sharing (DHS) model of McPeek and Strahs (1999), and use the estimated ancestral haplotypes as the “summary haplotypes” of the case haplotypes. We then define the “ancestral haplotype similarity (AHS) scores”, which quantify the similarity degree between the haplotypes of an individual to the case haplotypes by comparing them with these summary haplotypes of cases. The use of ancestral haplotype similarity scores allows us to utilize all sample information in the association testing. It also makes it straightforward to extend a similarity method from the 2-sample test to a regression model to incorporate covariates and allows for quantitative and qualitative traits. Finally, because all our method requires are that the estimated ancestral haplotypes be representative of the case haplotypes, we bypass the issue of whether the estimated ancestral haplotypes are actually the true ancestral disease haplotypes for complex traits.

In the following sections, we illustrate the procedure of the proposed AHS method, and present simulation results of type I errors and power. The results are compared to the standard haplotype score test of Schaid et al. (2002). We also applied our method to the simulated SNP data for Rheumatoid Arthritis (RA) from GAW (Genetic Analysis Workshop) 15 and examined whether this proposed method can detect the DR and D loci that affect the risk of RA.

2.3 Methods

2.3.1 Model and approach

Overall our method can be described in the following three major steps. First, we estimate the ancestral haplotypes of cases from the case genotypes using the DHS method of McPeek and Strahs (McPeek 1999, Strahs 2003). Next, the inferred ancestral haplotypes are used as reference haplotypes to compute the ancestral haplotype similarity (AHS) scores. The AHS score for an individual is obtained by directly comparing an individual's unphased genotypes to the inferred ancestral haplotypes. Finally, the phenotype of interest is regressed onto the AHS scores and covariates. A significant coefficient of the AHS scores indicates genetic association, as there is a different amount of sharing among the control haplotypes compared to the case haplotypes. The significance threshold will be determined by permutation to account for the estimation error induced from the use of the estimated ancestral haplotypes. Below we describe each of the steps in detail.

In the first step, we use the DHS method (McPeek 1999, Strahs 2003) to infer the ancestral case haplotypes. The DHS method models the decay process of haplotype sharing for haplotypes descended from certain common ancestors, taking into account recombination, mutation, and background LD. The model allows for multiple origins of case haplotypes, and incorporates haplotype correlation due to shared ancestry by applying a correction factor to the star-shaped genealogy of case haplotypes. The method has been implemented in the software of DHSMAP (Decay of Haplotype Sharing Mapping), and can take phased haplotypes (McPeek 1999) or unphased genotype data (Strahs 2003).

In the second step, a similarity score for each individual is computed based on the amount of similarity between that individual's haplotypes and the estimated ancestral haplotypes. To tackle the phase unknown problem, we adopt the count statistic of Tzeng et al. (2003a) and measure the level of similarity using the number of matching alleles between an individual’s haplotypes and ancestral haplotype. As shown in Schaid (2004) and Tzeng et al. (2008), such quantity can be obtained by counting the matching alleles between one’s genotypes and the ancestral haplotypes. In the case of multiple estimated ancestral haplotypes, we calculate the overall AHS score by taking a weighted average of the ancestral specific AHS score, with weights dependent on the likelihood of the ancestral haplotypes. Specifically, for individual i, define L_j the likelihood value obtained from DHSMAP for the ancestral haplotype j, and s_{ij}

the count of the matching alleles between the genotype of individual ***i*** and ancestral haplotype ***j***. The weight for ancestral haplotype ***j*** is $w_j = \frac{L_j}{\sum_k L_k}$, k is the number of all inferred ancestral haplotypes, and the overall AHS score for person ***i*** is $S_i = \sum_j s_{ij} w_j$.

In the third step of evaluating association between haplotypes and disease status, we fit a generalized linear model (McCullagh 1989): $g\left(m_i\right) = g^{t} X_i + b\, S_i$, where X_i is the design matrix of covariates including the intercept , $g\left(\cdot\right)$ is the link function, and $m_i = E\left(Y_i \mid X_i, S_i\right)$ with Y_i equaling to the trait value. The null hypothesis of no association corresponds to $\beta = 0$. To address the extra variability introduced into the procedure by using estimated ancestral haplotypes, we use permutation procedure to obtain the empirical p-values. The permutation datasets are obtained by randomly shuffling individual’s genotype instead of phenotype, so to maintain the potential association between the phenotype and the covariates. For each permutated dataset, we repeated the entire procedure of estimating the ancestral haplotypes, computing similarity scores and fitting the regression model. Empirical p-values were computed as the proportion of significant results from all the permutation datasets. All analysis procedures were implemented using R software (http://www.r-project.org/).

2.3.2 Simulation study

Simulations were performed to evaluate the power and type I error of the AHS method. Haplotype data were generated based on the HapMap CEU population. There are in total thirty trio families in the original CEU data set. To create an unrelated random population, we selected two parents from each family to form a sample pool and randomly drew cases and controls from this pool with replacement. We focus on the shortest chromosome 22 to facilitate data processing.

We consider six simulation scenarios, as listed in Table 2.1, with different allele frequency (AF) of the causal variant (0.1 and 0.4) and different LD between the causal variant and its surrounding LD (high ($R^2 > 0.8$), moderate ($R^2 \approx 0.5$) and low ($R^2 < 0.2$)) (Table 2.1). For each scenario of AF and LD, a SNP was chosen as the causal locus in accordance with the simulation setting. Given a causal locus, the three neighboring markers (one marker to its left and two to its right) form a haplotype of size 3 (Figure 2.1). We then randomly draw two haplotypes with replacement and convert it to the unphased genotype for an individual. Therefore, the observed genotype of an individual does not include the causal locus, though the trait value is determined by the genotype of the causal marker using the logistic model: Logit [Pr (Y=1 | G, E)] = $g_0 + g_1 \times E + b_g \times G$, where Y is the disease status coded as 0 and 1; E is for a binary covariate; G is the genotype of the causal marker coded as 0, 1 and 2. We randomly generated the value for covariate E from Bernoulli distribution with p = 0.5. Then, conditional on G and E, we used the logistic formula given above to determine the disease status of an individual. We repeated this process until we obtained enough cases and controls. Two different sample size, 200 and 400, were set to illustrate how sample size will affect power. In each sample, half of them are cases and the other half of them are controls. The logistic model parameters were set as $g_1 = 1.0$ and $g_0 = -4$ which corresponds the population prevalence rate of 0.03. For type I error analysis, β was set to be log(1.0), and for power analysis, β_g = log(1.5). We obtained the empirical p-value by 10^4 permutations.

2.3.3 Data application to GAW15 data

The Genetic Analysis Workshop (GAW) has provided several sets of data biannually since 1982. In 2006, GAW15 provided the simulated Rheumatoid Arthritis (RA) SNP data which includes the HLA DR locus and the D locus on chromosome 6. The DR locus directly affects risk of RA while D locus increases RA risk 5-fold. According to the simulation information provided by GAW, smoking status is a very important covariate which can affect RA risk directly.

We selected one child from each affected sib-pair family to create a case pool. We randomly selected N individuals from this case pool and N controls from all the controls in the simulated data. We considered N = 100 and 200 in the analysis. Among all the SNPs located on chromosome 6, we selected an 8cM region containing 103 SNPs covering the HLA DR locus located at 49.45cM and the D locus located at 54.57cM. We used three-SNP sliding windows of haplotypes to scan this region. For each sliding window, we estimated ancestral haplotypes of the cases using DHSMAP and calculated the AHS scores for each individual. We then regressed the disease status on the ASH scores, with the smoking status included as a covariate in the regression. We obtained the empirical p-value by 10^6 permutations.

2.4 Results

2.4.1 Simulation study

Table 2.2 and Table 2.3 displays the type I error of the AHS tests which was calculated on the basis of 1000 simulation replicates for 200 samples and 400 samples respectively. The values in Table 2.2 and Table 2.3 are close to the nominal level 0.05, indicating that the type I error is under controlled. Table 2.4 and Table 2.5 show the power at the nominal level 0.05 obtained on the basis of 500 simulation replicates for 200 samples and 400 samples respectively. The power of the AHS method is compared to the standard haplotype score test of Schaid et al. (2002) as implemented in the R function haplo.score. We also performed a McNemar test to examine whether the power difference is significant. First, as expected, when the disease locus in low LD with its nearby markers, there is no statistical power to detect the association for both of the method, and hence the power is around the nominal level. When the disease locus is in moderate or high LD with surrounding SNPs, we notice that the power performance depends on the disease allele frequency. The power of the two methods is not significantly different when the disease allele frequency is low. When the disease allele frequency is high, the power of AHS method is significantly higher than the standard haplotype method.

2.4.2 Data application to the GAW15 data

Under the 200 samples situation, the AHS method identified a fine-scale region of 1cM containing 6 SNPs that are significantly associated with RA disease status. Figure 2.2 shows the profile of empirical p-values over chromosomal locations. The peak of empirical p-values is surrounding 49.45cM where the true causal HLA DR locus lies. We also plotted p-values from haplo.score test in Figure 2.2a. It clearly shows that both methods are capable of detecting the disease causal signal.

Under the 400 samples situation, The AHS method identified two very close regions (Figure 2.2b). One region is exactly the same region as described above. The other one is a 1cM region containing 8SNPs and being highly linked with the first region.

On the other hand, both methods did not detect the second disease loci, the D locus located 5.12cM downstream from the DR locus on chromosome 6 whenever 200 samples or 400 samples were used for analysis.

2.5 Discussions

In this work, we proposed a new haplotype similarity method that is under the regression framework and uses the ancestral haplotype similarity scores, which are obtained by comparing individual haplotypes with the inferred ancestral haplotypes of the cases. Using the ancestral haplotypes as a reference for scoring similarity provides a possible mechanism to tackle several common issues encountered in current similarity-based approaches, including using partial of the sample similarity information, ignoring covariates, and applicable to binary traits. In the proposed AHS method, with respect to the reference haplotypes, we assign an AHS score to each individual. Through the AHS scores, the traditional similarity comparison between two samples (case-case pairs vs. control-control pairs or concordant pairs vs. discordant pairs) is now transformed to testing the association between the trait values and the AHS scores. With this transformation, the haplotype similarity information of each sample is retained and used, and the association can be examined naturally under a regression model so to account for covariates and various trait types. Finally, we define the similarity score based on the number of matching alleles between compared haplotypes, which allows for un-phased genotypes and avoids estimation of the haplotype phase for each individual.

As a quick note, we would like to point out that besides using AHS scores as proposed here, there also exist alternative approaches to tackle the common issues mentioned above in the similarity-based approaches, such as Tzeng et al. (2008). In these approaches, the use of the pairwise samples is extended to a model-based framework, and the focus is to study the correlation between the trait similarity and haplotype similarity. It may be interested to understand the relative efficacy of method of this type and the AHS methods.

The proposed AHS method also has its limitations. In the proposed method we used the decay of haplotype sharing method (DHSMAP) to infer the ancestral haplotypes of cases. Concerns may arise as to (a) whether the estimated ancestral haplotypes are actually the true ancestral disease haplotypes, and (b) how to deal with the variation of the estimated ancestral haplotypes. The former issue may be a less concern as what the AHS method requires is that the estimated ancestral haplotypes be representative of the case haplotypes, so that the similarity scores of each individual reflect similarity to case haplotypes. For the latter concern, we chose to perform permutation tests. In our permutation, we choose to permute the genetic information instead of the disease status among individuals. Such procedure retains the relationship between the disease status and the potential covariates.

Our choice of the DHS method to estimate the ancestral haplotypes is not only because it incorporates realistic complications in modeling the genealogy process and is recognized as a foundation method, but also because of its likelihood-based framework. This feature provides a potential opportunity to incorporate the estimation of ancestral haplotypes as an internal step. Currently, our AHS method requires separate steps for estimating the ancestral case haplotypes, calculating the individual AHS scores, and the tests for association. In future improvement, these three steps can be included as intermediate steps under an integrated likelihood of trait values and observed genotypes, and output the association test result that account for the uncertainty of the AHS scores.

