**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yunxuan Jiang Date

**Identification of the Effect of Population Stratification on Association Studies of Rare Variants**

**BY**

**Yunxuan Jiang**

**Master of Science in Public Health**

**Biostatistics**

---

**Karen N. Conneely, Ph.D.**

**Committee Chair**

---

**Michael P. Epstein, Ph.D.**

**Committee Chair**

**Identification of the Effect of Population Stratification on Association Studies of Rare Variants**

**BY**

**Yunxuan Jiang**

**Bachelor of Science**

**Beijing Forestry University**

**2009**

**Thesis Committee Chair: Karen N. Conneely, Ph.D**

**Michael P. Epstein, Ph.D.**

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2011

# Abstract


**Identification of the Effect of Population Stratification on Association Studies of Rare Variants**

**BY Yunxuan Jiang**

Human genome research, which aims to find the genetic etiology of the disease, is having a more and more profound influence on public health. And rare variants, which both have large effect size and can explain a great proportion of heritability, are becoming the focus of current human genome research. Although several statistical methods have developed to increase the power of detecting rare variants and reduce false positive rate, none of these methods address an important issue that often arises in genetic studies: false positives due to population stratification. Population stratification is a well-known problem that can substantially cause inflated false positive rate and decreased power to detect real association. We simulated several case-control studies with different sample size and population structure according to a series of disease prevalence for each population (Europea and Africa), and found that population stratification can have a significant influence on rare variants studies. The false positive rate increases dramatically as sample size increase and population structure become extreme. We applied principal component analysis to control for population structure. Our results showed that the principal component method performed very well even for highly structured data. The false positive rate remained around 0.05 in our simulation. Our results implicates that researchers need to carefully match case and control ancestry, in order of avoid false positive caused by population structure in rare variants study. If it is inevitable to recruit samples from different population, then researchers can correct for it with our easy implemented method.

# Identification of the Effect of Population Stratification on Association Studies of Rare Variants

**BY**

**Yunxuan Jiang**

**Bachelor of Science**

**Beijing Forestry University**

**2009**

**Thesis Committee Chair: Karen N. Conneely, Ph.D**

**Michael P. Epstein, Ph.D**

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2011

# Acknowledgements

I would especially thank Drs. Michael Epstein and Karen Conneely for their unconditional help throughout the thesis and two years of my master study. I would like to thank Dr. John Hanfelt for being the reader of the thesis. I also would like to thank Dr. Yun Li for suggestions on simulating haplotype data. Finally, I would like to thank all faculty members and staff in the department of Biostatistics and Bioinformatics. The excellent courses and research projects here make me better prepared for my future academic endeavor.

# Table of Contents

**Chapter 1 Introduction**.....**1**

**Chapter 2 Review of the Literatures** .....**6**

2.1 Common vs. Rare Variants .....7

2.2 Study Design.....11

2.2.1 Linkage Analysis.....12

2.2.2 Association Studies.....13

2.2.2.1 Candidate Gene Association Studies .....13

2.2.2.2 Genome Wide Association Studies.....14

2.3 Population Stratification .....16

2.3.1 Genomic Control.....18

2.3.2 Structure.....19

2.3.3 Principal Component Analysis .....19

2.4 Statistical Methods for Analyzing Rare Variants .....20

**Chapter 3 Methodology** .....**21**

3.1 Simulating Population Specific Haplotype .....23

3.1.1 Building the Genealogy ..... 23

3.1.2 Mutation ..... 25

3.1.3 Adding Neutral Mutations to the genealogy ..... 26

3.1.4 Migration ..... 27

3.1.5 Recombination ..... 29

3.2 Simulating a case-control Study ..... 30

3.3 Simulating GWAS Data ..... 32

3.4 Methods to Calculate Principal Components ..... 32

3.5 Testing for Association Between Rare Variants and Disease ..... 34

**Chapter 4 Results** ..... **36**

4.1 Simulated Study ..... 37

4.2 False Positive Rate ..... 37

4.3 Correcting for Population Stratification using Principal Components ..... 40

**Chapter 5 Conclusions, Implications and Recommendations** ..... **43**

**Reference** ..... **47**

# List of Tables


1. Study designs for simulating case-control study ..... **31**
2. Type 1 Error Rate before Correction ..... **41**
3. Type 1 Error Rate Corrected by Self-Reported Race ..... **41**
4. Type 1 Error Rate Corrected by Principal Components ..... **42**

# List of Figures


**1.** Odds Ratio for common vs. rare variants .....**9**

**2.** Study design for different allele frequency and effect size .....**11**

**3.** Type 1 error rate for different sample size .....**17**

**4.** Simulation flow chart.....**22**

# Chapter I Introduction

Genomics is playing a more and more important role in public health research. According to Center for Disease Control and Prevention (CDC), genetic factors are associated with nine of the ten leading causes of death in the United States, including heart disease, cancer, diabetes, and Alzheimer disease. There are already examples that pharmaceutical research has incorporated genomic findings, and Eric Green, the director of National Human Genome Research Institute (NHGRI), expects similar examples to appear in the next 5 to 10 years (Hayden, 2009). Genome wide association studies (GWAS), are a widely used design to identify susceptible single nucleotide polymorphisms (SNPs) that are associated with disease under the assumption that such diseases originate from the effects of common variants. During the past few years, GWAS have successfully identified a large number of SNPs that have strong associations (p-value smaller than $5\times10^{-8}$) with several diseases. According to NHGRI, 1212 genome wide associations have been identified to have strong association with 210 traits as of December 2010 (for a specific list, see (http://www.genome.gov/gwastudies)). These accomplishments are especially valuable for complex disease such as type 2 diabetes and schizophrenia.

However, most of these associated SNPs have very small effect sizes (odds ratio between 1.1-1.5), and the proportion of heritability (proportion of phenotypic variance in a population attributable to additive genetic factors) explained by these SNPs is at best modest for most traits. For example, type 2 diabetes has 21%-72% heritability while current findings can only explain 6% of it (Mathias et al., 2009; Manolio et al., 2009). Besides, there is not enough evidence to show that these common variants have a causal effect on the disease (Maher, 2008; Cirulli & Goldstein, 2010). Through a series of

review articles, researchers pointed out that common disease might be caused by a series of rare variants, each conferring a moderate but highly detectable increase in relative risk (Pritchard, 2001; Bodmer & Bonilla, 2008; Schork et al., 2009; Manolio et al., 2009). Rare variants have been shown to influence individual risk for autism, epilepsy and schizophrenia (Stankiewicz and Lupski, 2010). Furthermore, the odds ratios of rare variants are normally above 2, which are significantly higher than the odds ratios reported for common variants (Bodmer & Bonilla, 2008). There is also a chance that current findings of association between the common variants and the disease are actually caused by rare variants found nearby those common variants (Dickson et al., 2010). Recent development of cost-effective sequencing technologies, especially next-generation sequencing methods, has made direct sequencing of rare variants feasible.

As power to detect an individual rare variant is low (since power generally decreases with a decrease in frequency when the sample size and effect size are held constant), many statistical methods for rare-variant analysis develop tests that combine information from rare variants in a region into a composite variable and then test for association between the variable and disease (see the 'Literature Review' section for more detail on these methods). However, none of these methods address an important issue that often arises in genetic studies: false positives due to population stratification.

Population stratification is a well-known problem that can cause inflated false positive rates and decreased power to detect real association. Population stratification is a

systematic difference in allele frequencies between cases and controls caused by sampling of subjects from different populations whose disease prevalence and allele frequencies are significantly different from each other. This is because each population has a unique social and genetic background; and social or cultural events, such as the mating process, will greatly influence the genetic architecture of a population (Cardon, 2003). Marchini *et al*. (2004) showed that with the sample size required by genome wide association studies, even small fraction of admixture between different populations will lead to a greatly inflated type 1 error rate. And the type 1 error rate increases rapidly as the sample size grows large or the population structure becomes more extreme.

