Two Topics in Association Analysis of DNA Sequencing Data: Population Structure and Multivariate Traits

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yiwei Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advised by Wei Pan, Ph.D

August, 2013

Acknowledgements

I would like to express my gratitude to many people. Without their help, I would never imagine accomplishing my study. First, I want to sincerely thank my thesis advisor, Dr. Wei Pan. I respect his profound insights into the problems as well as keen foresight to future questions. In this research, he directs me in a strict and timely way, and spends substantive time discussing with me. It is such fun experience to work with him.

Thanks to my thesis committee members, Dr. Jim Neaton, Dr. Lin Yee Chen and Dr. Weihua Guan, for their time reviewing my thesis and offering constructive comments. Special thanks to Dr. Neaton for sharing the NvR study data and discussing with me about the analysis plan. I am also thankful to Dr. Saonli Basu, whom I did research with for two years and learned much from, and Dr. Ruzong Fan, my advisor in master program in Texas, for his encouragement through out my PhD study.

I also appreciate Minnesota Supercomputing Institute for providing us the advanced computing environment and their considerate assistance, which enables me to finish the genetic analysis on time.

Lastly, I owe my deepest gratitude to my husband, Lipeng Ning for his kind support, and our parents in China. Although they can not be here, I know their hearts are always with us.

Abstract

As the next-generation sequencing technologies become mature and affordable, we now have access to massive data of single nucleotides variants (SNVs) with varying minor allele frequencies (MAFs). This poses new opportunities, as more information from the human genome is available. However, new challenges also show up, such as how to utilize those SNVs with low MAFs. With current intensive efforts in association testing to detect genetic loci associated with common diseases and complex traits, two issues are of primary interest: reducing spurious findings and increasing power for true discoveries.

In association testing, a major cause to the elevated level of false positives is the confounding effect of population structure – the so-called population stratification.

As a remedy, one popular method is to add principal components (PCs) in a regression model, named principal component regression (PCR).

Yet, it is not clear how PCR will work in testing rare variants (RVs, with MAF<0.01), or with population stratification in a fine scale. More questions arise, like what types and what sets of SNVs should be used to construct PCs, and whether there are other better methods than principal component analysis (PCA) for constructing PCs. Utilizing the DNA sequencing data from the 1000 Genomes project, we first investigate whether PCR is adequate in adjusting for population stratification while maintaining high power when testing low frequency variants (LFVs with $0.01\leq$ MAF<0.05) and RVs. Furthermore, we compare the performance of two dimension reduction methods, PCA and spectral dimension reduction (SDR), as well as twelve different types and sets of variants for constructing PCs. The comparison is conducted with respect to controlling population stratification in a fine scale.

On the other hand, linear mixed models (LMM) have emerged with its superior performance in handling complex population structures. Herein, we examine the connection and difference between PCR and LMM based on the formulation of probabilistic PCA, and propose a hybrid method combining the two. Its outstanding performance in addressing both population structure and environmental confounders is established by simulations using the the Genetic Analysis Workshop (GAW) 18

data and the 1000 Genomes project data.

Lastly, we consider boosting power for association analysis of multivariate traits. A new class of tests, the sum of powered score tests (SPU), and an adaptive SPU (aSPU) test are extended to the generalized estimation equations (GEE) framework. We apply the new and some existing methods to association testing on both CVs and RVs with an HIV/AIDS dataset and the GAW 18 data.

Contents

List of Tables

List of Figures

List of Abbreviations

1. aRI: adjusted rand index
2. CV: common variants. It refers to the genetic variants with MAF$\geq$0.05.
3. EMMA: efficient mixed-model association
4. EMMAX: expedited version of EMMA
5. EREC: estimated regression coefficients
6. Fst: fixation index
7. GAW18: genetic analysis workshop 18
8. GWAS: genome-wide association study
9. GEE: generalized estimation equations
10. GEMMA: genome-wide efficient mixed-model association
11. IBS: identity-by-state matrix. It measures the genetic correlation between two subjects as the proportion of loci having the same allele.
12. LD: linkage disequilibrium. It refers to the non-random association of alleles at two or more loci.
13. LFV: low-frequency variant. It refers to the genetic variants with 0.01$\leq$MAF$<$ 0.05.
14. LMM: linear mixed model
15. MAF: minor allele frequency. It refers to the frequency of the genetic variant at which the least common allele occurs in a given population.
16. OLS: ordinary least squares
17. PC: principal component
18. PCA: principal component analysis

19. PCR: principal component regression. It refers to the model of regression with PCs added as covariates.

20. QTL: quantitative trait loci

21. RI: Rand Index

22. RVs: rare variant. It refers to the genetic variants with MAF$<$0.01.

23. SDR: spectral dimension reduction

24. SPU tests: the sum of powered score tests

25. SNV: single-nucleotide variant. It includes common variants, low frequency variants and rare variants.

26. SNP: single-nucleotide polymorphism, which are common variants.

27. λ: inflation factor. It refers to the ratio of the median of the χ^2 statistics corresponding to the observed p-values.

Chapter 1

Background

With the availability of next-generation sequencing data, there has been increasing interest in studying associations between complex traits and low-frequency variants (LFVs), which have minor allele frequencies (MAFs) between 1% and 5%, or rare variants (RVs) with MAF $< 1\%$ [1, 2]. RVs have been shown to play important roles in several diseases [3, 4] and are expected to account for some disease risk beyond common variants (CVs) with MAFs $\geq 5\%$. Plus, as pointed out by Henn et al. (2010) [5], "Rare variants are likely to have recently arisen and segregate between populations and are informative markers of ancestry". Accordingly, this raises research enthusiasm in examining how RVs would perform in detecting population structures [6, 7].

Due to the low MAFs of LFVs and RVs, statistical tests developed for CVs in genome-wide association studies (GWASs) may no longer be powerful, and there have been intensive efforts in developing new statistical tests for LFVs and RVs, see [8] for a comprehensive review and comparison of many existing methods. Although there is no uniformly most powerful test, the sum of squared score (SSU) test demonstrates a generally impressive performance. The SSU test has been shown [9] to be closely related to an empirical Bayes test for high-dimensional data [10], a kernel machine regression (KMR) and sequencing kernel association testing (SKAT) [11, 12, 13], a genomic-distance based regression (GDBR) [14] and C-alpha test [15]. Pooled association tests are another type of popular methods, which simply aggregate information across multiple variants. Their performances may deteriorate dramatically when variants have opposite effects or when many null variants are present. Typical tests in this category include but not confined to, the Sum test [16], the weighted Sum test [17], and the variable threshold (VT) test [18], etc. As the computation technology advances, data adaptive methods are also widely accepted, such as a kernel-based adaptive clustering (KBAC) test [19] and an estimated regression coefficients (EREC) developed by [20].

While there is a fast development in methods, the credibility of association testing is still somewhat blemished by spurious findings. This is partly caused by the confounding effect of population structure – the so-called population stratification. Population stratification occurs most frequently in the case-control study design, where different ancestral populations have varying disease risks and different distributions of genetic variants. Inference of population structure is also critical

in understanding different drug responses of different people [21, 22]. Several approaches prevail to account for population stratification in GWAS. Genomic control (GC) [23] aims to correct the statistics by having them divided by the inflation factor, which is estimated from many test statistics. This method is impaired by unrealistically assuming that all variants share one common inflation factor. The principal component regression (PCR) uses principal component analysis (PCA) to extract components related to ancestries and includes them in a regression model [24, 25]. A related method, structured association, uses variants or a few top principal components (PCs) to cluster the samples before carrying out a stratified test [26]. This approach is more suitable when the population groups are discrete. A recently emerging approach is genetic similarity score matching (GSM) that matches cases and controls based on a genetic similarity score before applying conditional logistic regression [27].

Although PCR is one of the most commonly used methods for accounting for population stratification, several questions still remain unclear. For example, how does the PCR method perform in association testing of RVs, or with population structure in a fine scale? Should the variants be pruned, as suggested by [24, 28], or not to construct a similarity matrix? Do RVs perform better than CVs as expected by [5]? Is PCA the best method to construct PCs? These questions will be discussed in detail in the following chapters.

To account for more complex population structures, Linear mixed models (LMMs) have emerged recently as most promising [29, 30]. Yu et al. (2005) first proposed a model with both a random effect for family correlatedness and a fixed effect for population structure. Recently it is claimed that the LMM without a fixed effect is also capable of addressing both population stratification and cryptic relatedness. Some fast algorithms have been developed for the LMM to overcome the computing bottleneck. Popular packages include efficient mixed-model association (EMMA) [31] and its expedited version EMMAX [32], and an exact method called genome-wide efficient mixed-model association (GEMMA) [33]. In this thesis, we will implement these methods and compare PCR and LMM for adjusting a structure.

Unknown environmental factors such as pollution or lifestyle preferences can cause differentiated disease risks in one or several neighborhood areas. Originally they are more of a concern in quantitative trait loci (QTL) studies of inbred plants

and animals. It is noted that in expression QTL (eQTL) studies, the hidden environmental factors such as batch information and cell culture conditions, can cause serious power loss as well as spurious associations [34, 35, 36]. Several methods have been proposed to address these problems in the framework of either PCR or LMM [31, 36, 37]. Here, we suggest that this issue should be brought more attention to in genetic association studies. Ignoring or inability of fully modeling the environmental factor may flaw the association testing results, especially when the cases and controls are collected from different regions [38].

Another topic of great consequence in association studies and will be addressed in this thesis, is improving power to detect associated disease loci. This is considered to be the ultimate goal for association testing. As we know, a complex disease usually exhibits its characters in several aspects, so combining multiple traits can be more informative than using a single trait to detect the underlying genetic mechanism. Various methods have been embraced for multivariate trait analysis, such as constructing a composite trait to represent all traits, or taking multiple testing adjustment for controlling the Type I error. Linear mixed models (LMM), the generalized linear mixed models (GLMM) or the generalized estimation equations (GEE) are also readily applicable.

Chapter 3 to 5 aimed to address the first topic – adjusting for population structure. By using next-generation sequencing data from the 1000 Genomes Project and simulated disease status, in Chapter 3 we showed that when testing LFVs or RVs, population stratification of two continental groups – Europeans (EUR) and Africans (AFR), could be largely controlled by PCR with a few top PCs of 10000 randomly selected CVs or LFVs. In Chapter 4, as a further step of Chapter 3, we assigned different subgroups in EUR and AFR groups to be cases and controls, in order to create the most extreme population stratification of a fine scale. We compared the performance of the PCs based on different types and sets of variants and constructed by two methods, PCA and spectral dimension reduction (SDR), with respect to adjusting for population stratification. We found that PCs based on all variants or all CVs without pruning, constructed by SDR, generally had a good control in Type I error, especially for testing RVs. In Chapter 5 we related the PCR with LMM under the framework of probabilistic PCA [39]. We pointed out that the PCR was an approximation to LMM: such an approximation was dependent on the

number of the top PCs used. Further, although the LMM outperformed PCR in several situations, the PCR can be better in the presence of unknown environmental confounders. As a result, we proposed to use a hybrid model – which incorporated both the PCs and the random effect.

In Chapter 6 we addressed the second topic by extending the sum of powered score (SPU) tests [40] to the generalized estimation equations (GEE) framework for multivariate trait analysis. We also applied our methods to analyze the NvR HIV/AIDS study data and the Genetic Analysis Workshop (GAW) 18 data. Although in simulation studies the SPU tests showed obvious advantages over other existing tests, this was not the case in the real data analysis. We have offered a detailed discussion regarding the results in the text.

Chapter 2

Data

2.1 The 1000 Genomes Project Data

The 1000 Genomes Project was launched in January 2008. Scientists planned to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups using the newly developed technology called next-generation sequencing. It parallelizes the sequencing process and generates thousands or millions of sequences at once. In 2010 the project finished its pilot phase, which was described in detail in [41]. We downloaded a low-coverage whole genome sequencing dataset released in August 2010 on the 1000 Genomes Project website and used it in Chapter 3, 4 and 5. The dataset included 629 individuals: 174 Africans (AFR), 283 Europeans (EUR) and 194 Asians. We focused on the European and African samples (Table 2.1). Across chromosomes 1-22, after 19938 monomorphic single-nucleotide polymorphisms (SNPs) were excluded, there were 8952087 variants, including 6227535 CVs, 1849693 LFVs and 854921 RVs. As it was usually suggested to use almost independent SNPs to construct PCs by PCA, we pruned all variants by PLINK [42] using a sliding window of size 50, shifted by 5 with threshold $r^2 \leq 0.05$. The same was done for all CVs, all LFVs and all RVs respectively. After pruning within each type of variants, 149324 CVs, 328813 RVs, and 384751 LFVs remained. In the European data, there were 27 individuals identified as admixed Americans. This would allow us to explore how PCA performed when small outlying clusters were present.

Table 2.1: Subgroups in European (EUR) and African (AFR) data in the 1000 Genomes Project. The MXL, PUR and PUR2 samples are also labeled as admixed Americans (AMRs).

Populations	EUR						AFR			
Subpopulation	CEU	TSI	GBR	FIN	MXL	PUR	PUR2	YRI	LWK	ASW
	Utah	Italy	British	Finnish	Mexican in LA	Puerto Rico	Puerto Rico	Nigeria	Kenya	African in SW US
#Samples	90	92	43	36	17	5	5	78	67	24

2.2 The GAW18 Data

The Genetic Analysis Workshop (GAW) 18 data contains DNA sequencing data from 20 Mexican American families, each with between 21 and 76 members who are part of the San Antonio Family Diabetes and Gallbladder Studies. Thus, there are familial correlations. Sequencing data for the odd numbered chromosomes, gender,

age, smoking status and medication were collected for 855 samples. Systolic and diastolic blood pressure (SBP and DBP) values were measured at up to four time points spanning 30 years, and hypertension (HTN) was defined as the use of anti-hypertensive medication or SBP>140 and DBP>90 (Table 2.1). Identities of 157 unrelated samples were also provided.

Besides, the GAW18 data simulated 200 replicates of the 3 phenotypes at 3 time points. These replicates were modeled after the real data with SBP and DBP distributions and frequencies of hypertension, medication use, and tobacco smoking. The simulation model used the real pedigrees and the cleaned imputed sequence data for each individual and was constructed to maintain the heritabilities of SBP and DBP and the observed correlations between them.

There were 8,348,674 single nucleotide variants (SNVs) across 11 (odd numbered) chromosomes, among which 2,791,923 SNVs were CVs and 3,977,003 were RVs. After pruning by PLINK [42] using a sliding window of size 50, moving step of 5 and $r^2 < 5\%$ and filtering out those with missing call>0.05, there were 63,157 CVs and 1,104,098 RVs left.

Figure 2.1: Summary of the real data at each time point

	Exam 1	Exam 2	Exam 3	Exam 4
N*	855	605	622	233
Year	1981# - 1996	1997 - 2000	1998 - 2006	2009 - 2011
Age	39.6 (16 – 94)	42.9 (17 – 97)	46.3 (18 – 95)	50.9 (30 – 81)
SBP	122 (80 – 216)	125 (90 – 211)	125 (76 – 220)	128 (93 – 233)
DBP	71 (40 -123)	72 (43 – 115)	71 (32 – 108)	78 (46 – 126)
Meds (%)	9.79	18.97	28.75	43.29
Hypertension (%)	18.13	28.38	34.77	51.93
Smoking (%)	22.90	18.25	20.00	11.16

* Number with blood pressure measurements.

2.3 The NvR Study Data

The NvR study was conducted by the Community Program for Clinical Research on AIDS (CPCRA) and the Canadian HIV Trials Network (CTN), named NvR study [43]. The NvR study is an open-labeled randomized trial testing the equivalence of two drugs, nelfinavir (NFV) or ritonavir (RTV) along with other protease inhibitors

(PIs), nucleoside and non-nucleosides reverse transcriptase (RT) inhibitors (NRTIs/NNRTIs) in treating advanced HIV disease. With the trial starting in January 1997, patients had visits at 1 month, 4 months, and every 4 month thereafter until December 31, 2001. Those whose disease progressed to an AIDs defining event or who became intolerant to their assign treatment could be switched to the alternated PIs.

Albeit it is recommended to use highly active antiretroviral therapy (HAART) for HIV infected patients to inhibit the virus, there is a high chance of disease progression. The treatment may lose its effectiveness later due to the high mutating rate of the viral RNA. This is the so-called drug-resistance. Here we are interested in investigating the association between drug-resistance mutations and disease progression.

One critical point is that the viral sequencing technique used in the NvR study was only capable of genotyping subjects with viral RNA load>2000 copies/ml. Thus, among the 610 patients who agreed to be genotyped, 549 of them were successfully genotyped. And in the follow-up visits, a large number of subjects had no genotypic drug resistance (Table 2.2) because their viral load was either below 2000 copies/ml, they had died, or they did not attend the exam. To utilize the maximum sample size, we adopted the strategy to use all the subjects alive at each time point and regarded those patients without genotypic data as having no new mutations since last visit. Once one viral mutation appeared in one subject, it was carried over afterwards. The viral mutation was coded as 1 for presence and 0 for non-presence.

We studied 3 traits related to disease progression: the binary response indicating whether the viral load at visit t increased more than one log-10 unit from that at $t-1$, denoted as V; whether the CD4+ count at period t was below 200 copies/ml, denoted as $CD4$; whether there were any new AIDS-defining illnesses at visit t, denoted as SYM. We did both cross-sectional multivariate trait analysis as well as longitudinal univariate trait analysis for t =4, 8, 12, 16, 20, 24 months. Common mutations (CVs) were defined as those with a relative frequency $\geq$0.1 among all samples genotyped at time t and rare mutations (RVs) were defined as those with a relative frequency <0.1.

We tested on the mutations at $t-1$. The covariates that we used include age, gender, the PI assigned at the start of the trial (basePI), an indicator whether the

subject is naive to AR drugs (NAIVEBL), an indicator whether AIDs was present at baseline (PODBL) and CD4 counts at baseline (CD4BL). These covariates were allowed to have different coefficients in multivariate trait analysis. We also included an indicator for whether the RNA viral load was above 2000 copies/ml at $t-1$, named "IND_RNA(t-1)" and an indicator for whether the subject was genotyped at visit $t-1$, "geno(t-1)".

Table 2.2: Sample size at different visits.

sample size	baseline	01	04	08	12	16	20	24	28	32	36	40
# in trial	610	546	531	519	507	497	483	470	460	446	432	387
# with RNA>=2000	549	299	305	348	346	334	334	317	317	325	320	320
# genotypic resistance	546	213	206	194	160	129	101	84				

Chapter 3

Adjustment for Population Stratification via Principal Components in Association Analysis of Rare Variants

3.1 Introduction

Population stratification occurs when population structure acts as a confounder in association testing. Investigation in population structures is of primary interest in genetic studies to infer human evolution history and to avoid spurious findings in association testing. Intuitively population stratification can arise in association studies of low-frequency variants (LFVs) with minor allele frequency (MAF) between 0.01 and 0.05 and rare variants (RVs, with MAF$<$ 0.01). Some existing techniques for common variants (CVs with MAF$\geq$ 0.05), e.g. principal component analysis (PCA), might be applicable to LFVs and RVs [20]. However, two recent studies [6, 7] achieved different conclusions on the relative effectiveness of CV- or RV-based principal components (PCs) in uncovering population structures. More importantly, to our knowledge, the issue has not been experimentally demonstrated in the context of association tests. Among the many existing techniques for CVs, Wu et al. (2011) [25] demonstrated that adding a few top PCs as covariates in a regression analysis, the so-called principal component regression (PCR), is a simple and effective approach to adjusting for population stratification for unrelated samples. Hence we adopt this approach throughout. Furthermore, with the availability of sequence data, as pointed out by [44], it is not completely clear whether LFVs or RVs can be used to infer genetic ancestry. If so, importantly, it is natural to ask whether using LFVs or RVs (or both LFVs/RVs and CVs) can perform better than using CVs alone in adjusting for population stratification.

