Adjusting for Population Stratification

in Longitudinal Quantitative Trait Locus Identification

A Dissertation Presented

by

Yifan Wang

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

Yifan Wang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dr. Stephen J. Finch – Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Dr. Nancy R. Mendell – Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Dr. Song Wu – Committee Member
Assistant Professor, Department of Applied Mathematics and Statistics

Dr. Derek Gordon – Outside Member
Associate Professor, Department of Genetics, Rutgers University

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Longitudinal Quantitative Trait Locus and Population Stratification

by

Yifan Wang

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2012

Genome-wide association studies (GWAS) are widely used to detect genotypes associated with complex diseases. Such GWAS studies of disease progression over time may be clinically significant. Longitudinal quantitative trait locus (LQTL) methods are used in these studies to simulate disease progression. However, population stratification (PS) can lead to false positive or negative findings when conducting a GWAS study. PS is induced by a candidate marker’s variation in allele frequency across ancestral populations. One of the approaches used to adjust for population stratification in GWAS is the global principal component analysis (PCA) approach.

In this thesis I examine the statistical properties of GWAS analysis procedures using principal component adjustments across the whole genome. I use additive risk allele models to test the association between rare genetic variants and the longitudinal quantitative phenotypes across the whole genome. The genotype data are taken from the

Hapmap 3 dataset for 1198 unrelated individuals. The simulated quantitative phenotype data are estimated using the Bayesian posterior probabilities (BPPs) that a participant belongs to a clinically important trajectory curve. The PCA method implemented in the EIGENSTRAT program is then used to reduce the data to ten variables containing most of the genetic variability information.

The power and rejection rates are evaluated based on 1000 simulated replicates. The association test follows a chi-square distribution with one degree of freedom under the null hypothesis of no association. The p-values of the test of the coefficient of a genotype with and without a PC adjustment for PS are documented. For each disease gene, I select 25 matching SNPs (the ones with high correlation coefficient of allele frequencies with the disease gene across population) and 25 non-correlated SNPs (the ones with low correlation coefficient of allele frequencies with the disease gene across population). All SNPs considered are in overall Hardy Weinberg equilibrium (HWE).

The additive risk allele model LQTL models have strong empirical power. The model with global PCA adjustment for PS is able to consistently maintain correct false positive rates.

With thanks to my beloved parents who allowed me to dream.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BIC: Bayesian information criterion ……………………………………………………56

BPP: Bayesian posterior probability ……………………………………………………14

GWAS: Genome-wide association studies ……………………………………………...13

HWE: Hardy-Weinberg equilibrium ……………………………………………………33

MAF: Minor allele frequency …………………………………………………………...17

LQTL: Longitudinal quantitative trait loci ……………………………………………...13

PCA: Principal component analysis …………………………………………………..14

PS: Population stratification ……………………………………………………………13

QT: Quantitative Trait …………………………………………………………………...19

SNP: Single-nucleotide polymorphism …………………………………………………16

Acknowledgments

I would like to express my deep appreciation and respect to my advisor Professor Stephen Finch for his trust, encouragement, support and invaluable instructions throughout the course of my graduate study.

I greatly appreciate the help from my committee members: Professor Nancy Mendell, Professor Derek Gordon and Professor Song Wu for their careful reading of this dissertation and helpful suggestions.

I am grateful to my dear friends in and out the department for their support: Dr. Tong Shen, Dr. Ruixue Wang, Dr. Jing Jin, Dr. Douglas Londono, Anthony Musolf, Tongfei Guo, Jiayin Sun and Yuan Sun.

Finally, I wish to express my special thanks to my father Jianfeng Wang and my mother Xiaowen Wei for their patience and love that made it possible for me to complete this dissertation.

Chapter 1 Introduction

Genome-wide association studies (GWAS) are widely used to detect genotypes associated with complex diseases. GWAS studies of disease progression over time are of increasing clinical significance. Longitudinal quantitative trait locus (LQTL) methods are used in these studies to assess disease progression. It is well documented that population stratification (PS) can lead to false positive or negative findings when conducting a GWAS study (Campbell and others, 2005; Deng, 2001; Ewens and Spielman, 1995; Heiman and others, 2004a; Heiman and others, 2004b; Marchini and others, 2004; Tian and others, 2008a; Tian and others, 2008c). PS is induced by a candidate marker's variation in allele frequency across ancestral populations. One of the widely used approaches to account for population stratification in GWAS is the global principal component analysis (PCA) approach (Menozzi and others, 1978; Novembre and Stephens, 2008; Patterson and others, 2006; Zhu and others, 2002) as calculated using the EIGENSTRAT software (Price and others, 2006). Bouaziz and his colleagues document that logistic regression using principal components as covariates is an effective tool to control the false positive rate in the study of a time-constant phenotype (Bouaziz and others, 2011b). The research questions in this dissertation are 1) Whether PS can induce false positive findings in the study of longitudinal traits? 2) If so, does the use of global PCs reduce or eliminate the effects of PS?

In my study, I used several additive risk allele models to test the association between genetic variants and the longitudinal quantitative phenotypes. The genotype data were taken from the Hapmap 3 dataset for 1198 unrelated individuals. The synthetic longitudinal disease data for an individual was generated by a trajectory group model, with trajectory group determined principally by the individual’s genotypes. The longitudinal data was then analyzed using the trajectory analysis software PROC TRAJ. The trajectory group with greatest estimated change was used as the “clinically important” group. The simulated quantitative phenotype traits were the estimated Bayesian posterior probabilities (BPPs) that a participant belonged to the clinically important trajectory group. The longitudinal trajectories were simulated to represent the observed progression in a disease such as Adolescent Idiopathic Scoliosis (AIS) study (Wise and others, 2008). The PCA method implemented in the EIGENSTRAT program was then used to reduce the genetic data to ten variables containing most of the genetic variability information.

1.1 Genome-wide Association Studies (GWAS)

During the past few decades, genetic research has focused more on complex human diseases such as asthma, Alzheimer's disease, cardiovascular disease, and diabetes. To better understand the pathogenesis of such complex diseases, researchers use GWAS to detect the genetic loci associated with a disease. Their aim is to improve prevention and treatment strategies by locating the genes that are implicated with the disease and its progression.

A single nucleotide polymorphism (SNP) is the simplest type of polymorphism and occurs when one nucleotide is substituted for another based on a single mutation. Nearly three million variants have been reported and are catalogued in a public database (http://www.ncbi.nlm.nih.gov/SNP/).

A GWAS seeks to assess the correlations between genotype frequencies of single nucleotide polymorphisms (SNPs) and genetic variants and disease trait levels across populations. There are at least three explanations for an association between an allele and a phenotype (Cardon and Palmer, 2003). A first is that the allele may directly affect the expression of the phenotype. Second, the allele may be correlated with a causative allele located nearby. Third, the association may be due to confounding or selection bias.

There are two commonly used study designs for GWAS. One is a family-based design, and the other is a design based on samples of unrelated individuals. A design based on sampling individuals can be more powerful in detecting weak genetic effects. These studies include the traditional case-control studies, which are commonly used and cohort studies.

Price et al. (2010) report that GWAS studies have identified hundreds of common variants associated with disease risk or related traits (Price and others, 2010). Since most genetic heritability remains unexplained, future work will increasingly focus on variants of low minor-allele frequency or rare variants (Manolio and others, 2009). I define a variant with low minor-allele frequency (MAF) as having MAF between 0.5% and 5%. I use the term “rare variant” to refer to a SNP with MAF less than 0.5%. In my study, I focus on variants with MAF between 1% and 5%.

The genetic analyses reported here were generated by PLINK. PLINK is a freely available program with open-source code. It is a whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner. The focus of PLINK is purely on analysis of genotype/phenotype data. The software deals with PS using a clustering approach and corresponding command codes.

1.2 Longitudinal Quantitative Trait Loci (LQTL)

A quantitative trait locus (QTL) is a gene that affects a quantitative trait. In a longitudinal study, observations on a participant are taken at more than one time point, and associations are considered temporally. Longitudinal studies may be prospective, retrospective, or, more commonly, partially retrospective. They are useful for studying the effects of new interventions or possible trends in behavior. Since a longitudinal study analyzes events at more than one point in time, it may suggest the causal direction of associations (Bowling, 2003).

The case-only study is a one approach to locate a longitudinal QTL (LQTL). A second approach is population based. The issues for the validity of a study are the same for both types of studies (Caspi and others, 2003). There is one well-recognized assumption for the validity of such studies. The susceptibility genotypes and each confounding variable must be independent in the population. Possible confounding factors include environmental variables and population stratification. When these are present, the study may be biased. Wang and Lee (2008) showed that hidden stratification in the study population could also severely bias a case-only study (Wang and Lee, 2008). They derived formulas for PS bias in a study using logistic regression. The bias involves three terms: 1) the coefficient of variation of the exposure prevalence odds, 2) the coefficient of variation of the genotype frequency odds, and 3) the correlation coefficient between the exposure prevalence odds and the genotype frequency odds.

1.3 Population Stratification (PS)

GWAS researchers reported that there were biases in GWAS studies and that only a few associations were consistently and convincingly replicated (Campbell and others, 2005; Tian and others, 2008a). That is, there were discoveries of spurious associations. Failure to account for the bias induced by population stratification (PS) is thought to be one of the main causes of spurious or incorrect findings (Campbell and others, 2005; Deng, 2001; Ewens and Spielman, 1995; Heiman and others, 2004a; Heiman and others, 2004b; Marchini and others, 2004; Tian and others, 2008a; Tian and others, 2008c). PS occurs when there is a systematic difference in allele frequencies between subpopulations in a population. There can then be admixture among populations in a sample due to demographic history, natural selection, and mating between subpopulations. For example, there is admixture of populations of African and European descent in the United States (Tiwari and others, 2008).

Enoch (2006) reported that population subdivision, recent admixture and sampling variance could lead to spurious associations between a phenotype and a marker locus, which may have masked true associations in case-control studies. Wang et al. (2006) reported that confounding by ethnicity (i.e., a type of PS) can result in bias and incorrect inferences in genotype-disease association studies (Wang and others, 2006). PS can lead to confounding in association studies, such as case-control studies, where the association found could be due to the underlying structure of the population rather than a genetic locus. Wacholder et al. (2000, 2002) noted that both the frequency of the marker and the background disease prevalence must vary substantially by ethnicity for PS to be an issue (Wacholder and others, 2000). Replication of genetic association studies may, therefore, be problematic partly because of PS (Ziv and Burchard, 2003).

Many genetic epidemiologists consider PS to be a manageable problem (Pritchard and Rosenberg, 1999). If the population structure is known or estimated, there are a number of ways to incorporate this structure into the association study to adjust for potential population bias (Tiwari and others, 2008). The four most widely used approaches to adjusting for PS are genomic control (GC), structured association methods (or as they are sometimes called PC based methods), regression models, and meta-analyses (Bouaziz and others, 2011a).

The GC approach is a nonparametric method for controlling the inflation of test statistics (Devlin and Roeder, 1999; Pritchard and Rosenberg, 1999; Reich and Goldstein, 2001). GC aims at correcting the null distribution of such statistics as the linear trend test by estimating an inflation factor using many markers. Researchers usually consider that an inflation factor less than 1.05 indicates that there is no important population stratification. The main assumption of GC is that the inflation factor is the same for all markers. Hao et al (2004) proposed GC approaches to detect and adjust for false findings partially due to PS (Hao and others, 2004). They tested the performance of two GC approaches in different scenarios including various numbers of GC markers and different degrees of population stratification and conducted extensive benchmark analyses on GC approaches using SNPs over the whole human genome and found that GC methods can cluster subjects into homogeneous subgroups if there is a substantial difference in genetic background. The inflation factor, estimated by GC markers, can effectively adjust for the confounding effect of PS regardless of its extent. They also suggest that as few as 50 random SNPs with heterozygosity >40% should be sufficient for effective GC adjustment.

The structured association methods (Pritchard and others, 2000) use genetic information to estimate and control for population structure. These approaches aim at inferring the structure of the population using parametric models. Currently, the most-widely used structured association PC software is EIGENSTRAT, developed by Price and colleagues (Price and others, 2006). See descriptions in the following section 1.4. Other software for this approach are the STRUCTURE software (Pritchard and others, 2000; Rosenberg and others, 2002), the STRAT software (Pritchard and others, 2000) and the ADMIXTURE software (Alexander and others, 2009).

Logistic regression models are used to adjust the results of the usual association test to correct for stratification using the PCs as independent variables. Tsai et al. (2005) reported an association study of Latin Americans that is an example of how to adjust PS using a logistic regression strategy (Tsai and others, 2005). They studied 362 Latino subjects with asthma and 359 ethnically matched controls. There were two groups of Latino participants—those from Mexico and those from Puerto Rico. Since they were concerned about PS, they genotyped each participant on 44 ancestry informative markers (AIMs). They compared allele frequencies of the 44 AIMs to assess whether there was any indication of PS. They found significant differences in allele frequencies between Puerto Rican cases and controls but no differences between Mexican cases and controls. They used logistic regression to test for associations between disease status and AIMs with age and gender entered as covariates. Having found evidence of PS, they adjusted for it by including ancestral proportions in the logistic regression model as covariates. They concluded that the assessment of stratification effects is critical to interpret case-control studies in admixed populations.

Wang et al. (2006) addressed the effect of PS in gene-gene or gene-environment interaction studies (Wang and others, 2006). They used logistic regression models to fit multiplicative interactions between two dichotomous variables that represented genetic and/or environmental factors for a binary disease outcome in a hypothetical cohort of multiple ethnicities. Biases in main effects and interactions due to PS were evaluated by comparing regression coefficients in models that were mis-specified because they ignored ethnicities with coefficients in models that accounted for ethnicities. They showed that biases in main effects and interactions were constrained by the differences in disease risks across the ethnicities. Therefore, large biases due to PS are not possible when baseline disease risk differences among ethnicities are small or moderate, which is consistent with Wacholder et al. (2000, 2002). Numerical examples of biases in genotype-genotype and/or genotype-environment interactions suggested that biases due to PS for main effects were generally small but could become large for studies of interactions, particularly when strong linkage disequilibriums between genes or large correlations between genetic and environmental factors existed. However, when linkage disequilibrium among genes or correlations among genes and environments were small, biases to main effects or interaction odds ratios were small to nonexistent.

There are also a number of less commonly used methods for adjusting PS: the qualitative semi-parametric test (Chen and others, 2003), the simultaneously correcting method (Cheng and Lin, 2007), a simple and improved correction in case-control studies (Epstein and others, 2007), the genotype-based matching method (Guan and others, 2009), matching strategies (Hinds and others, 2004), the variance component model (Kang and others, 2010), a randomization test (Kimmel and others, 2007), and the propensity score approach (Zhao and others, 2009).

Divers et al. (2007) used ancestry informative markers (AIMs) to obtain individual admixture proportion estimates (Divers and others, 2007). They used these estimates to reduce the false positive rate (type I error) or the loss of power due to PS or genetic admixture. They reported that the quadratic measurement error correction (QMEC) method maintains the type I error at its nominal level and controls for the confounding effect of admixture in genetic association tests.

The International HapMap Project has provided allele frequencies for approximately three million single nucleotide polymorphisms (SNPs) in Africans, Europeans and East Asians. SNP marker frequency variation is greatest in Africans. Statistical methods, such as structured association and genomic control for detecting and correcting for PS, use marker loci spread throughout the genome that are unlinked to the candidate locus to estimate the ancestry of individuals within a sample, and to test for and adjust the ethnic matching of cases and controls (Seldin and Price, 2008). Enoch and his colleague (2006) focused on the methods for selection of highly informative marker loci required to characterize populations that vary in substructure or the degree of admixture, and discussed how these theoretically desirable approaches can be put into practice effectively.

There are several comparative studies of approaches to correct for PS. Tsai et al. (2005) compared three different methods: maximum likelihood estimation, the program ADMIXMAP and the program STRUCTURE (Tsai and others, 2005). They used two simulated data sets and one real data set from a genetic study of asthma among Latino subjects. All three methods provided similar accuracy of ancestral estimates and similar control of type I error rate. They demonstrated that 100 AIMs were required since the main factor affecting the accuracy of individual ancestry estimates and controlling for the type I error rate was the number of AIMs.

Kosoy et al. (2008) organized a set of 128 AIMs in order to provide a resource for assessing continental ancestry in variety of genetic studies (Tian and others, 2008b). They chose markers for informativeness, genome-wide distribution, and genotype reproducibility on two platforms (TaqMans assays and Illumina arrays). They analyzed different ancestry for genotyping data from 825 subjects, including Europeans, East Asians, Amerindians, Africans, South Asians, Mexicans, and Puerto Ricans. A complete set of 128 AIMs and subset of 24 AIMs were found to be useful tools for identifying the origin of subjects from particular continents and to correct for PS in admixed population sample sets. Their findings can be used as general guidelines for the application of specific AIM subsets. The researchers concluded that Taqman assays could be used for the selected AIMs as a simple and relatively cheap tool to control for differences in continental ancestry when conducting association studies in ethnically diverse populations. Kosoy et al. reported that these 128 In4 AIMs and subsets of these SNPs are useful for characterizing sample sets from diverse population groups. Researchers can apply these markers either to identify those members of one continental population group from a particular study, or alternatively used to adjust for PS due to differences in continental population frequency in cases and controls.

1.4 Principal Components (PC)

Principal components (PC) analysis is an approach to correct for PS using methods that infer genetic ancestry (Menozzi and others, 1978; Novembre and Stephens, 2008; Patterson and others, 2006; Zhu and others, 2002). In 2006, Price reported that the EIGENSTRAT method, which is based on principal components analysis, could detect and correct for PS in genome-wide association studies (Price and others, 2006). PC analysis models ancestry differences between cases and controls along continuous axes of variation. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. The approach can easily be applied to disease studies with hundreds of thousands of markers. EIGENSTRAT was implemented as part of the EIGENSOFT package in December 2006. Researchers can get source code, documentation and executable program files for the EIGENSOFT package from Alkes Price’s web page (Price).

Chapter 2 Methodology

2.1 Dataset

2.1.1 Hapmap 3 and GAW17 Database

Hapmap 3 database

The International Hapmap project started in 2002 and is an international cooperation between Japan, the United Kingdom (UK), Canada, China, Nigeria, and the United States (USA). Its goals are to compare genetic sequences of people from different populations, to identify chromosomal regions with shared genetic variants, and to determine panels of tag SNPs across the whole genome. The Hapmap 3 database currently holds about 4 million SNP genotypes for the eleven populations listed in Table 1. Table 2 gives the distribution of founders and non-founders in Hapmap 3 across populations.

Table 1. The list of populations of Hapmap 3 database.

Populations		Note
1	**ASW**	African ancestry in Southwest USA
2	**CEU ***	Utah residents with Northern and Western European ancestry from the CEPH collection
3	**CHB ***	Han Chinese in Beijing, China
4	**CHD ***	Chinese in Metropolitan Denver, Colorado
5	**GIH**	Gujarati Indians in Houston, Texas
6	**JPT ***	Japanese in Tokyo, Japan
7	**LWK ***	Luhya in Webuye, Kenya
8	**MEX**	Mexican ancestry in Los Angeles, California
9	**MKK**	Maasai in Kinyawa, Kenya
10	**TSI ***	Toscani in Italia
11	**YRI ***	Yoruba in Ibadan, Nigeria

Note: * denotes a population that is also included in the GAW17 dataset.

Four populations are considered as the African group: ASW, LWK, MKK and YRI. Three populations are in the Asian group: CHB, CHD and JPT. Two populations are in the European group: CEU and TSI. The populations GIH and MEX are in none of the groups above according to researchers’ work.

Table 2. The distribution of Hapmap 3 participants.

Populations		Founder Counts	Non-founder Counts	Total Counts
1	**ASW**	53	34	**87**
2	**CEU ***	112	53	**165**
3	**CHB ***	137	0	**137**
4	**CHD ***	109	0	**109**
5	**GIH**	101	0	**101**
6	**JPT ***	113	0	**113**
7	**LWK ***	110	0	**110**
8	**MEX**	58	28	**86**
9	**MKK**	156	28	**184**
10	**TSI ***	102	0	**102**
11	**YRI ***	147	56	**203**
Overall Total		**1198**	**199**	**1397**

Note: * denotes a population that is also included in the GAW17 dataset. In Hapmap 3 database, six populations, CHB, CHD, GIH, JPT, LWK and TSI, are composed of unrelated individuals only (672 founders). The other five populations include both genetically unrelated individuals (526 founders) and their children (199 non-founders). I select the 1198 genetically unrelated founder participants from 11 populations to be analyzed in my sample.

I checked the genotype distribution and the extent of missing data for each of the 22 chromosomes in the Hapmap 3 dataset. Genotyping data was available for at least 99.7% of SNP genotypes for each chromosome as shown in Table 3.

Table 3. The number of markers and missing data information by chromosome in Hapmap 3 database.

Chromosome	*Number of Markers*	*Average Genotyping Rate*
1	*119487*	*99.7%*
2	*119502*	*99.7%*
3	*98971*	*99.7%*
4	*88135*	*99.7%*
5	*90368*	*99.7%*
6	*93671*	*99.7%*
7	*77377*	*99.7%*
8	*77111*	*99.7%*
9	*65251*	*99.7%*
10	*75616*	*99.7%*
11	*72993*	*99.7%*
12	*70482*	*99.7%*
13	*53293*	*99.7%*
14	*46655*	*99.7%*
15	*43309*	*99.7%*
16	*45778*	*99.7%*
17	*39329*	*99.7%*
18	*41942*	*99.7%*
19	*26953*	*99.7%*
20	*37159*	*99.7%*
21	*19802*	*99.7%*
22	*20649*	*99.7%*
Total	*1,423,833*	*99.7%*

Note: Genotyping data was available for at least 99.7% of SNP genotypes for each chromosome in Hapmap 3 database. There are almost one and a half million of markers recorded in Hapmap 3.

2.1.2 Sample

The Hapmap 3 Data contains the genotypes of 1397 individuals from 11 populations. It includes both individuals and small families consisting of one or two founders and children. I select the 1198 genetically unrelated founder participants from the 11 populations in Hapmap 3 database as my sample. The distribution of the populations is given in Table 2.

2.1.3 Genotype Data

For this research, I select 402,399 SNP markers from six chromosomes: chromosome 3, 6, 11, 12, 17 and 19. Chromosome 3 and 6 represent relatively large chromosomes. Chromosome 11 and 12 represent medium size chromosomes. Chromosome 17 and 19 represent smaller chromosomes. Chromosomes 1 and 2 are eliminated because they are too large.

The SNPs on these six selected chromosomes have low rates of missing genotypes and a large number of rare variants. Specifically, each of these SNPs has missing genotype rate less than 0.3%, as shown in Table 3. Each of these chromosomes has 924 or more SNPs with MAF<0.01 in GAW17 dataset as shown in Table 1 of Appendix A. Table 4 shows the number of markers for the six chromosomes chosen.

Table 4. Genotype data used in analysis.

Chromosome	*Number of Markers*
3	*98971*
6	*93671*
11	*72993*
12	*70482*
17	*39329*
19	*26953*
Total	***402,399***

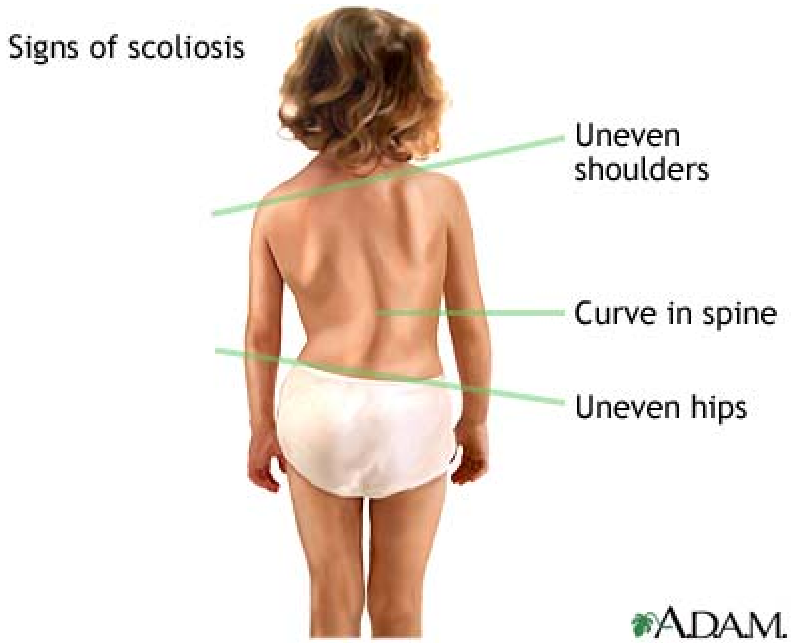
Genotype Data Cleaning

I checked the Hardy-Weinberg Equilibrium (HWE) condition on the disease SNPs, matching SNPs and non-correlated SNPs in my sample (see section 2.2 for more details) for the genetically unrelated individuals. A marker with the p-value for the HWE test less than 10^{-6} (that is, a highly significant deviation from HWE proportions) is removed using the HWE goodness of fit test command line options in PLINK v1.07. The PLINK commands are listed in Appendix B.

2.1.4 Phenotype Data

The longitudinal phenotype used my study is simulated to reflect the course of progression of a disease with every increasing severity such as adolescent idiopathic scoliosis (AIS). Following Gordon et al. (in press), I specify the trajectory curve parameters to model the development of the preliminary longitudinal scoliosis data (Wise et al. 2008). Figure 1 shows some of the symptoms of scoliosis. The Cobb angle of a patient is a quantitative longitudinal trait that has clinical relevance in that increasing Cobb angle indicates greater spinal deformity.

