MODELS AND METHODS FOR GENOME-WIDE ASSOCIATION STUDIES

by

Weihua Guan

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

Doctoral Committee:

Professor Gonçalo Abecasis, Co-Chair
Professor Michael Lee Boehnke, Co-Chair
Professor Roderick J. Little
Assistant Professor Jun Li
Associate Research Scientist Laura J. Scott

2010

To Hannah and Yu

Acknowledgements

It is a pleasure to acknowledge the group of people who have contributed to my thesis. Without their help, this dissertation would not have been possible.

First and foremost I want to thank Mike who has been very supportive throughout my graduate school study and guided me to grow up as an independent researcher. I would like to thank Gonçalo for his inspiring ideas and knowledgeable suggestions. I would like to thank Laura for her precious experiences of real studies and always helpful discussions. I also want to thank my other committee members: Rod and Jun, for their help and support during my writing of this dissertation.

I would like to thank my colleagues and fellow students in the Center for Statistical Genetics. They made my life as a graduate student joyful and unforgettable. I am especially grateful to Liming for our extensive collaborations in the first part of my dissertation. His knowledge and skills are essential to the success of our work.

I owe great gratitude to my family for their support and encouragement throughout my entire student life. Yu has always put great confidence on me and been supportive on everything. Hannah is such an angel that gives me enormous incentive to accomplish this work.

TABLE OF CONTENTS

List of Figures

Figure

List of Tables

Table

ABSTRACT

Genome-wide association (GWA) studies provide an extensive assessment of common genetic variants across the human genome for disease association. However, due to variation in allele frequencies and disease prevalence across populations, combining samples from different geographic or ethnic groups may lead to spurious evidence for association or diminish the true association signals. In part one of this dissertation, I propose a novel approach to correct for population stratification that makes use of the large amount of genetic information available in a GWA study. Based on allele-sharing identity-by-state (IBS) measures, I develop similarity scores that can describe genetic similarity between individuals, and match cases and controls accordingly. Association tests can then be performed conditional on the matched case-control groups. I apply our approach to the Pritzker bipolar GWA study.

In part two, I extend our matching approach to families of arbitrary structure. I first apply similarity score-based matching to selected members from each family and then assign other family members to the same matched group. I modify a corrected chi-square test [Bourgain et al., 2003] following the Mantel-Haenszel procedure to account for correlations both between the family samples and between the matched cases and controls.

The rapid advance in next-generation sequencing technologies allows a near-complete survey of genomic regions of interest and even whole genomes, enabling more extensive genetic association studies of rare variants. As we plan such re-sequencing studies of a complex disease, it is useful to consider the range of plausible genetic models, e.g., risk allele frequency (RAF) and genotype relative risk (GRR) of rare or less common causal variants, based on results of previous genetic linkage and association studies for the trait. In part three, I compute the power to detect linkage and/or association as a function of genetic model, and summarize the range of models likely to yield results that are consistent with existing GWA and/or linkage studies.

CHAPTER 1

INTRODUCTION

1.1 Complex Diseases and Previous Linkage Analysis

Unlike "Mendelian" diseases, most of which are rare and show a simple Mendelian inheritance pattern (dominant, recessive, X-linked), complex diseases often have relatively high population prevalence and show familial aggregation without following a simple Mendelian pattern. Complex diseases, such as coronary heart disease, hypertension, type 2 diabetes, obesity, various cancers, and Alzheimer's disease, are likely to be caused in part by a combination of multiple low-penetrance genetic variants together with environmental and behavioral factors, which makes mapping the disease predisposing variants difficult. Other factors that may contribute to disease susceptibility include rare genetic variants, structural variants including insertion-deletion polymorphisms or copy number variants (CNV), non-inheritable genetic effects, and epistatic interaction. For some diseases, such as psychiatric disorders, ambiguities in defining disease phenotype can also contribute to the difficulty of studying the genetic factors that underlie complex diseases.

Given the high incidence of many complex diseases, localizing and identifying the predisposing genetic variants could be of great public health importance. Geneticists have put

enormous efforts into understanding the underlying biology and etiology of such diseases. Earlier efforts focused on linkage analysis, which assesses the co-transmission within families of variant marker alleles and disease. A genome-wide linkage scan typically examines several hundred to a few thousand genetic markers on families ascertained to include multiple affected individuals. Due to the small number of recombination events on each chromosome per meiosis, there is noticeable dependence of transmission for loci as much as 10-20 cM (typically equivalent to ~10-20 Mb) apart. Therefore, even if linkage is found, it is necessary to carry out subsequent fine mapping and positional candidate gene studies, in which a denser set of genetic markers is scrutinized across the linked regions, to narrow down the search for the disease predisposing variant(s). Linkage analysis has achieved remarkable successes in the last 30 years in mapping the genetic defects for >1000 Mendelian traits [Bostein and Risch, 2003].

For the complex diseases, however, linkage analysis has not proven nearly as successful. For type 2 diabetes, for example, >20 genome-wide linkage studies have been carried out, but only a few studies have reported significant linkage results, and meta-analysis has failed to identify any single location in the genome with convincing evidence of linkage [Guan et al., 2008]. Even for other diseases and traits for which convincing evidence for linkage has been identified, there has been only limited success in identifying the true genetic causal variants under the linkage peak.

Complex diseases typically involve predisposing variants with small effects, which linkage is often underpowered to detect. In addition, large numbers of family samples are difficult to collect. Because of these difficulties Risch and Merikangas [1996] suggested given the same

sample size, association tests may have greater power to detect common variants predisposing to complex diseases than linkage analysis, if a large number of genetic markers across the genome could be identified and genotyped.

1.2 Genome-wide Association (GWA) Studies

High-throughput genotyping and DNA sequencing have made huge advances in the intervening years. With the completion of the Human Genome Project in 2003 [International Human Genome Sequencing Consortium, 2004] and the International HapMap Project in 2007 [The International HapMap Consortium, 2007], deep surveys of the human genome providing good coverage of the entire genome or specific genetic regions of interest have become feasible. These advances have made the sort of study envisioned by Risch and Merikangas a reality, allowing investigators to (with high probability) genotype genetic markers that are in linkage disequilibrium (LD) with the disease predisposing variants, or even the disease predisposing variants themselves. These technologies make it possible and cost effective to phenotype and genotype unrelated individuals and carry out large-scale association studies instead of linkage studies, providing more power and greater accuracy to pinpoint common disease-predisposing variants.

With these advances, genome-wide association (GWA) studies have become the standard approach to map genes for complex diseases [Altshuler et al., 2008; Manolio et al., 2008; McCarthy et al., 2008; www.genome.gov/gwastudies]. GWA studies typically assay hundreds of thousands of SNPs across the genome representing, directly or indirectly 80-90% of known

common human genetic variants. By comparing allele frequencies between different phenotypic groups (such as cases and controls), GWA studies can detect SNPs that are associated with the disease. GWA studies are preferred by many researchers not only because they potentially provide more accurate localization of causal genes than linkage studies, but because they provide an unbiased ("hypothesis-free") search of the entire genome for association compared to candidate-gene studies. Furthermore, GWA studies are more flexible in terms of study design, not being restricted to family samples as in linkage studies. It is often possible to collect thousands of samples with collaborative efforts from multiple research groups, increasing the chance to detect disease predisposing variants with small to modest effects.

1.3 The Scope of this Dissertation

Epidemiological and genetic studies have shown that disease prevalence and allele frequencies can vary across human populations. Combining samples from different geographic or ethnic groups may then lead to confounding due to population stratification, which can result in spurious evidence for association or mask true association signals. In practice, ethnicity information is often not collected or not accurately recorded. In Chapter 2, we propose a novel approach to correct for population stratification that can make good use of the large amount of relationship information provide by GWA studies. Based on allele-sharing identity-by-state (IBS) measures, we develop similarity scores that quantify the degree of genetic similarity between individuals, and match cases and controls accordingly. An association test can then be performed conditional on the matched case-control groups. This individual-based matching scheme is flexible in that it does not require pre-determining the number of subpopulations in the samples,

and has advantages in controlling for potential outliers.

GWA studies are not necessarily restricted to unrelated samples. Due to sample availability or historical reasons (such as association studies based on previous linkage studies), family samples are often used for association studies. In Chapter 3, I extend our matching approach proposed in Chapter 2 to adapt to families of arbitrary structure. The similarity score-based matching can be applied to selected members from each family and then be extended to other family members. Bourgain et al. [2003] proposed a corrected chi-squared test for disease-marker association when related samples are present. We modify this test statistic following a Mantel-Haenszel procedure to account for correlations both between the family samples and between the matched cases and controls. Results from this work will help to correct for population stratification in GWA studies while efficiently utilizing all available samples.

Although GWA studies have been successful in identifying hundreds of common variants that are very likely to be associated with real biological causation in complex diseases, for most diseases, these variants account for a small proportion of disease risk, in aggregate often <10% [Visscher et al., 2008]. While studies with larger sample sizes are being undertaken in hopes of finding additional common disease predisposing variants, the search for less common genetic variants has now accelerated, owing to the rapid development in DNA sequencing technology. It is therefore of great interest to explore the plausible range of genetic models underlying common diseases, particularly for variants not well covered in current GWA studies, such as those with minor allele frequency (MAF) of .05 or less. Genetic risk variants with very low allele frequency could be in linkage equilibrium (LD) with common SNPs and potentially be responsible for the

association results in GWA studies [Dickson et al., 2010]. They could also be directly detected by linkage analysis due to joint inheritance with the genotyped markers during meiosis. However, linkage studies of many complex diseases have not reported significant results or the findings could not be replicated. In Chapter 4, we address the range of genetic models, incorporating genotype relative risk (GRR) and risk allele frequency (RAF) of the rare or less common causal variants, consistent with results of existing linkage and GWA studies. We perform power calculations for linkage and association tests for a range of genetic models. We evaluate the range of models that could result in little evidence for linkage, and/or little or strong evidence for association. From the range of consistent models, we provide guidance for the design and analysis of future large-scale sequencing studies.

CHAPTER 2

GENOTYPE-BASED MATCHING TO CORRECT FOR POPULATION STRATIFICATION IN LARGE-SCALE CASE-CONTROL GENETIC ASSOCIATION STUDIES

Genome-wide association studies are helping to dissect the etiology of complex diseases. Although case-control association tests are generally more powerful than family-based association tests, population stratification can lead to spurious disease-marker association or mask a true association. Several methods have been proposed to match cases and controls prior to genotyping, using family information or epidemiological data, or using genotype data for a modest number of genetic markers. Here, we describe a genetic similarity score matching (GSM) method for efficient matched analysis of cases and controls in a genome-wide or large-scale candidate gene association study. GSM is comprised of three steps: 1) calculating similarity scores for pairs of individuals using the genotype data; 2) matching sets of cases and controls based on the similarity scores so that matched cases and controls have similar genetic background; and 3) using conditional logistic regression to perform association tests. Through computer simulation we show that GSM correctly controls false positive rates and improves power to detect true disease predisposing variants. We compare GSM to genomic control using computer simulations, and find improved power using GSM. We suggest that initial matching of cases and controls prior to genotyping combined with careful re-matching after genotyping is a method of choice for genome-wide association studies.

Joint Work with Liming Liang

This chapter was a joint work with Liming Liang, PhD candidate in the Department of Biostatistics at the University of Michigan. While all work were done interactively with discussion, exchanging ideas, motivations from findings of each other and sharing codes with each other, I have been focusing on the design and evaluation of different similarity scores, implementation of the optimal matching and conditional logistic regression for disease-marker association, evaluation of the simulation results, and application of our method to real data (e.g., HapMap and bipolar data).

2.1 Introduction

With the success of the International HapMap Project [The International HapMap Consortium, 2007], a dense set of single nucleotide polymorphisms (SNPs) throughout the human genome is now available for genetic studies of complex diseases, and many genome-wide association studies are being undertaken and published [Klein et al., 2005; Maraganore et al., 2005; Cheung et al., 2005; Sladek et al., 2007; Scott et al., 2007; Saxena et al., 2007; Zeggini et al., 2007].

Although case-control association tests are in principle more powerful for detecting disease variants than family-based association tests, population stratification can lead to spurious disease-marker association or mask true association [Li, 1972]. In genome-wide association studies, thousands of samples are typically used to ensure adequate power to identify disease predisposing variants, making it difficult to guarantee genetic homogeneity of the sample

[Freedman et al., 2004]. Ancestry information on the sampled individuals may be unavailable to the researchers, and even when available, may not fully specify the underlying population genetic structure, due to vague definitions of ancestry groups and imperfect accuracy of self-report information.

Several methods have been proposed to adjust for the possible confounding effects of population substructure. Family-based association tests, such as the transmission/disequilibrium test (TDT) [Spielman et al., 1993], assess the transmission of alleles from parents to affected offspring. Comparisons are made within parent-offspring trios, and the resulting association test is immune to potential genetic heterogeneity between families. However, collecting trios can be difficult and expensive, and may simply be impractical for late-onset diseases. For unrelated case-control samples, approaches have been proposed to adjust the standard chi-square contingency test statistics according to a non-central χ^2 distribution [Devlin et al., 1999; Gorroochurn et al., 2006], to infer population structure [Pritchard et al., 2000], or to cluster the similarity estimates into several components [Zhang et al., 2002]. A few more recent approaches [Price et al., 2006; Epstein et al., 2007; Kimmel et al., 2007; Luca et al., 2008] focus specifically on genome-wide association studies.

In this paper, we propose a different approach, genetic similarity score matching (GSM), to correct population stratification using individual-based matching rather than clustering. The huge amounts of data in genome-wide association studies have the potential to provide extremely accurate matching of individuals who share similar ancestries. We match cases with controls based on genetic (dis)similarity scores calculated from the genotype data available in a

genome-wide association study or a large-scale candidate gene study and test the resulting matched sets for disease-marker association by conditional logistic regression. This matching-association framework builds on our previous work [Guan et al. 2005] and is similar to that of Luca et al. [2008]. Luca et al. [2008] derive the dissimilarity (distance) scores based on principal components of the variance matrix of genotypes, while our approach obtains the dissimilarity scores based on identity-by-state (IBS) measures. Simulations show that GSM results in false positive rates at the desired nominal level while retaining high power to detect disease associated markers. We find that with large-scale association data, the calculated genetic similarity scores differentiate subpopulations well, and that matching can be done with high accuracy even for samples that are mixtures of genetically similar populations. We further demonstrate that when population stratification is present, association tests based on GSM-matched case-control data can have a higher power than those that rely on either the standard trend test or the genomic-control method.

2.2 Methods

2.2.1 Outline

GSM includes three basic components:

1) *Genetic similarity score:* We calculate genetic similarity scores between pairs of cases and controls across all loci. Large scores should reflect pairs with similar genetic backgrounds.

2) *Matching:* Based on the matrix of similarity scores calculated in 1), we conduct optimal full matching [Rosenbaum, 2002] which groups one case with one or more controls, or one

control with one or more cases to maximize the overall similarity of matched cases and controls.

3) *Association tests:* We use conditional logistic regression to assess the association between candidate markers and disease status. For ease of exposition, we consider here only single marker association tests, but other genetic or environmental factors can be easily incorporated into the regression.

2.2.2 Genetic Similarity Score

We define a genetic similarity score for a pair of individuals which measures the degree of similarity of their genotype data. Individuals with similar genetic backgrounds will generally have higher scores. For simplicity, we consider M biallelic genetic markers each with alleles "A" and "a"; the scores can easily be generalized to multiallelic markers. We consider three similarity scores.

The first score calculates the proportion of marker alleles shared identity-by-state (IBS), a simple measure for the number of alleles shared between pair of individuals. If IBS_k is the number of alleles shared at marker k (Table 2.1), then

$$S_{IBS} = \frac{1}{2M^*} \sum_{k=1}^{M^*} IBS_k \qquad (1)$$

where $1 \le M^* \le M$ is the number of markers that are successfully genotyped in both individuals.

While S_{IBS} has the virtue of simplicity, we may want to allow different markers to make different contributions to measure similarity. For example, we may wish to weight sharing a rare

allele more strongly than sharing a common allele. We define our second score as:

$$S_{freq} = -\frac{1}{2M^*}\sum_{k=1}^{M^*}\sum_{i\in\{A,a\}} IBS_{k,i}\cdot\log(q_{k,i}) \qquad (2)$$

where $q_{k,i}$ is the frequency of allele i at marker k, and $IBS_{k,i}$ is the number of copies of allele i at marker k shared by the pair of individuals (Table 2.1). We can estimate $q_{k,i}$ using our sample or from the results of previous studies.

In a random mating population, markers are expected to follow Hardy-Weinberg Equilibrium (HWE). When population subdivision is present, tests of HWE tend to be significant owing to excess homozygosity. Our third score takes advantage of this by weighting markers based on their one-sided (excess homozygosity) HWE test p-value p_k [Wigginton et al., 2005]:

$$S_{HWE} = -\frac{1}{2M^*}\sum_{k=1}^{M^*} IBS_k\cdot\log(p_k) \qquad (3)$$

To avoid the impact of genotyping error that may lead to strong deviation from HWE, we exclude the markers that fail quality control; practically speaking, this might mean using markers with HWE p-value satisfying $p>10^{-6}$.

As an example, suppose 3 cases and 3 controls are genotyped at 3 loci, as listed in Table 2.2. Then the similarity scores S_{IBS} are as listed in Table 2.3.

For matching, we may use all genotyped markers, or a selected subset. For example, we might pick the markers with the smallest p-values in a HWE test for excess homozygosity, excluding those that fail quality control, in the hope that the selected markers provide maximal information about population stratification in the sample. Further, to avoid selecting markers

which are highly correlated, we might choose at most one marker in every *n*-marker window or per linkage disequilibrium group.

In our analyses, matching relies on a transformed dissimilarity score, defined as:

$$D_{ij} = f(S_{ij}) = \left(\frac{\max - S_{ij}}{\max - \min} \right)^2 \quad (4)$$

where max = $\max_{i,j} S_{ij}$ and min = $\min_{i,j} S_{ij}$, the maximum and minimum similarity scores among all case-control pairs.

2.2.3 Matching

We use the chosen (dis)similarity score to identify optimal matches between cases and controls. The simplest matching scheme is a 1:1 match in which each case is matched to a unique control. This approach is widely used but has obvious drawbacks. For example, when the numbers of cases and controls are not equal, some subjects must be discarded, resulting in a loss of information. Further, samples from various subpopulations often are not equally represented among the cases and controls, leading to forced mismatches if only 1:1 matching is allowed.

Instead, we consider an optimal matching approach that minimizes the total dissimilarity score:

$$T = \sum_{s=1}^{S} \sum_{i \in A_s, j \in B_s} D_{ij}$$

Here, A_s and B_s are the sets of cases and controls in a matched set s, and S is the total number of matched sets. It has been shown that an optimal solution to this minimization problem is a full

matching, in which each matched set contains one case and one or more controls, or one control and one or more cases, that is, a 1:*m* or *m*:1 matching [Rosenbaum, 1991]. Given *n* cases and *n* controls, the summation can in principle contain as few as *n* terms for 1:1 matching to as many as 2(*n*-1) terms for 1:*n*-1 and *n*-1:1 matching. Since large sets result in larger numbers of terms, optimization tends to favor small matched sets. This helps mitigate any potential power loss due to unbalanced matching, i.e., 1:*m* or *m*:1 matching with $m>>1$ (see Discussion).