We used the 1000 Genomes Project data for two continental groups, 174 African (AFR) and 283 European (EUR) samples. Our result was in agreement with [7], that the top PC based on 10000 LFVs could better separate the two groups than that based on CVs. However, the PCs based on either CVs, LFVs or RVs could not separate the underlying subgroups. More interestingly and perhaps surprisingly, using PCs based on CVs maintained power while using PCs based on some randomly selected LFVs might suffer from substantial power loss in the absence of population stratification, which was likely due to the high linkage disequilibrium (LD) among the randomly selected LFVs.

3.2 Methods

3.2.1 Data

We used the sequencing data from the 1000 Genomes Project. On chromosome 1, among the 694231 common SNPs in both groups, there were 478208 CVs, 146353 LFVs and 69670 RVs. For the purpose of simulation, we selected a few regions of multiple LFVs or RVs associated with the continental group. As pointed out by [44], since spurious associations often arise at differentiated variants whose MAFs are unusually different between different ancestral groups, it is crucial to consider these SNPs when correcting for population stratification. We used sliding windows with various sizes on chromosome 1 and tested the association between the continental group and the LFVs or RVs inside each window using a few statistical tests (discussed below). We identified 3 regions, termed R1 to R3, as representatives for unusually differentiated LFVs or RVs with various characteristics.

For each region, based on a statistical model and the selected LFVs or RVs from the sequence data, we generated simulated datasets with a simulated disease status for each subject. We then tested possible association between the generated disease status and the observed SNPs in each region, based on which we assessed the performance of each test in terms of its statistical power and Type I error. Of particular interest was to investigate how the performance of a test depended on whether and how to use PCs constructed from the genome-wide sequence data.

3.2.2 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a classical method to find the pattern of the high dimensional data and find the major axes of variation. More specifically, it converts the high throughput genotype data by orthogonal transformation to a set of linearly uncorrelated variables called principal components (PCs). It was evidenced that the top few PCs should capture the variation from the ethnic groups. Using PCs was proposed to be an effective way to control population stratification [24].

Suppose $\tilde{X}$ is an n by p matrix, with n subjects and p SNPs, and $\tilde{X}_{im}$ denotes the genotype of the m^{th} SNP for i^{th} subject. We have $p >> n$. $\tilde{X}_{im}$ is coded 0, 1,

2 for the minor allele count. Before we apply any method to construct PCs, each SNP is standardized as $\frac{\tilde{X}_{.m}-2p_m}{\sqrt{p_m(1-p_m)}}$, where $\tilde{X}_{.m}$ is the m^{th} column of $\tilde{X}$, a vector of genotypes of SNP m for all subjects. The MAF, p_m, are obtained by PLINK. No outliers are removed. We denote the the SNP matrix after standardization as X. PCA preserves the pairwise distance by using a kernel proportional to the n by n sample covariance matrix of X,

$$A = XX^T.$$

The eigenvalues of A are denoted as δ_q and eigenvectors as w_q, for $q = 1, ..., n-1$. Eigenvalues should be sorted decreasingly and eigenvectors rearranged accordingly, denoted as δ_q^* and w_q^*. Then the l^{th} PC score as the l^{th} dimension to represent n subjects is:

$$PC_l = Xw_l^*.$$

Geometrically, this corresponds to the rescaled eigenvectors of A, which is $w_l^*\sqrt{\delta_l^*}$ [45]. Yet, through out this chapter, we will use the PC score to adjust for population stratification.

3.2.3 Statistical Tests

We applied two sets of representative statistical tests for association analysis of LFVs or RVs. The first set includes the score test, the sum of squared score (SSU) test, the weighted sum of squared score (SSUw) test, the Sum test and the univariate minimum p-value (UminP) test ([16]), while the second includes the T1, T5, Fp, VT and EREC tests [20]. We will first introduce the first set of the five tests. They were chosen based on the following reasons. The score test is a classical test in general statistical applications, asymptotically equivalent to the Wald test and likelihood ratio test. The UminP test is perhaps the most popular in association analysis of CVs, as used in GWASs. The Sum test is a representative of the so-called pooled association tests [46], similar to the well-known CAST [47] and CMC test [48]. The SSU test is closely related to GDBR [14], KMR [12, 25] and C-alpha test [15]. In an extensive simulation study, [8] found that the SSU test performed similarly to the KMR and C-alpha, and was an overall winner with the highest or close to the highest power in association analysis of RVs. With either CVs or RVs, the SSUw

test often performed similarly to the SSU test; however, with both RVs and CVs, the SSUw test might perform better [8]. All the five tests are based on the score vector of a regression model, e.g. a generalized linear model (GLM), hence only a reduced model under the null hypothesis H_0 needs to be fitted, leading to their being computationally faster and numerically more stable than those based on fitting a full model, e.g. the Wald or likelihood ratio test. Importantly, since the tests are formulated in the general regression framework, it is easy to incorporate covariates or extend them to other more complex studies, e.g. with censored event times as traits, correlated family data or multiple traits.

For a binary trait, Y_i, for subject i with k SNPs, $X_i = (X_{i1}, ..., X_{ik})'$, and covariates $Z_i = (Z_{i1}, ...Z_{iJ})'$, all five tests are based on the null model

$$\text{Logit}Pr(Y_i = 1) = \beta_0 + \sum_{j=1}^{J} Z_{ij}\gamma_j,$$

which is simpler than the full model:

$$\text{Logit}Pr(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j + \sum_{j=1}^{J} Z_{ij}\gamma_j.$$

We use the additive coding for each SNP; that is, $X_{ij} = 0$, 1 or 2 is the the number of minor alleles in SNP j subject i. Due to the extremely low MAF, it is unlikely to have two copies of the minor allele for a given RV, and thus there is little difference between various coding schemes for RVs. The PCR, which is the method used through out this thesis to adjust for population stratification, is simiply adding the top few PCs to the covariates.

All five tests are global tests with the null hypothesis H_0: $\beta = (\beta_1, ..., \beta_k)' = 0$ are global in the sense of not identifying specific zero subcomponents of β. Given a score vector $U = (U_{.1}, ..., U_{.k})'$ and its covariance estimate $V = Cov(U)$, the five

test statistics are respectively:

$$
\begin{aligned}
Score &= U'V^{-1}U, \\
SSU &= U'U = \sum_{j=1}^{k} U_{\cdot j}^2, \\
SSUw &= U'diag(V)^{-1}U = \sum_{j=1}^{k} U_{\cdot j}^2/V_{jj}, \\
Sum &= 1'U = \sum_{j=1}^{k} U_{\cdot j}, \\
UminP &= \max_{j=1}^{k} U_{\cdot j}^2/V_{jj},
\end{aligned}
$$

where $diag(V)$ is a diagonal matrix with diagonal elements (V_{jj}'s) of V.

Under H_0, based on the asymptotic Normality of U, $U \sim N(0, V)$, the asymptotic distribution of the first four tests can be easily derived and used to obtain their p-values ([16]), while numerical integrations with a multivariate Normal density can be used for the UminP test ([49]). For relatively small sample sizes, especially with RVs, the above asymptotics may not be applicable. Alternatively, as suggested by other authors ([20, 25]), we can apply the parametric bootstrap ([50]) in the following steps: 1) fit the null model; 2) use the fitted null model to generate $Y_{i,b}^*$'s as the bth bootstrap dataset with $b = 1, ..., B$; 3) calculate a test statistic T with the original data $(Y_i, X_i)'$s, and T_b with the bth bootstrap data $(Y_{i,b}^*, X_i)'$s; 4) the p-value is $\sum_{b=1}^{B} I(|T| > |T_b|)/B$. We used $B = 200$ throughout (and using $B = 1000$ gave similar results in all the simulations), though in practice we might need to use a much larger B to achieve a higher level of statistical significance. As to be shown, for the score test and to a lesser degree for the UminP test, the asymptotics might give inflated Type I error rates, while the bootstrap gave much better results; in contrast, the SSU and SSUw tests are more robust to small samples with Type I error rates always close to the nominal level in all our experiments.

For comparison, we also included the T1, T5, Fp, VT and EREC tests ([20]), all implemented in software SCORE-Seq available at

`http://www.bios.unc.edu/~dlin/software/SCORE-Seq/`

. As shown by [20], a general class of the score-based association tests can be

formulated as

$$T_G = \sum_{j=1}^{k} \zeta_j U_j,$$

where ζ_j is a weight for SNP j. Different choices of the weights lead to a variety of tests: 1) the T1 (or T5) test corresponds to $\zeta_j = 1$ if the MAF of SNP j is less than 1% (or 5%) and $\zeta_j = 0$ otherwise; 2) in the Fp test, we have $\zeta_j = 1/\sqrt{\tilde{p_j}(1-\tilde{p_j})}$, where $\tilde{p_j}$ is an estimate of the MAF of SNP j with pseudo counts from the pooled sample, giving higher weights to rarer SNPs [17]; 3) the VT test combines multiple tests based on multiple thresholds, and for each threshold, $\zeta_j = 1$ if the MAF of SNP j is less than the threshold and $\zeta_j = 0$ otherwise [18]; it is a form of the adaptive Neyman's test [51]; 4) the EREC test uses $\zeta_j = \tilde{\beta}_j \pm c$ with $\tilde{\beta}_j$ as the (univariate) maximum likelihood estimate of β_j and $c = 1$ for binary traits. Although an asymptotic null distribution is available for each of the first three tests, it is not available for the EREC test. Furthermore, the asymptotic approximations might result in inflated Type I error rates for RVs. Hence we only show the results of the second set of the tests with their p-values calculated by the parametric bootstrap with the minimum allowable $B = 10^6$ resamples.

We also note that the score, SSU, SSUw and Sum tests are also special cases of the general T_G test. In particular, the SSUw test uses the weight $\zeta_j = U_j/V_{jj} \approx \tilde{\beta}_j$ [16], suggesting its close connection to the EREC test, whose weight ζ_j can be regarded as shrinking $\tilde{\beta}_j$ towards constants c or $-c$.

3.3 Results

3.3.1 Data description

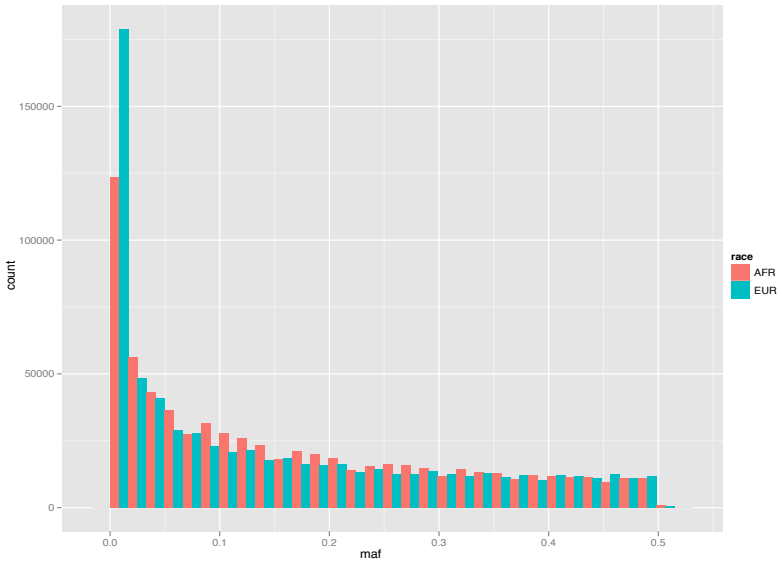
As shown in Figure 3.1, there are clear differences between the MAF distributions of the two continental groups. In particular, the difference seems to be larger for low MAFs than for high MAFs.

We selected 3 regions, named R1 to R3: the first two contained 19 and 40 consecutive LFVs (and only LFVs) respectively, while the third one consisted of 40 consecutive RVs (and only RVs); we calculated the MAF of any SNP based on

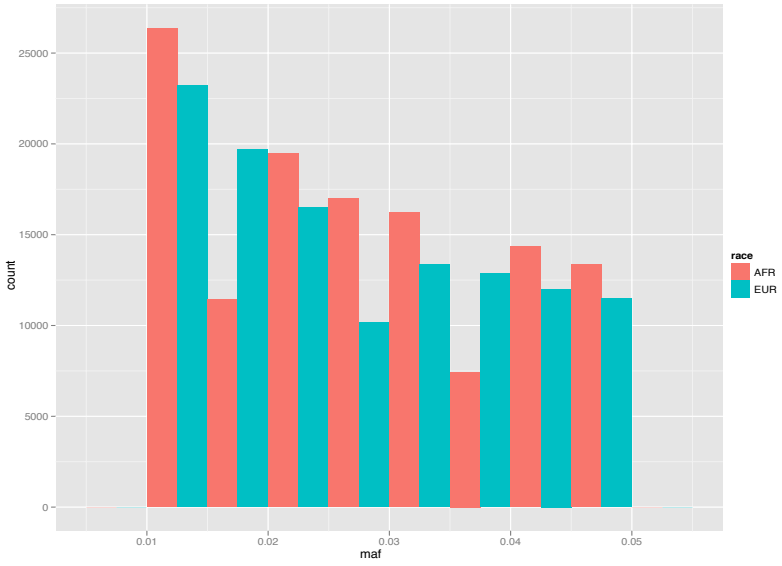
the pooled sample. The LFVs or RVs within each region were associated with the continental group; that is, the MAFs of the SNPs were different between the AFR and EUR groups (Figure 3.2). These 3 regions also showed different LD patterns (Figure 3.3): LD was weak in R1, moderate in R3, and strong in R2.

We randomly selected a large number of CVs (or LFVs) from chromosome 1 to construct PCs. As shown in Figure 3.3, the top PC based on CVs could largely separate the two AFR and EUR groups; however, perhaps surprisingly, the top PC based on LFVs did better in completely separating the two groups. When using some randomly selected SNPs, including CVs, LFVs and RVs, the results were between those based on either CVs or LFVs alone (not shown). We will present and discuss results based on RVs later. Since the results with 100000 CVs (or LFVs) (not shown) were similar, in the following, we used a few top PCs based on either 10000 CVs or 10000 LFVs.

Figure 3.1: Distributions of MAFs for the EUR and AFR groups.



(a) All SNPs



(b) Only LFVs

Figure 3.2: Comparison of the MAFs between the EUR and AFR groups for the SNPs in regions R1–R3.

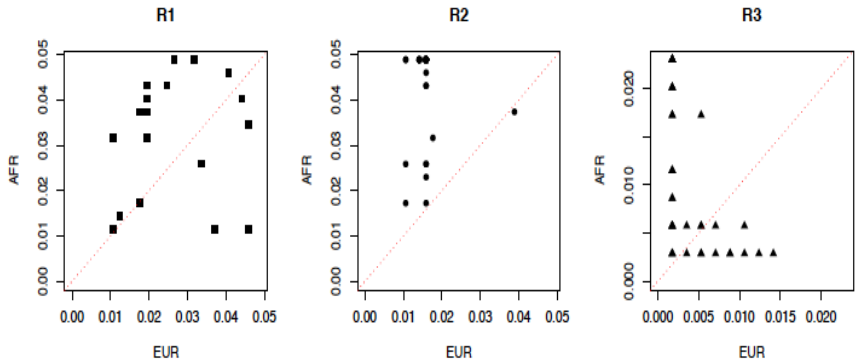


Figure 3.3: LD plots in r^2 for the EUR (top row) and AFR (bottom row) groups in regions R1–R3.

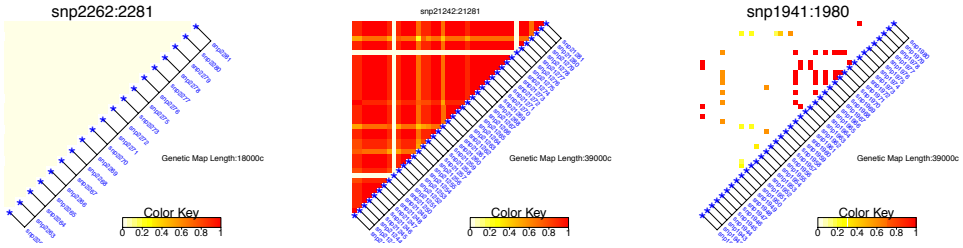
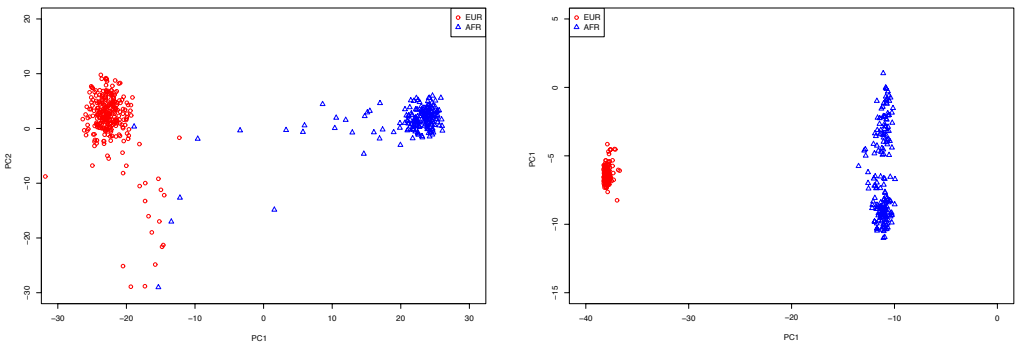


Figure 3.4: The top two PCs constructed with CVs or LFVs.



(a) Based on 10000 CVs

(b) Based on 10000 LFVs

3.3.2 Association testing with LFVs: Type I Error

We first generated simulated data under H_0 with population stratification. Specifically, we randomly selected 90% of the EUR samples and 10% of AFR samples as cases (i.e. $Y_i = 1$), while treating the remaining ones as controls (i.e. $Y_i = 0$). In this way, none of the SNPs caused the "disease", and there was a clear association between the continental group and the "disease" (i.e. population stratification). We applied the five tests to the two LFV regions with 1000 simulated datasets for each case. Since the results are similar for PCs based on either CVs or LFVs, we only show that for the former. Table 3.1 lists the Type I error rates at the nominal level $\alpha = 0.05$. It is clear that, without adjustment for population stratification, all the tests could have dramatically inflated Type I error rates (except the Sum test for R1), suggesting the necessity of adjusting for population stratification. With PCs, including with even just the single top PC (i.e. #PCs=1), the problem with inflated Type I error rates largely disappeared; there was almost no difference between using various numbers of PCs, as long as at least one PC was used. It is noted that the asymptotics-based score test could have severely inflated Type I error rates, even in the presence of PCs for region R2, and that the asymptotics-based UminP test could also have slightly inflated Type I error rates. The bootstrap-based tests all

had their Type I error rates better controlled.

