Principal Components Ancestry Adjustment

A Dissertation Presented

by

Jing Jin

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

August 2012

Stony Brook University

The Graduate School

Jing Jin

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Stephen Finch, Professor
Department of Applied Mathematics and Statistics

Nancy R. Mendell, Professor
Department of Applied Mathematics and Statistics

Wei Zhu, Professor
Department of Applied Mathematics and Statistics

Sun Jung Kang, Assistant Professor
Downstate Medical Center

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Principal Components Ancestry Adjustment

by

Jing Jin

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2012

A genome wide association study may have spurious or misleading results due to population stratification. This research evaluated the properties of global principal components and local principal components to adjust for population stratification. Principal components were calculated using both common variants (with minor allele frequency greater than 0.05) and rare variants (with minor allele frequency between 0.0005 and 0.05). One genetic model considered was from the Genetic Analysis Workshop 17 (GAW17). Additional genetic models developed in these analyses used the genotypes in the International Hapmap data. Phenotypes were simulated using these genotypes. Both type I error rates and powers of different models for identifying genetic variants associated with a phenotype were assessed. The four models in these analyses were: (1) using the number of minor alleles as the predictor variable for the phenotype; (2) using the number of minor alleles and 10 global principal components as the predictor variables for the phenotype; (3) using the number of minor alleles and 10 local principal components as the predictor variables for the phenotype; (4) using the number of minor alleles and the self-reported population of the participants as the predictor variables for the phenotype. Both the global PC adjustment model and local PC adjustment model had null hypothesis rejection rate roughly equal to the nominal significance level and comparable power to detect the causal genes. Both had better rejection rates than the model using the self-reported population indicators.

Contents

List of Figures

List of Tables

List of Abbreviations

SNP-Single Nucleotide Polymorphism

GWAS-Genome Wide Association Study

GAW-Genetic Analysis Workshop

PS-Population Stratification

LTT- Linear Trend Test

MCMC- Markov Chain Monte Carlo

LD-Linkage Disequilibrium

MAF-Minor Allele Frequency

PC-Principal Components

GPC-Global Principal Components

LPC-Local Principal Components

CEU-Utah residents with Northern and Western European ancestry from the CEPH collection

CHD-Chinese in Metropolitan Denver, Colorado

YRI-Yoruba in Ibadan, Nigeria

CHB-Han Chinese in Beijing, China

JPT-Japanese in Tokyo, Japan

LWK-Luhya in Webuye, Kenya

TSI - Tuscans in Italy(TSI)

Acknowledgments

GAW grant, R01 GM031575

Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org).

Chapter1 General Introduction

1-1 Research Background

The human genome contains millions of single-nucleotide polymorphisms (SNPs) which is a DNA sequence variation occurring when a single nucleotide — A, T, C or G — in the genome (or other shared sequence) differs among members of a biological species or paired chromosomes in an individual. These SNPs may either directly influence individual phenotype variations or indirectly cause changes by affecting nearby mutations that are associated with phenotypes, including disease. Many diseases have been identified as associated with genetic variation, such as cancer, autism, and schizophrenia. Thus, as more information about human genome is found, greater understanding of human disease can be developed. More importantly, researchers might be able to predict, control and prevent genetically related diseases.

However, genome databases are usually extremely large which makes the study of the human genome difficult. The Human Genome Project (launched in 2003) [1] and the International HapMap Project (launched in 2005) [2] contain research useful for deciphering the human genome. Scientists have developed a set of research tools that can be used to manipulate the gigabytes of data contained in genome database. These tools include computerized databases that contain the reference human genome sequence, a map of human genetic variation, and a set of new technologies that can quickly and accurately analyze whole-genome samples for genetic variations that contribute to the onset of a disease. [3] Using these modern tools, researchers all over the world are developing methods and algorithms to accelerate the speed and increase the accuracy of detecting specific polymorphisms whose variation is truly associated with a disease.

1-2 Genome-Wide Studies

1-2-1 Genome-Wide Association Studies

Genome-wide association studies (GWAS) are used to identify common genetic factors that influence health and disease [4]. The main task of GWAS includes: scanning a large number of markers across the complete set of DNA or genomes of large samples of affected and unaffected participants and searching for genetic variations associated with a specified disease. With the results of a GWAS as a foundation, researchers can develop better strategies to detect the presence of disease based on the genetic associations identified and to treat or prevent the disease. These studies are particularly useful in finding genetic loci that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses [3]. GWAS has discovered associations with as many as 40 common diseases since 2005[5]. These findings may have a substantial impact on medical care.

The typical GWAS approach includes four steps: (1) select a large number of participants with or without the disease or trait of interest; (2) get DNA from each participant, genotype the DNA with high genotyping quality; (3) perform statistical analyses to detect associations between the SNPs and the disease or trait of interest; and (4) repeat the same approach on an independent sample to confirm the results [5]. Among all the study designs, the most frequently one used in GWAS is the case-control design. Minor allele frequency is defined as the lowest allele frequency at a locus that is observed within a population. In case-control studies, genotype frequencies in the case group are compared to the frequencies in the control group for each gene with genotype data. Statistical techniques then evaluate whether any set of differences is large enough to merit further study.

1-2-2 Genetic Analysis Workshop

The Genetic Analysis Workshops (GAWs) are collaborative workshops among researchers from all over the world. Each GAW focuses on one of the hottest topic in genetic epidemiology and distributes a set of real or computer-simulated data to all registered researchers. Researchers apply different statistical approaches to the data set. Results of analyses are discussed and compared at conferences held in even-numbered years. [6]

1-2-3 Limitations of Recent Genome-Wide Studies

GWAS studies have three main challenges: obtaining adequate data, controlling the cost of sequencing or genotyping the DNA, and analyzing the statistical data to obtain correct conclusions. In principle, GWAS studies require a large number of participants to achieve statistical significance. However, the reality is that the number of cases with a disease may be relatively small, especially for a rare disease (i.e, one with low prevalence, such as schizophrenia). Although hospitals all over the world are collaborating more closely, it is still hard to collect enough cases to develop an effective GWAS. For some extreme scenarios, the number of cases available may be less than five. (i.e, for the case I studied about copy number variation in Mount Sinai School of Medicine, only three participants from one family were available) Thus, better statistical methods are needed. Additionally, the expense for sequencing the genome of a participant or processing a modern SNP platform is high. Improved technology and procedures can reduce the genotyping or sequencing error, thus improving data quality. However, even when a high quality data set with large sample size is available, there still can be a high false positive and false negative result rate in GWAS.

One possible cause for misleading results is population structure in the genotype data. Most genetic variations are associated with the geographical and historical populations in which the mutations first arose. Because of this, studies must take account of the geographical and historical background of participants—controlling for what is called population stratification. As the people of the world have migrated and inter-married over many generations, these geographical variations also have mixed over time and have become more complex. As a result, it is more challenging to interpreting the results of testing for association between candidate SNPs and disease.

1-3 Introduction to Population Stratification

1-3-1 Causes of Population Stratification (PS)

The basic cause of population stratification is non-random mating between groups, often due to their physical separation (e.g., for populations of African and European descent) followed by genetic drift of allele frequencies in each group. In some regions (e.g., in Europe), individuals in a population can be divided into mutually exclusive subpopulations by location. That is, the minor allele frequencies vary with location. In the modern world, however, population admixture is a realistic scenario. Admixture results in a new mixed population (as in children with an African American and Caucasian parents). PS is mainly due to the demographic history of a population, natural selection and random fluctuations resulting from admixture. From a statistical point of view, PS is the result of systematic ancestry differences in allele frequencies between cases and controls.

1-3-2 Impact on Association Studies

GWAS studies have identified hundreds of common variants associated with disease risk or related traits [7]. One limitation of GWAS studies is that PS can be a source of confounding. That is, the association found by some statistical models might be due to the underlying population structure. This kind of association is not a real association between disease and SNPs but a misleading signal caused by the mixture of genotype data from different populations. Meanwhile, some real associations between the genotyped SNPs and disease may not be found if the associations with genotype frequency have opposite signs in the different populations so that the test statistic would lead to a non-significant result when combining the results from multiple populations. Thus, if not corrected, PS may cause false-positive and/or false-negative findings [8] and produce spurious associations [9].

1-4 Methods to adjust for Population Stratification

1-4-1 Genomic Control

Devlin and Roeder[10] developed the genomic control method in 1999. Theoretically, this method extends the linear trend test (LTT) which has inflated Type I error rate when PS occurs. It corrects the test statistic by estimating an inflation factor using a Bayesian approach which can be appropriate in some cases dealing with a large number of candidate genes using the linear trend test. A second test that genetic researchers use is the chi-square test for allelic frequencies. These two statistics should be approximately equal if the sample is in Hardy-Weinberg equilibrium. Under this assumption, the allelic test statistic approximately follows a chi-square distribution with one degree of freedom. A population inflation factor is estimated by comparing the two test statistics. The null hypothesis is that there is no population stratification. When the null distribution of the LTT is inflated by a constant factor λ due to population stratification, the test statistics are multiplied by λ. If the value of λ is approximately 1, there is apparently no population stratification in the data. A value of λ greater than 1 indicates there exists population stratification or other confounders, such as family structure in the data. [10] Devlin and Roeder considered $\alpha = 0.05$ and did not study the properties of genomic control at genome wide level of significance (10^{-8}).

Besides the estimator mentioned above, scientists have proposed different estimator of λ. For example, Reich and Goldstein [11] suggested using the mean of the statistics instead of the median in the test statistic. With this correction the overall type I error rate should be approximately equal to α (nominal level) even when the population is stratified.

1-4-2 Structured Association

When PS is the only confounding issue, methods inferring genetic ancestry often lead to an effective adjustment. To do this, accurate separation of the samples into sub-populations is essential. The definition of a population may include aspects such as linguistic, cultural or physical characteristics and the geographic location of the population. However, it may be inaccurate to assign individuals to different subpopulation only using these factors. Genetic information should be considered as well.

Structured association was designed to assign each individual to a subpopulation. Pritchard and Rosenberg [12] considered the use of genetic information to detect population structure in 1999. Their method successfully tested association with good adjustment for PS using unlinked genetic markers.

Later, Pritchard and Stephens (2000) [13-14] suggested a Bayesian clustering approach based method to assign participants into different sub-populations. They assumed there were K populations (K may be unknown) in the sample and that each population was characterized by a set of allele frequencies at each locus. The Markov Chain Monte Carlo (MCMC) algorithm was used to calculate parameters involved. They also developed software called STRUCTURE [15-18] to perform the calculations.

In the same year, Pritchard and his collaborators [15] also developed a method for case-control association studies in structured populations. They used unlinked genetic markers. Individuals were assigned to unstructured subpopulations which did not have association between these unlinked markers. They used a two-stage procedure. The first step was to assign each individual to an unstructured subpopulation. The second step was to test association within subpopulations. They argued that there should not be association due to population structure in these subpopulations. Program STRAT [15] was developed to execute this method. However, since the computational cost was relatively high, the applicability of this approach to large genome-wide was limited.

Alexander (2009) [19] developed a method that used maximum likelihood estimates of underlying admixture coefficients and ancestral allele frequencies rather than the Bayesian estimates and created software called ADMIXTURE [19] to do association tests.

1-4-3 Principal Components Analysis

Principal components Analysis (PCA) was first applied to genetic data by Cavalli-Sforza[20] and colleagues. They published a paper discussing using principal component to identify population structure. Patterson [21] used global PCs calculated using all SNPs with a minor allele frequency greater than 0.05 across the whole genome in his paper. These PCs provide each participant's coordinates along axes of variation rather than classifying all participants into discrete population or linear combinations of populations. Novembre and Stephen [25] also published a paper about applying PCA on spatial genetic data based on Cavalli-Sforza et al.'s genetic maps.   Global PCs adjustment [20-25] is widely used to adjust for ancestry when dealing with admixture data. Researchers choose a relatively small number of global PCs that contain the most information about genetic ancestry, usually according to the linkage disequilibrium (LD) pattern of the population of interest. Linkage disequilibrium was one kind of association between markers that two nearby markers tend to be together on the same gamete with the disease allele. Therefore, if a marker was in LD with a disease marker, and association between this marker and disease was found, this particular marker can be used to test for association.

The general approach to calculate Principal Components using genetic data is as follows. Let $g_{ij\epsilon\{0,1,2\}}$ be a matrix of genotypes for SNP *i* and individual *j*, where *i*=1 to *m* and *j*=1 to *n*; g_{ij} is the minor allele count at SNP *i*. Each row represents an individual, and each column represents a SNP position. The mean of each SNP is:

$$\mu_{ij} = \frac{\left(\sum_j g_{ij}\right)}{n}$$

Normalize each entry by: $\frac{g_{ij}-\mu_{ij}}{\sqrt{p_i(1-p_i)}}$

where $p_i$is an estimate of the allele frequency of SNP *i* defined as:

$$p_i = \frac{\mu_{ij}}{2}$$

Missing entries are usually excluded when calculating the mean for each marker and set to be 0 when calculating PCs.[21] There are some other different but similar ways to normalize the sample matrix. For instant, Price [22] used $\frac{(1+\sum_j g_{ij})}{(2+2n)}$ as p_i.

Sankararaman and Sridhar[26] introduced local principal component ancestry adjustment, which was to calculate PCs using SNPs within a region on the genome. They proved that the local method was significantly more accurate and more efficient than existing methods for inferring locus-specific ancestries, enabling it to handle large-scale datasets. Kang and Larkin[27] used regression models with both global principal components and local principal components to adjust for ancestry and successfully reduced type I error for Framingham Heart Study. They suggested that local ancestry adjustment was especially useful for the scenarios where the ancestral populations in a region of genome are significantly different from the rest of the genome.

Chapter 2: Method

2-1 Global Principal Components Analysis

In previous studies using global ancestry adjustment, PCs were calculated using all SNPs with minor allele frequency (MAF) greater than 0.05, and ancestry was represented by the first few PCs. Minor allele frequency was defined as the percentage of the less common allele of the two alleles in the same loci. Most studies plotted the first two PCs to show the clusters of population. Other studies used the first few PC scores as covariates in regression models to adjust for population stratification. [26-28]

2-2 Local Principal Components Analysis

In previous studies using local ancestry adjustment, the genome was cut into several regions based on linkage disequilibrium (LD) pattern; e.g. 20MB length. Local PCs were calculated using all the SNPs with MAF>0.05 within these regions. Ancestry for each region is represented by the first few local PCs. These local PCs were used as covariates in regression models to adjust for PS within this region.

Sankararaman[26] introduced a method and developed a software program called Local Ancestry in adMixed Population (LAMP), which inferred the ancestry of each individual at every SNP. LAMP used overlapping windows of contiguous SNPs and used a majority vote to decide ancestry. They used LAMP to test a real dataset from the HapMap project. They used the SNPs of chromosome 1 from the 500K Affymetrix GeneChip assay from each of the four HapMap populations: Yorubans from Ibadan, Nigeria (YRI), Japanese from the Tokyo area

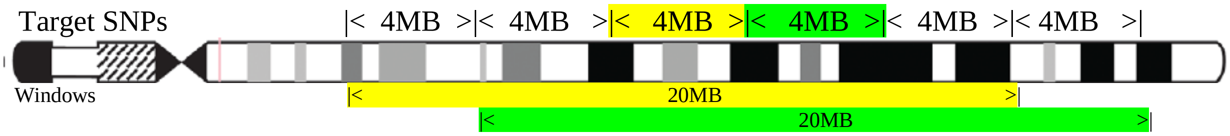
(JPT), Han Chinese from Beijing (CHB), and Utah residents with European ancestry (CEU). Sankararaman proved that LAMP was more accurate compared to other procedures. Sankararaman also suggested a way to calculate the length of the window. Breakpoint refers to a recombination event that results in a change in ancestry of the adjacent SNPs. In order for the local predictions to achieve reasonable accuracy, the length of the window should be short enough so that most individuals do not have a breakpoint in the window and long enough so that the SNPs provide sufficient information for the observation of a difference between the populations. He suggested a maximum window length is based on the breakpoints in the window.

Kang[27] et al. applied local principal component analysis to the 500 k single-nucleotide polymorphism data from the Framingham Offspring Cohort of GAW16 data. They selected unrelated adults from each family (i.e., spouses) based on an algorithm that prioritized individuals with higher genotyping rates, selecting individuals at random when needed. They regressed height on age and age squared across all visits separately for each gender, with the resultant average standardized residuals used as the outcome of interest. They excluded SNPs that failed the filtering criteria on completeness of each SNP and minor allele frequency. They calculated the first 10 local PCs to represents local ancestry. Chromosomes were divided into non-overlapping windows with a length of 20MB. Local PCs were calculated for each window and used as covariates in the regression model. They found that adjusting for local PC may control the false-positive rate due to the population stratification underlying the Framingham Heart Studies sample.

Local ancestry adjustment, however, had a major problem. For SNPs near the boundary, local PCs are calculated only using SNPs on its left or its right which may lead to a loss of information and might lead to spurious results on the boundaries of each region.

Qin et al. [28] modified this procedure to deal with this problem. Chromosomes were divided into overlapping windows with a length of 20MB as before. Local PCs were calculated using all the common SNPs in each window. However, the first 10 PCs are only used to represent the ancestry for the SNPs in the middle 4 MB as shown in figure 1 below. This sliding window approach would best ensure that there were enough SNPs surrounding the target SNP. The sliding window approach yielded better adjustment for SNPs near the boundary of each window in previous studies.