The problem of minimizing the total dissimilarity score *T* is analogous to the classic minimum cost flow (MCF) problem in computer science [Rosenbaum, 1991; Hansen, 2004; Hansen et al., 2006] (Appendix), and can be solved using the RELAX-IV algorithm [Bertsekas et al., 1994; Frangioni et al., 2006]. Given pre-calculated dissimilarity scores and an upper bound on *m*, determining the optimal matched set takes on the order of $n^3 \log n$ operations, where *n* is the total number of subjects. The choice of parameter *m* constrains the size of matched sets and is somewhat arbitrary; we typically require $m \leq 5$ when numbers of cases and controls are comparable (see Discussion). Prior to matching, we may exclude a few individuals with maximum similarity scores that are extremely small (this is the *caliper* parameter recommended by Hansen et al., 2006). In datasets including ~2,000 individuals, the matching typically takes <1 minute on a modern PC workstation.

To continue with the previous example, we calculate the dissimilarity scores in Table 2.3, and perform both 1:1 matching and optimal matching. In 1:1 matching, the best match yields three pairs: (1, 4), (2, 5), and (3, 6). The total dissimilarity score is 1/36 + 16/36 + 0 = 17/36. In contrast, the optimal full match has two matched sets: (1, 2, 4) and (3, 5, 6). The matched sets

include 4 case-control pairs: (1, 4), (2, 4), (3, 5), and (3, 6). The total dissimilarity score is 1/36 + 0 + 1/36 + 0 = 2/36. In this example, the individuals within group (1, 2, 4) and (3, 5, 6) are similar to each other, and less similar to the individuals in the other group. Full matching offers an obvious matching advantage over 1:1 matching here. In the general case, full matching is guaranteed to produce a total dissimilarity score that is no greater than that obtained using 1:1 matching.

2.2.4 Conditional Logistic Regression

Once matching is done, a natural choice for matched-set analysis is to use conditional logistic regression to test for disease-marker association. We employ an additive model for association by assigning values of 0, 1, and 2 to genotypes AA, Aa, and aa, respectively. Other genotyping coding schemes could be considered, corresponding for example to dominant, recessive, or general models. The regression can easily incorporate genotype, covariate, and interaction effects.

In a genome-wide association scan, we apply conditional logistic regression analysis to each marker separately. The multiple testing problem can be addressed using Bonferroni correction, permutation, or false-discovery rates.

2.3 Simulation

We simulated case-control data influenced by genotypes at a disease locus with alleles D and

d, under six additive disease models (Table 2.4). We assumed sampling from a population that consisted of two subpopulations. We randomly sampled 500 cases and 500 controls from this mixed population. For each model, the relative risk (RR) of the predisposing variant allele is set to be the same in different populations. For models 1 and 2, the disease prevalences $K_1=K_2$ and predisposing variant allele frequencies $q_1=q_2$; these models represent the scenario of no population stratification. For models 3 and 4, $K_1<K_2$, creating population stratification in the simulated data. For models 5 and 6, $K_1<K_2$ and $q_1\neq q_2$. For model 5, the first population has lower prevalence but higher predisposing variant allele frequency ($K_1=.07$, $q_1=.55$), than the second population ($K_2=.13$, $q_2=.45$). For model 6, the population with higher prevalence also has higher predisposing variant allele frequency ($K_2=.13$, $q_2=.55$) than the other population ($K_1=.07$, $q_1=.45$). For each model, we simulated 500 datasets.

We simulated autosomal SNPs using GENOME, a coalescent-based simulator [Hudson, 1983; Hudson, 1990; Donnelly et al., 1995; Liang et al., 2007]. Assuming discrete generations, GENOME simulates the genealogy of a sample of sequences. As the algorithm proceeds backwards in time, coalescence, recombination, and migration events are simulated. Multiple events can occur in the same generation. We set the effective population size as 10,000, the recombination rate as 10^{-8} per base pair, and the mutation rate as 10^{-9} per base pair, assuming the infinite-site mutation model [Kimura, 1969]. We set the rate of migration between subpopulations to .0025 per individual per generation, which resulted in a distribution of allele frequency differences similar to that observed when comparing HapMap Han Chinese (HCB) and Japanese (JPT) samples (www.hapmap.org). In particular, the mean allele frequency difference between the two simulated populations is .0470, compared to .0477 between the HCB

and JPT samples. The simulated genome scans surveyed autosomal genomes of ~2866 Mb comprised of 22 chromosomes, whose lengths approximate the actual lengths of the human autosomes (NCBI build 33, www.ncbi.nlm.nih.gov/genome/seq/). We randomly selected 300,000 SNPs with minor allele frequencies > .05, and choose a disease liability locus with the desired allele frequencies.

To calculate the similarity scores, we used 10,000 markers with the smallest one-sided HWE p-values, choosing no more than one marker from each 10-marker window. We set the maximum size of matched groups (m) to 6. We compared the type I error and power of GSM, the trend test, genomic control, and EIGENSTRAT for each simulated setting. Given that the simulated samples were drawn from two subpopulations, we used the first principal component to adjust for stratification in EIGENSTRAT; using additional principal components gave similar results. The estimated type I error rates are the proportion of simulated SNPs in which the association test p-value is less than the nominal value 10^{-6}, a significance threshold similar to that typically used in genome-wide scans. In this evaluation of type I error rates, we only considered SNPs that were effectively unlinked to the disease locus. We calculated power as the proportion of simulated replicates where the empirical p-value is $< 10^{-6}$ at the disease locus using a threshold obtained by inspection of test statistics at the null loci.

2.4 Bipolar data

We applied GSM to genome-wide association data from the Pritzker Consortium bipolar study [Scott et al. 2009]. We selected 717 independent bipolar I European American cases and

779 independent European American controls from NIMH Human Genetics Initiative (www.nimhgenetics.org); controls were carefully matched to cases by self-reported ethnicity prior to genotyping. In addition, we downloaded genotype data on 3,182 independent European American controls from Illumina iControlIDB database (www.illumina.com/pages.ilmn?ID=231). All individuals were genotyped using the Illumina HumanHap550 BeadChip. 505,796 autosomal SNPs passed quality-control criteria in the Prtizker bipolar study: 1) HWE p-values > 10^{-5}; 2) genotype call rate > 95%; and 3) no more than 1 non-mendelian inheritance or inconsistency among 15 father-mother-offspring trios and 30 duplicate samples. Of these, we excluded 1,632 SNPs due to allele frequency differences > .05 between the Illumina and Pritzker control samples. We applied GSM and trend tests for association on the Pritzker samples alone and then on the combined Pritzker and Illumina samples. In GSM, we used the 100,000 markers that passed quality control and have the smallest p-values from the one-sided HWE test to calculate the similarity scores. Given the relatively large control:case ratio of 3,961/717≈5.5, we set the upper limit of the group sizes (m) to 30.

2.5 Results

2.5.1 Similarity score performance in HapMap

We first examined the performance of our similarity scores in the HapMap dataset. We calculated our three similarity scores for all pairs of the 89 independent Han Chinese (CHB) and Japanese (JPT) individuals in the HapMap sample, using 100,000 HapMap phase I autosomal

SNPs with MAF>.05, selected based on one-sided Hardy-Weinberg equilibrium test p-values of 4.3×10^{-6} to .11. In Figure 2.1, we showed plots from using multidimensional scaling (MDS) on the similarity score matrices. All three scores showed good separation between the two populations, except for one JPT individual residing in between the two clusters in the plots. The same individual is at a similar position in principal component analysis when plotting the first two principal components. While S_{IBS} and S_{freq} provided similar separation, S_{HWE} provided less separation with that JPT individual much closer to the CHB cluster instead of the JPT cluster. The relatively poorer performance of S_{HWE} arises because of the heavy weighting of the small subset of markers with very small p-values from the one-sided HWE test, even after we have excluded markers with HWE p-value$<10^{-6}$.

Our experiences in simulations and real data (unpublished results) suggest that S_{freq} may perform slightly better than S_{IBS} in matching the samples. In the following simulations and analyses, we report results using S_{freq} as our measure of genetic similarity. Although the p-values from one-sided HWE test may not be the best weights for the similarity score as in S_{HWE}, they can still be employed to select a subset of markers for the score computation. In so doing we assume that markers with small HWE p-values but still passing quality control provide more information about population heterogeneity than randomly selected markers. In the following analyses, the matching is usually based on a subset of markers (10,000-100,000 markers) which had the smallest p-values from one-sided HWE test among those passing quality control filters.

2.5.2 False positive rate and power

For the six simulation models, mismatch rates are calculated as the proportion of individuals from population 1 matched to individuals from population 2. The minimal degree of mismatch in the simulations (Table 2.5) suggests accurate matching given the similarity measures and numbers of markers used.

In the absence of population stratification (models 1 and 2), all three methods give false positive rates close to the nominal value of 10^{-6}. The power of our GSM method is typically ~2% lower than the trend test and genomic control, assumedly due to the unnecessary grouping of samples. When population stratification is present (models 3-6), the type I error rate of the trend test is ~30 times greater than the nominal value, while GSM and genomic control maintain the type I errors at or lower than the nominal value. Using empirical type I error rates, the power of the trend test is equal to that of genomic control, but significantly lower than that of GSM for models 3-4. For models 5 and 6, where population stratification is present, the variation of disease variant frequency may mask the association (model 5) or increase the power to detect association (model 6). For model 5, power of the trend test and genomic control drop ~30% compared to model 3, while GSM maintains the same level of power. For model 6, although the type I error is inflated, the trend test has adjusted power comparable to that of GSM. EIGENSTRAT has power similar to GSM in all simulation settings examined.

We also compared the frequency with which the disease variant is the most strongly associated marker, or among the most strongly associated 10, 100, and 1000 markers, in the trend test or GSM (Figure 2.2). The results are consistent with the observations above. In the absence of population stratification (models 1 and 2), the trend test identifies the disease variant slightly

more frequently than GSM. When population stratification is present, GSM picks the correct disease variant more frequently for models 3-5. For model 6, GSM picks the correct disease variant almost as frequently as the trend test.

2.5.3 Bipolar data

We first applied standard trend tests to the Pritzker bipolar case and control samples. The estimated genomic control variance inflation factor λ of the test statistics was 1.03, close to the expected value of 1 when there is no population stratification [Devlin et al., 1999], arguing that the matching based on self-reported ethnicity resulted in a sample with only limited population stratification. Applying GSM reduced the estimated λ slightly to 1.02. However, when we added the Illumina control samples to the analysis, the estimated λ from standard trend tests became 1.51 (Figure 2.3), indicative of strong population stratification between the cases and controls. We then applied our GSM method on the combined samples, excluding one Illumina control sample that had a noticeably high similarity score with one Pritzker case sample (S_{IBS}=0.85), consistent with a first degree relationship. Using GSM, the estimated λ dropped to 1.072 when we used S_{freq} as our similarity measure (Figure 2.3) and 1.088 using S_{IBS}, suggesting that GSM using either score provided good correction for the stratification problem. Using S_{freq}, each of the 712 cases was matched to one or more controls (i.e., 1:*m* matching only): 316 cases were matched to 1 control, 207 cases to 2-5 controls, 79 cases to 6-10 controls, and 115 cases to 10-30 controls. To check the appropriateness of setting the maximum number of controls (*m*) at 30, we repeated our analysis by changing *m* to 10 or 50, resulting in estimated λ values of 1.23 and 1.067, respectively. This suggests that some controls may be matched to dissimilar cases when

we only allow up to 10 controls per case, while increasing *m* from 30 to 50 resulted in little improvement on the matching. Since the combined sample contains many more controls than cases, we considered removing some controls with relatively high dissimilarity by restricting the total number of controls to be matched from 3,960 to 3,500, and the estimated λ dropped slightly to 1.065. We also repeated the matching using 50,000 markers instead of 100,000, and in this setting the estimated λ increased slightly to 1.086, as expected.

As a comparison, we also applied EIGENSTRAT and another principal component-based method (Luca et al. [2008], GEM) to the bipolar data, using 10 principal components. Without removing any potential outliers, EIGENSTRAT gave an estimated λ of 1.074, comparable to our results. GEM removed 132 samples as outliers and gave a slightly better estimated λ of 1.063. When we applied our method to the same set of samples used in GEM, we obtained an estimate λ of 1.065. Although the removal of these samples decreased the inflation of type I error rates, its impact on power requires further investigation.

2.6 Discussion

Population stratification, which can result in high false-positive rates and mask true associations, poses a potential problem for case-control association studies. In this paper, we propose GSM, a practical approach to correct for population stratification for large-scale association studies that uses information at thousands of genotyped genetic markers to group case and control subjects according to their similarity. Simulation studies show that GSM can control the false positive rates in the presence of population substructure, while maintaining

power to detect disease loci.

GSM is computationally efficient. The computational time for similarity score calculation is linear in the number of markers used and in the number of all case-control pairs, and approximately cubic in the number of individuals.

We have compared the performance of GSM to the commonly used genomic control method [Devlin et al., 1999]. Genomic control assumes that a scaled test statistic (dividing the standard test statistic by a global correction factor λ) has an approximate central χ^2 distribution. When stratification is modest, the genomic control procedure is able to control the false-positive rate at the nominal level through λ, but does not change the relative order of the test statistics along the genome. As shown in our simulations (model 3-5), when stratification masks the association, genomic control can be quite conservative. Another popular approach to correct for population stratification is structured association [Pritchard et al., 2000] which infers population structure using a set of independent makers. We did not evaluate this method in our simulations due to its computational intensity. Structured association also requires an assumption about the number of underlying subpopulations in the sample. EIGENSTRAT [Price et al., 2006] is an approach for genome-wide association studies based on principal components analysis (PCA). It has been shown that the K-1 principal components can be related to the solution to the K-way clustering solution [Ding et al., 2004]. EIGENSTRAT is less sensitive to the number of components than structured association (if the number is sufficiently large) because of orthogonality of the axes of variation, but the interpretation of the axes is less intuitive.

Our new GSM method tackles the stratification problem by matching at the individual level, without assuming an explicit population structure. Effectively, it treats every sample as a single population and compares it to the most similar counterparts. For samples from clearly distinguished subpopulations, such as the HapMap HCB and JPT populations or the two subpopulations in our simulations, GSM performs almost as well as cluster-based matching or EIGENSTRAT, with little loss of power. In real GWA studies, where sampled individuals may often derive from continuous mixtures of ancestral populations, the individual-based matching in GSM should be more flexible than cluster-based matching. Luca et al. [2008] (GEM) also applied full matching to correct for population stratification, but used a different score calculated from the top eigenvectors from PCA. They showed that outliers may greatly inflate type I errors of association tests using EIGENSTRAT and need to be carefully removed beforehand. The similarity scores in GSM can be used like the GEM scores to identify outliers, but are more intuitive in measuring genetic similarity, compared to the abstract measures from eigenvectors used in GEM. In addition, PCA analysis is very sensitive to the independence of samples, while GSM can actually help to identify related samples through IBS scores. In our Pritzker study example, we checked the pairwise IBS measure between samples and found one pair of individuals with large similarity score of 0.85 (S_{IBS}), which strongly suggested a potential first-degree relative. Although the two samples showed strong correlation in their PC scores, they were not identified as outliers by EIGENSTRAT or GEM because their scores did not show strong deviation from the center of the score distributions in the top 10 PCs.

The success of our GSM procedure depends on the accuracy of matching. Incorrectly grouping individuals from different populations could inflate the type I error rate, decrease the

power to detect the susceptibility genes, or both. To ensure correct matching, a well-defined similarity measure and a substantial number of markers in which to compute this measure are both important. We have considered both IBS and IBD-based scores, which are commonly used in genetic studies to describe genetic similarity between pairs of individuals. IBS-based scores are simple to calculate and do not require much computing power. Our simulations analysis and practical experience show that such scores can provide an effective means of matching individuals. Furthermore, we found that weighting IBS estimates by a function of the marker allele frequencies (S_{freq}) improved the accuracy of matching. We also considered similarity scores based on pairwise IBD estimates calculated using an E-M algorithm; the average mismatch rates using IBD-based scores were slightly higher than those for IBS-based scores. A weakness of IBD based scores is that they are truncated at zero: when many pairs of individuals are assigned IBD ~0, it becomes difficult to select optimal pairings. Figure 2.4 demonstrates the relationship between the IBD scores and IBS scores (S_{freq}) computed on the HapMap HCB and JPT samples. Other score metrics to describe similarity of samples can also be easily incorporated into our approach.

The number of markers used in score calculation is another factor that affects the matching. We prefer to calculate the scores based on a large set of markers (typically including 10,000 – 100,000 SNPs). However, using too many markers increases the computational load while not necessarily improving the accuracy of matching. In our simulations, 10,000 markers with the smallest p-values from one-sided HWE test can correctly match the individuals from closely related populations such as Han Chinese and Japanese, with zero or almost zero mismatch (Table 2.5). In this example, using 30,000 markers worked as well as using 10,000 markers, while using

only 1,000 markers led to incorrect grouping of individuals from different populations with up to ~10% mispaired individuals. For samples with subtle differences in genetic ancestry, such as the European American samples in the bipolar data, more markers (50,000 to 100,000, passing quality control) may help to obtain better matching. Inspecting the genomic control parameter λ on its closeness to the expected value of 1 from different analysis strategies can help to determine the appropriate number of markers for controlling stratification. To select the subset of markers, we usually prefer those with smaller p-values from one-sided HWE tests, because they tend to be more informative about population structure. However, we need to be cautious regarding data quality, since markers with high error rates may show strong deviation from HWE and then give incorrect information about the genetic background of sampled individuals. A reasonable compromise is to exclude SNPs with extreme deviations from HWE (say, $p<10^{-6}$) but focus on those with mild deviations (say, $10^{-2}<p<10^{-6}$) to evaluate stratification. GSM does not require that *all* markers should be independent of disease status, since in a typical genome-wide setting the vast majority of markers will meet this criterion and the impact of disease-associated markers on the similarity scores is negligible and can be ignored. For the bipolar data, we have excluded the top 100 SNPs with the smallest p-values and re-ran our GSM program. The new results are highly correlated with the original ones (Pearson correlation coefficient = 0.98 between the test statistics). Furthermore, since our similarity scores are a function of the mean (weighted) IBS values across a large number of markers, it is also not critical that the assessed SNPs should be independent of each other.

The application of GSM to the bipolar data showed great reduction of the false positive signals (Figure 2.3). The estimate GC λ dropped from 1.51 of the trend test to 1.07 using GSM.

The PCA-based method implemented in EIGENSTRAT gave similar reduction. Although the reduction is significant, the estimated λ is still a little off from the ideal value of 1. One possible reason for this slight inflation of false positive rates is due to outliers in the sample, e.g., individuals with other ethnic background were mislabeled as European Americans. We have performed outlier removal procedures implemented in EIGENSTRAT and GEM, and caliper matching in GSM, but the further reduction on λ is minimal (from 1.07 to 1.06). We suspect that there might be slight systematic difference between the NIMH and Illumina samples that can cause more false positives than expected, since the two datasets were collected and genotyped by different groups at different time periods.

We chose not to include X-linked markers in our matching scheme to avoid any possible biases due to differences by gender. Given genome-wide association data, the autosomal markers provide ample information for accurate matching.

When there is no population stratification, our simulations showed a small loss of power in GSM due to unnecessary matching. Studies have shown that when the population is indeed homogeneous, random matching by pairs (1:1) can do almost as well as the unmatched test [Chase, 1968]. Additional power may be lost when the matching is not balanced, so that multiple controls are compared to a single case subject or multiple cases are compared to a single control (i.e., 1:*m* or *m*:1 when $m > 1$). However, when stratification is present, larger values of *m* are preferred to decrease the chance of matching errors. It is then a trade-off of efficiency and bias that we need to consider in practice. In our GSM method, the objective function (*T*) we choose for optimal matching favors smaller groups, minimizing loss of efficiency. Although the original

optimal matching [Rosenbaum, 1991] is unconstrained ($m = \infty$) so that all controls are allowed to be matched to a single case or all cases to a single control, Hansen [2004] showed that the matching with restriction on m can reduce the variance of estimated parameters with little increase in bias, and suggested a linear search for good values of m that are as close to 1 as possible. In our simulations, a large proportion of the matched sets are 1:1 matches even when the proportions of the two populations in cases and controls are not equal, and the average size of matched sets does not vary much for different values of the upper bound of m. For example, for simulated setting 3, the average matched set size is 2.44 and 2.47 when the upper limits of m are set as 2 and 5, respectively.