Tables

Table 2.1 The six simulation scenarios for causal locus

	Low LD (R^2<0.2)	Moderate LD (R^2 around 0.5)	High LD (R^2>0.8)
AF ≈ 0.1	rs134220	rs2858522	rs9614393
AF ≈0.4	rs133914	rs5768355	rs5750870

Table 2.2 Type I error at nominal level 0.05 (100 cases and 100 controls)

	Low LD	Moderate LD	High LD
AF ≈ 0.1	0.056	0.056	0.052
AF ≈ 0.4	0.052	0.040	0.042

Table 2.3 Type I error at nominal level 0.05 (200cases and 200 controls)

	Low LD	Moderate LD	High LD
AF ≈ 0.1	0.054	0.046	0.050
AF ≈ 0.4	0.054	0.042	0.046

Table 2.4 Power at nominal level 0.05 (100 cases and 100 controls)

		Low LD	Med LD	High LD
AF ≈ 0.1	haplo.score	0.054	0.133	0.175
	AHS method	0.058	0.133	0.168
	McNemar	(p = 0.777)*	(p = 0.951)*	(p = 0.746)*
AF ≈ 0.4	haplo.score	0.041	0.163	0.441
	AHS method	0.038	0.214	0.522
	McNemar	(p = 0.822)*	(p = 0.010)*	(p = 0.010)*

* is the p-value obtained by McNemar test.

Table 2.5 Power at nominal level 0.05 (200 cases and 200 controls)

		Low LD	Med LD	High LD
AF ≈ 0.1	haplo.score	0.092	0.230	0.394
	AHS method	0.091	0.252	0.374
	McNemar	(p = 1.00)*	(p = 0.339)*	(p = 0.493)*
AF ≈ 0.4	haplo.score	0.134	0.349	0.694
	AHS method	0.129	0.403	0.769
	McNemar	(p = 0.805)*	(p = 0.053)*	(p = 0.053)*

* is the p-value obtained by McNemar test.

Figures

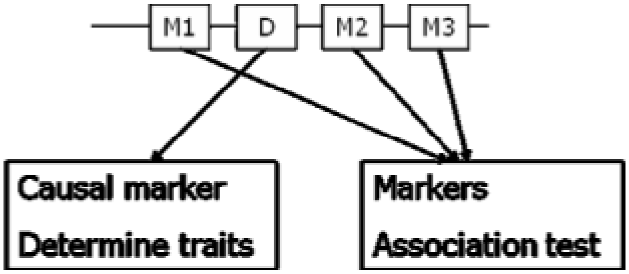
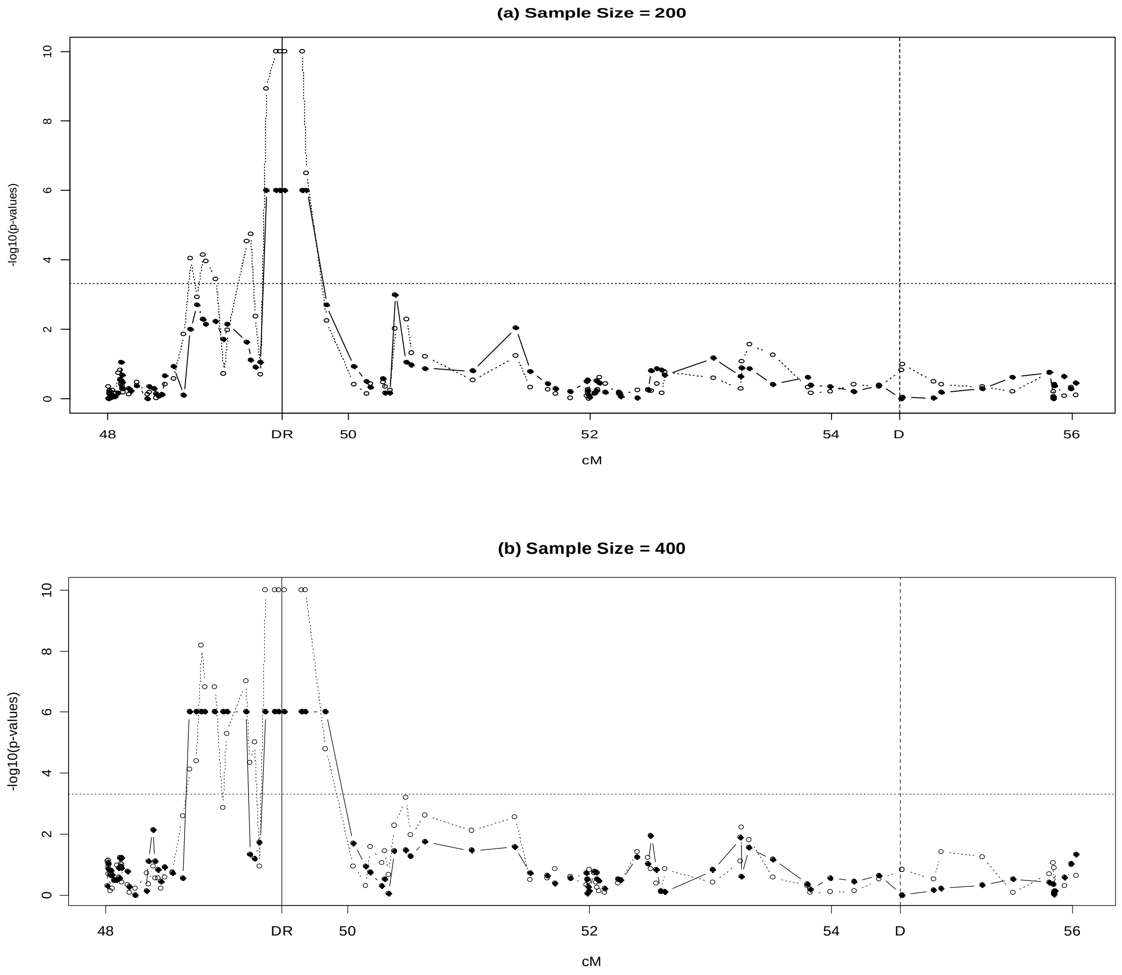


Figure 2.1 Simulation model
D is the causal marker and used for determining traits. M1, M2 and M3 are the markers highly linked with D and used for association test.



(a) Sample Size = 200

(b) Sample Size = 400

Figure 2.2 Data application to the GAW15 data
The negative base 10 logarithm of empirical p-values of the proposed AHS method (solid line with filled circles) and those of the standard haplotype score test (dashed line with open circles) around the DR locus. The p-values which were obtained by AHS method and smaller than 10-6 were replaced with 10-6 in this figure. The p-values which were obtained by the standard haplotype score test and smaller than 10-10 were replaced with 10-10. The dotted horizontal line is the bufferoni correction. The solid vertical line is the location of DR locus. The dashed vertical line is the location of D locus. Figure 2.2a is for the analysis study with sample size equaling to 200, which includes 100 cases and 100 controls. Figure 2.2b is for the analysis study with 400 sample size with 200 cases and 200 controls.

Chapter 3

Association Studies Using Intensity Data

Youfang Liu and Jung-Ying Tzeng

3.1 Abstract

Current genotyping technology produces two dimensional intensity data, from which genotypes are inferred by a scoring algorithm and genetic association are evaluated based on the scored genotypes and phenotypes. Genotyping scoring errors remain a major challenge for automated scoring programs and it renders a negative impact on association analysis. Here, we propose an alternative strategy that uses the intensity data to study gene-trait association. In the analysis, we treat the original two dimensional intensity data or their transformation as the observed genetic variables and regard genotypes as unobserved variables. The genotyping uncertainty is hence incorporated in the assessment of the association. Simulation studies demonstrate that intensity information based association test slightly outperforms other approaches that use inferred genotypes as input when mis-call rate is high.

3.2 Introduction

SNPs, single nuclear polymorphism, the most abundant and stable marker, are widely used in linkage analysis, association mapping and complex disease study (Risch 2000). With the completion of the Human Genome Project, a huge volume of SNPs have been discovered in the human genome (The international HapMap consortium 2005). With more SNPs becoming available, various statistical methods to assess the associations between SNP allelic variants and diseases have been proposed. Previous studies had demonstrated that ~ 200K - 300K tagging SNPs will be required to cover most of genetic variations in the whole genome (Gabriel 2002, Judson 2002, Stephens 2001). Therefore, high-throughput genome-wide SNP genotyping assay across many thousands of samples is required for association mapping study, and the performance of the abovementioned association studies depends on the high-fidelity genotyping technology.

Genotyping can be separated into two steps: allele discrimination and allele detection. Allele discrimination is the generation of allele-specific products for SNPs, which is done by allele-specific biochemical reaction (Syvanen 2001, Kim 2007). There are four different allele discrimination methods: enzymatic cleavage, hybridization with allele-specific probes, oligonucleotide ligation, and single primer extension (Syvanen 2001, kim 2007). Allele detection methods include indirect colorimetric, chemiluminescence, fluorescence, fluorescence resonance energy transfer, fluorescence polarization, mass spectrometry (Syvanen 2001, Kim 2007). Many whole genome scan SNP chips have been developed recently, such as Illumina BeadArray, Affymatrix, Perlegen, and Tagman, etc. They use different allele discrimination methods and allele detection methods (Syvanen 2001). For example, the TagMan assay involves hybridization with allele specific probes and detection by fluorescence resonance energy transfer (Syvanen 2001). Although genotyping technology has been considerably improved recently, further improvements are still necessary in order to improve the quality and efficiency of genotyping (Syvanen 2001, Kim 2007).

Recently, many genotyping scoring algorithms have been published. Those genotyping scoring algorithms can be divided into two groups: (a) classification based method (Olivier

2002, Liu 2003, Rabbee 2006, Bierut 2007), and (b) distribution based method (Fujisawa 2004, Di 2005, Hua 2006, Nicolae 2006, Moorhead 2006, Xiao 2007, Teo 2007, WTCCC 2007). Among the recently published classification based methods. Oliver's method and Bierut's method are both based on the K-means clustering strategy (Oliver 2002, Bierut 2007). Liu et al. proposed the modified partitioning around medoids as a classification method for relative allele signals (Liu 2003). RLMM (Rabbee 2006) and BRLMM (Cawley 2006) are very similar and both based on a robustly fitted linear model and use the Mahalanobis distance for classification. The difference between RLMM and BRLMM is the addition of a Bayesian step which provides improved estimates of cluster centroids and variances (Cawley 2006). RLMM and BRLMM both have been built in an Affymetrix software, GType. Although classification based algorithms are widely used in genotyping scoring, model and likelihood based methods are also developed recently. Fujisawa et al. (2004) proposed a model-based clustering method using a normal mixture model and a well-conceived penalized likelihood. Di et al. (2005) introduced a new dynamic model-based algorithm (DM) for screening over 3 million SNPs and genotyping over 100,000 SNPs. SNiPer-HD (Hua 2006) is based on Gaussian distribution and employs an expectation-maximization (EM) algorithm with parameters obtained from a training sample set. GEL (Nicolae 2006) uses likelihood calculations that are based on Gamma distributions inferred from the observed data. Moorhead (2006) proposed a new method based on normal mixture distribution. SAMS (Xiao 2007) and MAMS (Xiao 2007) are both based on multivariate normal distribution. Teo's method is based on multivariate truncated t distribution. CHIAMO, a newly developed algorithm for Affymetrix SNP chip data, is based on a Bayesian hierarchical mixture model and is one of the scoring algorithms used for WTCCC project.

For genetic association analysis, the major concern is the genotyping error. It could result in incorrect estimates of allele frequency, linkage disequilibrium, and genetic distance. It can also reduce the power and increase the false positives of association analysis (Goldstein 1997, Abecasis 2001, Akey 2001, Gordon 2002, Kang 2004, Hao 2004, Ahn 2006).Genotyping error is defined as the proportion of mistyping in all called genotypes that can be categorized into two groups: the technological error and the scoring error (Kang 2004). Those technological problems have been addressed in recent years by the improvement of genotyping technology whereas the scoring error is still a considerable problem.

Almost all the association studies performed currently would first determine genotype through genotyping scoring and then use the inferred genotype as input to do association mapping. Under such a procedure, it is difficult to avoid genotyping scoring error regardless of which scoring method used. The common strategy to cope with genotyping error is to model the error rates in the association tests (Hao 2004, Kang 2004, Ahn 2006, Cheng 2007 and Plagnol 2007). An alternative strategy is to use probe intensity data instead of the genotype data as input for association test. A few papers have been published trying to use this strategy to incorporate genotyping uncertainty in association test (Kang 2004; Zhu 2006). Kang et al. demonstrated a new method for haplotype inference by incorporating genotyping uncertainty (Kang 2004) by using a t-mixture model to calculate the probability of each genotype before doing the haplotype inference. They found that probabilistic scoring gives rise to more quantitative information and flexibility in the haplotype phasing step and can improve the accuracy in haplotype phasing, especially in high LD and high ambiguity situation (Kang 2004). Zhu moved one step further: they not only estimated possible genotypes thus the haplotype inference but also did the haplotype association test (Zhu 2006). Simulation studies show that their likelihood-based method reduced the impact by genotyping errors. These two papers discussed above are focusing on haplotype inference and haplotype association test. The similar principle can be applied to SNP-based analysis.