3.3.3 Association testing with LFVs: Power

We generated a disease status from the following logistic regression model:

$$\text{Logit}Pr(Y_i = 1) = \beta_0 + \sum_{j=1}^{k_1} X_{ij}\beta_j,$$

where X_{ij} was the jth SNP of the ith subject (AFR or EUR), $\beta_0 = -\log 3$ was chosen to generate a background disease incidence of 25% (when all $X_{ij} = 0$), and the causal effect sizes β_j were randomly generated from a uniform distribution $U(-a, a)$ or $U(0, a)$ for a constant $a > 0$. With $U(-a, a)$, some causal effects were deleterious while others were protective against disease; with $U(0, a)$, all causal effects were in the same direction of being deleterious. We used $k_1 \leq k$: if $k_1 < k$, we randomly selected a subset of the SNPs to be causal while others were neutral or non-causal, but a test was always applied to all the k SNPs; it is important to assess a test's robustness to the number of non-causal SNPs, since in practice we expect causal SNPs to be mixed with some neighboring non-causal ones. Each subject i's genotype was input to the above model to generate his/her disease status. In such a way, we generated a dataset of 457 subjects with various numbers of cases and controls.

Since the general conclusions remained the same, we chose only a small subset of results to present in Tables 3.2 and 3.3. Recall that there was weak and strong LD, and a small and a large number of LFVs, in the two regions R1 and R2 respectively. The PCs were constructed based on the 10,000 randomly selected CVs. First, since all the SNPs had MAF between 1% and 5%, the T1 test was not applicable (with all weights $\zeta_j = 0$), and the T5 test (with all weights $\zeta_j = 1$) was essentially the same as the Sum test. In addition, since the causal SNPs were selected randomly and were not correlated with lower or higher MAFs, the Fp and VT tests were not expected to improve over the T5 and Sum tests. Second, since there was no population stratification, it is good to see that using or not using PCs, or using different numbers of PCs, gave similar results for all the tests. We emphasize that

this is a desired property. In practice, for a given dataset, population stratification may or may not be present; to be safe and avoiding spurious associations, we might still want to apply an adjustment, e.g. based on PCs. Hence, it would be desirable to have no or minimum power loss when adjusting for population stratification. Third, the identity of the most powerful tests varied with the set-up. For example, in region R1, 1) with all 19 causal SNPs being deleterious, the Sum, T5, Fp and EREC tests performed similarly and were most powerful; 2) with the 19 causal SNPs with opposite association directions, as expected, the Sum, T5 and Fp tests were low powered, while the SSU, SSUw and score tests were most powerful. Although there was no uniformly most powerful test, the SSU and SSUw tests seemed to be the overall winners. Fourth, due to the small sample size, the asymptotics-based score test might lose power as compared to the bootstrap-based score test (not shown). In contrast, other tests seemed to be more robust to small samples: their asymptotics-based version and bootstrap-based version always gave similar results (not shown).

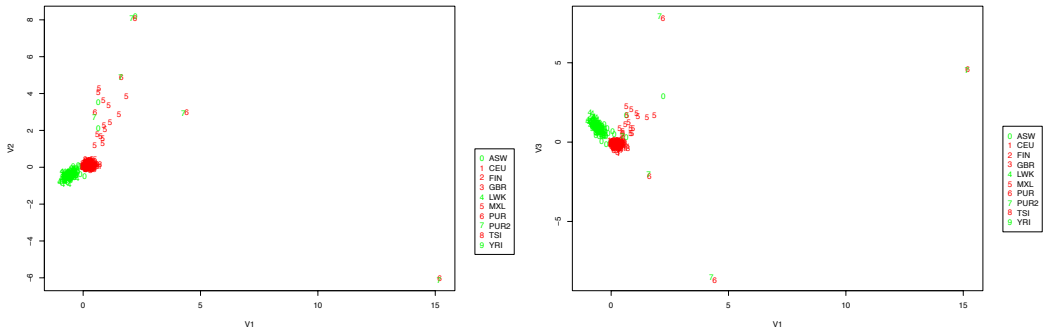
Since the PCs based on LFVs could better separate the AFR and EUR groups, it would be interesting to see the performance of the tests with PCs constructed from LFVs. In many situations, a test with LFV-based PCs and with CV-based PCs performed similarly; however, as shown in Tables 3.3 and 3.4, when all the causal effects were in the same direction, it is clear that adjusting with more than one PC led to power loss, which was often substantial. For example, in the set-up with 30 causal SNPs with positive effects in region R2 (Table 3.4), the SSU and SSUw tests were most powerful; however, with 1, 5 and 10 PCs, the power of the SSU test monotonically decreased from 0.871 to 0.865, 0.803 and 0.781, respectively. This is a case called over-adjustment in the sense of losing substantial power when adjusting for population stratification (or more generally, confounders). This phenomenon was not specific to the first set of the five tests shown in Table 3.4; it also appeared for the second set (Table 3.3): for example, for region R1 with $k_1 = 8$ causal SNPs with the same effect direction, the power of the T5, Fp, VT and EREC tests reduced respectively from 0.872, 0.872, 0.838 and 0.916 with no PC to 0.812, 0.805, 0.755 and 0.850 with 10 PCs constructed from 10000 LFVs. It is interesting to note that, in all our examples, using only the top PC could largely control the Type I error rate while maintaining the power (with no or negligible power loss).

We explored the reason for the over-adjustment. We first hypothesized that the LFV-based PCs might reflect some hidden ethnic structure. When adjusting for ancestry using either of the two reported continental groups or reported ethnic subgroups, the test results were similar to those with no or only 1 PC; in other words, there was no loss of power. Second, we regressed the binary trait on the top 10 PCs and referred to the corresponding linear combination of the top 10 PCs as a PCs-defined group score. We found that in the cases with over-adjustment, the PCs-defined group score was much more significantly associated with the sum of the LFVs to be tested than those in other cases. Although the LFVs were randomly selected to construct PCs, we found that surprisingly many of them were highly correlated, as shown by an exome sequence dataset from the 1000 Genomes Project [52]. For CVs, it is highly recommended to use only nearly independent SNPs to construct PCs [24, 28] to avoid the resulting PCs' representing some peculiar features of the data. Hence, we first tried to remove highly-correlated LFVs by using PLINK with a threshold of $r^2 \leq 0.5$ or $r^2 \leq 0.05$ respectively. Then using the top 10 PCs constructed with the remaining LFVs, we obtained the results (not shown) similar to those without adjustment (or to using CV-based PCs). In conclusion, the multiple PCs based on the original possibly highly correlated LFVs perhaps represented some unknown and possibly artificial structure in the data.

3.3.4 Subgroup analysis with RVs

We used a subset of 786487 non-monomorphic RVs from chromosomes 2 to 22 to construct PCs. After pruning, we had 305036 RVs. We then selected a random set of 10000 RVs to construct PCs. As shown in Figure 3.4, the first PC could largely separate the two continental groups, but not the 10 subgroups. Several PUR and PUR2 samples appeared to be outliers, which might have unduly influenced the PCA results; on the other hand, it may be argued that the RVs could better separate the PUR and PUR2 samples from other subgroups.

Figure 3.5: The top three PCs constructed with 10000 RVs. The red/dark points are the EUR samples and the green/grey points are the AFR samples.



(a) PC1 vs PC2

(b) PC1 vs PC3

To check the number of PCs based on RVs relating to population structure, we applied the Tracy-Widom (TW) proposed by [24] (implemented in R package `EigenCorr`, Lee et al 2011). As we know the largest eigenvalue of a Wishart matrix follows TW distribution after standardization, [24] proposed to use TW distribution to test if the PCs reflect the real structure in the data. In our data, the TW test yields a statistically significant p-value less than 0.05 for each of the top 19 eigenvalues (Table 3.5). We also applied one-way ANOVA to test the significance of each PC with varying mean values across the two continental groups or the 10 subgroups; in both cases, the most significant PCs were in the top 13.

A visual examination of the scatter plots of the top PCs did not reveal that the top PCs could separate the 10 subgroups. Hence, we applied finite Gaussian mixture model-based clustering (implemented in R package `mclust`) to top the 10, 20 and 50 PCs. Based on the Rand index (calculated using R package `clue`), using the top 20 PCs led to the highest agreement between the resulting 12 clusters and 10 true subgroups (with a Rand index value of 0.812 and an adjusted Rand index value of 0.408). As shown in Table 3.6, in agreement with Figure 3.4, although the two continental groups could be largely but not perfectly separated, the subgroups in the EUR group could not be distinguished: most of them were mixed into two clusters.

3.3.5 Association testing with RVs

We also conducted a simulation study with the 40 RVs in region R3. To assess the Type I error rates, we generated a binary trait as before under population stratification; for power, we randomly selected $k_1 = 15$ RVs as causal ones with their effect sizes $\beta_j \sim U(0, \log(3))$. We used the parametric bootstrap for p-value calculation for each test. Note that since all the 40 SNPs had MAF less than 1%, the results for the T1 and T5 tests were exactly the same. As shown in Table 3.7, under H_0, if no adjustment was made, all the tests resulted in dramatically inflated Type I errors. Since the first PC could not completely separate the two continental groups (Figure 3.4), using only the top PC still yielded largely inflated Type I error rates. In contrast, using the top 10 or 20 PCs could largely remedy the problem, though there were still some slightly inflated Type I error rates for some tests, which could be due to the fact that even the top PCs could not completely separate the two continental groups (Figure 3.4 and Table 3.6). For power, with the exception of the score, SSU and SSUw tests, all other tests seemed to have some power loss with 20 PCs.

3.4 Discussion

We have used a low-coverage whole-genome sequencing dataset generated by the 1000 Genomes Project to empirically investigate some characteristics of LFVs or RVs that are relevant to the association analysis. For example, some might argue that, due to the low MAFs, LFVs and RVs are expected to be independent; we have demonstrated that the neighboring LFVs or RVs in a region may be in either low, moderate or high LD, suggesting that future studies on the performance of any association test should consider varying LD as a factor. Furthermore, as a useful complement to the extensive simulation studies of [8], we have used real sequence data to demonstrate the power properties of the various tests with or without PCs, though it was not the main aim of the current study. In particular, it is confirmed that the Sum test, a representative of simple pooled association tests [53], is not powerful in the presence of different association directions or of many non-causal SNPs; in contrast, the SSU and SSUw tests are much more powerful in

these situations. It is also shown that the asymptotics of the Sum, SSU and SSUw tests seemed to work well with a reasonable sample size for LFVs, much more robust than the score test. Of course, with small samples sizes or RVs with extremely low MAFs, one has to be cautious in using asymptotics. As shown here and in other places [20, 25], the parametric bootstrap is a useful alternative. Given the generally good performance of the SSU and SSUw tests, we would recommend their use in practice; if the applicability of the asymptotics is of concern, a two-step procedure can be taken: one could first use the asymptotics-based SSU or SSUw test to quickly scan the genome, then apply the more computing-intensive bootstrap-based SSU or SSUw test to the more significant regions identified in the first step.

Perhaps the most interesting finding of this study is that, in accordance with [7] but differing from Baye et al (2011), PCs constructed with LFVs could potentially separate different continental or ethnic groups better than those with CVs, though either can be used to adjust for population stratification effectively. We note that Siu et al (2012) used a similar whole genome sequence dataset as ours while Baye et al (2011) used a smaller subset of the exome sequence dataset with much fewer LFVs or RVs. In addition, differing from Mathieson et al (2012), we focused on two relatively well-separated populations, i.e. AFR and EUR samples; further studies are warranted for other more challenging cases. In all our numerical examples, in contrast to that using PCs based on CVs led to no or little power loss in the absence of population stratification, surprisingly using multiple PCs based on LFVs might result in **over-adjustment** in the sense of substantial power loss. It is also interesting to note that, in all our examples, using only the top PC based on LFVs could largely control the Type I error rate while maintaining the power (with no or minimum power loss).

The over-adjustment with multiple PCs based on LFVs in our experiments was likely due to the use of many LFVs in high LD; once we used LFVs not in high LD, the problem largely disappeared. This is in agreement with two known results: first, it is highly recommended to use only almost independent CVs to construct PCs (Patterson et al 2006; Lee et al 2011); second, for unknown reasons, there seems to exist long-range correlations among LFVs or RVs in real sequence data (Tintle et al 2011). Hence, one has to be careful in selecting LFVs or RVs to construct PCs; in particular, a random subset of far-away LFVs or RVs may not

be sufficient. Furthermore, our preliminary analysis also shows that PCA of RVs with MAFs $< 1\%$ might not be effective in separating subpopulations. One possible reason is the sensitivity of PCA to outliers, which are present with some diverse subpopulations and largely varying numbers of subpopulation samples; it would be interesting to apply other more robust methods (e.g. Lee et al 2011). We also emphasize that our conclusions are based on the use of a low-coverage whole-genome sequencing dataset, which may be different from high-coverage sequencing data; for example, high-coverage sequencing tends to uncover more RVs [54]. Importantly, we only considered using CVs, LFVs or RVs, but not their combined use; it remains to be investigated how to select and combine CVs, LFVs and RVs to best capture population structures.

Table 3.1: Type I error rates with population stratification. The PCs were constructed using 10000 CVs.

		Asymptotics				Bootstrap			
Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
R1	Score	0.693	0.055	0.052	0.048	0.716	0.044	0.044	0.042
	SSU	0.618	0.062	0.039	0.035	0.647	0.061	0.039	0.044
	SSUw	0.620	0.052	0.036	0.035	0.642	0.053	0.038	0.037
	Sum	0.067	0.066	0.050	0.047	0.070	0.044	0.048	0.048
	UminP	0.201	0.084	0.065	0.067	0.232	0.068	0.039	0.047
R2	Score	0.155	0.180	0.192	0.171	0.624	0.062	0.071	0.061
	SSU	0.709	0.053	0.055	0.054	0.684	0.047	0.051	0.055
	SSUw	0.700	0.052	0.055	0.054	0.669	0.049	0.052	0.055
	Sum	0.684	0.054	0.055	0.055	0.652	0.049	0.059	0.056
	UminP	0.677	0.052	0.061	0.066	0.685	0.044	0.050	0.047

Table 3.2: Empirical power of various tests based on the parametric bootstrap for the two regions with k_1 causal SNPs. The PCs were constructed using 10000 CVs.

Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 3)$			
	Score	0.489	0.489	0.489	0.496	0.838	0.836	0.824	0.826
	SSU	0.500	0.492	0.497	0.501	0.882	0.880	0.891	0.881
R1	SSUw	0.507	0.480	0.481	0.486	0.883	0.883	0.889	0.886
$k_1 = 8$	Sum	0.240	0.230	0.234	0.230	0.860	0.860	0.856	0.852
	UminP	0.401	0.397	0.393	0.383	0.813	0.820	0.813	0.802
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.5)$			
	Score	0.483	0.467	0.479	0.475	0.504	0.504	0.493	0.479
	SSU	0.507	0.493	0.511	0.492	0.773	0.771	0.758	0.738
R1	SSUw	0.479	0.479	0.483	0.477	0.769	0.764	0.765	0.756
$k_1 = 19$	Sum	0.207	0.202	0.201	0.204	0.842	0.839	0.839	0.832
	UminP	0.288	0.286	0.280	0.287	0.558	0.544	0.546	0.541
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 2)$			
	Score	0.256	0.240	0.251	0.244	0.707	0.702	0.699	0.687
	SSU	0.401	0.406	0.400	0.406	0.786	0.787	0.785	0.793
R2	SSUw	0.404	0.404	0.403	0.408	0.783	0.787	0.787	0.795
$k_1 = 4$	Sum	0.405	0.406	0.406	0.409	0.784	0.785	0.789	0.793
	UminP	0.360	0.344	0.347	0.349	0.761	0.763	0.758	0.756
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.1)$			
	Score	0.364	0.365	0.362	0.366	0.776	0.761	0.765	0.761
	SSU	0.601	0.602	0.607	0.606	0.871	0.868	0.871	0.869
R2	SSUw	0.601	0.603	0.607	0.610	0.873	0.867	0.872	0.869
$k_1 = 30$	Sum	0.599	0.601	0.602	0.605	0.874	0.866	0.869	0.869
	UminP	0.563	0.550	0.546	0.544	0.848	0.834	0.836	0.835

Table 3.3: Empirical power of various tests based on the parametric bootstrap for region R1. The PCs were constructed using either 10000 CVs or 10000 LFVs.

		10000 CVs				10000 LFVs			
Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(-\log 3, \log 3)$			
	T5	0.200	0.207	0.195	0.188	0.200	0.205	0.193	0.182
R1	Fp	0.203	0.197	0.186	0.192	0.203	0.195	0.184	0.179
$k_1 = 8$	VT	0.214	0.212	0.208	0.199	0.214	0.211	0.202	0.195
	EREC	0.401	0.399	0.381	0.375	0.401	0.401	0.373	0.368
		$\beta_i \sim U(0, \log 3)$				$\beta_i \sim U(0, \log 3)$			
	T5	0.872	0.878	0.869	0.857	0.872	0.876	0.866	0.812
R1	Fp	0.872	0.873	0.864	0.852	0.872	0.872	0.856	0.805
$k_1 = 8$	VT	0.838	0.832	0.826	0.813	0.838	0.820	0.784	0.755
	EREC	0.916	0.914	0.903	0.891	0.916	0.912	0.873	0.850
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(-\log 2, \log 2)$			
	T5	0.222	0.225	0.223	0.216	0.222	0.232	0.215	0.193
R1	Fp	0.220	0.223	0.214	0.209	0.220	0.219	0.208	0.198
$k_1 = 19$	VT	0.230	0.235	0.228	0.229	0.230	0.225	0.233	0.200
	EREC	0.395	0.396	0.398	0.382	0.395	0.390	0.392	0.375
		$\beta_i \sim U(0, \log 1.5)$				$\beta_i \sim U(0, \log 1.5)$			
	T5	0.844	0.846	0.845	0.828	0.844	0.843	0.823	0.768
R1	Fp	0.849	0.852	0.837	0.827	0.849	0.852	0.806	0.762
$k_1 = 19$	VT	0.779	0.779	0.752	0.747	0.779	0.760	0.729	0.685
	EREC	0.826	0.820	0.807	0.796	0.826	0.817	0.782	0.698

Table 3.4: Empirical power of various tests based on the parametric bootstrap for the two regions with k_1 causal SNPs. The PCs were constructed using 10000 LFVs.