The problem caused by population stratification cannot be simply solved by using self-reported race. This is because race is not a comprehensive representation of one's ancestral make-up. Barholtz-Sloan et al. (2005) examined a case-control study on early onset lung cancer and found that ancestry information inferred by genetic markers do not exactly match self-reported race. And subjects from different races, such as Caucasian, non-Hispanics and African Americans have a significant proportion of ancestry that is overlapped. Researchers have developed several methods to correct for population structure (Marchini et al., 2004; Pritchard et al., 2000; Price et al., 2006); the basic idea of these methods is to infer the information about population structure through a set of genetic markers.

Although it is clear that population stratification is a severe problem in association studies of common variants and although the analysis of rare variants is becoming an increasingly hot topic, nothing is known about the effect of population stratification on rare variants, which inspired us to investigate this problem. We have two very clear aims in this thesis: 1) Examine whether population stratification affects studies of rare variants; 2) if there exists an effect, we aim to evaluate the effectiveness of existing methods (adjustment using principal components) to solve this problem.

# Chapter II Review of the literature

In this section, we will review current hypotheses about the genetic architecture of common disease and justify why we want to focus on rare variants studies. We will discuss different methods to identify disease associated alleles and the advantages and disadvantages for each of them. We will also review the population stratification problem in genetic association studies and available statistical methods to correct for population stratification.

## 2.1 Common vs. Rare Variants

Current studies have identified many alleles that contribute to several complex diseases, most of which have profound public health influence. For example, Hunter et al. have identified *FGFR2* alleles associated with risk of sporadic postmenopausal breast cancer (Hunter et al., 2007); Scott et al. (2007) have identified alleles associated with type 2 diabetes which is the 6$^{th}$ major cause of death in the US. These findings contribute to understanding of the disease etiology, which can further help prevention, diagnosis and treatment of disease (Manolio et al., 2009). However, a great proportion of heritability of the diseases cannot be explained by current findings, which is a major concern to researchers.

Current genome wide association study (GWAS) is carried out under the "common disease common variants" (CDCV) assumption. The CDCV theory assumes that common diseases are caused by alleles that have moderate frequency in the population. These common alleles identified by GWAS mostly have odds ratio between 1.1-1.5, and the proportion of heritability (proportion of phenotypic variance in a population attributable

to additive genetic factors) that can be explained by these variants is small (Manolio et al., 2009). Weedon et al. identified 20 common variants in a sample of 30,147 subjects associated with human height, which has overall heritability of 90%. However, these 20 variants can only explain 3% of height variation (Weedon et al., 2008). Since the proportion of heritability that can be explained by these variants is so small, it is very hard to build any prevention strategies based on these findings.

The common disease rare variants hypothesis (CDRV), as its name suggests, assumes that common human disease is attributed to a group of alleles with relatively low frequency but high penetrance (the probability of having the disease given that the person carries the allele). Each of these alleles acts independently and contributes moderately to the variation in disease risk. (Bodmer & Bonilla, 2008). There are several points that support the CDRV assumption. First, the odds ratio of rare variants identified in association studies is much larger than those of common variants identified in association studies. Bodmer & Bonilla (2008) summarized the odds ratios (OR) of 61 rare variants and 327 common variants and the results are shown in Figure 1. The authors noted that, for common variants, relatively few have OR values above 2, and the mean OR is 1.26; for the rare variants, most have OR above 2, and the mean OR is 3.74 (Bodmer & Bonilla, 2008).

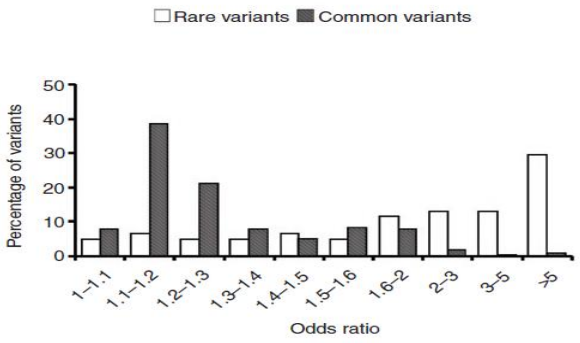


Figure 1. Odds Ratio for common vs. rare variants (Bodmer & Bonilla, 2008)

Second, rare variants are more likely to have causal effects than common variants. Genetics can benefit public health by providing more insight into the etiology of a disease. In this sense, only variants that have "causal effects" are meaningful findings. Alleles identified to be associated with the disease don't necessarily mean they will cause the disease. A GWAS is performed under the assumption that the identified alleles are in close proximity with the actual functional variants which have the causal effect. In theory, the effect sizes of these alleles are so small that it is hard to find the actual functional variants based on these alleles. In practice, researchers do rarely establish causal relations between these common alleles and the disease. Researchers further propose a hypothesis that these common alleles might be related to gene expression. However, the intersection set of alleles associated with gene expression and alleles associated with disease is almost empty (Circulli and Goldstein, 2010). Rare variants, on the contrary, are always identified to cause the functional effect themselves. They are expected to change amino acids and further influence the protein-protein interaction (Bodmer & Bonilla, 2008).

Pritchard (2001), supports the CDRV hypothesis from a population-genetics aspect. Pritchard pointed out a population-level process, such as random genetic drift, will select

against disease related mutations. Common variations are likely to be older and hence have been subjected to potential selective forces over time, such that they are likely to have less effect on disease risk. However, rare variants are either likely to be new and hence have not gone through a long period of negative selection, or are rare because they are selected against (Schork et al., 2009). Pritchard (2001) supports his point through solid statistical and simulation evidence, his simulation result shows that the allele frequencies of variants that influence common disease are unlikely to be moderate as assumed in CDCV.

As discussed above, although common variants have contributed a lot to understanding the genetic architecture of a disease, they provide limited information about the etiology of a disease. Their low penetrance also makes them less likely to benefit public health compared to rare variants. Public health practitioners implement prevention strategies, such as screening, based on the penetrance. As the penetrances of common variants are so small, it is hard to develop convincing strategies based on them. However, the penetrances of rare variants are normally large enough to justify preventative screening strategies, and thus, have more potential to influence public health (Bodmer & Bonilla, 2008).

Previous study designs and sequencing technologies limited the development of rare variants study. Linkage analyses, which try to identify the gene tha co-segregates with the disease, normally can identify alleles with very high effect size, and are helpful for

Mendelian disease (disease associated with only one gene, will discuss in detail later). GWAS use tagSNPs; although they can detect alleles with small effect size, the minor allele frequency needs to be relatively high, at least 5%. So rare variants are in a dilemma, since the effect of a single allele is not large enough to be identified through linkage analysis and the allele frequency is not common enough to be captured through genotyping in a GWAS study. (Figure 2, Manolio et al., 2009).

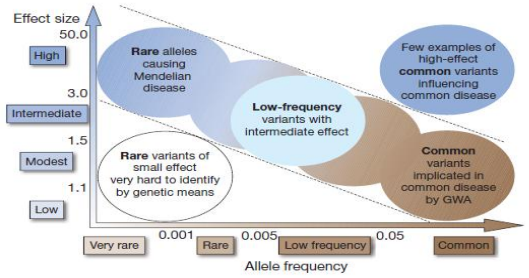


Figure 2. Study design for different allele frequency and effect size (odds ratio) (Manolio et al., 2009)

However, the development of sequencing technology, especially "next generation sequencing" has made sequencing rare variants feasible. "Next-generation sequencing" technology is economical and fast; it can process millions of sequence reads in parallel comparing to only 96 reads by previous technology. The identification of rare variants will also be facilitated by the 1000 Genomes project. Instead of genotyping the tagSNPs, the 1000 Genomes Project is trying to build a comprehensive catalogue of variants with minor allele frequency greater than 1% (some variants with lower frequencies are also identified). The project has successfully identified more than 11 million new SNPs in initially low-dept coverage of 172 individuals (Manolio et al., 2009).