Although the full matching scheme is flexible, cases (or controls) from a population without a corresponding partner among the controls (or cases) will decrease power and may lead to spurious association if matching is forced. Further, 1:1 matching is more efficient than m:1 for $m>1$. Therefore, we still strongly encourage careful sample selection during the study design. Skol et al. [2005] showed that the self-reported ethnicity can be a good predictor for population structure, consistent with our results based on the NIMH case and control samples alone.

In summary, we propose a new framework to match case and control samples by their genetic similarity and adjust for the underlying population substructure. Our GSM method is specifically designed to use the full information provided by the large number of genotypes in genome-wide association studies or large-scale candidate gene studies. Our method can correctly control the false positives, while maintaining considerable power to detect the disease-marker association. Our individual-based matching scheme can reflect the continuous mixing of

ancestral populations. By comparing each case to one or more controls sharing the most genetic backgrounds, we hope our method may increase the chance to identify the genetic variants that influence disease risk. Our GSM software is available freely with C++ source code at http://www.sph.umich.edu/csg/liang/gsm/. The package allows the users to automatically calculate matching score matrices, conduct full matching with a range of parameter choices, and carry out association analyses. We expect our method will aid analyses of large-scale genome-wide association studies.

Appendix

In a minimum cost flow (MCF) problem, we define a directed graph consisting of nodes, $i \in \mathcal{N}$, and arcs connecting the nodes, $(i, j) \in \mathcal{A}$. For each arc (i, j), an integer a_{ij} denotes the cost and a positive integer c_{ij} the capacity. For each node i, an integer s_i denotes the exogenous supply. A solution of the MCF problem is a set of arc flows x_{ij} that minimizes:

$$\sum_{(i,j)\in\mathcal{A}} a_{ij} x_{ij}$$

subject to the constraints on capacity:

$$\sum_{\{j|(i,j)\in\mathcal{A}\}} x_{ij} - \sum_{\{j|(j,i)\in\mathcal{A}\}} x_{ji} = s_i, \quad \text{for all } i \in \mathcal{N}$$
$$0 \le x_{ij} \le c_{ij}, \quad \text{for all } (i, j) \in \mathcal{A}$$

It is easy to see the equivalence between the MCF and the optimal matching (Figure 2.5). The nodes in a directed graph correspond to the cases and controls, a_{ij} is the dissimilarity measure between i and j, and the capacity of the flow, c_{ij}, is 1 between case and control nodes, and 0 between two cases or two controls. The optimal solution of the MCF problem is equivalent to an optimal matching. The nodes connected by arcs with non-zero flow are assigned to the same matched set.

In full matching, the numbers of case-control pairs vary across matched sets, so the supply of nodes (s_i) cannot be predetermined. To deal with this complication, we include an "overflow" node to the graph to balance the flows from or to the case or control nodes. Parameters U and U_c control the maximum flows going to "overflow" from each node, which correspond to the maximum number of cases or controls allowed in each matched set, i.e., the upper limit of m in 1:m or m:1 match. For each case node, there are m connected control nodes and U-m arcs connecting it to "overflow"; for each control node, there are m connected case nodes and m arcs

connecting to “overflow”. The cost for arcs entering “overflow” is set as 0, so these extra arcs do not affect the total cost. Similarly, another node, “sink”, may also be added to control the total number of controls to be matched, and the cost for arcs entering “sink” is also 0 (Hansen et al., 2006).

The translation is demonstrated in Figure 2.5. The MCF problem is then solved by iteratively updating a dual cost vector and the flow vector $\boldsymbol{x}$ (Bertsekas et al., 1994; Frangioni et al., 2006).

Table 2.1 Values of IBS_k and $IBS_{k,i}$ for calculation of similarity scores.

Genotype Pair	IBS_k	$IBS_{k,A}$	$IBS_{k,a}$
aa aa	2	0	2
aa Aa	1	0	1
aa AA	0	0	0
Aa Aa	2	1	1
Aa AA	1	1	0
AA AA	2	2	0

Table 2.2 Example genotypes.

Cases		Controls	
Individual	Genotype	Individual	Genotype
1	aa, aa, AA	4	aa, aa, Aa
2	aa, aa, Aa	5	Aa, AA, aa
3	AA, AA, aa	6	AA, AA, aa

Table 2.3 Similarity (dissimilarity) scores for individuals in Table 2.2.

	Controls		
Cases	4	5	6
1	5/6 (1/36)	1/6 (25/36)	0 (1)
2	1 (0)	2/6 (16/36)	1/6 (25/36)
3	1/6 (25/36)	5/6 (1/36)	1 (0)

Table 2.4 Characteristics of simulated disease models: samples drawn from two subpopulations in 1:1 ratio.

Model	Population 1			Population 2		
	K_1	p_1	RR_1	K_2	p_2	RR_2
1	.10	.5	1.6	.10	.5	1.6
2	.10	.2	1.6	.10	.2	1.6
3	.07	.5	1.6	.13	.5	1.6
4	.07	.2	1.6	.13	.2	1.6
5	.07	.55	1.6	.13	.45	1.6
6	.07	.45	1.6	.13	.55	1.6

K_i: disease prevalence in population *i*.
p_i: predisposing variant allele frequency in population *i*.
RR_i: relative risk of the predisposing variant allele in population *i*.

Table 2.5 Average false positive rate and power of GSM, trend test (Chisq), and genomic control (GC) given 500 cases and 500 controls, 300,000 SNPs with MAF > .05, significance level = 10^{-6}.

Setting	Mismatch (%)	$\lambda^{\S}$	Average false positive rate ($\times 10^{-6}$)				Power*		
			GSM	Chisq	GC	EIGEN	GSM	GC	EIGEN
1	0	1.01	1.08	1.29	1.19	0.93	.80	.82	.82
2	0	1.01	1.10	1.16	1.10	0.97	.55	.56	.56
3	0.016	1.39	1.17	31.8	0.73	1.03	.75	.53	.76
4	0.015	1.38	1.15	30.7	0.47	1.07	.54	.28	.55
5	0.010	1.37	1.14	31.2	0.64	0.90	.72	.22	.72
6	0.010	1.38	1.09	33.0	0.66	0.87	.79	.78	.81

§. The global correction parameter in genomic control (GC), averaged over simulation replicates.
*. Power adjusted for the nominal false positive rates.

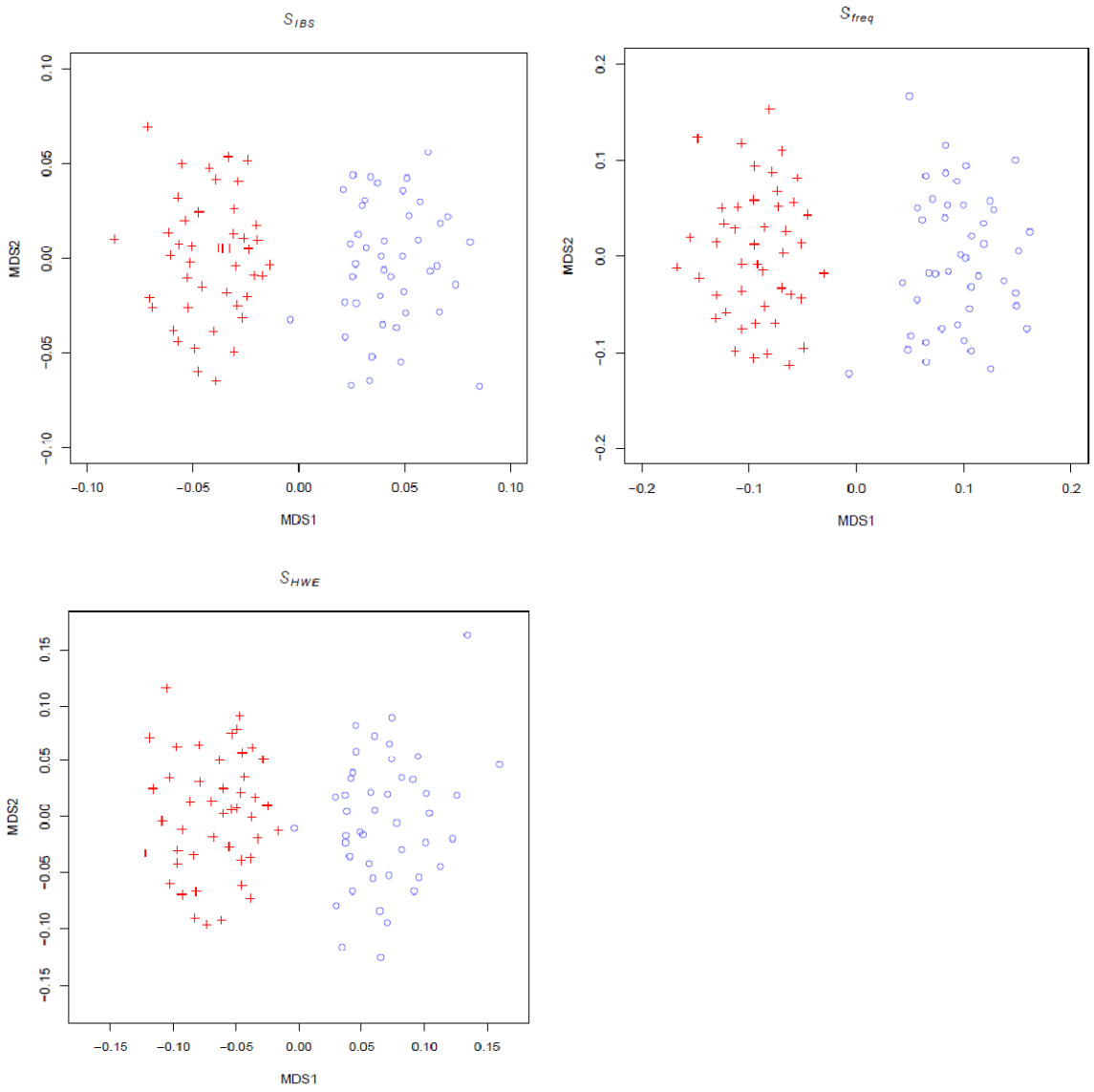


Figure 2.1 Multidimensional Scaling plots using dissimilarity scores as distance measure (calculated from 100,000 SNPs) for Han Chinese (HCB) and Japanese (JPT) HapMap samples. Red: HCB; blue: JPT.

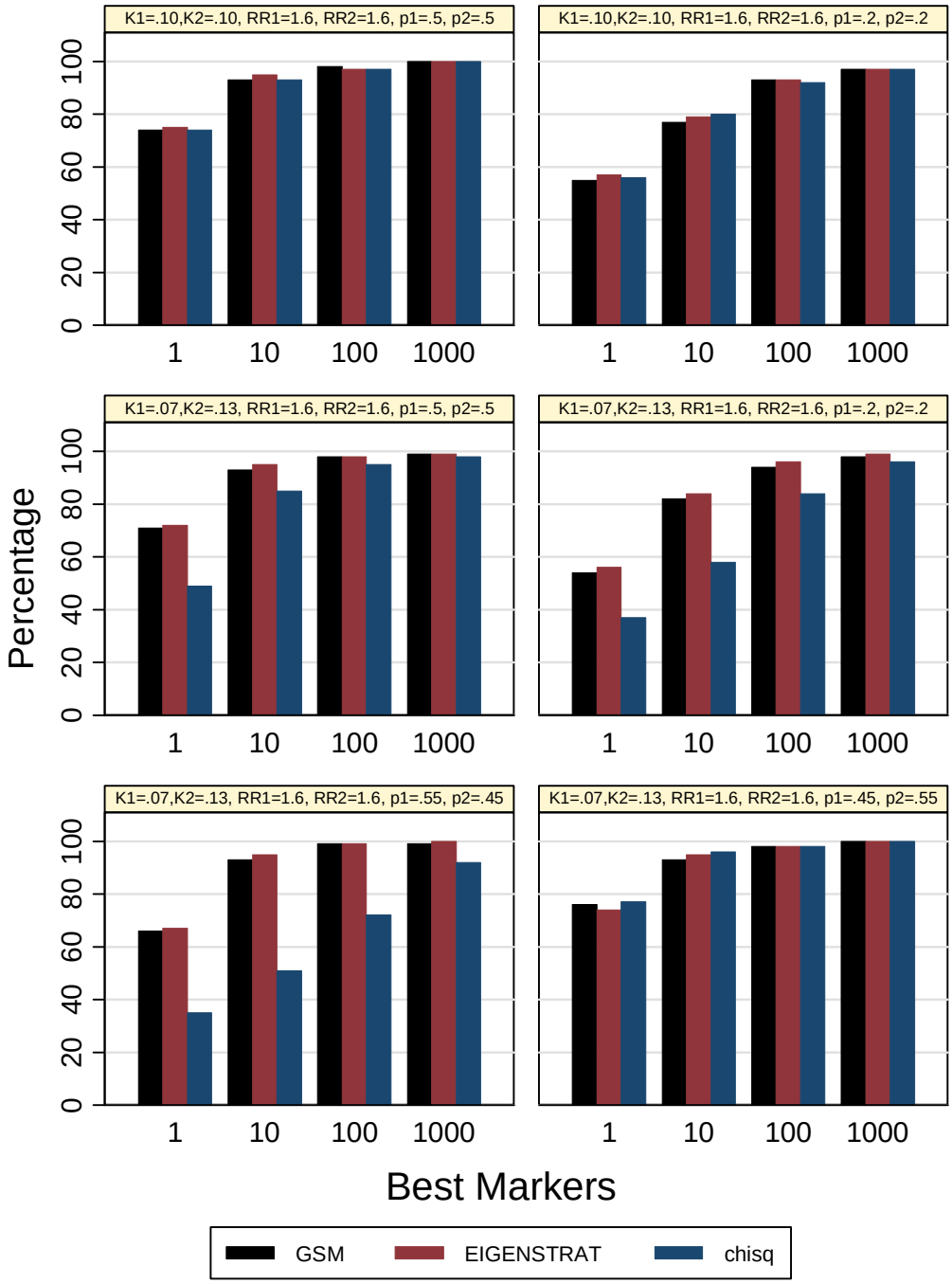


Figure 2.2 The frequencies of disease predisposing variant being identified among the best markers by similarity score matching method (GSM), EIGENSTRAT and trend test (Chisq).

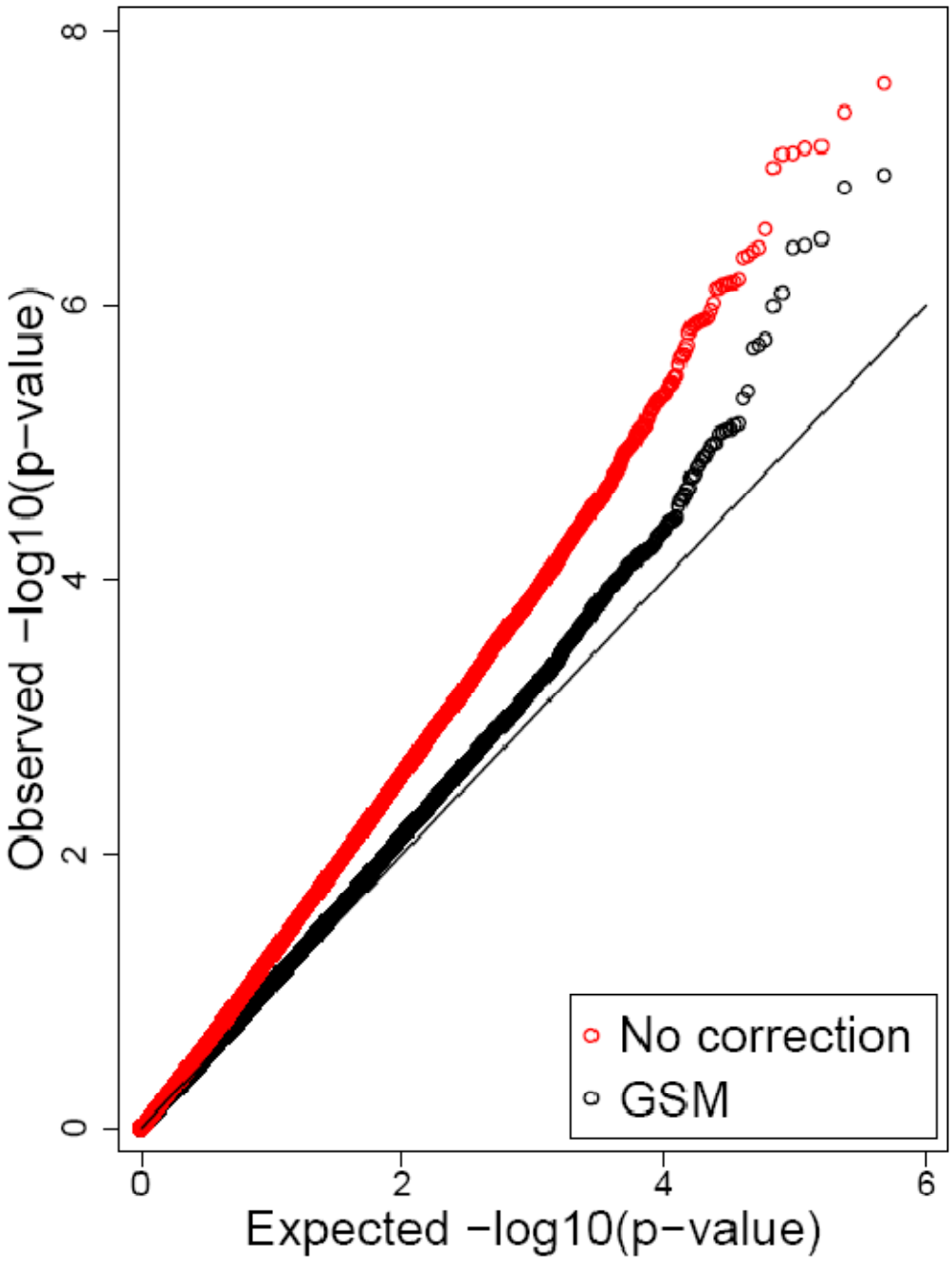


Figure 2.3 QQ-plot of the association results for the bipolar data (1 sibling per family) combined with Illumina controls. Red: the trend test; Black: GSM using 100,000 markers and S_{freq} as the similarity score.