Figure:1 An example of a local window.



Each color represents a window with its target SNPs. Color above the chromosome indicates the region for target SNPs, and the color below the chromosome indicates the region for SNPs which are used to calculated local PCs for its target SNPs. For example, all the common SNPs in the yellow window below the chromosomes are used to calculate local PCs to adjust for the SNPs under the yellow region above the chromosome. All the common SNPs in the green window below are used to calculate local PCs to adjust for the SNPs under the green region above.

Using this sliding window approach, each SNP was covered with the complete information in the local region around it except for the very beginning and ending SNPs in a chromosome. This approach successfully reduced most of the spurious association due to the insufficient of ancestry information for SNPs near boundary.

My study used these ideas with some modifications.

CHAPTER 3: METHODOLOGY

3-1 Part 1: Methods using Global PC adjustment

3-1-1 Data

The GAW17 (2010) genetic data was mini-exome data using 697 subjects from the 1000 Genomes Project [29]. The Exome was the coding region on the DNA sequences. The data provided 200 replicates of simulated phenotypes. It included a dichotomized disease status and quantitative risk factors named Q1, Q2, Q4, and smoking status. The disease status (affected/unaffected) and the values of Q1, Q2, and Q4 were generated by the SNP variants. The disease status was simulated using a liability threshold model that was a function of Q1, Q2, and Q4[30] . The top 30% of the liability values was declared affected and reported in the GAW17 data. Sex and age was also provided in the data. Sex was taken from the 1000 genomes project. Age was simulated in the family data set, the other data set provided by GAW17. The quantitative phenotypes Q1 and Q2 were generated as normally distributed phenotypes in the 200 replicates. I document the level of significance of the test of the coefficient of a genotype with and without population stratification adjustment in selected genes known not to be associated with the phenotypes Q1 and Q2.

The values of trait Q1 were simulated to be associated with the minor alleles of 39 SNPs in 9 genes. These genes were on Chromosomes 1, 4, 5, 6, 14 and 19. The values of trait Q2 were simulated to be associated with the minor alleles of 72 SNPs in 13 genes. These genes were on Chromosome 2, 3, 6, 7, 8, 9, 10, 11, 12 and 17. A full list of the SNPs that were associated with Q1 and Q2 can be found in appendices. I also studied selected non-associated SNPs to estimate null distribution properties. SNPs on chromosomes 12, 21, and 22 were used as SNPs not associated with Q1. SNPs on chromosomes 21 and 22 were used as SNPs not associated with Q2. Table 1 lists the distribution of minor allele frequencies of SNPs studied.

Table 1: Distribution of Minor Allele Frequencies of SNPs in the Genes studied in GAW17 Data.

Genes	SNPs	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
		Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2
Associated Genes	Non-associated	60	100	5	9	12	15	9	15	86	139
	Associated	32	61	0	5	5	4	2	2	39	72
Non-associated Genes	Non-associated	1422	532	189	57	295	90	288	80	2194	759

Note: *The Q1 associated genes were ARNT, ELAVL4, FLT1, FLT4, HIF1A, HIF3A, KDR, VEGFA and VEGFC. The Q2 associated genes were BCHE, GCKR, INSIG1, LPL, PDGFD, PLAT, RARB, SIRT1, SREBF1, VLDLR, VNN1, VNN3 and VWF. A non-associated gene was any gene on chromosome 21 , 22 and any other gene on chromosome 12 that were known to be not associated with Q1 and Q2.*

3-1-2 Modeling

I dichotomized the quantitative measures Q1 and Q2 so that the top 25% of each replicate was scored as affected (1) and others as unaffected (0). The dichotomized measures and quantitative measures were dependent variables in the analyses.

The independent variables in these analyses were selected from the number of minor alleles in the *i*th SNP genotype (SNP_i), the participant's age (*AGE*) and smoking status (*SMOKING*), and the ten ancestry adjustment principal component scores (called *GPC1, ..., GPC10*). I used the FamCC software [31] to calculate these ten PCs.

I used the PLINK software [32] to fit two logistic regression models to assess the association between each SNP in the genes studied and the dichotomized phenotype. The *i*th SNP was considered associated with the phenotype when the permutation p-value of the coefficient of SNP_i reported in the PLINK logistic regression analysis was less than 0.05. For Q1, since it was affected by age and smoking, the models considered were:

No adjustment:

$$\beta_0 + \beta_1 SNP_i + \beta_2 AGE + \beta_3 SMOKING$$

Global PC adjustment (SNP adjusted for age, smoking, and ancestry adjustment PCs):

$$\beta_0 + \beta_1 SNP_i + \beta_2 AGE + \beta_3 SMOKING + \beta_4 GPC_1 + \cdots + \beta_{13} GPC_{10}$$

Since Q2 was not associated with either age or smoking, age and smoking were not used in the Global PC adjustment model of Q2. I also fit the models above to the continuous phenotypes Q1 and Q2 using PLINK. Each model was fit to the 200 replicates provided.

Similar approach of Global PC adjustment was applied to the International HapMap III data for further comparison. The comparison results can be found in results part, section 4-2.

I also fit the model to both dichotomized and quantitative phenotype with population reported by participants as covariates for further comparison. The comparison results can be found in discussion part of section 5-1.

3-2 Part 2: Methods using Local PC adjustment

3-2-1 Data

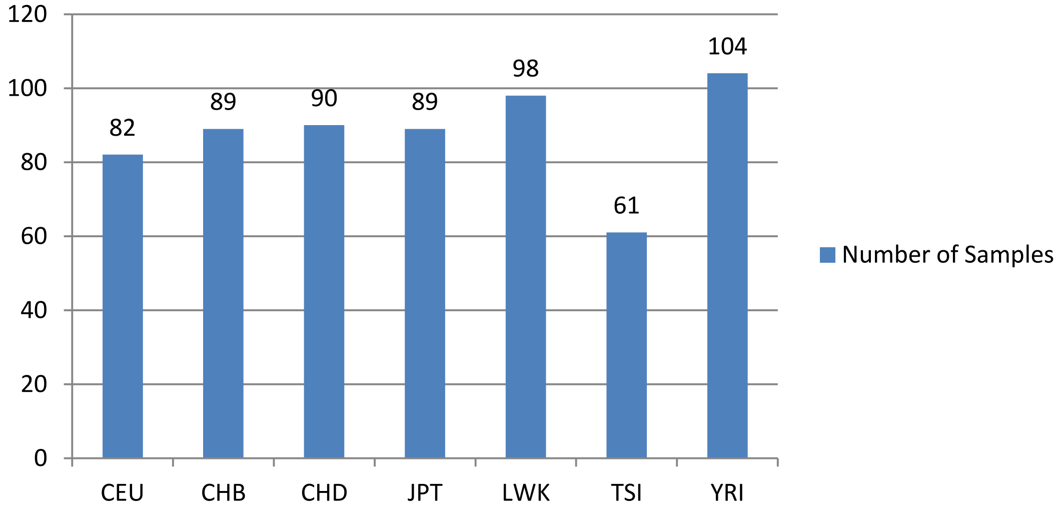
The HapMap III genotype data was downloaded from its website [33]. This SNP genotype data was generated from 1,397 participants using two platforms: the Illumina Human1M (by the Wellcome Trust Sanger Institute) and the Affymetrix SNP 6.0 (by the Broad Institute). The data from these two platforms were merged to create the HapMap III genotype data. The HapMap III genotype data had information on 613 participants from the 697 used in the GAW 17 data. The HapMap III data also had genotypes for a larger number of SNPs so that using local adjustment techniques were possible for the subset of 613. The genotyping rates of HapMap data were greater than 0.99. Table 2 contains summary information for the SNPs included in this research.

Table 2 Distribution of SNPs in HapMap III Genotype Data

	Overall	MAF>=0.05	MAF<0.05
Number of SNPs	1,457,897	1,256,096	201,801

The participants in this study were from seven populations. These were Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Chinese in Metropolitan Denver, Colorado (CHD), Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT), Luhya in Webuye, Kenya (LWK), and Tuscans in Italy(TSI). Figure 2 shows the distribution of participants.

Figure 2 Distribution of Overlapping Samples in HapMap Project and GAW 17 Project (n=613)



3-2-2 Modeling

Since only genotype data was available in the HapMap database, I had to generate disease models using the genotype data. I created both a rare causal SNP disease model and a common causal SNP disease model.

a. Choose Disease SNPs

In order to reduce the size of the database for this research, I focused on five chromosomes that supported local adjustment. I chose chromosome 12 since many GAW 17 conference participants found high false positive rates for SNPs on this chromosome. After I examined my GAW 17 analyses, I chose chromosomes 1, 6, 9, and 14 because the permutation results of my regression models with global PC as covariates had low type I error rates and relatively high power for some of the causal genes on these four chromosomes. Table 3 shows the distribution of MAFs on the 5 chromosomes chosen.

Table 3 Information of 5 chromosomes chosen to provide disease SNPs using HapMap III genotype data.

Chromosome	Gene in GAW17	Number of SNP with MAF<0.05	Number of SNP with MAF>0.05	Total number of SNPs
1	ARNT, ELAVL4	8,626	110,861	119,487
6	VEGFA	6,302	87,369	93,671
14	HIF1A	2,727	43,928	46,655
19	HIF3A	2,154	24,799	26,953
12	Null genes	5,155	65,327	70,482
Total		24,964	332,284	357,248

For each chromosome, I chose 200 rare SNPs (i.e., those with MAF less than 0.05) and 200 common SNPs (i.e., those with MAF greater than 0.05) so that I had a total of 1000 rare SNPs and 1000 common SNPs. I studied two sets of disease SNPs, common disease SNPs and rare disease SNPs. I chose 10 common disease SNPs and 10 rare disease SNPs from these 2000

SNPs. In order to test the effects of PC adjustment, population stratification should be an issue in the disease mechanism. My ideal 20 candidate disease SNPs should have the following two properties:

1. There was no interaction effect (SNP*i**SNP*j* or so) among multiple disease SNPs. That is, the disease SNPs were relatively independent with each other. Thus, when calculating the correlation based on the genotype matrix of the 10 SNPs (two sets, 10 common disease SNPs and 10 rare disease SNPs), the value should be low.
2. There was population stratification in the data so that population stratification was a potential source for false-positive and false-negative results, and ancestry adjustment was required when performing an association test. Thus, when calculating the correlation of any two of the 10 disease SNPs (two sets, 10 common disease SNPs and 10 rare disease SNPs) based on population MAF, the value should be high. Besides, the 10 disease SNPs (two sets, 10 common disease SNPs and 10 rare disease SNPs) should be correlated with some other non causal markers so that these correlated non causal markers might have significant results when performing association tests.

Table 4 lists the 10 rare SNPs that I selected to be causal with their overall MAF and MAFs by population. Table 5 lists the 10 common SNPs that were selected. They come from five different chromosomes. The distribution of MAF by population is quite different for different populations. Clearly population stratifications exist among the data. That is, for each selected SNP, MAF is high in some population while quite low, even zero, in others.

Table 4 Minor Allele Frequencies by Population for the Ten Rare SNPs chosen to be Disease SNPs using HapMap III Genotype Data.

SNPs	Chromosome	Minor Allele Frequencies							
		Overall	CEU	CHD	YRI	CHB	JPT	LWK	TSI
rs7553321	1	0.011	0	0	0.053	0	0	0.015	0
rs17035549	1	0.046	0.043	0.022	0.077	0.034	0.022	0.087	0.025
rs6693629	1	0.032	0	0.028	0.058	0.028	0.022	0.061	0.008
rs11811671	1	0.010	0	0	0.038	0	0	0.020	0
rs4512698	1	0.040	0.128	0.006	0.048	0	0	0.031	0.090
rs12401645	1	0.015	0.037	0.011	0.029	0	0.011	0.010	0
rs17358725	1	0.007	0.012	0.022	0	0.006	0	0	0.008
rs12215941	6	0.032	0.043	0.028	0	0.011	0.096	0.005	0.057
rs17044864	12	0.016	0.024	0	0.053	0	0	0.005	0.033
rs28651257	19	0.048	0	0	0.154	0	0	0.138	0

Table 5 Minor Allele Frequencies by Population for the Ten Common SNPs chosen to be Disease SNPs using HapMap III Genotype Data.

SNPs	Chromosome	Minor Allele Frequencies							
		Overall	CEU	CHD	YRI	CHB	JPT	LWK	TSI
rs1888991	1	0.235	0.073	0	0.726	0	0.006	0.571	0.090
rs1390333	1	0.223	0.364	0.178	0.236	0.146	0.169	0.189	0.328
rs10498665	6	0.299	0.402	0.478	0.043	0.433	0.444	0.010	0.385
rs9376923	6	0.201	0.329	0.194	0.139	0.208	0.191	0.077	0.353
rs7296827	12	0.393	0.610	0.261	0.375	0.332	0.348	0.362	0.533
rs7300747	12	0.223	0.427	0.017	0.351	0.017	0.006	0.383	0.393
rs17100963	14	0.259	0.055	0.356	0.240	0.343	0.348	0.280	0.131
rs3759670	14	0.314	0.177	0.264	0.456	0.278	0.172	0.510	0.271
rs6509333	19	0.265	0.512	0.211	0.216	0.185	0.202	0.153	0.484
rs17207579	19	0.179	0.366	0.122	0.178	0.112	0.073	0.158	0.303

Table 6 lists the MAF and MAFs by population for a random set of 10 general SNPs which do not have a population stratification effect and therefore not a good set for causal SNPs. For both rare and common SNPs in this table, MAF is relatively uniformly distributed across the seven populations. PS was less likely a major problem for these 10 randomly selected SNPs.

Table 6 MAF and MAF by Population for Ten Randomly Selected SNPs using HapMap III Genotype Data

SNPs	Chromosome	Minor Allele Frequencies							
		Overall	CEU	CHD	YRI	CHB	JPT	LWK	TSI
rs12749138	1	0.312	0.323	0.315	0.327	0.309	0.326	0.217	0.402
rs28583821	1	0.276	0.232	0.239	0.418	0.253	0.253	0.209	0.320
rs2395663	6	0.292	0.201	0.306	0.311	0.337	0.287	0.347	0.213
rs736795	6	0.222	0.171	0.239	0.279	0.273	0.267	0.134	0.172
rs9474391	6	0.064	0.018	0.078	0.048	0.084	0.051	0.117	0.033
rs4427625	12	0.046	0.030	0.011	0.115	0.028	0.022	0.061	0.041
rs1111662	12	0.484	0.463	0.467	0.572	0.567	0.478	0.561	0.475
rs2273708	14	0.017	0.018	0.011	0.038	0.017	0.006	0.015	0.008
rs1952151	14	0.427	0.494	0.411	0.346	0.449	0.483	0.357	0.492
rs4805440	19	0.353	0.329	0.361	0.365	0.376	0.438	0.327	0.238

***NOTE:** Among the 10 SNPs shown in this table, 8 are common SNPs and 2 are rare SNPs which do not have population stratification effect. That is the alleles are commonly uniformly distributed among different populations.*

Additionally, these selected disease SNPs were highly correlated with some non causal SNPs which was a relationship that might result in spurious association and misleading results. Table 7 shows the top five SNPs that are correlated with the 10 rare disease SNPs. The non causal SNPs were from the whole genome. Since most individuals would have two common homozygotes at two rare loci, only a few participants had score 1 or 2 at these rare SNPs. As a result, it was more common for a rare SNP to be correlated with a SNP that was far away from it, even from different chromosomes. Common causal SNPs were also correlated with some other non causal SNPs.

Table 7 Non Causal SNPs Highly Correlated with Causal SNPs for Rare Disease Model using HapMap III Genotype Data.

Causal SNPs	non- causal SNPs		
	SNP	CHR	MAF
rs11811671,chromosome 1,MAF=0.01	rs10137145	14	0.04
	rs10413708	19	0.04
	rs11807413	1	0.01
	rs12068423	1	0.01
	rs12083753	1	0.03
rs7553321,chromosome 1,MAF=0.01	rs17100268	14	0.31
	rs4902439	14	0.49
	rs1124417	12	0.37
	rs262825	6	0.49
	rs1045217	19	0.49
rs17035549,chromosome 1, MAF=0.05	rs10415814	19	0.17
	rs10424089	19	0.44
	rs10862007	12	0.34
	rs11882235	19	0.03
	rs1381029	12	0.27
rs12401645,chromosome 1, MAF=0.01	rs10499001	6	0.07
	rs10777181	12	0.21
	rs10860589	12	0.01
	rs11160249	14	0.05
	rs11571537	1	0.02
rs4512698, chromosome 1, MAF=0.05	rs1046248	14	0.04
	rs11084665	19	0.42
	rs11578776	1	0.29
	rs11610422	12	0.1
	rs11832143	12	0.25

Table 7 Non Causal SNPs Highly Correlated with Causal SNPs for Rare Disease Model using HapMap III Genotype Data (continued).