Loc	Test	#PCs=0	1	5	10	#PCs=0	1	5	10
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 3)$			
	Score	0.489	0.498	0.499	0.486	0.838	0.836	0.805	0.787
	SSU	0.500	0.502	0.495	0.506	0.882	0.881	0.880	0.851
R1	SSUw	0.507	0.490	0.489	0.493	0.883	0.885	0.882	0.857
$k_1 = 8$	Sum	0.240	0.223	0.226	0.219	0.860	0.857	0.850	0.821
	UminP	0.401	0.402	0.386	0.396	0.813	0.822	0.792	0.753
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.5)$			
	Score	0.483	0.474	0.481	0.490	0.504	0.497	0.465	0.452
	SSU	0.507	0.501	0.507	0.503	0.773	0.765	0.724	0.663
R1	SSUw	0.479	0.476	0.484	0.488	0.769	0.763	0.727	0.681
$k_1 = 19$	Sum	0.207	0.202	0.194	0.180	0.842	0.835	0.819	0.791
	UminP	0.288	0.286	0.273	0.283	0.558	0.547	0.498	0.441
		$\beta_i \sim U(-\log 3, \log 3)$				$\beta_i \sim U(0, \log 2)$			
	Score	0.256	0.242	0.233	0.229	0.707	0.703	0.628	0.616
	SSU	0.401	0.410	0.386	0.382	0.786	0.790	0.734	0.713
R2	SSUw	0.404	0.408	0.384	0.380	0.783	0.791	0.737	0.713
$k_1 = 4$	Sum	0.405	0.407	0.384	0.380	0.784	0.786	0.737	0.712
	UminP	0.360	0.341	0.330	0.317	0.761	0.762	0.703	0.672
		$\beta_i \sim U(-\log 2, \log 2)$				$\beta_i \sim U(0, \log 1.1)$			
	Score	0.364	0.361	0.347	0.337	0.776	0.761	0.657	0.634
	SSU	0.601	0.604	0.585	0.582	0.871	0.865	0.803	0.781
R2	SSUw	0.601	0.600	0.583	0.580	0.873	0.866	0.805	0.782
$k_1 = 30$	Sum	0.599	0.604	0.579	0.578	0.874	0.863	0.804	0.787
	UminP	0.563	0.546	0.511	0.503	0.848	0.833	0.770	0.737

Table 3.5: The p-values of the Tracy-Widom (TW) test and one-way ANOVA applied to the eigenvalues or PCs constructed from 10000 RVs.

# Eigenvalue or PC	TW	2-group ANOVA	10-subgroup ANOVA
1	0.000e+00	6.667e-24	2.955e-25
2	0.000e+00	3.061e-21	9.428e-28
3	0.000e+00	1.212e-28	6.830e-27
4	0.000e+00	0.1283	0.0393
5	0.000e+00	3.165e-33	8.635e-34
6	0.000e+00	6.745e-10	1.593e-27
7	0.000e+00	9.098e-08	5.310e-18
8	0.000e+00	0.0349	1.002e-25
9	0.000e+00	0.1118	1.221e-55
10	0.000e+00	0.0002	3.654e-09
11	0.000e+00	0.2260	1.376e-05
12	0.000e+00	0.9646	0.9393
13	0.000e+00	0.0013	9.934e-09
14	0.000e+00	0.1094	0.3517
15	0.000e+00	0.2797	0.3142
16	0.000e+00	0.8881	0.1215
17	0.000e+00	0.4054	0.0243
18	3.800e-03	0.9982	0.9916
19	4.230e-02	0.1583	0.7185
20	1.580e-01	0.7069	0.3861

Table 3.6: The numbers of samples assigned to each of the 12 clusters based on top 20 PCs constructed from 10000 RVs.

Subgroup/Cluster	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
CEU	72	13	0	0	0	0	5	0	0	0	0	0
FIN	9	27	0	0	0	0	0	0	0	0	0	0
GBR	29	13	0	0	0	0	1	0	0	0	0	0
TSI	73	14	0	0	0	0	0	1	1	1	1	1
MXL	5	0	0	0	0	3	9	0	0	0	0	0
PUR	0	0	0	0	0	0	0	1	1	1	1	1
YRI	0	0	68	1	9	0	0	0	0	0	0	0
LWK	0	0	0	38	24	5	0	0	0	0	0	0
ASW	0	0	1	0	14	9	0	0	0	0	0	0
PUR2	4	1	0	0	0	0	0	0	0	0	0	0

Table 3.7: Type I error and power for region R3. The PCs were constructed using 10000 RVs.

	Type I error				Power			
Test	#PCs=0	1	10	20	#PCs=0	1	10	20
Score	0.972	0.521	0.114	0.086	0.525	0.519	0.504	0.537
SSU	0.995	0.301	0.040	0.052	0.654	0.639	0.623	0.640
SSUw	0.992	0.542	0.056	0.070	0.659	0.652	0.634	0.652
Sum	0.995	0.818	0.076	0.050	0.671	0.664	0.628	0.630
UminP	0.900	0.561	0.108	0.088	0.492	0.476	0.427	0.434
T1, T5	0.995	0.821	0.061	0.038	0.663	0.673	0.624	0.608
Fp	0.993	0.816	0.064	0.041	0.658	0.654	0.615	0.606
VT	0.984	0.647	0.056	0.057	0.605	0.590	0.537	0.533
EREC	0.997	0.119	0.012	0.066	0.662	0.648	0.609	0.594

Chapter 4

Adjusting for Population Stratification in a Fine Scale with Principal Components and Sequencing Data

4.1 Introduction

In the last chapter, by using simulation studies with the 1000 Genomes Project data, we observed that population stratification could inflate the Type I error when testing on low frequency variants (LFVs) with MAFs between 1% and 5% or rare variants (RVs)with MAF$<$ 1%. However, including a few top PCs of a set of 10000 CVs or LFVs as covariates in a logistic regression model, the so called PCR model, could largely control the confounding effects of two continental groups. Generally, the power of PCR is maintained, except that there was a power loss with PCs of LFVs. PCs of 10000 RVs was a weaker performer with respect to control inflation.

Nonetheless, population stratification at the level of ethnic subgroups and the role of RVs in uncovering population structure, were still not fully investigated. Here subgroups, as opposed to continental groups like Europeans and Africans, are more confined in one area, where the geographic barrier and social or cultural differences limit the gene flow. Naturally, the following questions arise: Can PCs based on 10000 CVs or LFVs also adjust for population stratification in a fine scale? Can principal component analysis, or other more robust dimension reduction methods like spectral dimension reduction (SDR), satisfactorily improve PCs for uncovering subgroups? In doing so, which types of variants should be used, CVs, LFVs, RVs, or a mixture of all, and pruned or non-pruned ones? The main goal of this chapter is to address these questions using the same DNA sequencing data, the 1000 Genomes Project data. We still employ the PCR method, for its simplicity and good performance.

Several studies have touched on some of these issues, but with different types of data. Heath et al. (2008) [55] used autosomal SNPs selected from the Illumina HumanHap 300 chip to show that PCs based on CVs, LFVs or RVs had similar patterns for 13 subpopulations in Europe. The authors also observed that the top 2 PCs had strong correlations with geographic coordinates. However, Babron et al. (2012) [56] used the SNPs on the Affymetrix 500K chip to show some non-negligibly different influences by different types of variants on population stratification in 12 UK regions. It was suggested that the effect of population stratification was stronger when testing RVs. But in a simulated case-control study, they did not observe any major difference with the use of a few top PCs constructed from different types of variants when testing on variants with MAF $\geq$ 0.005. Both papers agreed that the

top PCs of CVs performed considerably well in controlling population stratification, and the number of variants used to construct PCs mattered in extracting ancestry information. Furthermore, the two papers showed that clustering analyses with either STRUCTURE [26] or Admixture [57] did not satisfactorily uncover the population subgroups. Using simulated sequencing data, Mathieson and McVean (2012) [38] clearly showed that none of the several commonly used adjustment methods worked well when testing on RVs in the presence of a local non-genetic risk; in contrast, all the methods performed well for testing on CVs. Due to the possible limitation with the use of the above genotyping or simulated data, it is not clear whether the above conclusions hold with real sequencing data.

Our study differs from the previous ones in the following aspects. First, instead of focusing only on one continent or one country, we used multiple subgroups of European and African ancestries, imposing a challenge to differentiating both dramatic and subtle genetic variations across the subgroups. Secondly, we compared PCA with another dimensional reduction method, spectral dimension reduction (SDR), in several scenarios with different types or sets of variants. We also simulated both binary traits and quantitative traits with varying population structures. Finally, we used a low-coverage whole genome sequencing dataset released in August 2010 by the 1000 Genomes Project (1000 Genomes Project Consortium 2010), not genotyping array data or simulated sequencing data. In particular, compared to genotyping array data, our data included much more variants of each type, facilitating a more realistic assessment on various approaches.

We tackled the problems by first visualizing the subpopulation structure using scatter plots of the top PCs. We also inspected how well we could uncover the subgroups by model-based clustering on a few top PCs. It helped us study which subgroups were genetically separable. The clustering result was evaluated mainly by the (adjusted) Rand Index ((a)RI) [58, 59]. Then we used three configurations of the subgroups to generate a binary trait with population stratification, and tested on pruned CVs, RVs and LFVs on chromosomes 1 and 2, with or without adjustment with a few top PCs constructed with various methods and types or sets of genome-wide variants. Lastly, a quantitative trait in the presence of a local non-genetic risk (i.e. spatially structured population) was simulated and studied. The estimated Type I error rates and inflation factors [23] were both used to evaluate to what

extent population stratification could be controlled.

4.2 Methods

4.2.1 PCA and SDR

Suppose $\tilde{X}$ is an n by p matrix, with n subjects and p SNVs, and $\tilde{X}_{im}$ denotes the genotype score of the m^{th} SNVs for the i^{th} subject. $\tilde{X}_{im}$ is coded 0, 1 or 2 as the minor allele count. Before we apply any method to construct PCs, each SNV is standardized as $(\tilde{X}_{im} - 2p_m)/\sqrt{p_m(1-p_m)}$ for all i and m, where p_m is the MAF for SNV m. No outliers are removed. We denote the standardized SNV matrix as X.

As widely known that PCA is sensitive to outliers, or unsuccessful in separating closely related subpopulations [60], Lee et al. (2009) [45] proposed a spectral clustering method, called SDR here. It is based on a normalized graph Laplacian matrix,

$$L = I - D^{-1/2} W D^{-1/2},$$

where W is a matrix measuring the similarities among subjects with elements

$$W_{ij} = \begin{cases} \sqrt{X_{i\cdot}' X_{j\cdot}} & \text{if } X_{i\cdot}' X_{j\cdot} \geq 0, \\ 0 & \text{if } X_{i\cdot}' X_{j\cdot} < 0, \end{cases}$$

where $X_{i\cdot}$ is the i^{th} row of X containing the standardized genotype scores of subject i, and $D = diag(D_{11}, ..., D_{nn})$ with $D_{ii} = \sum_{r=1}^{n} W_{ir}$. We sort the eigenvalues θ_q of $D^{-1/2} W D^{-1/2}$ decreasingly with the corresponding eigenvectors v_q for $q = 1, ...n-1$. As $D^{-1/2} W D^{-1/2}$ may not be positive semi-definite, we define $\lambda_q = \max\{0, \theta_q\}$. Hence, similar to PCA, the lth PC of SDR is

$$PC_l = \sqrt{\lambda_l} v_l.$$

Regarding how many PCs of PCA or SDR to use, we applied the Tracy-Widom (TW) test [24] or the eigengap heuristic method [45] respectively. Depending on the type of the variants used, it yielded varying numbers of the significant PCs

(p-value<0.05 for TW test) (Table 4.1). Note that using pruned RVs, the heuristic method identified a large number (453) of "significant" eigenvalues (or PCs) based on a calculated cut-off at 0.00575; if the cut-off was slightly increased to 0.00576, then only 48 eigenvalues were to be "significant". We also checked the top eigenvalues of the similarity matrices calculated based on different sets of SNVs: they appeared to be similar, but not exactly the same.

In summary, in most cases, 25 PCs appeared to suffice. We also tried adding more than 25 PCs in some simulations, but the performance was not improved.

Table 4.1: Numbers of significant eigenvalues from PCA by the Tracy-Widom test or those from SDR by a heuristic method.

	w/o pruning				with pruning				10000 pruned			
Methods	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
PCA	19	22	11	28	27	26	19	14	20	18	15	12
SDR	11	3	11	31	13	9	14	453	13	9	14	9

4.2.2 Data and association testing

We focused on 10 subgroups of 283 European and 174 African samples from the 1000 Genomes Project data. We constructed 12 different types of PCs based on 1) all variants, all CVs, all LFVs, and all RVs without pruning; 2) pruned set of all variants, CVs, LFVs and RVs; and 3) 10000 randomly selected pruned variants, CVs, LFVs, RVs.

For the purpose of association testing, all 10,848 pruned CVs with MAF>0.2, all 61,279 pruned LFVs and 50,476 pruned RVs were extracted from chromosomes 1 and 2 to be tested. We conducted a single SNP analysis by the score test on each CV. We scanned the RVs with 10092 overlapping sliding windows (with window size 20 and moving step 5) by the T1 and Fp tests implemented in software SCORE-Seq developed by Lin and Tang (2011) [20]. Both the T1 and Fp tests belong to the class of the burden tests, assessing the aggregated effects of a group of RVs (i.e. multiple RVs inside a sliding window here). Specifically, the T1 test only includes the RVs with MAF < 0.01 to be tested, while the Fp test gives each RV j a weight $1/\sqrt{f_j(1-f_j)}$, where f_j is a stabilized estimate of the MAF for RV j. The phenotypes were simulated independent of any SNVs, but solely related to ethnic subgroups to create population stratification.

Since the resampling method is computationally intensive, and as far as we observed, asymptotic p-values and resampling-based p-values were close in a few examples in this study, we only used asymptotic p-values. At the nominal significance level 0.05, the Type I error rate was estimated as the proportion of the test with a p-value less than 0.05 with single run over the set of the SNVs to be tested, as the simulated phenotypes were not associated with any SNVs to be tested. The inflation factor, λ, was defined as the ratio of the median of the χ^2 statistics corresponding to the observed p-values and that of the expected ones under the null hypothesis of no association, as proposed in genomic control to correct for the inflation of p-values caused by cryptic relatedness or population structure in case-control studies; we used the R package **gap** authored by Jinghua Zhao for its calculation.

4.2.3 Simulation of binary traits

To assess the performance of PCs constructed by SDR or PCA in association testing, we first designed three case-control studies with population structures as the following: we assigned different subgroups as "cases" and the rest as "controls", as summarized in Table 4.2. The presence of population stratification was expected given that the MAFs of some SNVs varied among the subgroups. At the same time, due to the arbitrary assignments of the "disease" status, there should be no association between any SNV and the "disease" after a proper adjustment for population stratification. Hence, the above simulation set-ups were used to investigate how well various PCs could control Type I error rates and λ's in association testing.

Table 4.2: Simulation set-ups with binary traits.

Set-up	Cases	Controls	all samples #Cases/#Controls	w/o AMRs #Cases/#Controls
1	CEU, FIN, MXL, PUR, LWK, ASW	GBR, TSI, PUR2, YRI	239/218	217/213
2	CEU, TSI, FIN, PUR, ASW	GBR, MXL,YRI, LWK, PUR2	247/210	242/188
3	CEU, TSI, ASW	GBR, FIN, MXL, YRI, LWK, PUR2	206/251	206/229

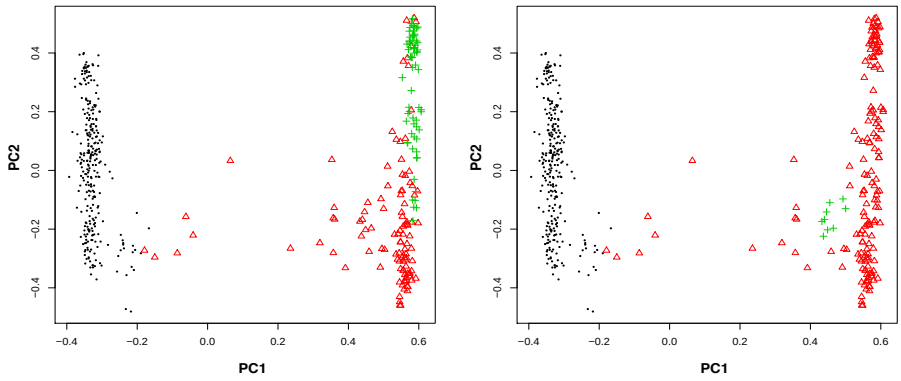
4.2.4 Simulation of quantitative traits with a local non-genetic risk

As in Mathieson and McVean (2012) [38], we were also interested in investigating how PCs could control inflations caused by spatially structured populations induced

by a local non-genetic risk. Specifically, we looked into the scenarios where only a neighborhood area endured a high environmental risk factor. To mimic the well-known correspondence between top PCs and geographical locations, we used the top two PCs constructed by PCA based on 10000 pruned CVs as the location coordinates and chose a medium or small region to have a high non-genetic risk. Two such local regions were selected: one with 65 out of 78 YRI samples to form the risk region R1, and the other with 10 out of 24 MXL samples to form R2 (Figure 4.1).

The phenotype y_i for subject i was simulated as $y_i = \mu_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$ and μ_i was the non-genetic risk. Here we used the so-called "square risk" such that only the samples in the risk region suffered from the elevated environmental risk: $\mu_i = 10$ for any sample i in the risk region, and $\mu_i = 0$ otherwise.

Figure 4.1: Two local risk regions R1 and R2 represented by the top two PCs of PCA based on the 10000 pruned CVs. The risk regions are marked in green crosses. EURs are in black dots and AFRs are in red triangles.



(a) R1 (b) R2

4.3 Results

4.3.1 Population structure

We first looked at Wright's Fst statistic (Wright, 1984) calculated in software EIGENSTRAT (Price et al., 2006; Patterson et al., 2006) to assess the genetic differences among the subgroups. The software was downloadable at

`http://www.hsph.harvard.edu/faculty/alkes-price/software/`.

Since Mathieson and McVean(2012) showed by simulations that Fst statistics varied dramatically when calculated with SNVs of different MAFs, we calculated Fst statistics based on all pruned variants (Table 4.3), all pruned CVs and all pruned RVs (Table A.1, A.2). We noticed that Fst statistics based on all pruned variants were very similar to those based on all pruned CVs. Even though these results were a little different from those from RVs, the general patterns were the same. Based on Table 4.3, we could see that the Fst statistic between any two European subgroups was usually <0.01, e.g. the Fst between FIN and GBR is 0.007. They were much smaller than the Fst's between two subgroups from the different continents, such as between FIN and LWK (0.145), or between two African subgroups. The smallest Fst was found between CEU and GBR (0.004), indicating a high genetic similarity between the two subgroups. The Fst statistics between African subgroups were also small, about $0.011\sim0.012$, but they were generally greater than those between European subgroups. The AMR subgroups, including MXL, PUR and PUR2, were genetically different from other European or African subgroups.