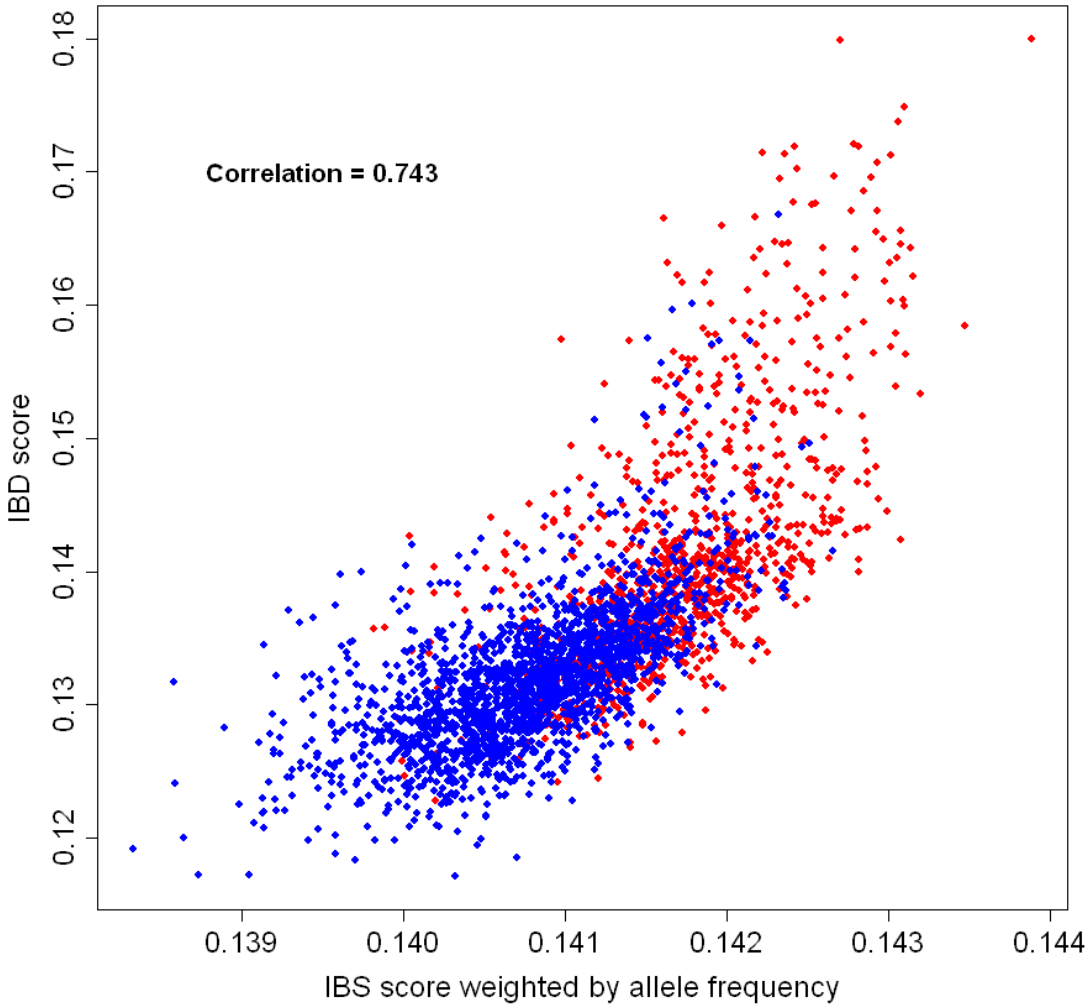


Figure 2.4 Similarity scores (calculated from 888,071 SNPs) between each pair of Han Chinese (HCB) and HCB-Japanese (JPT) in HapMap. Red: HCB-HCB pair; blue: HCB-JPT pair.

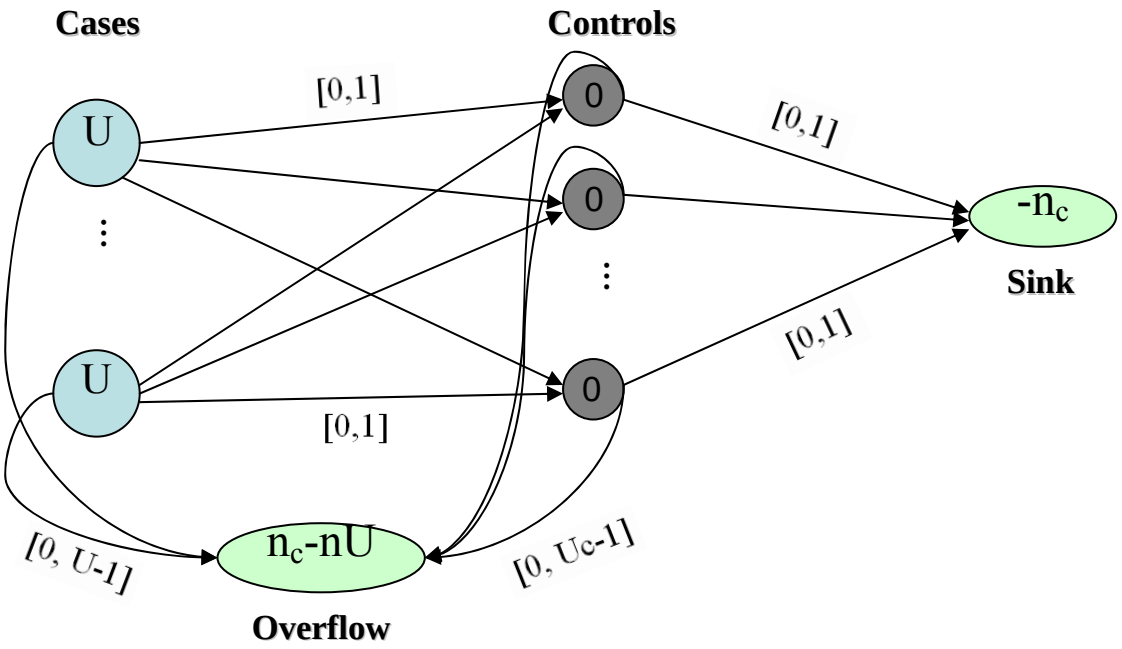


Figure 2.5 Solve optimal full matching problem as a minimum cost flow (MCF) problem. U denotes the maximal number of controls each case can match, U_c the maximal number of cases each control can match, n_c the number of controls to match, and n the total number of cases and controls.

CHAPTER 3

GENETIC SIMILARITY MATCHING FOR GENOME-WIDE ASSOCIATION STUDIES WITH RELATED INDIVIDUALS

In genome-wide association (GWA) studies, family samples are often used. We previously proposed a similarity score matching method that matches independent cases and controls based on their identity-by-state similarity scores using the large amount of genotype data from GWA studies. We now extend our method to the analysis of related cases and controls. We apply a new test statistic, based on our own work and that of Bourgain et al. [2003], to account for correlation both between the related samples and between matched case-control pairs. Through computer simulations, we demonstrate that in the context of family data, our method correctly controls type I error, and has improved power compared to genomic control in the presence of stratification. We illustrate our method with data from the Pritzker Consortium Bipolar GWA study.

3.1 Introduction

Genome-wide association (GWA) studies have identified >2000 common human genetic variants that are reproducibly associated with common diseases and traits [Altshuler et al., 2008; Manolio et al., 2008; www.genome.gov/gwastudies]. Unlike candidate gene studies which focus on pre-selected list of genes, GWA studies test for disease association in an unbiased way across

the entire genome. Current GWA studies typically genotype 500,000 to 1 million SNPs, covering the majority of the known common variants (~80-90%). For complex diseases, predisposing variants typically have modest effects on disease risk, requiring genotyping and analysis of large number of samples in GWA studies.

Most GWA studies focus on unrelated individuals – for example, cases and controls for a disease or a population cohort unselected for disease status. However, in some GWA studies, related individuals are genotyped. Ignoring sample relatedness can lead to spurious evidence for disease-marker association [Newman et al., 2001]. For continuous traits, variance component models can be used to test for association [Abecasis et al., 2000]. For dichotomous traits, (non-linear) variance components models may not always be appropriate due to impossibility in estimating within-family variance, when all members of the same family have concordant phenotype. For example, many studies sample affected sib-pairs and unrelated controls. As one alternative to variance component methods, Bourgain et al. [2003] proposed a corrected chi-square association test, in which they derived the correct variance term for allele frequency difference between cases and controls in presence of arbitrary family structures.

When the GWA samples are of heterogeneous ancestry, differences in allele frequencies and disease prevalence across ancestry groups may result in spurious association. Multiple approaches have been developed to correct for this population stratification problem. Devlin et al. [1999] developed the genomic-control (GC) method to rescale the test statistics according to a non-central χ^2 distribution. Price et al. [2006] and others developed methods based on principal components analysis (PCA), which adjust for the top principal components calculated from a

matrix of genotype data and account for population substructure. Alternatively, different study designs can be employed. For example, family-based association tests, such as the transmission/disequilibrium test (TDT) [Spielman et al., 1993], assessing the transmission of alleles between family members and so are immune to population substructure.

When family data are included in GWA studies, both these types of correlation must be considered when assessing disease-marker association: the relatedness of family members and the shared ethnic background within subpopulations. It is important to develop analysis tools that can deal with both population stratification and family samples. In Chapter 2, we proposed a genetic similarity score matching (GSM) method, to correct for population stratification in the analysis of case-control GWAS data using genotype-based matching. In this chapter, we describe how to extend this approach to handle related samples in GWA studies in presence of population substructure. Assuming all family members share the same ancestry, we perform the case-control matching using one member from each family and apply the matching result to other family members. Instead of conditional logistic regression, we propose a new chi-squared test based on the corrected chi-squared test following the Mantel-Haenszel procedure. Our simulations show that this new approach results in false positive rates at the desired nominal level while retaining high power to detect disease associated markers. We further demonstrate that when population stratification and related samples are present, association tests based on GSM-matched control data can have higher power than those that rely on either the corrected chi-squared test [Bourgain et al., 2003] or genomic-control method [Devlin et al., 1999].

3.2 Methods

For M genetic markers, we previously defined similarity scores between a pair of individuals based on allele sharing identity-by-state (IBS) measure:

$$S_{IBS} = \frac{1}{2M}\sum_{k=1}^{M} IBS_k \qquad (1)$$

or IBS measure weighted by allele frequency:

$$S_{freq} = -\frac{1}{2M}\sum_{k=1}^{M}\sum_{i\in\{A,a\}} IBS_{k,i}\cdot\log(q_{k,i}) \qquad (2)$$

where $q_{k,i}$ is the frequency of allele i at marker k, and $IBS_{k,i}$ is the number of copies of allele i at marker k shared by the pair of individuals.

Suppose our data are comprised of N families, with family j having k_j members, $j = 1, 2, \ldots, N$. Each family can contain arbitrary number of cases and controls. Assuming all individuals from the same family share similar genetic background, we first select N samples, one member from each family randomly, and calculate similarity scores between every case-control pair among the N samples.

We transform the similarity scores to dissimilarity scores using a monotone decreasing function, and then perform optimal full matching [Rosenbaum, 1991] to match cases and controls according to similarity for S_{IBS} or S_{freq} in the N selected samples. Optimal full matching results in G independent matched groups, each containing one case and one or more controls, or one control and one or more cases, that is, a 1:m or m:1 matching, by maximizing the overall similarity of matched cases and controls. If a selected individual $j = 1, 2, \ldots, N$, is assigned to group g, we then assign j 's relatives to the same group g. All samples in the data are therefore matched, and individuals from the same family are always assigned to the same group.

To test for disease-marker association, we first consider the situation without population substructure. Suppose we collect a total of n samples from N families, among which there are n_c cases . Let $\mathbf{1}$ denote a column vector of 1s with length n, $\mathbf{1}_c$ a column vector of length n with ith entry 1 if individual i is a case or 0 if i is a control, and Φ the kinship matrix. To test for association while taking into account relatedness of family members, Bourgain et al. [2003] proposed the corrected chi-squared ($W_{\chi^2_{corr}}$) statistic:

$$W_{\chi^2_{corr}} = \frac{\left(\hat{p}_{test} - \hat{p}_{null}\right)^2}{\mathrm{var}\left(\hat{p}_{test} - \hat{p}_{null}\right)} = \frac{\left(\hat{p}_{test} - \hat{p}_{null}\right)^2}{\frac{1}{2}\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_c^2}\mathbf{1}_c^T\Phi\mathbf{1}_c - 2\frac{1}{nn_c}\mathbf{1}^T\Phi\mathbf{1}_c + \frac{1}{n^2}\mathbf{1}^T\Phi\mathbf{1}\right)} \quad (3)$$

Here $\hat{p}_{test}$ is the estimated frequency of a specific allele in the case group, $\hat{p}_{null}$ is the estimated allele frequency in combined case-control data, and $\hat{p}$ is the estimated frequency in the population (often replaced by $\hat{p}_{null}$), all ignoring family relatedness.

When population substructure exists, we match cases and controls based on IBS information. We then need to consider the correlations both within the families and within the matched sets, in which samples share common ancestry, in the association test. We calculate a new test statistic following Mantel-Haenszel procedure:

$$\begin{aligned} T &= \frac{\left(\sum_{g=1}^{G}\left(\hat{p}_{test,g} - \hat{p}_{null,g}\right)\right)^2}{\sum_{g=1}^{G}\mathrm{var}\left(\hat{p}_{test,g} - \hat{p}_{null,g}\right)} \\ &= \frac{\left(\sum_{g=1}^{G}\left(\hat{p}_{test,g} - \hat{p}_{null,g}\right)\right)^2}{\frac{1}{2}\hat{p}\left(1-\hat{p}\right)\sum_{g=1}^{G}\left(\frac{1}{n_{c,g}^2}\mathbf{1}_{c,g}^T\Phi_g\mathbf{1}_{c,g} - 2\frac{1}{n_g n_{c,g}}\mathbf{1}_g^T\Phi_g\mathbf{1}_{c,g} + \frac{1}{n_g^2}\mathbf{1}_g^T\Phi_g\mathbf{1}_g\right)} \end{aligned} \quad (4)$$

where n_g is the number of samples in matched set g, $n_{c,g}$ is the number of cases in g, and $\hat{p}_{test,g}$, $\hat{p}_{null,g}$, $\mathbf{1}_g$, $\mathbf{1}_{c,g}$, Φ_g are the corresponding estimates as in (3), but calculated separately in each

set *g*. Given the matched sets are independent, this test statistic approximately follows a χ^2 distribution with 1 df, under the null hypothesis of no disease-marker association.

3.3 Simulations

We simulated case-control data influenced by genotypes at a single disease locus, under three additive disease models (Table 3.1). We assumed sampling from a population consisting of two subpopulations. We randomly sampled 300 affected sib-pairs and 300 unrelated controls from this mixed population. For each model, the relative risk (RR) of the predisposing variant allele is set to be the same in the two subpopulations (RR=1.6). For model 1, the disease prevalences K_1=K_2 and predisposing variant allele frequencies q_1=q_2; these models represent the scenario of no population stratification. For model 2, the first population has lower prevalence but higher predisposing variant allele frequency (K_1=.07, q_1=.55) than the second population (K_2=.13, q_2=.45). For model 3, the population with higher prevalence also has higher predisposing variant allele frequency (K_2=.13, q_2=.55) than the other population (K_1=.07, q_1=.45). For each model, we simulated 1000 datasets.

We simulated autosomal chromosomes using GENOME, a coalescent-based simulator [Liang et al., 2007]. We set the effective population size as 10,000 for each subpopulation, the recombination rate as 10^{-8} per base pair, and the mutation rate as 10^{-9} per base pair, assuming the infinite-site mutation model [Kimura, 1969]. We set the rate of migration between subpopulations to .0025 per individual per generation, which resulted in a distribution of allele frequency differences similar to that observed when comparing HapMap Han Chinese (HCB)

and Japanese (JPT) samples (www.hapmap.org). We randomly selected 300,000 SNPs with minor allele frequencies > .05 on 22 chromosomes whose lengths approximated the lengths of the human autosomes (NCBI build 33, www.ncbi.nlm.nih.gov/genome/seq/), and chose a disease liability SNP with the desired allele frequencies.

We first simulate the first affected siblings and the controls as independent cases and controls, by calculating the probability of each genotype at the disease locus given the affection status, based on the genotype penetrance, disease prevalence, and desired risk allele frequency. The parents of an affected sibling pair were then simulated by randomly picking one chromosome from the simulated sibling and pairing it with a randomly drawn chromosome from the general population. The second affected sibling was then generated conditional on the parental genotypes at the disease locus and the genotype penetrance.

To calculate the similarity scores, we chose 10,000 markers with the smallest one-sided HWE p-values, choosing no more than one marker from each 10-marker window. We set the maximum value of m in 1:m or m:1 matching to be 6. We compared the type I error and power of GSM, the $W_{\chi^2_{corr}}$ test, and the $W_{\chi^2_{corr}}$ test after genomic control for each simulation model.

We calculated estimated type I error rates as the proportion of simulated SNPs in which the association test p-value is less than the nominal value 10^{-6}, a significance threshold similar to that typically used in genome-wide scans. In this evaluation of type I error rates, we only considered SNPs that were effectively unlinked to the disease locus, defined operationally as 100 markers away from the disease locus. We calculated power as the proportion of simulated replicates where the empirical p-value is $< 10^{-6}$ at the disease locus using a threshold obtained by

inspection of test statistics at the null loci.

3.4 Bipolar data

We applied GSM to GWA data from the Pritzker Consortium bipolar study [Scott et al. 2009]. We selected 1217 independent bipolar I European American cases, which included 489 affected sib-pairs, and 779 independent European American controls from NIMH Human Genetics Initiative (www.nimhgenetics.org); controls were carefully matched to cases by self-reported ethnicity prior to genotyping. In addition, we downloaded genotype data on 3,182 independent European American controls from Illumina iControlIDB database (www.illumina.com/pages.ilmn?ID=231). All individuals were genotyped using the Illumina HumanHap550 BeadChip. 505,796 autosomal SNPs passed quality-control criteria in the Pritzker bipolar study: 1) HWE p-value > 10^{-5}; 2) genotype call rate > 95%; and 3) no more than 1 non-mendelian inheritance or duplicate genotype inconsistency among 15 father-mother-offspring trios and 30 duplicate samples. In addition, we excluded 1,632 SNPs due to allele frequency differences > .05 between the Illumina and Pritzker control samples. We applied GSM and the corrected chi-squared ($W_{\chi^2_{corr}}$) tests for association on the Pritzker samples alone and then on the combined Pritzker and Illumina samples. In GSM, we used the 100,000 markers that passed quality controls and had the smallest p-values from the one-sided HWE test to calculate the similarity scores. Given the relatively large control:case ratio of 5:1, we set the upper limit of the group sizes (m) to 30.

3.5 Results

3.5.1 False positive rate and power

Table 3.2 summarizes the average false positive rates and power for the three simulation models. In the absence of population stratification (model 1), GSM, $W_{\chi^2_{corr}}$, and $W_{\chi^2_{corr}}$ after genomic control all give false positive rates very close to the nominal value of 10^{-6}. The power of our GSM method is similar to that of the $W_{\chi^2_{corr}}$ test before and after genomic control. When population stratification is present (models 2 and 3), the type I error rate of the $W_{\chi^2_{corr}}$ test is ~30 times greater than the nominal value, while GSM and genomic control maintain the type I errors at or lower than the nominal level. For model 2, the $W_{\chi^2_{corr}}$ test has power comparable to that of GSM, after adjusting for inflated type I errors. For model 3, the risk allele frequency is lower in the subpopulation with higher disease prevalence, and population stratification may therefore mask the association. We observe that powers of the $W_{\chi^2_{corr}}$ test and genomic control drop ~30% compared to model 2, while GSM maintains the same level of power by correctly matching the samples with common ancestry.

3.5.2 Bipolar data

We first applied the $W_{\chi^2_{corr}}$ tests to the Pritzker bipolar case and control samples. The estimated genomic control variance inflation factor λ of the test statistics was 1.02, close to the expected value of 1 when there is no population stratification. However, when we added the Illumina control samples to the analysis, the estimated λ from the $W_{\chi^2_{corr}}$ tests became 1.44,

suggesting strong population stratification between the cases and controls. We then applied our GSM method on the combined samples. The GC λ dropped to 1.071 when we used S_{freq} (Figure 3.1) as our similarity measure and 1.080 when we used S_{IBS}, suggesting that GSM using either score provided useful correction for the stratification problem. To check the impact of setting the maximum number of controls per case or cases per control (m) to 30, we repeated our analysis by changing m to 10 and 50, resulting in estimated λ values of 1.20 and 1.065, respectively. This suggests that some controls may be matched to dissimilar cases when we only allow up to 10 controls per case, while increasing m from 30 to 50 resulted in little improvement on the accuracy of matching. We also repeated the matching using 50,000 markers instead of 100,000, and the estimated λ increased slightly to 1.079, consistent with the 100,000 markers providing slightly more information on the genetic background of samples than 50,000 markers.