Causal SNPs	non- causal SNPs		
	SNP	CHR	MAF
rs17358725,chromosome 1, MAF=0.01	rs10424162	19	0.3
	rs10498338	14	0.1
	rs11108860	12	0.03
	rs11157383	14	0.22
	rs11583984	1	0.1
rs6693629,chromosome 1, MAF=0.03	rs10415814	19	0.17
	rs10424089	19	0.44
	rs10862007	12	0.34
	rs11882235	19	0.03
	rs1381029	12	0.27
rs12215941,chromosome 6,MAF=0.04	rs10405607	19	0.13
	rs11062710	12	0.05
	rs1145813	6	0.33
	rs11624508	14	0.04
	rs11625625	14	0.19
rs17044864,chromosome 12, MAF=0.02	rs1004968	12	0.34
	rs10499001	6	0.07
	rs11063099	12	0.1
	rs11106394	12	0.47
	rs11107705	12	0.04
rs28651257,chromosome 19, MAF=0.04	rs7148786	14	0.46
	rs35672141	1	0.3
	rs2080087	14	0.32
	rs17305332	19	0.28
	rs4805131	19	0.2

NOTE: *Each causal SNP was correlated with hundreds of non causal SNPs. All the non causal SNPs listed in this table were highly correlated with the causal SNP that most of the absolute correlation value between the non causal SNP and causal SNP was greater than 0.95.*

Table 8 and Table 9 describe the correlations by genotype for common and rare causal SNPs respectively. The correlations shown have maximum absolute 0.2. Other words, these selected SNPs can be treated as independent SNPs. Criteria I was satisfied.

Table 8 Correlation of the Ten Rare Disease SNPs based on Genotype using HapMap III Genotype Data (n=613).

Corr	rs 4512698	rs 7553321	rs 17035549	rs 6693629	rs 11811671	rs 17358725	rs 12401645	rs 12215941	rs 17044864	rs 28651257
rs 4512698	1	-0.01	0.03	-0.05	0.04	0.07	-0.02	0.07	-0.02	0.04
rs 7553321		1	0.10	-0.04	0.14	-0.02	0.04	-0.04	0.03	0.10
rs 17035549			1	0.01	-0.01	0.01	0.01	-0.08	-0.03	0.08
rs 6693629				1	0.01	-0.03	0.03	-0.04	0.03	0.03
rs 11811671					1	-0.02	-0.03	-0.04	-0.03	0.07
rs 17358725						1	-0.02	-0.03	-0.02	-0.04
rs 12401645							1	-0.01	0.08	-0.06
rs 12215941								1	-0.01	-0.08
rs 17044864									1	0.03
rs 28651257										1

Note: *The max|corr|is 0.14; most correlations have absolute values less than 0.1. Each has MAF less than 0.05*

Table 9 Correlation of the Ten Common Disease SNPs based on Genotype using HapMap III Genotype Data (n=613).

	rs 1888991	rs 1390333	rs 10498665	rs 9376923	rs 7296827	rs 7300747	rs 17100963	rs 3759670	rs 6509333	rs 17207579
rs 1888991	1	0.03	-0.45	-0.16	-0.05	0.27	-0.05	0.26	-0.10	0.00
rs 1390333		1	0.00	0.13	0.03	0.14	-0.08	0.03	0.11	0.05
rs 10498665			1	0.09	0.05	-0.18	-0.02	-0.16	0.09	0.06
rs 9376923				1	0.01	0.00	-0.05	-0.14	0.09	0.10
rs 7296827					1	0.13	-0.09	-0.04	0.12	0.09
rs 7300747						1	-0.17	0.07	0.07	0.09
rs 17100963							1	0.01	-0.09	-0.11
rs 3759670								1	0.00	-0.03
rs 6509333									1	0.15
rs 17207579										1

Note: *Most correlations have absolute values less than 0.2. Each SNP has MAF greater than 0.05*

Table 10 and Table 11 are the correlation by population for common and rare causal SNPs respectively. Each SNP is highly correlated with one or more SNPs with an absolute correlation larger than 0.6. Some pairs have a correlation greater than 0.9 which indicates the two SNPs are highly associated by population, such as *rs7553321* and *rs11811671* in Table 9 and *rs1390333* and *rs3759670* in Table 10. Therefore, criteria II was satisfied.

Table 10 Correlation of the Ten Rare Disease SNPs based on Population using HapMap III Genotype Data(P=7).

SNPs	rs 4512698	rs 7553321	rs 17035549	rs 6693629	rs 11811671	rs 17358725	rs 12401645	rs 12215941	rs 17044864	rs 28651257
rs 4512698	1	0.01	0.10	-0.49	-0.02	0.14	***0.54***	0.06	***0.61***	-0.05
rs 7553321		1	***0.74***	***0.72***	***0.97***	-0.48	0.44	***-0.56***	***0.71***	***0.87***
rs 17035549			1	***0.79***	***0.86***	***-0.55***	0.37	***-0.69***	0.38	***0.95***
rs 6693629				1	***0.82***	***-0.50***	-0.01	***-0.64***	0.09	***0.89***
rs 11811671					1	***-0.54***	0.39	***-0.61***	***0.61***	***0.96***
rs 17358725						1	0.03	0.01	-0.22	***-0.57***
rs 12401645							1	-0.13	***0.50***	0.30
rs 12215941								1	-0.23	***-0.64***
rs 17044864									1	0.45
rs 28651257										1

***NOTE**: Any one of these 10 SNPs was highly correlated with one or more other SNPs, with absolute correlation greater than 0.5.The highlight values were high correlations (|corr|>0.5).*

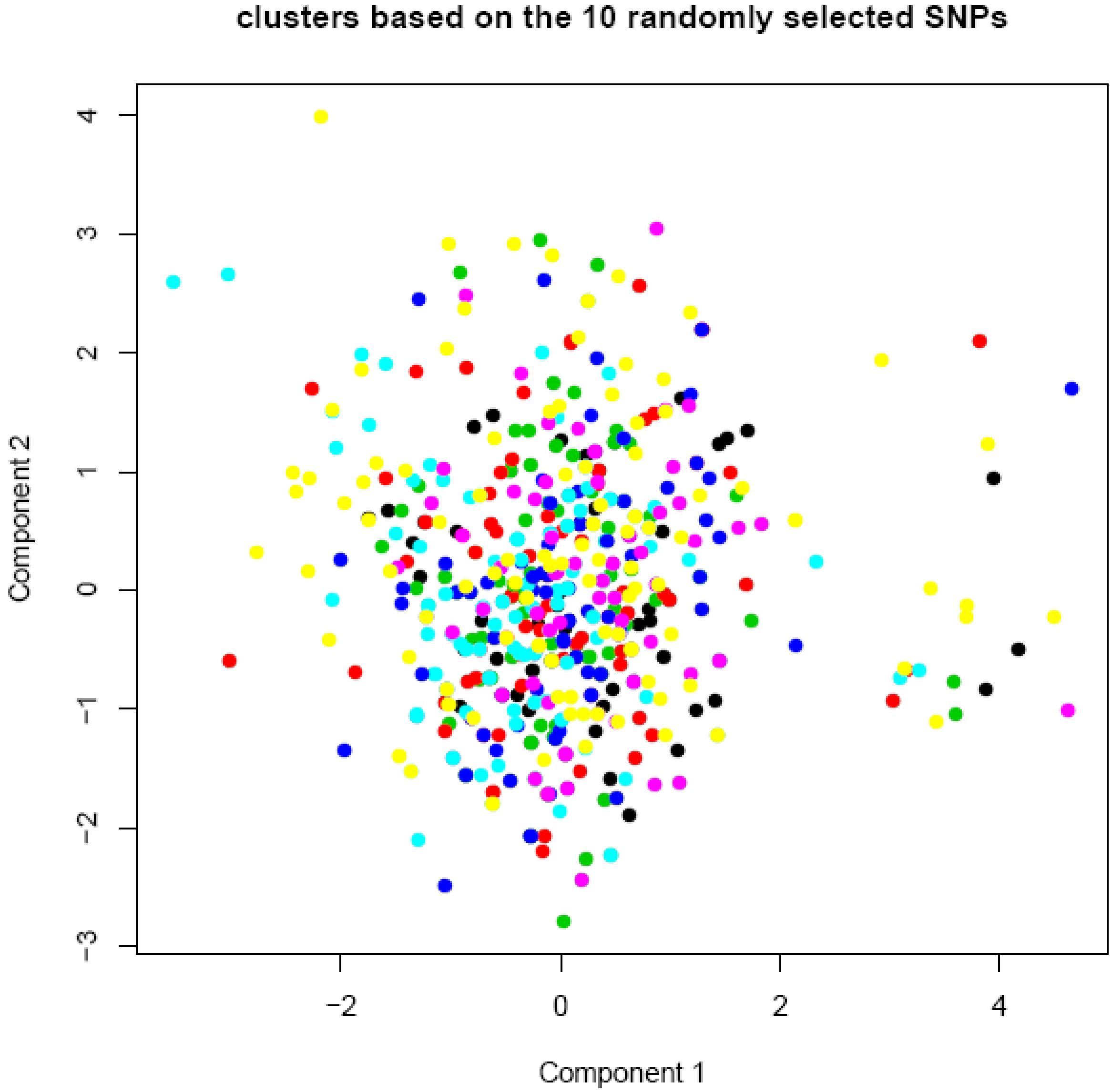
Table 11 Correlation of the Ten Common Disease SNPs based on Population using HapMap III Genotype Data(P=7).

	rs1888991	rs1390333	rs10498665	rs9376923	rs7296827	rs7300747	rs17100963	rs3759670	rs6509333	rs17207579
rs1888991	1	-0.01	***-0.98***	-0.61	-0.08	0.56	-0.07	***0.90***	-0.31	0.00
rs1390333		1	0.03	***0.75***	***0.94***	***0.78***	***-0.98***	-0.24	***0.95***	***0.98***
rs10498665			1	0.64	0.06	-0.58	0.06	***-0.93***	0.33	0.01
rs9376923				1	***0.75***	0.22	-0.68	***-0.71***	***0.91***	0.72
rs7296827					1	***0.74***	***-0.97***	-0.30	***0.92***	***0.94***
rs7300747						1	***-0.83***	0.40	0.57	***0.80***
rs17100963							1	0.15	***-0.91***	***-0.99***
rs3759670								1	-0.48	-0.17
rs6509333									1	***0.93***
rs17207579										1

***NOTE:** Any one of these 10 SNPs was highly correlated with one or more other SNPs, with absolute correlation greater than 0.5.The highlight values were high correlations (|corr|>0.7).*

Figures 3-5 below show the population structure for selected SNPs. Principle Component Analysis was performed on the set of 10 randomly selected SNPs shown in Table 6 as well as the set of selected common and rare causal SNPs. Figure 3 is a plot of the first two PCs calculated on the genotype matrix of the 10 randomly selected SNPs. Different color indicates different population. The dots with different colors are randomly distributed in the space. That is, no clusters exist for these 10 SNPs.

Figure 3 Plot of the first two PCs calculated using the 10 randomly selected SNPs

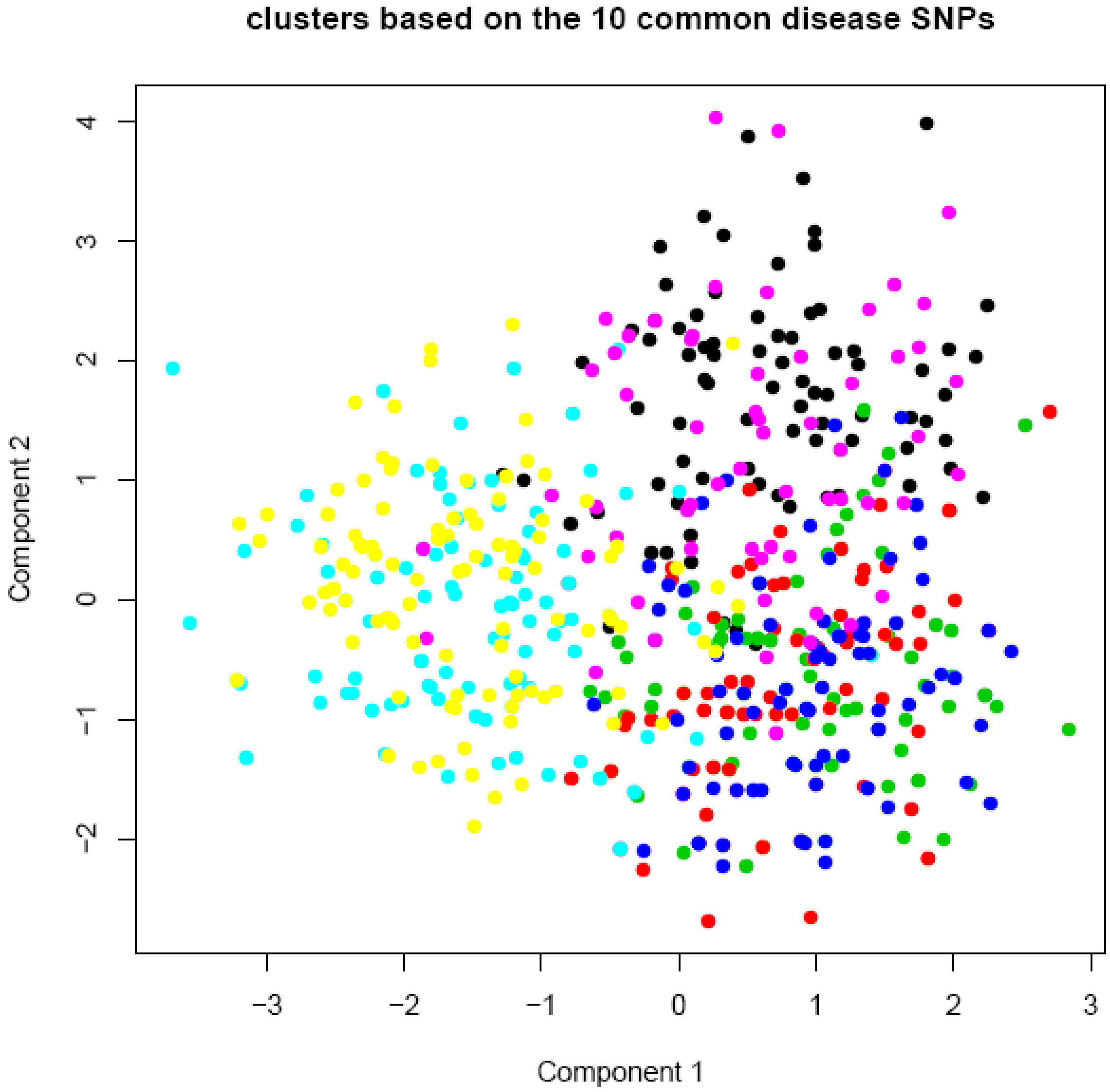


NOTE: Black-CEU; Pink-TSI; Light Blue-LWK; Yellow-YRI; Green-CHD; Red-CHB; Blue-JPT.

Figure 4 and Figure 5 are plots for the first two PCs calculated on the genotype of the common and rare disease SNPs. There are clear clusters present in the figure describing the disease SNPs.

In Figure 4, there are three major groups: 1st group contains Luhya in Webuye, Kenya (LWK, light blue in the plot) and Yoruba in Ibadan, Nigeria (YRI, yellow), both African; 2nd group contains Utah residents with Northern and Western European ancestry from the CEPH collection (CEU, black) and Tuscans in Italy (TSI, pink), both from Europe; and last group contains Chinese in Metropolitan Denver, Colorado (CHD, green), Han Chinese in Beijing, China (CHB, red) and Japanese in Tokyo, Japan (JPT, blue), all have Asian ancestry.