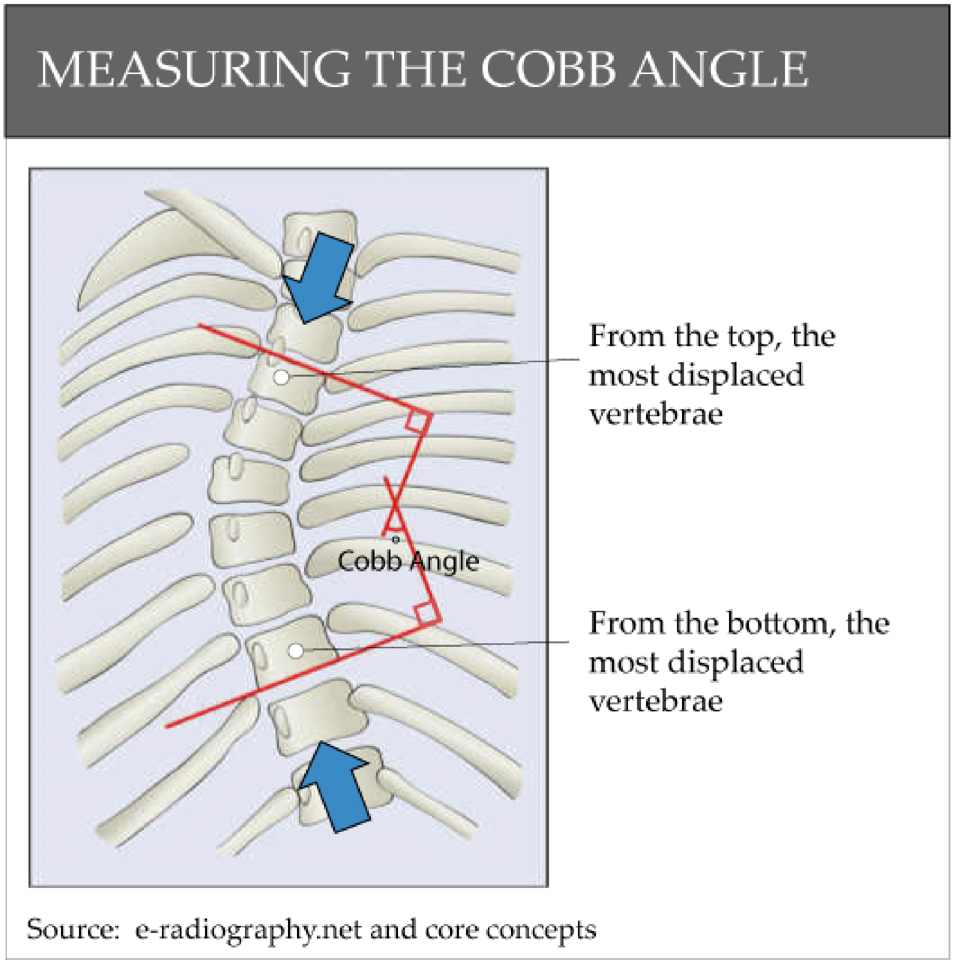
Figure 1. Signs of scoliosis.



Source: (Zieve, 2011)

The Cobb angle is measured by first identifying the upper and lower end vertebrae. Then lines are drawn extending the vertebral borders. The resulting angle is the Cobb angle and is measured as shown in Figure 2.

Figure 2. Measuring the Cobb angle.


MEASURING THE COBB ANGLE
Cobb Angle
From the top, the most displaced vertebrae
From the bottom, the most displaced vertebrae

Source: e-radiography.net and core concepts

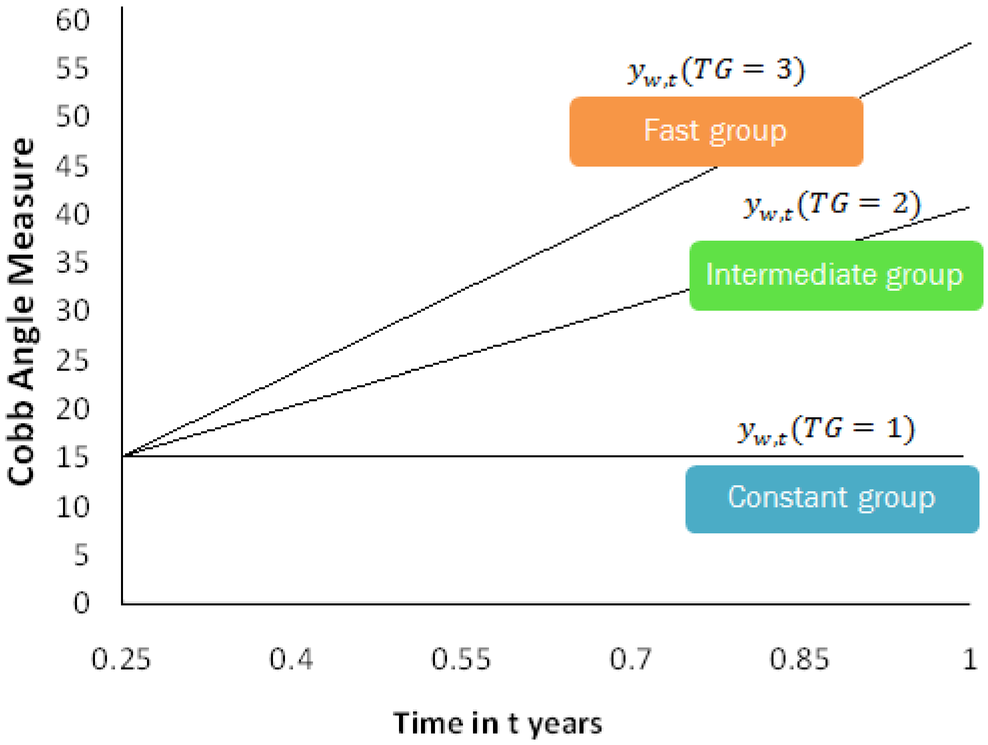
There are three linear trajectory groups based on a PROC TRAJ analysis (implemented in the SAS program). The linear growth mixture model with the three linear longitudinal trajectory equations for each participant w used in my simulation study is:

$$\boldsymbol{y_{w,t}} = \begin{cases} \boldsymbol{15 + N(0, \sigma^2)} & \boldsymbol{TG = 1} \\ \boldsymbol{15 + 28(t - 0.25) + N(0, \sigma^2)} & \boldsymbol{TG = 2} \\ \boldsymbol{15 + 56(t - 0.25) + N(0, \sigma^2)} & \boldsymbol{TG = 3} \end{cases} \qquad \textbf{Equation 1}$$

Here, $y_{w,t}$ refers to the Cobb angle of a participant w at time t for the trajectory group the participant was assigned to. The groups $TG = 1,2,3$ are modeled so that a participant w is in the constant, intermediate and fast groups, respectively. The genetic model (introduced in the next section) determines TG, the trajectory group. The time variable t ranges from 0.25 through 1 in intervals of 0.15 units. The random error follows a normal distribution with a mean of 0. The estimated standard deviation, σ, is set to 4 in my simulations.

Figure 3 presents the separation among the curves of each of the three polynomial trajectory functions in Equation 1 for one replicate.

Figure 3. The three trajectory curves for one replicate of simulated trajectory data ($n = 1198$).



Note: The group $TG = 3$ is the clinically important group. It is also called the fast group as it has the most rapid growth of the disease across time (with a slope of 56). The group $TG = 2$ has a slope of 28 and is the intermediate group. The group $TG = 1$ has no increase in the progression of the disease and is called the constant group.

2.2 Longitudinal Simulations

I use the Equation set 1 specified in section 2.1.4 to simulate longitudinal data for each participant. Each participant is assigned to one of the three trajectory groups: the slow (or constant), intermediate, or fast groups, according to the disease penetrance matrix (See section 2.3). The dependent variable at the last time point $t = 1$ presents the progression of the disease. A larger value indicates more rapid disease progression.

A SNP that generates the disease using the longitudinal trajectory functions is called a disease SNP or causal SNP. A SNP that is not related to the disease is called **a** non-causal SNP. I use non-causal SNPs results as the basis of my null simulations and use disease SNPs for my power simulations.

2.2.1 Null Simulations

The empirical type I error (false-positive rate) for any SNP is the proportion of p-values of the association test as given in PLINK that are less than the nominal significance level 0.05. The null hypothesis is that there is no association between this SNP's genotype and the participant's phenotype. That is, genotypes and phenotypes appear to be independent.

$$\begin{aligned} \boldsymbol{Type\ I\ error\ rate} &= \boldsymbol{P}(\boldsymbol{reject\ H_0}|\boldsymbol{H_0}) \\ &= \boldsymbol{P}(\boldsymbol{association\ test\ P-value} < \ 0.05|no\ association) \end{aligned}$$

Equation 2

For example, if a SNP matching one of the disease genes has 67 out of 1000 replicates significant in the association test at the 0.05 level, its estimated type I error rate is 0.067 ± 0.015.

For each of the combinations of population, penetrance, prevalence and PC setting, I generate a total of 1,000 replicates on the Hapmap dataset. The dependent variables in my statistical analyses are the empirical type I error rate, the empirical power and the lack of robustness of validity of the two statistics with or without 10 PCs adjustment. I define the lack of robustness of validity measure of a method on a null SNP as:

$$\boldsymbol{Lack\ of\ Robustness\ of\ Validity = \ (type\ I\ error\ rate - 0.05)^2} \quad \textbf{Equation 3}$$

A value of lack of robustness of validity close to 0 indicates that the type I error rate is close to the nominal value, while a larger value indicates a lack of robustness of validity.

In my study, I specify two null simulations according to the correlation between non-causal SNPs and the disease SNPs. Further information for the disease loci is given in the power simulation section. The Pearson correlation coefficients are calculated based on the MAFs of the non-causal SNPs across population using MATLAB software.

2.2.2.1 Pearson Correlation coefficient

In my research, I calculate the sample Pearson correlation coefficient between the non-causal SNPs and the disease SNPs across the 11 populations. For each pair of SNPs, I calculate eleven pairs of numbers that are population MAFs on 1198 participants. The definition is presented below.

$$\boldsymbol{r_{xy}} = \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2\sum_{i=1}^{m}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{11}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{11}(x_i-\bar{x})^2\sum_{i=1}^{11}(y_i-\bar{y})^2}} \qquad \textbf{Equation 4}$$

Here, x_i represents the MAF for population i of a non-causal SNP. y_i represents the MAF for population i of a disease SNP. $\bar{x}$ and $\bar{y}$ are the average MAF of the eleven population MAFs of the non-causal SNP and the disease SNP. s_x and s_y are the sample standard deviations of the eleven population MAFs of the non-causal SNP and the disease SNP.

2.2.2.2 Null Simulation I Using Uncorrelated SNPs

Under the null hypothesis, a participant's phenotype is independent of genotype. I calculate the matrix of correlation coefficients of the MAFs by populations with all the SNPs in my sample as the rows and the 18 disease SNPs as the columns. The MATLAB software is used for finding the correlation coefficients and the p-values. In my null simulation I, I identify a group of 25 SNPs whose MAF by population is least correlated with a disease locus. For example, for multi-locus simulations with 18 disease SNPs, I choose 450 different SNPs. Each set of 25 has the lowest absolute correlations with one of the 18 disease loci. The list of the low correlated SNPs is presented in the Appendix D.

2.2.2.3 Null Simulation II for SNPs Having MAF Correlated with Disease SNP

In my null simulation II, I identify a group of 25 SNPs with MAF by population most highly correlated with the MAFs of a disease locus. I call these "matching SNPs". They are the SNPs that might be confounded with the disease genes in analyses due to PS. For example, for multi-locus simulations with 18 disease SNPs, I choose 450 different SNPs that are most associated with the disease SNPs ($|r| \geq 0.99$). The list of the correlated matching SNPs is presented in the Appendix D.

Since population stratification is commonly considered an important confounding variable in a one sample study, I will apply the PC adjustment method to detect how effective this approach is in dealing with the PS problem. Specifically I focus on the extent to which the PC adjustment distinguishes SNPs having correlated population MAFs from the disease SNPs.

2.2.2 Power Simulations

Empirical power for a disease gene is defined here as the proportion of replicates that have p-value less than 0.05 for the disease gene. That is,

$$\begin{gathered}\boldsymbol{power} = \boldsymbol{P}(\boldsymbol{reject}\ \boldsymbol{H_0}|\boldsymbol{H_a}) \\ = \boldsymbol{P}(\boldsymbol{association}\ \boldsymbol{test}\ \boldsymbol{P-value} < \ 0.05|dependent/associated)\end{gathered}$$

Equation 5

I consider two types of scenarios under the alternative hypothesis, a single locus (gene) model and a model with multiple causal genes.

2.2.2.1 Single-locus models

I select the three loci with MAF 0.01 on chromosome 3 (representing the African populations), chromosome 17 (representing the Asian populations) and chromosome 11 (representing the European populations). I also select three loci on chromosome 3 representing the African populations with MAF 0.05, 0.15 and 0.30. All six selected single-locus disease SNPs are in apparent HWE. Each SNP has widely varying MAFs among the eleven populations in my sample database. Table 5 contains the list of the six single-locus disease SNPs.

Table 5. Selected single-locus disease gene.

MAF	Single-locus Disease Genes	Chromosome	Population
0.01	rs7355991	3	African
0.01	rs2073868	17	Asian
0.01	rs12790383	11	European
0.05	rs6792511	3	African
0.15	rs11924006	3	African
0.30	rs9810313	3	African

2.2.2.2 Multi-locus Simulations

I specify 18 rare disease SNPs that each has overall MAF less than 0.01 on chromosomes 3, 6, 11, 12, 17 and 19 for the multi-locus disease model. For each chromosome, three genes from three general populations (African, European and Asian respectively, as defined in section 2.2.1) are selected in my sample. Table 6 shows the distribution of the 18 disease genes by chromosome and general population. Table 7 contains the MAF statistics on these 18 SNPs. These SNP markers have p-value of HWE goodness of fit test greater than 0.10 and do not appear to deviate from HWE proportions.

Table 6. The 18 multi-locus simulation disease genes.

Chromosome	African	European	Asian
3	rs7355991	rs17195948	rs3733124
6	rs9459886	rs1259069	rs3761998
11	rs11825331	rs12790383	rs11217935
12	rs1696449	rs12822275	rs17117910
17	rs9899123	rs34742396	rs2073868
19	rs10411117	rs270771	rs3745465

Table 7. MAF by populations for 18 multi-locus simulation genes.

Chr	Disease SNPs	African				European		Asian			Indian	Mexican	Overall MAF
		ASW	LWK	MKK	YRI	CEU	TSI	CHB	CHD	JPT	GIH	MEX	
3	**rs7355991**	0.028	0.023	0.016	0.044	0	0	0	0	0	0	0	**0.011**
3	**rs17195948**	0.019	0	0	0	0.054	0.044	0	0	0	0	0.009	**0.010**
3	**rs3733124**	0	0	0	0	0	0	0.026	0.032	0.036	0	0.026	**0.010**
6	**rs9459886**	0.038	0.009	0.016	0.048	0	0	0	0	0	0	0	**0.010**
6	**rs1259069**	0.009	0	0	0	0.063	0.049	0	0	0	0	0.009	**0.011**
6	**rs3761998**	0	0	0	0.003	0	0	0.055	0.014	0.031	0	0	**0.011**
11	**rs11825331**	0.038	0.023	0.029	0.020	0	0	0	0	0	0	0	**0.010**
11	**rs12790383**	0	0	0	0	0.067	0.049	0	0	0	0	0.009	**0.011**
11	**rs11217935**	0	0	0	0	0	0	0.033	0.046	0.022	0	0	**0.010**
12	**rs17117910**	0	0	0	0	0	0	0.018	0.014	0.071	0	0	**0.010**
12	**rs12822275**	0	0	0	0	0.054	0.050	0	0	0	0.010	0	**0.010**
12	**rs1696449**	0.028	0.032	0.013	0.041	0	0	0	0	0	0	0	**0.011**
17	**rs9899123**	0.047	0.023	0.006	0.048	0	0	0	0	0	0	0	**0.011**
17	**rs2073868**	0	0	0	0	0.004	0	0.022	0.023	0.054	0	0	**0.010**
17	**rs34742396**	0	0	0	0	0.054	0.054	0	0	0	0.005	0.009	**0.010**
19	**rs3745465**	0	0	0	0	0	0	0.022	0.050	0.035	0	0	**0.010**
19	**rs10411117**	0.019	0.023	0.022	0.034	0	0	0	0	0	0	0	**0.010**
19	**rs270771**	0.009	0	0	0	0.058	0.054	0	0	0	0	0.009	**0.011**

2.3 Genetic Models

In my study, I used the genotype information of the selected disease SNPs as the principal determinant of trajectory group membership. I study both a single-locus gene model and models with multiple causal genes. In this section, I discuss models of complete penetrance with high prevalence, complete penetrance with low prevalence, partial penetrance with high prevalence, and partial penetrance with low prevalence.

2.3.1 Disease Prevalence and Penetrance Matrix

2.3.1.1 Disease Prevalence

The prevalence of a disease in epidemiology is defined as the proportion of cases in a population. That is, the number of individuals that are with the disease symptoms divided by the total number of people in the population. The law of total probability specifies the relation between the penetrance parameters and the trait prevalence. For example with G_1, G_2 and G_3 denoting the three SNP genotypes,

$$\boldsymbol{P(D) = P(D|G_1)P(G_1) + P(D|G_2)P(G_2) + P(D|G_3)P(G_3)} \qquad \textbf{Equation 6}$$

For the multi-locus model, a high prevalence model is considered in which there are two trajectory groups determined by the number of minor alleles. If there is any minor allele in a disease gene, the participant is assigned to the fast trajectory group. Otherwise, the participant is assigned to the constant trajectory group.

A low prevalence model is considered in which there are three trajectory groups determined by the number of minor alleles. If there are only a few minor alleles of the disease gene or genes for a participant, the participant is assigned to the intermediate trajectory group. If there are more minor alleles, the participant is assigned to the fast trajectory group. Otherwise, if there are no minor alleles of the disease genes for a participant, the participant is assigned to the constant trajectory group.

2.3.1.2 Penetrance Matrix

In genetics, trait penetrance of a genotype is defined as the conditional probability that a participant with the specified genotype has the trait being studied. A large penetrance value for a genotype indicates that an individual who has the genotype is likely to have the trait. Conversely, a small penetrance value means that an individual with the genotype is not likely to have the trait. For example, let *d* represent the total number of risk alleles in a given disease gene, and TG represent the trajectory group an individual is assigned to. Here, $TG = 3$ indicates that the individual is assigned to the group with fast disease development. Then, an example of complete penetrance for participants with two minor alleles is that $P(TG = 3|d = 2) = 1$.

The penetrance matrix for specified disease gene or genes is defined as:

$$\begin{pmatrix} \boldsymbol{P(TG=1|G \le B_1)} & \boldsymbol{P(TG=2|G \le B_1)} & \boldsymbol{P(TG=3|G \le B_1)} \\ \boldsymbol{P(TG=1|B_1 < G \le B_2)} & \boldsymbol{P(TG=2|B_1 < G \le B_2)} & \boldsymbol{P(TG=3|B_1 < G \le B_2)} \\ \boldsymbol{P(TG=1|G > B_2)} & \boldsymbol{P(TG=2|G > B_2)} & \boldsymbol{P(TG=3|G > B_2)} \end{pmatrix} =$$

$$\begin{pmatrix} \boldsymbol{1-\rho} & \boldsymbol{\frac{3}{4}\rho} & \boldsymbol{\frac{1}{4}\rho} \\ \boldsymbol{\frac{1}{2}\rho} & \boldsymbol{1-\rho} & \boldsymbol{\frac{1}{2}\rho} \\ \boldsymbol{\frac{1}{4}\rho} & \boldsymbol{\frac{3}{4}\rho} & \boldsymbol{1-\rho} \end{pmatrix} \qquad \textbf{Equation 7}$$

Here, G refers to the total number of risk alleles of the disease gene or genes for a participant. B_1 and B_2 are the limit bounds determined by the prevalence of the disease gene or genes. The penetrance parameter $\rho = 0$ indicates complete penetrance, $\rho = 0.1$ means a high penetrance and $\rho = 0.4$ means a low penetrance. For example, a complete penetrance matrix for a single-locus model is defined as:

$$\begin{pmatrix} \boldsymbol{P(TG=1|G=0)} & \boldsymbol{P(TG=2|G=0)} & \boldsymbol{P(TG=3|G=0)} \\ \boldsymbol{P(TG=1|G=1)} & \boldsymbol{P(TG=2|G=1)} & \boldsymbol{P(TG=3|G=1)} \\ \boldsymbol{P(TG=1|G=2)} & \boldsymbol{P(TG=2|G=2)} & \boldsymbol{P(TG=3|G=2)} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix} \qquad \textbf{Equation 8}$$

Further information is provided in the next section.

2.3.2 Additive Model

2.3.2.1 Single-locus Additive Model

First, I consider a single-locus additive model in which there are three trajectory groups determined by the effect of a single-locus gene. The most common trajectory group has intercept equal to 15, and slope equal to 0 (that is, $\beta_1 = 15$ and $\beta_2 = 0$). This group is called the flat or constant group. The second has intercept 15 and slope 28 (that is, $\beta_1 = 15$ and $\beta_2 = 28$); this group is called the intermediate group. The third has intercept 15 and slope 56 (that is, $\beta_1 = 15$ and $\beta_2 = 56$); this group is called the fast group. It is the clinically important group. The model is

$$\boldsymbol{y_{w,t}} = \begin{cases} \mathbf{15} + \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}) & \boldsymbol{TG} = \mathbf{1} \\ \mathbf{15} + \mathbf{28}(\boldsymbol{t} - \mathbf{0.25}) + \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}) & \boldsymbol{TG} = \mathbf{2} \\ \mathbf{15} + \mathbf{56}(\boldsymbol{t} - \mathbf{0.25}) + \boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}) & \boldsymbol{TG} = \mathbf{3} \end{cases} \qquad \textbf{Equation 9}$$

I consider three penetrance settings for a single-locus additive model: low, high and complete. As discussed above, the penetrance matrix is given by:

$$\begin{pmatrix} \boldsymbol{P}(\boldsymbol{TG} = \mathbf{1}|\boldsymbol{G} = \mathbf{0}) & \boldsymbol{P}(\boldsymbol{TG} = \mathbf{2}|\boldsymbol{G} = \mathbf{0}) & \boldsymbol{P}(\boldsymbol{TG} = \mathbf{3}|\boldsymbol{G} = \mathbf{0}) \\ \boldsymbol{P}(\boldsymbol{TG} = \mathbf{1}|\boldsymbol{G} = \mathbf{1}) & \boldsymbol{P}(\boldsymbol{TG} = \mathbf{2}|\boldsymbol{G} = \mathbf{1}) & \boldsymbol{P}(\boldsymbol{TG} = \mathbf{3}|\boldsymbol{G} = \mathbf{1}) \\ \boldsymbol{P}(\boldsymbol{TG} = \mathbf{1}|\boldsymbol{G} = \mathbf{2}) & \boldsymbol{P}(\boldsymbol{TG} = \mathbf{2}|\boldsymbol{G} = \mathbf{2}) & \boldsymbol{P}(\boldsymbol{TG} = \mathbf{3}|\boldsymbol{G} = \mathbf{2}) \end{pmatrix} =$$

$$\begin{pmatrix} \mathbf{1} - \boldsymbol{\rho} & \frac{\mathbf{3}}{\mathbf{4}}\boldsymbol{\rho} & \frac{\mathbf{1}}{\mathbf{4}}\boldsymbol{\rho} \\ \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\rho} & \mathbf{1} - \boldsymbol{\rho} & \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\rho} \\ \frac{\mathbf{1}}{\mathbf{4}}\boldsymbol{\rho} & \frac{\mathbf{3}}{\mathbf{4}}\boldsymbol{\rho} & \mathbf{1} - \boldsymbol{\rho} \end{pmatrix} \qquad \textbf{Equation 10}$$

Where $\rho \in [0,1]$. The penetrance matrix with parameter $\rho = 0$ indicates complete penetrance. That is, it is

$$\begin{pmatrix} P(TG=1|G=0) & P(TG=2|G=0) & P(TG=3|G=0) \\ P(TG=1|G=1) & P(TG=2|G=1) & P(TG=3|G=1) \\ P(TG=1|G=2) & P(TG=2|G=2) & P(TG=3|G=2) \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix} \qquad \textbf{Equation 11}$$

Here, a participant with major homozygote genotype ($G = 0$) is in the constant trajectory group with probability 1, in the intermediate increase group with probability 0 and in the fast increase group with probability 0. A participant with heterozygote genotype ($G = 1$) is in the intermediate increase group with probability 1, in the constant trajectory group with probability 0 and in the fast increase group with probability 0. A participant with minor homozygote genotype ($G = 2$) is in the fast increase group with probability 1, in the constant trajectory group with probability 0 and in the intermediate group 0.

The penetrance matrix with $\rho = 0.1$ is called the high penetrance model, that is:

$$\begin{pmatrix} \boldsymbol{P(TG=1|G=0)} & \boldsymbol{P(TG=2|G=0)} & \boldsymbol{P(TG=3|G=0)} \\ \boldsymbol{P(TG=1|G=1)} & \boldsymbol{P(TG=2|G=1)} & \boldsymbol{P(TG=3|G=1)} \\ \boldsymbol{P(TG=1|G=2)} & \boldsymbol{P(TG=2|G=2)} & \boldsymbol{P(TG=3|G=2)} \end{pmatrix} = \begin{pmatrix} \mathbf{0.9} & \mathbf{0.075} & \mathbf{0.025} \\ \mathbf{0.05} & \mathbf{0.9} & \mathbf{0.05} \\ \mathbf{0.025} & \mathbf{0.075} & \mathbf{0.9} \end{pmatrix} \qquad \textbf{Equation 12}$$

Here, a participant with major homozygote genotype ($G = 0$) is in the constant trajectory group with probability 0.9, in the intermediate increase group with probability 0.075 and in the fast increase group with probability 0.025. A participant with heterozygote genotype ($G = 1$) is in the intermediate increase group with probability 0.9, in the constant trajectory group with probability 0.05 and in the fast increase group with probability 0.05. A participant with minor homozygote genotype ($G = 2$) is in the fast increase group with probability 0.9, in the constant trajectory group with probability 0.025 and in the intermediate group 0.075.

The penetrance matrix with $\rho = 0.4$ is called the low penetrance model, that is:

$$\begin{pmatrix} \boldsymbol{P(TG = 1|G = 0)} & \boldsymbol{P(TG = 2|G = 0)} & \boldsymbol{P(TG = 3|G = 0)} \\ \boldsymbol{P(TG = 1|G = 1)} & \boldsymbol{P(TG = 2|G = 1)} & \boldsymbol{P(TG = 3|G = 1)} \\ \boldsymbol{P(TG = 1|G = 2)} & \boldsymbol{P(TG = 2|G = 2)} & \boldsymbol{P(TG = 3|G = 2)} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0.6} & \mathbf{0.3} & \mathbf{0.1} \\ \mathbf{0.2} & \mathbf{0.6} & \mathbf{0.2} \\ \mathbf{0.1} & \mathbf{0.3} & \mathbf{0.6} \end{pmatrix} \qquad \textbf{Equation 13}$$

This matrix indicates that a participant with major homozygote genotype will be in the constant trajectory group ($TG = 1$) with probability 0.60, in the intermediate group ($TG = 2$) with probability 0.30, and in the fast increase group ($TG = 3$) with probability 0.10. For a participant with heterozygote genotype, the probability of being in the intermediate trajectory group is 0.60, in the fast increase trajectory group 0.20, and in the constant trajectory group 0.20. For a participant with minor homozygote genotype, the probability of being in the fast increase trajectory group is 0.60, in the intermediate trajectory group 0.30, and in the constant group 0.10.