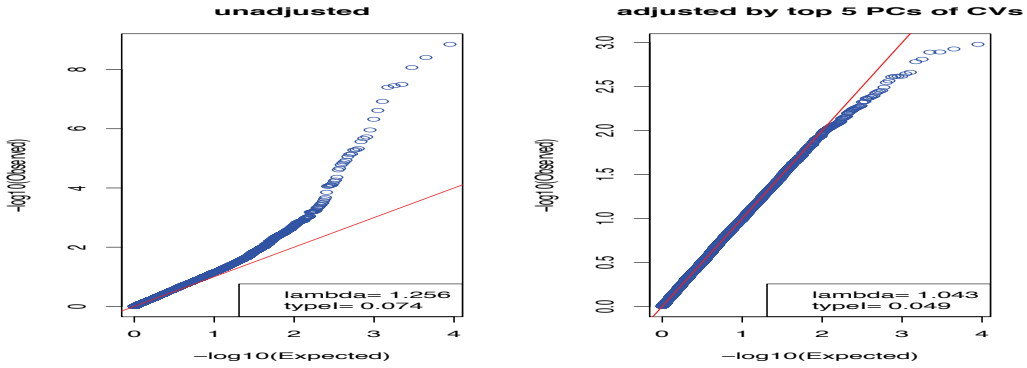
Given the small genetic difference between CEU and GBR subgroups as shown by their Fst statistic, one might wonder whether they could be treated as a homogeneous population without causing any problem. We tested all the pruned CVs on chromosomes 1 and 2 between the two subgroups by the score test (Figure 4.2). If indeed there was no genetic difference between the two subgroups, we would expect to have a rejection rate of 0.05 and an inflation factor around 1. In contrast, we had an inflated rejection rate of 0.074 and an estimated inflation factor of 1.427, indicating the presence of population stratification for these two genetically similar subgroups. Furthermore, an adjustment with the top 5 PCs of SDR based on all CVs reduced the inflation factor to 1.041. These results suggested that CEU and GBR samples were still genetically different enough to warrant a careful controlling

for population stratification, and importantly the population stratification could be largely controlled by using a few top PCs.

Table 4.3: Fst statistics between subgroups calculated with all pruned variants.

	GBR	FIN	PUR	CEU	MXL	TSI	PUR2	YRI	LWK	ASW
GBR	0.000	0.007	0.010	0.004	0.022	0.008	0.025	0.153	0.140	0.098
FIN	0.007	0.000	0.014	0.010	0.026	0.015	0.029	0.159	0.145	0.104
PUR	0.010	0.014	0.000	0.013	0.013	0.013	-0.087	0.116	0.104	0.071
CEU	0.004	0.010	0.013	0.000	0.024	0.004	0.027	0.150	0.140	0.099
MXL	0.022	0.026	0.013	0.024	0.000	0.025	0.025	0.134	0.122	0.083
TSI	0.008	0.015	0.013	0.004	0.025	0.000	0.026	0.145	0.135	0.095
PUR2	0.025	0.029	-0.087	0.027	0.025	0.026	0.000	0.099	0.087	0.057
YRI	0.153	0.159	0.116	0.150	0.134	0.145	0.099	0.000	0.012	0.012
LWK	0.140	0.145	0.104	0.140	0.122	0.135	0.087	0.012	0.000	0.011
ASW	0.098	0.104	0.071	0.099	0.083	0.095	0.057	0.012	0.011	0.000

Figure 4.2: Q-Q plots for the score test in the simulated case-control study where CEUs are "cases" and GBRs are "controls".



4.3.2 Pairwise PC-plots for PCA and SDR

To visualize the population structure, we examined the scatter plots of a few top PCs of PCA and SDR. First we look at SDR. The top two PCs based on any type of variants could successfully separate the two continental groups, EURs and AFRs (Figure 4.3). It is also not difficult to notice the similarity between the plots using all variants and using all CVs. This was probably due to that the number of CVs was 3∼4 times as large as that of LFVs or of RVs. Based on all variants and all

CVs, the subgroups of AFRs were also generally distinguishable.

The scatter plots of PCs of PCA (Figure 4.4) appear to show different patterns from those of SDR. While the second PC of SDR separated continental groups, the 1st PC of PCA separated them. The top 2 PCs based on all variants were again very similar to those of all CVs. The subgroups of AFRs were separable by the PCs of all variants, all CVs or all LFVs. However, based on the PCs of all RVs we were not able to differentiate any subgroups; the top 2 PCs separated out two outliers (PURs) from the main body of other samples (Table A.6). This confirms that PCA may be heavily influenced by outliers. The pairwise plots of the 3rd and 4th PCs (i.e. PC3 and PC4) constructed by SDR or PCA are included in the supplementary material (Suppl Figure A.1).

We were interested in seeing whether excluding the AMRs would improve the performance of PCs since the AMRs appeared different from other EUR and AFR subgroups. The plots without AMRs are shown in Suppl Figure A.3 and A.4. Excluding the AMRs barely changed the pairwise plots of the top 2 PCs of either SDR or PCA as it did not help differentiate the EURs subgroups. This might not be surprising according to Patterson et al. (2006); they discussed that adding an admixed population will not change the significant number of PCs. In our case, considering AMRs as admixed American samples founded by Europeans, native Americans and some other ancestry groups, these samples had coordinates joining the centers of the EURs and other founding populations. In other words, with or without the AMRs, the significant number of axes of variation stayed the same. This situation differed from that in Lee et al. (2009), in which it was shown that in a dataset of 580 Europeans, all self-identified as Italians and British, adding 1 African American, 1 East Asian, 1 Indian and 1 Mexican sample had a great impact on the top 4 PCs of PCA.

As the 2nd PC of SDR accounted for the variation between continental groups, we projected the data onto the panel of the 2nd and other PCs in order to find the best visualization of the subgroups by a pairwise PC plot. Figure 4.5 shows that with the 2nd and 5th PCs of SDR based on all CVs, we could almost separate all the subgroups except CEUs and GBRs. But based on all pruned CVs, all RVs or all pruned RVs, EUR subgroups were still mixed together.

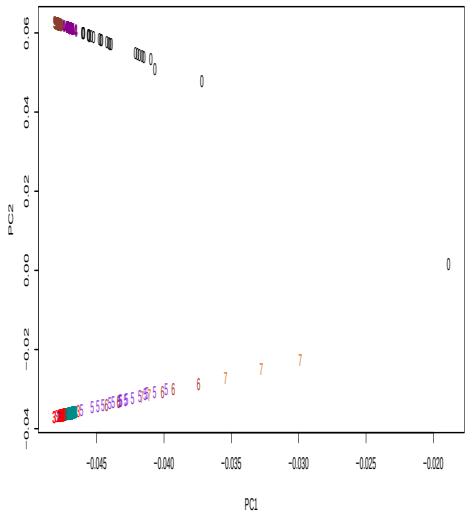
It is suggested that instead of using all variants, we should prune variants to infer

the ancestry information, as it was claimed that linkage disequilibrium (LD) "will seriously distort the eigenvector/eigenvalue structure" and thus result in misleading PCs (Patterson et al., 2006), though we could not find any theoretical justification for such a claim. It is also common to randomly select a large number of, e.g. 10000, pruned variants to construct PCs for its computational convenience. We show a comparison among the top 2 PCs of PCA based on all pruned CVs, 10000 pruned CVs, all pruned RVs or 10000 pruned RVs (Figure 4.6); a comparison of PCs of SDR is shown in Suppl Figure A.5. We notice that the plot of the top 2 PCs based on all pruned CVs resembled that of 10000 pruned CVs. This suggests that probably we did not lose much information by randomly select a large number of pruned CVs as compared to using all pruned CVs.

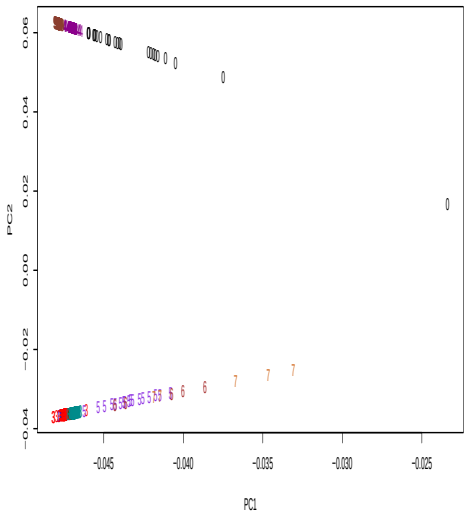
In summary, the continental groups were generally differentiable by a few top PCs of either SDR or PCA. With a few top PCs of PCA or SDR based on all variants, all CVs or all LFVs, the subgroups of AFRs were scattered apart. Even better, most samples represented by the top PCs of SDR based on all CVs could be divided into separate subgroups. However, the top PCs based on RVs were weaker in separating the subgroups.

Figure 4.3: The top 2 PCs of SDR. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), MXL=5(blueviolet), PUR=6(brown), PUR2=7(chocolate), TSI=8(cyan), YRI=9(coral).

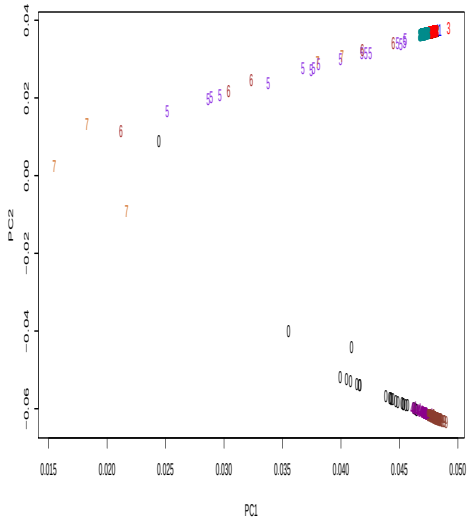
all variants



all CVs



all LFVs



all RVs

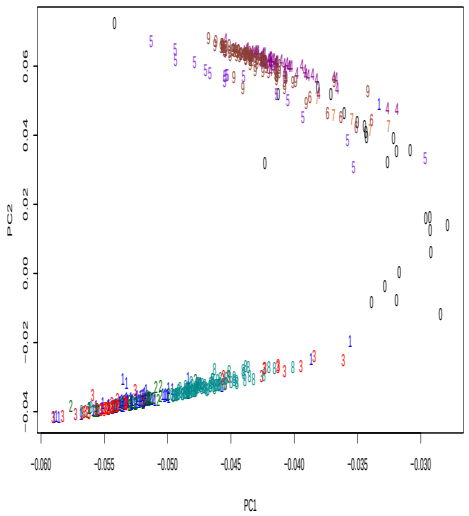
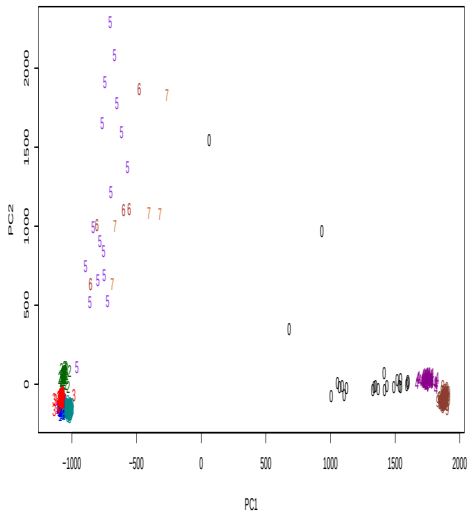
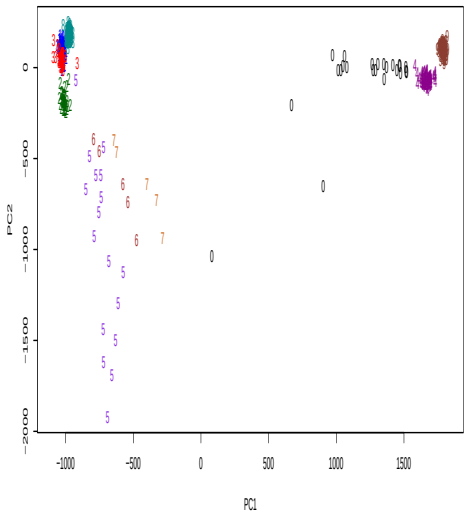


Figure 4.4: The top 2 PCs of PCA. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), MXL=5(blueviolet), PUR=6(brown), PUR2=7(chocolate), TSI=8(cyan), YRI=9(coral).

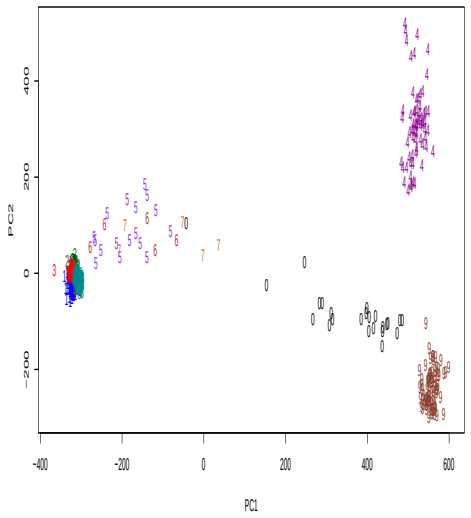
all variants



all CVs



all LFVs



all RVs (zoom in)

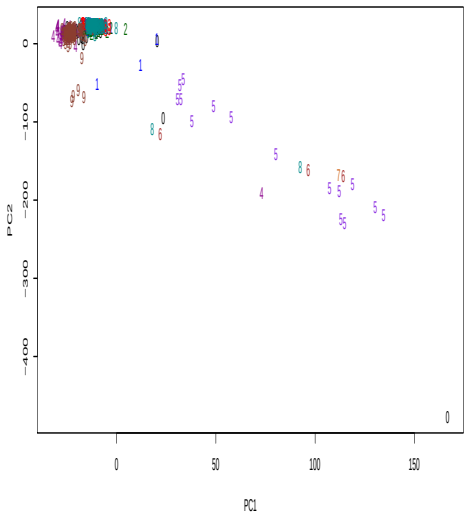
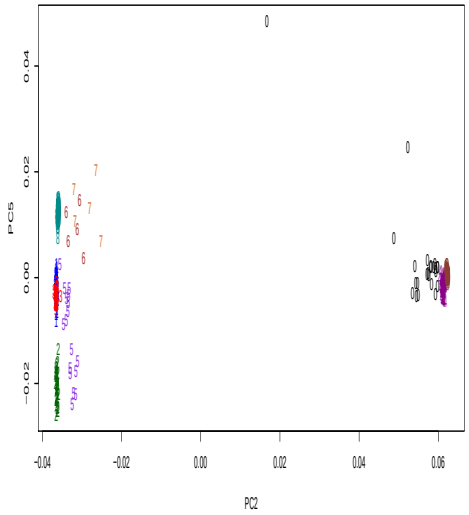
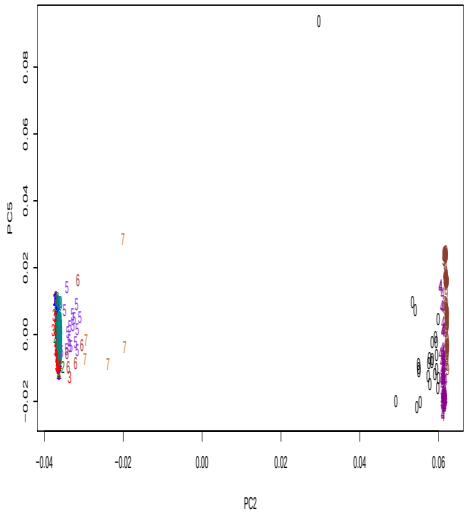


Figure 4.5: The subgroups represented by two PCs of SDR based on all CVs, all pruned CVs, all RVs and all pruned RVs. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), MXL=5(blueviolet), PUR=6(brown), PUR2=7(chocolate), TSI=8(cyan), YRI=9(coral).

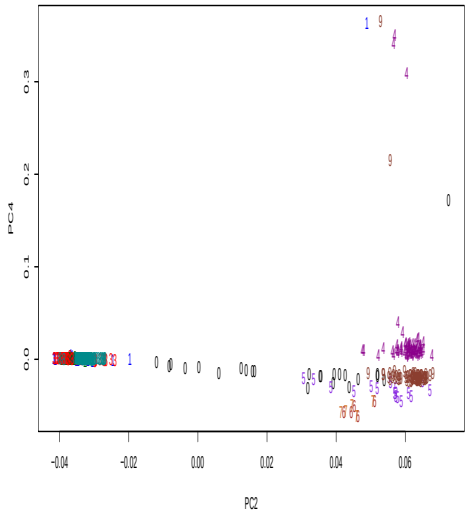
all CVs



all pruned CVs



all RVs



all pruned RVs

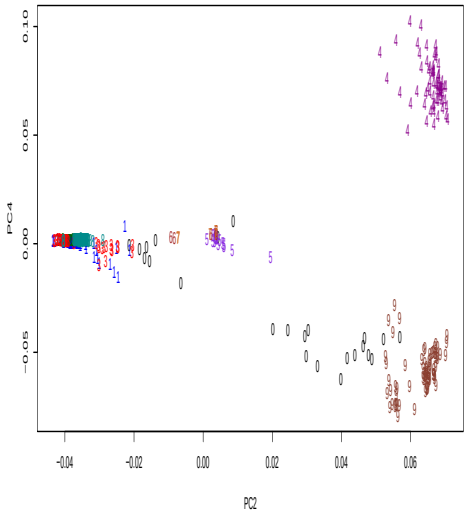
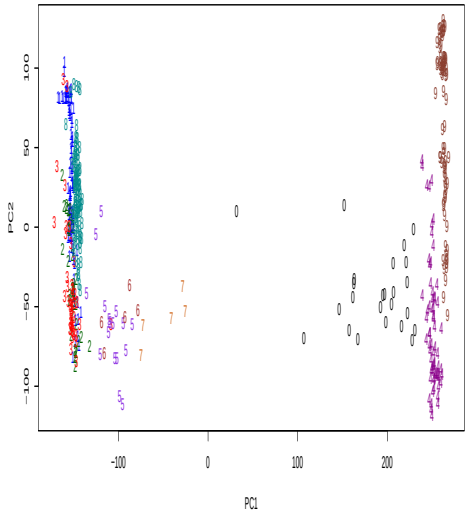
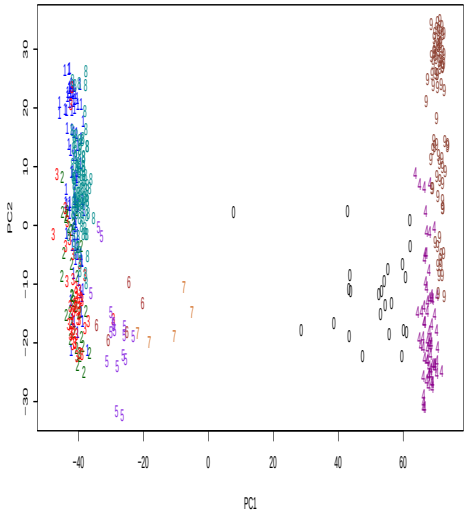


Figure 4.6: Comparison of the top 2 PCs of PCA based on all (or 1000) pruned CVs (or RVs). ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), MXL=5(blueviolet), PUR=6(brown), PUR2=7(chocolate), TSI=8(cyan), YRI=9(coral).