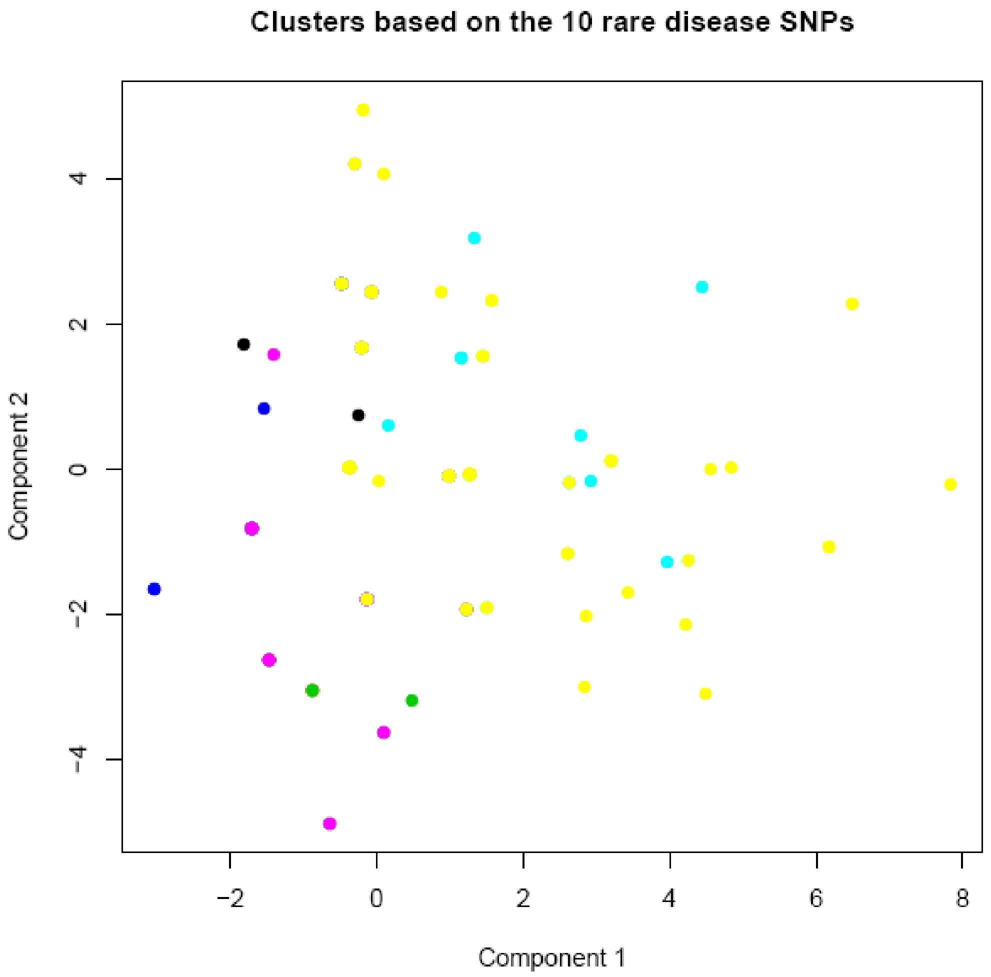
Figure 4: **Plot of the first two PCs calculated using the 10 common causal SNPs**



NOTE: *Light Blue-LWK; Yellow-YRI; Black-CEU; Pink-TSI; Green-CHD; Red-CHB; Blue-JPT*

Figure 5 show population structure for the 10 rare causal SNPs. For rare SNPs, only a few participants carry the disease alleles so that most people have a score 0, which indicates the number of disease allele, in one or more of these positions. As a result, there are many dots overlapping with each other. Yoruba in Ibadan, Nigeria (YRI, yellow) and Luhya in Webuye, Kenya (LWK, light blue) make the major group. It overlaps with some Europeans and Asians. Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) and Tuscans in Italy(TSI) are near each other. Chinese (CHD, CHB) and Japanese (JPT) are close to each other although TSI seems also close to JPT.

Figure 5: **Plot of the first two PCs calculated using the 10 rare causal SNPs**



NOTE: *Yellow-YRI; Light Blue-LWK; Black-CEU; Pink- TSI; Blue-JPT ; Green- CHD; Red-CHB;*

Unlike the random 10 SNPs, these two set of selected SNPs carry ancestry information. As a result, the ancestry distance among participants can be recognized using PCs

b. Disease Model

Assume that the probability a person has a disease when he or she has a single disease allele is 0.03, and this probability is the same for each of the 10 disease SNPs. Additionally, assume that the probability a person has the disease is proportional to the number of alleles he or she has. Then, the probability that a person with *n* disease alleles has the disease is 0.03*n*. For example, suppose that a person has genotypes *Aa*, *AA*, *AA*, *AA*, *AA*, *AA*, *aa*, *Aa*, *AA*, *AA*,where *a* is the disease allele. Then the probability that this person, who has four *a* (disease) alleles, has the disease is 0.12.

For each of the 613 GAW17 participants who were also in the HapMap III data, I found the number of disease alleles. There were 372 participants who had 0 disease alleles; 177 participants who had 1 disease allele; 57 participants who had 2 disease alleles; 4 participants who had 3 disease alleles; and 3 participants who had 4 disease alleles. There were no individuals who have more than 4 disease alleles.

For each number of disease alleles, I used the R [34] statistical package rbern(n,p) function to generate the disease phenotype. Table 12 contains the expected number of affected participants in each category and the parameter settings for generating the phenotypes. The expected number of cases among the whole population was 9.45, which was 1.5% of the 613 participants.

Table 12 Parameter Settings for Phenotype Generating for Rare Disease SNPs using Hapmap Genotype Data. (N=613)

Number of disease alleles	Number of participants	P(Disease|number of alleles)	Expected number of cases	Phenotype generator
0	372	0	0	0
1	177	0.03	5.31	Binomial(177,0.03)
2	57	0.03*2=0.06	3.42	Binomial(57,0.06)
3	4	0.03*3=0.09	0.36	Binomial(4,0.09)
4	3	0.03*4=0.12	0.36	Binomial(3,0.12)

***Note**: Participants are grouped into 5 categories based on the number of disease alleles a person has. Phenotypes are generated in each category using the R package.*

I generated 100 replicates using these 10 SNPs.

I used the same approach for the common disease SNP model. There were 10 disease SNPs. The minor allele of each contributed to the risk of having the disease. I counted the number of minor alleles among the 613 individuals. The counts ranged from 0 to 12. The probability that an individual with 12 disease alleles has the disease was 0.36. Table 13 shows the parameter settings for phenotype generating for this model. The expected number of cases among the whole population of 613 was 95.13 which was a prevalence of 15.5%.

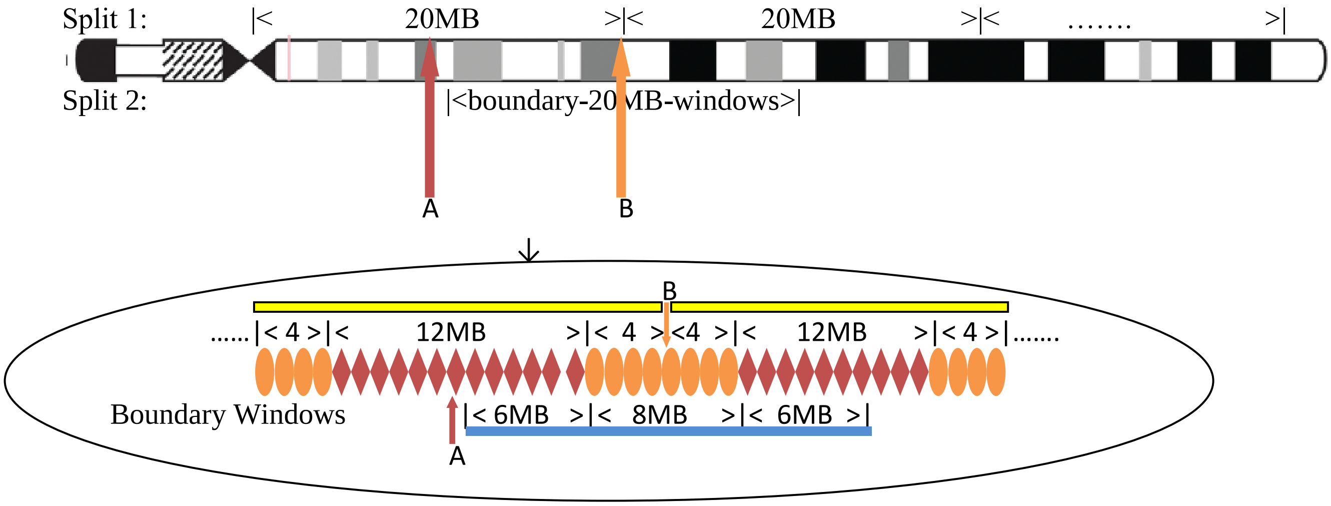
Table 13 Parameter Settings for Phenotype Generating for Common Disease SNPs using Hapmap Genotype Data. (N=613)

Number of disease alleles	Number of samples	P(Disease|number of alleles)	Expected number of cases	Phenotype Generator
0	3	0	0	0
1	17	0.03	0.51	Binomial(17,0.03)
2	39	0.06	2.34	Binomial(39,0.06)
3	81	0.09	7.29	Binomial(81,0.09)
4	99	0.12	11.88	Binomial(99,0.12)
5	111	0.15	16.65	Binomial(111,0.15)
6	92	0.18	16.56	Binomial(92,0.18)
7	86	0.21	18.06	Binomial(86,0.21)
8	54	0.24	12.96	Binomial(54,0.24)
9	20	0.27	5.4	Binomial(20,0.27)
10	6	0.3	1.8	Binomial(6,0.3)
11	4	0.33	1.32	Binomial(4,0.33)
12	1	0.36	0.36	Binomial(1,0.36)

c. Local PC Calculation

I used the software FamCC [31] to calculate local PCs, using all SNPs available in the five chromosomes chosen, (i.e., chromosomes 1, 6, 12, 14 and 19), including rare variants. Local PC required that I split a chromosome into "windows," which were non-overlapping segments of a chromosome. I first created a set of windows of length 20 MB. Since SNPs near the boundary of a window may not have the same quality of adjustment, I created a second set of windows with length 20 MB for those SNPs near boundary as shown in Figure 6. I call this set the set of boundary windows. Any SNP that was near the boundary of a window was then in the middle of a boundary window. Splits started from the front part of the chromosome (according to genomic position) and proceeded from left to right. The precise endpoints were set such that each SNP was in the middle region of at least one window. For SNPs in each window, ten PCs were calculated using all SNPs within the window. For SNPs that were in a 4MB window around either the right or left boundary (that is, for a total of 8MB), the analysis PCs were set to be the PCs calculated using the boundary windows. Figure 6 is an example of how local PCs were calculated using the two sets of windows.

Figure 6: An example of Local PC Calculation.



***NOTE:** This piece of genome was split into two sets of windows. PCs were calculated using all SNPs within a window of 20 MB (as point A). For SNPs within a band of 4 MB on each side of the window, the results from the boundary window were used. More specifically, for all SNPs in dark red, PCs were calculated using SNPs under yellow bar; for all SNPs in orange, PCs were calculated using SNPs under blue bar.*

d. Test for Genetic Association

The independent variables in these analyses were the number of minor alleles in the *i*th SNP genotype (SNP_i), and the ten local principal component scores (LPCs). The dependent variable was the disease status, which was 1 for an affected individual and 0 for an unaffected individual. The *i*th SNP was considered associated with the phenotype when the p-value of the coefficient of SNP$_i$ reported in the PLINK logistic regression analysis was less than 0.05.

I used two logistic regression models which included local PCs as covariates as specified below.

No Adjustment: $\beta_0 + \beta_1 SNP_i$

Local PC Adjustment: $\beta_0 + \beta_1 SNP_i + \beta_2 LPC_1 + \cdots + \beta_{11} LPC_{10}$

3-3 Part 3: Comparison of GPC adjustment and LPC adjustment

In order to make the results of global PC adjustment and local PC adjustment comparable, I applied the same approach I used for analysing GAW17 dataset to HapMap phase III data set, that was global PC adjustment, but with the disease SNPs I chose in previous analysis for local PC adjustment.

Global PCs were calculated using Eigensoft software [35] and applied to the same five chromosomes of HapMap data. I analyzed both the rare SNP disease model and the common SNP disease model generated above. The Global PC adjustment model I used was

$$\beta_0 + \beta_1 SNP_i + \beta_2 GPC_1 + \cdots + \beta_{11} GPC_{10}$$

The comparison results of global PC adjustment and local PC adjustment can be found in result part of section 4-2.

Chapter 4 Results

4-1 Results of Global PC adjustment for GAW 17 phenotype

The Type I error rate (i.e., false-positive rate) for non-associated genes was the fraction of replicates that had permutation p-value for the non-associated genes less than 0.05. Table 14 and Table 15 show the Type I error rates for Q1 and Q2 respectively. The model with global PC adjustment had a Type I error rate closer to 0.05 than the Type I error rate for SNP only model. For Q2, Type I error rates were relatively close to the nominal value 0.05 for each model.

Table 14: Type I error rates for Q1 in GAW17 Data Using All Non-Associated SNPs in Non-Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
No Adjustment	7.6%	5.6%	8.6%	9.0%	16.2%	19.0%	18.6%	21.7%	10.3%	9.8%
PC Adjustment	6.6%	5.7%	6.3%	6.4%	6.8%	6.9%	5.7%	6.7%	6.5%	6.0%

Note: *D represents the dichotomized phenotype, Q represents the quantitative phenotype. Non-associated SNPs came from chromosome 12, 21 and 22.*

Table 15: Type I error rates for Q2 in GAW17 Data Using All Non-Associated SNPs in Non-Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
No Adjustment	5.9%	5.6%	6.8%	7.0%	5.9%	5.8%	6.7%	7.9%	6.1%	6.0%
PC Adjustment	5.7%	5.4%	6.0%	6.0%	5.3%	5.0%	4.8%	4.9%	5.6%	5.4%

Note: *D represents the dichotomized phenotype, Q represents the quantitative phenotype. Non-associated SNPs came from chromosome 21 and 22.*

Table 16 contains the results for Q1, using all associated and non-associated SNPs in genes that determine Q1. Table 17 shows the parallel results for Q2. For non-associated SNPs in associated genes in both Q1 and Q2, Global PC adjustment had permutation Type I error rates which were closer to 0.05, although the Type I error rates were slightly above the nominal value of 0.05. In Q1 the PC adjustment model had the lowest rejection rate for associated SNPs, possibly due to better control of the rejection rate. For Q2, where all rejection rates were relatively close to the nominal rate of 0.05, the rejection rates for associated SNPs were slightly lower for the PC adjustment model.

Table 16: Type I error rates for Q1 in GAW17 Data Using All Non-Associated and Associated SNPs in Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
Non-Associated SNPs										
SNP Only	9.4%	6.9%	10.6%	13.3%	14.3%	12.4%	13.5%	16.1%	10.6%	9.0%
PC Adjustment	5.7%	6.3%	5.4%	9.3%	6.1%	4.2%	7.6%	8.9%	5.9%	6.4%
Associated SNPs										
SNP Only	14.1%	19.2%	NA	NA	87.5%	93.9%	76.3%	91.5%	26.7%	32.5%
PC Adjustment	14.3%	19.8%	NA	NA	64.0%	75.6%	64.5%	85.5%	23.2%	30.3%

Note: *D represents the dichotomized phenotype, and Q represents the quantitative phenotype.*

Table 17: Type I error rates for Q2 in GAW17 Data Using All Non-Associated and Associated SNPs in Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
Non-Associated SNPs										
SNP Only	5.2%	5.9%	7.3%	6.1%	5.8%	6.5%	5.5%	6.0%	5.4%	6.0%
PC Adjustment	5.4%	5.6%	6.9%	4.1%	5.2%	5.3%	4.9%	5.2%	5.4%	5.4%
Associated SNPs										
SNP Only	11.9%	13.3%	30.4%	31.5%	39.8%	45.6%	55.8%	80.8%	16.0%	18.2%
PC Adjustment	11.2%	12.7%	26.5%	24.5%	36.5%	44.4%	45.3%	69.8%	14.6%	16.9%

Note: *D represents the dichotomized phenotype, and Q represents the quantitative phenotype.*

For the genes reported here, Global PC adjustment had an empirical Type I error rate apparently closer to the nominal level for SNPs in genes not associated with the phenotype and for non-associated SNPs in associated genes, especially for genes determining Q1. The level of significance for Q2 was much closer to the nominal 0.05 level for each of the two models. This may be due to the way the Q2 phenotype was generated.

The power of the PC adjustment model was relatively strong and increased as the MAF increased, as expected. The power of regression modeling for the quantitative phenotype was greater than the power of logistic regression modelling of the dichotomized phenotype for both Q1 and Q2, as expected.

These findings of GAW17 data had been published. [36-37]

4-2 Comparison results of Local PC adjustment to Global PC adjustment using generated phenotype and HapMap III genotype

The type I error rate was the fraction of non-associated SNPs that had p-value less than 0.05. Table 18 and Table 19 show the type I error rates for the model with common disease SNPs and the model with rare disease SNPs respectively. For the model with common disease SNPs, the type I error rate without adjustment for population stratification was above 8%, higher than the nominal 5% rate. Both the global PC adjustment and local PC adjustment with common causal SNPs had type I error rates below the nominal level of 5%. Global PC adjustment had a type I error rate between 4.6% and 4.8%, while local PC adjustment also had a type I error rate between 4.6% and 4.8%. Similar results held for the model with rare causal SNPs. The type I error rate without adjustment for population stratification was almost 7%, higher than nominal level 5% while the type I error rate for the model with global PC adjustment controlled the type I error rate to be 3.5. The type I error rate for the model with local PC adjustment had a type I error rate to be between 3.1% and 3.6%. For the model with common causal SNPs, both global

PC adjustment and local PC adjustment had 95% confidence intervals that contain the nominal level of 5% for chromosome 14 and chromosome19.

Table 18 Comparison of Type I Error Rate for Common Causal SNPs using Generated Phenotype - HapMap III Genotype Data.

Common Causal SNPs (α=0.05)									
	no adjustment			Global PC adjustment			Local PC adjustment		
Chromosome	Observed Type I error	Confidence Interval		Observed Type I error	Confidence Interval		Observed Type I error	Confidence Interval	
1	8.2%	8.1%	8.4%	4.7%	4.6%	4.8%	***4.6%***	***4.2%***	***5.0%***
6	8.1%	8.0%	8.3%	4.6%	4.4%	4.7%	4.7%	4.5%	4.9%
12	8.5%	8.3%	8.7%	***4.8%***	***4.7%***	***5.0%***	4.8%	4.6%	4.9%
14	8.3%	8.0%	8.5%	***4.8%***	***4.6%***	***5.0%***	***4.8%***	***4.6%***	***5.0%***
19	8.1%	7.6%	8.6%	***4.7%***	***4.4%***	***5.1%***	***4.7%***	***4.3%***	***5.1%***

NOTE: The observed p_value and confidence interval are calculated for 100 replicates. The highlighted intervals contain the nominal level 0.05.