2.3.2.2 Multi-locus Additive Model

There are 1198 unrelated participants in this Hapmap data base, and 1187 have non-missing genotypes on the 18 disease genes. I calculate R, the number of minor alleles, for the 18 disease genes for each of the 1187 participants in my sample with complete genotype information. I use this as the number of risk alleles R.

The distribution of R from my sample is given below. As shown in Table 8, 833 participants (70.2%) have no minor alleles, 270 participants (22.8%) have only one minor allele and 84 participants (7.1%) have more than 2 minor alleles.

Table 8. The distribution of the risk alleles, R, for 18 disease genes.

R	**Frequency**	**Percentage (%)**	**Cumulative Percentage (%)**
0	833	70.18	70.18
1	270	22.75	92.92
2	75	6.32	99.24
3	8	0.67	99.92
4	1	0.08	100.00
Total	1187	100	----

Note: Frequency missing =11.

As in the single-locus model, each participant is assigned to a trajectory group according to the following general additive model and the penetrance matrix.

The model specifies trajectory class by

$$\begin{cases} \boldsymbol{TG = 1} & \boldsymbol{R = 0} \\ \boldsymbol{TG = 2} & \boldsymbol{R = 1} \\ \boldsymbol{TG = 3} & \boldsymbol{R} > 1 \end{cases}$$ **Equation 14**

The penetrance matrix for multi-locus disease genes model is the same as in single-locus disease genes model, given by

$$\begin{pmatrix} \boldsymbol{P(TG = 1|R = 0)} & \boldsymbol{P(TG = 2|R = 0)} & \boldsymbol{P(TG = 3|R = 0)} \\ \boldsymbol{P(TG = 1|R = 1)} & \boldsymbol{P(TG = 2|R = 1)} & \boldsymbol{P(TG = 3|R = 1)} \\ \boldsymbol{P(TG = 1|R} > 1) & \boldsymbol{P(TG = 2|R} > 1) & \boldsymbol{P(TG = 3|R} > 1) \end{pmatrix} =$$

$$\begin{pmatrix} \boldsymbol{1-\rho} & \boldsymbol{\frac{3}{4}\rho} & \boldsymbol{\frac{1}{4}\rho} \\ \boldsymbol{\frac{1}{2}\rho} & \boldsymbol{1-\rho} & \boldsymbol{\frac{1}{2}\rho} \\ \boldsymbol{\frac{1}{4}\rho} & \boldsymbol{\frac{3}{4}\rho} & \boldsymbol{1-\rho} \end{pmatrix}$$ **Equation 15**

where $\rho = 0$ for complete penetrance, $\rho = 0.1$ for high penetrance and $\rho = 0.4$ for low penetrance as described in the last section.

2.3.3 Complete Penetrance with High Disease Prevalence Model

There are only two trajectory groups in the complete penetrance high prevalence model. If there is any minor allele among the disease genes, the individual goes to the fast group. Otherwise, the individual is in the constant trajectory group.

$$\boldsymbol{G_j = \begin{cases} 1 & R \geq 1 \\ 0 & R = 0 \end{cases}} \qquad \textbf{Equation 16}$$

Here, j represents the participant and R is the number of minor alleles.

2.4 Methods

2.4.1 BPP Association Testing Method

After generating the longitudinal data, I apply the SAS TRAJ procedure to the simulated data. There were 1198 vectors of Y values, one vector for each participant in my small sample dataset. The vectors are the input to SAS PROC TRAJ analysis. I run PROC TRAJ with number of trajectory classes being 1, 2, 3, 4, 5 and 6, and a linear trajectory pattern for each trajectory class. I use the global maximum Bayesian Information Criterion (BIC) scores to select the number of trajectory classes. That is, I chose as the number of trajectory classes the model with the largest BIC score.

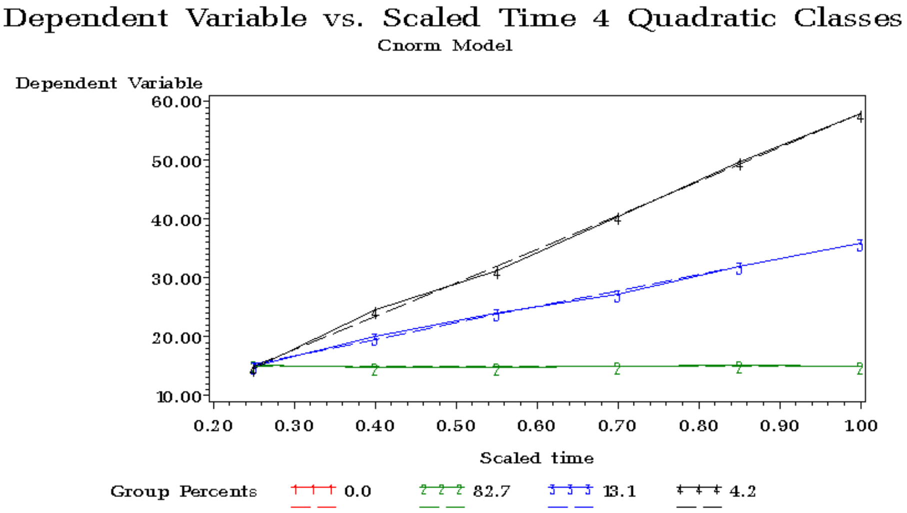
For example, I analyzed one replicate with SNP rs13322354 on Chromosome 3 as the disease gene. The data set contained 1198 subjects observed at 6 time points. There were three trajectory groups in the data for this replicate. The largest BIC score came from four trajectory classes. Table 9 presents the PROC TRAJ BIC scores and estimated trajectory group prevalence for settings of number of trajectory classes from 1 to 6.

Table 9. BIC scores with estimated trajectory class prevalence for marker rs13322354 on chromosome 3.

Trajectory Groups	**Linear Order**	
	BIC	**Group Membership (%)**
1	-12714.10	100.00
2	-11172.64	82.71/17.29
3	-11185.48	0.00/82.71/17.29
4	**-10707.72**	**0.00/82.71/13.05/4.24**
5	-10720.55	0.00/0/82.71/13.05/4.24
6	-10731.57	0.00/0.75/81.95/0/13.05/4.24

The graph of the four trajectories is presented in Figure 4. The four group model had three trajectory groups with non-zero fractions. Each model with four or more groups had only three groups with percentage greater than 1%. Note that only three groups had participants assigned in Figure 4.

Figure 4. Four trajectory groups plot for marker rs13322354 on chromosome 3.



Note: There are 0.0% of participants found in trajectory group 1.

The corresponding contingency table of the four trajectories with the genotype of the disease gene is presented in Table 10.

Table 10. Contingency row percentage table of genotype by trajectory group.

G	Trajectory Group Membership (%)			
	1 (flat)	**2 (flat)**	**3 (intermediate)**	**4 (fast)**
0	0.00	90.70	6.62	2.68
1	0.00	0.00	87.50	12.50
2	0.00	0.00	16.67	83.33

Here, trajectory group 4 is clinically important.

Having chosen the number of classes, I select the trajectory group that is most important clinically. It is the one with the fastest increase in expected value of Y.

Follow Datta and Satten et al. (2000), I use the Bayesian Posterior Probability (BPP) of being in the most clinically important group as my quantitative trait in the association test in PLINK. The p-values of the association test are recorded. The procedures were repeated for each of 1000 replicates.

Here the BPP method instead of modal BPP is used for assigning individuals to a particular trajectory group because Lubke and Muthen (2007) documented that modal BPP assignment has low accuracy for simple models with groups that are close together.

For each scenario and model in my study, I apply an association test with the BPP of the clinically important group as the quantitative trait. Then I apply the linear regression PC adjustment method to account for PS. The procedure is done by a SAS program and PLINK as given in Appendix B.

2.4.2 PC Adjustment for PS and Linea Regression

I use EIGENSTRAT software in EIGENSOFT to calculate 10 PCs. All 402,399 SNPs on the six chromosomes and 1198 unrelated participants are used. These 10 PCs and the SNP genotypes are considered as independent variables. The simulated quantitative traits are recorded as the phenotypes. A linear regression model is fit using PLINK v1.07 command lines to evaluate the association between each SNP studied and the phenotype. A SNP is considered associated with the quantitative trait phenotype when the p-value of the coefficient of the SNP is less than 0.05.

The original model fit to the quantitative trait phenotypes using an association test in PLINK is:

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 SNP_{ij} + \varepsilon_{ij}} \qquad \textbf{Equation 17}$$

The PC adjustment model fit to the quantitative trait phenotypes using a linear regression option in PLINK is:

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 SNP_{ij} + \beta_2 PC_1 + \beta_3 PC_2 + \cdots + \beta_{11} PC_{10} + \varepsilon_{ij}} \qquad \textbf{Equation 18}$$

Here, j represents the jth individual. i represents the ith SNP, and $\mathrm{PC_i}$ represents the *i*th global principal component.

2.5 Factorial Designs

2.5.1 PC Adjustment

For each simulation setting, I have two levels of the PC adjustment setting (no PC adjustment and with PC adjustment) using 10 PCs calculated from the entire sample space to account for population stratification.

2.5.2 Disease Penetrance

I specify three levels of disease penetrance for each simulation. They are low, high and complete penetrance. A model with low or high penetrance is also called a reduced or partial penetrance model. A model with complete penetrance is sometimes called a complete penetrance model. See descriptions in section 2.3 genetic models.

2.5.3 Disease Prevalence

I specify two disease prevalence models for each of the single-locus and multi-locus models. One is the high prevalence model. That is, if an individual has at least one minor allele among the genes, then the participant is in the fast trajectory group under a complete penetrance model. Otherwise, the participant is in the constant trajectory group. See descriptions in section 2.3 Genetic Models. I also specify low prevalence models. There are three possible trajectory groups in a low prevalence model: constant, intermediate and fast groups. A participant is assigned to one of the three groups according to the penetrance matrix and the total number of minor alleles R.

2.5.4 General Population

For both the single-locus and multi-locus model, I consider a three level population factor. The three levels are African, Asian and European.

2.5.5 MAF (for single-locus model only)

For a single-locus model, I also consider an overall MAF factor which has four levels: 0.01, 0.05, 0.15 and 0.30.

2.6 Experiments

I consider three experiments under the single-locus model and two experiments under the multi-locus model for predicting (y) the robustness of validity of null simulations and the rejection rates of the power simulations. Table 11 below shows the settings for each experiment.

Table 11. Factors used in the single and multi-locus model experiments table.

Settings	Degree of Freedom	Single-locus Experiments			Multi-locus Experiments	
		Exp I	Exp II	Exp III	Exp IV	Exp V
PC	1	+	+	+	+	+
Penetrance	2	+	+	-	+	-
Population	2	+	-	-	+	+
MAF	2	-	+	+	-	-
Chromosome	5	-	-	-	+	+

Note: “+”: the settings are in an experiment; “-”: the settings are not in an experiment.

2.6.1 Single-locus Experiments

Experiment I has three single-locus disease SNPs with $MAF \approx 0.01$ on chromosome 3 representing the African population, chromosome 11 representing the European population, and chromosome 17 representing the Asian population factors. Experiment I has three factors: population, penetrance and use of PC adjustment. The population factor, x_1, has three levels (Af = African, As = Asian, Eu = European). The penetrance factor, x_2, has three levels (low, high and complete). The use of PC factor, x_3, has two levels (nopc = no PC adjustment, pc = 10 PCs adjustment). Chromosome is not considered here as a factor because it is confounded with population. The model also contains all two factor interactions. I fit the generalized linear model:

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \varepsilon} \qquad \textbf{Equation 19}$$

Experiment II has three single-locus disease SNPs with MAF = 0.05, 0.15 and 0.30 on chromosome 3 representing the African population using an additive model. Experiment II has three factors: MAF, penetrance and use of PC adjustment. The MAF factor, x_1, has three levels ($MAF = 0.05, 0.15, 0.30$). The penetrance factor, x_2, has three levels (low, high and complete). The use of PC, x_3, has two levels (nopc = no PC adjustment, pc = 10 PCs adjustment). The model also contains all two factor interactions. I fit the generalized linear model given in Equation 19 above.

Experiment III uses the high prevalence complete penetrance model with the same SNPs as in experiment II. Experiment III has two factors: MAF and use of PC adjustment. The MAF factor, x_1, has three levels ($MAF = 0.05, 0.15, 0.30$). The use of PC, x_2, has two levels (nopc = no PC adjustment, pc = 10 PCs adjustment). The model also contains the interaction between MAF and PC. I fit the generalized linear model:

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon} \qquad \textbf{Equation 20}$$

2.6.2 Multi-locus Experiments

Experiment IV has 18 multi-locus disease SNPs with MAF = 0.01 on chromosome 3, 6, 11, 12, 17 and 19 representing the African, Asian and European population. The additive model is used for experiment IV. Experiment IV has four factors: population, penetrance, chromosome and use of PC adjustment. The population factor, x_1, has three levels (Af = African, As = Asian, Eu = European). The penetrance factor, x_2, has three levels (low, high and complete). The use of PC factor, x_3, has two levels (nopc = no PC adjustment, pc = 10 PCs adjustment). The chromosome factor, x_4, has six levels (chromosome 3, 6, 11, 12, 17 and 19). The model also contains all two factor interactions. I fit the generalized linear model:

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3 + \beta_9 x_2 x_4 + \beta_{10} x_3 x_4 + \varepsilon} \qquad \textbf{Equation 21}$$

Experiment V has the same settings but using the high prevalence complete penetrance model. Experiment V has three factors: population, chromosome and use of PC adjustment. The population factor, x_1, has three levels ($MAF = 0.05, 0.15, 0.30$). The chromosome factor, x_4, has six levels (chromosome 3, 6, 11, 12, 17 and 19). The use of PC, x_2, has two levels (nopc = no PC adjustment, pc = 10 PCs adjustment). The model also contains the interaction between population and PC. I fit the generalized linear model given in Equation 19 above.

2.7 Software

I used four programs in my study: PLINK, MATLAB, SAS PROC TRAJ, and EIGENSTRAT as described below.

2.7.1 PLINK

PLINK is a freely available program with open-source code. It is a whole genome association analysis toolset designed to perform a range of basic, large-scale analyses in a computationally efficient manner. The focus of **PLINK** is purely on the analysis of genotype/phenotype data.

I use PLINK v1.07 software to extract the data on the six chromosomes selected. I also use PLINK to perform the association analysis of the quantitative phenotype extracted from the SAS PROC TRAJ analysis described below. Specifically, I use the association and linear association functions in PLINK to analyze the data. PLINK –assoc option calculates the P-value of the association chi-square test for the disease SNPs selected and also the matching and non-correlated SNPs without any PC adjustments. PLINK –linear option calculates the P-value of linear association test for the disease SNPs selected and also the matching and non-correlated SNPs with 10 PCs added as covariates.

2.7.2 MATLAB

MATLAB is a language for technical computing. It is a programming environment for algorithm development, data analysis, visualization, and numerical computation. It can solve technical computing problems faster than the traditional programming languages, such as C, C++, and FORTRAN. I use MATLAB to calculate the correlation coefficients matrix with 402,399 rows and 18 columns. Here, the 402,399 rows are all the SNPs' MAFs by population. The 18 columns represent the 18 multi-locus disease SNPs.

2.7.3 *SAS* TRAJ procedure

SAS TRAJ procedure is widely used to model longitudinal data (Jones et al. 2001). For each replicate, each participant is assigned to a particular trajectory group according to the genotype data of the simulation (null or alternative) and scenario (single or multi-locus) model described in the above sections. A total of 1000 replicates are generated for each simulation under each scenario. The longitudinal data is generated using one of the three linear equations mentioned in 2.1.4 according to the trajectory group a participant is in. There are n vectors of Y values, one vector for each participant in the sample. Here, n is the sample size. The vectors are the input to SAS PROC TRAJ analysis.

I then perform PROC TRAJ analysis with number of trajectory classes being 1, 2, 3, 4, 5 and 6, and a linear trajectory pattern for each trajectory class to estimate each model. The time points range from 0.25 through 1 in intervals of 0.15 and are used as independent variables. I use the global maximum BIC score to select the number of trajectory classes. That is, I choose as the number of trajectory classes the model with the largest BIC score. Other selection rules are used in practice.

The clinically important group is identified as the group with highest disease progression speed, that is, the one with the greatest slope. The BPP of the clinically important group is recorded to be used in the PLINK software as the quantitative trait phenotype input.

2.7.4 EIGENSTRAT

Price et al. (2006) created the EIGENSTRAT stratification correction software. The software EIGENSOFT 4.2 performs the computations and can be downloaded from the website: http://www.hsph.harvard.edu/faculty/alkes-price/software/. The software uses principal components analysis to detect and model ancestry differences, and correct for population stratification in genome-wide association studies. It supports several file formats including the PLINK PED format. The PC adjustment is specific to a candidate marker's variation in frequency across ancestral populations. It can be applied to disease studies with hundreds of thousands of markers.

I use CONVERTF software in EIGENSOFT to convert PLINK PED format data sets to EIGENSTRAT formats. Then I use EIGENSTRAT software to calculate 10 PCs using all 402,399 SNPs on the six chromosomes in my sample.

Chapter 3 Results

3.1 Single-locus Model Results

3.1.1 Experiment I Results

Table 11 presents the average empirical type I error rate and empirical power observed in the AIS simulations under experiment I, which has disease SNP MAF = 0.01, representing African, Asian and European population as described in Chapter 2 Methodology. The table also includes 95% confidence intervals. Each single-locus disease SNP in the table has 25 uncorrelated SNPs (MAFs across population less correlated with the disease SNP) and 25 matching SNPs (MAFs across population as correlated as possible with the disease SNP MAFs).

The empirical type I error rate using the uncorrelated SNPs has a lack of robustness of validity appearing in the complete penetrance without PC adjustment model. For example, on Table 11, using nominal level of significance 0.05 and analyzing data from the complete penetrance model, 16% (**0.160 ±0.005**) of the replicates for which the 25 SNPs uncorrelated with the disease SNP rs7355991 on chromosome 3 representing the African population are found to be significant in the association test when there is no PC adjustment. With PC adjustment, the number of replicates that are significant at the nominal 0.05 level in a linear regression test decreases to 4% (**0.040 ±0.002**).

The empirical type I error rate using the matching SNPs shows a robustness of validity for the low penetrance settings both with and without PC adjustment. For example, the 25 SNPs matching the disease SNP rs7355991 on chromosome 3 representing African population have an average type I error rate 0.053 ±0.003 without PC adjustment at the nominal 0.05 level. With PC adjustment, the rate is 0.052 ±0.003. There is a failure of robustness of validity for complete penetrance both with and without PC adjustment. For example, the 25 SNPs that match the disease SNP rs7377991 have an average type I error rate **0.560 ±0.006** without PC adjustment. With PC adjustment, the rate decreases to **0.280 ±0.006**, but is still far above the nominal 0.05 level. In general, PC adjustment improves robustness of validity.

The rejection rate for disease SNPs with PC adjustment is close to the rate without

PC adjustment. For example, the rejection rate without PC adjustment is **0.549 ±0.031** for the disease SNP rs2073868 on chromosome 17 representing the Asian population at the nominal 0.05 level. With PC adjustment, the rejection rate is **0.523 ±0.031**.

Table 12. **Empirical rejection rates with 95% confidence interval for the single-locus model under experiment I (disease SNP MAF = 0.01) at the nominal 0.05 level (1000 replicates)**

MAF	Single-locus Disease Genes	Chr	Pop	Penetrance					
				Complete		High		Low	
				No PC	PC	No PC	PC	No PC	PC
				Uncorrelated SNPs (null I)					
0.01	rs7355991	3	Af	**0.160 ±0.005**	**0.040 ±0.002**	0.053 ±0.003	0.053 ±0.003	0.051 ±0.003	0.050 ±0.003
	rs2073868	17	As	**0.120 ±0.004**	**0.040 ±0.002**	**0.055 ±0.003**	0.053 ±0.003	0.050 ±0.003	0.052 ±0.003
	rs12790383	11	Eu	**0.000 ±0.000**	**0.040 ±0.002**	**0.045 ±0.003**	0.047 ±0.003	0.050 ±0.003	0.051 ±0.003
				Matching SNPs (null II)					
0.01	rs7355991	3	Af	**0.560 ±0.006**	**0.280 ±0.006**	**0.074 ±0.003**	**0.059 ±0.003**	0.053 ±0.003	0.052 ±0.003
	rs2073868	17	As	**0.520 ±0.006**	**0.160 ±0.005**	**0.079 ±0.003**	**0.067 ±0.003**	**0.055 ±0.003**	0.051 ±0.003
	rs12790383	11	Eu	**0.520 ±0.006**	**0.440 ±0.006**	**0.070 ±0.003**	**0.065 ±0.003**	0.052 ±0.003	0.049 ±0.003
				Disease SNPs (Power)					
0.01	rs7355991	3	Af	1.000 ±0.000	1.000 ±0.000	0.259 ±0.027	0.248 ±0.027	0.362 ±0.030	0.352 ±0.030
	rs2073868	17	As	1.000 ±0.000	1.000 ±0.000	0.716 ±0.028	0.729 ±0.028	0.549 ±0.031	0.523 ±0.031
	rs12790383	11	Eu	1.000 ±0.000	1.000 ±0.000	0.165 ±0.023	0.173 ±0.023	0.236 ±0.026	0.232 ±0.026

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = General populations, including African (Af), Asian (As) and European (Eu). The complete penetrance model is with high prevalence (two trajectory groups) since the MAF = 0.01 is too small. The high and low penetrance models are with low prevalence (additive model is used, three trajectory groups). See Methodology – Genetic Models. Type I error rates that are significantly different from the nominal 0.05 level are in bold. For each setting 1000 replicates are generated. Africans n = 466, Asians n = 359, Europeans n = 214.

3.1.1.1 Null I Simulation Using Uncorrelated SNPs

The ANOVA of the measure of lack of robustness of validity for the uncorrelated SNPs is shown in Table 12 below. The model has an *R-square* of 0.8924. These results indicate that the overall model is not statistically significant ($F = 2.55,\ p - value = 0.1892$). Because the smallest *p-value* in the ANOVA is the *p-value* for the factor penetrance and is equal to 0.0770, I conclude that there are no significant factors for the lack of robustness of validity. That is, the statistical analysis did not confirm the apparent failure of robustness of validity for data from the complete penetrance model analyzed without PC adjustment.

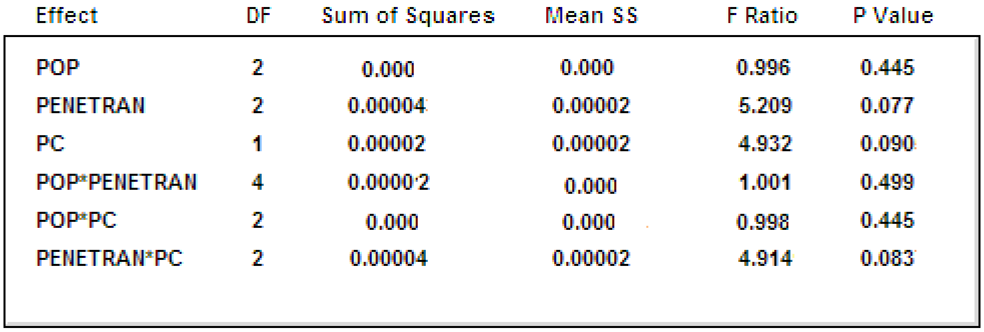
Table 13. The ANOVA table for the single-locus model uncorrelated SNPs under experiment I at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lack of robustness of validity

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	0.00014	0.000	2.55	0.189
Error	4	0.00002	0.000		
Corrected Total	17	0.00015			

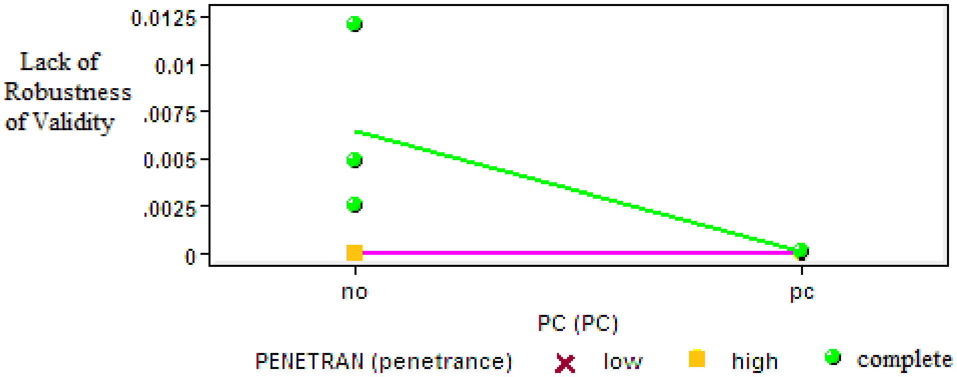
R-Square	Coeff Var	Root MSE	Lack of Robustness of Validity Mean
0.892393	184.658	0.002	0.001



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
POP	2	0.000	0.000	0.996	0.445
PENETRAN	2	0.00004	0.00002	5.209	0.077
PC	1	0.00002	0.00002	4.932	0.090
POP*PENETRAN	4	0.00002	0.000	1.001	0.499
POP*PC	2	0.000	0.000	0.998	0.445
PENETRAN*PC	2	0.00004	0.00002	4.914	0.083

To assist in the interpretation of these results, Figure 5 presents the plot of PC and penetrance interaction. The lack of robustness of validity appears in the complete penetrance model without PC adjustment.

Figure 5. Lack of robustness of validity of single-locus model uncorrelated SNPs under experiment I at the nominal 0.05 level.



3.1.1.2 Null II Simulation Using Matching SNPs

The ANOVA for the lack of robustness measure for the matching SNPs is shown in Table 13 below. The model has an *R-square* of 0.9702. These results indicate that the overall model is statistically significant ($F = 10.02,\ p - value = 0.0194$). The value of the statistic $F = 12.224$ ($p - value = 0.0250$) for the factor PC and $F = 43.585$ ($p - value = 0.0019$) for the factor penetrance are statistically significant. Furthermore, because the test statistic for the interaction of PC and penetrance is $F = 12.131$ ($p - value = 0.0200$), I conclude that this interaction is also significant. This analysis documents that the use of PC adjustment significantly improves robustness of validity.