## 2.2 Study Design

### 2.2.1 Linkage Analysis

In the early stage of genetic research, researchers tried to identify disease genes under monogenic 'Mendelian disease' through linkage analysis. "Linkage" presents when the co-segregation of a chromosomal region and the disease appears more than expected by chance. Linkage analysis is performed by comparing the likelihood of observing data given that the loci are linked to the likelihood of observing data given that the loci are unlinked. There are two types of linkage analysis: parametric linkage analysis and non-parametric linkage analysis. The parametric linkage analysis needs specification of a disease model such as frequency of the allele and penetrance. The non-parametric linkage analysis does not need any models; it just tests whether the co-segregation deviates from independent assortment. Genes identified by linkage analysis are normally quite rare and have very large effect size. Linkage analysis can be successful under the condition that markers linked with the disease gene segregate with the disease in families. These characteristics of linkage analysis give it several limitations. First, it needs to be studied within families, which can make the recruitment of research subjects difficult. Also, parametric linkage analysis requires the disease gene to be linked with markers and correct specification of the penetrance model. More importantly, it is limited to 'Mendelian disease' (disease caused by a single gene with high penetrance), which is not a proper assumption for complex disease. Due to the nature of linkage analysis, it is very hard to identify the real variants that cause the complex disease (Altmuller, J., 2001). Although linkage analyses have achieved some success in identifying variants that contribute to complex disease, such as type 1 diabetes (Bennett et al., 1994), in most

cases mutations identified by linkage analysis can only explain a small fraction of the overall heritability of the disease (Altmuller, 2001).

### 2.2.2 Association Studies

As an alternative approach, association studies have been proposed as a powerful means of identifying genetic factors contributing to complex disease (Risch & Merikangas, 1996; Hirschorn and Daly, 2005), and have demonstrated power by successfully identifying alleles associated with many common diseases. For example, Need et al. have identified alleles and copy number variants associated with schizophrenia (Need et al., 2009). Unlike linkage analysis, association studies aim to find disease-predisposing alleles at the population level. They test whether a particular allele, genotype or haplotype will be seen more often than expected by chance in a disease subject. The simplest way to do this is to compare the frequency of alleles or genotypes of a variant between cases and controls. As association studies are not restricted to families, their research subjects are more convenient to recruit. More importantly, since association studies use unrelated cases and controls, they have more power than linkage analysis to identify disease associated alleles. Here we present characteristics of different association study designs as well as the advantages and disadvantages of each of them.

#### 2.2.2.1 Candidate Gene Association Study

Candidate-gene association studies have identified several genes that are associated with common disease. Barroso et al. identified alleles associated with type 2 diabetes through their influence on insulin action (Hirschorn and Daly, 2005; Barroso et al., 2003). However, candidate gene studies are not as powerful as researchers expected. Only

"candidate" genes are examined in the study; however, as discussed above, common disease has a complex genetic architecture such that each candidate gene has limited effect on the disease. More importantly, candidate-gene association studies are hypothesis tests. The hypotheses generally come from previous linkage analysis studies or biological inference. The success of candidate-gene association studies is under the premise that these hypotheses are the right ones. Even if these hypotheses are right, and the hypotheses are broad enough to embrace several genes, their findings can explain a limited fraction of heritability.

#### 2.2.2.2 Genome Wide Association Study

The completion of the Human Genome Project, the deposition of millions of SNPs into public database, rapid improvements in SNP genotype technology and the International HapMap Project have made the genome-wide association approach become feasible. A genome-wide association study is a hypothesis free test that normally tests 300,000 or more markers (SNPs, single-nucleotide polymorphisms) that are spread evenly across the genome (Hardy and Singleton, 2009), and looks for variations between individuals with and without disease. Unlike candidate-gene association studies, genome wide association studies are hypothesis free tests, where no "candidates" were proposed. As a result, we can examine a large set of genes with no need to worry about whether we made the right assumption about the "candidates" or not. This approach has become more and more popular as the genotyping technology improved and the cost of genotyping reduced rapidly. In summary, a genome-wide association study is a comprehensive approach to identify disease associated alleles, especially when we don't have any solid evidence

about the “candidate” genes (Hirschhorn and Daly, 2005). Genome-wide association studies have been successful; identifying common alleles associated with over 200 traits.

Many study designs are available for association analyses, which can be broadly broken down into family-based designs and population-based studies. In this thesis, we focus on the population-based case-control study. Case-control design has many favorable characteristics in genetic association studies. Cases and controls are easier to enroll than family-based subjects, which can save time, energy and expenses. Also, cases and controls are not genetically related, which can increase the power to identify disease associated alleles. Besides, case-control study is a well-understood study design, due to its wide application in epidemiology. Finally, disease-allele frequency, penetrance, and population attribute risk can be estimated all through this study design (Cardon and Palmer, 2003).

However, like other case-control studies, this type of study design assumes that the detected differences in allele frequencies are the real cause of the disease outcome, or equivalently, that there are no confounding effects. Unfortunately, this is normally not the truth, especially for large scale genome wide association studies. One of the most common types of possible confounding is caused by population stratification, which is discussed below.

## 2.3 Population Stratification

Genome-wide association studies identify disease associated alleles by comparing allele frequencies between cases and controls. However, some allele frequencies are different between cases and control, and have nothing to do with the disease; population stratification is one of the major reasons to cause this "spurious association". Population stratification is the differences of allele frequencies between cases and controls caused by sampling of subjects from different populations whose disease prevalence and allele frequencies are significantly different from each other (Price et al., 2006). It is a major cause of spurious differences of allele frequencies between cases and controls. There are several reasons that ancestry differences can lead to differences in allele frequencies. Each population has their unique genetic and social history. And the social events, such as migration, agricultural, and mating practice largely influence the genetic background of a population (Cardon and Palmer, 2003).

The most frequently cited example comes from a study of the association between an HLA haplotype and diabetes on a Pima Indian reservation. This study tried to identify the association between haplotype Gm3:5,13,14 with reduced risk of non-insulin-dependent diabetes mellitus (NIDDM). Only 1% of full heritage American Indian population have Gm3:5,13,14, while 66% of Caucasian population have the haplotype. And the prevalence of NIDDM in American Indian is 40%, while the prevalence is 15% in Caucasian. In this case, both allele frequencies and disease prevalence are different in two populations. When the two populations are mixed together, the results show that there is an significant association between the haplotype and the reduced risk of NIDDM, the odds ratio is 0.27 with 95% CI (0.18, 0.40). However, when the analysis was

restricted to full-Papago Indians, this association disappeared. Another example is about a study trying to identify the relation between CYP3A4 variant and prostate cancer. Although instead of sampling from different populations, this study restricted to African-Americans, a false positive still occurred. Before correcting for population stratification, the result shows a significant relation between the variation and the disease (p-value=0.0007). However, after using genomic control (discussed in detail later) the significant result disappeared; the p-value is 0.254.

Marchini et al. (2004) also identified population stratification's effect on large scale genetic association studies through simulation. They showed that even small amounts of population admixture can undermine an association study and lead to false positive results. These adverse effects increase as the sample size grows large (Figure 3). For the size of study required for many complex diseases, relatively modest levels of structure within a population can have serious consequences. Population structure can also lead to missed real associations, so it cannot safely be ignored.

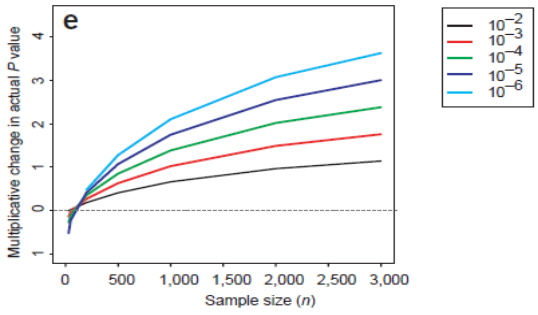


Figure 3. Type 1 error rate for different sample size (Marchini et al. 2004)

Researchers have noticed that population stratification can result in a great proportion of false positive association. As a result, grant applications and manuscripts are all required to show that population stratification issue is carefully taken care of. The concern for population stratification caused false positive has lead researchers to change the study design from association studies back to the inefficient family-based designs. However, as discussed above, these family based studies will cause difficulty in recruiting research subjects and, in general, be less powerful in identifying disease associated alleles than case-control studies (Cardon & Palmer, 2003). Therefore, there has been substantial work in trying to adjust for population stratification in case-control studies.