3.6 Discussion

In Chapter 2, we proposed a new approach, genetic similarity matching (GSM), to correct for population stratification for large-scale association studies with independent cases and controls. In this chapter, we extended GSM for association studies that allow any type of family relationship. Simulation studies showed that the extended method can control the false positive rate in the presence of population substructure, while maintaining power to detect disease loci, when related samples are included in association studies. We applied our approach to the Pritzker bipolar GWA study in which an external control dataset was included, and showed that it can substantially improve control of the false positive rate.

We have compared the performance of the extended GSM to the commonly used genomic control method [Devlin et al., 1999]. As shown in our simulations (model 3), when stratification masks the association, genomic control showed less power than our approach to detect the disease-marker association. Another approach is based on the principal components analysis (PCA) as implemented in the EIGENSTRAT software [Price et al., 2006]. A typical strategy is to calculate the principal components for the independent samples first, and impute the scores for the other family members (e.g., Zhu et al., 2008), analogous to the approach we propose that starts the matching on unrelated samples. We applied this PCA-based approach to the bipolar data, and performed association tests using the robust variance ("sandwich") estimator [White, 1982] to account for the violation of independence in the family samples. We observed similar reduction of the false positives to our approach. Using the top 10 principal components, we obtained an estimated λ of 1.074.

When we match family members, we randomly select an individual to represent the entire family. When large families are collected, such random selection scheme may lose information. To reduce the variability caused by this selection scheme, one alternative is to calculate the similarity scores using all family samples. We calculate the pairwise similarity scores between families, which are defined as the average score between all pairs of members of the two families. We can then match families by treating each family as a single node in Figure 2.5. This modified approach should increase the accuracy of matching by paying extra cost on computing. In addition, for large families with roughly equal numbers of cases and controls, it could be more efficient to treat them as individuals groups without being matched with others.

It has been shown that 1:1 matching will generally lead to higher power than other matching schemes [Chase, 1968]. In addition to information loss, another disadvantage of picking one sample per family for matching is that an unbalanced number of cases and controls could occur when other family members are added to the matched sets. Some adjustments could be added to the matching scheme to reduce the power loss from unbalanced matching. For example, we could preferentially match families that contain more cases to families that contain more controls.

Although our method can be applied to families of arbitrary size and phenotypes, it is most useful for matching families with homogeneous phenotypes, i.e., case-only or control-only families. For families containing both cases and controls, especially when the numbers of cases and controls are approximately equivalent, each family can serve as a matched set without being matched to other samples, assuming members of these families sharing the same ancestry.

The test statistic we employed in our association test is based on a chi-squared test corrected for correlation between family members. Although it is computationally simple, it is difficult to incorporate other covariates, such as environmental factors. When the trait is continuous, or the family members are discordant in their affection status, we may consider variance component models that can take into account multi-level correlations. However, when all family members have the same affection status, it becomes difficult to estimate the variance component within families. An alternative is to treat the genotypes as a continuous response variable and the disease status along with other covariates as explanatory variables. We can then apply linear mixed model to handle the multi-level correlations.

In our approach, we assume that all members of each family share the same ancestry. This is often but not always the case. When the family members are of different ancestry, such as one parent with European ancestry while another parent with African ancestry, we may have to consider either the independent samples only, such as the parents in nuclear families, or alternative approaches that can deal with the complicated correlation between the samples that are related and between the samples that are from the same ancestral populations.

In summary, we extend our GSM approach to adjust for the underlying population substructure when related samples are collected in GWA studies. It inherits the merits of our original approach, with a chi-squared test statistic for the association that is easy to compute. We expect our method will aid analyses of large-scale GWA studies for arbitrary types of samples.

Table 3.1 Characteristics of simulated disease models: samples drawn from two subpopulations in 1:1 ratio.

Model	Population 1			Population 2		
	K_1	p_1	RR_1	K_2	p_2	RR_2
1	.10	.50	1.6	.10	.50	1.6
2	.07	.45	1.6	.13	.55	1.6
3	.07	.55	1.6	.13	.45	1.6

K_i: disease prevalence in population *i*.
p_i: predisposing variant allele frequency in population *i*.
RR_i: relative risk of the predisposing variant allele in population *i*.

Table 3.2 Average false positive rate and power of similarity score matching (GSM), $W_{\chi^2_{corr}}$ test, and $W_{\chi^2_{corr}}$ test after genomic control (GC) given 300 affected sib-pairs and 300 controls, 300,000 SNPs with MAF > .05, significance level = 10^{-6}.

Setting	$\lambda^{\S}$	Average false positive rate ($\times 10^{-6}$)			Power*	
		GSM	$W_{\chi^2_{corr}}$	GC	GSM	GC
1	1.01	1.06	1.18	1.10	.48	.49
2	1.34	1.12	26.7	0.70	.48	.48
3	1.36	1.10	28.3	0.65	.49	.23

§. The global correction parameter in genomic control (GC), averaged over simulation replicates.
*. Power adjusted for the nominal false positive rates.

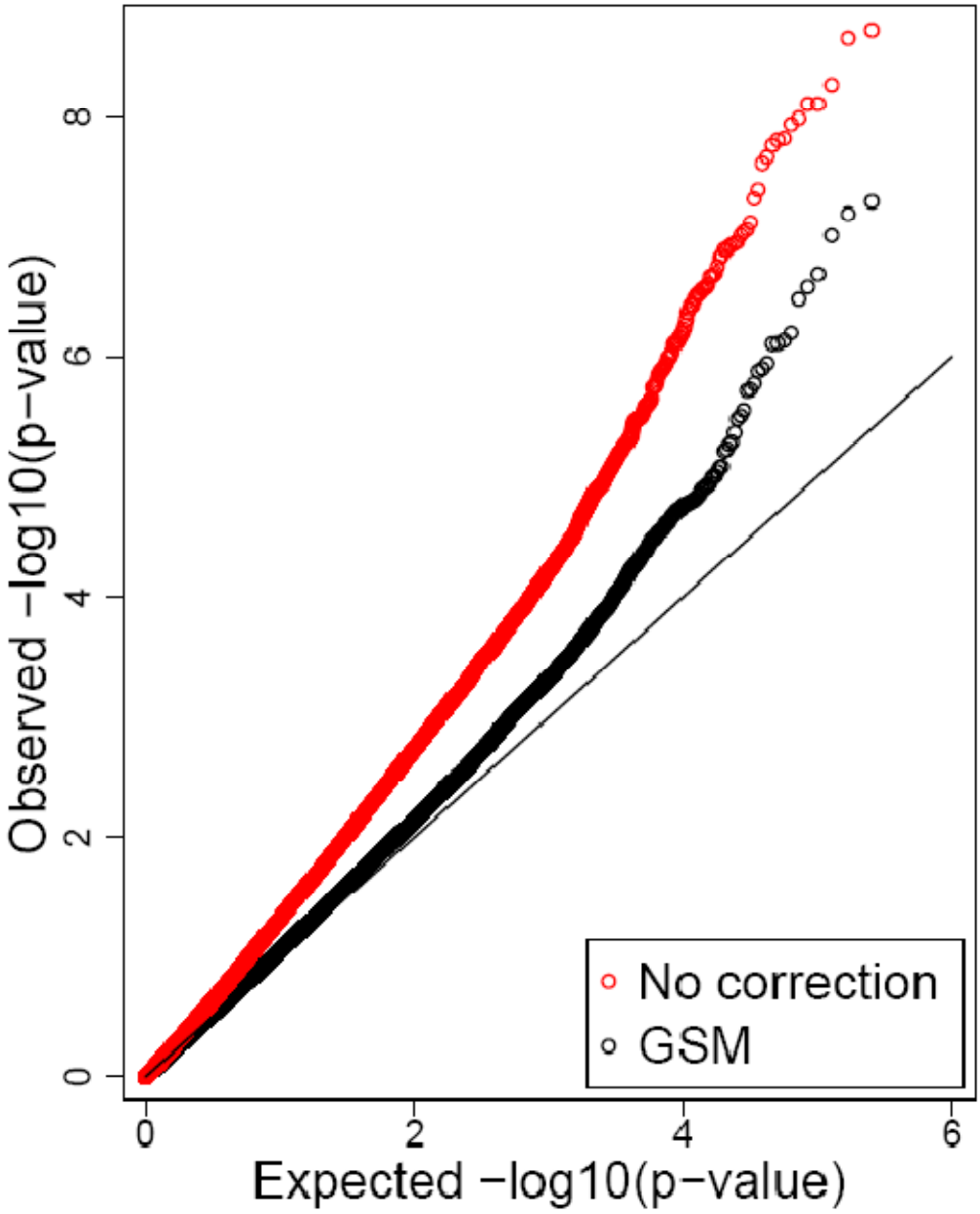


Figure 3.1 QQ-plot of the association results for the bipolar data combined with Illumina controls. Red: the trend test; Black: GSM using 100,000 markers and S_{freq} as the similarity score.

CHAPTER 4

ESTIMATION OF THE CONTRIBUTION OF RARE CAUSAL VARIANTS TO COMPLEX DISEASES

Rapid advances in next-generation sequencing technologies are providing ever more complete surveys of genomic regions of interest and even whole genomes, and these advances are facilitating genetic association studies of an increasingly wide array of common and rare variants. As we plan re-sequencing studies, it is interesting to consider constraints on the range of plausible genetic models based on results of previous genetic studies of complex traits. In particular, results of genome-wide association (GWA) studies and previous linkage studies provide information on the range of causal variant risk allele frequencies (RAF) and effect sizes (genotype relative risk, or GRR) that might underlie common GWA signals and might, more broadly, exist in a given genomic region. In this study, we use power calculations to evaluate the range of models that would result in little evidence for linkage in a linkage scan (a typical finding) and different degrees of evidence for association at genotyped markers in an association study, assuming a multiplicative model of disease risk. We find that prior negative linkage or association results can help to restrict the plausible range of GRR, and significant evidence for association can reduce the plausible ranges of both GRR and RAF given little or no evidence for linkage in the corresponding region. We apply our approach to large-scale linkage and GWA meta-analyses for type 2 diabetes, and suggest ranges of plausible models for 15 loci identified as genome-wide significant.

4.1 Introduction

GWA studies allow investigators to test for disease or trait association with common single nucleotide polymorphisms (SNPs) throughout the human genome. For common variants (say minor allele frequency (MAF) >.05), today's commercial GWA platforms typically cover 80-90% of known common genetic variants when combined with genotype imputation [Li and Abecasis, 2006; Browning and Browning, 2007; Marchini et al., 2007; Purcell et al., 2007]. In recent years, GWA studies have been conducted for many complex diseases or traits [Altshuler et al., 2008; Manolio et al., 2008, Hindroff et al., 2009, and http://www.genome.gov/gwastudies/]. For many diseases or traits, such as type 2 diabetes, large numbers of significant disease-SNP associations have been reported [Scott et al., 2007; Saxena et al., 2007; Zeggini et al., 2007, Zeggini et al., 2008] while for some others, such as bipolar disorder, fewer significant associations have been identified [Sklar et al., 2008; Scott et al., 2009]. For most diseases and traits, the combined effects of associated SNPs explain only a small proportion of the genetic variation [Manolio et al., 2009]. Explaining this "missing heritability" remains an important challenge.

Results to date suggest that most common complex traits associated variants have modest effect. Other types of genetic variants including rare (MAF < .005) and less common (.005 < MAF <.05) polymorphisms have not yet been studied extensively and may be associated with larger effects. With the recent advances in sequencing technology, it has become feasible to identify and genotype these variants in individual genomes. While many re-sequencing studies are being undertaken, our understanding of the contribution of rare variants to complex diseases remains limited.

Most linkage analyses of complex diseases have reported limited evidence for linkage, and even in studies with positive linkage results, most of the genome provides no evidence for linkage. These negative linkage results may limit the range of plausible effect sizes for disease predisposing variants. Rare or less common variants may in principle be responsible for some of the significant evidence for association observed in GWA studies through linkage disequilibrium (LD). Evidence for association or lack of such evidence in a region of interest could also reduce the set of plausible models, particularly if we assume high LD (e.g., D' = 1) between causal and genotyped variants.

In this chapter, we seek to identify the range of genetic models for rare or less common variants, parameterized in terms of genotype relative risk (GRR) and risk allele frequency (RAF), and consider the following scenarios:

1) Little or no evidence for linkage;
2) Significant evidence for association;
3) Little or no evidence for association;
4) Significant association, but little or no evidence for linkage;
5) Little or no evidence for either linkage or association.

To do so, we calculate the power to detect linkage and association analytically for given genetic models and sample sizes. We then summarize the ranges of models that appear consistent with results from linkage and/or association studies, as listed above in 1) - 5).

Our results show that for scenarios 1, 3, and 5), all RAFs are in principle plausible, but the effect sizes (GRR) of causal variants are bounded. For scenario 2 and 4), causal variants with

small RAF can be identified as implausible, assuming high LD (D'=1) between the causal and genotyped variants. For 15 T2D loci identified through large scale GWA and meta-analysis [Scott et al., 2007; Zeggini et al., 2008], taking advantage of linkage results from a large linkage study [Guan et al., 2008] allows us to identify association signals that are unlikely to be explained by rare or less common variants and suggest ranges of allele frequencies and effect sizes for all these signals.

4.2 Methods

To understand the genetic architecture underlying a complex disease, we seek to obtain a set of models parameterized by the genotype relative risk (GRR_C) and risk allele frequency (RAF_C) of a causal variant (C) which are plausible given prior results from linkage and association, usually GWA, studies. Specifically, we consider the situation of a prior affected sib pair (ASP) linkage study, and/or a case-control association study. We assume a multiplicative model for disease penetrance [Risch, 1990a] and a dense set of linkage markers so that IBD can be observed for the ASPs, and initially assume a standardized LD coefficient D'=1 between the causal SNP C and a nearby genotyped SNP M, and disease prevalence 10%. We restrict GRR_C to be smaller than 10, because for a disease with 10% prevalence, $GRR_C > 10$ suggests that genotype with 1 or 2 rare or less common risk alleles are fully penetrant, which is unlikely for complex diseases.

4.2.1 Power to detect linkage using ASP test

Given a sufficiently dense panel of genotyped markers, we assume that identity-by-state (IBD) sharing between siblings is known throughout genome. Let N_i be the number of ASPs sharing i alleles IBD. Under the null hypothesis of no linkage, the number of alleles shared IBD at a specific locus between N affected sibs is distributed as binomial on $2N$ trials with probability of success 0.5. We calculate the maximum LOD score (MLS) [Risch, 1990b] as:

$$MLS = \begin{cases} (N_1 + 2N_2)\log_{10}\left(\frac{N_1 + 2N_2}{N}\right) + (N_1 + 2N_0)\log_{10}\left(\frac{N_1 + 2N_0}{N}\right), & \text{if } \frac{1}{2}N_1 + N_2 > N \\ 0, & \text{if } \frac{1}{2}N_1 + N_2 \le N \end{cases}$$

Assuming a multiplicative model, the power to detect linkage using ASPs depends on the locus-specific relative risks, λ_{1R}, for relatives of type R (Risch, 1990b). Specifically, under the alternative hypothesis, $N_1 + 2N_2$ is distributed as binomial($2n$, $.5z_1+z_2$), where z_i = P(IBD=i | ASP), and

$$z_1 = \frac{1}{2} \cdot \frac{\lambda_{1O}}{\lambda_{1S}}$$

$$z_2 = \frac{1}{4} \cdot \frac{\lambda_{1M}}{\lambda_{1S}}$$

where O = parent-offspring pair, M = monozygote (MZ) twins, and S = sibling pair. The three types of relative risk, λ_{1O}, λ_{1M} and λ_{1S}, are all functions of GRR_C and RAF_C.

We calculate power to detect linkage for studies of 500, 1000, and 5000 ASPs. For the causal variant C, we assume RAF_C in range of .001 to .05, but the calculation can be generalized to more common frequencies. Given RAF_C, we determine the value of GRR_C that leads to 95% power to detect linkage with specific *MLS* cut-off using the false position method [Press et al., 1992], a computer algorithm for root finding. We report results for *MLS* cut-off values of .05 and

1, representing essentially no evidence for linkage or modest evidence, respectively.

4.2.2 Power to detect association at genotyped variant that is in high LD with causal variant

We assume a GWA study of common variants with equal numbers (*n*) of cases and controls. Let C be a causal variant, which is in high LD (D' = 1) with a genotyped variant M in the GWA study. Denote the risk allele frequency at M as RAF_M and genotypes at C and M as g_C and g_M, respectively. Given RAF_C, RAF_M, and D', we calculate the conditional probability of genotypes, $P(g_C \mid g_M)$. For specific genetic model and disease prevalence, we can then compute the penetrance of g_M:

$$P(Y \mid g_M) = \sum_{g_C} P(Y \mid g_M, g_C) P(g_C \mid g_M) = \sum_{g_C} P(Y \mid g_C) P(g_C \mid g_M)$$

and genetic relative risk at M, GRR_M. Note that we bound the penetrance, $P(Y \mid g_C)$, at 1 in our calculation, when the assumed GRR_C is large.

We test for disease association at M using the Cochran-Armitage trend test. The power of the trend test can be calculated at a specific significance level α, by estimating the variance of test statistic under the alternative hypothesis [Freidlin et al., 2002].

We consider GWA studies with *n* = 1000, 10000, and 50000 cases and the same number of controls. For causal variant C and genotyped variant M, we assume RAF_C and RAF_M in range of .001 to .05, and .05 to .95, respectively. Given RAF_C and RAF_M, we estimate the value of GRR_C that leads to 5% or 95% power to detect association. We focus on significance level

5×10^{-8}, representing genome-wide significant evidence for association in a GWA studies, or 10^{-4} representing no strong significance evidence for association in a GWA study.

4.2.3 D'=1 assumption and 1000 Genomes Project data

The 1000 Genomes Project (http://www.1000genomes.org/) is an international research effort to catalog human genetic variation. As of March, 2010, 56 European samples have been sequenced at ~4X coverage as part of the first pilot project. To assess the plausibility of our assumption of the existence of a marker M in strong LD (D'=1) with the causal variant C, we evaluate the values of D' between less common variants identified in the 1000 Genomes Project and common variants that are likely genotyped or imputed in GWA studies. Specifically, we evaluate D' between variants of frequency .01 to .03 (2 or 3 copies of the rarer allele in the 56 1000-Genomes samples) and common variants (MAF > 5%) genotyped in Phase 2 of the International HapMap Project (http://www.hapmap.org/). We limit our attention to common variants within a 1500-SNP window (approximately ~500 kb) of each less common variant. To reduce the computational load, we further limit our analysis to chromosome 1.

4.2.4 Application to type 2 diabetes (T2D)

To illustrate how existing linkage and GWA studies could provide information on rare or less common variants underlying complex diseases, we use type 2 diabetes as an example. We carried out a joint linkage analysis of 23 individual studies as part of the International Type 2 Diabetes Linkage Analysis Consortium [Guan et al., 2008, Nancy Cox, personal communication], with a

combined sample of 9,455 affected individuals and the equivalent of ~6500 ASPs, using the approximation that *m* affected siblings correspond to *m*-1 ASPs [Hodge 1984]. No genome-wide significant linkage result was found, and the largest MLS is approximately 1.7. Two meta-analyses of T2D GWA results [Scott et al., 2007; Zeggini et al., 2008] identified 15 T2D-associated loci (Table 4.1), using total samples of up to 14586 cases and 17968 controls and 28645 cases and 39397 controls with multiple-stage study designs, respectively, approximately equivalent in information to case/control samples of $n \sim 16000$ and ~32000, respectively. To estimate genetic models consistent with these results, we calculate power to detect linkage using 6500 ASPs and to detect association using *n* cases and *n* controls, where *n* corresponds to the sample size of GWA at each locus. In addition, we use the observed association p-values from the corresponding GWA as significance level α, and observed *MLS* scores from the linkage study as *MLS* cut-off for power calculation at each locus.