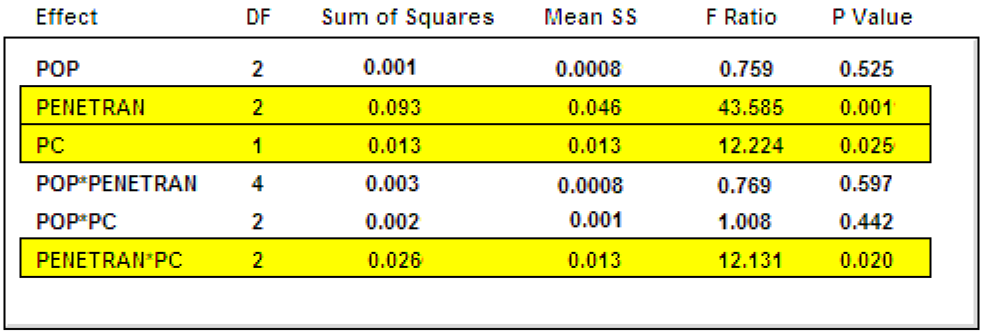
Table 14. The ANOVA table for the single-locus model matching SNPs under experiment I at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lack of robustness of validity

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	0.140	0.011	10.02	0.019
Error	4	0.004	0.001		
Corrected Total	17	0.144			

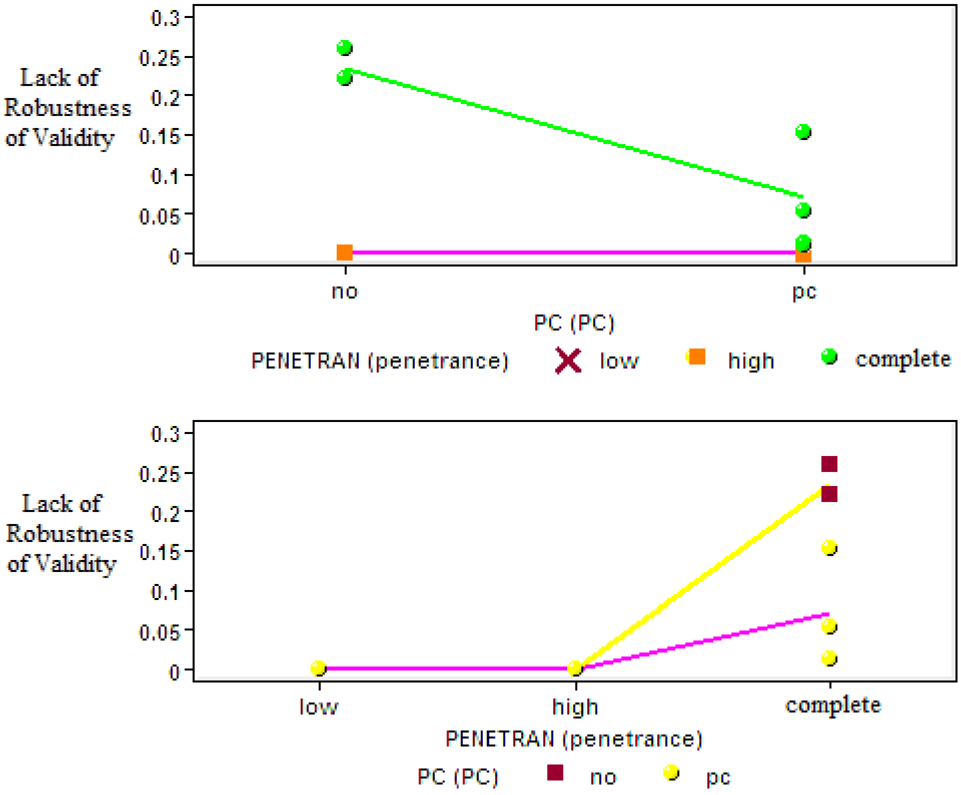
R-Square	Coeff Var	Root MSE	Lack of Robustness of Validity Mean
0.970209	64.003	0.033	0.051



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
POP	2	0.001	0.0008	0.759	0.525
PENETRAN	2	0.093	0.046	43.585	0.001
PC	1	0.013	0.013	12.224	0.025
POP*PENETRAN	4	0.003	0.0008	0.769	0.597
POP*PC	2	0.002	0.001	1.008	0.442
PENETRAN*PC	2	0.026	0.013	12.131	0.020

For interpreting the results, I present the graph of the average lack of robustness of validity measure of the empirical type I error rate at each combination of the penetrance and PC in Figure 6. The lack of parallelism of the lines indicates that there is a significant interaction between these two factors. In general, better robustness of validity is attained at low and high penetrance, regardless of the use of PC adjustment. With PC adjustment, the robustness of validity improves for the complete penetrance model. PC adjustment should be applied since it maintains better robustness of validity as penetrance level changes.

Figure 6. Lack of robustness of validity of single-locus model matching SNPs under experiment I at the nominal 0.05 level.



3.1.1.3 Power Simulation Using single-locus disease SNPs

The ANOVA of the rejection rates for the disease SNPs is shown in Table 14 below. The model has an *R-square* of 0.9999. These results indicate that the overall model is statistically significant ($F = 2605.22,\ p - value < 0.0001$). Since the interaction of population and penetrance has $\mathrm{F} = 774.95$ ($\mathrm{p - value} < 0.0001$), I conclude that there is a significant interaction between population and penetrance. Furthermore, $\mathrm{F} = 2256.5$ ($\mathrm{p - value} < 0.0001$) for population and $\mathrm{F} = 13125$ ($\mathrm{p - value} < 0.0001$) for penetrance. Hence the main effects of population and penetrance are also significant. The most important finding is that the use of PC adjustment and interactions involving PC adjustment are not significant. That is, the fitted rejection rate with PC adjustment is statistically equal to the fitted rate without PC adjustment.

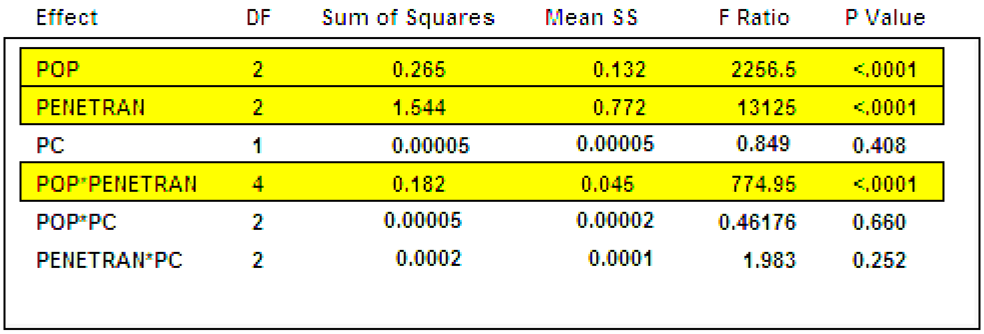
Table 15. The ANOVA table for the single-locus model disease SNPs under experiment I at the nominal 0.05 level.

The GLM Procedure

Dependent Variable: rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	1.993	0.153	2605.22	<.0001
Error	4	0.0002	0.0001		
Corrected Total	17	1.993			

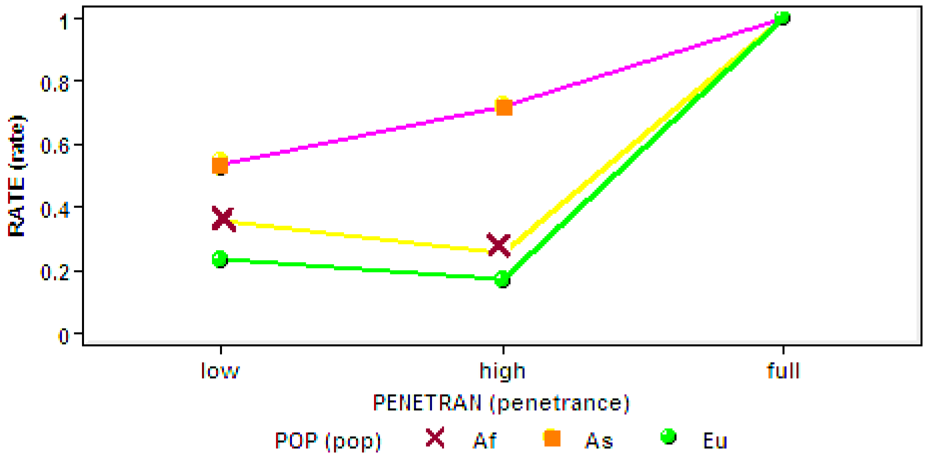
R-Square	Coeff Var	Root MSE	rate Mean
1.000	1.309	0.008	0.586



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
POP	2	0.265	0.132	2256.5	<.0001
PENETRAN	2	1.544	0.772	13125	<.0001
PC	1	0.00005	0.00005	0.849	0.408
POP*PENETRAN	4	0.182	0.045	774.95	<.0001
POP*PC	2	0.00005	0.00002	0.46176	0.660
PENETRAN*PC	2	0.0002	0.0001	1.983	0.252

For interpreting the results, I present the graph of the average empirical power at each combination of the population and penetrance in Figure 7. The lack of parallelism of the lines indicates that there is a significant interaction between these two factors. In general, the power of the Asian population is greater than that of the African population. The power of the African population is also greater than that of the European population. Higher power is attained at complete penetrance, regardless of population.

Figure 7. Power of Single-locus model disease SNPs under experiment I at the nominal 0.05 level.



3.1.1.4 Statistical Analysis on Experiment I

Table 15 reports the averages of the rejection rate for the disease SNPs and lack of robustness of validity measure with or without PC adjustment. Each entry is the average over nine settings (three disease models time three penetrance settings). The results document the value of PC adjustment. When power is less than 1, PC adjustment is effective and has lower but still strong power. Without PC adjustment, the type I error rate cannot be controlled. Overall, the type I error rate with PC adjustment has substantial improvement but is significantly higher than the nominal significance level for matching SNPs.

Table 16. Average rejection rates and lack of robustness of validity on Factor PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
Uncorrelated SNPs (null I)	0.065 ±0.001	0.047 ±0.001	0.029 ±0.001	0.013 ±0.001
Matching SNPs (null II)	0.220 ±0.002	0.136 ±0.001	0.162 ±0.002	0.089 ±0.001
Disease SNPs (Power)	0.587 ±0.002	0.584 ±0.002		

I also calculate the average of type I error, rates power and lack of robustness of validity with or without PC by the three general populations: African, Asian and European. My hypothesis is that African population needs PC adjustment more than European population but less than Asian population because of the population immigration. The results in Table 16 indicate that global PC method adjust the type I error rate well overall, especially for the African and Asian population, but less effectively for the European population. The Asian population has the best power.

Table 17. Average rejection rates and lack of robustness of validity on Factors population and PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
African				
Uncorrelated SNPs (null I)	0.088 ±0.001	0.048 ±0.001	0.049 ±0.001	0.013 ±0.001
Matching SNPs (null II)	0.229 ±0.002	0.130 ±0.002	0.169 ±0.002	0.085 ±0.001
Disease SNPs (Power)	0.540 ±0.003	0.533 ±0.003		
Asian				
Uncorrelated SNPs (null I)	0.075 ±0.001	0.048 ±0.001	0.037 ±0.001	0.013 ±0.001
Matching SNPs (null II)	0.218 ±0.002	0.093 ±0.001	0.158 ±0.002	0.050 ±0.001
Disease SNPs (Power)	0.755 ±0.002	0.751 ±0.002		
European				
Uncorrelated SNPs (null I)	0.032 ±0.001	0.046 ±0.001	0.001 ±0.000	0.013 ±0.001
Matching SNPs (null II)	0.214 ±0.002	0.185 ±0.002	0.157 ±0.002	0.133 ±0.002
Disease SNPs (Power)	0.467 ±0.003	0.468 ±0.003		

I further calculate the average of the type I error rate, power and lack of robustness of validity with or without PC adjustment by the three penetrance levels: low, high and complete. As shown in Table 17, the reduced penetrance models (low and high) maintain a correct type I error rate compared to the rates for the complete penetrance models. The complete penetrance model has a larger power than the reduced penetrance models.

Table 18. Average rejection rates and lack of robustness of validity on Factors penetrance and PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	**PC**	**No PC**	**PC**
Low Penetrance				
Uncorrelated SNPs (null I)	0.050 ±0.001	0.051 ±0.001	0.000 ±0.000	0.000 ±0.000
Matching SNPs (null II)	0.053 ±0.001	0.050 ±0.001	0.000 ±0.000	0.000 ±0.000
Disease SNPs (Power)	0.382 ±0.002	0.369 ±0.002		
High Penetrance				
Uncorrelated SNPs (null I)	0.050 ±0.001	0.051 ±0.001	0.000 ±0.000	0.000 ±0.000
Matching SNPs (null II)	0.074 ±0.001	0.064 ±0.001	0.002 ±0.000	0.001 ±0.000
Disease SNPs (Power)	0.380 ±0.002	0.383 ±0.002		
Complete Penetrance				
Uncorrelated SNPs (null I)	0.093 ±0.001	0.040 ±0.001	0.087 ±0.001	0.038 ±0.001
Matching SNPs (null II)	0.533 ±0.003	0.293 ±0.002	0.483 ±0.003	0.267 ±0.002
Disease SNPs (Power)	1.000 ±0.000	1.000 ±0.000		

I calculate the average of type I error, power and lack of robustness of validity with or without PC by the two prevalence levels: low and high. The low prevalence model maintains a correct type I error rate than the high prevalence model while the high prevalence model has a larger power.

Table 19. Average rejection rates and lack of robustness of validity on Factors prevalence and PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
Low Prevalence				
Uncorrelated SNPs (null I)	0.051 ±0.001	0.051 ±0.001	0.000 ±0.000	0.000 ±0.000
Matching SNPs (null II)	0.064 ±0.001	0.057 ±0.001	0.001 ±0.000	0.001 ±0.000
Disease SNPs (Power)	0.381 ±0.002	0.376 ±0.002		
High Prevalence				
Uncorrelated SNPs (null I)	0.093 ±0.001	0.040 ±0.001	0.087 ±0.001	0.038 ±0.001
Matching SNPs (null II)	0.533 ±0.003	0.293 ±0.002	0.483 ±0.003	0.267 ±0.002
Disease SNPs (Power)	1.000 ±0.000	1.000 ±0.000		

3.1.2 Experiment II Results

Table 20 presents the average empirical type I error rate and empirical power observed in the simulations under experiment II, which has disease SNP MAF = 0.05, 0.15, 0.30, representing the African population as described in Chapter 2 Methodology. The additive model is used in experiment II. The table also includes 95% confidence intervals. Each single-locus disease SNP in the table has 25 uncorrelated SNPs (MAFs across population less correlated with the disease SNP) and 25 matching SNPs (MAFs across population as correlated as possible with the disease SNP MAFs).

As in experiment I, the empirical type I error rate using the uncorrelated SNPs has a lack of robustness of validity appearing in the complete penetrance without PC adjustment analysis. The high penetrance model does not show much of lack of robustness of validity and does not need much PC adjustment. The low penetrance model does not show lack of robustness of validity and does not need any PC adjustment. For example, on Table 20, using nominal level of significance 0.05 and analyzing data from the complete penetrance model, there are 11.5% (**0.115 ±0.004**) of the replicates for which the 25 SNPs uncorrelated with the disease SNP rs6792511 on chromosome 3 representing the African population are found significant in the association test when there is no PC adjustment. With PC adjustment, the number of replicates that are significant at the nominal 0.05 level in a linear regression test decreases to 4% (**0.040 ±0.002**). There are two unexpected results using uncorrelated SNPs. Analyzing data from the complete penetrance model, the SNP rs11924006 with MAF 0.15 has a type I error rate 0 (0.000 ±0.000) with PC adjustment, which shows a lack of robustness of validity. From the low penetrance model, the SNP rs9810313 with MAF 0.30 has a type I error rate 13.9% (**0.139 ±0.004**) without PC adjustment.

PS becomes a bigger problem as MAF increases for the matching SNPs. For example, without PC adjustment, the disease SNP rs6792511 with a low MAF 0.05 has an average type I error rate **0.695 ±0.006**. The disease SNP rs11924006 with MAF 0.15 has an average type I error rate **0.920 ±0.003**. The average type I error rate increases to **0.960 ±0.002** as MAF increases to 0.30 for the disease SNP rs9810313. The empirical type I error rate shows a lack of robustness of validity for all the three penetrance settings: complete, high and low. PC adjustment does not help much using the complete and high penetrance models. For example, the 25 SNPs matching the disease SNP rs9810313 on chromosome 3 representing the African population have an average type I error rate **0.960 ±0.002** without PC adjustment at the nominal 0.05 level. With PC adjustment, the rate is **0.520 ±0.006**. Even the low penetrance model shows a problem using PC adjustment. For example, the 25 SNPs that match the disease SNP rs9810313 have an average type I error rate **0.779 ±0.005** without PC adjustment. With PC adjustment, the rate decreases to **0.335 ±0.006**, but is still far above the nominal 0.05 level. In general, PC adjustment improves robustness of validity.

The rejection rate for disease SNPs with PC adjustment is close to the rate without PC adjustment. For example, the rejection rate without PC adjustment is **0.849 ±0.022** for the disease SNP rs6792511 on chromosome 3 representing the African population at the nominal 0.05 level. With PC adjustment, the rejection rate is **0.824 ±0.024**.

Table 20. Empirical rejection rates with 95% confidence intervals of the single-locus model under experiment II (additive model with disease SNP MAF = 0.05, 0.15, 0.30) at the nominal 0.05 level.

MAF	Single-locus Disease Genes	Chr	Pop	Penetrance					
				Complete		High		Low	
				No PC	PC	No PC	PC	No PC	PC
				Uncorrelated SNPs (null I)					
0.05	rs6792511	3	Af	**0.115 ±0.004**	**0.040 ±0.002**	**0.061 ±0.003**	**0.055 ±0.003**	0.052 ±0.003	0.050 ±0.003
0.15	rs11924006	3	Af	**0.120 ±0.004**	**0.000 ±0.000**	**0.115 ±0.004**	**0.045 ±0.003**	**0.079 ±0.003**	0.048 ±0.003
0.30	rs9810313	3	Af	**0.160 ±0.005**	**0.037 ±0.002**	**0.167 ±0.005**	**0.025 ±0.002**	**0.139 ±0.004**	**0.067 ±0.003**
				Matching SNPs (null II)					
0.05	rs6792511	3	Af	**0.695 ±0.006**	**0.334 ±0.006**	**0.277 ±0.006**	**0.167 ±0.005**	**0.145 ±0.004**	**0.119 ±0.004**
0.15	rs11924006	3	Af	**0.920 ±0.003**	**0.280 ±0.006**	**0.891 ±0.004**	**0.163 ±0.005**	**0.732 ±0.005**	**0.104 ±0.004**
0.30	rs9810313	3	Af	**0.960 ±0.002**	**0.520 ±0.006**	**0.951 ±0.003**	**0.446 ±0.006**	**0.779 ±0.005**	**0.335 ±0.006**
				Disease SNPs (Power)					
0.05	rs6792511	3	Af	1.000 ±0.000	1.000 ±0.000	0.998 ±0.003	0.998 ±0.003	0.849 ±0.022	0.824 ±0.024
0.15	rs11924006	3	Af	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
0.30	rs9810313	3	Af	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	0.979 ±0.009	0.994 ±0.005

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = African (Af). The complete, high and low penetrance models are with low prevalence (additive model is used, three trajectory groups). See Methodology – Genetic Models. Type I error rates that are significantly different from the nominal 0.05 level are in bold. For each setting 1000 replicates are generated. Africans n = 466.

3.1.2.1 Null I Simulation Using Uncorrelated SNPs

The ANOVA of the measure of lack of robustness of validity for the uncorrelated SNPs is shown in Table 21 below. The model has an *R-square* of 0.7422. These results indicate that the overall model is not statistically significant ($F = 0.89,\ p - value = 0.6147$). Because the smallest *p-value* in the ANOVA is the *p-value* for the factor PC and is equal to 0.2732, I conclude that there are no significant factors for the lack of robustness of validity. That is, the statistical analysis did not confirm the apparent failure of robustness of validity for data from the complete penetrance model analyzed without PC adjustment.

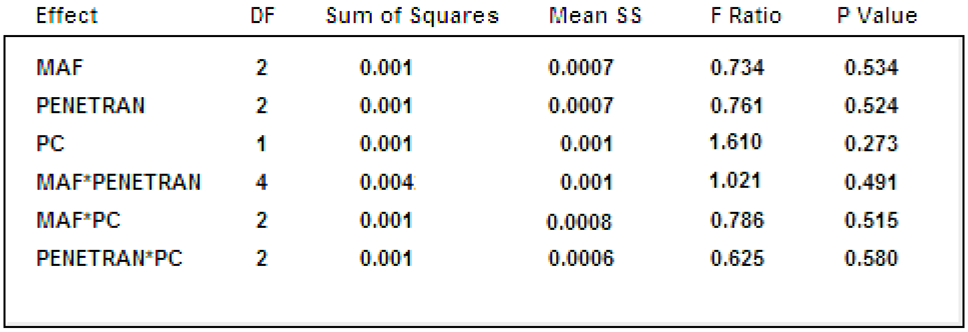
Table 21. The ANOVA table for the single-locus model uncorrelated SNPs under experiment II at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lackrobust

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	0.012	0.0009	0.89	0.6147
Error	4	0.004	0.001		
Corrected Total	17	0.016			

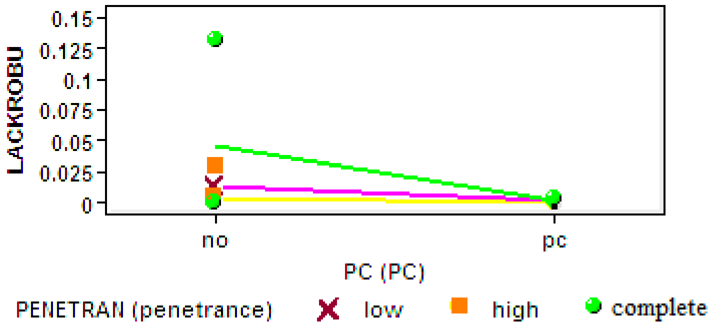
R-Square	Coeff Var	Root MSE	lackrobust Mean
0.742	295.749	0.032	0.011



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
MAF	2	0.001	0.0007	0.734	0.534
PENETRAN	2	0.001	0.0007	0.761	0.524
PC	1	0.001	0.001	1.610	0.273
MAF*PENETRAN	4	0.004	0.001	1.021	0.491
MAF*PC	2	0.001	0.0008	0.786	0.515
PENETRAN*PC	2	0.001	0.0006	0.625	0.580

To assist in the interpretation of these results, Figure 8 presents the plot of PC and penetrance, PC and MAF interaction. The lack of robustness of validity appears in the complete penetrance model without PC adjustment.

Figure 8. Lack of robustness of validity of single-locus model uncorrelated SNPs under experiment II at the nominal 0.05 level.



3.1.2.2 Null II Simulation Using Matching SNPs

The ANOVA of the measure of lack of robustness of validity for the matching SNPs is shown in Table 22 below. The model has an *R-square* of 0.9523. These results indicate that the overall model is statistically significant ($F = 6.14,\ p - value = 0.0466$) at the nominal 0.05 level. The value of the statistic $F = 51.282$ ($p - value = 0.0020$) for the factor PC is statistically significant. This analysis thus documents that the use of PC adjustment significantly improves robustness of validity.

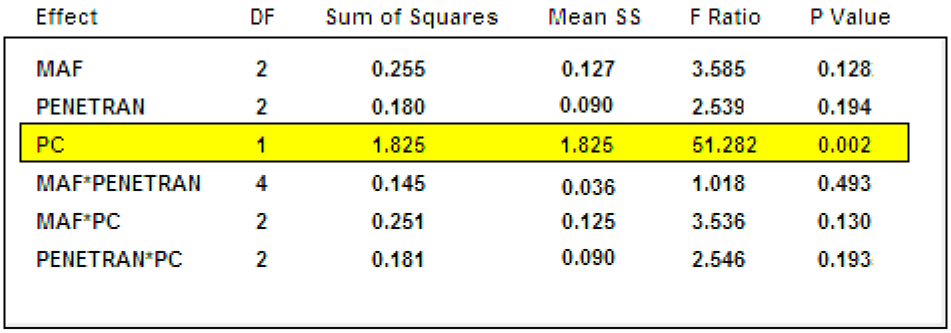
Table 22. The ANOVA table for the single-locus model matching SNPs under experiment II at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lackrobust

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	2.839	0.218	6.14	0.047
Error	4	0.142	0.036		
Corrected Total	17	2.982			

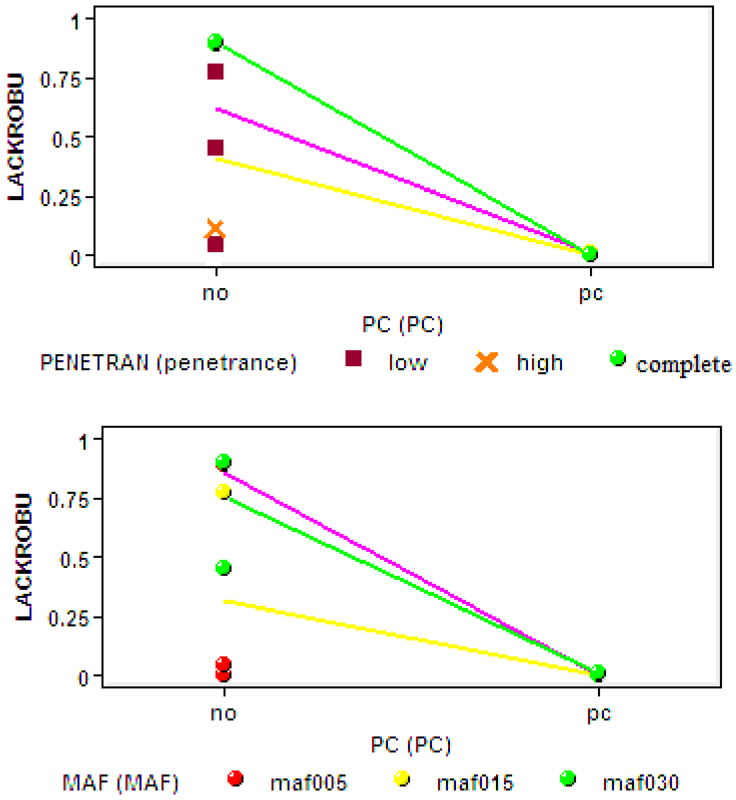
R-Square	Coeff Var	Root MSE	lackrobust Mean
0.952	58.620	0.189	0.322



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
MAF	2	0.255	0.127	3.585	0.128
PENETRAN	2	0.180	0.090	2.539	0.194
PC	1	1.825	1.825	51.282	0.002
MAF*PENETRAN	4	0.145	0.036	1.018	0.493
MAF*PC	2	0.251	0.125	3.536	0.130
PENETRAN*PC	2	0.181	0.090	2.546	0.193

For interpreting the results, I present the graph of the average lack of robustness of validity measure of the empirical type I error rate at each combination of the PC and penetrance, and PC and MAF in Figure 9. In general, better robustness of validity is attained at low MAF (0.05) and reduced penetrance (low and high). With PC adjustment, the robustness of validity improves for all three penetrance models: low, high and complete. The PC model gives the best results since it maintains better robustness of validity as MAF and penetrance levels change.