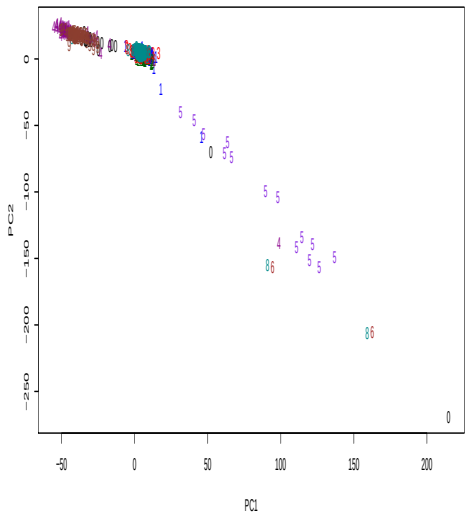
all pruned CVs



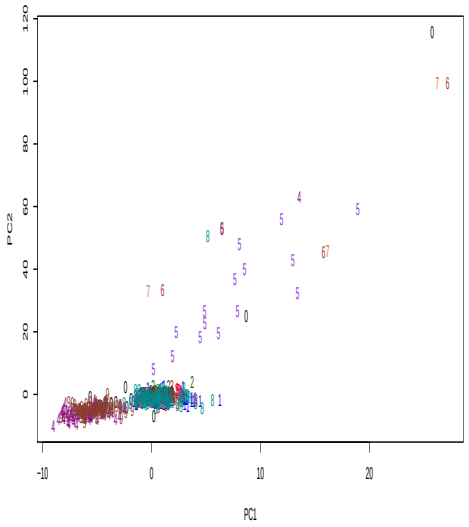
10000 CVs



all pruned RVs



10000 pruned RVs (zoom in)



4.3.3 Clustering analysis

Since it is difficult to visualize high-dimensional scatter plots, we applied Gaussian model-based cluster analysis (Banfield and Raftery, 1993) to the top 25 PCs to see whether we could reach the same conclusions as before.

Due to the importance of a specified number of clusters in clustering analysis, we first studied how the RI and aRI changed with the number of clusters (Suppl Figure A.7). This allowed us to choose the cluster number to achieve the best separation of the samples. Suppl Figure A.7 shows that based on all variants, all LFVs and all RVs, the maximum (a)RI values of clustering based on the top 25 PCs of SDR outperformed those of PCA. But based on all CVs, PCs of SDR and PCA performed equally well. Overall, the (a)RI based on all variants and all CVs were higher than those based on all LFVs and all RVs.

Table 4.4 shows the results with the optimal number of clusters (giving the largest RI), and with 10 clusters, which was the number of the true subgroups. The highest RI (0.957) and aRI (0.830) with all samples included were obtained using the PCs of SDR based on all variants at 8 clusters. It was followed by clustering using the PCs of SDR based on all CVs (RI=0.953 and aRI=0.805) at 10 clusters. The PCs of PCA based on all CVs also did a good job. The PCs of SDR based on the pruned RVs achieved the 4th highest (a)RI (RI=0.941 and aRI=0.777). After excluding the AMRs (Suppl table A.3), the best clustering results were still obtained when using the top PCs of SDR with all variants (RI=0.963 and aRI=0.865) or all CVs (RI=0.959 and aRI=0.842).

In conclusion, in agreement with the earlier visual inspection of the PC plots, the best clustering results were obtained by using the top PCs of SDR with all variants or with all CVs. These two types of variants were anticipated to be able to adjust for population stratification. The results also confirmed that we could not perfectly estimate every sample's subgroup identity based on the top 25 PCs of the whole-genome data.

Table 4.4: Clustering results with the top 25 PCs for all samples.

		w/o pruning				with pruning				10000 pruned			
		all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
SDR	best cluster #	8	10	9	10	13	17	14.	7	10	26	28	15
	best RI	0.957	0.953	0.928	0.889	0.907	0.876	0.901	0.941	0.891	0.880	0.889	0.897
	best aRI	0.830	0.805	0.719	0.532	0.554	0.383	0.510	0.777	0.526	0.348	0.421	0.562
	RI at 10 clusters	0.947	0.953	0.917	0.889	0.888	0.863	0.882	0.892	0.891	0.849	0.836	0.836
	aRI at 10 clusters	0.776	0.805	0.662	0.532	0.535	0.438	0.480	0.578	0.526	0.382	0.442	0.511
PCA	best cluster #	13	8	22	6	15	12	15	23	8	9	16	13
	best RI	0.929	0.947	0.900	0.755	0.888	0.890	0.897	0.850	0.904	0.917	0.888	0.828
	best aRI	0.675	0.803	0.508	0.404	0.435	0.474	0.479	0.394	0.629	0.643	0.467	0.375
	RI at 10 clusters	0.925	0.931	0.884	0.730	0.879	0.869	0.881	0.832	0.881	0.914	0.868	0.822
	aRI at 10 clusters	0.682	0.698	0.558	0.312	0.478	0.432	0.471	0.376	0.502	0.620	0.464	0.435

4.3.4 Association testing

Binary traits and CVs

We started with controlling population stratification in testing CVs. In general $\lambda \leq 1.05$ was the recommended criterion to determine that there was no population stratification (Price et al., 2010). However, due to the relatively small sample size and possibly correlated variants (and thus correlated p-values), here we would use a less restrictive criterion of $\lambda \leq 1.15$. Across all three simulation set-ups with all samples included (Table 4.5, Suppl Tables A.6, A.7, A.8, A.9, A.10), the best adjustments were obtained by using the top 25 PCs of SDR based on all variants, all CVs and all pruned CVs. For example, in simulation set-up 2 (Table 4.5), without adjustment, the Type I error rate was far beyond the nominal level 0.05 while the inflation factor λ was as high as 23.755. After adjusted with a few top PCs, the Type I error rates and λ were much reduced. With the top 10 PCs of SDR based on all variants, we obtained Type I error 0.066 and λ 1.126; with the top 15 PCs of SDR based on all CVs, we reduced Type I error further to 0.064 and λ to 1.098; with the PCs based on the pruned CVs, of either SDR or PCA, the Type I error was about 0.06 and λ was about 1.166. With the AMR samples excluded (Supple Table A.6), the results were similar. Overall, the three types of PCs named above – the PCs of SDR based on all variants, all CVs or all pruned CVs were the best performers.

Suppl Figure A.8 shows the Q-Q plots of the p-values of the tests with or without adjustment by the top PCs of SDR in all three simulation set-ups. We could see that without adjustment, the p-values were far above the 45 degree identity line, while

after adjustment with SDR, the p-values were almost uniformly distributed along the identity line. The top PCs of SDR based on all variants, all CVs or all pruned CVs consistently controlled population stratification across all three simulation set-ups.

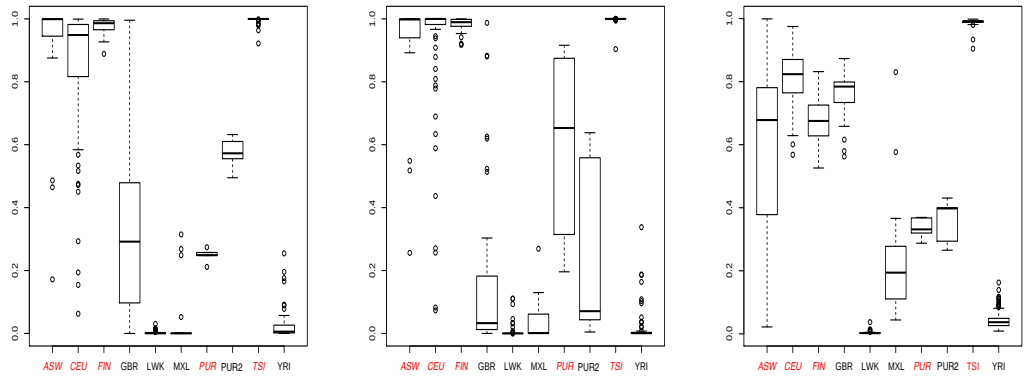
Table 4.5: Results of association testing on CVs with a binary trait in simulation set-up 2.

			w/o pruning				with pruning				10000 pruned			
		#PCs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677	0.677
		10	0.066	0.069	0.092	0.094	0.077	0.077	0.093	0.093	0.083	0.099	0.089	0.102
		15	0.069	0.064	0.091	0.097	0.086	0.066	0.097	0.097	0.092	0.092	0.092	0.102
		20	0.068	0.064	0.086	0.101	0.090	0.079	0.084	0.084	0.091	0.078	0.097	0.102
	PCA	10	0.083	0.084	0.085	0.088	0.083	0.087	0.081	0.081	0.088	0.084	0.080	0.091
		15	0.090	0.070	0.090	0.096	0.094	0.070	0.090	0.090	0.085	0.074	0.083	0.092
		20	0.092	0.070	0.093	0.103	0.098	0.062	0.089	0.089	0.088	0.073	0.089	0.093
λ	SDR	0	23.755	23.755	23.755	23.755	23.755	23.755	23.755	23.755	23.755	23.755	23.755	23.755
		10	1.126	1.136	1.302	1.349	1.225	1.205	1.331	1.331	1.307	1.410	1.304	1.414
		15	1.164	1.098	1.319	1.391	1.280	1.168	1.376	1.376	1.383	1.355	1.360	1.426
		20	1.174	1.150	1.312	1.434	1.315	1.255	1.310	1.310	1.290	1.233	1.440	1.427
	PCA	10	1.318	1.340	1.325	1.273	1.309	1.362	1.310	1.310	1.361	1.350	1.283	1.323
		15	1.308	1.167	1.317	1.347	1.372	1.216	1.374	1.374	1.297	1.276	1.327	1.324
		20	1.360	1.210	1.368	1.443	1.402	1.166	1.363	1.363	1.340	1.241	1.375	1.357

One interesting observation was that, while the PCs of SDR based on the pruned RVs achieved higher (a)RI than those based on the pruned CVs in clustering analysis, the former ones could not, but the latter could, effectively control Type I errors in association testing. To explore the reasons for this discrepancy, we first compared the clustering results by using the top PCs of SDR based on the pruned CVs or the pruned RVs (Suppl Tables A.4, A.5). However, both mis-classified several subgroups such that we could not determine which type of variants yielded a higher accuracy. Next we fitted a logistic regression model on the simulated disease status (in simulation set-up 2) against the top 25 PCs of SDR and plotted the predicted probability of each sample's having disease versus his/her subgroup membership (Figure 4.7). When we used the PCs based on the pruned RVs, the predicted probabilities of having disease for both CEU and GBR samples were close to 1, while in fact all the GBRs were assigned to be controls. But when we used the PCs based on the CVs or the pruned CVs, in agreement to the truth, the predicted probabilities of CEUs were close to 1 and those of GBRs were close to 0. Furthermore, with the PCs of all CVs or the pruned CVs, we could even separate the 5 PUR samples and 5 PUR2 samples. These results indicated that with the PCs of SDR based on the pruned RVs, we could not differentiate GBRs from CEUs, but by PCs of all CVs or the

pruned CVs we could largely distinguish the two genetically similar subgroups.

Figure 4.7: Distributions of the estimated probabilities of having disease for subjects in each subgroup based on the top 25 PCs of SDR in simulation iset-up 2. The subgroups marked in red are cases.



(a) all CVs (b) all pruned CVs (c) all pruned RVs

Binary traits and RVs/LFVs

To investigate controlling population stratification when testing variants with lower MAFs, we scanned all the pruned RVs (or all the pruned LFVs) on chromosomes 1 and 2 with sliding windows using the T1 (or T5) and Fp tests (Table 4.6) and Suppl Tables A.11, A.12, A.13, A.14, A.15).

Across all three simulation set-ups, the top PCs of SDR based on all variants, all CVs or all pruned CVs could generally control the Type I error rates around 0.06 and λ's below 1.15 as in testing CVs. For example, in Table 4.6, with the top 10 PCs of all variants the Type I error of the T1 test was 0.040 and λ=1.009; with the top 10 PCs of all CVs the Type I error of the T1 test was 0.044 and $\lambda = 1.025$, while with the top 10 PCs of the pruned CVs the Type I error was 0.047 and λ=0.988. However, different from testing CVs, none of the top PCs of PCA could effectively control the inflation factor below 1.15 when testing RVs.

In conclusion, our simulation results demonstrated that a few top PCs based on all variants, all CVs or all pruned CVs constructed by SDR could generally control population stratification in testing multiple RVs or multiple LFVs.

Table 4.6: Results of association testing on RVs with a binary trait in simulation set-up 2.

				w/o pruning				with pruning				10000 pruned			
		#PCs	Test	all	CV	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T1	0.417	0.417	0.417	0.417	0.417	0.417	0.417	0.417	0.417	0.417	0.417	0.417
			Fp	0.392	0.392	0.392	0.392	0.392	0.392	0.392	0.392	0.392	0.392	0.392	0.392
		10	T1	0.040	0.044	0.058	0.071	0.065	0.047	0.076	0.092	0.066	0.060	0.076	0.112
			Fp	0.041	0.045	0.058	0.072	0.063	0.048	0.074	0.094	0.066	0.059	0.074	0.111
		25	T1	0.087	0.064	0.071	0.090	0.074	0.081	0.146	0.123	0.071	0.066	0.075	0.113
			Fp	0.087	0.066	0.071	0.088	0.073	0.082	0.145	0.120	0.074	0.069	0.075	0.116
	PCA	10	T1	0.114	0.114	0.124	0.099	0.101	0.112	0.077	0.068	0.138	0.111	0.079	0.091
			Fp	0.116	0.114	0.126	0.103	0.101	0.113	0.078	0.065	0.141	0.115	0.078	0.091
		25	T1	0.066	0.098	0.072	0.063	0.074	0.120	0.098	0.066	0.085	0.125	0.084	0.074
			Fp	0.065	0.097	0.071	0.063	0.075	0.124	0.099	0.063	0.085	0.131	0.085	0.071
λ	SDR	0	T1	6.114	6.114	6.114	6.114	6.114	6.114	6.114	6.114	6.114	6.114	6.114	6.114
			Fp	5.665	5.665	5.665	5.665	5.665	5.665	5.665	5.665	5.665	5.665	5.665	5.665
		10	T1	1.009	1.025	1.122	1.246	1.161	0.988	1.297	1.470	1.237	1.098	1.309	1.669
			Fp	1.004	1.006	1.137	1.253	1.187	0.986	1.289	1.472	1.222	1.107	1.308	1.705
		25	T1	1.463	1.206	1.214	1.431	1.293	1.336	1.841	1.707	1.257	1.147	1.228	1.693
			Fp	1.459	1.198	1.202	1.440	1.284	1.334	1.844	1.731	1.212	1.170	1.238	1.701
	PCA	10	T1	1.699	1.708	1.854	1.530	1.491	1.656	1.314	1.205	2.002	1.610	1.311	1.454
			Fp	1.774	1.763	1.892	1.556	1.525	1.687	1.347	1.192	2.027	1.690	1.339	1.479
		25	T1	1.191	1.482	1.342	1.199	1.230	1.624	1.456	1.213	1.350	1.781	1.308	1.254
			Fp	1.218	1.487	1.368	1.199	1.241	1.623	1.465	1.191	1.343	1.786	1.305	1.276

Testing RVs in the presence of a local non-genetic risks

Population stratification can be a more severe issue in association testing for RVs when the samples collected were exposed to some localized environmental risks, as discussed in Mathieson and McVean (2011)[38]. One of their main observations was that, when the non-genetic risk region was small sized, the inflation of Type I error could be controlled for testing CVs, but not for RVs, after adjusting with a few top PCs of PCA. Since they used simulated sequencing data while acknowledging the need to use real sequencing data, we addressed the problem with real sequencing data.

First, when we tested CVs without any adjustment, there were severe inflations of both Type I error and λ for both local regions R1 and R2. As expected, the inflation could be well controlled with the use of a few top PCs of either SDR or PCA based on all variants, all CVs or all LFVs without pruning (Table 4.7). The top PCs of SDR, but not of PCA, based on RVs also worked well.

When testing multiple RVs, again we saw inflated Type I error and λ without adjustment. In Table 4.8 for local region R1, using the top PCs of SDR based on all variants, all CVs or all RVs could largely, but not perfectly, control the inflation,

where Type I error rates were around 0.075 and λ around 1.16. The top PCs of SDR based on the pruned CVs performed less well in controlling inflations (with Type I error 0.080 and $\lambda = 1.211$ for the T1 test). Similar results were obtained for sliding windows of size 20 (Suppl Table A.16). For example, when testing with window size 10, the top 10 PCs of all CVs could reduce Type I error to .074 and λ to 1.170 for the T1 test, while testing with window size 20, they could be reduced to 0.078 and 1.209 respectively.

The results for local region R2 were a little different. While the top PCs based on all variants or all CVs worked a bit poorer in controlling the inflation, the top PCs of SDR based on all or some RVs could still maintain the Type I error and λ at an acceptable level. For example, when testing with window size 10 (Table 4.9), with adjustment of the top 25 PCs of SDR based on all CVs, the Type I error was only reduced from 0.109 to 0.096 for the T1 test; in contrast, adjusting with the top 25 PCs of SDR based on all RVs, Type I error was only around 0.069 and λ around 1.20. On the other hand, the top PCs of RVs performed worse than those of CVs if PCA was used. Similar results were obtained for testing with window size 20 (Suppl Table A.17). In summary, for R2 it seemed that using the top PCs of RVs in SDR performed better, though it was the reverse if PCA was used.

Overall, in the presence of a local non-genetic risk, although a few top PCs of any type of variants without pruning could control the inflation effectively when testing CVs, it was not the case when testing RVs. Again we found that a few top PCs of SDR based on all variants or all CVs were among the best candidates for controlling the inflation, though the top PCs of SDR based on all RVs also performed well or even slightly better.

Table 4.7: Results of association testing on CVs for two local non-genetic risk regions R1 and R2. The PCs were constructed with unpruned variants.