Table 19 Comparison of Type I Error Rate for Rare Causal SNPs using Generated Phenotype - HapMap III Genotype Data.

Rare Causal SNPs (α=0.05)									
	No Adjustment			Global Adjustment			Local Adjustment		
Chromosome	Observed Type I error	Confidence Interval		Observed Type I error	Confidence Interval		Observed Type I error	Confidence Interval	
1	7.0%	6.8%	7.1%	3.5%	3.3%	3.6%	3.6%	3.5%	3.8%
6	6.7%	6.3%	7.0%	3.5%	3.3%	3.6%	3.3%	3.2%	3.5%
12	6.8%	6.6%	7.0%	3.5%	3.3%	3.7%	3.2%	3.5%	3.4%
14	7.0%	6.7%	7.2%	3.5%	3.3%	3.7%	3.3%	3.1%	3.5%
19	7.0%	6.5%	7.4%	3.5%	3.0%	3.9%	3.1%	2.7%	3.6%

NOTE: *The observed p_value and confidence interval are calculated for 100 replicates.*

Figure 7 and Figure 8 show the qqplot of observed p_values against the expected uniform distribution for all non causal SNPs in chromosome 19 for the model with common causal SNPs and the model with rare causal SNPs for a randomly selected replicate respectively. For the model with common causal SNPs, the observed p_values without any adjustment for population were divergent from the uniform distribution. Both the model with global adjustment and the model with local adjustment had a uniform distribution as expected. For the model with rare causal SNPs, the observed p_value with no adjustment for population stratification was not uniform at all. The model with global PCs and local PCs adjusted the distribution so that it was closer to uniform. That is, local PC adjustment seems to be better than global PC adjustment with regard to the distribution of the observed p_values.

Figure 7 QQPlot for observed p_value against expected uniform distribution for all non causal SNPs in chr19 with common causal SNPs.

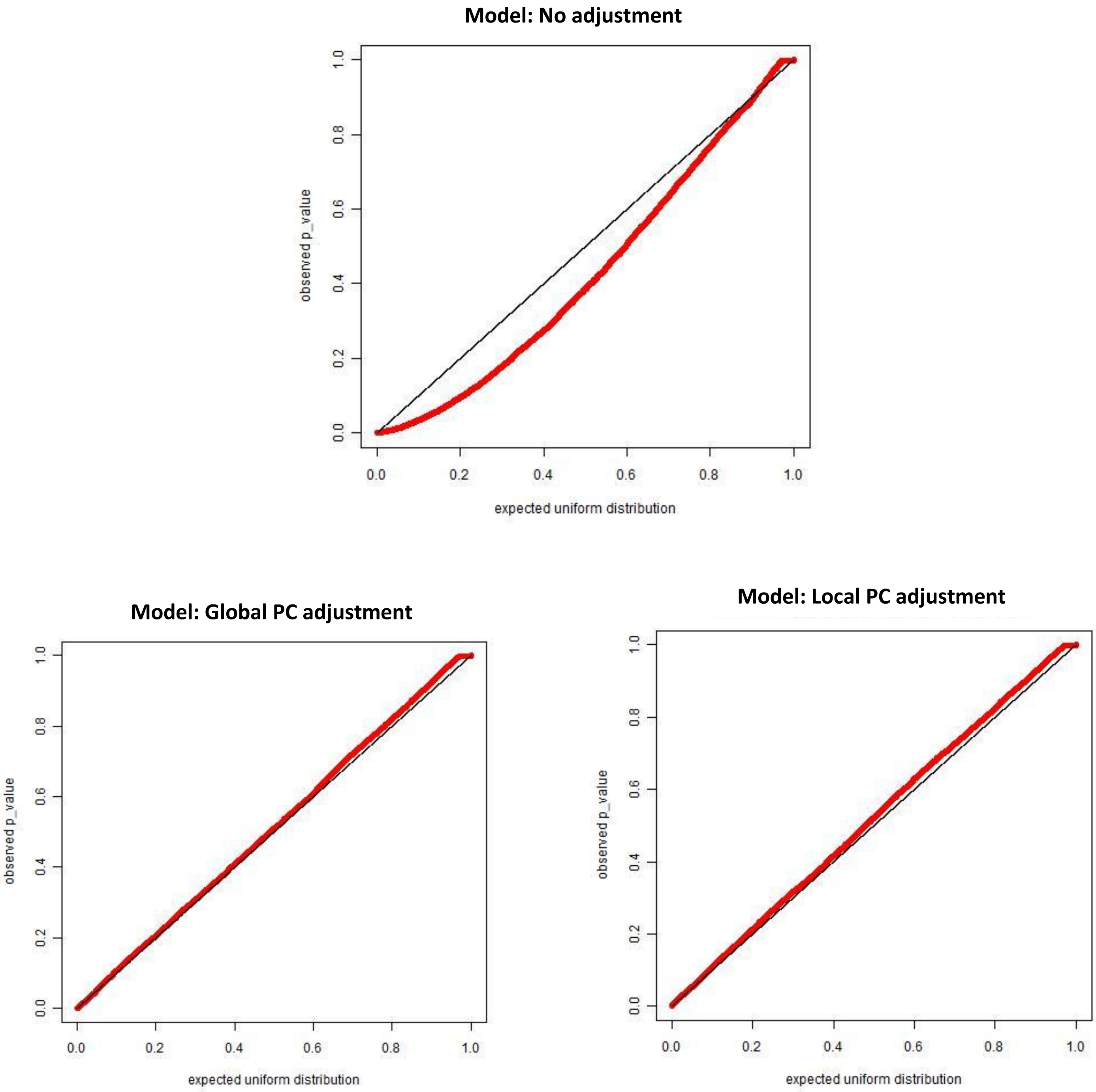
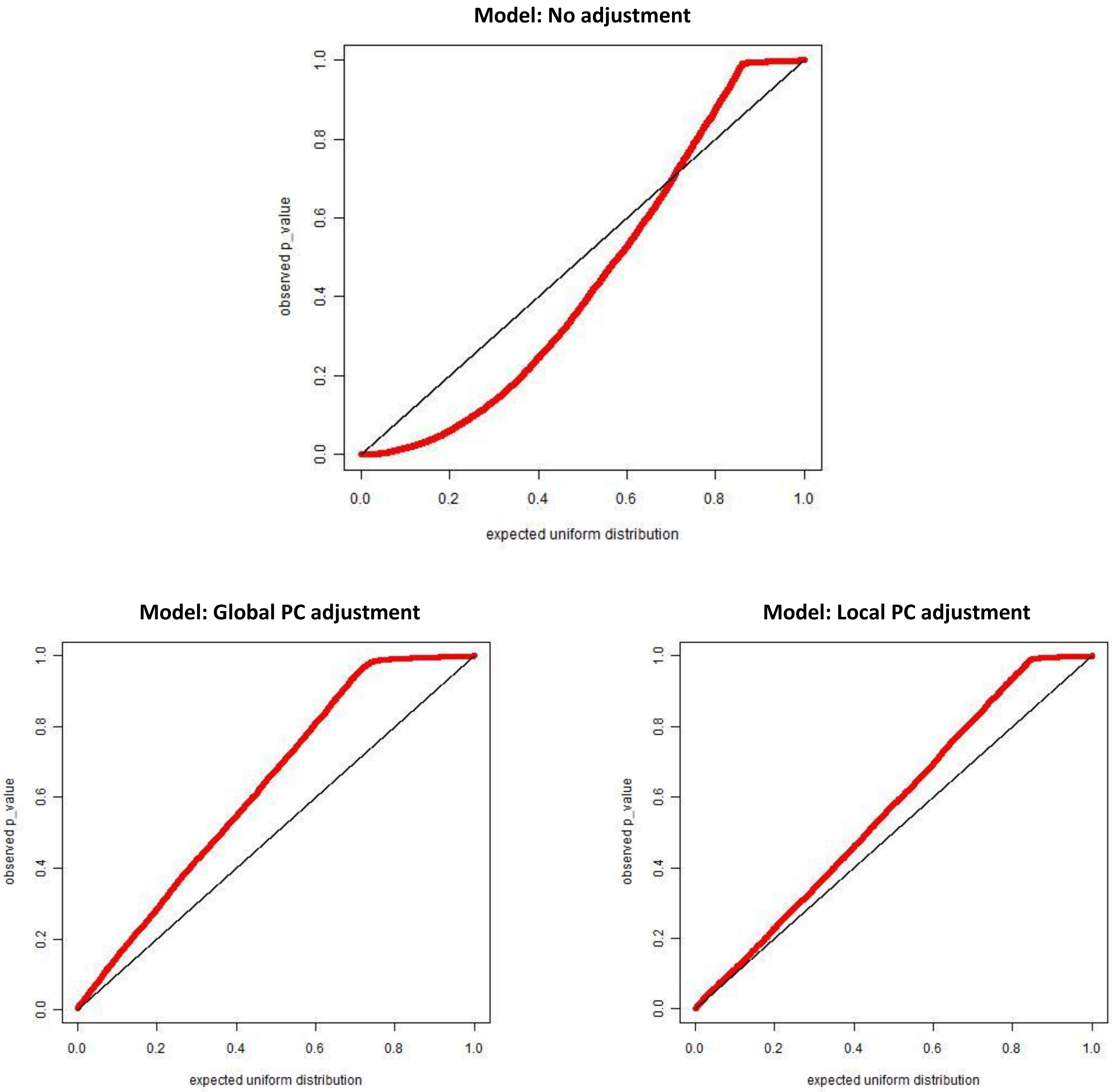


Figure 8 QQPlot for observed p_value against expected uniform distribution for all non causal SNPs in chr19 with rare causal SNPs.



Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

The qqplot for other chromosomes can be found in appendices.

Table 20 and Table 21 show the power of no adjustment, global PC adjustment, and local PC adjustment.

Table 20 Comparison of Power for Rare Causal SNPs using Generated Phenotype - HapMap III Genotype Data. α=0.05

Rare Causal SNPs					
model	chr1	chr6	chr12	chr14	chr19
number of causal SNPs	7	1	1	0	1
no adjustment	26.9%	22.0%	***27.0%***	NA	***43.0%***
Global PC	18.3%	***25.3%***	18.0%	NA	16.0%
Local PC	***31.4%***	0.0%	16.0%	NA	13.1%

NOTE: *The power was calculated based on 100 replicates. There are no causal SNPs on chromosome 14. The highlighted value was the largest power among the results for the three models.*

Table 21 Comparison of Power for Common Causal SNPs using Generated Phenotype - HapMap III Genotype Data. α=0.05

Common Causal SNPs					
model	chr1	chr6	chr12	chr14	chr19
number of SNPs	2	2	2	2	2
no adjustment	***28.5%***	15.0%	***50.0%***	14.5%	***42.5%***
Global PC	14.5%	22.1%	19.0%	***19.5%***	21.0%
Local PC	20.0%	***28.0%***	21.5%	17.5%	22.0%

NOTE: *The power was calculated based on 100 replicates. The highlighted value was the largest power among the results for the three models.*

The power was not impressive generally. The model without any adjustment seems to have higher power according to the tables above. However, considering the type I error rate performance, it was not as good as models with global or local adjustment. Table 22 shows the results of McNemar’s test for the 20 causal SNPs for global and local models.

Table 22 Results for McNemar's Test, Comparing Local PC Performance to Global PC Performance using Generated Phenotype - HapMap III Genotype Data

Type	chromosome	SNP	McNemar's Test Statistic	p_value of McNemar's Test
Common	1	rs1888991	0.27	0.61
	1	rs1390333	0.21	0.64
	6	***rs10498665***	***13.08***	***3E-04***
	6	rs9376923	0	1
	12	rs7296827	0.06	0.8
	12	rs7300747	0.44	0.51
	14	rs17100963	0	1
	14	rs3759670	1.13	0.29
	19	rs6509333	0.31	0.58
	19	rs17207579	0	1
Rare	1	rs7553321	0.9	0.34
	1	rs17035549	0.06	0.8
	1	rs6693629	0.13	0.7
	1	rs11811671	1.8	0.18
	1	rs4512698	1.9	0.17
	1	rs12401645	0.6	0.4
	1	rs17358725	2.1	0.15
	6	rs12215941	0.57	0.4
	12	rs17044864	0.13	0.72
	19	rs28651257	0.8	0.4

For common causal SNPs, nine out of ten SNPs had a p_value for McNemar's test above 0.05. For rare causal SNPs, all of the ten SNPs had a p_value above 0.05. No significant difference exists between global adjustment and local adjustment. That is, global PC adjustment and local PC adjustment had similar power, except for rs10498665 on chromosome 6.

Chapter 5 Discussion

5-1 PC Adjustment vs. Population Adjustment

The population of origin for each participant was reported in the GAW17 data. I estimated the effects of adjustment for the self reported populations. The dependent variables in these analyses were the same as in the global test for GAW17 data (see section 3-1-2). The independent variables were the number of minor alleles in the *i*th SNP genotype (SNP_i), the participant's age (*AGE*) and smoking status (*SMOKING)*, and the seven indicator variables of the populations (called POP_1 … POP_6 respectively). I used the population adjustment model given by:

$$\beta_0 + \beta_1 SNP_i + \beta_2 AGE + \beta_3 SMOKING + \beta_4 POP_1 + ... + \beta_9 POP_6$$

For Q2, age and smoking were not used as covariates.

Table 23 and Table 24 shows the Type I error rates for Q1 and Q2 respectively, comparing PC adjustment to population adjustment. For Q2, all three models have similar Type I error rate. For Q1, the model with no adjustment resulted in a Type I error rate between 5.6% and 21.7%; and the model with population adjustment had a Type I error rate between 5.8% and 23.6%. Both models showed a high inflation of type I error rate. The model with global PC adjustment, however, had a Type I error rate between 5.7% and 6.9% which was much closer to the nominal level of 5%.

Table 23: Type I error rates for Q1 in GAW 17 Data using All Non-Associated SNPs in Non-Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
No adjustment	7.6%	5.6%	8.6%	9.0%	16.2%	19.0%	18.6%	21.7%	10.3%	9.8%
Population Adjustment	9.1%	5.8%	9.0%	8.5%	16.3%	17.9%	20.0%	23.6%	11.5%	10.0%
Global PC Adjustment	6.6%	5.7%	6.3%	6.4%	6.8%	6.9%	5.7%	6.7%	6.5%	6.0%

Note: *D represents the dichotomized phenotype, Q represents the quantitative phenotype. Non-associated SNPs came from chromosome 12, 21 and 22.*

Table 24: Type I error rates for Q2 Using All Non-Associated SNPs in Non-Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
SNP Only	5.9%	5.6%	6.8%	7.0%	5.9%	5.8%	6.7%	7.9%	6.1%	6.0%
Population Adjustment	6.1%	5.6%	7.7%	7.1%	5.9%	5.6%	5.6%	5.7%	6.2%	5.7%
Global PC Adjustment	5.7%	5.4%	6.0%	6.0%	5.3%	5.0%	4.8%	4.9%	5.6%	5.4%

Note: *D represents the dichotomized phenotype, Q represents the quantitative phenotype. Non-associated SNPs came from chromosome 21 and 22.*

Table 25 contains the results for Q1, using all associated and non-associated SNPs in genes that determine Q1. Table 26 shows the corresponding results for Q2. With regard to power, the rejection rates with both population adjustment and global PC adjustment had slightly decreases from the rejection rates without adjustment.

Table 25: Rejection rates for Q1 in GAW17 Data using All Non-Associated and Associated SNPs in Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
Non-Associated SNPs										
SNP Only	9.4%	6.9%	10.6%	13.3%	14.3%	12.4%	13.5%	16.1%	10.6%	9.0%
Population Adjustment	11.3%	7.8%	10.0%	11.7%	15.8%	14.4%	8.1%	9.7%	11.5%	9.1%
Global PC Adjustment	5.7%	6.3%	5.4%	9.3%	6.1%	4.2%	7.6%	8.9%	5.9%	6.4%
Associated SNPs										
SNP Only	14.1%	19.2%	NA	NA	87.5%	93.9%	76.3%	91.5%	26.7%	32.5%
Population Adjustment	22.6%	21.7%	NA	NA	91.3%	97.7%	85.3%	98.3%	34.7%	35.4%
Global PC Adjustment	14.3%	19.8%	NA	NA	64.0%	75.6%	64.5%	85.5%	23.2%	30.3%

Note: *D represents the dichotomized phenotype, and Q represents the quantitative phenotype.*

Table 26: Rejection rates for Q2 in GAW17 Data using All Non-Associated and Associated SNPs in Associated Genes. α=0.05

MAF	MAF<0.005		0.005<MAF<0.01		0.01<MAF<0.05		0.05<MAF<0.5		Total	
Phenotypes	D	Q	D	Q	D	Q	D	Q	D	Q
Non-Associated SNPs										
SNP Only	5.2%	5.9%	7.3%	6.1%	5.8%	6.5%	5.5%	6.0%	5.4%	6.0%
Population Adjustment	6.3%	5.7%	11.2%	7.8%	6.1%	5.4%	4.9%	5.7%	6.5%	5.8%
Global PC Adjustment	5.4%	5.6%	6.9%	4.1%	5.2%	5.3%	4.9%	5.2%	5.4%	5.4%
Associated SNPs										
SNP Only	11.9%	13.3%	30.4%	31.5%	39.8%	45.6%	55.8%	80.8%	16.0%	18.2%
Population Adjustment	11.2%	12.9%	31.1%	29.1%	39.4%	46.5%	48.0%	70.8%	15.2%	17.5%
Global PC Adjustment	11.2%	12.7%	26.5%	24.5%	36.5%	44.4%	45.3%	69.8%	14.6%	16.9%

Note: *D represents the dichotomized phenotype, and Q represents the quantitative phenotype.*

As a conclusion, global PC adjustment worked better than adjustment with the indicators of the original populations. Both population adjustment and global PC adjustment showed a slightly decrease in power compared to using the nominal level of 0.05 and making no adjustment. However, considering the type I error rate performance, global PC adjustment reported a comparable power while the population adjustment worked badly here.