4.3 Results

Here we address the range of plausible model parameters (RAF_C and GRR_C) for rare or less common variants (MAF < .05), assuming a multiplicative genetic model for a genomic region and given results from prior linkage and/or association studies. To do so, we compute the power to detect linkage and/or association as a function of genetic model, for the five scenarios listed in "Introduction". To initially explore the plausible models, we assume a single causal variant C in a region of interest which is in D'=1 with the genotyped variant M.

4.3.1 Range of plausible models given no evidence for linkage

Linkage studies generally reveal little or no evidence for linkage for most of the genome, and in the case of complex diseases, often all of the genome. We explore the range of genetic model parameters consistent with this observation. Figure 4.1 displays values for GRR_C that result in 95% power to detect at least modest evidence for linkage (MLS > 1) given analysis of N = 500 to 5000 ASPs as a function of RAF_C. Assuming a causal variant exist in such region, models (RAF_C, GRR_C) above the lines have high probability (p > 95%) of being detected by linkage at MLS > 1 given N ASPs, and therefore are unlikely in regions in which MLS < 1. For example, given N = 5000 ASPs, causal variants with RAF_C = .01 or .05 have ≥95% power to achieve MLS>1 if GRR_C > 3.9, or 2.4, respectively, suggesting these models are unlikely given no evidence for linkage in the region. As expected, larger linkage studies with little or no evidence for linkage (MLS < 1) result in a more extensive set of implausible models than do smaller studies.

Although in Figure 4.1, all models under the lines are mathematically consistent with little or no evidence for linkage, some of these models may still be questionable given our knowledge of complex diseases. For example, when RAF_C = .001, the power calculation suggests that a single variant with GRR_C ~ 9.9 may not be identified in a linkage study using 5000 ASPs. However, for a complex disease with population prevalence of 10%, such high GRR_C implies that any genotype carrying at least one risk allele has penetrance close to 1, unlikely for a complex disease.

4.3.2 Range of plausible models given significant evidence for GWA association

Significant associations have been reported for many common diseases/traits (http://www.genome.gov/gwastudies/). We explore the range of models that could explain these findings by single rare or less common causal variants. Figure 4.2 shows values of GRR_C that lead to 5% power to detect association ($p < 5\times10^{-8}$) at M with RAF_M = .05-.95, assuming a study of *n* cases and *n* controls (*n* = 1000 to 50000). The models below the lines have low probability (< 5%) to achieve significant association ($p < 5\times10^{-8}$) for a genotyped marker M in high LD (D'=1) with the causal variant C. For a given RAF_C, significant associations are more likely to be observed at genotyped variants M with low RAF_M. For example, when *n* = 10000, RAF_C = .01 and GRR_C = 3.9, a significant association at M that can be explained by C would require RAF_M < .4. Note that due to the "winner's curse", GRR_M may be overestimated (particularly for markers where the association test p-value is close to significance thresholds) that results in an excessively narrow estimate for bound of RAF_M (See "Discussion").

Not surprisingly, Figure 4.2 also shows that significant evidence for association observed in a smaller scale GWA study leads to a more limited set of plausible models than does the same association evidence in a larger study. In addition, given fixed GRR_C and RAF_M, variants with RAF_C ~ .01-.05 are more likely than those with RAF_C ~ .001-.005, to be detected via GWA.

4.3.3 Range of plausible models given no evidence for GWA association

Over most of the genome, little or no evidence for association is observed in a GWA study (say $p > 10^{-4}$). Figure 4.3 shows values of GRR_C that lead to 95% power to detect association (at $p < 10^{-4}$) at genotyped variant M, for given RAF_C and sample size. Compared to Figure 4.2, we

have changed the power cut-off from .05 to .95 and the significance threshold from 5 x 10^{-8} to 10^{-4}, the combination of which yields similar curves. Assuming a single causal variant exists in such region with no SNP showing association p-value < 10^{-4}, and the causal variant is in D'=1 with the genotyped GWA variants, the models above the lines are unlikely for a rare or less common causal variant C given little or no evidence for association in GWA. For example, when n = 10000, RAF_C = .01 and RAF_M = .2, causal variants with GRR_C > 3.6 are unlikely to be present in regions with little or no evidence ($p > 10^{-4}$) for association.

4.3.4 Range of plausible models given evidence for association but not for linkage

For complex diseases for which linkage and association scans have been carried out, we expect no or limited evidence for linkage (MLS < 1) in most regions of the genome. In some such regions, there will be genome-wide significant association results ($p < 5\times10^{-8}$). Figure 4.4 shows results for this scenario, specifically, values of GRR_C that lead to 5% power to detect association ($p < 5\times10^{-8}$) at genotyped variant M or 95% power to detect linkage (MLS < 1), for given RAF_C and sample size. The models above the line for association and below the line for linkage, for specific sample size, are consistent with strong evidence for evidence but little evidence for linkage in the corresponding region. For example, a significant association ($p \sim 5\times10^{-8}$) observed at M with RAF_M = .5 in a GWA study (n = 10000) would require RAF_C > .015, given no linkage signal of MLS > 1 using 5000 ASPs.

For a causal variant C with specific RAF_C, it is interesting to find the maximum RAF_M, for which significant association at M could be explained by C in D'=1, conditional on MLS < 1.

Table 4.2 gives the upper bounds of RAF_M for GWA studies with n = 1000 to 50000, given negative linkage results using N = 500 to 5000 ASPs. The lower the RAF_C, the lower the upper bound on RAF_M that is consistent with the lack of linkage and presence of GWA association results. For example, in regions where MLS < 1 in a linkage study using 5000 ASPs, significant associations at M with $RAF_M \leq .25$ using 50000 cases and 50000 controls could be consistent with underlying C with $RAF_C = .001$. Again, RAF_M in Table 4.2 could be overestimated due to "winner's curse" (see "Discussion").

4.3.5 Range of plausible models given no evidence for association or linkage

Most of the genome typically shows little or no evidence for association ($p > 10^{-4}$) or linkage (MLS < 1). In these regions, we again combine information from linkage and GWA studies to infer the set of plausible models. Figure 4.5 shows GRR_C that lead to 95% power to detect association ($p < 10^{-4}$) at genotyped variant M or 95% power to detect linkage (MLS < 1), for given RAF_C and sample size. Assuming a single causal variant C is present in the region of interest, the models below both the linkage and association lines are consistent with no evidence for linkage and association. For example, lack of evidence for linkage (MLS < 1, N = 5000) and no significant association observed at any M with $RAF_M < .3$ ($p > 10^{-4}$, n = 10000) suggest that $GRR_C < 2.9$ for C with $RAF_C = .01$, assuming C is present in the region and in high LD (D'=1) with M.

4.3.6 D' for common and less common variants in the 1000 Genomes Data

In our analyses, we have assumed D'=1 between the causal variant C and the GWA genotyped variant M. To explore the validity of this assumption, we used sequencing data on 54 European subjects currently available from the 1000 Genomes Project (http://www.1000genomes.org/). Specifically, we calculated D' between less common SNPs (.01 < MAF < .03) and common SNPs (MAF > .05) that are also present in HapMap Phase 2 (HapMap2) data (http://hapmap.ncbi.nlm.nih.gov/) (Figure 6) to mirror the imputation and testing of HapMap SNPs in published GWA studies. On chromosome 1, there are 80,408 less common SNPs and 173,964 common HapMap2 SNPs. Among the less common SNPs, 79,676 (99.1%) are within ~500kb of at least one common HapMap2 SNP, and the majority of these SNP pairs (~72%) have standardized LD coefficient D' 1. This suggests that less common SNPs are very likely to be in high LD with common HapMap2 SNPs.

There could be multiple common HapMap2 SNPs in high LD with a single less common SNP. We examine the common SNPs that would be more likely to be identified (i.e., higher D' and smaller RAF_M) in GWA, if the less common SNPs were causal. Specifically, for each less common SNP, we identify the best pairing SNP that has the highest value of D' and smallest pairing allele frequency (PAF). Figure 6b shows that among the 79,676 less common SNPs on chromosome 1, ~53%, 31%, and 10% SNPs have a best pairing SNP with PAF in .05-.1, .1-.2, and .2-.3, respectively. This suggests that, if we genotype or impute all common HapMap2 SNPs with high accuracy in a GWA study, most of the less common SNPs (~94%) will be in D'=1 with at least one common SNP with PAF < .3 (but see "Discussion").

4.3.7 Example: type 2 diabetes (T2D)

Many linkage and association studies have been carried out for type 2 diabetes. A linkage meta-analysis using ~6500 ASPs [Guan et al., 2008; Nancy Cox, personal communication] did not find genome-wide significant evidence for linkage (no MLS > 3), while two GWA-based meta-analysis [Scott et al., 2007, Zeggini et al., 2008] using total samples approximately equivalent to 16000/16000 and 32000/32,000 cases/controls reported SNPs in 15 genomic regions significant at $p < 5\times10^{-8}$ (Table 4.1). For these T2D-associated loci, we use the corresponding GWA sample size at each locus to find the range of plausible RAF_C. Signals with less frequent risk alleles (*CDC123* and *TSPAN8*) could plausibly be explained by a wide range RAF_C that may require sequencing studies to pinpoint the causal variants. For other associated SNPs, the associations are unlikely to be explained by a single causal variant with $RAF_C < .01$, and for two (*CDKN2A/B* and *THADA*), RAF_C would need to be larger (> .05). In these regions, a GWA study with dense marker set (say, a good coverage for SNPs with MAF > .01) might well be sufficient to insure that the causal variant would be genotyped.

4.4 Discussion

In this paper, we have sought to determine the range of genetic models consistent with existing linkage and/or GWAS results, focusing on less common and rare variants. Specifically, we examine the range of plausible models assuming weak or no evidence for linkage, in regions with or without strong evidence of association from GWA studies. Our results show that, under the assumption of a single causal variant in the region of interest, little or no evidence for either linkage or association restricts the likely magnitude of GRR_C, although all RAF_C in the range we

consider (.001 - .05) are possible. Combining results of linkage and GWA can substantially reduce the set of plausible models. In loci with significant evidence for association, negative results in a linkage study typically makes unlikely causal variants with small risk allele frequencies (say $RAF_C < .01$), especially when the association is detected with a marker M with high RAF_M.

We applied our approach to explore plausible models for type 2 diabetes (T2D). Given the linkage results from a combined study with equivalent sample size of ~6500 ASPs, we estimated the range of plausible models for T2D-associated loci reported in two GWA meta-analyses. Our results suggest that several association signals can only be explained by relatively common causal variants ($RAF_C > .01$), while other signals are consistent with a wide range of RAF_C and GRR_C.

Our analyses show that for a fixed D' value with the same rare or less common causal variant C, a genotyped variant M with low RAF_M will have higher power to be identified in association test than variant with high RAF_M. However, we do not observe excess number of T2D-associated SNPs with $.1 < RAF_M < .3$ compared to $.3 < RAF_M < .9$, a range in which GWA study well suited to detect. This suggests that common causal variants are likely to underlie at least some of the T2D signals.

In our analysis, we assume a single causal variant that is in high LD with genotyped variant M. When multiple causal variants are present within a short region, multiple risk alleles could lie on the same haplotype with a specific ("risk") allele of M and therefore increase the effective

value of RAF_C. However, Dickson et al. [2010] showed, through a computer simulation of genealogical trees, that when a large number of causal variants (say ≥ 5) exist, the risk alleles are likely to be distributed on different haplotypes with "risk" allele of M, which leads to decreased power to detect association at M, compared to that with a smaller number of variants. In contrast to tests of association, linkage analysis tests the co-transmission of risk allele and genotyped marker, and we would see increased effective value of RAF_C when multiple causal variants exist in the region of interest. If each individual variant considered has the same frequency, when RAF_C is increased due to the presence of multiple causal variants, the plausible range of genetic model will become wider compared to our results based on single-variant assumption.

Another assumption we make when assessing the results of association is that the causal variant C and the genotyped variant M have a standardized LD coefficient D' = 1. We examined sequence data on 54 individuals currently sequenced as part of the 1000 Genomes Project. We find that among 80,408 less common SNPs (.01 < MAF < .03) on chromosome 1,>94% have at least one pairing common SNP (MAF > .05) in D' = 1 with pairing allele frequency (PAF) < .3, so the assumption of D' = 1 is likely true for a large proportion of the identified less common SNPs. However, the current 1000 Genomes Project has a relatively small sample size, which will have resulted in an overestimate of D', for example in Figure 4.6(a). In Figure 4.6(b), we estimate D' between the less common SNP (C_r) and best pairing SNP (M_p) that has the lowest PAF. When most of PAFs are small, as suggested in this figure, the overestimation is likely to be less severe. For example, when PAF = .1, the chance of two minor alleles of C_r sharing a haplotype with the same pairing allele of M_p is only 10%×10% = 1%, if the two minor alleles of C_r are randomly assigned. But in Figure 4.6(b), we observed ~53% best pairing alleles with PAF

< .1, much greater than 1% due to randomness. As more sequencing data becomes available, we will obtain more accurate estimate for the distribution of D' between less common and common variants. For a given allele frequency, if the best pairing SNP has D' < 1, significant associations in GWA studies would require higher RAF_C and/or GRR_C to be consistent models. In contrast, negative association results would correspond to a wider range of models than our current conclusion. For example, 1000 Genomes Project data show that ~0.9% of less common variants (.01 < MAF < .03) on chromosome 1 have D' = 0 with all common HapMap2 SNP in a ~500 kb window. We will obtain no information for such variants given current GWA results.

When we explore the plausible range of models based on significant GWA results, the estimated RAF_C and/or GRR_C are conditional on the association tests passing a given p-value cut-off. Particularly if analyses result in p-values just beyond these cutoffs, the resulting effect size estimates are likely to be inflated due to the "winner's curse". To address this issue will require a fully specified formulation for the distribution of genetic models and the probability of observing significant association given models, similar to that of Xiao and Boehnke [2009].

For significant evidence of association (scenarios 2 and 4), we used a power value of 5% to estimate the plausible range of models. Because there almost certainly exist many causal variants across the genome, many of these of very modest effect, associations may often be detected even for low power models. Whether 5% power is sufficiently low for a definition of plausible models is therefore uncertain. Choosing a lower power value (say, 1%) rather than 5% will of course result in a wider range of plausible models.

In our analyses here, we assume the minor allele of variant C, denoted by A, is causal. When RAF_C is small (say < .01), C is relatively unlikely to be consistent with significant association result at genotyped variant M that has large risk allele frequency (say $RAF_M > .5$). However, if allele A is protective rather than causal, it will appear more in the controls than in the cases in a GWA study. The allele at M that does not share the same haplotype with A will appear to be the causal allele in association test, and can have a large RAF_M.

The set of plausible models depends on sample size. Evidence for linkage or association based on small samples more strongly restricts the range of plausible disease models (increased lower bound for RAF_C and decreased upper bound for GRR_C) than evidence based on large samples. Similarly, evidence against linkage or association based on large samples more strongly restricts the range of plausible models. In particular, as we carry out association studies based on increasingly large samples, our ability to restrict the range of models for newly detected loci will decrease.

In our analysis of negative linkage results, we calculate the power of the ASP test. In fact, many linkage studies collected nuclear families or larger pedigrees. Given the same number of samples, larger families often provide higher power to detect linkage than ASPs. The set of plausible models therefore might be reduced given little or no evidence for linkage in a similar sized study. Our work could be extended to calculate power for linkage based on the equivalent number of fully informative meioses [Edwards, 1988].

For linkage analysis, we assume IBD sharing between affected siblings is known. We

therefore ignore the uncertainty during IBD estimation, especially when the set of markers is not sufficiently dense. This estimation uncertainty reduces the power to detect linkage, and therefore raises the upper bound of plausible GRR_C than our current estimation. We also assume a multiplicative model [Risch, 1990a] for disease penetrance, which allows us to focus on locus-specific GRR_C and RAF_C when calculating power to detect linkage. Impact of this modeling assumption needs to be further explored.

Exploring genetic architecture underlying complex diseases/traits is of great interest to researchers. Purcell et al. [2009] used a score analysis and simulations to estimate the consistent models for the entire genome, specifically a combination of large numbers of common and rare causal variants, which could explain the observed association results in a GWA study of schizophrenia and the disease heritability. Their results suggested that rare or less common causal variants cannot be the sole source to explain genetic variation observed in schizophrenia. Dickson et al. [2010] studied the genetic architecture of rare or less common variants ($.005 < MAF < .02$) through simulations of genealogical trees, and showed that common variants identified in GWA studies could have been caused by multiple rare variants in high LD. Both of these studies mainly focused on the results from GWA studies. Purcell et al. [2009] did not use information provided by linkage analysis, while Dickson et al. [2010] used knowledge of linkage result to impose a general upper bound for GRR_C, specifically, $GRR_C < 6$, for all causal variants in their simulations.

To determine if disease etiology is due to rare or less common variant, our ideal goal would be to estimate the probability of genetic models given results from GWA and/or linkage studies,

say P(model | data). In this paper, we calculate the power to detect association and/or linkage for assumed genetic models, i.e., P(data | model), and use it as proxy to study P(model | data). To achieve this goal, we would need to examine the marginal distribution of models, i.e., P(model). Our knowledge about genetic models is still limited, but more information will be available over time as we learn more about the effect sizes and allelic spectrum of less common and rare causal variants.

In summary, we estimate ranges of plausible genetic models based on results from GWA and/or linkage studies for complex diseases. Given negative linkage or association results for a region of interest, we show that the effect size of rare or less common variants can be bounded, assuming such variants exist. Presence of both significant association and negative linkage results can further exclude genetic models for causal variants with small frequency. Our results may serve as a starting point for modeling genetic architecture of complex diseases as more information on disease etiology becomes available in the near future. The knowledge of plausible genetic models for a given region may allow more efficient design of sequencing studies to identify causal variants.

Table 4.1 T2D susceptibility loci reported in two GWA meta-analyses [Scott et al., 2007; Zeggini et al., 2008]. The value of RAF_C is an approximate lower bound, for which there is 5% power to detect association at observed p-value at M in a GWA study of corresponding sample size, and 95% power to detect linkage at observed MLS value using 6500 ASPs.

Nearby gene(s)	**RAF_M**	**OR**	**p**	**MLS**	**Minimum RAF_C**	**GRR_C†**
CDC123, CAMK1D	0.18	1.11	1.20×10^{-10}	0.0	.004	3.90
TCF7L2	0.18	1.34	1.30×10^{-08}	1.0	.019	2.95
TSPAN8, LGR5	0.27	1.09	1.10×10^{-09}	0.5	.004	4.51
IGF2BP2	0.30	1.14	8.90×10-^{16}	0.3	.018	2.72
CDKAL1	0.36	1.12	4.10×10^{-11}	0.0	.019	2.44
FTO	0.38	1.17	1.30×10^{-12}	1.5	.013	3.43
KCNJ11	0.46	1.14	6.70×10^{-11}	0.0	.027	2.23
JAZF1	0.50	1.10	5.00×10^{-14}	1.6	.012	3.56
HHEX	0.52	1.13	5.70×10^{-10}	0.2	.024	2.45
SLC30A8	0.61	1.12	5.30×10^{-08}	1.0	.017	3.05
NOTCH2	0.73	1.13	4.10×10^{-08}	0.0	.026	2.26
ADAMTS9	0.76	1.09	1.20×10^{-08}	0.3	.025	2.50
PPARG	0.82	1.14	1.70×10^{-06}	0.5	.038	2.31
CDKN2A/B	0.85	1.20	7.80×10^{-15}	0.5	.154	1.74
THADA	0.90	1.15	1.10×10^{-09}	0.2	.087	1.82

*. For MLS = 0, we use 0.05 as cut-off to calculate power to detect linkage

§. The TCF7L2 signal and corresponding power calculation are based on a smaller GWA (Fusion 1+2), which contains 2,376 cases and 2,432 controls.