			R1				R2			
		#PCs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	0.640	0.640	0.640	0.640	0.101	0.101	0.101	0.101
		2	0.141	0.142	0.135	0.258	0.063	0.062	0.061	0.082
		10	0.053	0.055	0.066	0.048	0.055	0.055	0.062	0.066
		25	0.057	0.057	0.056	0.054	0.061	0.062	0.057	0.070
	PCA	2	0.128	0.128	0.051	0.645	0.057	0.056	0.057	0.134
		10	0.054	0.055	0.053	0.520	0.061	0.060	0.058	0.105
		25	0.066	0.065	0.059	0.211	0.059	0.059	0.057	0.104
λ	SDR	0	17.813	17.813	17.813	17.813	1.573	1.573	1.573	1.573
		2	1.748	1.748	1.710	3.023	1.099	1.094	1.081	1.339
		10	0.990	1.010	1.116	1.006	1.059	1.059	1.087	1.140
		25	1.035	1.036	1.021	1.006	1.069	1.094	1.061	1.167
	PCA	2	1.580	1.582	0.994	18.142	1.070	1.070	1.059	1.789
		10	1.058	1.066	1.041	9.348	1.087	1.081	1.078	1.499
		25	1.083	1.097	1.048	2.389	1.059	1.070	1.039	1.487

Table 4.8: Results of association testing on RVs with window size 10 in the presence of a local non-genetic risk region R1.

				w/o pruning				with pruning			
		#PCs	Test	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T1	0.190	0.190	0.191	0.190	0.190	0.190	0.190	0.190
			Fp	0.171	0.171	0.172	0.171	0.171	0.171	0.171	0.171
		2	T1	0.134	0.135	0.140	0.156	0.140	0.136	0.140	0.135
			Fp	0.130	0.130	0.136	0.153	0.135	0.131	0.136	0.131
		10	T1	0.074	0.074	0.080	0.073	0.141	0.094	0.075	0.077
			Fp	0.072	0.075	0.082	0.071	0.139	0.091	0.076	0.076
		25	T1	0.080	0.080	0.084	0.079	0.081	0.080	0.076	0.082
			Fp	0.076	0.076	0.081	0.078	0.078	0.078	0.076	0.080
	PCA	2	T1	0.144	0.143	0.095	0.202	0.153	0.092	0.090	0.203
			Fp	0.157	0.154	0.107	0.205	0.168	0.094	0.103	0.207
		10	T1	0.099	0.092	0.102	0.156	0.085	0.098	0.086	0.114
			Fp	0.099	0.094	0.102	0.147	0.084	0.097	0.083	0.115
		25	T1	0.083	0.084	0.092	0.119	0.092	0.084	0.080	0.088
			Fp	0.083	0.083	0.093	0.117	0.090	0.083	0.081	0.085
λ	SDR	0	T1	2.100	2.100	2.110	2.100	2.100	2.100	2.100	2.100
			Fp	1.928	1.928	1.941	1.928	1.928	1.928	1.928	1.928
		2	T1	1.745	1.747	1.767	1.978	1.768	1.729	1.756	1.725
			Fp	1.682	1.676	1.744	1.896	1.742	1.709	1.726	1.715
		10	T1	1.153	1.170	1.236	1.152	1.767	1.298	1.178	1.180
			Fp	1.141	1.173	1.229	1.143	1.735	1.310	1.168	1.176
		25	T1	1.222	1.204	1.204	1.220	1.205	1.211	1.207	1.203
			Fp	1.230	1.205	1.191	1.210	1.209	1.219	1.186	1.212
	PCA	2	T1	1.822	1.781	1.300	2.291	1.898	1.377	1.291	2.263
			Fp	1.781	1.754	1.312	2.110	1.906	1.311	1.281	2.133
		10	T1	1.376	1.320	1.376	1.868	1.303	1.457	1.277	1.522
			Fp	1.350	1.296	1.354	1.751	1.288	1.459	1.265	1.514
		25	T1	1.193	1.224	1.200	1.646	1.339	1.264	1.208	1.289
			Fp	1.183	1.212	1.218	1.655	1.362	1.238	1.181	1.291

Table 4.9: Results of association testing on RVs with window size 10 in the presence of a local non-genetic risk region R2.

				w/o pruning				with pruning			
		#PCs	Test	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T1	0.109	0.109	0.109	0.109	0.109	0.109	0.109	0.109
			Fp	0.109	0.109	0.109	0.109	0.109	0.109	0.109	0.109
		2	T1	0.094	0.095	0.095	0.087	0.093	0.093	0.094	0.089
			Fp	0.095	0.096	0.096	0.088	0.096	0.095	0.095	0.090
		10	T1	0.097	0.097	0.096	0.089	0.099	0.097	0.099	0.088
			Fp	0.096	0.098	0.096	0.091	0.099	0.097	0.100	0.089
		25	T1	0.092	0.096	0.097	0.069	0.095	0.095	0.095	0.073
			Fp	0.092	0.097	0.099	0.070	0.095	0.096	0.095	0.073
	PCA	2	T1	0.101	0.100	0.100	0.120	0.100	0.095	0.100	0.123
			Fp	0.105	0.104	0.102	0.121	0.102	0.097	0.104	0.124
		10	T1	0.108	0.107	0.109	0.120	0.106	0.100	0.097	0.117
			Fp	0.109	0.109	0.111	0.122	0.109	0.101	0.098	0.115
		25	T1	0.091	0.093	0.096	0.125	0.085	0.097	0.090	0.096
			Fp	0.093	0.095	0.099	0.125	0.086	0.100	0.095	0.097
λ	SDR	0	T1	1.334	1.334	1.334	1.334	1.334	1.334	1.334	1.334
			Fp	1.360	1.360	1.360	1.360	1.360	1.360	1.360	1.360
		2	T1	1.395	1.408	1.385	1.299	1.394	1.388	1.381	1.373
			Fp	1.418	1.424	1.385	1.320	1.404	1.407	1.406	1.395
		10	T1	1.361	1.337	1.329	1.294	1.328	1.329	1.354	1.275
			Fp	1.355	1.362	1.339	1.302	1.323	1.345	1.352	1.289
		25	T1	1.360	1.420	1.410	1.216	1.407	1.389	1.384	1.204
			Fp	1.364	1.431	1.453	1.197	1.414	1.446	1.439	1.186
	PCA	2	T1	1.359	1.362	1.359	1.414	1.354	1.354	1.346	1.424
			Fp	1.370	1.375	1.371	1.437	1.378	1.368	1.374	1.444
		10	T1	1.387	1.378	1.399	1.434	1.374	1.328	1.320	1.416
			Fp	1.390	1.388	1.406	1.406	1.378	1.345	1.341	1.427
		25	T1	1.397	1.432	1.409	1.529	1.261	1.446	1.398	1.313
			Fp	1.425	1.451	1.421	1.557	1.306	1.444	1.393	1.315

4.3.5 Why not only RVs

Based on our association testing results, in most situations, a few top PCs based on CVs performed better than those based on RVs in SDR for controlling population stratification. This might be surprising given that RVs are expected to be more recent and population-specific. Babron et al. (2012) [56] and Moore et al. (2012) [61] both pointed out that RVs were more likely to cluster in a few subpopulations than to be distributed uniformly across all subpopulations.

As there were 457 samples in total, we could observe at most 9 copies of the minor allele for any RV, meaning that the minor allele of the RV could appear in

at most 9 subgroups across all the samples. 705 RVs were singletons, which were population-specific but could also have resulted from sequencing errors. Table 4.10 shows the distribution of the minor alleles of all RVs across the subgroups from all the 457 samples. The proportion of the RVs whose minor alleles were present in varying numbers of the subgroups are reported. We see only a small proportion of the RVs with their minor alleles appearing in 1 or 2 subgroups. For example, for the RVs with 6 copies of the minor allele, only 7.8% of them appeared in only 1 or 2 subgroups while the rest appeared in 3-6 subgroups. Albeit inconclusive, these results might suggest that most of the RVs were not population-specific. Next, we applied Fisher's exact test (as implemented in R function **fisher.test**) to test whether the minor allele of each RV was randomly distributed across (or independent of) the subgroups. The results are given in Table 4.10; in total, only 25.62% of 68,434 RVs were significant with their minor alleles distributed subgroup-specifically.

As a comparison, we also tested the random distribution across the 10 subgroups of the minor alleles of of CVs or of LFVs in chromosome 1 (with R function **prop.test**). In total, there were 82.53% of 478,208 CVs shown to be significant (p-value$<$0.05) while 74.28% of 146,350 LFVs significant. After pruning CVs and LFVs respectively, there were 79.25% out of 12,269 pruned CVs significant and 72.40% out of 32,080 LFVs significant. Although Fisher's exact test could be conservative, these results lent some support for our observation that PCs of CVs were more capable of adjusting for population stratification than those of RVs. In summary, we found that more CVs and LFVs were subgroup-specific, which, in combination with a RV's less informativeness due to its extremely low MAF, offers a preliminary explanation to better performance of CVs when used to construct PCs to adjust for population stratification than that of RVs.

Table 4.10: Distribution of the RVs with minor alleles present in a given number of the subgroups. In the first 9 rows, the number in cell (i, j) is the proportion of the RVs each with j copies of its minor allele and present in i subgroups; the last two rows give the total number of RVs and the proportion of significant ones by Fisher's exact test.

	# of the minor allele across the 457 samples								
# of subgroups	1	2	3	4	5	6	7	8	9
1	1.0000	0.0225	0.0116	0.0056	0.0026	0.0021	0.0010	0.0043	0.0004
2		0.9775	0.4442	0.2426	0.1399	0.0764	0.0539	0.0368	0.0276
3			0.5441	0.5197	0.4478	0.3801	0.2962	0.2527	0.2161
4				0.2321	0.3430	0.3972	0.4343	0.4031	0.3938
5					0.0667	0.1293	0.1751	0.2290	0.2435
6						0.0149	0.0362	0.0635	0.0942
7							0.0033	0.0103	0.0230
8								0.0003	0.0013
9									0.0000
# of RVs	705	9653	9972	9234	8785	8258	7792	7197	6838
prop. of sig. RVs	0.000	0.111	0.163	0.222	0.230	0.290	0.314	0.379	0.442

4.4 An example using the GAW18 data

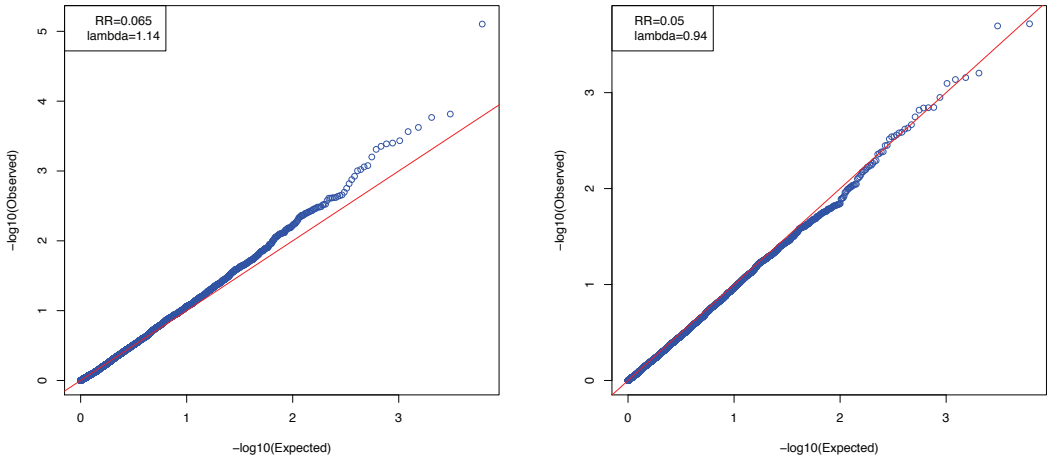
We used the methods discussed above to analyze the Genetic Analysis Workshop (GAW)18 data with a pedigree structure, which can be regarded as a population structure in a fine scale. The dataset was introduced in the Data section. We used the measurements of systolic blood pressure at time point 1, SBP_1, and the hypertension diagnosis at time point 1, HTN_1. The former one is a quantitative trait and the latter one is a binary trait. There are 855 samples available. Gender, smoking and age are the covariates.

We carried out single SNP analysis on a set of 6228 CVs randomly selected from all pruned CVs. According to the findings of previous GWASs that the majority of the SNPs are not significantly associated with hypertension, we can assume these 6228 CVs are null SNPs. Since some subjects are from the same families and thus correlated, we expect to observe an inflated Type I error if we treat the samples as independent. If PCs are effective in adjustment, the p-values should follow a uniform distribution. This also means that the proportion of the tests with p-value$<$0.05 should be close to 0.05 and the inflation factor λ close to 1. We did not only compare CVs and RVs for constructing PCs, but also used two different similarity matrices – the covariance matrix and Identity-By-State (IBS) matrix. The IBS

matrix characterizes the genetic correlation between two subjects as the proportion of loci having the same allele. We extracted 10837 pruned CVs and 11103 pruned RVs for constructing the similarity matrices and all PCs were constructed by PCA.

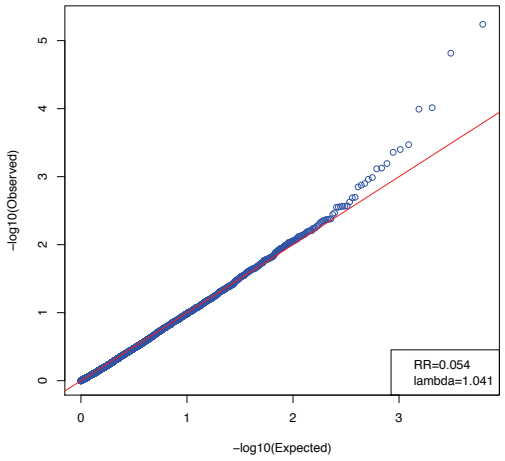
Figure 4.8 shows that without adjustment, for SBP_1 the observed p-values deviate from the theoretical uniform distribution (λ=1.14). Yet, the p-values of HTN_1 seem to follow the uniform distribution ($\lambda = 0.94$). This observation indicates the heritability may be mild for HTN_1 in the GAW18 dataset. From Figure 4.9 we can see that PCs based on CVs can better control the inflation than those based on RVs. For example, when using PCs of the covariance matrix based on CVs (Panel (b)), λ is reduced to 1.068, smaller than that with PCs based on RVs (Panel (d)). For HTN_1, we observe a similar result, although the contrast is not so strong (Figure 4.10). This analysis result is consistent with our conclusion from simulation studies that PCs based on CVs are more effective than those based RVs for adjusting for population structure.

Figure 4.8: Q-Q plots of p-values without considering the correlation among samples. RR stands for rejection rate.

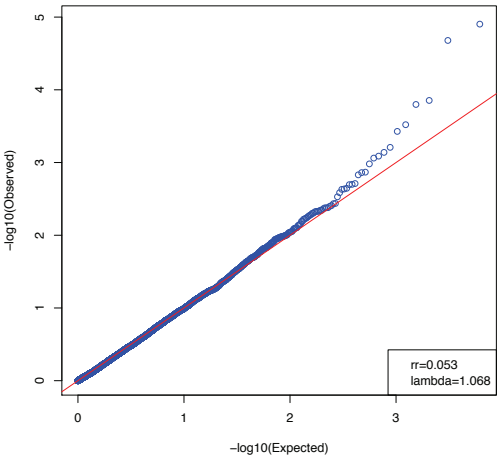


(a) SBP_1 (b) HTN_1

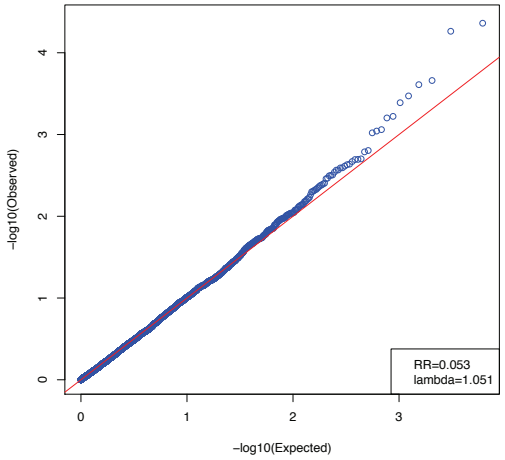
Figure 4.9: Q-Q plots of p-values for analyzing SBP_1 with adjustment of PCs of (a) the IBS matrix based on CVs (b) the covariance matrix based on CVs (c) the IBS matrix based on RVs (d) the covariance matrix based on RVs. RR stands for rejection rate.



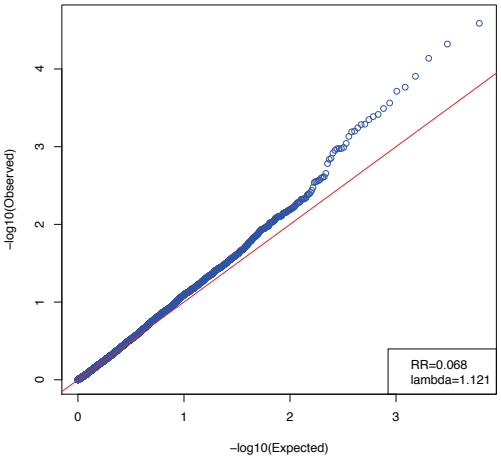
(a)



(b)



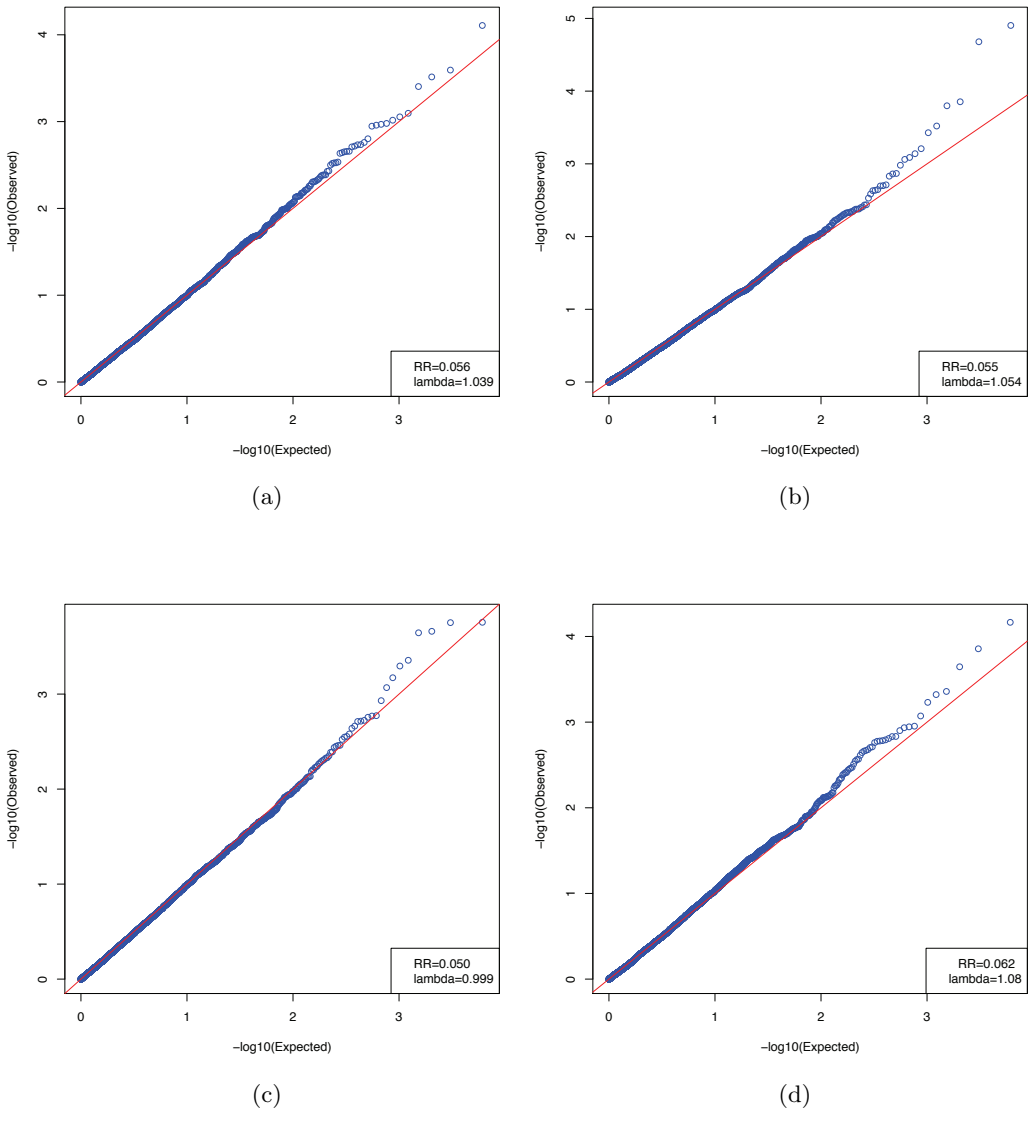
(c)



(d)

Figure 4.10: Q-Q plots of p-values for analyzing HTN_1 with adjustment of PCs of (a) the IBS matrix based on CVs (b) the covariance matrix based on CVs (c) the IBS matrix based on RVs (d) the covariance matrix based on RVs. RR stands for rejection rate.