Here, we propose an algorithm that incorporates the genotyping uncertainty to assess the association between trait data and SNPs. In this strategy, we use the original two-dimensional intensity data or the transformed one-dimensional intensity data as input and regard genotypes as unobserved variables. We also considered alternatively strategies that are commonly used in practice to reach a better understanding on how different strategies would optimally be applied to various scenarios.

3.3 Methods

3.3.1 Transformation algorithms of the two-dimensional intensity data

The raw intensity data are two-dimensional. Transformation is usually used to create a one-dimensional normal variable, with which a certain scoring algorithm is then applied to determine genotypes. Below we list four commonly used transformation algorithms:

Algorithm 1 (Cawley 2006): $X = \text{asinh}[\ 4\ (\ I_a – I_b\ )\ /\ (\ I_a + I_b\ )\ ]\ /\ \text{asinh}(\ 4\ )$,

Algorithm 2 (Teo 2007): $X = (\ I_a - I_b\ )\ /\ (\ I_a + I_b\ )$,

Algorithm 3 (Bierut 2007): $X = I_a\ /\ (\ I_a + I_b\ )$,

Algorithm 4 (Moorhead 2006): $X = \sinh[\ 2\ (\ I_a – I_b\ )\ /\ (\ I_a + I_b\ )\ ]\ /\ \sinh(\ 2\ )$,

* $\sinh(m)=(\exp(m)-\exp(-m))/2$.

In the algorithms above, I_a stands for probe intensity of allele “a” and I_b stands for probe intensity of another allele “A”. After transformation, X follows a normal distribution, $N(\mu_g,\sigma_g^2)$, with g stands for three different genotypes AA, Aa, and aa.

Many scoring algorithms have been proposed in recent years. In this work we particularly focus on the genotype determination algorithm of Moorhead et al. (2006). The algorithm firstly transforms the two dimensional intensity data into one dimensional data using Algorithm 4 and then fit the transformed data using a mixture normal distribution. It estimates the parameters by EM algorithm, and determines the genotype based on the likelihood value. We choose Moorhead’s method because it is easier to write the likelihood for one dimension data and it is easy to achieve convergence for EM algorithm when there is less parameter in the model.

3.3.2 Likelihood of the complete data

The observed data are the original or transformed intensity data (denoted by R), environmental covariates (denoted by E), and trait value (denoted by Y). The complete data are (G, Y, E) where G is the genotype. We first specify the complete data likelihood as follows, from which we can use EM algorithm to obtain MLEs of the observed-data likelihood or the score function of the observed-data likelihood:

(1) $L = f(Y, G, R \mid E) = f(Y \mid G, R, E)\, f(G, R) = f(Y \mid G, E)\, f(G, R)$, where Y is the trait data, R is the input data which could be the original two dimensional intensity data or the transformed one dimensional data, G is genotype (AA = 0, Aa = 1, aa = 2) and E is the covariate;

(2) $f(G, R)$ = mixture of normal distribution, which will be discussed in the next paragraph;

(3) $f(Y \mid G, E) = \exp\{ (Y\eta - b(\eta)) / a(\varphi) + c(Y, \varphi) \}$, which is expressed as an exponential family data, where a, b and c are known functions, φ is the dispersion parameter and η is the link function;

(4) $\eta = \beta_0 + \beta_g G + \beta_e E$, where $\beta_{0,}$ β_g and β_e are the regression parameters for the intercept, the genotype factor and environmental factor, respectively.

We fit two different normal mixture models for $f(G, R)$: (1) univariate normal mixture distribution model, which uses transformed data X of Algorithm 4 as input data; and (2) bivariate normal mixture distribution, which uses original two dimensional probe intensity data I_a and I_b as input. For the ***i***-th sample, we can write its un-normalized probability of belonging to the ***j***-th cluster, with respect to the two normal mixture models abovementioned, as follows:

(1) $f_{ij}(G = j, R_i = X_i) = (\lambda_j / \sigma_j) \exp(-1/2\, ((X_i - m_j) / \sigma_j)^2)$, where m_j, σ_j and λ_j are the mean, sigma, and weight of the jth cluster, respectively;

(2) f_{ij} (G = j, R_i = (I_{ai}, I_{bi})) = (λ_j / (σ_{xj} σ_{yj} (1 - ρ^2)$^{1/2}$)) exp { -1 / (2 (1 - ρ^2)) [(I_{ai} - m_{xj})^2 / σ_{xj}^2 + (I_{bi} - m_{xj})^2 / σ_{xj}^2 + (I_{ai} - m_{xj}) (I_{bi}- m_{yj}) / (σ_{xj} σ_{yj})]}, where m_{xj}, m_{yj}, σ_{xj}, σ_{yj}, ρ_j and λ_j are the mean, sigma, and weight of the *j*th cluster, respectively.

3.3.3 Score test for H_0: $\beta_g = 0$

Let $q = \left(b_{\mathrm{g}}, y\right)$, where b_{g} is of our interest and $y = \left(b_0, b_e, f_0, f_1, f_2\right)$ is the nuisance parameter. In the one-dimensional intensity based score test approach, we define $f_j = \left(m_j, s_j, l_j\right) \ \forall\, j = 0,1,2$ for each genotype cluster. In the two-dimensional intensity based score test, we have $f_j = \left(m_{xj}, m_{yj}, s_{xj}, s_{yj}, r_j, l_j\right) \ \forall\, j = 0,1,2.$

We are interested in testing the hypothesis $H_0 : b_g = 0$. Denote y% as the maximum likelihood estimates (MLEs) of the nuisance parameter under the null hypothesis. First, we obtain the score and Fisher information matrix from the complete data likelihood as follows:

$$S\left(0,y\%\right)=\begin{pmatrix} S_{b_g}\left(0,y\%\right) \\ S_y\left(0,y\%\right) \end{pmatrix}=\begin{pmatrix} \dfrac{\partial \log L\left(0,y\%\right)}{\partial b_g} \\ \dfrac{\partial \log L\left(0,y\%\right)}{\partial y} \end{pmatrix}=\left.\begin{pmatrix} \dfrac{\partial \log L\left(b_g,y\right)}{\partial b_g} \\ \dfrac{\partial \log L\left(b_g,y\right)}{\partial y} \end{pmatrix}\right|_{b_g=0,y=y\%} \quad \text{and}$$

$$I\left(0,y\%\right)=\begin{pmatrix} I_{b_g b_g} & I_{b_g y} \\ I_{y b_g} & I_{y y} \end{pmatrix}$$

where

$$I_{b_g b_g}=E\left[-\frac{\partial^2 \log L\left(0,y\%\right)}{\partial b_g \partial b_g}\right],$$

$$I_{b_g y}=E\left[-\frac{\partial^2 \log L\left(0,y\%\right)}{\partial b_g \partial y}\right],$$

$$I_{y y}=E\left[-\frac{\partial^2 \log L\left(0,y\%\right)}{\partial y \partial y}\right].$$

Next we use Louis's method (Louis 1982) to compute the score statistic of the observed likelihood (note the asterisks that denote statistics for observed data) as follows:

$$T^*_{b_g}=S^*_{b_g}\left(0,y\%\right)V^{-1}S^*_{b_g}\left(0,y\%\right)$$

where $S^*_{b_g}\left(0,y\%\right)=E\left[S_{b_g}\left(0,y\%\right)\right]$ and

$$V=\operatorname{var}\left[S^*_{b_g}\left(0,y\%\right)\right]=I^*_{b_g b_g}\left(0,y\%\right)-I^*_{b_g y}\left(0,y\%\right)\left[I^*_{y y}\left(0,y\%\right)\right]^{-1}I^*_{y b_g}\left(0,y\%\right).$$

In the above equations,

$$S^*\left(0,y\%\right)=E\left[S\left(0,y\%\right)\right],\ I^*\left(0,y\%\right)=E\left[I\left(0,y\%\right)\right]-E\left[S\left(0,y\%\right)S^T\left(0,y\%\right)\right]+S^*\left(0,y\%\right)S^*\left(0,y\%\right).$$

Note that $I^*_{b_g b_g}\left(0,y\%\right)$ is the element in the matrix $I^*\left(0,y\%\right)$ that corresponds with b_g.

3.3.4 Five strategies for testing association

We will discuss five different association tests using either genotype or intensity data as input.

Test 1: Score test using true genotypes.

Test 2: Score test using estimated genotype

Test 3: Score test using estimated genotype probability

Test 4: Score test using the one dimensional transformed intensity data

Test 5: Score test using the two dimensional original intensity data

3.3.5 Simulation schemes

To compare the performance of using intensity data with that of using genotypes, we simulated probe intensity data for each SNP such that each genotype cluster's shape and size are similar to that of the real probe intensity data. We assumed Hardy-Weinberg equilibrium (HWE) for each SNP in the simulated population and multivariate *t* distribution for each cluster of genotype. Each of the three different genotypes has a multivariate *t* distribution with different mean and variance matrix. Each genotype cluster has a center and spreads in two dimensions with a constant variance. The ambiguity level is controlled by changing the correlation coefficient ρ, the correlation coefficient of the variance matrix. Many scenarios were simulated with different ρ, allele frequencies and sample sizes. Detailed simulation procedure is as follow:

1. Simulate the true genotype using a binomial distribution with p equaling to the allele frequency.

2. Simulate the two dimension probe intensity data using a multivariate t distribution.

3. Determine the disease status for each individual based on the logistic regression model: Logit [Pr (Y=1 | G, E)] = $\beta_0 + \beta_g G + \beta_e E$.

4. Repeat step 1, 2, and 3 until obtaining enough cases and controls. Two different sample size, 500 and 1000, were set to illustrate how sample size will affect power.

5. Generate 1000 replicates for power and type I error analysis.

3.3.6 Application to the WTCCC data

In 2007, the Wellcome Trust Case Control Consortium (WTCCC) provided several data sets including one control data set from the 1958 UK Birth Cohort, another control data from the UK National Blood Service and case data sets for seven diseases, such as Type 1 diabetes, Type 2 diabetes, rheumatoid arthritis, inflammatory bowel disease, bipolar disorder, hypertension and coronary artery disease. In our real data application, we focused on Type 2 diabetes. The molecular mechanism involved in the development of Type 2 diabetes is still poorly understood. According to several association studies reported to date, the number of Type 2 diabetes susceptibility signals increased from three to nine recently (Zeggini 2007). We constructed a data set containing 1,500 samples from the 1958 British Birth Cohort, 1504 samples from the UK Blood Service Control Group and 1,999 samples from the Type 2 diabetes collection. Genotype data and normalized intensity data are both available in WTCCC. Two different association tests were performed in this real data analysis: 1) genotype based score test; 2) transformed one dimensional intensity based score test.

3.4 Results

3.4.1 Comparisons of the four different transformation algorithms

We compared the performance of four different transformation algorithms described in methods under different ambiguity levels from low to high. In our simulation, bivariate *t* distributions were used to generate the fluorescence intensity (FI) scatter plots (see Figure 3.1 shows an example of simulated two dimensional intensity data). We determined the genotype for each individual through Moorhead's method and calculate the miscall rate for each transformation algorithm. We also performed the SNP single marker association test with genotype uncertainty based on four different transformed data and calculated the power.

Table 3.1 shows miscall rates for all the transformation algorithms under different ambiguity level. To make a fair comparison, we gave the same starting points for the centroids for all the algorithms. In the mixed normal model, we picked the cluster with the highest probability. We counted the number of erroneous calls (defined as the calls different from the true calls) in each simulation scenario. At every ambiguity level, the Algorithm 4 outperformed other algorithms.

In addition to performing better in clustering, Algorithm 4 also got higher power for the one dimensional intensity base score test for association study (Table 3.1) while the type I error rate of the four algorithms remain around the nominal level. All above considered, we decided to use the Algorithm 4 in the subsequent simulations.

3.4.2 Ambiguity level and mis-call rate

The ambiguity level of the two dimension intensity data is controlled by changing the correlation coefficient for the covariates matrix, ρ, which is also called the ambiguity parameter. Mis-call rate is the percentage of the mis-classified genotypes among all the genotypes. To describe the relationship between ρ and mis-call rate, we calculated the mis-call rates based on five different ρ values ranging from 0.1 to 0.9. As the ambiguity parameter increases, the mis-call rate decreases (Figure 3.2).