†. GRR_C values correspond to minimum RAF_C.

Table 4.2 An approximate upper bound of RAF_M, for which genome-wide significant association ($p < 5\times10^{-8}$) at M in a GWA of *n* cases and *n* controls could be explained by a rare or less common variant (with given RAF_C) in D'=1, conditional on no or modest linkage evidence (MLS < 1) using *N* ASPs.

RAF_C	Linkage (*N*)								
	500			1000			5000		
	GWA (*n*)								
	1000	10000	50000	1000	10000	50000	1000	10000	50000
0.001	<0.05	0.18	0.53	<0.05	0.13	0.43	<0.05	<0.05	0.25
0.005	0.09	0.53	0.85	0.06	0.43	0.79	<0.05	0.24	0.62
0.01	0.17	0.7	0.92	0.12	0.61	0.88	0.05	0.4	0.77
0.05	0.54	0.92	0.95	0.44	0.89	0.95	0.24	0.78	0.94

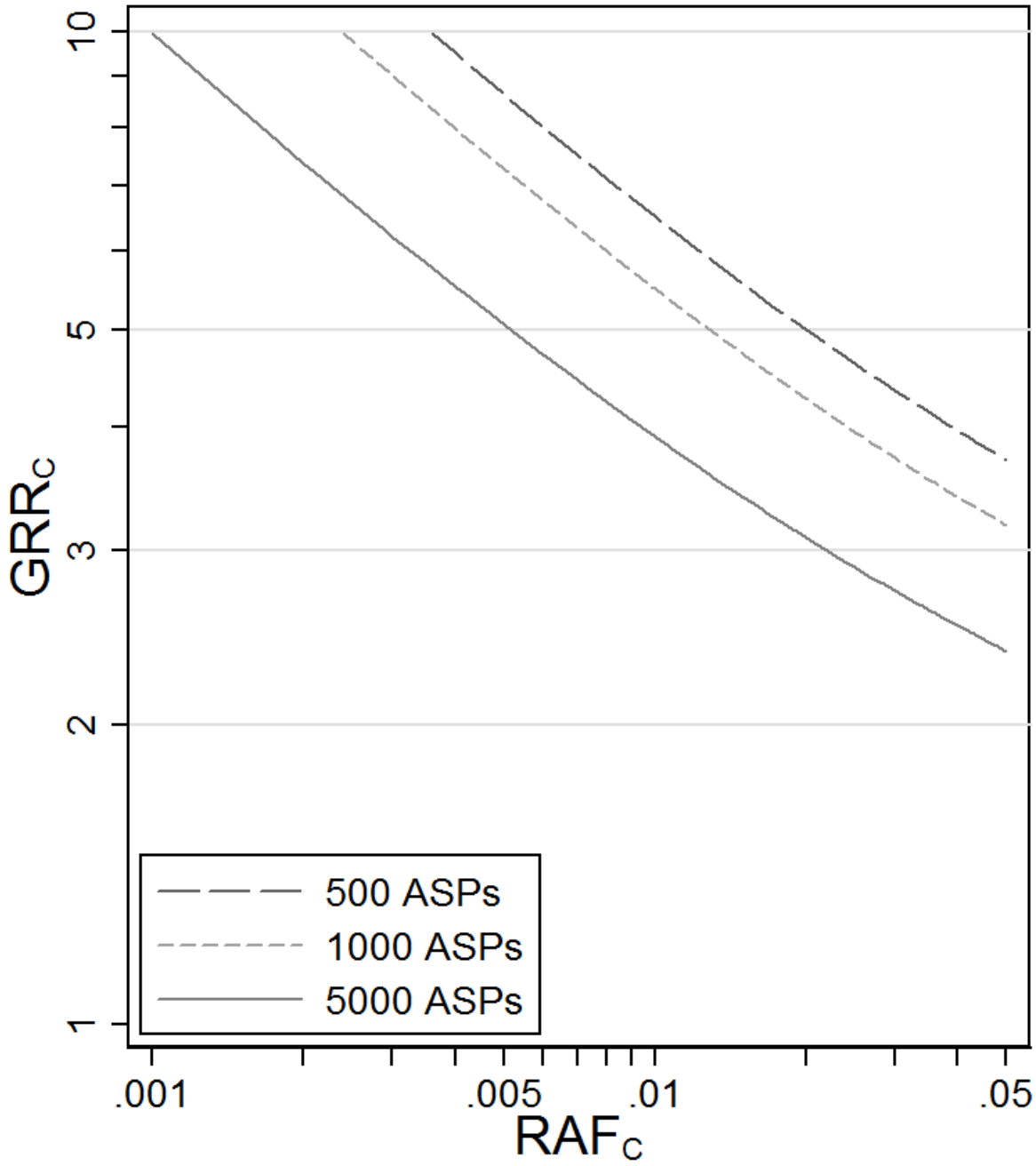


Figure 4.1 Genotype relative risks at causal variant C (GRR_C) that result in 95% power to detect linkage (MLS>1). The graphs are truncated at $GRR_C = 10$.

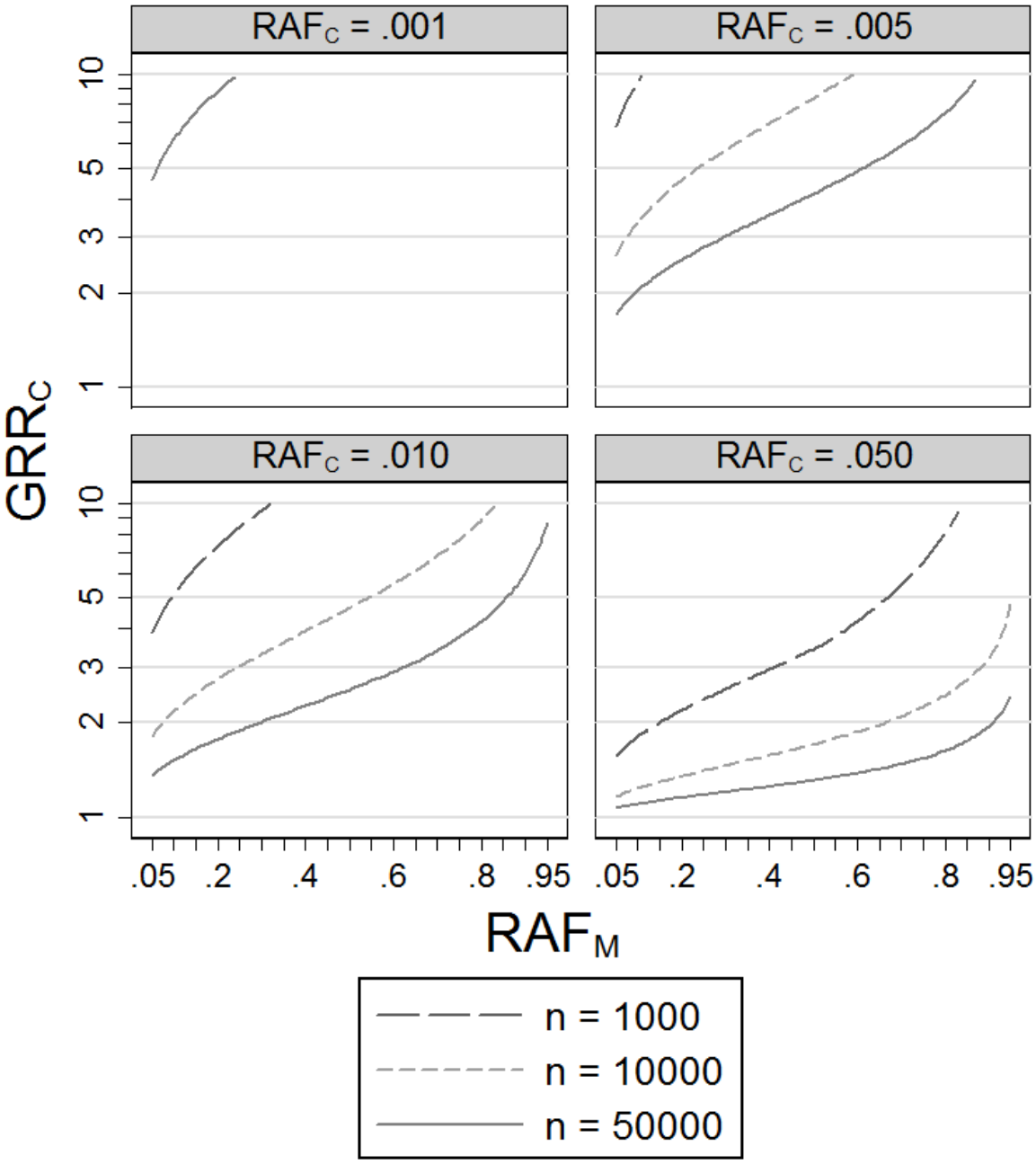


Figure 4.2 Genotype relative risks at causal variant C (GRR_C) that result in 5% power to detect association ($p < 5\times10^{-8}$) at genotyped variant M using *n* cases and *n* controls. We assume disease prevalence 10%, and D'=1 between M and C. The graphs are truncated at $GRR_C = 10$.

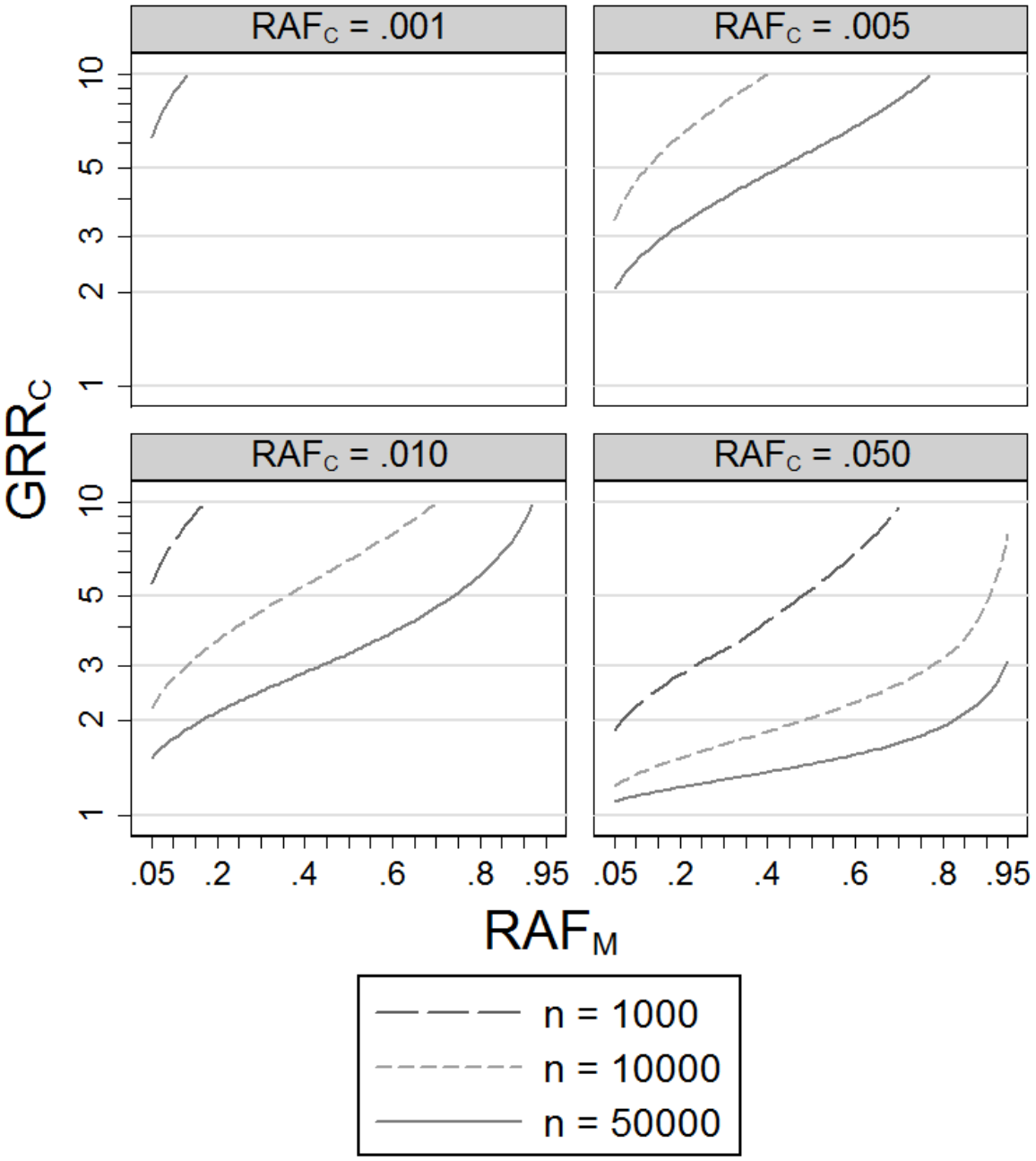


Figure 4.3 Genotype relative risks at causal variant C (GRR_C) that result in 95% power to detect association ($p < 10^{-4}$) at genotyped variant M using *n* cases and *n* controls. We assume disease prevalence 10%, and D'=1 between M and C. The graphs are truncated at $GRR_C = 10$.

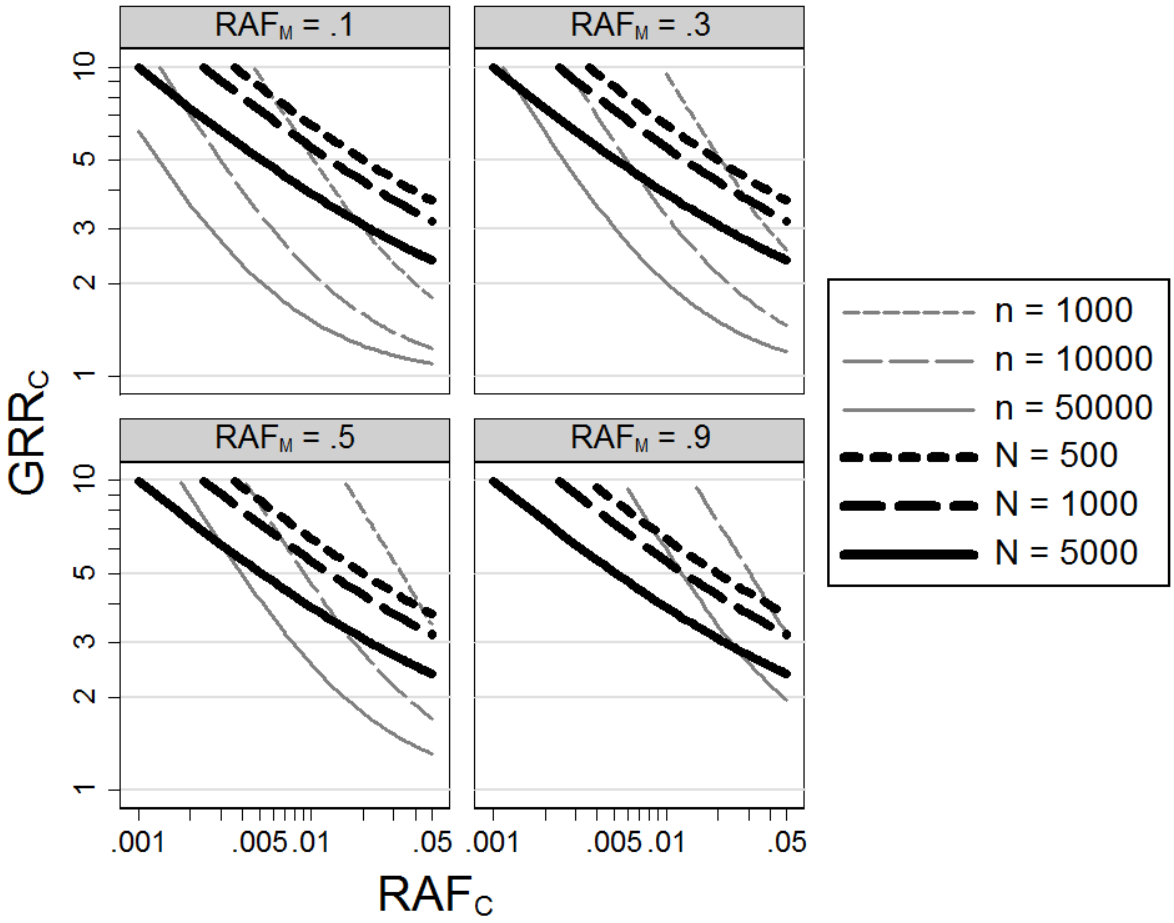


Figure 4.4 Genotype relative risks at causal variant C (GRR_C) that result in 95% power to detect linkage (MLS > 1) using *N* ASPs and 5% power to detect association ($p < 5\times10^{-8}$) at genotyped variant M using *n* cases and *n* controls. We assume disease prevalence 10%, and D'=1 between M and C. The graphs are truncated at $GRR_C = 10$.