### 2.3.1 Genomic control

The key idea of genomic control is that, without stratification, the test statistics, $Y^2$ (test statistics of Armitage's trend test) is distributed with 1 degree of freedom under the null hypothesis: $Y^2\sim$ . In the presence of population structure, instead of following a distribution, $Y^2$ follows a $\lambda$ distribution. And the value of $\lambda$ is related to population structure in the sample. Genomic control methods use a series of L anonymous markers to estimate the inflation factor, $\lambda$. One drawback of this method is that the result is sensitive to L, the number of genomic control markers genotyped. Marchini et al. (2004) showed that when L is relatively small (<100), and the sample size is large, the correction is not effective. In some studies, especially those with large sample size, it probability not realistic to genotype as many markers as needed for genomic control method to work. In these cases, the genomic control method will not work. Also, when L is large (>500), the result is overcorrection and conservative which also lead to reduced power (Marchini et al., 2004). Another main issue of the genomic control method is that it assumes that each

allele contributes the same to the population structure. However, this is normally not the case. Some allele frequencies vary more across different populations than others. So the assumption that λ is uniform and constant is not an appropriate assumption for all studies.

### 2.3.2 Structure

This method assumed that the underlying subpopulations could be identified by the allele frequencies at each locus, and assigns each individual to different subpopulations based on the information told by their allele frequencies. Individuals can be assigned to more than one subpopulation if their genotype shows that they are admixed (Pritchard et al., 2000). However, when more than one subpopulation is assigned, this method will result in a huge computational cost, especially when analyzing genome-wide data (Price et al., 2006).

### 2.3.3 Principal component analysis

Price et al. developed the "Eigenstrat" method to address previous limitations. Their software first applied principal component analysis to genotype data to get the "components" of the data. Then, they adjusted both genotype and phenotype by the "component" underlying the data and computed the test statistics. Their results show that for random SNPs whose allele frequencies do not vary much between cases and controls, both genomic control and EIGENSTRAT can correct for inflation of type I error due to population stratification. But for highly differentiated SNPs, genomic control could not appropriately correct for this inflation, but EIGENSTRAT could. Moreover, for causal SNPs, genomic control loses nearly all power while EIGENSTRAT only suffers a partial power loss. (Price et al., 2006).

## 2.4 Statistical methods for analyzing rare variants

Since the power of statistical methods decreases as allele frequency goes down, researchers developed several statistical methods to increase the power of rare variants studies. The central idea of these methods is to collapse the rare variants in a region together into a composite variable, and then test the association between the composite variable and the disease. The disadvantages of the collapsing method are that it combines the functional and nonfunctional variants together and is not sensitive to misclassification. Li and Leal polished this method by developing a "Combined Multivariate and Collapsing" (CMC) method. It collapses the rare variants in a region to several composite variables according to whether the variant is functional or not, and then does a multiple marker test. Their simulation results show that the CMC method has high power to identified disease associated rare variants and is sensitive to misclassification (Li and Leal, 2008). Madsen and Browning improved this method by giving each collapsed group a weight, which is based on the standard deviation of the allele frequency (Madsen and Browning, 2009). Price *et al*. (2009) have provided a method for detecting association of multiple variants in protein-coding genes with a quantitative or dichotomous trait. The idea of their method is that different genes affect the disease at different allele frequencies: some associate with the disease when the allele frequency is high while others associate with the disease at a very low frequency. So instead of arbitrarily pooling these alleles together, their pooling method has a variable threshold for different genes.

# Chapter III Methodology

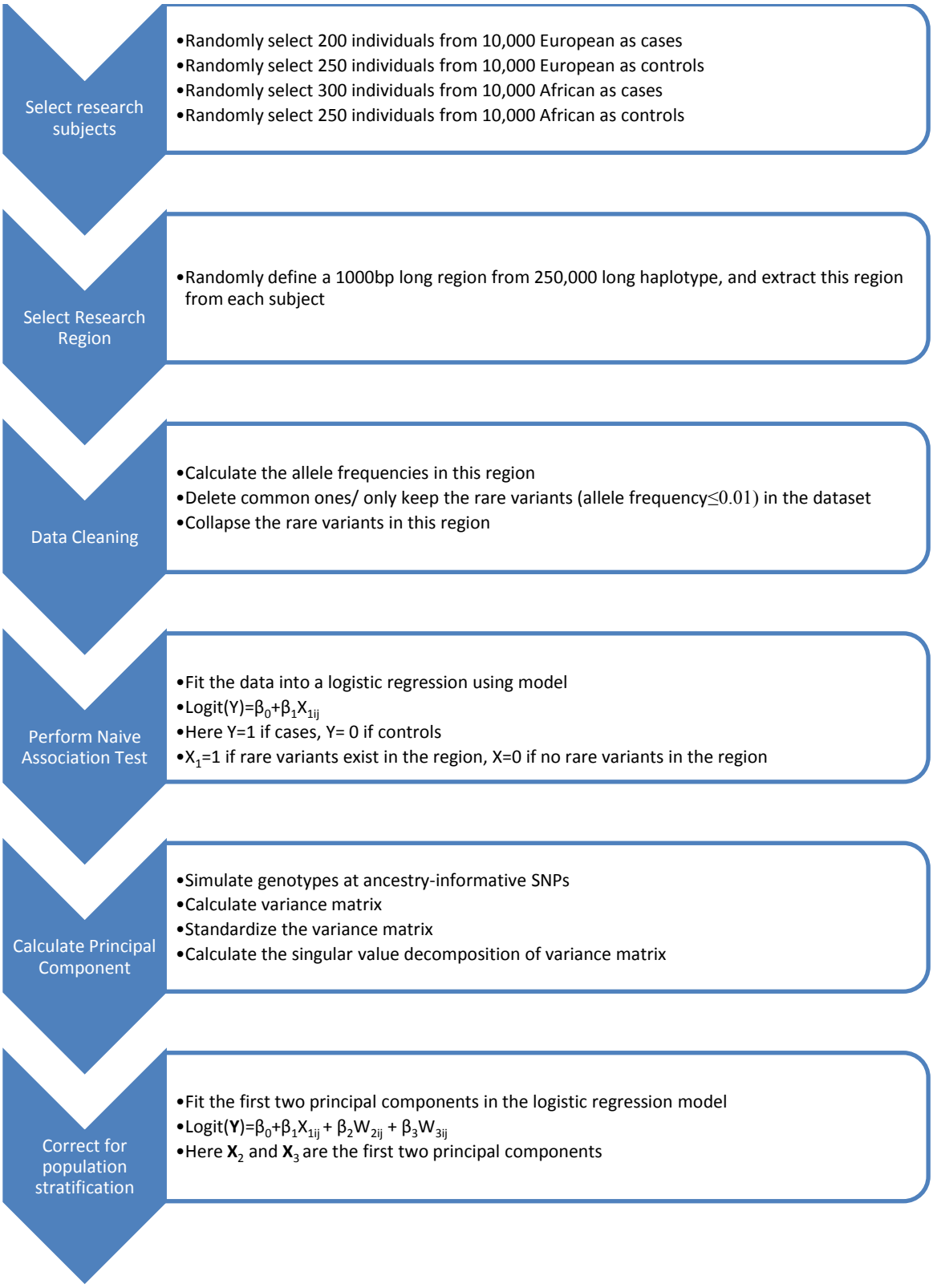


Figure 4. Simulation flow chart, using 500 Cases and 500 Controls, number of European vs. African=3:2 in cases as example