I have some conjectures about the causes of these patterns. Firstly, the population information was collected by an employee or reported by the patients themselves. There might have been misleading information or errors in reporting. Secondly, the population the participants reported only reflected their racial or national information. It was not satisfactory for adjustment due to stratification on genotype level. That is, the physical population did not fully explain the genotype stratification due to population drift. The participant's genotype data better reflected the population structure. Since the PCs were calculated directly from the genotypes, they were a more effective representation of the distance of genotype among the participants.

5-2 Computing Time

For global PC adjustment, the most difficult task was to get the PCs from the whole genome matrix, which had one row for each participant and one column for each SNP. Sometimes, this matrix was so large that the computer ran out of memory and crashed. For example, FAMCC software did not work for HapMap Data due to the large number of SNPs in the database. Even Eigenstrat took a whole day to read in the data and calculate the PCs.

The local PC adjustment software worked well because there were a relatively small number of SNPs within each local window. The hash time for Local PC adjustment was the time you had to wait until PLINK returned the results of logistic regression. The hash time for local PC adjustment was much longer than for global PC adjustment simply because local PC adjustment splits the genome into windows so that the work to be done in one window (the whole genome) was to be applied to multiple windows, and the time associated with this was multiplied.

In this analysis, local PC adjustment usually took about a week to get all the logistic regression results out of PLINK for one chromosome. Global PC adjustment could finish five chromosomes within one day.

Chapter 6 Conclusion

PC adjustment using logistic regression had controlled type I error rate when population stratification was an issue. PC adjustment was better than population adjustment and was able to reduce the type I error rate to the nominal level of 0.05 while achieving some power. For SNPs with MAF less than 0.05 (rare SNPs), however, PC adjustment had a type I error rate less than the nominal level.

Generally, there was no significant difference between local PC adjustment and global PC adjustment, on both type I error rate and power. Global PC adjustment was easy to understand and manipulate, although it may take a long time to calculate the global PCs. On the other hand, local PC adjustment had a similar adjustment and less demands on computer. Therefore, for a study focused on local region, considering the computing time and cost, local PC adjustment may be better than global PC adjustment.

Chapter 7 Future Work

The real meaning of these PCs can be explored. In most scenarios, the first two or three PCs were able to describe more than 90% genetic information and successfully recognized the clusters. The use of 10 PCs worked better than 3 PCs in these analyses. The extra seven PCs may hold important ancestry information. To better understand the meaning of these PCs is necessary. Further approaches on choosing best number of PCs will be explored.

Since both GPCs and LPCs can be used to adjust for PS and had similar results, combining both GPCs and LPCs as covariates in regression model might have a better adjustment. The model would be:

$$\beta_0 + \beta_1 SNP_i + \beta_2 GPC_1 + \cdots + \beta_{11} GPC_{10} + \beta_{12} LPC_1 + ... + \beta_{21} GPC_{10}$$

The disease model used in the analyses of HapMap data was simple multilocus model that unrelated causal SNPs were selected as disease SNPs. More complex model might reflect more realistic scenarios. Applying global PC adjustment and local PC adjustment to a more complex model might enable us to explore GWAS significance level and have a more significant comparison results.

Reference

1. Human Genome Project website: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
2. International HapMap Project website, http://HapMap.ncbi.nlm.nih.gov
3. National Human Genome Research Institute, http://www.genome.gov
4. Genome-Wide Association Studies (GWAS), http://gwas.nih.gov/index.html
5. Pearson TA, Manolio TA, How to Interpret a Genome-wide Association Study,*JAMA*,2008;299 (11):1335-1344
6. Genetic Analysis Workshop website, http://www.gaworkshop.org
7. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN, Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Nature Reviews Genetics* 9, 356-369, 2008
8. Zhu X, Tang H and Risch N: Admixture mapping and the role of population structure for localizing disease genes. *Adv Genet* 2008, 60:547–569.
9. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN, Demonstrating stratification in a European American population. *Nature Genet* 2005, 37, 868–872.
10. Devlin B, Roeder K, Genomic control for association studies, *Biometrics*. 1999 Dec;55(4):997-1004
11. Reich DE, Goldstein DB, Detecting association in a case-control study while correcting for population stratification, *Genet Epidemiol. 2001*, 20(1): 4–16
12. Pritchard JK, Rosenberg NA, Use of unlinked genetic markers to detect population stratification in association studies, *Am J Hum Genet 1999*, 65:220-228

13. Pritchard JK, Stephens M, Donnelly P, Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959 (2000)

14. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW, Genetic structure of human population. *Science* 298, 2381-2385 (2002)

15. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P, Association mapping in structured populations. *Am J Hum Genet* 2000, 67: 170-181

16. Falush D, Stephens M, Pritchard, Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, 2003 Aug; 164(4):1567-87

17. Falush D, Stephens M, Pritchard JK, Inference of population structure using multilocus genotype data: dominant markers and null alleles, *Mol Ecol Notes*. 2007 Jul 1; 7(4): 574-578

18. Hubisz MJ, Falush D, Stephens M, Pritchard JK, Inferring weak population structure with the assistance of sample group information, *Molecular ecology resources* (2009) **9**,1322-1332

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

20. Menozzi PP,Piazza A, Cavalli-Sforza L, Synthetic maps of human gene frequencies in Europeans. *Science* 201, 786-792 (1978)

21. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12): e190.

22. Price AL, Patterson NJ, Plenge RM, Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics* **38**, 904-909 (2006)

23. Cavalli-Sforza L. L., Menozzi P. &Piazza A. The History and Geography of Human Genes (Princeton Univ. Press, 1994)

24. Cavalli-Sforza LL, Feldman MW, The application of molecular genetic approaches to the study of human evolution, *Nature Genetics* **33**, 266 - 275 (2003)

25. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* 40, 646-649 (2008)

26. Sankararaman S, Sridhar S,Kimmel G, Halperin E, Estimating Local Ancestry in Admixed Populations, *Am J Hum Genet* 82, 290-303, (2008)

27. Kang SJ, Larkin EK, Song Y, Barnholtz-Sloan J, Baechle D, Feng T, Zhu X: Assessing the impact of global versus local ancestry in association studies. *BMC Proceedings 2009*, 3(Suppl 7):S107.

28. Qin H, Morris N, Kang SJ, Li M, Tayo B, Lyon H, Hirschhorn J, Cooper RS, Zhu X, Interrogating local population structure for fine mapping in genome-wide association studies, *Bioinformatics*, 2010 Dec 1;26 (23): 2961-8

29. 1000 Genome Project: http://www.1000genomes.org

30. GAW17 DATA

31. Zhu X, Li S, Cooper RS, Elston RC: A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet* 2008, 82(2):352-365.

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a toolset for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, 81(3): 559-575.

33. HapMap website: http://HapMap.ncbi.nlm.nih.gov/downloads/genotypes/2010-05_phaseIII/, plink format

34. R software: http://www.r-project.org/

35. Eigensoft software website: http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html

36. Jin, J., Cerise,J. E., Kang, S. J., Yoon, E. J., Yoon, S., Mendell, N. R. and Finch, S. J., Principal Components Ancestry Adjustment for Genetic Analysis Workshop 17 Data, *BMC Proceedings* 2011, 5(Suppl 9):S66

37. Thomas, A., Abel, H. J., Di, Y., Faye, L. L., Jin, J., Liu, J., Wu, Z. and Paterson, A. D. (2011), Effect of linkage disequilibrium on the identification of functional variants, *Genetic Epidemiology*, 35: S115–S119

Appendices

Table 27: Effects on Q1 in GAW17 Data

GENE	SNP	MAF	BETA
ARNT	C1S6533	0.011478	0.5619
ARNT	C1S6537	0.000717	0.64454
ARNT	C1S6540	0.001435	0.24129
ARNT	C1S6542	0.002152	0.46026
ARNT	C1S6561	0.000717	0.65721
ELAVL4	C1S3181	0.000717	0.76911
ELAVL4	C1S3182	0.000717	0.30432
FLT1	C13S320	0.001435	0.19605
FLT1	C13S399	0.000717	0.39602
FLT1	C13S431	0.017217	0.74136
FLT1	C13S479	0.000717	0.75946
FLT1	C13S505	0.000717	0.4485
FLT1	C13S514	0.000717	0.56643
FLT1	C13S522	0.027977	0.6183
FLT1	C13S523	0.066714	0.64997
FLT1	C13S524	0.004304	0.62223
FLT1	C13S547	0.000717	0.52601
FLT1	C13S567	0.000717	0.17493
FLT4	C5S5133	0.001435	0.15986
FLT4	C5S5156	0.000717	0.4301
HIF1A	C14S1718	0.000717	0.15382
HIF1A	C14S1729	0.002152	0.28532
HIF1A	C14S1734	0.012195	0.21203
HIF1A	C14S1736	0.000717	0.21716
HIF3A	C19S4799	0.000717	0.28351
HIF3A	C19S4815	0.000717	0.53114
HIF3A	C19S4831	0.000717	0.29287
KDR	C4S1861	0.002152	0.56311
KDR	C4S1873	0.000717	0.58301
KDR	C4S1874	0.000717	0.47262
KDR	KDR	0.000717	1.07706
KDR	C4S1878	0.164993	0.13573
KDR	C4S1879	0.000717	0.6183
KDR	C4S1884	0.020803	0.29558
KDR	C4S1887	0.000717	0.29558
KDR	C4S1889	0.000717	0.94133

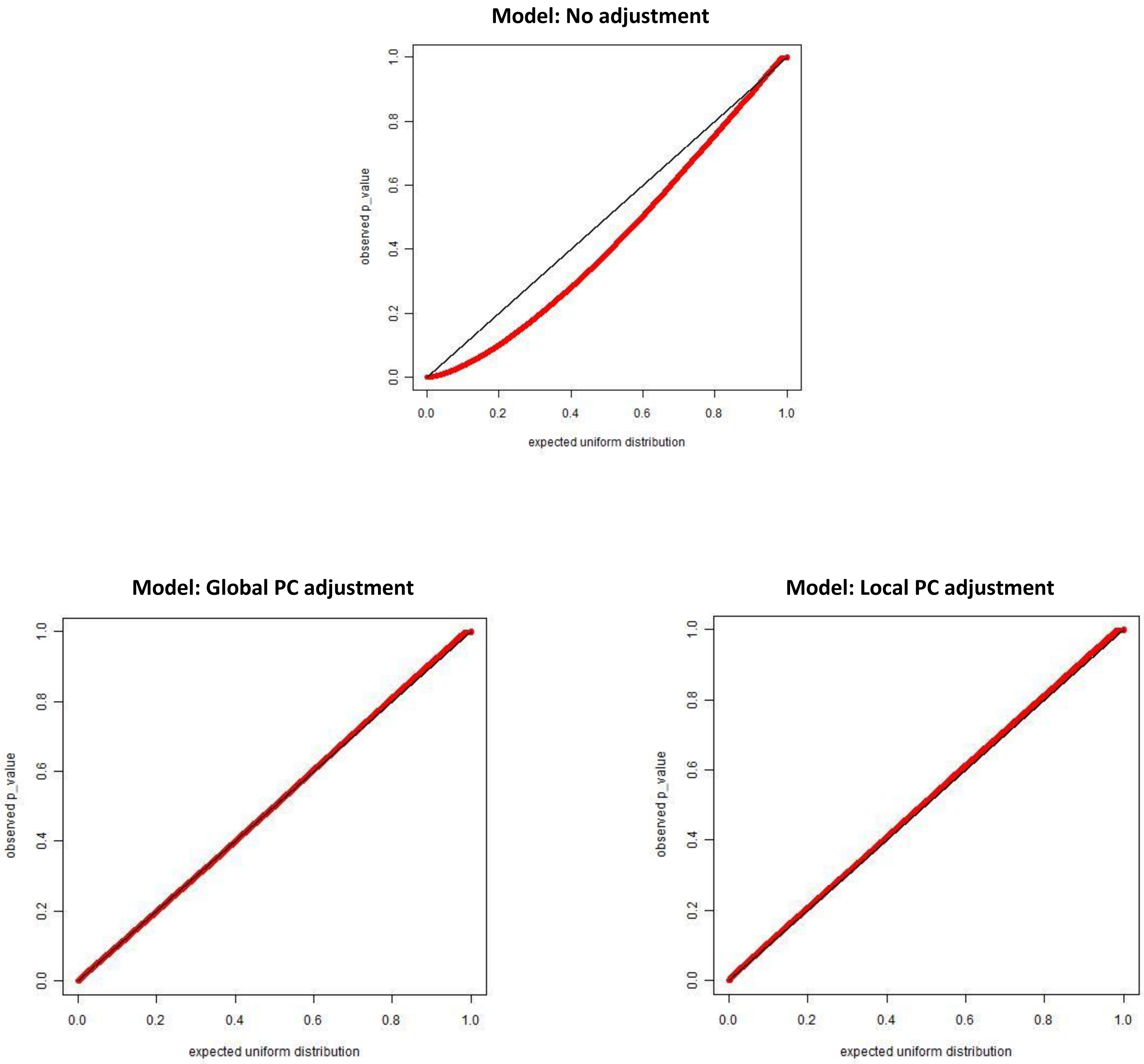
KDR	C4S1890	0.002152	0.42407
VEGFA	C6S2981	0.002152	1.20645
VEGFC	C4S4935	0.000717	1.35726

Table 28: Effects on Q2 in GAW17 Data

GENE	SNP	MAF	BETA
BCHE	C3S4834	0.000717	0.24092
BCHE	C3S4836	0.000717	0.23749
BCHE	C3S4856	0.000717	0.22027
BCHE	C3S4859	0.002152	0.59302
BCHE	C3S4860	0.000717	0.25057
BCHE	C3S4862	0.000717	1.01672
BCHE	C3S4867	0.000717	0.65326
BCHE	C3S4869	0.000717	1.01569
BCHE	C3S4873	0.002869	0.59096
BCHE	C3S4874	0.000717	1.0057
BCHE	C3S4875	0.000717	1.09484
BCHE	C3S4876	0.000717	0.75583
BCHE	C3S4880	0.001435	0.20651
GCKR	C2S354	0.012195	0.37757
INSIG1	C7S5132	0.000717	0.19962
INSIG1	C7S5133	0.000717	0.19618
INSIG1	C7S5144	0.000717	0.19275
LPL	C8S442	0.015782	0.49459
LPL	C8S476	0.000717	0.63365
LPL	C8S530	0.001435	0.72864
PDGFD	C11S5292	0.008608	0.5827
PDGFD	C11S5299	0.000717	0.82157
PDGFD	C11S5301	0.000717	0.87904
PDGFD	C11S5302	0.001435	0.81502
PLAT	C8S1741	0.003587	0.68079
PLAT	C8S1742	0.000717	0.8491
PLAT	C8S1758	0.001435	0.92516
PLAT	C8S1770	0.000717	0.62916
PLAT	C8S1772	0.001435	0.26296
PLAT	C8S1773	0.001435	0.55792
PLAT	C8S1799	0.005739	0.20651
PLAT	PLAT	0.001435	0.13767

RARB	C3S635	0.000717	0.70936
RARB	C3S679	0.005022	0.63502
SIRT1	C10S3048	0.002152	0.83224
SIRT1	C10S3050	0.002152	0.9706
SIRT1	C10S3058	0.000717	0.36621
SIRT1	C10S3092	0.000717	0.43608
SIRT1	C10S3093	0.000717	0.5352
SIRT1	C10S3107	0.000717	0.93549
SIRT1	C10S3108	0.000717	0.5328
SIRT1	C10S3109	0.000717	0.51421
SIRT1	C10S3110	0.002152	0.10326
SREBF1	C17S1007	0.002152	0.53073
SREBF1	C17S1009	0.000717	0.64568
SREBF1	C17S1024	0.004304	0.45329
SREBF1	C17S1030	0.000717	0.80366
SREBF1	C17S1043	0.004304	0.49941
SREBF1	C17S1045	0.003587	0.33524
SREBF1	C17S1046	0.002869	0.62779
SREBF1	C17S1048	0.001435	0.28739
SREBF1	C17S1055	0.001435	0.87767
SREBF1	C17S1056	0.000717	0.51524
VLDLR	C9S367	0.000717	0.58476
VLDLR	C9S376	0.002869	0.5328
VLDLR	C9S377	0.001435	1.21565
VLDLR	C9S391	0.000717	0.52694
VLDLR	C9S430	0.000717	0.55551
VLDLR	C9S443	0.001435	0.62642
VLDLR	C9S444	0.001435	0.86528
VLDLR	C9S497	0.000717	0.65808
VNN1	C6S5378	0.005739	0.45811
VNN1	C6S5380	0.170732	0.24437
VNN3	C6S5412	0.000717	0.64431
VNN3	C6S5426	0.032999	0.10326
VNN3	C6S5439	0.000717	0.10326
VNN3	C6S5441	0.098278	0.27053
VNN3	C6S5446	0.000717	0.48014
VNN3	C6S5448	0.000717	0.54036
VNN3	C6S5449	0.010043	0.66909
VWF	C12S181	0.000717	0.74757
VWF	C12S211	0.005739	0.33661