Figure 9. Lack of robustness of validity of Single-locus model matching SNPs under experiment II at the nominal 0.05 level.



3.1.2.3 Power Simulation Using single-locus disease SNPs

The ANOVA of the rejection rates for the disease SNPs is shown in Table 23 below. The model has an *R-square* of 0.9942. These results indicate that the overall model is statistically significant ($F = 52.98,\ p - value = 0.0008$). Since $F = 82.737$ ($p - value = 0.0006$) for MAF and $F = 101.16$ ($p - value = 0.0004$) for penetrance, the factors of MAF and penetrance are also significant. Furthermore, since $F = 79.666$ ($p - value = 0.0005$) for the interaction of MAF and penetrance, this interaction is also significant. PC is not a significant factor in this power simulation ($F = 0.081633,\ p - vlaue = 0.7893$). The most important finding is that the use of PC adjustment and interactions involving PC adjustment are not significant. That is, as before, the fitted rejection rate with PC adjustment is statistically equal to the fitted rate without PC adjustment.

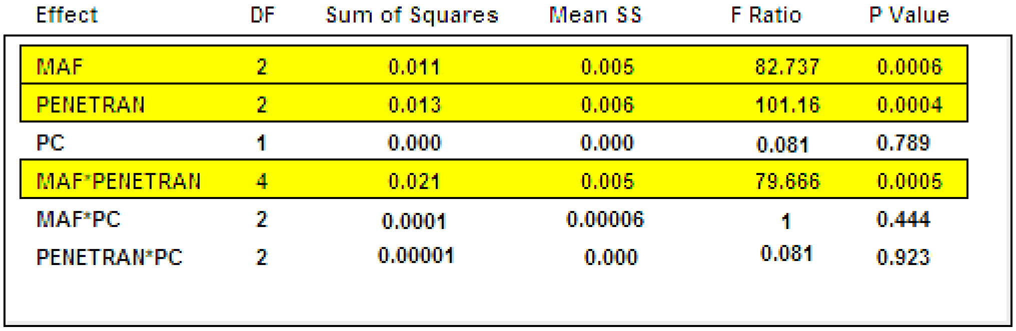
Table 23. The ANOVA table for the single-locus model disease SNPs under experiment II at the nominal 0.05 level (dependent variable: power).

The GLM Procedure

Dependent Variable: rate rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	0.047	0.004	52.98	0.001
Error	4	0.0003	0.0001		
Corrected Total	17	0.047			

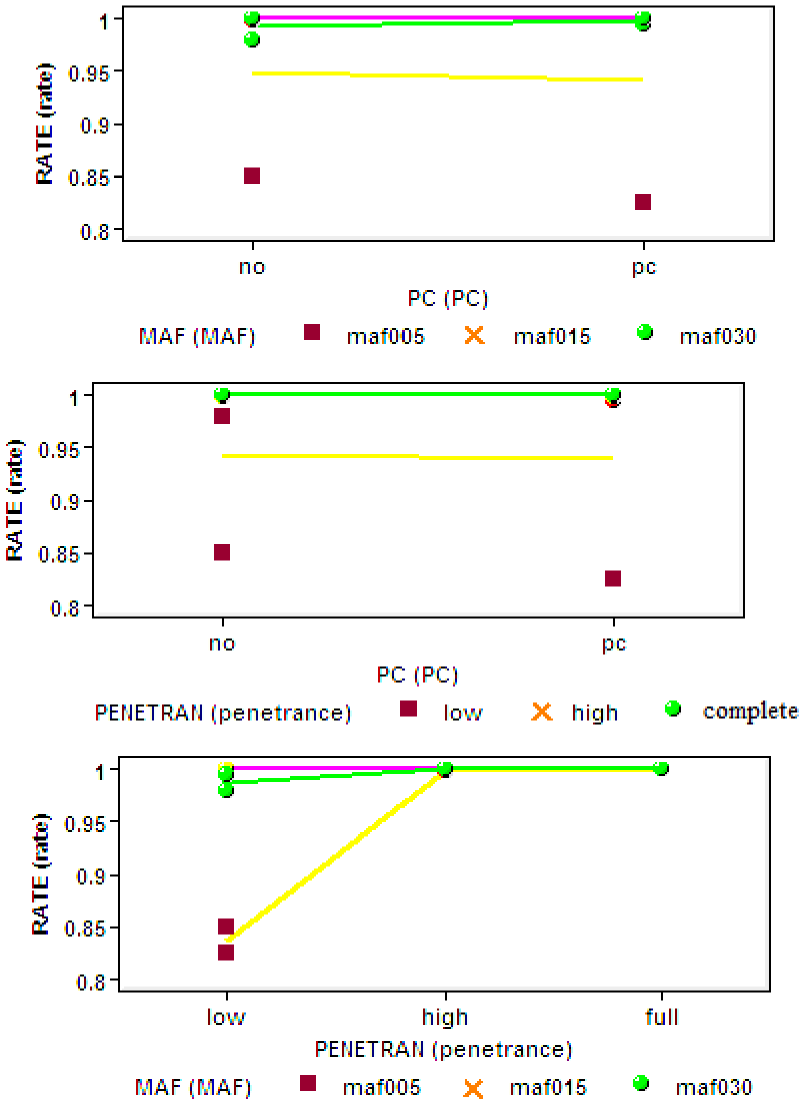
R-Square	Coeff Var	Root MSE	rate Mean
0.994	0.842	0.008	0.980



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
MAF	2	0.011	0.005	82.737	0.0006
PENETRAN	2	0.013	0.006	101.16	0.0004
PC	1	0.000	0.000	0.081	0.789
MAF*PENETRAN	4	0.021	0.005	79.666	0.0005
MAF*PC	2	0.0001	0.00006	1	0.444
PENETRAN*PC	2	0.00001	0.000	0.081	0.923

To assist in the interpretation of these results, Figure 10 presents plots of the MAF-PC, penetrance-PC and MAF-penetrance interactions. The interaction between PC and MAF, PC and penetrance are fairly small, as shown by the similar shape of the curves in Figure 10. I conclude that the rejection rate (power) with PC adjustment is close to the rate without PC adjustment. The powers of the high and complete penetrance models are greater than that of low penetrance model, regardless of MAF. The power is lower when MAF is small: i.e. MAF = 0.05.

Figure 10. Power of Single-locus model disease SNPs under experiment II at the nominal 0.05 level.



3.1.2.4 Statistical Analysis on Experiment II

I calculate the average of the overall power of Experiment II, null rejection rate and lack of robustness of validity with or without PC. Without PC adjustment, the null II type I error rate has a mean of 0.706 at the nominal 0.05 level, and the lack of robustness of validity has a mean of 0.565. After the PC adjustment, the average of overall type I error decreases to 0.274 and the lack of robustness of validity decreases to 0.183. The results indicate that the PC effect is smaller than no PC effect and that PC helps to maintain the power.

Table 24. Average rejection rates and lack of robustness of validity on Factor PC.

Overall Average	**Rejection Rate**		**Lack of robustness of validity**	
	No PC	**PC**	**No PC**	**PC**
Uncorrelated SNPs (null I)	0.112	0.041	0.062	0.007
Matching SNPs (null II)	0.706	0.274	0.565	0.183
Disease SNPs (Power)	0.981	0.980		

I also calculate the average of type I error and lack of robustness of validity with or without PC by the three general populations: African, Asian and European. My hypothesis is that African population needs PC adjustment more than European population but less than Asian population because of the population immigration. The results indicate that global PC methods adjust the type I error rate well overall, especially for the African and Asian populations.

Table 25. Average rejection rates and lack of robustness of validity on Factors MAF and PC.

Overall Average	Rejection Rate		Lack of robustness of validity	
	No PC	PC	No PC	PC
	MAF = 0.05			
Uncorrelated SNPs (null I)	0.076	0.048	0.022	0.006
Matching SNPs (null II)	0.372	0.207	0.221	0.117
Disease SNPs (Power)	0.949	0.941		
	MAF = 0.15			
Uncorrelated SNPs (null I)	0.105	0.031	0.059	0.002
Matching SNPs (null II)	0.848	0.182	0.704	0.098
Disease SNPs (Power)	1.000	1.000		
	MAF = 0.30			
Uncorrelated SNPs (null I)	0.155	0.043	0.106	0.012
Matching SNPs (null II)	0.897	0.434	0.768	0.333
Disease SNPs (Power)	0.993	0.998		

I further calculate the average of power and lack of robustness of validity with or without PC by the three penetrance levels: low, high and complete. The reduced penetrance models (low and high) maintain a correct type I error rate than the complete penetrance models.

Table 26. Average rejection rates and lack of robustness of validity on Factors penetrance and PC.

Overall Average	Rejection Rate		Lack of robustness of validity	
	No PC	PC	No PC	PC
Low Penetrance				
Uncorrelated SNPs (null I)	0.090	0.055	0.026	0.000
Matching SNPs (null II)	0.552	0.186	0.385	0.086
Disease SNPs (Power)	0.943	0.939		
High Penetrance				
Uncorrelated SNPs (null I)	0.115	0.041	0.054	0.002
Matching SNPs (null II)	0.707	0.259	0.571	0.141
Disease SNPs (Power)	0.999	0.999		
Complete Penetrance				
Uncorrelated SNPs (null I)	0.132	0.026	0.107	0.018
Matching SNPs (null II)	0.858	0.378	0.737	0.320
Disease SNPs (Power)	1.000	1.000		

3.1.3 Experiment III Results

Table 27 presents the average empirical type I error rate and empirical power observed in the simulations under experiment III, which has disease SNP MAF = 0.05, 0.15, 0.30, representing the African population as described in Chapter 2 Methodology. The high prevalence complete penetrance model is used in experiment III. The table also includes 95% confidence intervals. Each single-locus disease SNP in the table has 25 uncorrelated SNPs (MAFs across population less correlated with the disease SNP) and 25 matching SNPs (MAFs across population as correlated as possible with the disease SNP MAFs).

As in experiment I and II, the empirical type I error rate using the uncorrelated SNPs has a lack of robustness of validity appearing in the high prevalence complete penetrance without PC adjustment model. For example, on Table 27, using nominal level of significance 0.05, there are 16% (**0.160 ±0.005**) of the replicates for which the 25 SNPs uncorrelated with the disease SNP rs6792511 on chromosome 3 representing the African population are found significant in the association test when there is no PC adjustment. With PC adjustment, the number of replicates that are significant at the nominal 0.05 level in a linear regression test decreases to 4% (**0.040 ±0.002**).

There is a failure of robustness of validity for high prevalence complete penetrance model both with and without PC adjustment. For example, the 25 SNPs that match the disease SNP rs6792511 have an average type I error rate **0.960 ±0.002** without PC adjustment. With PC adjustment, the rate decreases to **0.240 ±0.005**, which is still far above the nominal 0.05 level. In general, PC adjustment improves robustness of validity.

The rejection rate for disease SNPs with PC adjustment is equal to the rate without PC adjustment. They are all equal to 1.000 ±0.000 at the nominal 0.05 level.

Table 27. Empirical rejection rates with 95% confidence intervals of the single-locus model under experiment III (disease SNP MAF = 0.05, 0.15, 0.30) at the nominal 0.05 level.

MAF	Single-locus Disease genes	Chr	Pop	Complete Penetrance High Prevalence			
				Rate		Lack of Robustness of Validity	
				No PC	PC	No PC	PC
				Uncorrelated SNPs (null I)			
0.05	rs6792511	3	Af	**0.160 ±0.005**	**0.040 ±0.002**	0.146 ±0.004	0.038 ±0.002
0.15	rs11924006	3	Af	**0.120 ±0.004**	**0.040 ±0.002**	0.110 ±0.004	0.038 ±0.002
0.30	rs9810313	3	Af	**0.320 ±0.006**	**0.040 ±0.002**	0.290 ±0.006	0.038 ±0.002
				Matching SNPs (null II)			
0.05	rs6792511	3	Af	**0.960 ±0.002**	**0.240 ±0.005**	0.867 ±0.004	0.219 ±0.005
0.15	rs11924006	3	Af	**1.000 ±0.000**	**0.240 ±0.005**	0.903 ±0.004	0.219 ±0.005
0.30	rs9810313	3	Af	**0.960 ±0.002**	**0.240 ±0.005**	0.867 ±0.004	0.218 ±0.005
				Disease SNPs (Power)			
0.05	rs6792511	3	Af	1.000 ±0.000	1.000 ±0.000		
0.15	rs11924006	3	Af	1.000 ±0.000	1.000 ±0.000		
0.30	rs9810313	3	Af	1.000 ±0.000	1.000 ±0.000		

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = African (Af). The complete penetrance model is with high prevalence (complete model, two trajectory groups). See Methodology – Genetic Models. Type I error rates that are significantly different from the nominal 0.05 level are in bold. For each setting 1000 replicates are generated. Africans n = 466.

3.1.3.1 Statistical Analysis on Experiment III

I report the average of the overall power and lack of robustness of validity with and without PC in Table 28. Without PC adjustment, the type I error rate cannot be controlled. Overall, the type I error rate of the PC analysis for matching SNPs has substantial improvement but is much greater than the nominal level of significance.

Table 28. Average rejection rates and lack of robustness of validity on Factor PC (nominal significance level 0.05).

Overall Average	**Rejection Rate**		**Lack of Robustness of Validity**	
	No PC	**PC**	**No PC**	**PC**
Uncorrelated SNPs (null I)	0.200	0.040	0.182	0.038
Matching SNPs (null II)	0.973	0.240	0.879	0.218
Disease SNPs (Power)	1.000	1.000		

3.2 Multi-locus Model Results

3.2.1 Experiment IV Results

Table 29 shows the empirical type I error rate and 95% confidence interval of the 450 uncorrelated SNPs (null I) for the 18 disease genes. The results show that the low penetrance model does not need PC adjustment. The high penetrance model does not need much PC adjustment. The complete penetrance model benefits substantially from PC adjustment. When the type I error rate and its 95% confidence interval are in bold, the target $\alpha = 0.05$ is not contained in the confidence interval. The number of intervals not containing the target α decreases when using the PC adjustment model. It also decreases as the penetrance decreases.

Table 30 shows the empirical type I error rate and 95% confidence interval of the 450 matching SNPs (null II) for the 18 disease genes. The results show that the low penetrance model does not need PC adjustment. The high penetrance model needs PC adjustment, and PC adjustment is effective. PC adjustment matters in the complete penetrance model and helps adjust for the PS, but the rejection rates for the matching SNPs are well above the nominal significance level for many SNPs. As before, when type I error rate and 95% confidence interval in bold, the target $\alpha = 0.05$ is not contained in the confidence interval. As the penetrance becomes smaller, the number of intervals not containing the target α decreases.

Table 31 contains the power to detect the disease gene in the complete, high and low penetrance additive models (three trajectory groups are used). The power decreases as the penetrance is less. There is no substantial change in power after PC adjustments. For the complete penetrance models, all the 18 disease genes have a power of 1, that is, 100% disease genes are detected before and after PC adjustment. For the partial high penetrance model, among the 18 disease genes, 17 had power greater than 0.945 ± 0.014 before the PC adjustment. One of the 18 disease genes, rs3761998 from chromosome 6, with relatively high MAF in the Asian population, has a relatively good power, 0.876 ± 0.02. After the PC adjustment, the power of this disease gene rs3761998 increases to 0.900 ± 0.019. The other 17 disease SNPs have a power greater than 0.952 ± 0.013. For the partial low penetrance model without PC adjustment, the power ranges from 0.463 ± 0.031 to 0.953 ± 0.013. Two of the 18 genes have low power less than 0.50. They are rs3761998 with a power of 0.463 ± 0.0310, which is the disease gene from chromosome 6 having a high MAF for an Asian population, and rs12822275 with a power of 0.464 ± 0.031, which is the disease gene from chromosome 12 that has a high MAF for the European population. After the PC adjustment, the power range stays the same. One additional disease SNP rs270771 appears to have power less than 0.50. Its rejection rate is 0.476 ± 0.031. It is on chromosome 19 and has a high MAF for a European population.

Figure 11 displays the trends and histograms of the type I error rates of the two partial penetrance models for the 450 non-correlated SNPs under the null I with and without PC adjustments. Figure 12 presents the scatter plot of the type I error rates for the 450 non-correlated SNPs under the null I with and without PC adjustment.

The models with PC adjustment have an average type I error rate closer to 0.05 than the models without PC adjustment. The PC adjustment models have more intervals containing the target α. Specifically, the partial high penetrance model has a type I error rate range from 0.026 ± 0.01 to 0.097 ± 0.018 without PC adjustment. Using PC adjustment, the range shifts to 0.016 ± 0.008 to 0.073 ± 0.016. For the partial low penetrance model, the type I error rates ranging from 0.044 ± 0.013 to 0.066 ± 0.015 before PC adjustments. The range shrinks to 0.042 ± 0.012 to 0.058 ± 0.014 after PC adjustments. All intervals in the partial low penetrance PC adjustment model contain the target $\alpha = 0.05$.

Some of the matching SNPs in the high partial penetrance model, especially those representing the African population, have rejection rates above the nominal value of 0.05 (the disease SNPs representing African population have type I error rates that range from 0.095 to 0.165) without PC adjustment. Disease SNPs number 1, 5, 8, 10, 13 and 16 that represent the Asian population have lower type I error rate in high partial penetrance model, ranging from 0.013 ± 0.007 to 0.049 ± 0.013 without PC adjustment.

The models with PC adjustment have an average type I error rate closer to 0.05 than the models without PC adjustment. PC models also have more intervals containing the target α. Specifically, the low penetrance PC adjustment model has type I error rates that range 0.044 to 0.059. Before adjustment, they ranged from 0.039 to 0.077. All matching SNPs for this model have confidence intervals containing the target $\alpha = 0.05$ with PC adjustment.

Table 29. Empirical rejection rates and 95% confidence interval of the multi-locus additive models for the uncorrelated SNPs (null I) under experiment IV at the nominal 0.05 level.

#	Disease Genes	Chr	Pop	Penetrance					
				Complete		High		Low	
				No PC	PC	No PC	PC	No PC	PC
1	rs3733124	3	As	**0.080 ±0.017**	**0.000 ±0.000**	0.055 ±0.014	**0.034 ±0.011**	0.052 ±0.014	0.046 ±0.013
2	rs7355991	3	Af	0.040 ±0.012	0.040 ±0.012	**0.034 ±0.011**	0.049 ±0.013	0.049 ±0.013	0.050 ±0.014
3	rs17195948	3	Eu	**0.080 ±0.017**	**0.079 ±0.017**	0.063 ±0.015	0.052 ±0.014	0.055 ±0.014	0.051 ±0.014
4	rs1259069	6	Eu	0.039 ±0.012	0.040 ±0.012	**0.029 ±0.010**	0.039 ±0.012	0.044 ±0.013	0.042 ±0.012
5	rs3761998	6	As	0.040 ±0.012	0.040 ±0.012	0.044 ±0.013	**0.032 ±0.011**	0.044 ±0.013	0.042 ±0.012
6	rs9459886	6	Af	0.040 ±0.012	0.040 ±0.012	0.053 ±0.014	0.046 ±0.013	0.051 ±0.014	0.051 ±0.014
7	rs12790383	11	Eu	**0.000 ±0.001**	**0.000 ±0.000**	0.045 ±0.013	**0.021 ±0.009**	0.052 ±0.014	0.045 ±0.013
8	rs11217935	11	As	0.040 ±0.012	0.040 ±0.012	**0.067 ±0.015**	0.043 ±0.013	**0.066 ±0.015**	0.058 ±0.014
9	rs11825331	11	Af	0.040 ±0.012	**0.080 ±0.017**	0.038 ±0.012	0.050 ±0.014	0.048 ±0.013	0.046 ±0.013
10	rs17117910	12	As	**0.000 ±0.000**	**0.000 ±0.000**	**0.026 ±0.010**	**0.016 ±0.008**	0.047 ±0.013	0.043 ±0.013
11	rs12822275	12	Eu	**0.000 ±0.000**	**0.000 ±0.000**	**0.036 ±0.012**	**0.035 ±0.011**	0.049 ±0.013	0.047 ±0.013
12	rs1696449	12	Af	0.040 ±0.012	0.040 ±0.012	0.057 ±0.014	0.050 ±0.014	0.053 ±0.014	0.051 ±0.014
13	rs2073868	17	As	**0.080 ±0.017**	**0.160 ±0.023**	**0.068 ±0.016**	**0.073 ±0.016**	0.054 ±0.014	0.054 ±0.014
14	rs9899123	17	Af	0.040 ±0.012	0.040 ±0.012	0.054 ±0.014	0.053 ±0.014	0.051 ±0.014	0.048 ±0.013
15	rs34742396	17	Eu	**0.120 ±0.020**	**0.081 ±0.017**	**0.097 ±0.018**	**0.068 ±0.016**	0.059 ±0.015	0.050 ±0.014
16	rs3745465	19	As	**0.120 ±0.020**	**0.079 ±0.017**	0.063 ±0.015	0.058 ±0.014	0.055 ±0.014	0.054 ±0.014
17	rs10411117	19	Af	**0.080 ±0.017**	0.040 ±0.012	0.064 ±0.015	0.051 ±0.014	0.051 ±0.014	0.049 ±0.013
18	rs270771	19	Eu	**0.000 ±0.000**	0.040 ±0.012	**0.029 ±0.010**	**0.031 ±0.011**	0.047 ±0.013	0.045 ±0.013

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = General populations, including African (Af), Asian (As) and European (Eu). For the complete penetrance low prevalence model, two trajectory groups instead of three are used (see Methodology – Genetic Models). For each setting 1000 replicates are generated. Africans n = 466, Asians n = 359, Europeans n = 214.

Table 30. **Empirical rejection rates and 95% confidence interval of the multi-locus gene models for the matching SNPs (null II).**

#	**Disease genes**	**Chr**	**Pop**	**Penetrance**					
				Complete		**High**		**Low**	
				No PC	**No PC**	**No PC**	**PC**	**No PC**	**PC**
1	rs3733124	3	As	**0.000 ±0.000**	**0.000 ±0.000**	**0.017 ±0.008**	**0.023 ±0.009**	0.042 ±0.012	0.044 ±0.013
2	rs7355991	3	Af	**0.200 ±0.025**	**0.160 ±0.023**	**0.154 ±0.022**	**0.095 ±0.018**	**0.072 ±0.016**	0.059 ±0.015
3	rs17195948	3	Eu	**0.120 ±0.020**	**0.160 ±0.023**	**0.123 ±0.020**	**0.113 ±0.020**	**0.067 ±0.015**	0.059 ±0.015
4	rs1259069	6	Eu	**0.240 ±0.026**	**0.120 ±0.020**	**0.116 ±0.020**	**0.105 ±0.019**	**0.066 ±0.015**	0.058 ±0.014
5	rs3761998	6	As	**0.000 ±0.000**	**0.000 ±0.000**	**0.013 ±0.007**	**0.018 ±0.008**	0.039 ±0.012	0.039 ±0.012
6	rs9459886	6	Af	**0.200 ±0.025**	0.040 ±0.012	**0.132 ±0.021**	0.058 ±0.014	**0.068 ±0.016**	0.051 ±0.014
7	rs12790383	11	Eu	**0.160 ±0.023**	**0.120 ±0.020**	**0.107 ±0.019**	0.049 ±0.013	**0.067 ±0.015**	0.059 ±0.015
8	rs11217935	11	As	**0.000 ±0.000**	**0.000 ±0.000**	**0.027 ±0.010**	**0.036 ±0.012**	0.050 ±0.014	0.051 ±0.014
9	rs11825331	11	Af	**0.080 ±0.017**	0.040 ±0.012	**0.110 ±0.019**	**0.070 ±0.016**	0.064 ±0.015	0.054 ±0.014
10	rs17117910	12	As	**0.000 ±0.000**	**0.000 ±0.001**	0.049 ±0.013	**0.037 ±0.012**	0.060 ±0.015	0.051 ±0.014
11	rs12822275	12	Eu	0.080 ±0.017	**0.080 ±0.017**	**0.075 ±0.016**	0.046 ±0.013	0.060 ±0.015	0.051 ±0.014
12	rs1696449	12	Af	**0.160 ±0.023**	**0.000 ±0.000**	**0.095 ±0.018**	0.040 ±0.012	0.060 ±0.015	0.052 ±0.014
13	rs2073868	17	As	**0.040 ±0.012**	0.040 ±0.012	0.042 ±0.012	**0.037 ±0.012**	0.052 ±0.014	0.048 ±0.013
14	rs9899123	17	Af	**0.240 ±0.026**	**0.120 ±0.020**	**0.165 ±0.023**	**0.093 ±0.018**	**0.069 ±0.016**	0.054 ±0.014
15	rs34742396	17	Eu	**0.120 ±0.020**	0.040 ±0.012	**0.082 ±0.017**	0.059 ±0.015	0.059 ±0.015	0.050 ±0.014
16	rs3745465	19	As	**0.000 ±0.000**	**0.000 ±0.000**	**0.020 ±0.009**	**0.025 ±0.010**	0.043 ±0.013	0.044 ±0.013
17	rs10411117	19	Af	0.160 ±0.023	0.040 ±0.012	**0.145 ±0.022**	0.063 ±0.015	**0.077 ±0.017**	0.058 ±0.014
18	rs270771	19	Eu	0.240 ±0.026	**0.080 ±0.017**	**0.120 ±0.020**	**0.067 ±0.015**	**0.075 ±0.016**	0.060 ±0.015

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = General populations, including African (Af), Asian (As) and European (Eu). For the complete penetrance low prevalence model, two trajectory groups instead of three are used (see Methodology – Genetic Models). For each setting 1000 replicates are generated. Africans n = 466, Asians n = 359, Europeans n = 214.