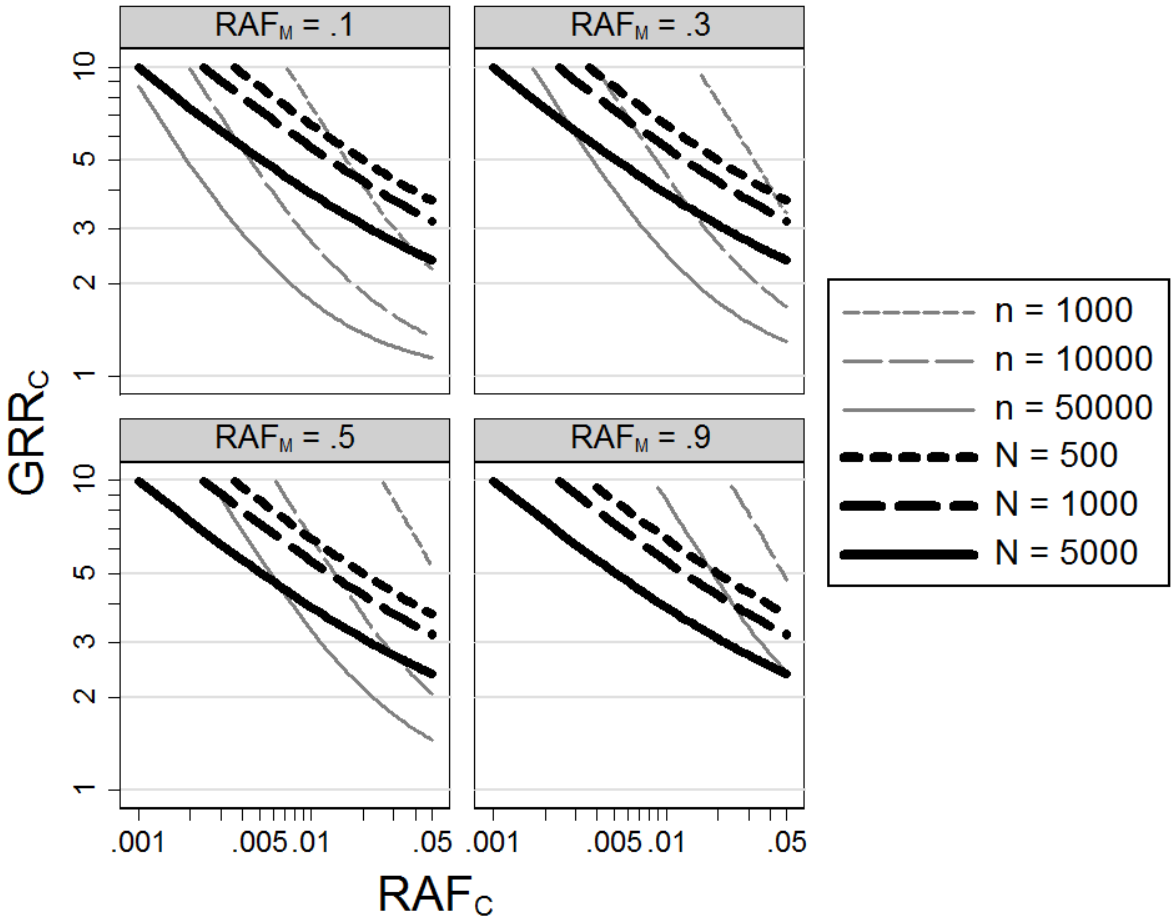
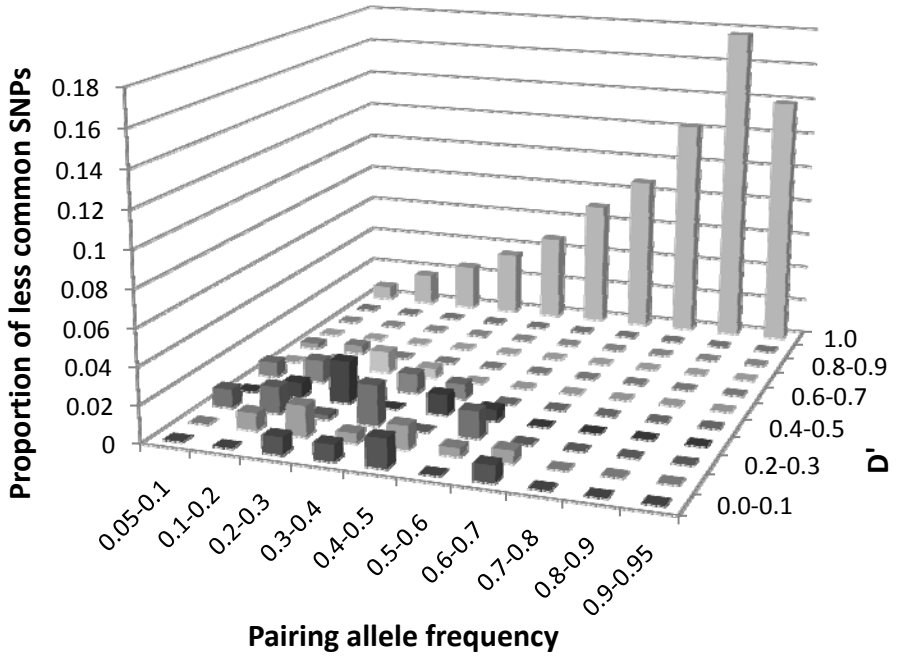


Figure 4.5 Genotype relative risks at causal variant C (GRR_C) that result in 95% power to detect linkage (MLS > 1) using *N* ASPs and 95% power to detect association ($p < 10^{-4}$) at genotyped variant M using *n* cases and *n* controls. We assume disease prevalence 10%, and D'=1 between M and C. The graphs are truncated at $GRR_C = 10$.

(a)



(b)

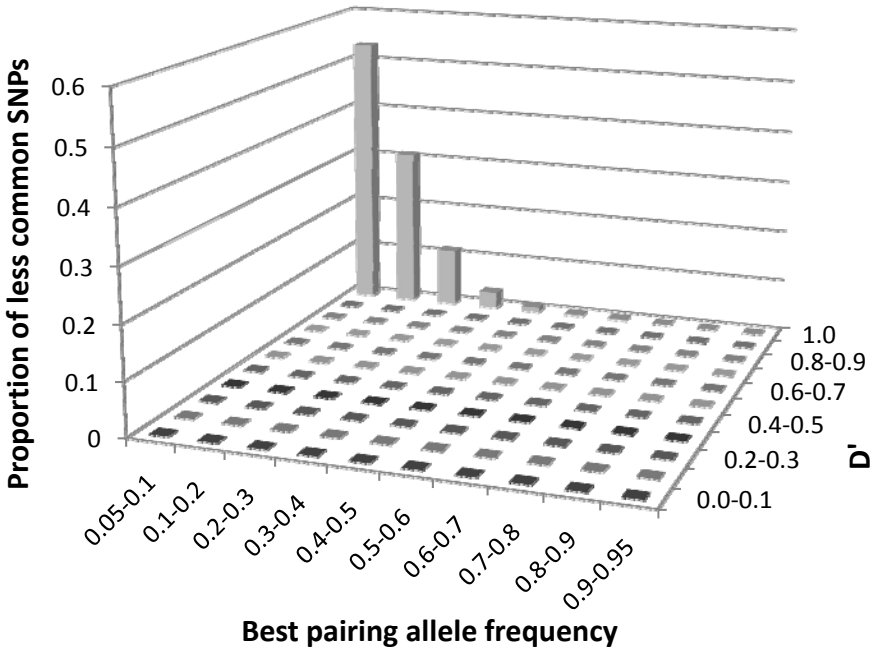


Figure 4.6 Distribution of D' between less common SNPs (.01<MAF<.03) in 1000 Genomes Project data and common SNP (MAF>.05) in HapMap2 on chromosome 1: (a) evaluate for all less common-common (HapMap) SNP pairs; (b) for each less common SNP, choose one paired common SNP with the highest D' and smallest allele frequency.

CHAPTER 5

SUMMARY AND DISCUSSION

A major goal of human genetic studies is to identify and characterize the variants that predispose to complex diseases or are responsible for variation of human quantitative traits. The basic strategy is to examine the relationship between the trait of interest and the genetic markers across human genome from sampled individuals. However, complex diseases may involve many genetic and environmental factors. Previous linkage and association studies have identified some but a far from complete set of the genetic variants underlying the complex diseases/traits. To achieve our goal, we need to employ efficient study designs, control for confounding factors such as environments and ethnicity, carry out powerful statistical analyses, and have good understanding of underlying genetic architecture. The work presented in this dissertation focuses on two aspects of the problems: confounding due to population stratification, and better understanding of genetic architecture and more efficient resequencing study design.

5.1 Summary of previous chapters

In Chapter 2, I presented a novel approach that can match unrelated cases and controls based on their genetic similarity calculated from genotypes in large-scale association studies. This approach takes advantage of the large amount of genetic information provided in GWA studies to make adequate assessment of ancestry. It is especially useful when samples are collected from

different geographic regions but their ethnicity information is not reported or may be inaccurate. I demonstrated that a carefully designed similarity score can distinguish individuals from different ancestry groups. Simulation results showed that with optimal matching our method can efficiently match individuals with similar genetic background, and therefore can achieve correct type 1 errors while maintaining the power to detect genetic predisposing variants in the presence of population substructure. I then applied our method to a GWA study of bipolar disorder. Using the unrelated case and control samples from this study, I showed that inflated type 1 error rates can be controlled when the samples were collected from two different sources. I believe that our approach will aid researchers in controlling for population stratification in GWA studies and reducing potential false positive results.

Many GWA studies include not only unrelated cases and controls, but family samples. To address potential population stratification in such a study design, in Chapter 3 I extended our matching approach to general family structures. With the extended matching, I proposed a test statistic that takes into account correlations between family members and between matched cases and controls that share the same ancestry. Using simulations of genetic data from affected sibling pairs, I showed that this approach can control for false positive rates and maintain the power of the association test in the presence of family data. I applied this extended approach to our GWA study of bipolar disorder which includes affected siblings. This approach greatly reduces the inflated false positive signals observed in the original study, and can make full use of the available samples. This approach adds flexibility to our previous genetic similarity matching method, and can be applied to any type of family samples to correct for population stratification in GWA studies.

Although GWA studies have achieved great success in identifying > 600 common genetic variants associated with common diseases and traits [http://www.genome.gov/gwastudies/], for most diseases the proportion of genetic variation explained by those variants is limited. With the rapid advances in biotechnology, especially the recently emerging next-generation sequencing, geneticists now can genotype and analyze a large variety of genetic variants that cover most of the human genome. It is of great interest to explore the genetic architecture underlying the complex diseases/traits. In Chapter 4, I describe power calculations for linkage and association tests, and estimate the range of plausible models (GRR and RAF) for rare or less common variants (.001 < RAF < .05) which are consistent with the results from previous linkage and/or GWA studies. Our results showed that given little or no evidence for linkage or association we can put upper bounds on the GRR_C of causal variants for a given MAFc (examined MAFc =.001-.05). Presence of significant association in GWA study combined with negative linkage results in the region of interest can reduce the plausible ranges of both GRR_C and RAF_C. I also evaluated results from a large-scale linkage study and two GWA meta-analyses for type 2 diabetes, and suggested ranges of plausible GRR_C and RAF_C for 15 loci identified as genome-wide significant in GWA. Our findings may provide insights on the plausible models underlying complex diseases, and help to design more efficient genetic studies.

5.2 Future work

For our genetic similarity matching method, when family data are present, multiple levels of correlation raise challenges for appropriate analysis. Our current approach in Chapter 3 can be

further improved. First, we propose a chi-square test statistic to compare the allele frequency between cases and controls. While this test can adjust the correlations between family members and between matched samples sharing the same ancestry, it cannot easily incorporate other covariates, such as environmental factors, into the analysis. One alternative is to test the association using a regression-based method. For example, we may regress the genotypes on the affection status and other covariates, using a linear mixed model framework. We will also compare its power to that of the proposed chi-square test in absence of other covariates. Alternatively, we can consider another regression model that regresses the estimated IBD sharing on the trait squared sums and squared differences among all pairs of relatives, similar to that proposed by Sham et al. [2001].

Secondly, when we match the family data, we currently choose one individual from each family randomly which loses information from other family members. There can be other matching schemes that will make better use of the data. For example, we might compute the average similarity from all pair-wise scores between samples in two families, and use the average scores as the distance measure for matching families. We will carry out simulations to evaluate the potential gains on matching accuracy using the new type of scores, and the expense on computing pair-wise scores using all family samples.

In chapter 4, we estimate the range of plausible models (GRR and RAF) for rare or less common variants underlying complex diseases and other human traits, given existing results from linkage and GWA studies. In power calculations for association tests at a genotyped variant, we assume that the association could be explained by a nearby single causal variant in high LD

(D’ = 1). Compared to the single-variant assumption, it is possible that multiple causal variants exist in a region of high LD. We would like to extend our analysis to the case of multiple variants, most likely by computer simulation. In addition, we examine the assumption of D’ = 1 between causal and genotyped variants using data from ongoing 1000 Genomes Project. Due to the small sample size currently available, we are only able to evaluate less common variants with MAF > .01, and may overestimate D’. As more sequencing data become available soon, we would like to re-run our analysis to obtain more accurate estimate for D’ and extend the analysis to variants with smaller frequencies. Alternatively, we could evaluate the distribution of D’ using computer simulations based on coalescence theory, similar to that in Dickson et al. [2010].

Ideally, we would like to calculate the probability of specific genetic model given linkage and/or association results. To do so, we would need knowledge about the marginal distribution of models, specifically, the frequency spectrum for rare or less common variants and distribution of GRR. It is of interest to explore different assumptions on such distributions, for example, uniform/exponential distribution for GRR and U-shaped distribution for RAF, and evaluate their impact on our conclusions.

5.3 Conclusion

In summary, I proposed new approach to correct for population stratification in GWA studies. My approach can make use of the large amounts of genotype data available in GWA studies and correctly match individuals with similar genetic background. With the extended version of the method, we can also analyze family data. I also explored the range of plausible genetic models

(in terms of GRR and RAF) for rare or less common variants underlying complex diseases and traits, which are consistent with existing GWA and/or linkage results. We hope that our findings will contribute to more effective design of future genetic studies, and inspire the development of novel statistical tools to identify disease causal variants.

REFERENCES

Abecasis GR, Cardon LR and Cookson WO. 2000. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279-292.

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. Science 322:881-888.

Bertsekas DP, Tseng P. 1994. RELAX-IV: A faster version of the RELAX code for solving minimum cost flow problems. Technical report LIDS-P-2276, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365-1369.

Boustein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33:228-237.

Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS. 2003. Novel case-control test in a founder population identifies p-selectin as an atopy-susceptibility locus. Am J Hum Genet 73:612-626.

Broeckel U, Hengstenberg C, Mayer B, Holmer S, Martin LJ, Comuzzie AG, Blangero J, Nurnberg P, Reis A, Riegger GA, Jacob HJ, Schunkert H. 2002. A comprehensive linkage analysis for myocardial infarction and its related risk factors. Nat Genet 30:210-214.

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

Ding C, He X. 2004. K-means clustering via principal component analysis. Proc Intl Conf Machine Learning (ICML 2004). p 225-232.

Donnelly P, Tavaré S. 1995. Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401-421.

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

Epstein MP, Allen AS, Satten GA. 2007. A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 80:921-930.

Edwards JH. 1988. The analysis of close linkage in large families. Ann Hum Genet 52:299-307.

Falconer DS. 1965. The inheritance of liability to certain diseases, estimated from the incidences

among relatives. Ann Hum Genet 29:51-79.

Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK, Smoller JW, Grozeva D, Stone J, Nikolov I, Chambert K, Hamshere ML, Nimgaonkar VL, Moskvina V, Thase ME, Caesar S, Sachs GS, Franklin J, Gordon-Smith K, Ardlie KG, Gabriel SB, Fraser C, Blumenstiel B, Defelice M, Breen G, Gill M, Morris DW, Elkin A, Muir WJ, McGhee KA, Williamson R, MacIntyre DJ, MacLean AW, St CD, Robinson M, Van Beck M, Pereira AC, Kandaswamy R, McQuillin A, Collier DA, Bass NJ, Young AH, Lawrence J, Ferrier IN, Anjorin A, Farmer A, Curtis D, Scolnick EM, McGuffin P, Daly MJ, Corvin AP, Holmans PA, Blackwood DH, Gurling HM, Owen MJ, Purcell SM, Sklar P, Craddock N; Wellcome Trust Case Control Consortium. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet 40:1056-8.

Frangioni A, Manca A. 2006. A computational study of cost reoptimization for min cost flow problem. INFORMS J Comput 18:61-70.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388-393.

Freidlin B, Zheng G, Li Z, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53:146-152.

Goldstein DB. 2009. Common genetic variation and human traits. N Engl J Med 360:1696-1698.

Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. 2006. Centralizing the non-central chi-square: a new method to correct for population stratification in genetic case-control association studies. Genet Epidemiol 30:277-289.

Gu XS, Rosenbaum PR. 1993. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat 2:405-420.

Guan W, Liang L, Boehnke M, Abecasis GR. 2003. Matching cases and controls using genotype data from a whole genome association study. ASHG 2005 Annual Meeting #2395 (poster).

Guan, W., Pluzhnikov, A., Cox, N.J., Boehnke, M. for the International Type 2 Diabetes Linkage Analysis Consortium. 2008. Meta-analysis of 23 type 2 diabetes linkage studies from the international type 2 diabetes linkage analysis consortium. Hum Hered 66:35-49.

Hansen BB. 2004. Full matching in an observational study of coaching for the SAT. J Am Stat Assoc 99:609-618.

Hansen BB, Klopfer SO. 2006. Optimal full matching and related designs via network flows. J Comput Graph Stat 15:609-627.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362-9367.

Hodge SE. 1984. The information contained in multiple sibling pairs. Genet Epidemiol 1:109-22.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183-201.

Hudson RR. 1990. Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7:1-44.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931-945.

International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748-752.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137.

Khlat M, Cazes MH, Genin E, Guiguet M. 2004. Robustness of case-control studies of genetic factors to population stratification: magnitude of bias and type I error. Cancer Epidemiol Biomarkers Prev 13:1660-1664.

Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. 2007. A Randomization Test for Controlling Population Stratification in Whole-Genome Association Studies. Am J Hum Genet 81:895-905.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893-903.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308:385-389.

Li CC. 1972. Population subdivision with respect to multiple alleles. Ann Hum Genet 33:23-29.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311-321.

Li Y, Abecasis GR. 2006. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet S79:2290.

Liang L, Zöllner S, Abecasis GR. 2007. GENOME: a rapid coalescent-based whole genome simulator. Bioinfomatics 23:1565-1567.

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. 2008. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453-463.

Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118:1590-1605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747-753.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PVK, Frazer KA, Cox DR, Ballinger DG. 2005. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77:685-693.

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

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

Newman DL, Abney M, McPeek MS, Ober C, Cox NJ. 2001. The importance of genealogy in determining genetic associations with complex traits. Am J Hum Genet 69:1146-1148.

NIMH Genetics Workgroup. Genetics and mental disorders. *NIH Publication* No. 98-4268. Rockville, MD: National Institute of Mental Health, 1998.

Pankow JS, Heiss G, Evans GW, Sholinsky P, Province MA, Coon H, Ellison RC, Miller MB, Qaqish B. 2004. Familial aggregation and genome-wide linkage analysis of carotid artery plaque: the NHLBI family heart study. Hum Hered 57:80-89.

Press W, Teukolsky S, Vetterling W, Flannery B. 1992. Numerical recipes for C, 2nd ed. Cambridge: Cambridge University Press.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured

populations. Am J Hum Genet 67:170-181.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and populationbased linkage analyses. Am J Hum Genet 81:559–575.

Risch, N. 1990a. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46: 222-228.

Risch, N. 1990b. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46: 229-241.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516-1517.

Rosenbaum PR. 1991. A characterization of optimal designs for observational studies. J Roy Statist Soc Ser B 53:597-610.

Rosenbaum PR. 2002. Observational Studies, 2nd ed. New York: Springer.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331-1336.

Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, Upmanyu R, Tozzi F, Li JZ, Burmeister M, Absher D, Thompson RC, Francks C, Meng F, Antoniades A, Southwick AM, Schatzberg AF, Bunney WE, Barchas JD, Jones EG, Day R, Matthews K, McGuffin P, Strauss JS, Kennedy JL, Middleton L, Roses AD, Watson SJ, Vincent JB, Myers RM, Farmer AE, Akil H, Burns DK, Boehnke M. 2009. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc Natl Acad Sci USA 106:7501-7506.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341-1345.

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

Skol AD, Xiao R, Boehnke M, Veterans Affairs Cooperative Study 366 Investigators. 2005. An algorithm to construct genetically similar subsets of families with the use of self-reported ethnicity information. Am J Hum Genet 77:346-354.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881-885.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-513.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Xiao R, Boehnke M. 2009. Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol 33:453-462.

Visscher PM, Hill WG, and Wray NR. 2008. Heritability in the genomics era–concepts and misconceptions. Nat Rev Genet 9:255-266.

White H. 1982. Maximum likelihood estimation of misspecified model. Econometrica 50:1-25.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76:887-893.

Wray NR, Goddard ME, Visscher PM. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17:1520-1528.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. 2007. Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. Science 316:1336-1341.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Boström KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson

AU, Jørgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjögren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ; Wellcome Trust Case Control Consortium, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638-645.

Zhang S, Kidd KK, Zhao H. 2002. Detecting genetic association in case-control studies using similarity-based association tests. Statistica Sinica 12:337-359.

Zhu X, Li S, Cooper RS, Elston RC. 2008. A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 82:352-365.