Figure 9 qqplot for all non causal SNPs in chromosome 1 with common causal SNPs

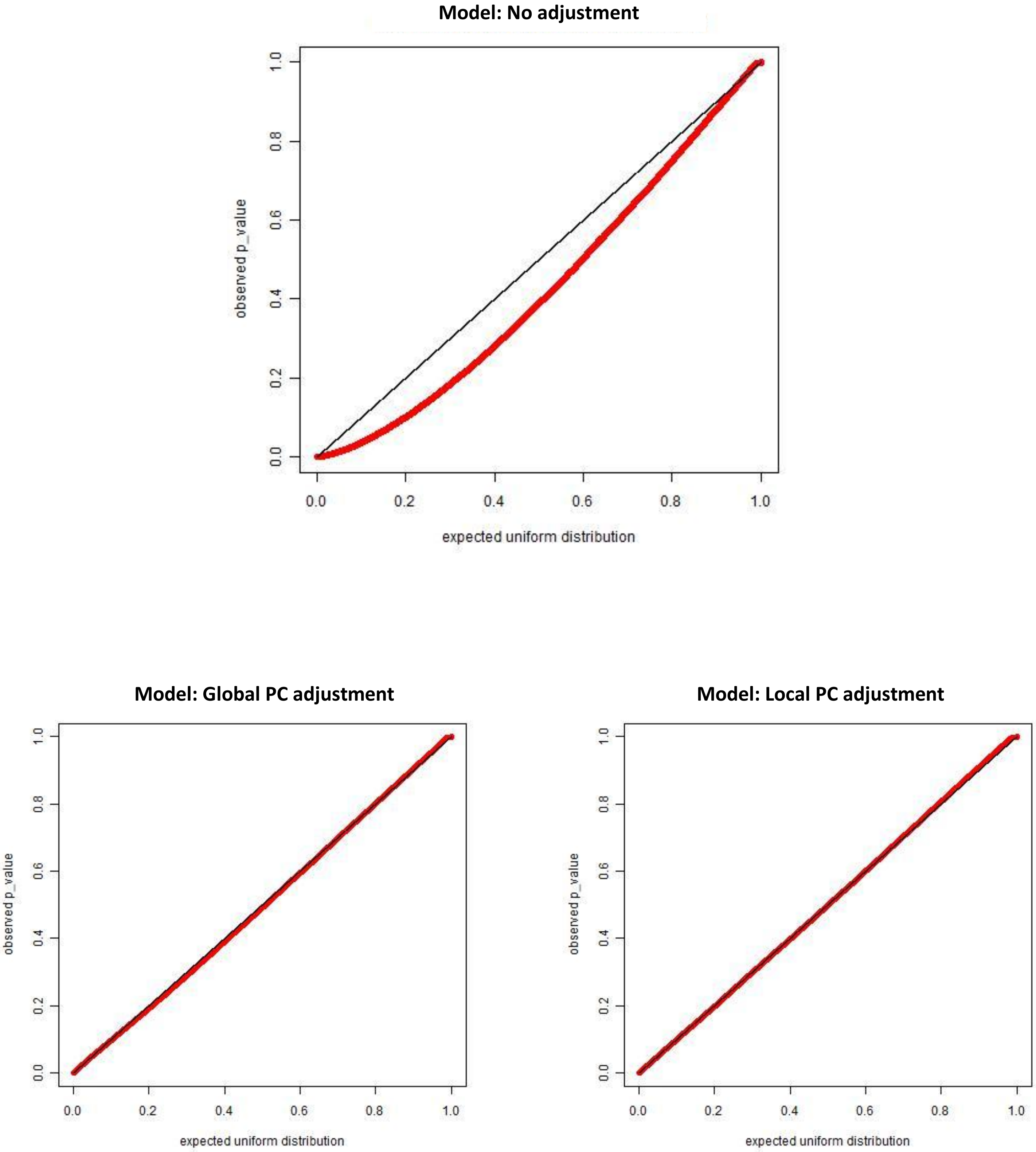


Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

Figure 10 qqplot for all non causal SNPs in chromosome 6 with common causal SNPs

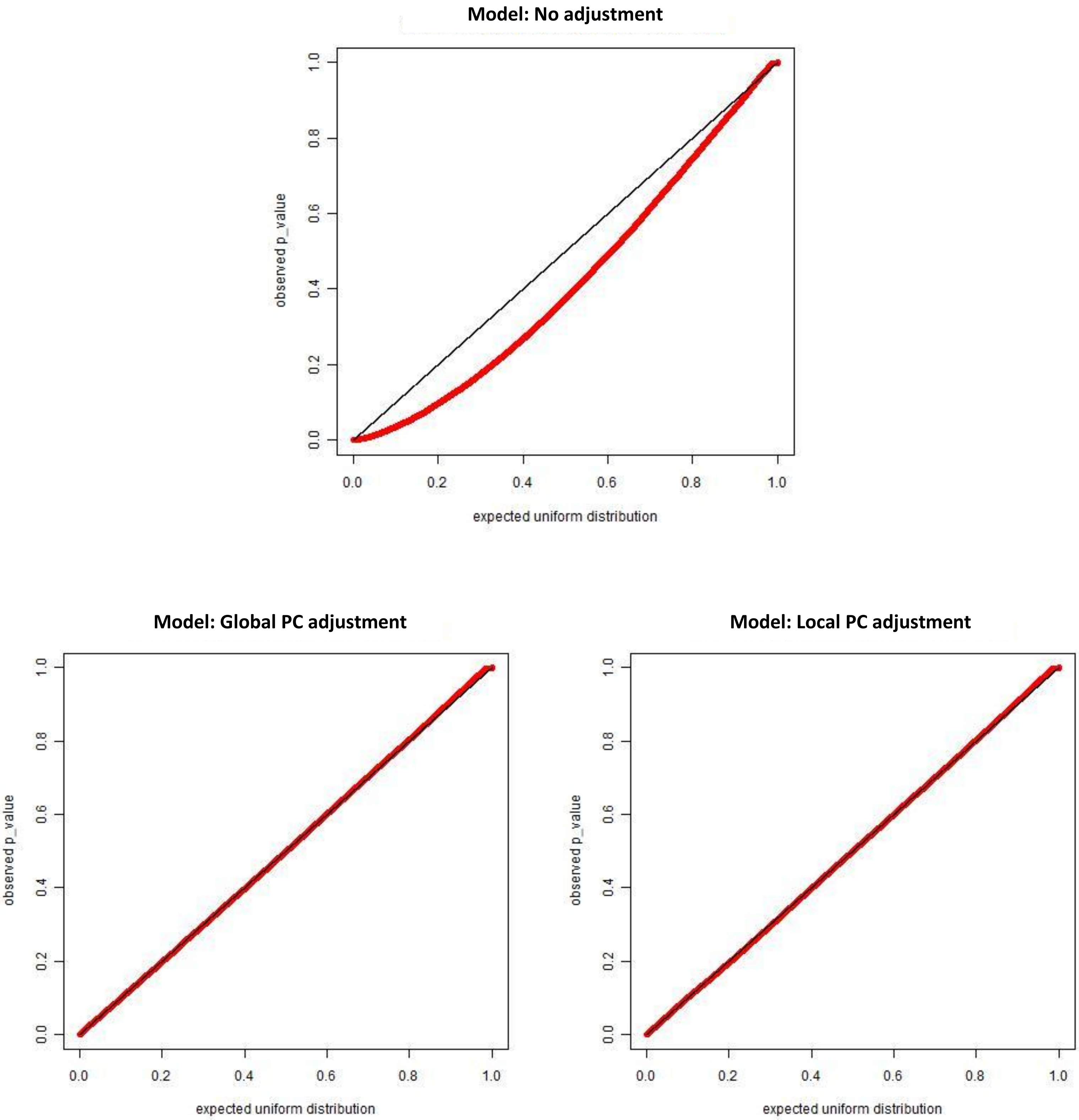


Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

Figure 11 qqplot for all non causal SNPs in chromosome 12 with common causal SNPs

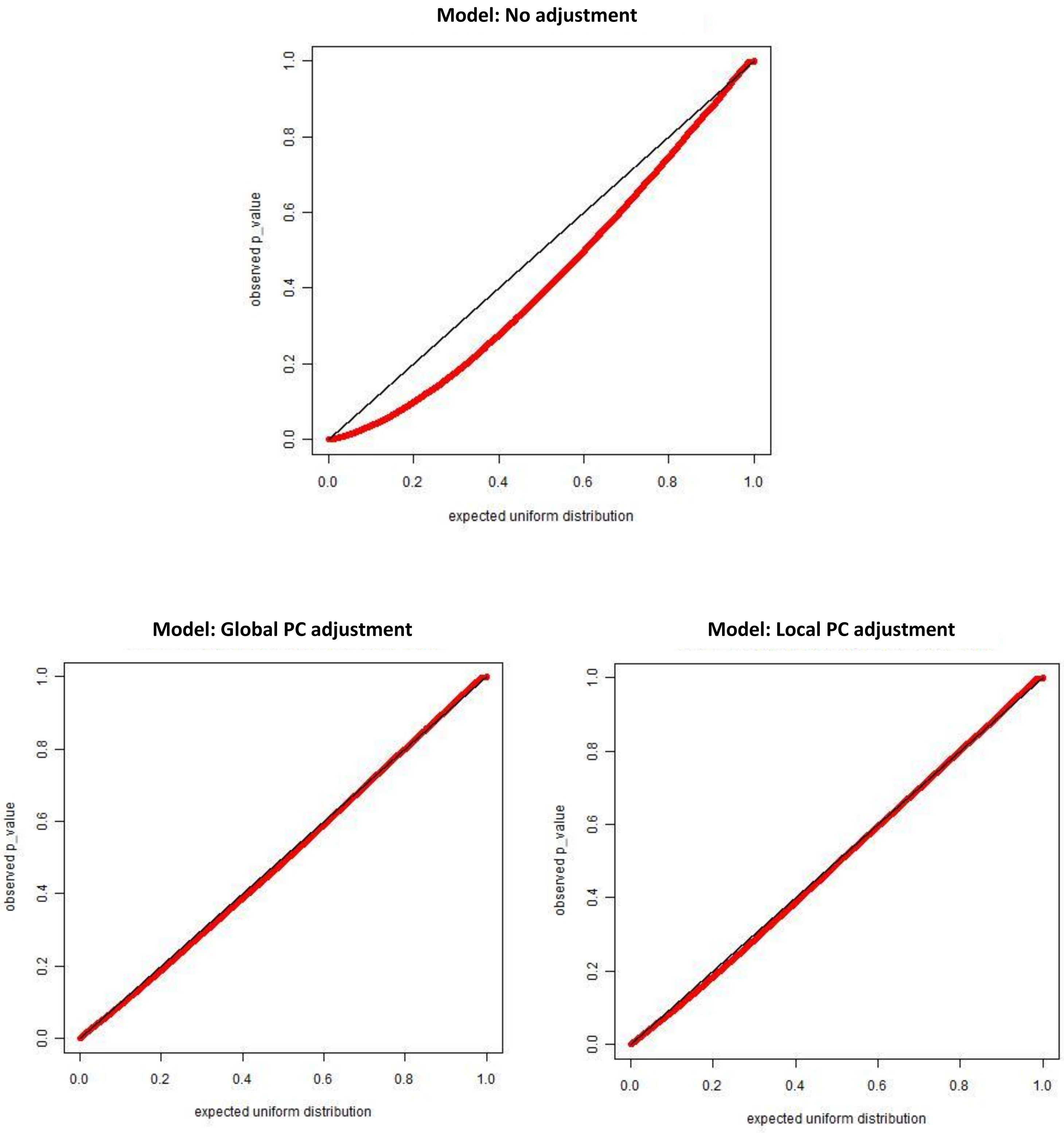


Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

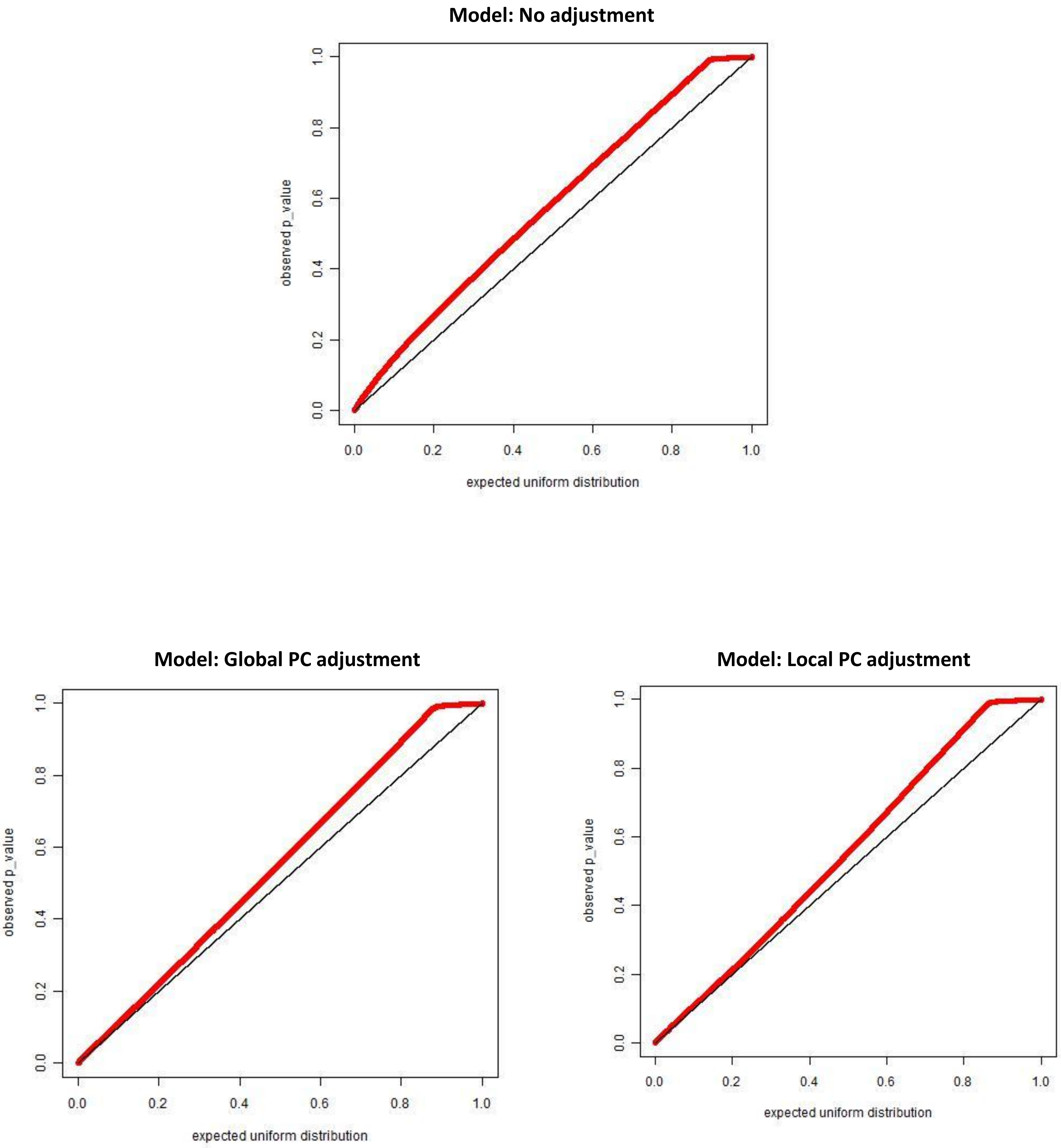
Figure 12 qqplot for all non causal SNPs in chromosome 14 with common causal SNPs


Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

Figure 13 qqplot for all non causal SNPs in chromosome 1 with rare causal SNPs

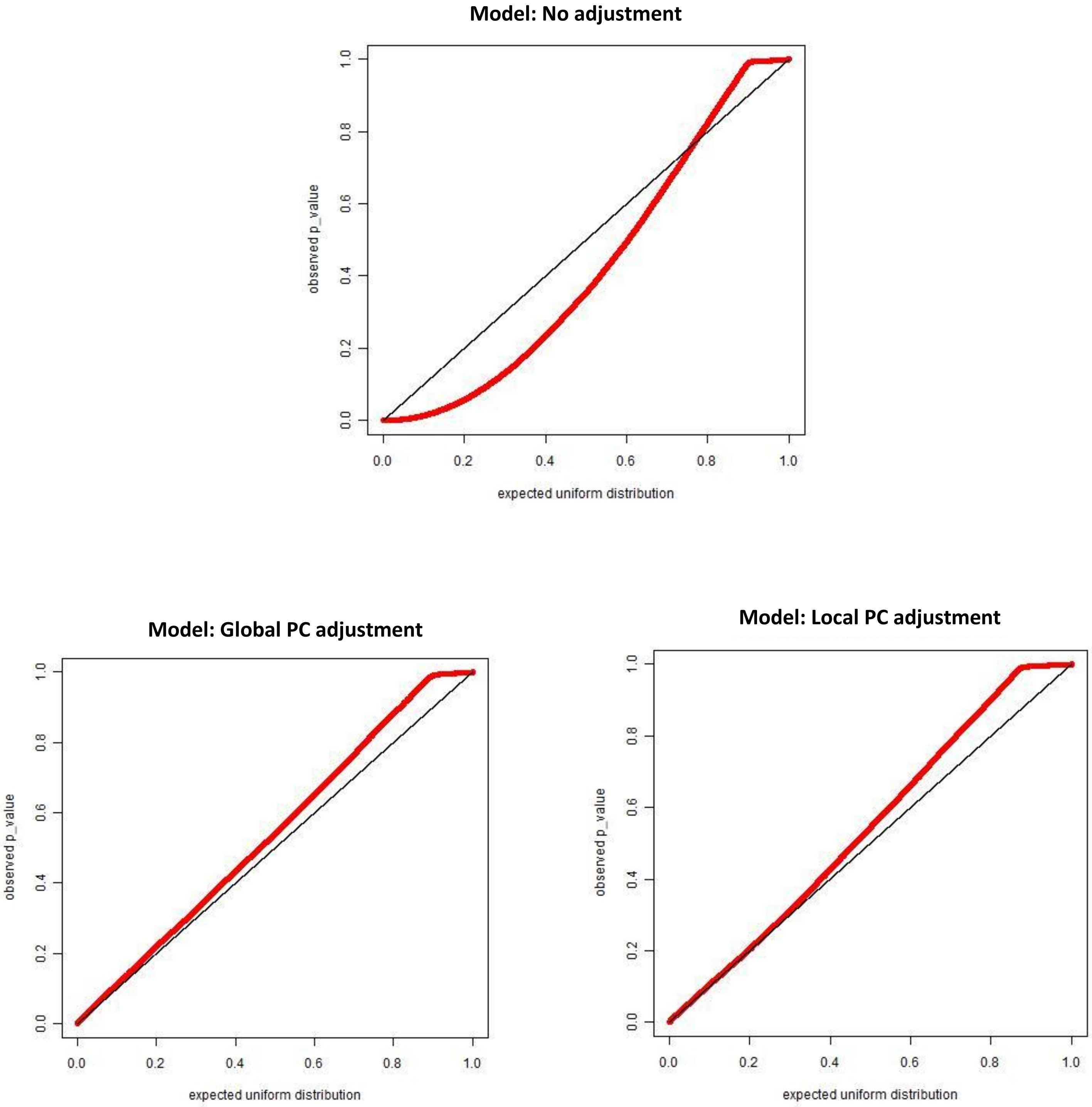


Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

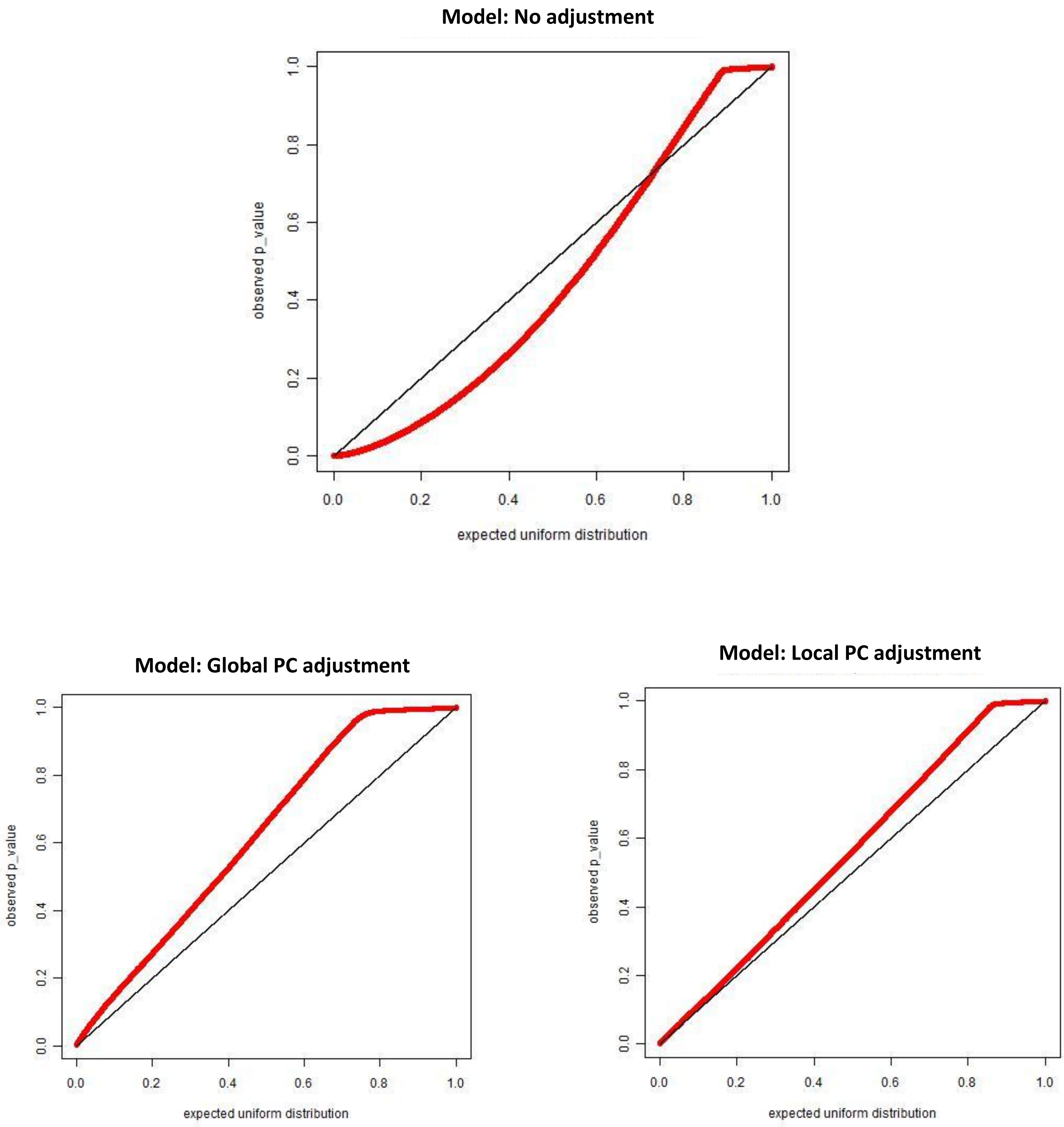
Figure 14 qqplot for all non causal SNPs in chromosome 6 with rare causal SNPs


Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

Figure 15 qqplot for all non causal SNPs in chromosome 12 with rare causal SNPs

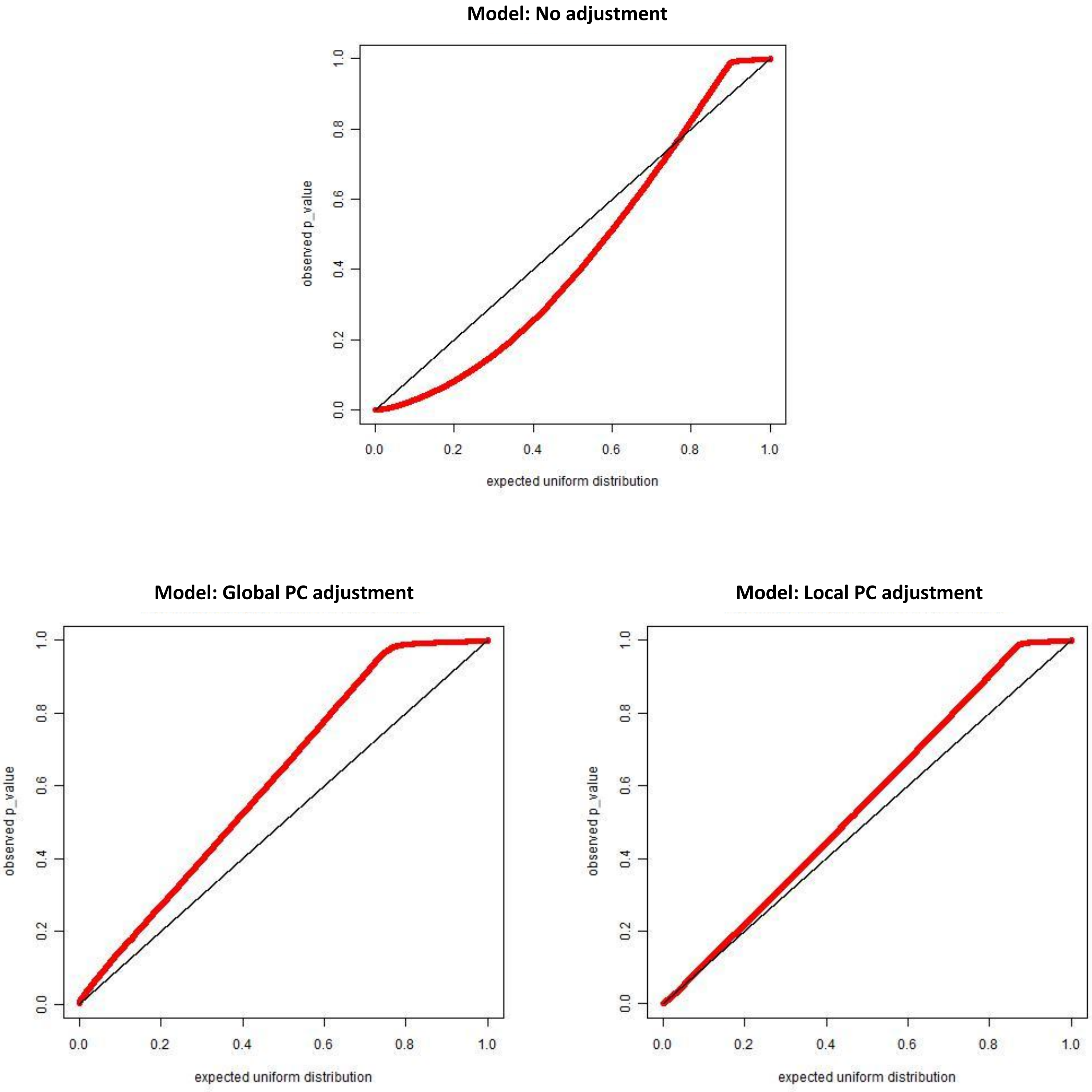


Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

Figure 16 qqplot for all non causal SNPs in chromosome 14 with rare causal SNPs



Model: No adjustment

Model: Global PC adjustment

Model: Local PC adjustment

Code-PLINK

```
--file filename --from startSNP --to endSNP2 --covar covariates, such as age, smoking, sex,
GPC, PLC --logistic –out outfile
```

Code-R

a. transform genotype matrix ATCG to 0,1,2

```
X=read.table("D:/Research/HapMap_data/rare1000.frq",header=T,sep="");

Y=read.table("D:/Research/HapMap_data/rare1000.ped",header=F,sep="");

R=length(X[,1]);

C=length(Y[,1]);

ind=factor(Y[,1]);

for (i in 1:R){

snpcol=as.vector(factor(X[i,2]))

m=X[i,3]

M=X[i,4]

    for (j in 1:C){

                if (identical(factor(Y[j,6+(2*i-1)]),factor(m))==T) k1=1 else k1=0;

                if (identical(factor(Y[j,6+2*i]),factor(m))==T) k2=1 else k2=0;

                count=k1+k2;

                snpcol=rbind(snpcol,count)

                   }

ind=cbind(ind,snpcol)

}

snp=ind[2:(C+1),2:(R+1)]
```

```
six=Y[,1:6];

geno=cbind(six,snp)


write.table(geno,

        file=paste("D:/Research/HapMap_data/rare1000.txt",sep=" "),

        quote=FALSE,sep=" ",row.names=FALSE,col.names=FALSE)
```

b. generate global adjustment results

```
chr<-6

type<-"common"

path0<-"c:/Users/Jing/Perl/perl/bin/result/"

data<-read.table(paste(path0,type,"/global/chr",chr,"/chr",chr,"_global_1.assoc.logistic",sep=""),header=T,sep="")

data<-data[,c("CHR","SNP","BP","A1","TEST","P")]

                for (i in 2:100){

add<-read.table(paste(path0,type,"/global/chr",chr,"/chr",chr,"_global_",i,".assoc.logistic",sep=""),header=T,sep="")

add<-add[,c("P")]

data<-cbind(data,add)

}

data<-data[which(data$TEST=="ADD"),]

k<-dim(data)[1]

cat(paste("number of SNP is:",k,sep=""))

data<-data[!is.na(data$P),]

m<-dim(data)[1]
```

```
cat(paste("number of SNP after filtering is:",m,sep=""))

write.table(data,paste(path0,"result_summary/",type,"_chr",chr,"_global100.txt",sep=""),quote=F,col.names=T,sep="\t")
```

c. generate local adjustment result

```
##

chr<-6

type<-"common"

path0<-paste("c:/Users/Jing/Perl/perl/bin/",sep="")

##spli1+chr19

split<-"split1"

#split<-"split2"

##read window plan

window<-read.table(paste(path0,"LPC/chr",chr,"/",split,"_chr",chr,".txt",sep=""),header=F,sep="\t")

colnames(window)<-c("start","end")

##read result

start0<-window$start[1]

end0<-window$end[1]

##chr19

##path<-paste(path0,"result/",type,"/chr",chr,"/chr",chr,"_",start0,"_",end0,sep="")

##others

path<-paste(path0,"result/",type,"/chr",chr,"/chr",chr,"_",split,"_",start0,"_",end0,sep="")

data0<-read.table(paste(path,"_1.assoc.logistic",sep=""),header=T,sep="")

data0<-data0[,c("CHR","SNP","BP","A1","TEST","P")]

                    for (i in 2:100){
```

```
add<-read.table(paste(path,"_",i,".assoc.logistic",sep=""),header=T,sep="")
add<-add[,c("P")]
data0<-cbind(data0,add)
}
data0<-data0[which(data0$TEST=="ADD"),]


for (j in 2:dim(window)[1]){
        start<-window$start[j]
        end<-window$end[j]
##chr19
##path<-paste(path0,"result/",type,"/chr",chr,"/chr",chr,"_",start,"_",end,sep="")
##others
path<-paste(path0,"result/",type,"/chr",chr,"/chr",chr,"_",split,"_",start,"_",end,sep="")
data<-read.table(paste(path,"_1.assoc.logistic",sep=""),header=T,sep="")
data<-data[,c("CHR","SNP","BP","A1","TEST","P")]
                for (i in 2:100){
add<-read.table(paste(path,"_",i,".assoc.logistic",sep=""),header=T,sep="")
add<-add[,c("P")]
data<-cbind(data,add)
}
data<-data[which(data$TEST=="ADD"),]
data0<-rbind(data0,data)
}
data0<-data0[!is.na(data0$P),]
##chr19
```

```
##write.table(data0,paste(path0,"result/",type,"_chr",chr,"_local_100.txt",sep=""),quote=F,col.names=T,sep="\t")

##others

write.table(data0,paste(path0,"result/",type,"_chr",chr,"_local_100",split,".txt",sep=""),quote=F,col.names=T,sep="\t")
```

d. Power comparison

```
Performance <-

matrix(c(13,5,3,79),

       nrow = 2,

       dimnames = list("Local" = c("True", "False"),

                       "Global" = c("True", "False")))

Performance

mcnemar.test(Performance)
```

e. Type I error rate comparison

```
chr<-6

type<-"rare"

type1<-"Rare"

data<-read.table(paste("c:/Users/Jing/Desktop/",type,"_chr",chr,".txt",sep=""),header=T,sep="\t")

m<-dim(data)[1]

data$x<-data$index/m

jpeg(paste("c:/Users/Jing/Desktop/",type,"_chr",chr,"_general.jpeg",sep=""))

qqplot(data$x,data$general,xlab="expected uniform distribution",ylab="observed p_value",main=paste("Model: No adjustment,\n",type1," Causal SNPs, Chromosome ",chr,sep=""),col=2)
```

```
lines(data$x,data$x)

dev.off()


jpeg(paste("c:/Users/Jing/Desktop/",type,"_chr",chr,"_global.jpeg",sep=""))

qqplot(data$x,data$global,xlab="expected uniform distribution",ylab="observed
p_value",main=paste("Model: Global adjustment,\n",type1," Causal SNPs, Chromosome
",chr,sep=""),col=2)

lines(data$x,data$x)

dev.off()


jpeg(paste("c:/Users/Jing/Desktop/",type,"_chr",chr,"_local.jpeg",sep=""))

qqplot(data$x,data$local,xlab="expected uniform distribution",ylab="observed
p_value",main=paste("Model: Local adjustment,\n",type1," Causal SNPs, Chromosome
",chr,sep=""),col=2)

lines(data$x,data$x)

dev.off()
```