3.4.3 Power comparisons of different association tests

To find which association test performs better, we compared different association tests under many scenarios with different sample sizes, allele frequencies and ambiguity levels. A thousand replicates under each scenario were simulated for type I error and power analyses.

Table 3.2 and Table 3.3 show the type I errors for all the association test methods with sample sizes of 500 and 1000, respectively. All the type I errors are close to the nominal level of 0.05, which verifies the validity of the test statistics constructed here.

Figure 3.3 and Figure 3.4 show the power for all the methods with sample sizes of 500 and 1000, respectively. As expected, the test using true genotypes as inputs for the association test has the highest power in every scenario. Overall, genotype based method has the lowest power. In the low ambiguity cases, genotype based test, genotype probability based test, transformed intensity based test and two dimensional intensity based test yielded similar power. When ambiguity level increases, transformed one-dimensional intensity based test has the highest power among all the methods.

We set four different minor allele frequencies at 0.01, 0.05, 0.10 and 0.25. The power trends for all the allele frequencies are similar. We noticed that it is very difficult for two dimensional intensity based test to estimate the centroid and variance of the minor allele homozygote genotype group when sample size is equal to 500 and minor allele frequency is lower than 0.1. After increasing the sample size to 1000, two dimensional intensity based test obtained the ability to handle simulated data with allele frequency 0.05. Thus, with enough sample size, such as 5000 or more, two dimensional intensity based test would possibly overcome the problem for small sample size and low allele frequency.

3.4.4 Gene-Environment interaction

Increased attention has been paid on gene-environment interaction for complex disease association study. The above framework can be extended to incorporate the gene-environmental interaction:

$$Y = \beta_0 + G + \beta_e E + \beta_{ge} GE,$$

where Y stands for the trait, G for the genotype, E for the environment effect, and GE for the interaction term. We tested the genetic effect $\beta_g = 0$, interaction effect $\beta_{ge} = 0$, and combined effect $\beta_g = \beta_{ge} = 0$ separately. To avoid the simulation complexity, we fixed the sample size to 500, set the ambiguity parameter at 0.7 and the allele frequency to 0.25. The results for four different scenarios are in the Table 3.4 and discussions are as follows:

(1) β_e = log(1.4), β_g = log(1.0), β_{ge} = log(1.0). Under this null scenario, there is no genetic effect and no gene-environment interaction effect. All the tests successfully controlled their type I errors around the nominal level of 0.05.

(2) β_e = log(1.4), β_g = log(1.4), β_{ge} = log(1.0). Since this is the genetic effect only scenario, it is reasonable that we couldn't detect the power by testing β_{ge} alone. The simulation results demonstrated that intensity based method outperformed the genotype based method when testing β_g effect alone or β_g and β_{ge} combined.

(3) β_e = log(1.4), β_g = log(1.0), β_{ge} = log(1.4). Under this scenario, the model contains the gene-interaction effect only. It is interesting to see that the performance of intensity based tests was very similar with that of genotype based test.

(4) β_e = log(1.4), β_g = log(1.4), β_{ge} = log(1.4). This is the most general scenario which contains both the genetic effect and the interaction effect. When testing $\beta_g = 0$ or $\beta_{ge} = 0$, both testing methods obtained similar level of power. However, when testing β_g and β_{ge} jointly, the intensity based test outperformed the genotype based test.

3.4.5 Application to the WTCCC data

We applied the one dimensional intensity based score test and genotype based score test to a real WTCCC data set. It contains 1,999 cases from the Type 2 diabetes collection, 1500 controls from the 1958 Birth Cohort and 1504 controls form the National Blood Service, in which both genotype data determined by Chiamo clustering method and normalized two dimensional intensity data were available. We performed the intensity based test and the genotype based test, respectively. According to several association studies to date, there are totally nine susceptive genes and 11 causal loci for Type 2 Diabetes (Zeggini 2007). However, two loci (rs1111875 and rs13266634) were missing from the WTCCC dataset and hence there were only eight susceptive genes and 9 causal loci included in our analysis. Genotype-based score test and intensity-based score test both identified all the 9 loci and eight susceptive genes. The results are summarized in Table 3.5 that also included the p-values from a previous association study (Zeggini 2007) using WTCCC data set with different number of cases and controls. Comparing our results with the p-values provided by Zeggini's paper, we found that both genotype-based test and intensity-based test got very similar p-values as the previous study.

3.5 Discussions

We constructed an alternative association score test using the original intensity data instead of genotype as input. Based on extensive simulations, we compared the performance of this new score test with another genotype based score test. Our findings are in agreement with two previous studies that intensity data based association test reduce the power impact induced by genotyping error (Kang 2004 and Zhu 2006). Kang' paper demonstrated a new method for haplotype inference by incorporating the probability of each genotype instead of inferred genotype. Zhu's paper only focused on the haplotype association test (Zhu 2006). Their simulation studies also showed that their intensity data based likelihood-based method can reduce the impact by genotyping errors. To date, there is no published method using intensity data as input for SNP based association test. Single marker based association test is generally more widely used than haplotype test in genome-wide association studies due to its simplicity and computational efficiency. In this paper, we proposed a new score test that incorporates the genotyping uncertainty to assess the association between traits and SNPs. In this method, we directly used the original intensity data and regard genotypes as unobserved variables such that genotyping scoring errors would not be a problem. Our simulation studies showed that association analysis using intensity data can improve the power comparing to other approaches using inferred genotypes.

Poor separation between genotype clusters always increases the mis-call rate and thus impacts the association test. The results of power comparison in association test showed that all intensity based test (genotype probability based score test, one dimensional data based score test and two dimensional probe intensity data based score test) had power improvements comparing with inferred genotype based score test when mis-call rate is high.

Two dimensional original intensity data is supposed to contain more quantitative information than transformed one dimensional data. It was surprising that one dimensional data based score test has higher power than two dimensional data based score test under almost all the simulation scenarios. One possible cause could be that two dimensional data based association test approach involves more parameters than one dimensional data based test, which, during Expection-Maximization iterations, could make it more difficult to achieve the convergence.

Since the two dimensional intensity based score test faces the convergence problem in the EM algorithm, it is limited in sample size and allele frequency. When sample size is small and allele frequency is low, there will be few individuals with homozygous minor alleles. Under such situation, it is very difficult for EM algorithm to estimate its centroid and variance. Based on the reasons above, two dimensional intensity based score test can not be applied to a data set with small sample size and low allele frequency simultaneously. However, one dimensional intensity based score test has the ability to overcome the convergence problem encountered during the EM procedure.

In recent years, more and more attention is paid to the copy number variation (CNV), a new type of genetic variation. There are two technology platforms to assess CNV: comparative genome hybridization (CGH) using whole genome TilePath array and comparative intensity analysis using common SNP chip. The raw intensity data generated during SNP genotyping can be mined for copy-number information, making such studies a potential source of data for CNV-disease association studies. In the future, it is possible to extend the proposed intensity data based score test to detect the causal SNP and CNV simultaneously.

Tables

Table 3.1 Comparing four different transformation algorithms

ρ	Tests	Algorithm 1	Algorithm 2	Algorithm 3	Algorithm 4
0.3	MisCall Rate	0.318	0.257	0.352	0.235
	Type I Error	0.045	0.040	0.053	0.051
	Power	0.588	0.603	0.546	0.622
0.6	MisCall Rate	0.250	0.186	0.315	0.152
	Type I Error	0.055	0.049	0.056	0.043
	Power	0.596	0.623	0.573	0.636
0.9	MisCall Rate	0.163	0.077	0.291	0.037
	Type I Error	0.055	0.047	0.044	0.045
	Power	0.668	0.726	0.631	0.732

Table 3.2 Type I error when sample size = 500

AF	ρ	0.9	0.7	0.5	0.3	0.1
0.01	MisCall Rate	0.008	0.082	0.147	0.193	0.227
	Test 1	0.045	0.039	0.046	0.055	0.048
	Test 2	0.044	0.040	0.054	0.054	0.052
	Test 3	0.055	0.041	0.046	0.053	0.048
	Test 4	0.055	0.041	0.048	0.055	0.047
	Test 5	-	-	-	-	-
0.05	MisCall Rate	0.014	0.090	0.153	0.200	0.236
	Test 1	0.048	0.049	0.050	0.044	0.047
	Test 2	0.054	0.055	0.040	0.042	0.050
	Test 3	0.052	0.054	0.045	0.052	0.045
	Test 4	0.052	0.054	0.045	0.052	0.047
	Test 5	-	-	-	-	-
0.10	MisCall Rate	0.020	0.098	0.162	0.210	0.245
	Test 1	0.053	0.045	0.044	0.047	0.041
	Test 2	0.051	0.052	0.049	0.042	0.051
	Test 3	0.041	0.053	0.041	0.048	0.047
	Test 4	0.044	0.053	0.036	0.039	0.048
	Test 5	0.030	0.035	0.033	0.038	0.040
0.25	MisCall Rate	0.035	0.116	0.181	0.230	0.270
	Test 1	0.052	0.043	0.052	0.050	0.047
	Test 2	0.048	0.051	0.050	0.046	0.058
	Test 3	0.046	0.043	0.045	0.049	0.047
	Test 4	0.048	0.044	0.044	0.049	0.046
	Test 5	0.039	0.036	0.043	0.041	0.041

Table 3.3 Type I error when sample size = 1000

AF	ρ	0.9	0.7	0.5	0.3	0.1
0.01	MisCall Rate	0.009	0.082	0.147	0.193	0.226
	Test 1	0.052	0.048	0.055	0.040	0.050
	Test 2	0.046	0.042	0.052	0.046	0.055
	Test 3	0.040	0.045	0.054	0.040	0.054
	Test 4	0.042	0.046	0.054	0.041	0.053
	Test 5	-	-	-	-	-
0.05	MisCall Rate	0.016	0.091	0.155	0.202	0.239
	Test 1	0.048	0.049	0.050	0.056	0.051
	Test 2	0.052	0.047	0.048	0.042	0.048
	Test 3	0.051	0.041	0.050	0.048	0.054
	Test 4	0.048	0.041	0.052	0.048	0.054
	Test 5	0.044	0.045	0.040	0.042	0.048
0.10	MisCall Rate	0.023	0.101	0.166	0.214	0.249
	Test 1	0.047	0.055	0.056	0.048	0.055
	Test 2	0.050	0.044	0.044	0.042	0.047
	Test 3	0.052	0.052	0.044	0.051	0.053
	Test 4	0.051	0.053	0.044	0.052	0.052
	Test 5	0.040	0.039	0.049	0.038	0.040
0.250	MisCall Rate	0.035	0.116	0.182	0.231	0.268
	Test 1	0.047	0.042	0.046	0.050	0.046
	Test 2	0.050	0.049	0.051	0.054	0.043
	Test 3	0.050	0.045	0.050	0.051	0.047
	Test 4	0.051	0.044	0.051	0.050	0.046
	Test 5	0.044	0.042	0.047	0.055	0.040

Table 3.4 Power for genetic model including gene-environment interaction

				$Y = \beta_g G + \beta_e E$	$Y = \beta_g G + \beta_e E + \beta_{ge} GE$		
β_e	β_g	β_{ge}	Tests	$\beta_g = 0$	$\beta_g = 0$	$\beta_{ge} = 0$	$\beta_g = 0$ & $\beta_{ge} = 0$
0.33	0.00	0.00	Test 2	0.054	0.053	0.055	0.054
			Test 4	0.048	0.050	0.048	0.050
0.33	0.33	0.00	Test 2	0.512	0.548	0.043	0.435
			Test 4	0.550	0.570	0.046	0.500
0.33	0.00	0.33	Test 2	0.096	0.058	0.554	0.428
			Test 4	0.102	0.056	0.554	0.427
0.33	0.33	0.33	Test 2	0.668	0.511	0.575	0.726
			Test 4	0.688	0.512	0.574	0.754

Table 3.5 Data application to the WTCCC data

Rs	Chr	Gene	p-values from previous paper	Test 2	Test 4
Rs1801282	3	PPARG	1.3e-03	3.8e-04	1.1e-03
Rs4402960	3	IGF2BP2	1.7e-03	1.7e-03	7.9e-05
Rs10946398	6	CDKAL1	2.5e-05	1.6e-05	7.9e-06
Rs564398	9	CDKN2B	3.2e-04	8.3e-04	7.2e-03
Rs10811661	9	CDKN2B	7.6e-04	1.2e-03	5.4e-03
Rs5015480	10	HHEX	5.4e-06	2.4e-05	1.2e-06
Rs7901695	10	TCF7L2	6.7e-13	8.6e-14	4.9e-14
Rs5215	11	KCNJ11	1.3e-03	1.7e-03	5.3e-03
Rs8050136	16	FTO	2.0e-08	1.6e-08	3.8e-08