### 3.1 Simulating population specific haplotype

The haplotypes for samples from Europe and Africa were simulated through the "COSI" package developed by Schaffner et al. (2005). The package was the first one to simulate the haplotype data with high resemblance to an empirical population according to wide range of criteria. Schaffner's simulation package is based on Hudson's method of building gene genealogies through a coalescent process (Hudson, 2002). The rationale of his method is that the genotype sequences of individuals can provide information about how closely individuals in a population are related with each other in evolution. Also, by building up the genealogy through these sequences we can gain the information about their ancestry, such as when they have the most recent common ancestor (MRCA); or if they were from different populations, when their ancestor migrated from one to another. A genealogy is composed of points that represent individuals and lines between points across generations that represent their parent-offspring relationship. There are several parameters to consider when building up a genealogy, like the mutation rate and the migration rate, as discussed below. Schaffner et al. (2005) took advantage of publicly available SNP data (dbSNP and International Hapmap Project) to calibrate these parameters and then used these parameters to simulate data, to make the simulated data consistent with empirical data on a wide range of measurements. Here we present the basic idea of Hudson's method to build up the genealogy and how Schaffner et al. (2005) calibrate the parameters through the empirical data.

#### 3.1.1 Building the genealogy

Hudson's simulation is processed by first building up the genealogy; and then adding mutations and other population events on the genealogy. The simulation is built up by

tracing back in time. That, the sampled generation is labeled as generation 0; their parent generation is labeled as generation 1; and the population t generations back referred as generation t. The population where the n research subjects were sampled from has a size N, which is large and constant across the generations. Hudson also assumes N is generated by sampling N times with replacement from their parent generation. Since the simulation was built by tracing back in time, the first step to build a genealogy tree is to calculate how many distinct lineages are there in generation 1. This is equivalent to examining whether two individuals in generation 0 have the same parents or not in generation 1. As sample size n is relatively small compared to population size N, the probability that three or more individuals have the same parent is very small and will be ignored. Hudson defined the probability that two individuals have different parent as 1-1/N. In analogy, the probability that all n samples have distinct parents in generation 1 is

$$P(n)= \qquad - \qquad (1)$$

This is the probability that there are still n distinct lineages in generation 1. Hudson defines the event that two samples have the same parent in previous generation as "coalescence", as shown before, the probability that two of n samples coalesced in the previous generation is 1/N. Tracing back in time, we will finally find a point at which the coalescent process happens. Hudson defined the time until the first coalescent event happens distributed as

$$f(t)= \quad P(n)^{t}\,[1-P(n)] \quad — \quad ¯ \qquad (2)$$

We can tell from this equation that Hudson assumes the time when the first coalescence happens is exponentially distributed with mean N/ . After this coalescent event happens, there will be n-1 distinct lineages left. Following the same principle, the time when the second coalescent event happens is exponentially distributed with mean N/ . The process of building up the genealogy ends when there is only one genealogy left, or equivalently, we find the common ancestor of all of our samples. Hudson summarized the process of building a gene tree through the time T(i), the generation when i distinct lineages exists. Measured in the unit of N generations, the T(i) is exponentially distributed with mean expected value 1/ .

$$E[T(i)]=1/ \qquad (3)$$

### 3.1.2 Mutation

In Hudson's simulation, the number of mutations that differentiate the offspring from their parents follows the Poisson distribution. Under the constant rate mutation model, Hudson set the rate of mutation, μ, to be constant across the generations. With this property, Hudson calculated S, the number of mutations on the genealogy of the sample, through the total length of the genealogy,

$$S| T_{tot} \sim \text{Poisson}(\mu T_{tot}) \qquad (4)$$

Schaffner et al. set the mutation rate, μ, to be 1.5× per base pair per generation. Under this Poisson model, we derive the expected value of mutations on the genealogy E(S), as well as the variance Var(S), in terms of $T_{tot}$.

$$E(S)= \mu E(T_{tot}) \quad (5)$$

$$Var(S)=\mu E(T_{tot})+\mu^2 Var(T_{tot}) \quad (6)$$

So we can derive the distribution of S once we know the distribution of $T_{tot}$ which is simulated in the first step.

### 3.1.3 Adding neutral mutations to the genealogy

From equation 5, we can calculate the expectation of S from the expectation of $T_{tot,}$ the total length of the genealogy. $T_{tot}$, sum of the lengths of the branches of the genealogy, is equal to . Similarly, the expected value of T(i) can be calculated through equation 3. With these quantities, Hudson defined the expectation of S, measuring time in units of 2N generations as

$$E(S)= \quad = \theta \quad (7)$$

where $\theta=4N\mu$. N, the population size, is set to be 100,000 in the "COSI" package (Schaffner et al.). resulting in $\theta=6\times$ . And from equation 6 and 7, var(S) is also easily obtained

$$Var(S) = \theta \quad + \theta^2 \quad (8)$$

In order to learn the probability of getting each value of S, Hudson also defined a way to calculate the entire distribution of S. He started this simulation from the probability that S=0, the event that two sampled alleles are identical, denoted as E(F). The two alleles are identical if no mutations occurred since their most recent common ancestor. Hudson pointed out that one way to calculate this is to trace the genealogy of these two alleles

back, until either the most recent common ancestor (MRCA) event or a mutation event happens. In each generation, the probability of MRCA, denoted as , is 1/2N; and the probability of a mutation is 2μ. The probability that S=0 can be calculated as given that either MRCA or mutation happens, the probability that the first event is MRCA event, so E(F) can be expressed in following way,

$$E(F) \quad \frac{\quad}{\quad} = \frac{\quad}{\quad} \qquad (9)$$

The complement of F, the probability that the first event is a mutation event, is θ/(1+θ). The probability that j mutations happened since the most recent common ancestor is

$$P_2(j) = \left(\frac{\quad}{\quad}\right)^j \frac{\quad}{\quad} \qquad (10)$$

Expanding this; Hudson further derived the probability of having j mutations while n lineages exists as

$$Q_n(j) = \left(\frac{\quad}{\quad}\right)^j \frac{\quad}{\quad} \qquad (11)$$

In order to put the mutations on the entire genealogy, we also need the probability of having j mutations while n-1, n-2, n-3,……1 lineages exist. Hudson defined the probability as

$$P_n(j) = \qquad Q_n(i) \qquad (12)$$

The two parameters in this equation are θ and n. As noted before, Schaffner et al. (2005) set θ to be 6× , once we specified the sample size n we needed, we can get the entire distribution of S.

### 3.1.4 Migration

In Hudson's simulation, migration is also an important way to influence the genetics of the population. He supposed that in each generation, the population is composed of a fraction m from the other subpopulation, and 1-m from the same population. For our two sampled alleles, the probabilities of interest are $P_s(\theta)$, the probability that they are from the same population; and $P_d(\theta)$, the probability that they are from different subpopulations. Hudson calculated the two probabilities in the same fashion as he calculated the distribution of number of mutations. He expected that there are three possible events when we trace back in time: coalescence, mutation, and migration. If the first event is coalescence with probability——— (M is defined as 4Nm), the probability that the two identical alleles are from the same population is 1. If the first event is mutation with probability ———, the probability that they are from the same population is 0. If the first event is a migration, the probability that they are same is $P_d(\theta)$. In this way, the probability that the two identical alleles are from the same population is

$$P_s(\theta) = \text{———}*1 + \text{———}*0 + \text{———}*P_d(\theta) \qquad (13)$$

If two identical alleles come from different population, then the first event of these two alleles must be migration. The probability of migration is (M/n) /( θ+M/n), and the probability that they are identical is $P_s(\theta)$. As such, Hudson defined $P_d(\theta)$ as

$$P_d(\theta) = \frac{\_}{\_} P_s(\theta) \qquad (14)$$

Solving the previous two equations, we get

$$P_s(\theta) = \frac{\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad} \qquad (15)$$

$$P_d(\theta) = \frac{\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad} \qquad (16)$$

In the "COSI" package, m, is calibrated by the empirical data. In COSI's best fit model, the rate of migration from African to European is $3.2 \quad 10^{-5}$.