Table 31. **Empirical rejection rates and 95% confidence interval of the multi-locus gene complete penetrance and partial penetrance models.**

#	Disease genes	Chr	Pop	Penetrance					
				Complete		High		Low	
				No PC	PC	No PC	PC	No PC	PC
1	rs3733124	3	As	1.000 ±0.000	1.000 ±0.000	0.992 ±0.006	0.997 ±0.003	0.652 ±0.030	0.663 ±0.029
2	rs7355991	3	Af	1.000 ±0.000	1.000 ±0.000	0.992 ±0.006	0.987 ±0.007	0.636 ±0.030	0.595 ±0.030
3	rs17195948	3	Eu	1.000 ±0.000	1.000 ±0.000	0.992 ±0.006	0.984 ±0.008	0.598 ±0.030	0.536 ±0.031
4	rs1259069	6	Eu	1.000 ±0.000	1.000 ±0.000	0.987 ±0.007	0.968 ±0.011	0.591 ±0.030	0.554 ±0.031
5	rs3761998	6	As	1.000 ±0.000	1.000 ±0.000	0.876 ±0.020	0.900 ±0.019	0.463 ±0.031	0.469 ±0.031
6	rs9459886	6	Af	1.000 ±0.000	1.000 ±0.000	0.996 ±0.004	0.996 ±0.004	0.876 ±0.020	0.872 ±0.021
7	rs12790383	11	Eu	1.000 ±0.000	1.000 ±0.000	0.992 ±0.006	0.974 ±0.010	0.542 ±0.031	0.505 ±0.031
8	rs11217935	11	As	1.000 ±0.000	1.000 ±0.000	0.945 ±0.014	0.967 ±0.011	0.567 ±0.031	0.583 ±0.031
9	rs11825331	11	Af	1.000 ±0.000	1.000 ±0.000	0.995 ±0.004	0.995 ±0.004	0.710 ±0.028	0.711 ±0.028
10	rs17117910	12	As	1.000 ±0.000	1.000 ±0.000	0.981 ±0.008	0.979 ±0.009	0.630 ±0.030	0.609 ±0.030
11	rs12822275	12	Eu	1.000 ±0.000	1.000 ±0.000	0.973 ±0.010	0.957 ±0.013	0.464 ±0.031	0.457 ±0.031
12	rs1696449	12	Af	1.000 ±0.000	1.000 ±0.000	0.955 ±0.013	0.952 ±0.013	0.513 ±0.031	0.499 ±0.031
13	rs2073868	17	As	1.000 ±0.000	1.000 ±0.000	0.996 ±0.004	0.997 ±0.003	0.707 ±0.028	0.690 ±0.029
14	rs9899123	17	Af	1.000 ±0.000	1.000 ±0.000	0.999 ±0.002	1.000 ±0.000	0.819 ±0.024	0.800 ±0.025
15	rs34742396	17	Eu	1.000 ±0.000	1.000 ±0.000	0.974 ±0.010	0.975 ±0.010	0.558 ±0.031	0.519 ±0.031
16	rs3745465	19	As	1.000 ±0.000	1.000 ±0.000	0.997 ±0.003	0.996 ±0.004	0.953 ±0.013	0.962 ±0.012
17	rs10411117	19	Af	1.000 ±0.000	1.000 ±0.000	0.994 ±0.005	0.994 ±0.005	0.691 ±0.029	0.677 ±0.029
18	rs270771	19	Eu	1.000 ±0.000	1.000 ±0.000	0.970 ±0.011	0.959 ±0.012	0.509 ±0.031	0.476 ±0.031

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = General populations, including African (Af), Asian (As) and European (Eu). For the complete penetrance low prevalence model, two trajectory groups instead of three are used (see Methodology – Genetic Models). For each setting 1000 replicates are generated. Africans n = 466, Asians n = 359, Europeans n = 214.

Figure 11. Empirical rejection rates of the two multi-locus partial penetrance models for the 450 non-correlated SNPs with and without PC adjustment.

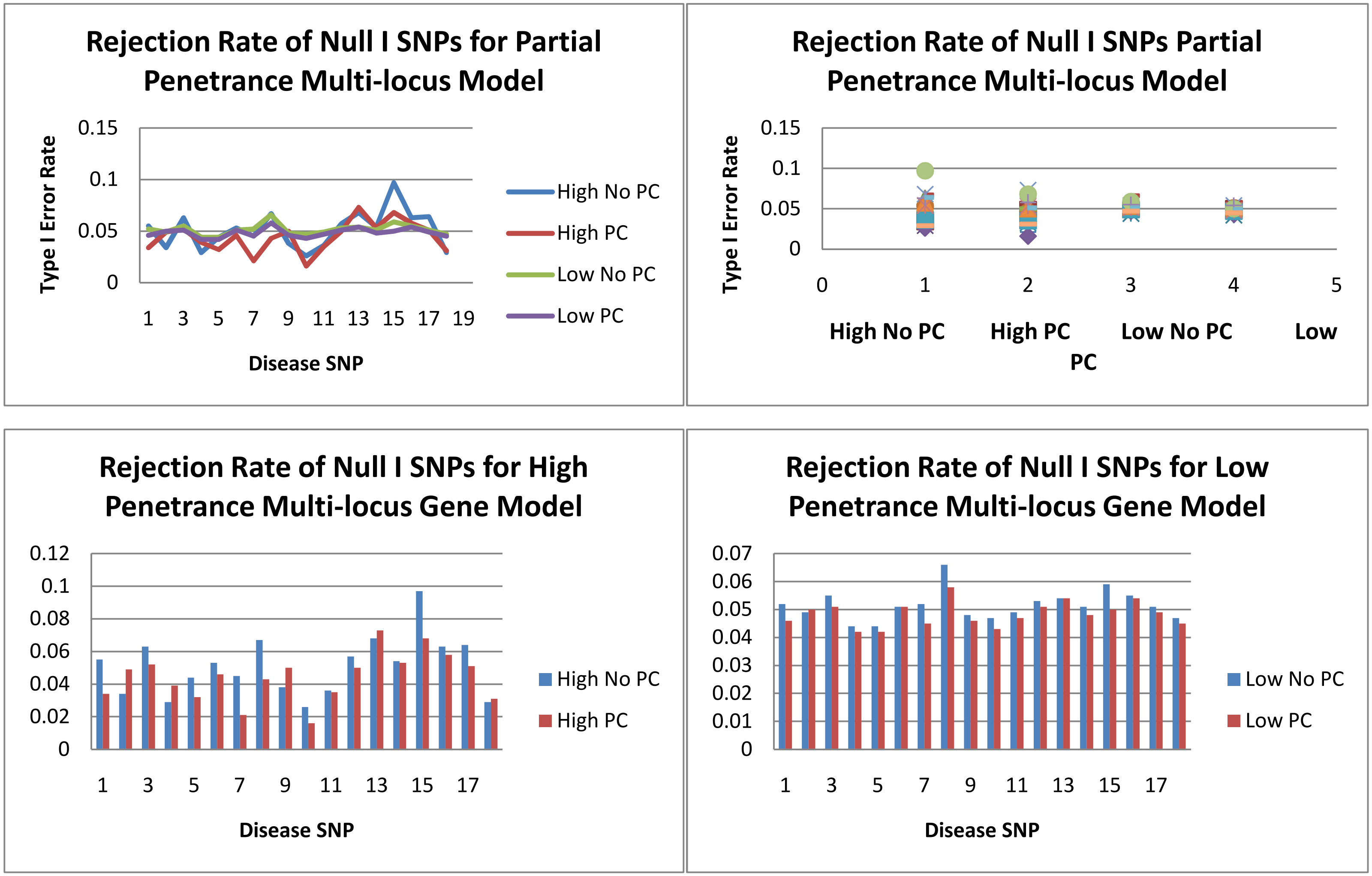
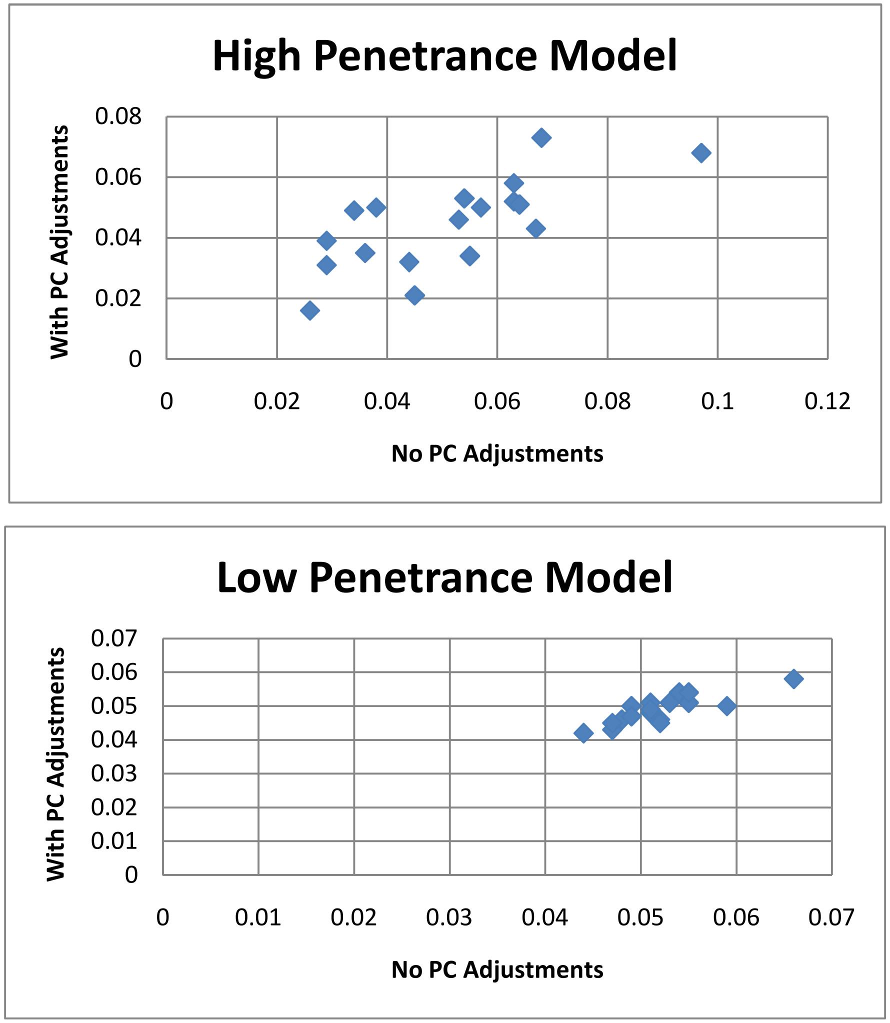


Figure 12. Empirical rejection rate scatter plot of 450 non-correlated SNPs for each of the 18 disease SNPs at the 0.05 nominal level.



3.2.1.1 Null Simulation Results Using Uncorrelated SNPs

The ANOVA of the measure of lack of robustness of validity for the multi-locus model uncorrelated SNPs is shown in Table 32 below. Although the model has an *R-square* of 0.0457, the overall model is statistically significant ($F = 2.96,\ p - value < 0.0001$). The results indicate that the statistic $F = 2.69,\ p - value = 0.0199$ for the chromosome factor and $\mathrm{F} = 40.35,\ \mathrm{p} - \mathrm{value} < 0.0001$ for penetrance are both significant at the 0.05 significance level. The factor PC is not significant in this model. That is, the statistical analysis did not confirm the failure of robustness of validity for data from the additive multi-locus null simulation model analyzed without PC adjustment using uncorrelated SNPs.

Table 32. The ANOVA table for the multi-locus model uncorrelated SNPs under experiment IV at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lackrobust lackrobust

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	43	1.606	0.037	2.96	<.0001
Error	2656	33.546	0.013		
Corrected Total	2699	35.152			

R-Square	Coeff Var	Root MSE	lackrobust Mean
0.046	616.281	0.112	0.018

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Chromosome	5	0.170	0.034	2.69	0.020
Population	2	0.018	0.009	0.71	0.494
Penetrance	2	1.019	0.510	40.35	<.0001
PC	1	0.001	0.001	0.08	0.775
pene*pc	2	0.0005	0.0002	0.02	0.981
chr*pop	10	0.142	0.014	1.13	0.338
chr*pene	10	0.221	0.022	1.75	0.065
chr*pc	5	0.013	0.003	0.21	0.958
pop*pene	4	0.020	0.005	0.39	0.816
pop*pc	2	0.002	0.001	0.07	0.931

3.2.1.2 Null Simulation Results for SNPs Having MAF Correlated with Disease SNP

The ANOVA of the measure of lack of robustness of validity for the matching SNPs is shown in Table 33 below. The model has an *R-square* of 0.1036. These results indicate that the overall model is statistically significant ($F = 7.14,\ p - value < 0.0001$) at the 0.05 level. The value of the statistic $F = 14.969$ $(p - value = 0.0001)$ for the factor PC, $F = 65.11$ $(p - value < 0.0001)$ for the factor penetrance and $F = 27.638$ $(p - value < 0.0001)$ for the factor population are statistically significant. The interaction between PC and penetrance, PC and population, population and penetrance are also statistically significant. This analysis documents that the use of PC adjustment is significantly associated with robustness of validity.

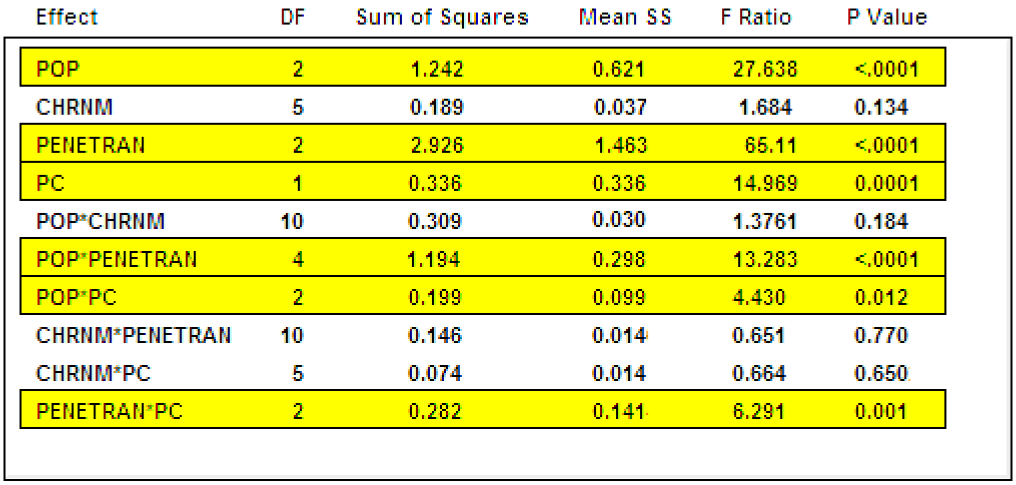
Table 33. The ANOVA table for the multi-locus model matching SNPs under experiment IV at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lackrobust lackrobust

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	43	6.901	0.160	7.14	<.0001
Error	2656	59.696	0.022		
Corrected Total	2699	66.597			

R-Square	Coeff Var	Root MSE	lackrobust Mean
0.104	436.422	0.150	0.034



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
POP	2	1.242	0.621	27.638	<.0001
CHRNM	5	0.189	0.037	1.684	0.134
PENETRAN	2	2.926	1.463	65.11	<.0001
PC	1	0.336	0.336	14.969	0.0001
POP*CHRNM	10	0.309	0.030	1.3761	0.184
POP*PENETRAN	4	1.194	0.298	13.283	<.0001
POP*PC	2	0.199	0.099	4.430	0.012
CHRNM*PENETRAN	10	0.146	0.014	0.651	0.770
CHRNM*PC	5	0.074	0.014	0.664	0.650
PENETRAN*PC	2	0.282	0.141	6.291	0.001

3.2.1.3 Power Simulation Using single-locus disease SNPs

The ANOVA of the rejection rates for the disease SNPs is shown in Table 34 below. The model has an *R-square* of 0.9471. These results indicate that the overall model is statistically significant ($F = 26.66,\ p - value < 0.0001$). Because $F = 11.637$ ($p - value < 0.0001$) for population, $F = 493.92$ ($p - value < 0.0001$) for penetrance and $F = 3.2751$ $(p - value = 0.0107)$ for chromosome, the factors of population, penetrance and chromosome are statistically significant. Furthermore, there is a significant interaction between population and penetrance, population and chromosome, penetrance and chromosome. PC is not a significant factor in the power analysis ($F = 0.3224,\ p - vlaue = 0.5721$). The most important finding is that the use of PC adjustment and interactions involving PC adjustment are not significant. That is, the fitted rejection rate with PC adjustment is statistically equal to the fitted rate without PC adjustment.

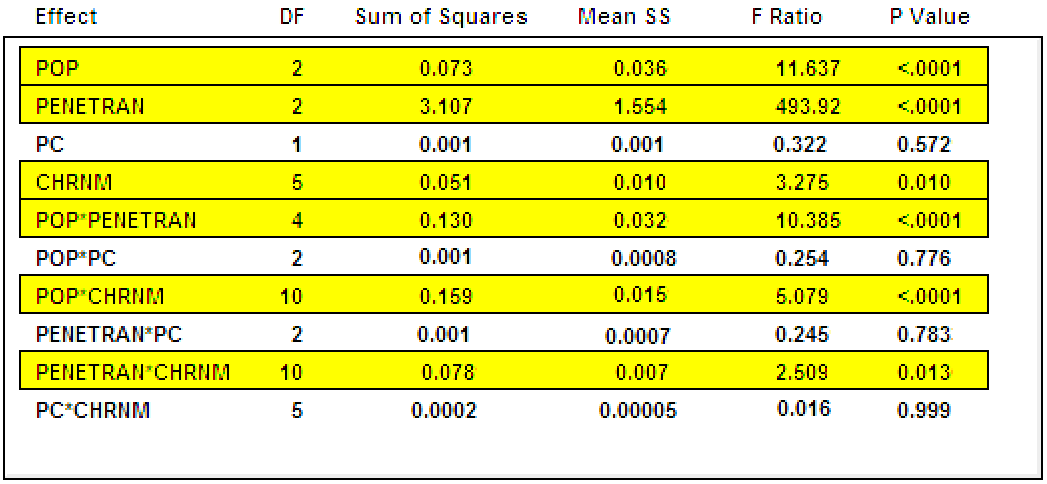
Table 34. The ANOVA table for the multi-locus model disease SNPs under experiment IV at the nominal 0.05 level (dependent variable: power).

The GLM Procedure

Dependent Variable: rate rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	43	3.607	0.084	26.66	<.0001
Error	64	0.201	0.003		
Corrected Total	107	3.808			

R-Square	Coeff Var	Root MSE	rate Mean
0.947	6.456	0.056	0.869



Effect	DF	Sum of Squares	Mean SS	F Ratio	P Value
POP	2	0.073	0.036	11.637	<.0001
PENETRAN	2	3.107	1.554	493.92	<.0001
PC	1	0.001	0.001	0.322	0.572
CHRNM	5	0.051	0.010	3.275	0.010
POP*PENETRAN	4	0.130	0.032	10.385	<.0001
POP*PC	2	0.001	0.0008	0.254	0.776
POP*CHRNM	10	0.159	0.015	5.079	<.0001
PENETRAN*PC	2	0.001	0.0007	0.245	0.783
PENETRAN*CHRNM	10	0.078	0.007	2.509	0.013
PC*CHRNM	5	0.0002	0.00005	0.016	0.999

3.2.1.4 Statistical Analysis on Experiment IV

Table 35 reports the average of the rejection rate for the disease SNPs and lack of robustness of validity measure with or without PC. Each entry is the average over 54 settings (three population settings times six chromosome settings, times three penetrance settings). The results indicate that the PC null rejection rates are closer to the nominal level than the rates without PC adjustment. For disease SNPs, the rejection rate with PC adjustment is only slightly less than the rejection rate without PC adjustment.

Table 35. Average rejection rates and lack of robustness of validity on Factor PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	**PC**	**No PC**	**PC**
Uncorrelated SNPs (null I)	0.051	0.046	0.019	0.018
Matching SNPs (null II)	0.087	0.056	0.046	0.023
Disease SNPs (Power)	0.872	0.866		

I also report in Table 36 the average of type I error rate, power and lack of robustness of validity with or without PC by the three general populations: African, Asian and European. My hypothesis is that the African population has a greater change with PC adjustment than European population but less than the Asian population because of the population immigration. The results indicate that global PC method adjusts the type I error rate well overall, especially for the African and Asian population, but less effectively for the European population. The rejection rates for the disease genes are roughly the same for each population. More importantly, the rejection rate with PC adjustment is roughly equal to the rate without adjustment.

Table 36. Average rejection rates and lack of robustness of validity on Factors population and PC.

Overall Average	Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
		African		
Uncorrelated SNPs (null I)	0.049	0.049	0.018	0.019
Matching SNPs (null II)	0.125	0.064	0.069	0.027
Disease SNPs (Power)	0.899	0.893		
		Asian		
Uncorrelated SNPs (null I)	0.056	0.048	0.023	0.019
Matching SNPs (null II)	0.027	0.027	0.004	0.004
Disease SNPs (Power)	0.876	0.878		
		European		
Uncorrelated SNPs (null I)	0.047	0.043	0.015	0.015
Matching SNPs (null II)	0.110	0.076	0.063	0.038
Disease SNPs (Power)	0.842	0.826		

I further report in Table 37 the average of type I error rate, power and lack of robustness of validity with or without PC by the three penetrance levels: low, high and complete. The reduced penetrance models (low and high) maintain a correct type I error rate with PC adjustment. With the complete penetrance model, the matching SNP rejection rate is above the nominal rate without PC adjustment. The complete and high penetrance models have larger power than the low penetrance model. As before, the rejection rate for disease SNPs with PC adjustment is less than but roughly equal to the rate without adjustment.

Table 37. Average rejection rates and lack of robustness of validity on Factors penetrance and PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
Low Penetrance				
Uncorrelated SNPs (null I)	0.052	0.048	0.001	0.000
Matching SNPs (null II)	0.060	0.052	0.002	0.001
Disease SNPs (Power)	0.638	0.621		
High Penetrance				
Uncorrelated SNPs (null I)	0.051	0.044	0.010	0.008
Matching SNPs (null II)	0.088	0.058	0.030	0.014
Disease SNPs (Power)	0.978	0.977		
Complete Penetrance				
Uncorrelated SNPs (null I)	0.049	0.047	0.046	0.044
Matching SNPs (null II)	0.113	0.058	0.104	0.054
Disease SNPs (Power)	1.000	1.000		

I report in Table 38 the average of type I error rate, power and lack of robustness of validity with or without PC by the six chromosomes: chromosome 3, 6, 11, 12, 17 and 19. The medium size chromosomes, chromosome 11 and 12, have type I error rate better controlled with the PC adjustment. The small size chromosomes, chromosome 17 and 19, have a larger power both with and without PC adjustment.

Table 38. Average rejection rates and lack of robustness of validity on factor chromosome.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
Chromosome 3				
Uncorrelated SNPs (null I)	0.056	0.044	0.024	0.015
Matching SNPs (null II)	0.088	0.079	0.049	0.042
Disease SNPs (Power)	0.874	0.862		
Chromosome 6				
Uncorrelated SNPs (null I)	0.043	0.041	0.015	0.016
Matching SNPs (null II)	0.097	0.054	0.056	0.021
Disease SNPs (Power)	0.865	0.862		
Chromosome 11				
Uncorrelated SNPs (null I)	0.044	0.043	0.012	0.015
Matching SNPs (null II)	0.074	0.053	0.034	0.021
Disease SNPs (Power)	0.861	0.859		
Chromosome 12				
Uncorrelated SNPs (null I)	0.034	0.031	0.007	0.006
Matching SNPs (null II)	0.071	0.040	0.030	0.010
Disease SNPs (Power)	0.835	0.828		
Chromosome 17				
Uncorrelated SNPs (null I)	0.069	0.070	0.030	0.034
Matching SNPs (null II)	0.097	0.060	0.053	0.027
Disease SNPs (Power)	0.895	0.887		
Chromosome 19				
Uncorrelated SNPs (null I)	0.057	0.050	0.024	0.019
Matching SNPs (null II)	0.098	0.049	0.051	0.017
Disease SNPs (Power)	0.902	0.896		

3.2.2 Experiment V Results

Table 39 shows the empirical type I error rate and 95% confidence interval of the 450 uncorrelated SNPs (null I) for the 18 disease genes. Table 40 shows the empirical type I error rate and 95% confidence interval of the 450 matching SNPs (null II) for the 18 disease genes. The results show that PC adjustment helps adjust for PS overall, but the null rates are higher than the nominal level. Table 41 shows the rejection rates for the 18 disease SNPs. The rejection rate for disease SNPs with PC adjustment is equal to the rate without PC adjustment. They are all equal to 1.000 ±0.000 at the nominal 0.05 level.

Table 39. **Empirical rejection rates and 95% confidence interval of the multi-locus high prevalence complete penetrance models for the uncorrelated SNPs (null I) under experiment V at the nominal 0.05 level.**

#	**Disease Genes**	**Chr**	**Pop**	**Complete Penetrance**	
				High Prevalence (Eli-2traj)	
				No PC	**PC**
1	rs3733124	3	As	**0.000 ±0.000**	0.040 ±0.012
2	rs7355991	3	Af	**0.120 ±0.020**	0.040 ±0.012
3	rs17195948	3	Eu	**0.080 ±0.017**	0.040 ±0.012
4	rs1259069	6	Eu	**0.000 ±0.000**	**0.000 ±0.000**
5	rs3761998	6	As	**0.000 ±0.000**	**0.000 ±0.000**
6	rs9459886	6	Af	**0.080 ±0.017**	**0.000 ±0.000**
7	rs12790383	11	Eu	**0.120 ±0.020**	**0.080 ±0.017**
8	rs11217935	11	As	**0.120 ±0.020**	0.040 ±0.012
9	rs11825331	11	Af	**0.080 ±0.017**	**0.000 ±0.000**
10	rs17117910	12	As	**0.000 ±0.000**	0.040 ±0.012
11	rs12822275	12	Eu	0.040 ±0.012	0.040 ±0.012
12	rs1696449	12	Af	**0.080 ±0.017**	0.040 ±0.012
13	rs2073868	17	As	0.040 ±0.012	0.040 ±0.012
14	rs9899123	17	Af	**0.120 ±0.020**	0.040 ±0.012
15	rs34742396	17	Eu	**0.080 ±0.017**	0.040 ±0.012
16	rs3745465	19	As	**0.120 ±0.020**	**0.160 ±0.023**
17	rs10411117	19	Af	**0.080 ±0.017**	**0.000 ±0.000**
18	rs270771	19	Eu	**0.080 ±0.017**	0.040 ±0.012

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = General populations, including African (Af), Asian (As) and European (Eu). For the complete penetrance low prevalence model, two trajectory groups instead of three are used (see Methodology – Genetic Models). For each setting 1000 replicates are generated. Africans n = 466, Asians n = 359, Europeans n = 214.