Figures

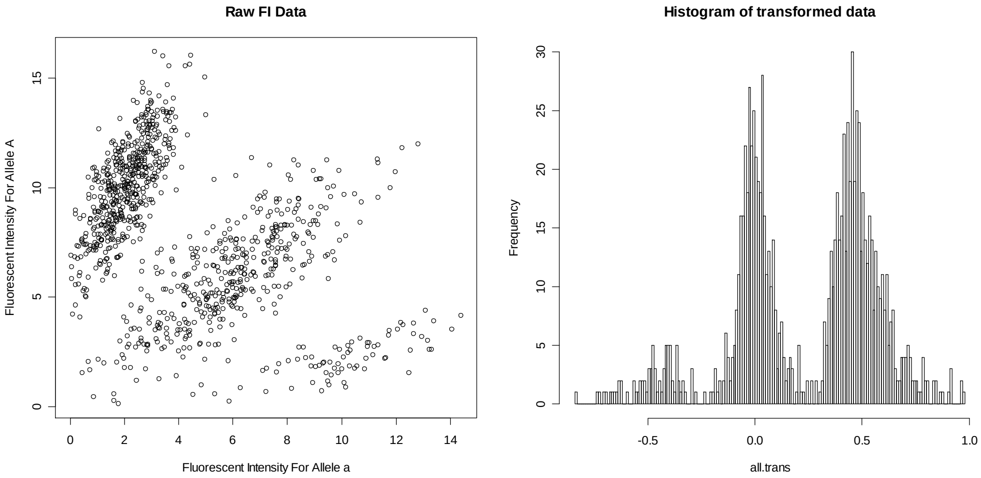


Figure 3.1 An example of simulated data and transformed data. a. The left one: plot for the two dimensional original data. b. The right one: histogram for the one dimensional transformed data.

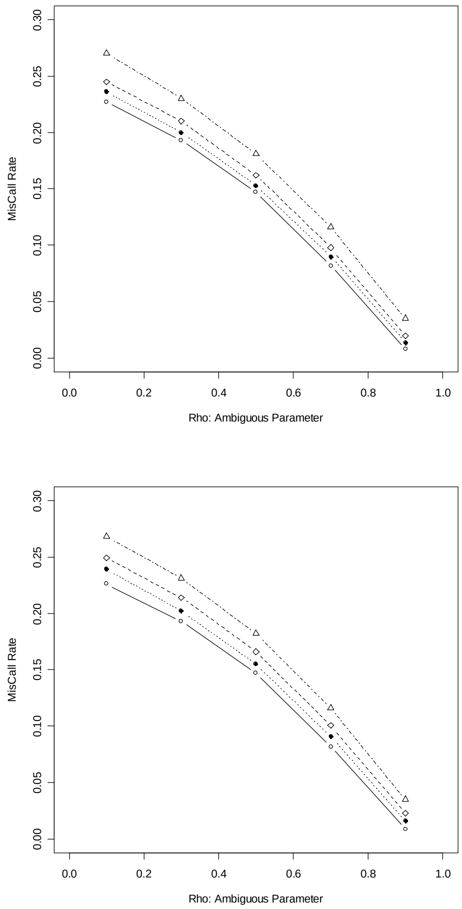


Figure 3.2 Mis-call rate and ambiguity level
Triangle and dot-dashed line stands for allele frequency equaling to 0.25, diamond and dashed line stands for allele frequency equaling to 0.10, solid circle and doted line stands for allele frequency equaling to 0.05, open circle and solid line stands for allele frequency equaling to 0.01. a. The top one: results from simulation with 500 individuals. b. The bottom one: results from simulation with 1000 individuals.

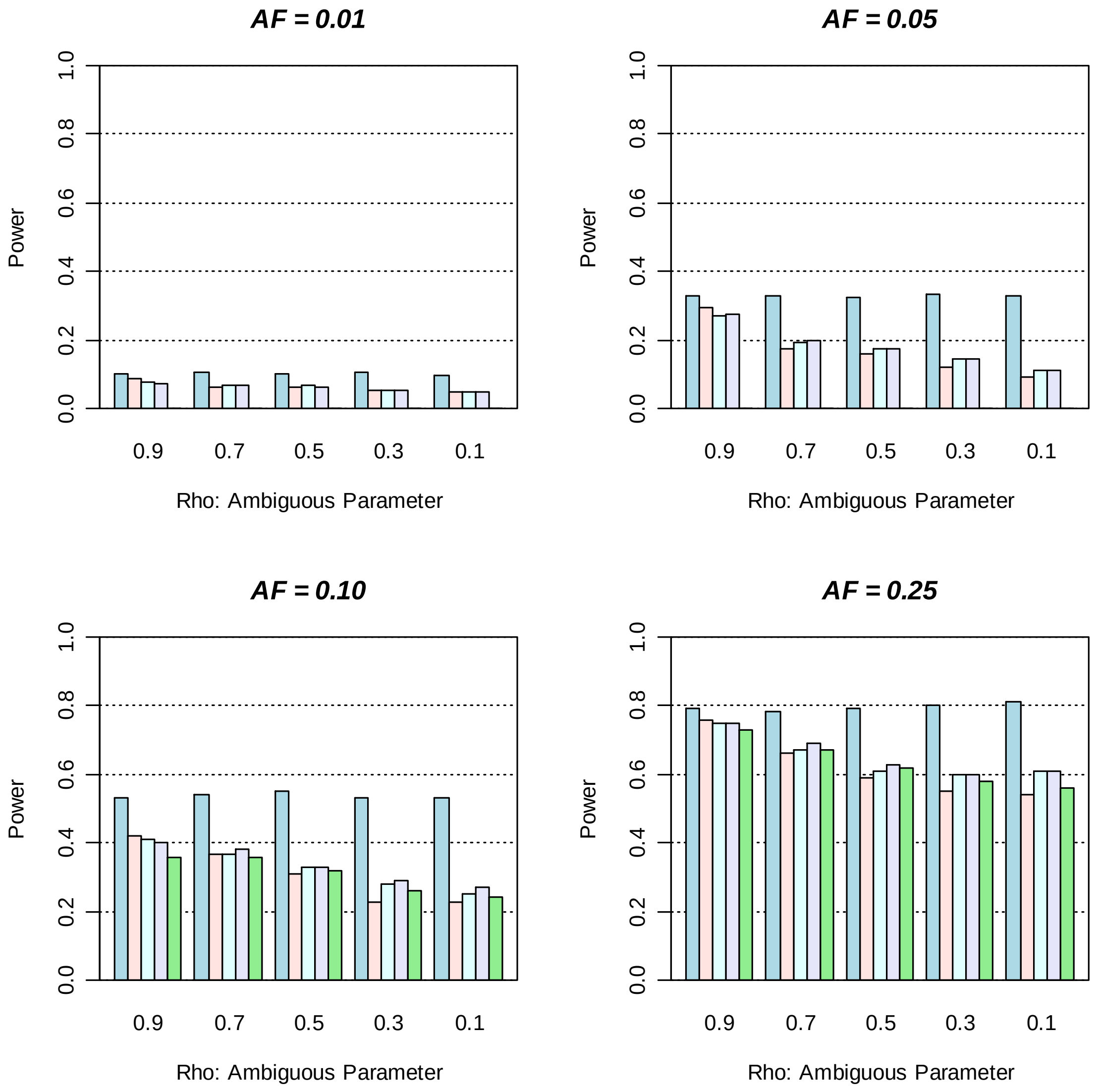


Figure 3.3 Power when sample size = 500
Lightblue bar stands for test 1, mistyrose bar stands for test 2, lightcyan bar stands for test 3, lavender bar stands for test 4, lightgreen bar stands for test 5.

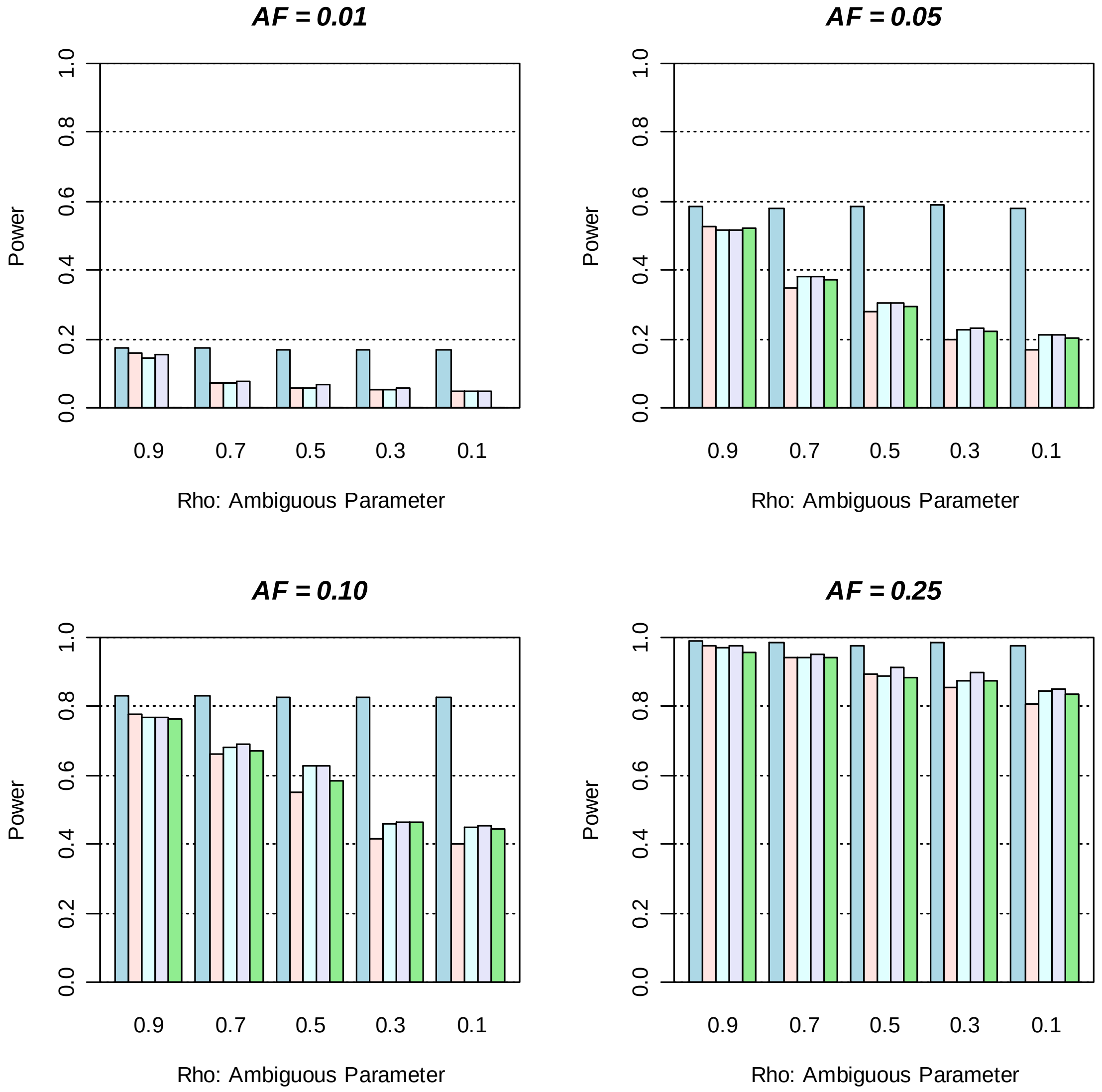


Figure 3.4 Power when sample size = 1000
Lightblue bar stands for test 1, mistyrose bar stands for test 2, lightcyan bar stands for test 3, lavender bar stands for test 4, lightgreen bar stands for test 5.

Chapter 4

SimuGeno: simulation software for genome wide case-control association study

Youfang Liu and Mike Weale

4.1 Abstract

Summary: SimuGeno is a package to simulate large scale genomic data for Case-Control study. SimuGeno use the logistic regression model which allows single causal locus, multiple causal loci and gene-gene interaction. SimuGeno is a real data based simulation software. It can take HapMap data or any other similar data sets as input. The advantage of SimuGeno is that SimuGeno can keep similar allele frequency and LD pattern as the original data.