### 3.1.5 Recombination

In Hudson's simulation, the recombination rate is uniformly distributed and is fixed in the entire sequence. Schaffner et al. added an additional feature to this: the recombination rate can be varied through the region. Schaffner et al.'s simulation set a new recombination rate for every "window", whose size is pre-specified before simulation; this rate follows a gamma distribution. Within each window, there are "hotspots" of recombination. The recombination is more frequent around these "hotspots". The intensity of recombination and spacing of these hotspots are also gamma-distributed.

Schaffner et al. set the mutation rate equal to $1.5\times10^{-8}$ before calibration and maintained it as a constant. They set Europe split from Africa in 3500 generations ago. In 200 generation ago, Africa experienced an agricultural event which expanded the population size; and Europe agricultural event happened at 350 generations. The population size after the agricultural event is set to be constant and equal to 100,000 for both European and African. Schaffner et al. (2005) gave initial values to parameters such as migration

rate, spacing of recombination hotspots, shape parameters of hotspots, fraction of recombination in hotspots; and then calibrated these parameters based on the empirical dataset. In iteration, they used these parameters to simulate haplotype data; and then calculated the deviation of root-mean-square (RMS) between the simulated data and the empirical data. They set the threshold of the deviation to be no more than 1.5 times larger than pure sampling deviation of empirical data. The measurements being compared are 1) allele frequency distributions, 2) the probability of being an ancestor allele given allele frequency, 3) genetic distance 4) linkage disequilibrium and 5) the total distribution of heterozygosity.

When simulating haplotype data with COSI package (Schaffner et al., 2005), we use the parameters in their "best fit model". According to Schaffner et al. (2005), the simulated results have a high resemblance with empirical population. The root mean square error between the simulated value and the empirical value was 1.35, averaging over all 5 measurements as discussed above. This is the only package so far that can provide simulated sequence with high resemblance with empirical data from a wide range of criteria. We simulated 10,000 (n=10,000) haplotypes for European and African, assuming that the samples were drawn from a 100,000 (N=100,000) large population. In the best fit model, the migration rate from African to European is $3.2 \quad 10^{-5}$; the recombination hotspot spacing is 8500bp long with a spacing shape parameter equals to 0.35; and 88% of the region in hotspots experienced a recombination.

**3.2 Simulating a case-control study**

In this thesis, the exposures of interest are genotypes in the sequenced region and the outcome of interest is disease status. We simulate case-control studies with three different sample sizes:

100 cases/100 controls, 500 cases/500 controls, 1000 cases/1000 controls

We set the number of European versus African individuals to be 1:1 in controls and hold this constant for all studies. However, we set the proportion of European and African individuals varied in cases according to their disease prevalence. We simulate across cases in four different proportions:

50% European/50% African, 40% European/60% African, 25% European/75% African, 10% European/90% African

To summarize, we have 12 study designs in total as listed in the following table

| | | | |
|---|---|---|---|
| 100 Cases/100 Controls<br>European: African=1:1 in cases | 100 Cases/100 Controls<br>European: African=1:1.5 in cases | 100 Cases/100 Controls<br>European: African=1:3 in cases | 100 Cases/100 Controls<br>European: African=1:9 in cases |
| 500 Cases/500 Controls<br>European: African=1:1 in cases | 500 Cases/500 Controls<br>European: African=1:1.5 in cases | 500 Cases/500 Controls<br>European: African=1:3 in cases | 500 Cases/500 Controls<br>European: African=1:9 in cases |
| 1000 Cases/<br>1000 Controls<br>European: African=1:1 in cases | 1000 Cases/<br>1000 Controls<br>European: African=1:1.5 in cases | 1000 Cases/<br>1000 Controls<br>European: African=1:3 in cases | 1000 Cases/<br>1000 Controls<br>European: African=1:9 in cases |

Table 1. Study designs for simulating case-control study

For each study design, we randomly select haplotypes for case and control individuals from European and African population in the proportions described above and use two haplotypes to form a subject's genotype.

### 3.3 Simulating GWAS data

Principal component analysis normally needs several thousand SNPs to infer information about population stratification; the haplotype generated by COSI (Schaffner et al., 2005) is not long enough to fit the requirement. We use the publicly available dataset provided by HapMap project to simulate the genome wide data. The HapMap Project genotyped over 3.1 million markers and provide allele frequencies for each of them for both European and Yoruba population. We screened the markers to leave only those that provide information on ancestry. The prune is based on variance inflation factor (VIF) using the "Plink" package (Purcell et al., 2007). We set the VIF threshold to be 0.05, and 35,000 SNPs left after the prune. We generated genome wide data under a binomial model for each allele, with the probability of success equals to allele frequencies provided by Hapmap project.

### 3.4 Methods to Calculate Principal Components

$$Z=$$

Here we use matrix Z to represent genotype data of our study subjects, containing both cases and controls. Each row represents the genotype of the research subject m, and each column represents the $n_{th}$ genotype of all m subjects. $Z_{ij}$=0,1,2 according to the genotype simulated in the first step. Let V denotes the variance covariance matrix of Z.

In this thesis, we use principal components analysis to identify the underlying variability within the matrix Z. Principal component analysis is one of the oldest and popular methods used in multivariate analysis to identify the variation among the variables. Since the data are simulated under the scenario that these SNPs are not associated with the disease, the only difference between cases and controls in our simulations is their population structure (i.e. they are composed by different proportion of Europeans and Africans). In this case, the majority of the variation in Z is caused by population structure. Our goal is to investigate whether principal component analysis has enough power to identify the variation caused by population structure. In this case we can avoid the inflated type 1 error rate by including principal components in our model. The basic idea of principal component analysis is to summarize the variation of the data as a sequence of uncorrelated "components". These components are linear combinations of the variables in the original dataset. In our Z matrix, we define each of n column as $\mathbf{z_i}$, where i=1,2,,,,n. $\mathbf{z_i}$ contains the information of the $i_{th}$ SNPs for all m subjects. So the first component is calculated as

$$w_1=\mathbf{a_{11}z_1+a_{12}z_2+\ldots\ldots\ldots+a_{1n}z_n} \quad (17)$$

denoted as $w_1=\mathbf{a_1\ Z.}$ These components are ordered by their ability to summarize the variation of X. So, the first component, $w_1$, will account for as much as possible of the variation in the original data; the second component, $w_2$, will account for as much as possible of the remaining variation and similar fashion for other components. The calculation of $\mathbf{a_i}$s are under the restriction that

$$\mathbf{a_i'a_i=1}$$

and

$$\mathbf{a_i}' \mathbf{a_j}=0$$

where i  j, and i,j<n. The challenge is how to obtain $\mathbf{a_i}$. Note that the variance of $y_1$ can be expressed as

$$\mathrm{Var}(w_1)=\mathrm{Var}(\quad)=\quad Va_1$$

Then, $a_1$ is the engienvector corresponding to the largest eigenvalue of V, $a_2$ corresponds to the second largest eigenvalue of V, and so on. So the process of calculating principal components is equivalent to calculating the eigenvector of the variance covariance matrix V. In this thesis, we use singular value decomposition (SVD) to get the principal components. The singular value decomposition method decomposes the matrix V as

$$V= W$$

and the left singular vector W, is a m×m matrix containing the eigenvector of $VV^T$. So the columns of W are $a_i$s. Singular value decomposition method requires the matrix V to be standardized. So after we get the variance covariance matrix V, we first subtract the column mean from each column, and divide each column by its standard deviation. Now **V** is an m×n dimensional matrix, with each column mean equal to 0 and each column variance equal to 1. As there are only two populations in the dataset, we only need the first two components, i.e. the first two columns of W, denoted as $\mathbf{W_1}$,$\mathbf{W_2}$.