Table 40. **Empirical rejection rates and 95% confidence interval of the multi-locus gene models for the matching SNPs (null II).**

#	**Disease genes**	**Chr**	**Pop**	**Complete Penetrance**	
				High Prevalence (Eli-2traj)	
				No PC	**No PC**
1	rs3733124	3	As	**0.000 ±0.000**	**0.000 ±0.000**
2	rs7355991	3	Af	**0.120 ±0.020**	**0.080 ±0.017**
3	rs17195948	3	Eu	**0.200 ±0.025**	**0.160 ±0.023**
4	rs1259069	6	Eu	**0.080 ±0.017**	**0.080 ±0.017**
5	rs3761998	6	As	**0.000 ±0.000**	**0.080 ±0.017**
6	rs9459886	6	Af	**0.200 ±0.025**	0.040 ±0.012
7	rs12790383	11	Eu	**0.200 ±0.025**	**0.160 ±0.023**
8	rs11217935	11	As	**0.080 ±0.017**	**0.000 ±0.000**
9	rs11825331	11	Af	0.040 ±0.012	**0.080 ±0.017**
10	rs17117910	12	As	**0.360 ±0.030**	**0.120 ±0.020**
11	rs12822275	12	Eu	**0.160 ±0.023**	**0.080 ±0.017**
12	rs1696449	12	Af	0.040 ±0.012	0.040 ±0.012
13	rs2073868	17	As	**0.120 ±0.020**	**0.080 ±0.017**
14	rs9899123	17	Af	**0.080 ±0.017**	**0.080 ±0.017**
15	rs34742396	17	Eu	**0.080 ±0.017**	**0.080 ±0.017**
16	rs3745465	19	As	**0.160 ±0.023**	**0.080 ±0.017**
17	rs10411117	19	Af	**0.200 ±0.025**	**0.120 ±0.020**
18	rs270771	19	Eu	**0.160 ±0.023**	**0.120 ±0.020**

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = General populations, including African (Af), Asian (As) and European (Eu). For the complete penetrance low prevalence model, two trajectory groups instead of three are used (see Methodology – Genetic Models). For each setting 1000 replicates are generated. Africans n = 466, Asians n = 359, Europeans n = 214.

Table 41. **Empirical rejection rates and 95% confidence interval of the multi-locus high prevalence complete penetrance model under experiment V at the nominal 0.05 level.**

#	**Disease genes**	**Chr**	**Pop**	**Complete Penetrance**	
				High Prevalence	
				No PC	**PC**
1	rs3733124	3	As	1.000 ±0.000	1.000 ±0.000
2	rs7355991	3	Af	1.000 ±0.000	1.000 ±0.000
3	rs17195948	3	Eu	1.000 ±0.000	1.000 ±0.000
4	rs1259069	6	Eu	1.000 ±0.000	1.000 ±0.000
5	rs3761998	6	As	1.000 ±0.000	1.000 ±0.000
6	rs9459886	6	Af	1.000 ±0.000	1.000 ±0.000
7	rs12790383	11	Eu	1.000 ±0.000	1.000 ±0.000
8	rs11217935	11	As	1.000 ±0.000	1.000 ±0.000
9	rs11825331	11	Af	1.000 ±0.000	1.000 ±0.000
10	rs17117910	12	As	1.000 ±0.000	1.000 ±0.000
11	rs12822275	12	Eu	1.000 ±0.000	1.000 ±0.000
12	rs1696449	12	Af	1.000 ±0.000	1.000 ±0.000
13	rs2073868	17	As	1.000 ±0.000	1.000 ±0.000
14	rs9899123	17	Af	1.000 ±0.000	1.000 ±0.000
15	rs34742396	17	Eu	1.000 ±0.000	1.000 ±0.000
16	rs3745465	19	As	1.000 ±0.000	1.000 ±0.000
17	rs10411117	19	Af	1.000 ±0.000	1.000 ±0.000
18	rs270771	19	Eu	1.000 ±0.000	1.000 ±0.000

Note: The headings for each column are defined as follows: Chr = Chromosome on which SNP marker is located (see Methodology – Power Simulations). Pop = General populations, including African (Af), Asian (As) and European (Eu). For the complete penetrance low prevalence model, two trajectory groups instead of three are used (see Methodology – Genetic Models). For each setting 1000 replicates are generated. Africans n = 466, Asians n = 359, Europeans n = 214.

3.2.2.1 Null Simulation Results Using Uncorrelated SNPs

The ANOVA of the measure of lack of robustness of validity for the multi-locus model uncorrelated SNPs is shown in Table 42 below. The model has an *R-square* of 0.0350. The overall model is not statistically significant ($F = 1.27,\ p - value = 0.1725$). The results indicate that the statistic $F = 4.34,\ p - value = 0.0375$ for factor PC is significant at the nominal level 0.05. That is, the PC adjustment method helps improve the robustness of validity in the complete penetrance high prevalence model for uncorrelated SNPs.

Table 42. The ANOVA table for the multi-locus model uncorrelated SNPs under experiment V at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lackrobust lackrobust

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	1.287	0.051	1.27	0.1725
Error	874	35.519	0.041		
Corrected Total	899	36.806			

R-Square	Coeff Var	Root MSE	lackrobust Mean
0.035	399.196	0.202	0.051

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Chromosome	5	0.356	0.071	1.75	0.120
Population	2	0.005	0.003	0.07	0.936
PC	1	0.176	0.176	4.34	0.038
chr*pop	10	0.481	0.048	1.18	0.299
chr*pc	5	0.072	0.014	0.35	0.880
pop*pc	2	0.196	0.098	2.41	0.090

3.2.2.2 Null Simulation Results for SNPs Having MAF Correlated with Disease SNP

The ANOVA of the measure of lack of robustness of validity for the matching SNPs is shown in Table 43 below. The model has an R-*square* of 0.0453. These results indicate that the overall model is statistically significant ($F = 1.66,\ p - value = 0.0228$) at the nominal 0.05 level. The value of the statistic $F = 4.83\ (p - value = 0.0282)$ for the factor PC, and the value $F = 2.52\ (p - value = 0.0054)$ for the interaction between PC and chromosome are both statistically significant. This analysis documents that the use of PC adjustment significantly improves robustness of validity.

Table 43. The ANOVA table for the multi-locus model matching SNPs under experiment V at the nominal 0.05 level (dependent variable: lack of robustness of validity).

The GLM Procedure

Dependent Variable: lackrobust lackrobust

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	3.087	0.123	1.66	0.023
Error	874	65.101	0.074		
Corrected Total	899	68.188			

R-Square	Coeff Var	Root MSE	lackrobust Mean
0.045	282.820	0.273	0.097

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Chromosome	5	0.396	0.079	1.06	0.379
Population	2	0.239	0.120	1.61	0.201
PC	1	0.360	0.360	4.83	0.028
chr*pop	10	1.877	0.188	2.52	0.005
chr*pc	5	0.191	0.038	0.51	0.767
pop*pc	2	0.023	0.012	0.16	0.855

3.2.2.3 Statistical Analysis on Experiment IV

Table 44 reports the average of the rejection rate for the disease SNPs and lack of robustness of validity measure with or without PC adjustment. Each entry is the average over 18 settings (six chromosome settings times three population settings). The results indicate that the PC null rejection rates are closer to the nominal level than the rates without PC adjustment. For disease SNPs, the rejection rate with PC adjustment is only slightly less than the rejection rate without PC adjustment.

Table 44. Average rejection rates and lack of robustness of validity on Factor PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
Uncorrelated SNPs (null I)	0.069	0.038	0.065	0.037
Matching SNPs (null II)	0.127	0.082	0.116	0.077
Disease SNPs (Power)	1.000	1.000		

I also calculate the average of type I error rate, power and lack of robustness of validity with or without PC by the three general populations: African, Asian and European. As before, the results indicate that global PC method adjusts the type I error rate well overall, especially for the African and Asian population, but less effectively for the European population. The rejection rates for the disease genes are roughly the same for each population. More importantly, the rejection rate with PC adjustment is roughly equal to the rate without adjustment.

Table 45. Average rejection rates and lack of robustness of validity on Factors population and PC.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
African				
Uncorrelated SNPs (null I)	0.093	0.020	0.086	0.020
Matching SNPs (null II)	0.113	0.073	0.104	0.069
Disease SNPs (Power)	1.000	1.000		
Asian				
Uncorrelated SNPs (null I)	0.047	0.053	0.044	0.051
Matching SNPs (null II)	0.120	0.060	0.110	0.057
Disease SNPs (Power)	1.000	1.000		
European				
Uncorrelated SNPs (null I)	0.067	0.040	0.063	0.039
Matching SNPs (null II)	0.147	0.113	0.135	0.104
Disease SNPs (Power)	1.000	1.000		

I calculate the average of type I error rate, power and lack of robustness of validity with or without PC by the six chromosomes: chromosome 3, 6, 11, 12, 17 and 19. The small chromosome, chromosome 19, needs PC adjustment more than other chromosomes. Overall, PC adjustment improves robustness of validity.

Table 46. Average rejection rates and lack of robustness of validity on factor chromosome.

Overall Average	Rejection Rate		Lack of Robustness of Validity	
	No PC	PC	No PC	PC
Chromosome 3				
Uncorrelated SNPs (null I)	0.067	0.040	0.062	0.038
Matching SNPs (null II)	0.107	0.080	0.099	0.075
Disease SNPs (Power)	1.000	1.000		
Chromosome 6				
Uncorrelated SNPs (null I)	0.027	0.000	0.026	0.003
Matching SNPs (null II)	0.093	0.067	0.087	0.062
Disease SNPs (Power)	1.000	1.000		
Chromosome 11				
Uncorrelated SNPs (null I)	0.107	0.040	0.099	0.038
Matching SNPs (null II)	0.107	0.080	0.099	0.075
Disease SNPs (Power)	1.000	1.000		
Chromosome 12				
Uncorrelated SNPs (null I)	0.040	0.040	0.038	0.038
Matching SNPs (null II)	0.187	0.080	0.170	0.075
Disease SNPs (Power)	1.000	1.000		
Chromosome 17				
Uncorrelated SNPs (null I)	0.080	0.040	0.075	0.038
Matching SNPs (null II)	0.093	0.080	0.087	0.075
Disease SNPs (Power)	1.000	1.000		
Chromosome 19				
Uncorrelated SNPs (null I)	0.093	0.067	0.087	0.063
Matching SNPs (null II)	0.173	0.107	0.158	0.099
Disease SNPs (Power)	1.000	1.000		

Chapter 4 Discussion and Conclusion

In this dissertation, I assessed whether PC adjustment was necessary in longitudinal data and whether PC adjustment reduced the inflation of the significance level resulting from PS. The BPP of participants of the clinically important group was used as the quantitative trait. I simulated two types of disease models, the single-locus disease model and the multi-locus disease model. In the single-locus disease model, I assumed that the disease was caused by a single gene, and I used six SNPs across three general populations (African, Asian and European) as disease SNPs: with three MAFs at 0.01, one MAF at 0.05, one at 0.15 and one at 0.30 respectively. In the multi-locus disease model, I assumed that the disease was caused by 18 SNPs, each with MAF smaller than 0.01. I conducted null simulations and power simulations. I considered data simulated under five experiments: 1. the single-locus complete penetrance high prevalence model experiment with disease

SNPs MAFs at 0.01; 2. the single-locus additive model experiment with disease SNPs MAFs at 0.05, 0.15 and 0.30; 3. the complete penetrance high prevalence model experiment with disease SNPs MAFs at 0.05, 0.15 and 0.30; 4. the multi-locus additive model experiment; 5. the multi-locus complete penetrance high prevalence model experiment. I reported the empirical type I error rates and powers to detect the disease SNPs using these genetic models.

The null simulations suggested that the global PC adjustment method helped adjust for PS. The PC method significantly improved robustness of validity of this association procedure. The PC method maintained correct type I error rates with SNPs that have MAFs across population uncorrelated with the disease SNPs. However, the PC method may be problematic with the SNPs that have MAFs across populations that match the disease SNPs. The PC adjustment method had rejection rates above the nominal level for these correlated non-causal SNPs when the genetic association was strong.

The power simulations in this work indicated that multi-locus models both with and without PC adjustment had high power to detect the disease SNPs (>86.6% for multi-locus model). For the single-locus models, the power to detect the disease SNPs increased as the MAF increased. For example, with MAFs 0.01, the power of both models was greater than 56.4%, while with MAFs 0.05, 0.15 and 0.30, the powers were greater than 98.0%.

The questions in my research were: 1. Is PS an issue in longitudinal studies? 2. Does PC correct it? My conclusions are: 1. Yes, PS was an issue in longitudinal studies. 2. Yes, PC corrected PS quite substantially, but not completely. The PC adjustment method helps improve the robustness of validity in the uncorrelated null simulations. The use of PC adjustment significantly improves robustness of validity in a null simulation with SNPs that have MAF across population matching the disease SNPs. The use of PC adjustment and interactions involving PC adjustment in the power simulations are not significant. That is, the fitted rejection rate with PC adjustment is statistically equal to the fitted rate without PC adjustment in a power simulation.

There are alternate rules for choosing the number of trajectory groups. For example, many researchers only consider models in which each trajectory group exceeds a threshold, often 10% of the sample. These rules were not used here.

In future work, there are multiple issues that I plan to explore. Specifically, is the inflation of rejection rate for matching SNPs a practical problem? Many correlated SNPs may appear marginally significant. But will there be a uniquely strong association that obscures the true association? Secondly, I will consider the use of genomic control or other methods of adjustment for further study. For example, admixture methods, propensity scores and the local PC adjustment method could also be used. Further study may also observe the effect of gene-environment interactions through the use of environmental covariates.

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19(9):1655-1664.

Bouaziz M, Ambroise C, Guedj M. 2011a. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. PLOS ONE 6.

Bouaziz M, Ambroise C, Guedj M. 2011b. Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. Plos One 6(12).

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. 2005. Demonstrating stratification in a European American population. Nature Genetics 37(8):868-872.

Cardon LR, Palmer LJ. 2003. Population stratification and spurious allelic association. Lancet 361(9357):598-604.

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, McClay J, Mill J, Martin J, Braithwaite A and others. 2003. Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. Science 301(5631):386-389.

Chen HS, Zhu X, Zhao H, Zhang S. 2003. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Annals of Human Genetics 67:250-264.

Cheng KF, Lin WJ. 2007. Simultaneously correcting for population stratification and for genotyping error in case-control association studies. American Journal of Human Genetics 81(4):726-743.

Deng H. 2001. Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 159:1319-1323.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics

55(4):997-1004.

Divers J, Vaughan LK, Padilla MA, Fernandez JR, Allison DB, Redden DT. 2007. Correcting for measurement error in individual ancestry estimates in structured association tests. Genetics 176(3):1823-1833.

Epstein MP, Allen AS, Satten GA. 2007. A simple and improved correction for population stratification in case-control studies. American Journal of Human Genetics 80(5):921-930.

Ewens WJ, Spielman RS. 1995. THE TRANSMISSION DISEQUILIBRIUM TEST - HISTORY, SUBDIVISION AND ADMIXTURE. American Journal of Human Genetics 57(2):455-464.

Guan WH, Liang LM, Boehnke M, Abecasis GR. 2009. Genotype-Based Matching to Correct for Population Stratification in Large-Scale Case-Control Genetic Association Studies. Genetic Epidemiology 33(6):508-517.

Hao K, Li C, Rosenow C, Wong WH. 2004. Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip (R) Human Mapping 10K array. European Journal of Human Genetics 12(12):1001-1006.

Heiman GA, Hodge SE, Gorroochurn P, Zhang J, Greenberg DA. 2004a. Effects of population stratification on false positive rates in association analysis: A simulation study. American Journal of Epidemiology 159(11):S25-S25.

Heiman GA, Hodge SE, Gorroochurn P, Zhang JY, Greenberg DA. 2004b. Effect of population stratification on case-control association studies - I. Elevation in false positive rates and comparison to confounding risk ratios (a simulation study). Human Heredity 58(1):30-39.

Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR, Ballinger DG. 2004. Matching strategies for genetic association studies in structured populations. American Journal of Human Genetics 74(2):317-325.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42(4):348-U110.

Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. 2007. A Randomization test for controlling population stratification in whole-genome association studies. American Journal of Human Genetics 81(5):895-905.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-753.

Marchini J, Cardon MS, Phillips P, Donnelly P. 2004. The effects of human population structure on large genetic asociation studies. Nat Genet 36:512-517.

Menozzi P, Piazza A, Cavallisforza L. 1978. SYNTHETIC MAPS OF HUMAN GENE-FREQUENCIES IN EUROPEANS. Science 201(4358):786-792.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40(5):646-649.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. Plos Genetics 2(12):2074-2093.

Price A. Eigensoft software.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904-909.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11:459-463.

Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65(1):220-228.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2):945-959.

Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. Genetic Epidemiology 20(1):4-16.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science 298(5602):2381-2385.

Seldin MF, Price AL. 2008. Application of ancestry informative markers to association studies in European Americans. Plos Genetics 4(1).

Tian C, Gregersen PK, Seldin MF. 2008a. Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics 17:R143-R150.

Tian C, Kosoy R, Lee A, Ransom M, Belmont JW, Gregersen PK, Seldin MF. 2008b. Analysis of East Asia Genetic Substructure Using Genome-Wide SNP Arrays. Plos One 3(12).

Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi LH, Gregersen PK and others. 2008c. Analysis and application of European genetic substructure using 300 KSNP information. Plos Genetics 4(1).

Tiwari HK, Barnholtz-Sloan J, Wineinger N, Padilla MA, Vaughan LK, Allison DB. 2008. Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. Human Heredity 66(2):67-86.

Tsai HJ, Choudhry S, Naqvi M, Rodriguez-Cintron W, Burchard EG, Ziv E. 2005. Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. Human Genetics 118(3-4):424-433.

Wacholder S, Rothman N, Caporaso N. 2000. Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. J Natl

Cancer Inst 92(14):1151-1158.

Wang LY, Lee WC. 2008. Population stratification bias in the case-only study for gene-environment interactions. American Journal of Epidemiology 168(2):197-201.

Wang YT, Localio R, Rebbeck TR. 2006. Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions. Cancer Epidemiology Biomarkers & Prevention 15(1):124-132.

Wise CA, Gao X, Shoemaker S, Gordon D, Herring JA. 2008. Understanding genetic factors in idiopathic scoliosis, a complex disease of childhod. Curr Genom, 9:51-59.

Zhao HQ, Rebbeck TR, Mitra N. 2009. A Propensity Score Approach to Correction for Bias due to Population Stratification Using Genetic and Non-Genetic Factors. Genetic Epidemiology 33(8):679-690.

Zhu XF, Zhang SL, Zhao HY, Cooper RS. 2002. Association mapping, using a mixture model for complex traits. Genetic Epidemiology 23(2):181-196.

Zieve D. 2011. Signs of scoliosis. ADAM.

Ziv E, Burchard EG. 2003. Human population structure and genetic association studies. Pharmacogenomics 4(4):431-441.

Appendix A

1. Table 1. Number of rare variants by chromosome in GAW 17 database.

Chromosome	***Number of markers with MAF<0.01***	***Total number of markers***	***Percentage of markers with MAF<0.01***
1	*1713*	*2237*	*76.6%*
2	*1225*	*1599*	*76.6%*
3	***970***	*1211*	*80.1%*
4	*745*	*944*	*78.9%*
5	*814*	*1074*	*75.8%*
6	***1059***	*1425*	*74.3%*
7	*794*	*1063*	*74.7%*
8	*705*	*982*	*71.8%*
9	*794*	*1166*	*68.1%*
10	*1003*	*1396*	*71.8%*
11	***1408***	*2102*	*67.0%*
12	***1022***	*1435*	*71.2%*
13	*330*	*425*	*77.6%*
14	*638*	*795*	*80.3%*
15	*738*	*933*	*79.1%*
16	*641*	*844*	*75.9%*
17	***924***	*1223*	*75.6%*
18	*496*	*634*	*78.2%*
19	***1092***	*1649*	*66.2%*
20	*431*	*591*	*72.9%*
21	*195*	*251*	*77.7%*
22	*394*	*508*	*77.6%*
Total	***18131***	***24487***	**----**

2. **Table 2. Missing data information for the Hapmap 3 database.**

Chromosome	***Total number of Markers***	***Average Genotyping rate in remain individuals***
1	*119487*	*0.997255*
2	*119502*	*0.997144*
3	*98971*	*0.997186*
4	*88135*	*0.997014*
5	*90368*	*0.997191*
6	*93671*	*0.997199*
7	*77377*	*0.997078*
8	*77111*	*0.99704*
9	*65251*	*0.997159*
10	*75616*	*0.997296*
11	*72993*	*0.997152*
12	*70482*	*0.997273*
13	*53293*	*0.997071*
14	*46655*	*0.996875*
15	*43309*	*0.997363*
16	*45778*	*0.997478*
17	*39329*	*0.997446*
18	*41942*	*0.997068*
19	*26953*	*0.997271*
20	*37159*	*0.997184*
21	*19802*	*0.997007*
22	*20649*	*0.997386*
Overall	*1423833*	*0.997188*

Appendix B

PLINK Codes

1. Basic Information

PLINK is a C/C++ command line program. At the command prompt, one should type in "plink" followed by “--options” to specify the data inputs or analysis to be used. The references for all options in PLINK is given in the link below:

http://pngu.mgh.harvard.edu/~purcell/plink/reference.shtml

To read in data files, the command is "plink --file mydata”. The data files are in two formats: the ped file and the map file. i.e., mydata.ped and mydata.map. If the PED and MAP file names are different, one should specify them separately using the command: "plink --ped mydata1.ped --map mydata2.map".

The PED file contains the pedigree and gene information for each individual in a sample. It is a space/tab delimited file including the following columns: family ID, individual ID, paternal ID, maternal ID, sex (1=male; 2=female; other=unknown), phenotype (0=missing; 1=unaffected; 2=affected). The phenotype could be quantitative traits (QT) instead of case-controls. Researchers could specify QTs with certain options in PLINK if the phenotypes are not case-controls. The first six columns are fixed and required in PLINK. But one can use commands to indicate certain missing fields. i.e., "--no-fid" indicates there is no Family ID column; "--no-parents" indicates there are no paternal and maternal ID columns; "--no-sex"indicates there is no sex field; "--no-pheno" indicates there is no phenotype column.

The MAP file contains the genotype location information. Each line of the MAP file describes a single marker and must contain exactly 4 columns: chromosome (1-22, X, Y or 0 if unplaced), marker name, genetic distance (in morgan), base-pair position. If the genetic distance is missing, a flag of "--map3" can be added.

2. Summary Statistics

(1) Hardy-Weinberg Equilibrium

The command of testing HWE is "plink --file data --hardy". An output file of plink.hwe will be created. It has the following columns:

SNP: SNP identifier; TEST: code indicating sample; A1: minor allele code; A2: major allele code; GENO: genotype counts: A1A1/A1A2/A2A2; O(HET): observed heterozygosity; E(HET): expected heterozygosity; P: HW p-value.

If the p-value of HWE test is significant, the SNP considered is not in HWE.

(2) Minor Allele Frequency (MAF)

PLINK could generate the MAF for each SNP under study using the commands "plink --file data --freq". An output file of plink.frq will be created with five columns:

CHR: chromosome; SNP: SNP identifier; A1: allele 1 code (minor allele); A2: allele 2 code (major allele); MAF: minor allele frequency; NCHROBS: non-missing allele count.

3. Association Analysis

(1) Case-control Association Test

The case-control association test could be performed using commands "plink --file mydata --assoc". An output file of plink.assoc will be created with columns:

CHR: chromosome; SNP: SNP ID; BP: base-pair; A1: minor allele name; F_A: frequency of this allele in cases; F_U: frequency of this allele in controls; A2: major allele name; CHISQ: basic allelic test chi-square (1df); P: asymptotic p-value for this test; OR: estimated odds ratio.

A SNP with significant p-value is considered to be associated with the disease. In addition, when the option "--ci 0.95" is added, the columns "L95: lower bound of 95% CI for odds ratio" and "U95: upper bound of 95% CI for odds ratio" will be included in the output.

(2) Quantitative Trait Association

If the phenotype in the 6th column of the PED file is quantitative, with the same commands given in a case-control association study, a quantitative trait analysis will be automatically performed in PLINK. An output file of plink.assoc will include the following columns:

CHR: chromosome; SNP: SNP ID; BP: base-pair; NMISS: # of non-missing genotypes; BETA: regression coefficient; SE: standard error; R2: regression r-squared; T: Wald test t-statistic; P: Wald test asymptotic p-value.

In my study, I used BPPs of the clinically important group as the quantitative traits.