Availability: http://www4.ncsu.edu/~yliu7/SimuGeno

Contact: yliu7@ncsu.edu

4.2 Introduction

Genetic association analysis has become a powerful and important tool in the study of genetic complex disease. Many novel methods for testing association have been developed. One key issue is how to evaluate the power of each method under realistic settings. Simulation is an efficient way to evaluate the ability of novel methods to detect the disease markers.

There are three main approaches for simulation (Liu 2008): 1) “backwards”, which starts with the samples that will form your simulated dataset, then works backwards in time to construct the genealogical information; 2) “forwards”, which starts with the entire population of individuals and then follows how all the genetic data are passed on from one generation to the next; 3) “Sidewards”, which starts with a collection of real genetic data, and uses these as a template for generating new simulated data with similar properties.

With the steady increase in public-available genomewide SNP data, such as the HapMap project and the 1000 genomes project, the potential advantage of the “sidewards” simulation approach has been realized recently and HapMap data based simulations have been already widely used in association study (Bakker 2005, Pe’er 2006). Dudbridge proposed forming random diploid chromosomes from phased HapMap data followed by a single round of artificial meiosis (Dudbrige, 2007). This idea has been put to use in the HAP-SAMPLE software. Durrant et al. proposed an alternative idea based on sliding windows for introducing new variation into simulated data. This method has been implemented in the GWAsimulator software. Jonathan Marchini’s hapgen software applies an approximation to the coalescent-with-recombination to generate new simulated data from existing phased HapMap data, but is slower than the other two sideways simulators.

SimuGeno is a package to simulate large scale genomic data for case-control association studies. It can take HapMap data or any other similar data sets as a starting point. Causal loci can be either random or user-defined. SimuGeno provides two different ways to generate genotype data: boostrapping and Dudbridge’s method. For large data set, we also provide causal region simulation to fasten the program. After generating genotype data, we use logistic model to determine the disease status for each individual.

SimuGeno simulated data maintains allele frequencies and LD structure that are similar to the original data. As an example, we applied SimuGeno to HapMap CEU chromosome 21 and 22 data. We found that the causal region simulation is more time saving and the simulated data do indeed have very a similar allele frequency pattern and LD structure compared to the original HapMap data.

4.3 methods

4.3.1 Generate genotype by bootstrapping or Dudbridge method

Two options we provide for the simulation of case and control datasets are: bootstrapping of haplotypes and a method proposed by Dudbridge (2007).

(1) Bootstrapping of haplotypes

In this method, two haplotypes are randomly selected firstly, then genotype data is formed by pairing those two haplotypes. We will keep select and pair haplotypes until we get enough cases and controls.

(2) Dudbridge’s model

In this method, new haplotypes are generated based on the recombination rate and random mating assumption (Dudbridge 2007). Two chromosomes are randomly selected, grouped in pairs, and gametes are constructed using HapMap recombination maps. New genotypes are constructed by random union of gametes.

4.3.2 Generate genotype by causal region simulation

Computational time for Dudbridge's method will be a problem when we simulate whole genome data containing about 500k SNPs (the current popular size for genome-wide association studies). To solve this problem, SimuGeno undertakes what we call causal region simulation. The rationale behind this causal region simulation is that only the SNPs close to causal loci will show differences between cases and controls. In causal region simulation, SimuGeno selects, for each causal SNP, a causal region with that causal SNP located at its center. The edges of each region are determined by recombination rates or hotspots. Dudbridge's model based simulation takes place within these regions only. Finally, the newly constructed causal regions are plugged back into the original chromosomal data to create new chromosomes.

4.3.3 Generate disease status by logistic model

SimuGeno use the logistic regression model to determine the disease status:

$\text{Logit}[\Pr(D|G)] = \alpha + \beta1*g1 + \beta2*g2 + \beta3*g3 + \ldots + \beta i*gi$, where D stands for diseases status, g stands for genotype of each causal locus coding in 0,1,2 and i stands for the number of disease loci. In Pr(D|G), the G here stands for the combination of all disease markers.

We define all K as the probability of cases in the whole population. α is solved by the following equation to:

$K = \Pr(D|G) * F(G1) + \Pr(D|G2) * F(G2) + \ldots + \Pr(D|Gj) * F(Gj)$

Here, F stands for genotype frequency and j stands for number of all the possible genotypes. Logistic model allows single causal locus, multiple causal loci and gene-gene interactions among them, and thus allows for complex disease simulation.

4.3.4 Calculate average LD

Pairwise LD block is a common way to depict LD structure. However, it is not easy to show all the LD blocks when the simulated data is quite large. To compare the LD structures between different simulated data sets, we chose to use average LD. Average LD can be calculated as follows: 1) LD between one marker and every marker in its 50kb neighborhood are calculated, 2) the averaged LD value is assigned to the marker.

4.4 Results for an example

4.4.1 Allele frequency, LD structure and running time

We compared the allele frequency and LD pattern among the original HapMap data, the simulated data by bootstrapping and simulated data by Dudbridge model to find out whether the simulated data could keep similar LD pattern and allele frequency as the original one. The original data used in the test was the HapMap chromosome 21 CEU data. There are in total thirty trio families in the original CEU data set. To create an unrelated random population, we selected two parents from each family to form a sample pool.

Minor allele frequencies of data sets simulated by bootstrapping, Dudbridge's method or causal region simulation were compared with the original HapMap data (Figure 4.2). The results show that simulated data by either method had very similar allele frequency as the original one.

Average LD was calculated as discussed in 2.4. Average LD of simulated data by either method was compared with the original data (Figure 4.2). The results show that simulated data sets by all the methods share very similar LD structure with the original HapMap data.

Finally, we found that causal region simulation does save time for large data simulation (Table 4.1). When sample size is small, the running times are quite similar for different simulation methods. However, when sample size is as large as 1000, the time needed for the causal region simulation is only half of the time for Dudbridge's method.

4.4.2 Type I error and power

The original data used in the test was the HapMap chromosome 21 and 22 combined data. CEU population was the only population to be used in the test to avoid potential population structure problem. Three causal SNPs were selected for the simulation. They were all separated far away and there was no LD between any two of them. Interaction between causal SNP1 and causal SNP2 was designed in the model. The logistic model used is listed as below:

$\text{Logit}[\Pr(D|G)] = \alpha + 0.4*g1 + 0.4*g2 + 0.4*g3 + 0.2*g1*g2$.

Detailed information of causal markers is listed in Table 4.2. Originally, minor allele frequency of each causal SNP was 0.25. After simulation, the minor allele frequency of each causal SNP changed a little bit.

To calculate type I error and power, logistic test was applied to the simulated data sets. Table 4.3 shows that simulated data by either model has type I error close to 0.05 as expected. Table 4.3 shows that logistic test has high power to detect the causal locus for simulated data sets by all simulation methods.

4.5 Operating systems

All the source codes are written in R. They can be run on many operating systems as long as the R-2.2.1 or higher version of R is installed. Many parameters, such as sample size and number of disease markers, all can be determined by user. A detailed documentation for how to use this package is available at the website.

Tables

Table 4.1 Running time for different simulation methods

Sample Size	Bootstrapping	Dudbridge	Causal Region
200	2´ 03″	3´ 31″	2´ 21″
1000	4´ 34″	12´ 31″	5´ 36″

Table 4.2 Causal SNPs information

	Chr	Position	MAF
SNP1	21	18053748	0.25
SNP2	21	40181010	0.25
SNP3	22	46323517	0.25

Table 4.3 Type I error and power at nominal level 0.05

		Bootstrapping	Dudbridge	Causal Region
Type I Error	SNP1	0.056	0.044	0.056
	SNP2	0.052	0.050	0.052
	SNP3	0.053	0.049	0.046
Power (without inter action term)	SNP1	0.999	0.995	0.998
	SNP2	0.987	0.996	0.996
	SNP3	0.983	0.994	0.952
Power (with inter action term)	SNP1	0.999	0.997	0.999
	SNP2	0.993	0.998	0.999
	SNP3	0.970	0.996	0.944
	SNP1*SNP2	0.182	0.210	0.204

Figures

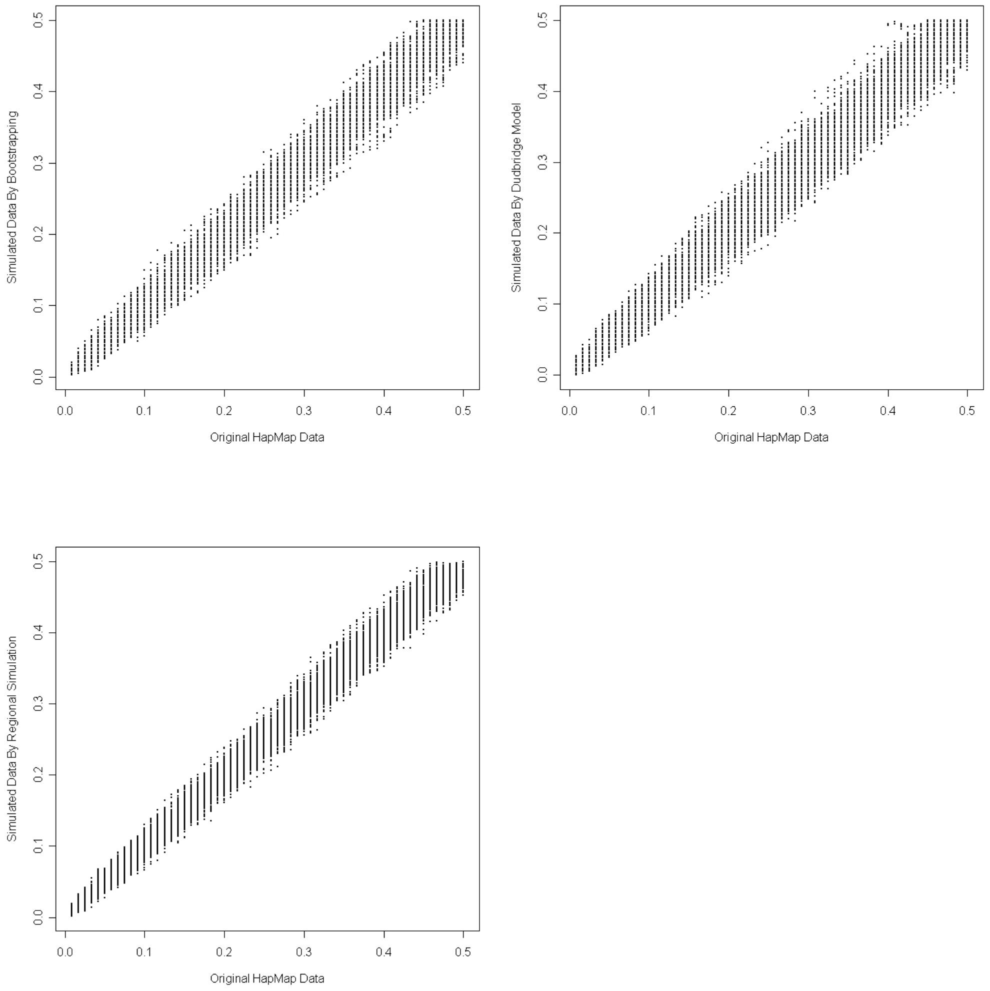


Figure 4.1 Allele frequency comparisons
Minor allele frequencies for original Hapmap data, simulated data by bootstrapping, simulated data by Dudbridge model, and simulated data by regional simulation.