### 3.5 Testing for Association Between Rare Variants and Disease

The outcome of interest in this study is disease/non-disease, which is a binary variable. We use logistic regression to summarize characteristics of this response variable. Logistic regression is the most important model for categorical data and is widely used in genetics studies. We model the the probability of subject *j* having the disease $Y_j$ as a function of the genotypes in the sequenced region. We model

$$\text{Logit}(Y_j)=\text{Log}\left(\frac{\quad}{\quad}\right)=\beta_0+\beta_1 X_{1j} \quad (18)$$

where $X_{1j}$ is a composite variable that collapse the SNPs in this region (1000bp long) together, so that if there are rare variants (rare variants are defined as alleles with minor allele frequency smaller than 0.01) in this region, regardless of how many, then $X_{1j}=1$, otherwise $X_{1j}=0$.

When we include principal components to adjust for population stratification, we model the outcome as:

$$\text{Logit}(Y)=\text{Log}\left(\frac{\quad}{\quad}\right)=\beta_0+\beta_1 X_{1j}+\beta_2 W_{2j}+\beta_3 W_{3j} \quad (19)$$

Here, $W_2$ and $W_3$ are the first two principal components. In both models, we define X as a binary variable for genotype. Under models (18) and (19), if the p-value for $\beta_1$ is smaller than 0.05, then the variants are considered to be significantly associated with the disease. However, we generate the data based on the assumption that the variation doesn't cause the disease, a p-value smaller than 0.05 is actually a false positive. We do each simulation 5000 times, and count how many times we observe a false positive result.

# Chapter IV Results

As discussed above, with the large sample sizes needed for association studies, even a mild mixture of subpopulations can lead to a substantially inflated false positive rate. Here we present results to answer two questions regarding population stratification's effect on association studies involving rare variants: 1) whether or not population structure can lead to inflated false positive rate in studies of rare variants, and 2) whether principal components analysis, a widely used correction method for common variants, is powerful enough to correct for population structure in rare variants studies.

### 4.1 Simulated study

Our simulated case-control studies were carried out in two steps under the null hypothesis of no genetic effect on disease status. We first used the COSI package (see Methods) to generate the genotype data at 25,000 SNPs for 10,000 Europeans and 10,000 Africans. Then we randomly assigned the individuals in European and African population to case or control groups according to disease prevalence in each subpopulation. When simulating sequencing data with the "COSI" package, we used the parameters in their "best fit model". According to Schaffner et al., the simulated results highly resemble data from an empirical population. The root mean square error between the simulated value and the empirical value was 1.35, averaging over all measurements (see Methods for details). This is the only package so far that can provide simulated sequence with high similarity to empirical data based on a wide range of criteria.

### 4.2 False positive rate

We first simulated the study of 500 cases and 500 controls (second row of table 2), with 60 percent of cases and 50 percent of controls from the African population, and 40 percent of cases and 50 percent of controls from the European population. For each time of simulation, we randomly selected a region of 1000bp long and collapsed the rare variants in the region (see Methods for details) into a single composite exposure variable. Then we fit the exposure variable and outcome variable into a logistic regression model, and calculated how many times the correlation coefficient of the exposure variable had a p-value smaller than 0.05. Again, the study was simulated under the assumption of no genetic effects on the disease status, so any p-value smaller than 0.05 was considered as false positive. After 5000 simulations, the false positive rate was 0.0632 (95% Confidence Interval 0.0565, 0.0699). This result shows that population stratification can lead to inflated false positive rate in the studies of rare variants.

As a comparison; we also simulated a study free of stratification (50 percent of cases and 50 percent of controls drawn from the African population, and the same for the European population). The false positive rate was 0.047 (95% CI 0.0411, 0.0529), which means that correctly matching the population structure in cases and controls for a rare variants study can avoid inflated false positive rates. To study the stratification issue under a more extreme case, we then simulated the study with the ratio of Africans versus Europeans equal to 3:1 in cases and 1:1 in controls. The false positive rate rises up to 0.174 (95%CI 0.163, 0.185) in this case. Finally, the false positive rate was 0.352 (95%CI 0.339, 0.365) when we simulated 90 percent of cases from African and 10 percent of cases from European, but 50 percent from each population for controls. In these cases, population

stratification caused a highly inflated false positive rate that increased as the population structure became more extreme. At the sample size of 500 cases and 500 controls, even a small fraction of structure can have a large effect.

We also simulated the study with 100 caes/100 controls and 1000 cases/1000 controls. If we read Table 2 by column, it is clear that the false positive rate increases as sample size increases. When the sample size is small (100 cases/100 controls) the population stratification will not be a problem until the ratio of African: European cases reaches 3:1 (first row, third column of Table 1). But when the sample size is large, even a small fraction of population structure cannot be safely ignored (third row, second column of Table 2). This is because the false positive rate is actually the power to detect the real difference between cases and controls, which is population structure. It is reasonable that as the population size increases, or population structure becomes more different between cases and controls, the power increases.

In some situations, when recruiting the sample, researchers know the races of the subjects. We next performed simulations to see whether adding race as a covariate could adequately solve the inflated false positive rate problem. We defined a covariate $X_2$, and if the individual was from the African population, $X_2=1$; if the individual was from the European population then $X_2=0$. We simulated the study under the same sample size and disease prevalence as before, and the results of 5000 simulations are shown in Table 3. We can tell that by adding an indicator variable of race in the logistic regression model, we can adequately control for population structure and reduce the false positive rate to an

acceptable level. This suggests that, if researchers are able to record research subjects' race in practice, they can then use it as a covariate when calculating odds ratios with logistic regression models.

### 4.3 Correcting for population stratification using principal components

Because genome-wide data is necessary to control for population structure with principal component analysis, we simulated based on publicly available HapMap data to infer principal components. There are total 403,0562 SNPs in both European and African population in the HapMap dataset. We used PLINK to prune the data by deleting SNPs that are in high linkage disequilibrium with each other. We used the VIF (Variance Inflation Factor) as a threshold by setting the $r^2$ to be less than 0.5, and had 35,000 SNPs left after pruning. We use allele frequencies of these SNPs to generate genome wide data. We assume each SNP of our simulated data is binomial distributed with probability of success equals to allele frequency of that SNP provided by Hapmap. We then used the singular value decomposition method to get the principal components and included the first two components as covariates in the logistic regression model. We simulated the study under the same scenario as before; the results are shown in Table 4. We can see that in a study with 500 cases/500 controls, principal component analysis performs very well. It still behaves well even when the study has a large sample size and the population structure is extreme (right corner of Table 4). Although it is somewhat conservative in a smaller study with 100 cases/100 controls, the overall performance is very good and the false positive rate is around 0.05. This result suggests that principal component analysis is adequate to correct for population stratification in a rare variants association study.

| | 50%African vs. 50%European in cases | 60%African vs. 40%European in cases | 75%African vs. 25%European in cases | 90%African vs. 10% European in cases |
|---|---|---|---|---|
| 100case/100control | 0.0308 (0.0260,0.0356) | 0.0384 (0.0331,0.437) | 0.0694 (0.0624,0.764) | 0.150 (0.140,0.159) |
| 500case/500control | 0.0470 (0.0411,0.0529) | 0.0632 (0.0565,0.0699) | 0.174 (0.163, 0.185) | 0.352 (0.339, 0.365) |
| 1000case/1000control | 0.0612 (0.0546,0.0678) | 0.0920 (0.0839,0.100) | 0.242 (0.230,0.254) | 0.464 (0.450,0.478) |