Appendix C

SAS Codes

```
/* MULTI LOCI PARTIAL PENETRANCE MODEL WITH 2 TRAJ GROUPS: C(50%), F(50%) */

LIBNAME TRAJ "C:\hapmap\yifan";

/* READ IN 'COUNTALLIID' WITH JUST IID AND SUMCT VARIABLES */

DATA CURVE;
    SET TRAJ.COUNTALLIID;
    U=UNIFORM(0);
    IF SUMCT=. THEN DELETE;
    ELSE IF SUMCT=0 OR SUMCT=1 AND U<=0.9 THEN GRP='C';
    ELSE IF SUMCT=0 OR SUMCT=1 AND 0.9<U=<1 THEN GRP='F';
    ELSE IF SUMCT>1 AND U<=0.1 THEN GRP='C';
    ELSE IF SUMCT>1 AND 0.1<U<=1 THEN GRP='F';
    T1=0.25;
    T2=0.4;
    T3=0.55;
    T4=0.7;
    T5=0.85;
    T6=1;
    IF GRP='C' THEN DO;
        CURVE1=15+4*RANNOR(0);
        CURVE2=15+4*RANNOR(0);
        CURVE3=15+4*RANNOR(0);
        CURVE4=15+4*RANNOR(0);
        CURVE5=15+4*RANNOR(0);
        CURVE6=15+4*RANNOR(0);
        END;
    ELSE IF GRP='F' THEN DO;
        CURVE1=15+56*(T1-0.25)+4*RANNOR(0);
        CURVE2=15+56*(T2-0.25)+4*RANNOR(0);
        CURVE3=15+56*(T3-0.25)+4*RANNOR(0);
        CURVE4=15+56*(T4-0.25)+4*RANNOR(0);
        CURVE5=15+56*(T5-0.25)+4*RANNOR(0);
```

```
          CURVE6=15+56*(T6-0.25)+4*RANNOR(0);
          END;
RUN;


/* PROC TRAJ FOR 2 CLASSES */
PROC TRAJ DATA=CURVE OUTPLOT=OP OUTSTAT=OS OUT=OF2 OUTEST=OE2 ITDETAIL;
    ID IID; VAR CURVE1-CURVE6; INDEP T1-T6;
    MODEL CNORM; MAX 1000; MIN -1000; NGROUPS 2; ORDER 2 2;
RUN;
/*%TRAJPLOT(OP,OS,'MULTI CMPLT PENETRANCE MODEL II-1 2Q CLASSES','Cnorm
Model','Dependent Variable','Scaled time')

/* PROC TRAJ FOR 3 CLASSES */
PROC TRAJ DATA=CURVE OUTPLOT=OP OUTSTAT=OS OUT=OF3 OUTEST=OE3 ITDETAIL;
    ID IID; VAR CURVE1-CURVE6; INDEP T1-T6;
    MODEL CNORM; MAX 1000; MIN -1000; NGROUPS 3; ORDER 2 2 2;
RUN;
/*%TRAJPLOT(OP,OS,'MULTI CMPLT PENETRANCE MODEL II-1 3Q CLASSES','Cnorm
Model','Dependent Variable','Scaled time')

/* PROC TRAJ FOR 4 CLASSES */
PROC TRAJ DATA=CURVE OUTPLOT=OP OUTSTAT=OS OUT=OF4 OUTEST=OE4 ITDETAIL;
    ID IID; VAR CURVE1-CURVE6; INDEP T1-T6;
    MODEL CNORM; MAX 1000; MIN -1000; NGROUPS 4; ORDER 2 2 2 2;
RUN;
/*%TRAJPLOT(OP,OS,'MULTI CMPLT PENETRANCE MODEL II-1 4Q CLASSES','Cnorm
Model','Dependent Variable','Scaled time')

/* PROC TRAJ FOR 5 CLASSES */
PROC TRAJ DATA=CURVE OUTPLOT=OP OUTSTAT=OS OUT=OF5 OUTEST=OE5 ITDETAIL;
    ID IID; VAR CURVE1-CURVE6; INDEP T1-T6;
    MODEL CNORM; MAX 1000; MIN -1000; NGROUPS 5; ORDER 2 2 2 2 2;
RUN;
/*%TRAJPLOT(OP,OS,'MULTI CMPLT PENETRANCE MODEL II-1 5Q CLASSES','Cnorm
Model','Dependent Variable','Scaled time')

/* PROC TRAJ FOR 6 CLASSES */
PROC TRAJ DATA=CURVE OUTPLOT=OP OUTSTAT=OS OUT=OF6 OUTEST=OE6 ITDETAIL;
    ID IID; VAR CURVE1-CURVE6; INDEP T1-T6;
    MODEL CNORM; MAX 1000; MIN -1000; NGROUPS 6; ORDER 2 2 2 2 2 2;
```

```
RUN;
/*%TRAJPLOT(OP,OS,'MULTI CMPLT PENETRANCE MODEL II-1 6Q CLASSES','Cnorm
Model','Dependent Variable','Scaled time')


/* MAKE PHENOTYPE FILE TO INPUT TO PLINK */
DATA FIDIID;
    SET TRAJ.FIDIID;
    KEEP FID IID;
RUN;

/* READ IN ALL 'OF' DATA, KEEP BPP COLUMN---GRPPROB. ADD 'GRP' COLUMN TO KEEP
A COLUMN POSITION BEFORE MERGE */
DATA PHENO;
    MERGE FIDIID OF2 (KEEP = IID GRP2PRB) OF3 (KEEP = IID GRP3PRB)
                        OF4 (KEEP = IID GRP4PRB) OF5 (KEEP = IID GRP5PRB) OF6 (KEEP =
IID GRP6PRB);
    BY IID;
    GRP=0;
RUN;

/* TAKE THE FIRST ROW, PARMS, OF THE _TYPE_ COLUMN, DATASET OEE ONLY HAS
ONE ROW */
/* SAS CAN'T RECOGNIZE _TYPE_ IN 'OE', SO I USED 'OBS=1' TO READ ONLY THE VERY
FIRST ROW */
DATA OEE;
    MERGE OE2 (KEEP=_BIC1_ RENAME=(_BIC1_=BIC2) OBS=1)
          OE3 (KEEP=_BIC1_ RENAME=(_BIC1_=BIC3) OBS=1) OE4 (KEEP=_BIC1_
RENAME=(_BIC1_=BIC4) OBS=1)
          OE5 (KEEP=_BIC1_ RENAME=(_BIC1_=BIC5) OBS=1) OE6 (KEEP=_BIC1_
RENAME=(_BIC1_=BIC6) OBS=1);
RUN;

/* FIND THE LARGEST BIC FROM 1 TO 6 CLASSES. THERE IS ONLY ONE VALUE IN
VARIABLE 'BIG' */
DATA BIG;
    SET OEE;
                           BIG=BIC1; N=1;
    IF BIC2>BIG THEN DO BIG=BIC2; N=2; END;
    IF BIC3>BIG THEN DO BIG=BIC3; N=3; END;
        IF BIC4>BIG THEN DO BIG=BIC4; N=4; END;
        IF BIC5>BIG THEN DO BIG=BIC5; N=5; END;
        IF BIC6>BIG THEN DO BIG=BIC6; N=6; END;
        KEEP N;
RUN;

/* MAKE A 613 ROWS MATRIX, WITH ID AND N VALUES */
DATA BIG613;
        SET BIG;
        DO ID=1 TO 613;
        OUTPUT;
        END;
RUN;

/* MAKE A QT TABLE, PUT THE BPP--GRPPRB DATA INTO 'GRP', KEEP THE 'N' COLUMN
FOR REFERENCE */
DATA QT;
        MERGE PHENO BIG613;
                          GRP= GRP1PRB;
        IF N=2 THEN    GRP= GRP2PRB;
        IF N=3 THEN    GRP= GRP3PRB;
        IF N=4 THEN    GRP= GRP4PRB;
        IF N=5 THEN    GRP= GRP5PRB;
        IF N=6 THEN    GRP= GRP6PRB;
        IF GRP = . THEN GRP = -9;
        KEEP FID IID GRP N;
RUN;

PROC EXPORT DATA= WORK.QT
                OUTFILE= "C:\hapmap\yifan\QT.txt"
                DBMS=TAB REPLACE;
RUN;
```

Appendix D

List of Uncorrelated and Matching SNPs for Multi-locus Model

num	group	uncorrelated SNPs	matching SNPs
1	snp01-01	rs9384246	rs16887596
2	snp01-02	rs9496769	rs766797
3	snp01-03	rs8182554	rs16889893
4	snp01-04	rs483574	rs9353978
5	snp01-05	rs2397132	rs9342347
6	snp01-06	rs956952	rs16870107
7	snp01-07	rs9834682	rs2034153
8	snp01-08	rs16884048	rs17062802
9	snp01-09	rs1551524	rs16963260
10	snp01-10	rs6798416	rs16964551
11	snp01-11	rs3772547	rs4955653
12	snp01-12	rs750438	rs17071850
13	snp01-13	rs9365263	rs12490747
14	snp01-14	rs1603537	rs1362525
15	snp01-15	rs12226382	rs3018622
16	snp01-16	rs11032481	rs3960851
17	snp01-17	rs4932691	rs2362191
18	snp01-18	rs4431401	rs4363010
19	snp01-19	rs4858960	rs12488302
20	snp01-20	rs11129414	rs17148932
21	snp01-21	rs10877383	rs12485909
22	snp01-22	rs7966445	rs16935493
23	snp01-23	rs9397313	rs3794077
24	snp01-24	rs11025588	rs4688741
25	snp01-25	rs562516	rs4688682
26	snp02-01	rs9404115	rs6445618
27	snp02-02	rs4387423	rs16860782
28	snp02-03	rs202069	rs11833193
29	snp02-04	rs937761	rs6926129
30	snp02-05	rs4801702	rs699637
31	snp02-06	rs4959793	rs6791183
32	snp02-07	rs6939425	rs12311968

33	snp02-08	rs687660	rs16829583
34	snp02-09	rs11918801	rs1366244
35	snp02-10	rs11867497	rs9447191
36	snp02-11	rs9813221	rs9877433
37	snp02-12	rs2356046	rs4992086
38	snp02-13	rs379977	rs7258703
39	snp02-14	rs6937313	rs12293932
40	snp02-15	rs523179	rs1776450
41	snp02-16	rs807858	rs9848710
42	snp02-17	rs4470547	rs13434278
43	snp02-18	rs13091924	rs2700221
44	snp02-19	rs12451743	rs7111830
45	snp02-20	rs4796835	rs12284508
46	snp02-21	rs9472686	rs17035243
47	snp02-22	rs2876586	rs1349434
48	snp02-23	rs814022	rs17079769
49	snp02-24	rs515246	rs6505497
50	snp02-25	rs726610	rs17026647
51	snp03-01	rs16933427	rs12201208
52	snp03-02	rs3884325	rs12201692
53	snp03-03	rs1596071	rs17660589
54	snp03-04	rs4767174	rs11130981
55	snp03-05	rs7954843	rs12208647
56	snp03-06	rs9284357	rs6937229
57	snp03-07	rs11104708	rs11922676
58	snp03-08	rs9873052	rs4686787
59	snp03-09	rs9273012	rs497704
60	snp03-10	rs7138898	rs1146240
61	snp03-11	rs17068440	rs11023888
62	snp03-12	rs2143071	rs17365525
63	snp03-13	rs3138289	rs12977468
64	snp03-14	rs11552205	rs1542123
65	snp03-15	rs7610823	rs2327748
66	snp03-16	rs9504044	rs2044124
67	snp03-17	rs2495964	rs17526236
68	snp03-18	rs4789846	rs17606030
69	snp03-19	rs2303146	rs17280334
70	snp03-20	rs6807356	rs1013426
71	snp03-21	rs1687310	rs11669191
72	snp03-22	rs1502380	rs35765580
73	snp03-23	rs2201438	rs12806315

74	snp03-24	rs885398	rs17443031
75	snp03-25	rs2236543	rs17421687
76	snp04-01	rs307223	rs17517058
77	snp04-02	rs4889835	rs17260403
78	snp04-03	rs16937972	rs33988791
79	snp04-04	rs2061185	rs1047841
80	snp04-05	rs9484448	rs2184925
81	snp04-06	rs12575969	rs17730847
82	snp04-07	rs12790182	rs41457949
83	snp04-08	rs12227286	rs11042572
84	snp04-09	rs13059911	rs11222105
85	snp04-10	rs1841704	rs9268219
86	snp04-11	rs7639226	rs17377726
87	snp04-12	rs6502546	rs13212023
88	snp04-13	rs10936033	rs1934793
89	snp04-14	rs7213831	rs12948969
90	snp04-15	rs4688381	rs4122113
91	snp04-16	rs789224	rs11047534
92	snp04-17	rs563385	rs17194345
93	snp04-18	rs3826301	rs12950551
94	snp04-19	rs9311833	rs11651302
95	snp04-20	rs12577984	rs7252322
96	snp04-21	rs33936986	rs1788279
97	snp04-22	rs1535708	rs568131
98	snp04-23	rs332496	rs3132453
99	snp04-24	rs9736016	rs8100439
100	snp04-25	rs2061907	rs17303478
101	snp05-01	rs1101834	rs4685047
102	snp05-02	rs719365	rs17144371
103	snp05-03	rs1945318	rs6266
104	snp05-04	rs16881458	rs12580498
105	snp05-05	rs870601	rs2306882
106	snp05-06	rs10501851	rs3744234
107	snp05-07	rs11033093	rs2234376
108	snp05-08	rs1547589	rs2280523
109	snp05-09	rs9364689	rs310467
110	snp05-10	rs11229425	rs2302644
111	snp05-11	rs28360477	rs11871642
112	snp05-12	rs12190869	rs2286406
113	snp05-13	rs2099015	rs3800370
114	snp05-14	rs11819769	rs2306260

115	snp05-15	rs11236449	rs11025368
116	snp05-16	rs13061863	rs12631683
117	snp05-17	rs2144425	rs17831672
118	snp05-18	rs2116984	rs11068493
119	snp05-19	rs4235835	rs12582287
120	snp05-20	rs12709501	rs12665305
121	snp05-21	rs9366653	rs13306166
122	snp05-22	rs10505891	rs16928868
123	snp05-23	rs2971566	rs2293433
124	snp05-24	rs35684970	rs588048
125	snp05-25	rs8178408	rs3799931
126	snp06-01	rs12216227	rs4389808
127	snp06-02	rs9481950	rs6915517
128	snp06-03	rs9841691	rs470414
129	snp06-04	rs9828938	rs7633529
130	snp06-05	rs4513814	rs7207508
131	snp06-06	rs11079280	rs7507584
132	snp06-07	rs2685054	rs11030936
133	snp06-08	rs2633703	rs1280826
134	snp06-09	rs2878960	rs16923710
135	snp06-10	rs577298	rs9503462
136	snp06-11	rs986819	rs10423596
137	snp06-12	rs673547	rs10426529
138	snp06-13	rs2071468	rs584884
139	snp06-14	rs11659000	rs4447137
140	snp06-15	rs11048450	rs12306371
141	snp06-16	rs1569355	rs1400965
142	snp06-17	rs4923544	rs9911743
143	snp06-18	rs3861401	rs7120161
144	snp06-19	rs7950646	rs12321864
145	snp06-20	rs7101994	rs7248452
146	snp06-21	rs11177946	rs11212915
147	snp06-22	rs12205626	rs9466392
148	snp06-23	rs10861761	rs9917063
149	snp06-24	rs1045764	rs8189127
150	snp06-25	rs12611099	rs7218045
151	snp07-01	rs6511401	rs11023498
152	snp07-02	rs8065026	rs17823624
153	snp07-03	rs1413524	rs12812119
154	snp07-04	rs2238099	rs11068315
155	snp07-05	rs6809063	rs35679149

156	snp07-06	rs3851994	rs12193743
157	snp07-07	rs11706015	rs2012061
158	snp07-08	rs6918702	rs11057392
159	snp07-09	rs1237027	rs4151031
160	snp07-10	rs12427294	rs4135255
161	snp07-11	rs4234594	rs3101943
162	snp07-12	rs1433127	rs9658069
163	snp07-13	rs3730363	rs17301388
164	snp07-14	rs17009261	rs16940655
165	snp07-15	rs12495014	rs742310
166	snp07-16	rs11219534	rs17216646
167	snp07-17	rs4294874	rs10832384
168	snp07-18	rs1317850	rs7301331
169	snp07-19	rs9845429	rs1805753
170	snp07-20	rs6444386	rs11023253
171	snp07-21	rs9914748	rs11235726
172	snp07-22	rs1491631	rs1148551
173	snp07-23	rs12207182	rs12205241
174	snp07-24	rs2053623	rs34166957
175	snp07-25	rs7932866	rs3134712
176	snp08-01	rs2102928	rs11220691
177	snp08-02	rs9472773	rs242560
178	snp08-03	rs10412222	rs11218980
179	snp08-04	rs2508822	rs12574726
180	snp08-05	rs4458393	rs3762701
181	snp08-06	rs13192116	rs3823112
182	snp08-07	rs6941603	rs3759383
183	snp08-08	rs1516715	rs16868725
184	snp08-09	rs7356965	rs11042961
185	snp08-10	rs324555	rs11039260
186	snp08-11	rs6919521	rs3778487
187	snp08-12	rs11654323	rs12574939
188	snp08-13	rs1866822	rs12573925
189	snp08-14	rs6922753	rs12661985
190	snp08-15	rs3815612	rs2276347
191	snp08-16	rs10773486	rs12578246
192	snp08-17	rs2214449	rs3826972
193	snp08-18	rs17709409	rs11023825
194	snp08-19	rs12812640	rs17179770
195	snp08-20	rs11609192	rs17164598
196	snp08-21	rs1063155	rs17147781

197	snp08-22	rs28897680	rs17087009
198	snp08-23	rs2879889	rs9388409
199	snp08-24	rs7939071	rs9395524
200	snp08-25	rs2068192	rs17073524
201	snp09-01	rs9381118	rs13320561
202	snp09-02	rs6900447	rs11057084
203	snp09-03	rs3781998	rs17069461
204	snp09-04	rs10834358	rs12313698
205	snp09-05	rs2510757	rs11835574
206	snp09-06	rs534858	rs9891296
207	snp09-07	rs385203	rs1274494
208	snp09-08	rs12216323	rs237967
209	snp09-09	rs13064262	rs6503958
210	snp09-10	rs6903998	rs4490677
211	snp09-11	rs267482	rs16924252
212	snp09-12	rs2511509	rs1883734
213	snp09-13	rs9459954	rs2872833
214	snp09-14	rs13064823	rs4623860
215	snp09-15	rs7128974	rs7926603
216	snp09-16	rs7610345	rs16922432
217	snp09-17	rs1453584	rs7962923
218	snp09-18	rs11226057	rs7406278
219	snp09-19	rs10501367	rs9870541
220	snp09-20	rs16929851	rs2041458
221	snp09-21	rs2069214	rs1869548
222	snp09-22	rs11214769	rs11836913
223	snp09-23	rs4386846	rs12305245
224	snp09-24	rs11652704	rs11112801
225	snp09-25	rs851987	rs7110328
226	snp10-01	rs6800770	rs4714999
227	snp10-02	rs7480678	rs2269347
228	snp10-03	rs11220520	rs11219461
229	snp10-04	rs2699061	rs2270966
230	snp10-05	rs2099048	rs12224420
231	snp10-06	rs9636146	rs4147617
232	snp10-07	rs441116	rs3741668
233	snp10-08	rs2397215	rs11104947
234	snp10-09	rs9866640	rs3772208
235	snp10-10	rs9451194	rs3864111
236	snp10-11	rs1920610	rs16910660
237	snp10-12	rs12460798	rs35773539

238	snp10-13	rs2388788	rs3741506
239	snp10-14	rs9275595	rs2074176
240	snp10-15	rs902557	rs1077521
241	snp10-16	rs12208401	rs11168709
242	snp10-17	rs9480502	rs2285060
243	snp10-18	rs9347140	rs3828741
244	snp10-19	rs9834678	rs9384252
245	snp10-20	rs1204331	rs3745327
246	snp10-21	rs810912	rs2291931
247	snp10-22	rs2066951	rs3763944
248	snp10-23	rs11177368	rs16881056
249	snp10-24	rs7108229	rs3734265
250	snp10-25	rs4964963	rs2306800
251	snp11-01	rs6414595	rs35191042
252	snp11-02	rs7117433	rs11021266
253	snp11-03	rs6916028	rs17578530
254	snp11-04	rs1532720	rs17207518
255	snp11-05	rs1939066	rs1463298
256	snp11-06	rs953730	rs1607394
257	snp11-07	rs4857926	rs12192975
258	snp11-08	rs8078764	rs485118
259	snp11-09	rs2245897	rs7647281
260	snp11-10	rs9386508	rs17769930
261	snp11-11	rs12819780	rs10510943
262	snp11-12	rs2455799	rs1489107
263	snp11-13	rs7383248	rs1493593
264	snp11-14	rs11172162	rs8109030
265	snp11-15	rs10791048	rs4134950
266	snp11-16	rs11825966	rs11039758
267	snp11-17	rs10492338	rs12804520
268	snp11-18	rs3891724	rs41420445
269	snp11-19	rs41463648	rs11542187
270	snp11-20	rs9876212	rs3687
271	snp11-21	rs2458304	rs12817914
272	snp11-22	rs12529487	rs11650611
273	snp11-23	rs845891	rs12426207
274	snp11-24	rs8073910	rs11064748
275	snp11-25	rs537225	rs17787343
276	snp12-01	rs9344315	rs7134845
277	snp12-02	rs225676	rs7760797
278	snp12-03	rs9847394	rs16924539

279	snp12-04	rs6590683	rs10877755
280	snp12-05	rs4687420	rs9825867
281	snp12-06	rs1868500	rs6767873
282	snp12-07	rs759679	rs17136321
283	snp12-08	rs760827	rs7955386
284	snp12-09	rs2227371	rs7968828
285	snp12-10	rs669776	rs7216862
286	snp12-11	rs7741797	rs7956459
287	snp12-12	rs6793160	rs17045159
288	snp12-13	rs7254543	rs9310100
289	snp12-14	rs16830730	rs711176
290	snp12-15	rs4931083	rs6782165
291	snp12-16	rs12419421	rs9880895
292	snp12-17	rs7124639	rs7135641
293	snp12-18	rs6539676	rs28438465
294	snp12-19	rs2617688	rs9893918
295	snp12-20	rs1861419	rs16827421
296	snp12-21	rs7953305	rs6806316
297	snp12-22	rs7252828	rs7108570
298	snp12-23	rs6457737	rs4611190
299	snp12-24	rs10049314	rs1454014
300	snp12-25	rs10945649	rs6907480
301	snp13-01	rs11111391	rs12222269
302	snp13-02	rs149942	rs4684678
303	snp13-03	rs3016500	rs12632557
304	snp13-04	rs8113142	rs3815405
305	snp13-05	rs7771441	rs17045968
306	snp13-06	rs4275668	rs11035319
307	snp13-07	rs9385629	rs2273012
308	snp13-08	rs6344	rs3748261
309	snp13-09	rs11825972	rs11040226
310	snp13-10	rs7224296	rs2277634
311	snp13-11	rs6901717	rs2241775
312	snp13-12	rs4792900	rs2301755
313	snp13-13	rs9328513	rs3814729
314	snp13-14	rs4789352	rs3792292
315	snp13-15	rs17834692	rs11525594
316	snp13-16	rs2301570	rs3744451
317	snp13-17	rs7950661	rs28372821
318	snp13-18	rs7224601	rs2291932
319	snp13-19	rs407056	rs11218773

320	snp13-20	rs9275374	rs16924079
321	snp13-21	rs7302533	rs3773639
322	snp13-22	rs588952	rs2279449
323	snp13-23	rs2166909	rs12660257
324	snp13-24	rs328486	rs9397997
325	snp13-25	rs6483748	rs1077646
326	snp14-01	rs17021512	rs7637778
327	snp14-02	rs9856266	rs16967585
328	snp14-03	rs385521	rs17079656
329	snp14-04	rs1729594	rs4767988
330	snp14-05	rs8078351	rs12284342
331	snp14-06	rs3744132	rs16967580
332	snp14-07	rs11234925	rs9676310
333	snp14-08	rs9367137	rs12318030
334	snp14-09	rs1879883	rs963803
335	snp14-10	rs12286721	rs1481459
336	snp14-11	rs10431559	rs9897023
337	snp14-12	rs901816	rs12327812
338	snp14-13	rs7119188	rs11925097
339	snp14-14	rs9819360	rs41426851
340	snp14-15	rs7224763	rs11829043
341	snp14-16	rs4485669	rs7971725
342	snp14-17	rs9322528	rs7948832
343	snp14-18	rs833670	rs12111139
344	snp14-19	rs11666111	rs7940727
345	snp14-20	rs1265067	rs9894979
346	snp14-21	rs6456397	rs9889821
347	snp14-22	rs7967594	rs9897155
348	snp14-23	rs812149	rs9303627
349	snp14-24	rs2272891	rs9913571
350	snp14-25	rs3898124	rs17077048
351	snp15-01	rs1784500	rs11228158
352	snp15-02	rs5004021	rs12194408
353	snp15-03	rs7751661	rs7246367
354	snp15-04	rs7312492	rs11216799
355	snp15-05	rs7750841	rs10440825
356	snp15-06	rs17077267	rs11658083
357	snp15-07	rs2117018	rs2847182
358	snp15-08	rs2000560	rs11058042
359	snp15-09	rs7937892	rs11752309
360	snp15-10	rs11060863	rs17513072

361	snp15-11	rs12270081	rs387233
362	snp15-12	rs6549049	rs2564915
363	snp15-13	rs9871261	rs4764628
364	snp15-14	rs11111146	rs17593921
365	snp15-15	rs3781839	rs11717208
366	snp15-16	rs10502030	rs17455493
367	snp15-17	rs4930597	rs214585
368	snp15-18	rs17061176	rs13199373
369	snp15-19	rs9381402	rs7178
370	snp15-20	rs6458958	rs1925791
371	snp15-21	rs17009623	rs3935910
372	snp15-22	rs1388206	rs2420543
373	snp15-23	rs3821525	rs13218115
374	snp15-24	rs11753865	rs35598292
375	snp15-25	rs12361586	rs11708369
376	snp16-01	rs6936123	rs2272325
377	snp16-02	rs8105273	rs3777674
378	snp16-03	rs9818739	rs2045018
379	snp16-04	rs2304819	rs3746293
380	snp16-05	rs1347104	rs9362034
381	snp16-06	rs4711738	rs3811068
382	snp16-07	rs2268846	rs3814444
383	snp16-08	rs6485673	rs16860184
384	snp16-09	rs6458115	rs17289925
385	snp16-10	rs1043898	rs12637276
386	snp16-11	rs3782120	rs11066280
387	snp16-12	rs11880539	rs11114646
388	snp16-13	rs7935908	rs2232217
389	snp16-14	rs2139077	rs2273568
390	snp16-15	rs1229933	rs3751958
391	snp16-16	rs6804121	rs28372785
392	snp16-17	rs7303170	rs3759313
393	snp16-18	rs12193001	rs4327695
394	snp16-19	rs4686667	rs12665064
395	snp16-20	rs10511022	rs3782886
396	snp16-21	rs16964191	rs2290054
397	snp16-22	rs3760843	rs11066132
398	snp16-23	rs1466235	rs9658068
399	snp16-24	rs9302994	rs279808
400	snp16-25	rs9843433	rs11236583
401	snp17-01	rs2359367	rs7611753

402	snp17-02	rs4680887	rs9830017
403	snp17-03	rs11718449	rs669300
404	snp17-04	rs11219688	rs9474681
405	snp17-05	rs3773935	rs864461
406	snp17-06	rs12494282	rs6932090
407	snp17-07	rs12494110	rs16957396
408	snp17-08	rs7313688	rs6916736
409	snp17-09	rs613197	rs16932733
410	snp17-10	rs6799479	rs4002154
411	snp17-11	rs12213009	rs9465456
412	snp17-12	rs11966093	rs7207237
413	snp17-13	rs3903688	rs2610736
414	snp17-14	rs9296733	rs11021700
415	snp17-15	rs9348055	rs6596984
416	snp17-16	rs1711957	rs11828768
417	snp17-17	rs1793051	rs10422231
418	snp17-18	rs4794291	rs9463381
419	snp17-19	rs3794970	rs16944991
420	snp17-20	rs34430583	rs4297462
421	snp17-21	rs308194	rs17066711
422	snp17-22	rs3860828	rs7117182
423	snp17-23	rs4894850	rs7769577
424	snp17-24	rs2623945	rs11966336
425	snp17-25	rs10744889	rs1864897
426	snp18-01	rs4549	rs11068535
427	snp18-02	rs12202106	rs11653020
428	snp18-03	rs9898132	rs1483121
429	snp18-04	rs543876	rs12818313
430	snp18-05	rs159268	rs13205266
431	snp18-06	rs11657217	rs6936632
432	snp18-07	rs7109038	rs17646359
433	snp18-08	rs2659610	rs1782449
434	snp18-09	rs13068339	rs1802668
435	snp18-10	rs9826798	rs4607423
436	snp18-11	rs17375799	rs10444560
437	snp18-12	rs1397881	rs2243368
438	snp18-13	rs4856154	rs2272450
439	snp18-14	rs12215495	rs11171407
440	snp18-15	rs10501960	rs12211424
441	snp18-16	rs12290811	rs17802736
442	snp18-17	rs10848907	rs1482442

443	snp18-18	rs4945151	rs12825841
444	snp18-19	rs7613917	rs10513541
445	snp18-20	rs13320885	rs9403910
446	snp18-21	rs7980273	rs12818059
447	snp18-22	rs1401454	rs11673632
448	snp18-23	rs4802723	rs1124303
449	snp18-24	rs8108921	rs17657473
450	snp18-25	rs334535	rs12487951