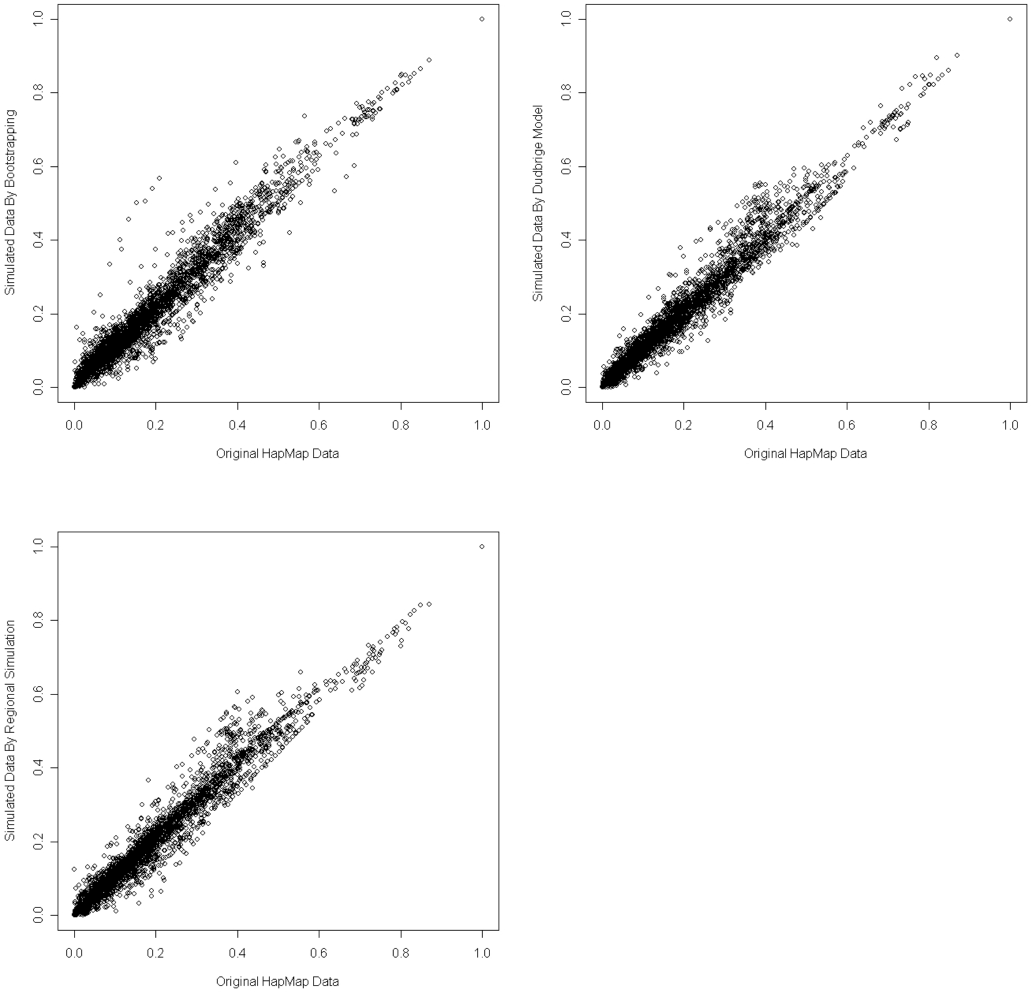


Figure 4.2 LD comparisons
LD for original Hapmap data, simulated data by bootstrapping, simulated data by Dudbridge Model, and simulated data by regional simulation.

Chapter 5

Summary

5.1 Summary and discussion

There is always a debate about whether the haplotype based test is better than the single marker based test. Generally, haplotype-based association tests have greater power when SNPs are in strong linkage disequilibrium with the disease locus (Akey 2001, Nielsen 2004, Zaitlen 2007) and are helpful in identifying rare causal variants (HapMap 2003, de Bakker 2005). The reasons for thinking that haplotype gains more power than single marker test include: 1) haplotypes are comprised of the functional units of genes and 2) haplotypes incorporate LD information (Allen 2008). In addition, comparing with haplotype test, single marker tests can lost significant power when multiple loci located at a same disease region affect disease simultaneously (Morris 2002). However, the large dimensionality of haplotypes often leads to high degrees of freedom and the existence of rare haplotypes results in power loss in haplotype-based analyses (Seltman 2001, Molitor 2003(a), Thomas 2003, Zhang 2003, Durrant 2004, Sha 2005, Tzeng 2005, Yu 2005, Browning 2006). In real data practice, single marker based test is still widely used because it is simple, fast and efficient. Thus, both the single marker based methods and haplotype based methods have their advantages and disadvantages, and both deserve further research.

For population-based association studies, many haplotype based tests have been developed. Haplotype global tests had been proposed by Zhao (2000) and Fallin (2001). They treated different haplotype as categorical data and tested them all together. These methods do not have the ability to identify the specific causal haplotype. To address this problem, haplotype specific tests had been developed by Schaid (2002) and Zaykin (2002). These methods incorporated the haplotypes into a regression model and had the ability to test the effect of each haplotype. However, haplotype diversity became a new problem for the haplotype specific tests because high haplotype diversity would use large number of degree freedom and result in low power. To tackle this haplotype dimensionality problem, many haplotype similarity methods had been developed (McPeek 1999, Tzeng 2003). There are two types of haplotype similarity tests: evolutionay based test (McPeek 1999) and case-control based test (Tzeng 2003). McPeek' method modeled the geneology of the case data and estimated the ancestral haplotype. The rational behind McPeek's method is that, in the cases, the haplotype similarity level should be higher in causal region than in other regions. Instead of comparing current haplotypes with ancestral haplotypes, Tzeng's method compared the haplotype similarity between cases and controls. However, neither McPeek's method nor Tzeng's method can adjust covariate. This became a big limitation for application of both methods because covariates play an important role in complex disease mapping. To tackle the covariates incorporation problem, we proposed the ancestral haplotype similarity based association method. By combining the strengths from the evolutionary based haplotype similarity test and the case-control based haplotype similarity test, our proposed method successfully solves the covariate-incorporation problem and the haplotype diversity problem simultaneously. In our method, we 1) use McPeek's method to estimate ancestral haplotypes and treat them as reference haplotypes, 2) calculate similarity score through Tzeng's method, 3) incorporate covariates through a logistic regression framework, 4) and calculate empirical p-values through permutation. However, there are limitations for our proposed method, too.

First of all, since the empirical p-values are obtained through permutations, time consumption becomes an unavoidable problem, which will limit its application to GWAS (genomewide association scan). Secondly, DHSMAP (McPeek 1999), the software that was used to estimate ancestral haplotypes, will be extremely slow when the sample size is large, which also limits the application of our proposed method to a large data set. Thus, it may be worth researching how to combine the ancestral haplotype estimation step and the testing step together and avoid the permutation and the sample size limitation.

Almost all the association studies conducted currently would first determine genotype through genotyping scoring and then use the inferred genotype as input to do association mapping. Under such a procedure, it is difficult to avoid genotyping scoring error regardless of the scoring method used. The common strategy to cope with genotyping error is to model the error rates in the association tests (Hao 2004, Kang 2004, Ahn 2006, Cheng 2007 and Plagnol 2007). An alternative strategy is to use probe intensity data instead of the genotype data as input for association test. A few papers have been published trying to use this strategy to incorporate genotyping uncertainty in association test (Kang 2004; Zhu 2006). Kang (2004) demonstrated a new method for haplotype inference by incorporating genotyping uncertainty (Kang 2004) by using a t-mixture model to calculate the probability of each genotype before doing the haplotype inference. They found that probabilistic scoring gives rise to more quantitative information and flexibility in the haplotype phasing step and can improve the accuracy in haplotype phasing, especially in high LD and high ambiguity situation (Kang 2004). In Zhu's method, they not only estimated possible genotypes thus the haplotype inference but also did the haplotype association test (Zhu 2006). Kang's method and Zhu's method motivated us to move one step further. Instead of using genotype probability as input, we proposed a new strategy to use intensity data as input. Under such strategy, we can incorporate the genotyping scoring algorithm and take all the genotyping uncertainty into consideration. Extensive simulation studies demonstrated that intensity information based association test outperforms genotype based approaches when ambiguity level of the intensity data is high. However, based on the running time observed during simulation study, intensity data based test consumes longer time than regular genotype based test. Although the running time difference is not so significant for single marker association analysis, time consummation could be a serious problem when the intensity based test is applied to GWAS. To tackle this potential running time problem, recoding the intensity based test using C^{++} instead of R language could be the possible solution. The reason is that R is usually slower than C^{++} due to the memory limitation in R environment.

Genetic association analysis has become a powerful and important tool in the study of genetic complex disease. Many novel methods for association testing have been developed. One key issue is how to evaluate the power of each method under realistic settings. Simulation is an efficient way to evaluate the ability of novel methods to detect the disease markers.

There are three main approaches for simulation (Liu 2008): 1) “backwards”, which starts with the samples that will form the simulated dataset, then works backwards in time to construct the genealogical information; 2) “forwards”, which starts with the entire population of individuals and then follows how all the genetic data are passed on from one generation to the next; 3) “Sidewards”, which starts with a collection of real genetic data, and uses these as a template for generating new simulated data with similar properties.

With the steady increase in publicly-available genomewide SNP data, such as the HapMap project and the 1000 genomes project, the potential advantage of the real data based simulation approach has been realized recently. Dudbridge (2006) proposed forming random diploid chromosomes from phased HapMap data followed by a single round of artificial meiosis, governed by empirical recombination rates also estimated from HapMap. This idea has been put to use in the ***HAP-SAMPLE*** software. Durrant (2004) proposed an alternative idea based on sliding windows for introducing new variation into simulated data. This method has been implemented in the ***GWAsimulator*** software. Jonathan Marchini’s ***hapgen*** software (2007), based on the same underlying principles as his genotype imputation software ***impute***, applies an approximation to the coalescent-with-recombination to generate new simulated data from existing phased HapMap data, but is slower than the other two sideways simulators.

Motivated by Dudbridge’s method, we proposed the causal region simulation which combines the strength from Bootstrapping method and Dudbridge’s strategy. In chapter 4, we described a real data based simulation program, SimuGeno, which can simulate large scale genomic data for case-control association studies by using HapMap data as a starting point.

SimuGeno provides three different ways to generate genotype data: bootstrapping, Dudbridge’s method and causal region simulation. Bootstrapping is fast, but it only provides limited genetic variation. Dudbridge provides more genetic variations which are generated from the simulated recombination and random mating. However, it is very slow when simulating whole genome data with large sample sizes. Causal region simulation, by combining the strength from Bootstrapping method and Dudbridge’s method, provides enough genetic variation and meanwhile keeps the running time short. SimuGeno, which currently only takes HapMap data as the original input data, can be easily extended to take other real data as input as long as that data contains a recombination map.

5.2 Future work

There are couples of topics which I am very interested in and may put some efforts on in the future. I will discuss two topics, gene-environment interaction and integrated association studies for SNP and CNV, in the following paragraphs.

CNV stands for a DNA segment that is 1 kb or larger and present at variable copy number in comparison with a reference genome (Feuk 2006). CNV can be created by deletions, insertions, duplications and complex multi-site variants (Redon 2006). By disrupting genes or altering gene dosage, CNV could effect gene expression, induce phenotypic variation and cause diseases. CNV associated diseases include CHARGE syndrome (Jongmans 2006), Parkinson’s (Singleton 2003) and Alzheimer’s disease (Rovelet-Lecrux 2006). Currently, there are two technology platforms widely used to assess CNV: comparative genome hybridization (CGH) and regular SNP chip (Redon 2006). The raw intensity data generated during SNP genotyping can be mined for copy-number information, making such studies a potential source of data for CNV-disease association studies (MacCarroll 2007). When I worked on my second project (Chapter 3), I noticed that it is possible to integrate association studies for SNPs and CNVs together. I am interested in exploring new methods in this direction.

Most of the complex diseases usually result from the interplay of genetic and environmental factors instead of genetic factors along or environmental factors along. One example is the much stronger effect of sunlight exposure on skin cancer risk in fair-skinned humans than in individuals with darker skin (Green 2002). Another example could be the ethnic differences in response to the exposure to the cigarette smoking in lung cancer (Haiman 2006). A recent report also demonstrated that the association between childhood asthma and exposure to large roadways was affected by genetic variation (McConnell 2006). There are different approaches to detect the disease susceptibility loci which are involved in gene-environmental interaction: 1) ignore the environmental exposure and detect the marginal effect of the loci (Clayton 2001); 2) screen markers for deviation from a odds ratio model for gene-environment interaction (Botto 2004); 3) use data mining methods to find disease predictors from genetic and environmental inputs (Cupples 2005). Kraft (2007) proposed a new likelihood ratio association test, allowing gene-environmental interaction. In their logistic regression model, they detect the G (gene only test), GE (gene-environmental interaction test), G-GE (joint test) separately. They compared the power and sample size requirements of different tests. They found that the joint test is nearly optimal across all penetrance models. In Kraft's study, they only considered the simplest situation which contains binary covariates for case-control study. In real world, continuous covariates and continuous response variables are more common, especially in pharmacogenetic study. For example, the drug response for obesity disease should be the weight lose, which is obviously a continuous variable. Thus extending Kraft's method to continuous response and continuous covariates would be an interesting research direction to go.

References

Abecasis GR, Cherny SS and Cardon LR (2001). The impact of genotyping error on family-based analysis of quantitative traits. Eur J Hum Genet 9: 130-134.

Affymetrix (2006). BRLMM: an improved genotype calling method for the GeneChip human mapping 500k array set. Affymetrix website.