Table 2. Type 1 Error Rate before Correction

| | 50%African vs. 50%European in cases | 60%African vs. 40%European in cases | 75%African vs. 25%European in cases | 90%African vs. 10% European in cases |
|---|---|---|---|---|
| 100case/100control | 0.0328 (0.0279,0.0377) | 0.0320 (0.0271,0.0369) | 0.0332 (0.0282,0.0382) | 0.0350 (0.0299,0.0401) |
| 500case/500control | 0.0500 (0.0439,0.0560) | 0.0556 (0.492,0.619) | 0.0506 (0.445,0.0567) | 0.0466 (0.0408,0.0524) |
| 1000case/1000control | 0.0518 (0.0457,0.0579) | 0.0508 (0.0447,0.0569) | 0.0440 (0.0383,0.0497) | 0.0480 (0.0421,0.0539) |

Table 3. Type 1 Error Rate Corrected by Self-Reported Race

| | 50%African vs. 50%European in cases | 60%African vs. 40%European in cases | 75%African vs. 25%European in cases | 90%African vs. 10% European in cases |
|---|---|---|---|---|
| 100case/100control | 0.0346 (0.0295,0.0397) | 0.0330 (0.0280,0.0379) | 0.0336 (0.0286,0.0386) | 0.0346 (0.0295,0.0397) |
| 500case/500control | 0.0490 (0.0430,0.0549) | 0.0470 (0.0411,0.0529) | 0.0539 (0.0477,0.0602) | 0.0544 (0.0481,0.0607) |
| 1000case/1000control | 0.0626 (0.0559,0.0693) | 0.0694 (0.0624,0.0764) | 0.0486 (0.0426,0.0546) | 0.0608 (0.0542,0.0674) |

Table 4. Type 1 Error Rate Corrected by Principal Components

# Chapter V Conclusions, Implications, and Recommendations

Human genome research, which aims to find the genetic etiology of the disease, is having a more and more profound influence on public health. The focus of recent genetic association studies is shifting from common variants to rare variants. Compared to common variants, rare variants can have much larger effect sizes, and can explain a larger percentage of heritability of the disease. Because of this, rare variant studies are meaningful for public health and can provide more valid background when making prevention strategies. As rare variants data have their own characteristics, statistical methods used for analyzing common variants may not be exactly suitable for rare variants. As such, developing statistical methods both powerful enough to identify real associations and conservative enough to avoid false positive are urgently needed for rare variants studies. Previous researchers have successfully developed statistical methods to meet the challenges of rare variants data. Although these methods improved the power, they all ignore a major cause of false positives---population stratification.

Population stratification is a widely known cause of false positives in common variants studies. If population structure is not adequately addressed, genetic variants identified by the study with population structure are very likely not genuinely associated with the disease. Prevention strategies, such as screening based on these "superficially" disease associated variants instead of real causal ones is a waste of both money and time. In this thesis, we examined the extent of inflated false positive rate under a variety of population-specific disease prevalences and sample sizes. We also applied the principal component method to see whether it can correct for population structure.

We simulated the study with different sample size and population structure according to a series of disease prevalence for each population, and found that population stratification can have a significant influence on rare variants studies. The false positive rate increases dramatically as sample size increase and population structure become extreme. When sample size reaches 1000 cases/1000 controls, even a small fraction of mixture will lead to highly inflated false positive rate. Current genome wide studies normally recruit several thousand subjects, which means researchers need to carefully match case and control ancestry, in order of avoid false positive caused by population structure.

For large scale studies, it may not be practical to recruit subjects from a single population, so we investigated correction for population structure in a rare variants association study. We applied principal component analysis to control for population structure. We inferred the first two principal components of the data, and added them as covariates in a logistic regression model. Our results showed that the principal component method performed very well even for highly structured data. The false positive rate remained around 0.05 in our simulation. These results suggest that researchers can use this method to correct for population structure in association studies involving collapsing of rare variants.

One minor disadvantage of the principal components method is that it needs several thousand SNPs to infer the components. This characteristic makes it unsuitable for studies that only genotype a few markers. Further methodological research should investigate methods that do not depend on the availability of a number of SNPs. As the development of sequencing technology makes detection of rare variants become more

accurate and cost effective, with the help of powerful statistical methods, we can expect rare variants will contribute more in understanding disease etiology.

# References

Altmuller, Janine et al. (2001). Genomewide scans of complex human disease: True linkage is hard to find. American Journal of Human Genetics, 69, 936-950.

Barnholtz-Sloan, Jill S. et al. (2005). Examining population stratification via individual ancestry estimates versus self-reported race. Cancer Epidemiology, Biomarkers Prevention, 14, 1545-1551.

Barroso, I. et al. (2003). Candidate Gene Association Study in Type 2 Diabetes Indicates a Role for Genes Involved in β-Cell Function as Well as Insulin Action. PLoS Biology, 1(1): e20.

Bodmer, Walter & Carolina Bonilla. (2008). Common and rare variants in multifactorial susceptibility to common disease. Nature Genetics, 40, 695-701.

Cardon, Lon R and Palmer, Lyle J. (2003). Population stratification and spurious allelic association. The Lancet, 361, 598-604.

Circulli, Elizabeth T. and Goldstein, David B. (2010). Uncovering the roles of rare variants in the common disease through whole-genome sequencing. Nature Reviews Genetics,11, 415-425.

Dickson, Samuel P. et al. (2010). Rare variants create synthetic genome-wide associations. PLOS Biology, 8, e1000294.

Hardy, John and Singleton Andrew. (2009). Genomewide association studies and human disease. The New England Journal of Medicine, 360, 1759-1768.

Hayden, Erick C. (2009). Diagnosing the future of genomics. Nature news, doi:10.1038/news.2009.1102.

Hirschhorn, Joel N. and Daly, Mark J. (2005). Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics, 6, 95-108.

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

Hunter et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics, 29, 870-874.

Kittles R. A. et al. (2002). CYP3A4-V and prostate cancer in African Americans: Causal or confounding association because of population stratification? Human Genetics, 110: 553–560

Li, Bingshan and Leal, Suzanne M. (2008). Methods for detecting associations with rare variants for common disease: application to analysis of sequence data. The American Journal of Human Genetics, 83, 311-321.

Madsen BE, Browning SR, (2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genetics 5(2):e1000384.

Maher B. (2008). Personal genomes: The case of missing heritability. Nature, 6:456(7218):18-21.

Manolio, Teri A. *et al*. (2009). Finding the missing heritability of complex disease. Nature, 461, 747-753.

Marchini, Jonathan., Cardon, Lon R., Phillips Michael S & Peter Donnelly. (2004). The effects of human population structure on large genetic association studies. Nature Genetics, 36, 512-517.

Mathias, Rasika A. (2009). Heritability of quantitative traits associated with type 2 diabetes mellitus in large multiplex families from South India. Metabolism Clinical and Experimental, 58 (2009) 1439-1445.

Need, AC et al. (2009). A Genome-Wide Investigation of SNPs and CNVs in Schizophrenia. PLoS Genetics (2):e1000373.

Price, Alkes L.et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38, 904-909.

Price, Alkes L. et al. (2009). Pooled association tests for rare variants in exon-resequencing studies. American Journal of Human Genetics, 86, 832-838.

Pritchard, Jonathan K. and Rosenberg, Noah A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics, 65, 220-228.

Pritchard, J.K. et al. (2000) Inference of population structure using multilocus genotype data. Genetics, 155, 945–959.

Pritchard, Jonathan K. (2001). Are rare variants responsible for susceptibility to complex disease? American Journal of Human Genetics, 69, 123-137.

Risch, Neil and Merikangas, Kathleen. (1996). The future of genetic studies of complex human disease. Science, 273, 1516-1517.

Schaffner, Stephen F. *et al.* (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Research, 15, 1576-1583.

Schork, Nicholas J. *et al* .(2009). Common vs. rare allele hypotheses for complex diseases. Science Direct, 19, 212-219.

Scott et al. (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science, 316, 1341-1345.

Stankiewicz, Pawel and Lupski, James R. (2010). Structural variation in the human genome and its role in disease. Annual review of medicine, 61: 437-455.

Weedon et al. (2008). Genome-wide association analysis identified 20 loci that influence adult height. Nature Genetics, 40, 573-583.