Ahn K, Haynes C, Kim W, St. Fleur R, Gordon D and Finch SJ (2006). The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. Annals of Human Genetics 71: 249-261.

Akey JM, Zhang K, Xiong M, Doris P and Jin L (2001). The effect that genotyping errors have on te robustness of common linkage-disequilibrium measures. Am J Hum Genet 68: 1447-1456.

Akey J, Jin L, and Xiong M (2001). Haplotypes vs single marker linkage disequilibrium tests: What do we gain? Eur J Hum Genet **9**:291-300.

Allen AS and Satten Glen (2008). Robust estimation and testing of haplotype effects in case-control studies. Genetic Epidemiology 32:29-40.

Bakker, P., Yelensky R., Pe’er I., Gabriel S. B., Daly, M. J. and Altshuler, D. (2005). Efficiency and power in genetic association studies. Nature Genetics 37, 1217–1223.

Bierut LJ, Madden PAF, Breslau N, Johnson EO et al (2007). Novel genes identified in a high-density genome wide association study for nicotine dependence. Human Molecular Genetics 16: 24-35.

Botto L and Khoury M (2004). Human genome epidemiology: a scientific foundation for using genetic information to improve the health and prevent disease. Oxford, Oxford University Press.

Browning SR (2006). Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78(6): 903-13.

Cardon LR and Bell JI (2001). Association study designs for complex diseases. Nature reviews 2: 91-99.

Cawley S, Di X, Hubbell E, Lincoln S, Moorhead M, Short W, Speed TP, Sugnet C, Veitch J, Webster T, Williams A and Yang G (2006): BRLMM: an improved genotype calling method for the GeneChip human mapping 500k array set. ASHG Annual Meeting

Cheng KF and Lin WJ (2007): Simultaneously correcting for population stratification and for genotyping error in case-control association studies. The American Journal of Human Genetics 81:726-743

Clayton D and McKeigue PM (2001). Epidemiological methods for studying genes and environmental factors in complex disease. Lancet 358:1356-1360

Cordell HJ and Clayton DG (2005). Genetic association studies. Lancet 366:1121-1131.

Cupples LA, Bailey J and Cartier KC (2005). Data mining. Genet Epidemol 29(suppl 1):s103-s109.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005). Efficiency and power in genetic association studies. Nat Genet 37(11): 1217-23.

Di X, Matsuzaki H, Webster TA Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM, Kulp D, Kennedy GC, Mei R, Jones KW and Cawley S (2007). Dynamic model based algorithms for screening and genotyping over 100k SNPs on oligonucleotide microarrays. Bioinformatics 21: 1958-1963.

Dudbridge F. (2007). A note on permutation tests in multistage association scans. Am J Hum Genet 78(6):1094-1095.

Dudek, S. M., Motsinger A. A., Velez, D. R., Williams S. M. and Ritchie M. D. (2006). Data simulation software for whole-genome association and other studies in human genetics. Pacific Symposium on Biocomputing 11, 499-510.

Durham LK, Feingold E (1997). Genome scanning for segments shared identical by descent among distant relatives in isolated populations. Am J Hum Genet 61:830–842.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, et al. (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35–43.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfield M, Cohen D, Schork N (2001). Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer’s disease. Genome Res11:143-151

Feuk L, Marshall CR, Wintle RF and Scherer SW (2006). Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 (Suppl 1):R57-R66

Fujisawa H, Eguchi S, Ushijima M, Miyata S, Miki Y, Muto Y and Matsuura M (2004). Genotyping of single nucleotide polymorphism using model-based clustering. Bioinformatics 20: 718-726.

Gabriel SB et al (2002). The structure of haplotype blocks in the human genome. Science 296: 2225-2229.

Goldstein DR, Zhao H and Speed TP (1997). The effects of genotyping errors and interference on estimation of genetic distance. Hum Hered 47: 86-100.

Gordon D, Finch SJ, Nothnagel M and Ott J (2002). Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. Hum Hered 54: 22-33.

Green A and Trichopoulos D (2002). Textbook of Cancer Epidemiology 281-300.

Haiman CA, Stram DO and Wilkens LR (2006). Ethnic and racial differences in the smoking –related risk of lung cancer. N Engl J Med 354:333-342.

Hao K and Wang X (2004). Incorporating individual error rate into association test of unmatched case-control design. Human Heredity 58: 154-163.

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA and Freimer NB (1994). Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. Nature Genetics 8:380-386.

Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER and Stephan DA (2007). SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. Bioinformatics 23(1): 57-63.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.

Jongmans MC (2006). CHARGE syndrome: the phenotypic spectrum of mutations in the CHD7 gene. J Med Genet 43:306-314

Judson R, Sallisbury B, Schneider J, Windemuth A and Stephens JC (2002). How many SNPs does a genome-wide haplotype map require? Pharmacogenomics 3: 379-391.

Kang H, Qin ZS, Niu T and Liu JS (2004). Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. Am J Hum Genet 74:495-510.

Kang SJ, Gordon D and Finch SJ (2004). What SNP genotyping errors are most costly for genetic association studies? Genet Epidemiol 26: 132-141.

Kim S and Misra A (2007). SNP genotyping technologies and biomedical application. Annu Rev Biomed Eng 9:289-320.

Kraft P, Yen Y, Stram DO, Morrison J and Gauderman WJ (2007). Exploiting gene-environment interaction to detect genetic associations. Human Heredity 63:111-119

Laird NM and Lange C (2006). Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7:385-394.

Liu WM, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder TB, Webster TA, Dong S, Liu G, Jones KW, Kennedy GC and Kulp D (2003). Algorithms for large-scale genotyping mircoarrays. Bioinformatics 19: 2397-24.

Liu Y., Athanasiadis G., and Weale M. E. (2008). A survey of genetic simulation software for population and epidemiological studies. (Submitted)

Louis AL (1982). Finding the observed when using the EM algorithm. J. R. Statist Soc B 44: 226-23.

McCarroll SA and Altshuler DM (2007). Copy-number variation and association studies of human disease. Nature Genetics 39: s37-s42

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA and Hirschhorn JN (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature 9:356-369.

McConnell R, Berhane K and Yao L (2006). Traffic, susceptibility, and childhood asthma. Environ Health Perspect 114:766-772

McCullagh P and Nelder JA (1989). Generalized Linear Model.

McPeek M and Strahs A (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping. Am J Hum Genet 65:858-875.

Molitor J, Marjoram P, Thomas D (2003a). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73:1368–1384

Molitor J, Marjoram P, Thomas D (2003b). Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol 25:95-105.

Moorhead M, Hardenbol P, Siddiqui F, Falkowski M, Bruckner C, Ireland J, Jones HB, Jain M, Willis TD and Faham M (2006). Optimal genotype determination in highly multiplexed SNP data. European Journal of Human Genetics 14: 207-215.

Morris AP, Whittaker JC, Balding DJ (2002). Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet 70(3):686-707.

Morris AP, Whittaker JC, Xu CF, Hosking LK, Balding DJ (2003). Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. Proc Natl Acad Sci USA 100(23):13442-6.

Morris AP (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. Genet Epidemiol 29(2):91-107.

Morris RW ad Kaplan NL (2002). On the advantage of haplotype analysis in presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221-233.

Neale BM and Sham PC (2004). The future of association studies: gene-based analysis an replication. Am J Hum Genet 75:353-362.

Nicolae DL, Wu X, Miyake K and Cox NJ (2006). GEL: a genotype calling algorithm using empirical likelihood. Bioinformatics, Genome analysis 22: 1942-1947.

Nielsen DM, Ehm MG, Zaykin DV, and Weir BS (2004). Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics 168(2): 1029–1040.

Pe’er, I., Bakker, P., Maller, J., Yelensky, R., Altshuler, D. And Daly M. J. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nature Genetics 38, 663-667.

Plagnol V, Cooper JD, Todd JA and Calyton DG (2007): A method to address differential bias in genotyping in large-scale association studies. PLOS Genetics 3(5):759-767

Rabbee N and Speed TP (2006). A genptype calling algorithm for affymetrix SNP arrays. Bioinformatics Genome analysis 22: 7-12.

Redon R, Ishikawa S, Fitch KR and Feuk L (2006). Global variation in copy number in the human genome. Nature 23:444-454

Risch N (2000). Searching for genetic determinants in the new millennium. Nature 405:847-856.

Rovelet-Lecrux A (2006). APPlocus duplication causes autosomal dominant early onset Alzheimer disease with cerebral amyloid angiopathy. Nature Genet 38: 24-26

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425-434

Schaid D (2004). Evaluating associations of haplotypes with traits. Genetic Epidemiology 27:348-364.

Seltman H, Roeder K, Devlin B: Transmission/disequilibrium test meets measured haplotype analysis (2001). Family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 68:1250–1263.

Seltman H, Roeder K, Devlin B (2003). Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48-58.

Service SK, Temple Lange DW, Freimer NB, Sandkuijl LA (1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am J Hum Genet 64:1728–1738.

Sha Q, Dong J, Jiang R, Zhang S (2005). Tests of association between quantitative traits and haplotypes in a reduced-dimensional space. Ann Hum Genet 69:715–732.

Sha Q, Chen HS, Zhang S (2007). A new association test using haplotype similarity. Genet Epidemiol 31(6):577-93.

Singleton AB (2003). α-Synuclein locus triplication causes Parkinson's disease. Science 302:841.

Stephens JC et al (2001). Haplotype variation and linkage disequilibrium in 313 human genes. Science 293: 489-493.

Strahs A and McPeek M (2003). Multipoint fine-scale linkage disequilibrium mapping: importance of modeling background LD. Notes Monograph Series Volume 40 Science and Statistics: 343-366.

Su SY, Balding DJ and Coin JM (2008). Disease association tests by inferring ancestral haplotypes using a hidden markov model. Genetics and Population Analysis 24:972-978.

Syvanen AC (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature 2:930-942.

Teo YY, Lnouye M, Small KS, Gwillian R, Deloukas P, Kwiatkowski DP, Clark TG (2007). A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics, Genetics and population analysis 23: 2741-2746.

The International HapMap Consortium (2003). The International HapMap Project. Nature 426: 789-796.

The international HapMap consortium (2005). A haplotype map of the human genome. Nature 437: 1299-1320.

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

Thomos DC, Morrison JL, Clayton DG (2001). Bayes estimates of haplotype effects. Genet Epidemol Suppl 21:S712-S717.

Thomas DC, Stram DO, Conti D, Molitor J, Marjoram P (2003). Bayesian spatial modeling of haplotype associations. Hum Hered 56:32–40.

Tzeng J, Devlin B, Wasserman L, and Roeder K (2003a). On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J. Hum Genet 72:891-902.

Tzeng J.Y., Byerley W., Devlin B., Roeder K. and Wasserman L (2003b). Outlier detection and false discovery rates for whole-genome DNA matching. Journal of the American Statistical Association 98:236-246.

Tzeng JY (2005). Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol 28:220–231.

Tzeng JY, Wand CH, Kao JT, and Hsiao CK (2006). Regression-based association analysis with clustered haplotypes through use of genotype. Am J Hum Genet 78:231-242.

Tzeng J.Y., Chang S.M., Zhang D, Thomas DC, Davidian M (2008). Regression-based Multi-marker Analysis for Genome-wide Association Studies Using Haplotype Similarity. Institute of statistics mimeo series #2606.

Van der Meulen M and Meerman G (1997). Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genetic Epidemiology 14:915-920.

Xiao Y, Segal MR, YangYH and Yel RF (2007). A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. Bioinformatics, Genome analysis 23: 1459-1467.

Yu K, Xu J, Rao DC, Province M (2005). Using tree-based recursive partitioning methods to group hapltypyes for increased power in association studies. Ann Hum Genet 69: 577-589.

Zaitlen N, Kang HM, Eskin E, and Halperin E (2007). Leveraging the HapMap Correlation Structure in Association Studies. Am J Hum Genet 80(4): 683-691.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79-91

Zeggini E and Weedon MN et al (2007). Replication of genome-wide association signals in UK samples reveals risk loci for Type 2 Diabetes. Science 316: 1336-1341.

Zhang S, Sha Q, Chen HS, Dong J, Jiang R (2003). Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet 73:566–579.

Zhao JH, Curtis D and Sham PC (2000). Model-free analysis and permutation tests for allelic associations. Hum Hered 50:133-139

Zhu W and Guo J (2006). A likelihood-based method for haplotype association studies of case-control data with genotyping uncertainty. Science in China 49: 130-144.

Zöllner S, Pritchard JK (2005). Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169(2):1071-92.