Figure 4.11: Q-Q plots of p-values with adjustment of PCs of (a)



4.5 Conclusions and Discussions

As a continuation of our previous study on adjusting for population stratification at the continental group level, we used the 1000 Genomes Project data to study the issue in a fine scale. We were interested in investigating what types or sets of variants and what dimension reduction method, SDR or PCA, could best find the axes of genetic variations among multiple continental subgroups. We first used the Fst statistic and a few top PCs to study population structure. We also confirmed that even the genetically similar subgroups, like CEU and GBR samples, could still induce population stratification in association testing. We have also observed that, with or without the AMR samples from admixed populations, the top PCs barely changed. Our observations in these exploratory analyses were later confirmed in clustering analysis and association testing. The main conclusions are the following.

First, in association testing of CVs, RVs or LFVs, we found that a few top PCs based on all variants, all CVs or all pruned CVs constructed by SDR could consistently control the Type I error around the nominal 0.05 level, and reduce the inflation factor, λ, to around 1, in most situations. The difficulty in perfectly uncovering ethnic subgroups did not appear to notably impair the effectiveness of a few top PCs in adjusting for population stratification. Second, in the presence of a local non-genetic risk (i.e. spatially structured populations), while it was rather easy to control inflations when testing CVs, it was much harder when testing RVs. For RVs, using a few top PCs of SDR based on all variants or all CVs could be largely, but not perfectly, effective; a few top PCs of SDR, but not of PCA, based on all RVs also showed good performance in controlling inflations.

Third, our study confirmed that SDR was more robust to outliers and noises, producing PCs more informative for subgroups than PCA. This led to better performance in both association testing and clustering. The difference was the largest and thus most clearly seen when RVs were used to construct PCs. In particular, it is noted that when testing RVs or LFVs, none of the top PCs of PCA were sufficient to control inflated false positives. Lastly, we observed that there was a larger proportion of CVs or of LFVs significantly distributed non-randomly across the subgroups than that of RVs, offering an explanation of why we witnessed better adjustment performance of the PCs based on CVs than that on RVs.

There are some limitations in our study. Firstly, although the dataset we used is one of the largest sequencing datasets available nowadays with multiple population subgroups, the sample size is still relatively small, especially at the subgroup level. As a consequence, we only considered some more extreme simulation set-ups with each whole subgroup belonging to one of the case and control groups. In fact, we could not randomly assign any subgroup to be cases or controls because there might be a complete separation between cases and controls. For example, in the simplest case, if we had all the EUR subgroups as the cases and all AFRs as the controls, then the second PC of SDR could perfectly separate the cases from the controls, leading to non-existence of the MLE and a convergence problem when fitting the logistic regression model. One remedy was, as we did here, to assign inseparable subgroups as cases and controls respectively. Another consequence of the small sample size is possibly inflated Type I error rates (and inflation factor) when an increasing number of PCs were included in a logistic regression model. This phenomenon was also shown in Bouaziz et al., 2011 [62] : when the authors included 20 PCs in the logistic regression to adjust for population stratification, the Type I error was around the 5% level; however, when 5 more PCs were added, the estimated Type I error increased to the upper bound of the 95% confidence interval. More generally, this important question is related to that of how to determine the number of the top PCs to use for adjusting for population stratification. Secondly, in this study we considered association testing on only single CVs or only multiple RVs (or LFVs). It would be of interest to check how population stratification could be controlled when testing both multiple CVs and multiple RVs at the same time. Finally, it is also most important to assess statistical power when adjusting for population stratification. As shown by Zhang et al. (2012) and others, there could be power loss with an adjustment by a few top PCs of PCA. One would wonder to what extent this could happen when using a few top PCs of SDR with all variants, all CVs or all RVs.

In sum, based on our study using the 1000 Genomes project data, using PCs based on all variants and all CVs constructed by SDR in PCR is generally an effect way to control population stratification in a fine scale. However, the over-adjustment problem should always be alerted.

Chapter 5

Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements?

5.1 Introduction

In the last two chapters, we inspected the performance of one popular method, the PCR, for adjusting for population stratification in association testing of CVs, LFVs and RVs. With PCR, the top few principal components (PCs) are included as covariates in regression [63, 25]. We showed that population stratification could damage the association testing with a dramatically inflated Type I error. While it was easy to use PCs of 10000 randomly selected common variants (CVs) with minor allele frequencies (MAFs) ≥ 0.05 or low frequency variants (LFVs) with $0.01 \leq \text{MAF} < 0.05$ to handle population stratification from continental groups, it was more difficult to resolve this problem with subgroups. We compared two dimension reduction methods, principal component analysis (PCA) and spectral dimension reduction (SDR), and different variants for constructing PCs to adjust for population stratification in a fine scale. In general, PCs of all variants and all CVs constructed by SDR could largely control inflation.

In practice, genetic correlations among subjects can arise from either population heterogeneity, familial relatedness, or cryptic relatedness, all of which can be regarded as population structure. Furthermore, due to the observational nature of GWAS, unknown environmental and non-genetic risk factors may arise as confounders. Failure to account for these correlations and confounders can produce both false positives and false negatives in GWAS.

When population structure acts like a confounder in GWAS, it is also called population stratification. Population stratification occurs most frequently in the case-control study design, where different ancestral populations have varying disease risks and different distributions of genetic variants. Many methods have been proposed, such as genomic control (GC) [23], structured association [26] and genetic matching and stratification [64, 27]. One of the most appealing method, and also the one we discussed about in last two chapters, is principal component regression (PCR) based on principal component analysis (PCA) of a large number of genetic variants across the genome [63]. It includes a few top principal components (PCs) as covariates in a regression model, and has proven competent. The top PCs also offer a way to visualize the spatial locations of subjects, though this may be over-interpreted [65, 66]. To account for more complex population structures, including familial

correlations or cryptic relatedness, linear mixed models (LMMs) have emerged recently as most promising [29, 30]. To overcome the computing bottleneck, several fast algorithms have been developed for LMM, including efficient mixed-model association (EMMA) [67] and its expedited version EMMAX [32], and genome-wide efficient mixed-model association (GEMMA) [33]. EMMA and GEMMA offer exact calculations, while EMMAX is an approximate method in estimating the variance components by ignoring the covariate (i.e. genetic variant) effect of interest. Kang et al. (2010), by using the 1966 Northern Finland Birth Cohort (NFBC66) and the Wellcome Trust Case Control Consortium (WTCCC), showed that EMMAX can better control inflated false positives than PCR in GWAS. On the other hand, Wu et al. (2011) [25] reported some simulated data showing that EMMAX could be "anticonservative" while PCR seemed to perform best. Wang and Peng (2012) [68] offered some theoretical and numerical properties of the two methods.

Given the popularity of PCR and the emerging promise of LMM, in view of these discrepant comparisons between the two methods, it is natural to ask which one is preferred in practice. To address this important question, we aim to point out their connections and differences. First, based on probabilistic PCA (Tipping and Bishop, 1999 [39]), we show that the PCR method can be regarded as an approximation to a LMM; the degree of the approximation depends on the number of the top PCs used, the choice of which is a difficult question in practice. This connection offers a *theoretical* explanation on why LMM performed better than PCR in several real studies. This may give an impression that LMM can completely replace PCR, which however is not true. Due to the use of the fixed effects in PCR and random effects in LMM, there are other implications from model fitting. First, as is well known, when a larger number of PCs are used in PCR to approximate better a LMM, due to an increasing number of parameters to be estimated, the PCR method will lose power. Second, more surprisingly, in the presence of unknown environmental (and non-genetic) confounders, PCR may outperform LMM. The reason is that, because PCs can represent geographical locations of human populations, use of the PCs may be able to adjust for environmental confounders that are spatially distributed. Hence, to account for both population stratification and unknown environmental confounders, we propose a hybrid method combining PCR and LMM. We use a real genotype dataset to confirm the above points.

5.2 Methods

Suppose $Y = (Y_1, ..., Y_n)^T$ is the quantitative trait vector for n subjects, and $g^* = (g_1^*, g_2^*, ..., g_n^*)^T$ is the genotype score vector of a single nucleotide polymorphism (SNP) of interest, where g_i^*=0, 1, 2 is the minor allele count for the i^{th} subject. We have $g = (g_1, ..., g_n)$ as the normalized genetic scores with $g_i = (g_i^* - 2p_0)/\sqrt{p_0(1-p_0)}$, where p_0 is the minor allele frequency (MAF) of the SNP.

5.2.1 LMM and PCR methods

A linear mixed model (LMM) accounting for population structure is

$$Y = \beta_0 \mathbf{1} + g\beta_1 + u + \epsilon, \tag{5.1}$$

where β_0 is the intercept, $\mathbf{1}$ is a vector of all 1's, $u \sim N(0, \sigma_g^2 K)$ is the so-called polygenic effect, K is a similarity matrix measuring the similarity or relatedness between any two subjects, and $\epsilon \sim N(0, \sigma^2 I)$ is the error term. σ_g^2 is the polygenic variance and σ^2 is the individual variance. The marginal covariance of Y is $\text{var}(Y) = \Omega = \sigma_g^2 K + \sigma^2 I$. The goal of an association analysis is to test the null hypothesis H_0: $\beta_1 = 0$.

The PCR model is

$$Y = \gamma_0 \mathbf{1} + g\gamma_1 + Z\gamma_2 + \epsilon \tag{5.2}$$

where γ_0 is the intercept, and Z is the matrix with each column as one of a few top PCs constructed by PCA from a large number of genetic variants, or more generally, as a few top eigen vectors of a similarity matrix measuring similarities among the subjects based on the genetic variants [45]. $\epsilon \sim N(0, \sigma^2 I)$ is the error term. The goal of an association analysis is to test the null hypothesis H_0: $\gamma_1 = 0$.

5.2.2 A connection between PCR and LMM

In the LMM (5.1), we can regard the polygenic effect u as a collapsed effect of many genetic variants, say X^* with p genetic variants. $X = (X_{ij})$ is the matrix after normalizing X^*: for each SNP j of subject i, $X_{ij} = (X_{ij}^* - 2p_j)/\sqrt{p_j(1-p_j)}$

with p_j as the MAF of SNP j. Then the LMM can be written as

$$Y = \beta_0 \mathbf{1} + g\beta_1 + \sum_{j=1}^{p} X_{\cdot j}\eta_j + \delta = \beta_0 \mathbf{1} + g\beta_1 + X\eta + \delta,$$

where $X_{\cdot j} = (X_{1j}, ..., X_{nj})'$, $\eta \sim N(0, \sigma_g^2 I)$ and $\delta \sim N(0, \sigma^2 I)$. Note that in the LMM, $K = XX^T/p$, the covariance matrix, can be used to measure the similarities among the n subjects. In probabilistic PCA [39], similar to factor analysis, each $X_{\cdot j}$ is modeled iid as

$$X_{\cdot j}|\zeta_j \sim N(W\zeta_j + \zeta_{\mathbf{0}}, \sigma_x^2 I),$$

where it is assumed that $\zeta_j \sim N(0, I)$ and σ_x^2 is the variance "lost" after dimension reduction in PCA. Since each SNP is already centered at 0, we take $\zeta_{\mathbf{0}} = 0$. With any chosen dimension $n \times q$ for W, the maximum likelihood estimators are

$$\hat{W} = U_q(\Lambda_q - \sigma_x^2 I)^{1/2} R, \text{ and } \hat{\sigma}_x^2 = \frac{1}{n-q}\sum_{j=q+1}^{n} \lambda_j,$$

where U_q is a matrix with columns as the top q eigenvectors of the similarity or sample covariance matrix $K = XX^T/p$, λ_j's are eigenvalues of K, and R is an arbitrary orthogonal rotation matrix. In other words, $\hat{W}$ contains the top q (scaled and rotated) PCs based on X; the scalings and rotations of the PCs have no effect in regression, and can be ignored therein. Taking $\zeta = (\zeta_1, ..., \zeta_p)$ and ϵ_x as the corresponding matrix for the error term in the probabilistic PCA model, we approximate the LMM as

$$Y = \beta_0 \mathbf{1} + g\beta_1 + (\hat{W}\zeta + \epsilon_x)\eta + \delta = \gamma_0 \mathbf{1} + g\beta_1 + \hat{W}\gamma_2 + \epsilon,$$

where $\gamma_0 = \beta_0$, $\gamma_2 = \zeta\eta$ and $\epsilon = \epsilon_x\eta + \delta$. If q is the number of the top PCs that we use in PCR, $\hat{W}$ is Z in Equation (5.2). Hence the above approximate LMM reduces to the PCR model in Equation (5.2). Note however that in the PCR model γ_2 is treated as a fixed (i.e. non-random) effect, while in the LMM u (or γ_2) is random.

The above derivation is for $K = XX^T/p$. For other K, e.g. calculated as the IBS matrix, due to its positive semi-definiteness as (a part of) the covariance matrix for the random effect u, we can have a decomposition $K = AA^T/p_A$, where A is

a $n \times p_A$ matrix. Denote the jth column of A as $A_{\cdot j}$. Now replace the $X_{\cdot j}$ by $A_{\cdot j}$ and then proceed as before, e.g. by assuming $A_{\cdot j}|\zeta_j \sim N(W\zeta_j + \zeta_{\mathbf{0}}, \sigma_x^2 I)$ and $\zeta_j \sim N(0, I)$, we can reach the same PCR model as an approximation to the LMM, where the PCs are generalized to the eigenvectors of any symmetric and positive semi-definite similarity matrix K. Hence our above conclusion holds for any positive semi-definite similarity matrix K.

5.2.3 An environmental confounder

In observational studies like GWAS, unobserved environmental and non-genetic factors may arise as confounders, which may not be fully captured by the similarity matrix K estimated from genetic variants [38]. A model with both a sample structure u and an environmental confounder μ is

$$Y = \beta_0\mathbf{1} + \mu + g\beta_1 + u + \delta = \beta_0\mathbf{1} + D\theta + g\beta_1 + u + \delta, \tag{5.3}$$

where $\mu = (\theta_1, ...\theta_1, ...\theta_k, ...\theta_k)^T$, $D = \text{diag}\{(\mathbf{1_{n_1}}^T, \mathbf{1_{n_2}}^T, ...\mathbf{1_{n_k}}^T)^T\}$, $\theta = (\theta_1, \theta_2, \cdots, \theta_k)^T$, $\mathbf{1_{n_j}}$ is a vector of all 1's of length n_j, the number of samples in cluster j, and $\text{diag}\{a\}$ is a matrix with vector a on the diagonal and all other elements 0. Here we assume that the samples are ordered into clusters with each cluster containing the samples sharing the same environmental risk; this assumption is not necessary, but only for simplicity and concreteness for the purpose of presentation.

Now suppose $\theta_h \sim f(.)$, $h = 1, ..., k$, where $f(.)$ is the unknown distribution density of θ_h with variance σ_θ^2. Then

$$\text{var}(Y) \quad = \quad DD^T\sigma_\theta^2 + \sigma_g^2 K + \sigma^2 I = \sigma_g^2 \left(\frac{\sigma_\theta^2}{\sigma_g^2} DD^T + K + \frac{\sigma^2}{\sigma_g^2} I \right),$$

and $DD^T = \text{diag}(\mathbf{11^T_{n_1}}, \mathbf{11^T_{n_2}}, ...\mathbf{11^T_{n_k}})$. One potential issue with EMMAX or GEMMA is that their only using K to model the covariance among samples. Due to the commonality of the human genomes, the K matrix has a more "smooth" structure that may not approximate well a block diagonal matrix like DD^T (or other more general matrix induced by environmental confounders). Consequently, with a relatively large σ_θ, using K alone may fail to capture the phenotype covariance structure, leading to a lack of fit of the standard LMM (5.1).

On the other hand, if μ can be well approximated by a linear combination of the top PCs, say $\mu \approx Z\rho$, then the PCR model is approximated by

$$Y = \gamma_0 \mathbf{1} + Z\rho + Z\gamma_2 + g\beta_1 + \epsilon = \gamma_0 \mathbf{1} + Z(\rho + \gamma_2) + g\beta_1 + \epsilon,$$

which can be well fitted by the standard PCR model (5.2). In practice, this assumption on $\mu \approx Z\rho$ may be plausible if environmental confounders are spatially distributed, because the top PCs of genetic variants can represent geographic coordinates (Wang et al., 2012).

5.2.4 A hybrid model

As discussed, neither the (standard) LMM nor PCR is a complete winner in adjusting for both population structure and environmental confounders, but with different advantages. Hence we propose a hybrid model including both a few PCs and a random effect:

$$y \quad = \quad \gamma_0 \mathbf{1} + g\beta + Z\rho + u + \delta, \tag{5.4}$$

where Z is the top q PCs from the similarity matrix K, and $u \sim N(0, \sigma_g^2 K)$, $\delta \sim N(0, \sigma^2 I)$. The PCs aim to capture a major part of population structure and environmental confounders, while the random effect account for the remaining and more subtle effects of population structure.

Since the PCs are extracted from the similarity matrix K, there may be some concerns on the repeated use of K for both the PCs and random effect. As an alternative, we can also use K_2 as the similarity matrix for u, where K_2 is a "residual" matrix of K after excluding the covariance explained by the top q PCs in Z. That is, by eigen decomposition we have

$$K = Q\Lambda Q^T = (Q_1,\ Q_2)\mathrm{diag}\{\Lambda_1,\ \Lambda_2\}(Q_1,\ Q_2)^T,$$

where Q is the eigenvectors and Λ is the eigenvalues, Q_1 is the top q PCs that we include in Z, Λ_1 is the corresponding eigenvalues of the top PCs. Q_2 is the remaining eigenvectors and Λ_2 is the corresponding eigenvalues. We define $K_2 = Q_2\Lambda_2 Q_2^T$.

It turns out that the restricted likelihood of these two hybrid models are exactly the same (see the proof in Appendix B), so will be their estimation and inference. Therefore, in the hybrid model the random effect (either u or u_2) is only used to capture the "residual" effects of the PCs.

The hybrid model has been used to account for population structure alone [30]. Here we use it in the presence of both population structure and environmental confounders. In particular, as a main contribution here, we explicitly separate out the effects of population structure and environmental confounders, based on which we can better illuminate the respective advantages (and disadvantages) of the PCR and LMM approaches.

5.3 Simulations

We use the Genetic Analysis Workshop (GAW) 18 data with a familial structure to illustrate our points. We first pruned all the common variants (CVs) by PLINK [42] with a sliding window of size 50, a moving step 5 and $r^2 \leq 0.05$. We randomly selected 31544 pruned CVs with MAF $\geq$0.05 from all autosomes, and use them to estimate the similarity matrix K; both the covariance matrix and IBS matrix were calculated, however, due to the increasing popularity of the IBS matrix in LMM for its better performance [67], we show the results with the IBS matrix.

We simulated quantitative traits following the sample structure shown in estimated IBS matrix from the data, with or without an environmental risk, under both null and alternative hypotheses. We compared the LMM, implemented in EMMAX [67] and GEMMA [33], PCR [24] and the hybrid method, with respect to their ability to correct for inflated type I errors as well as their power performance. We applied the F-test in EMMAX, and the Wald test in GEMMA, the PCR model and hybrid model. As a benchmark, we also applied the ordinary least squares (OLS) (i.e. assuming $K = I$ in the LMM) with the t-test.

We have also carried out similar simulations using the 1000 Genomes Project data, for comparing the PCR and LMM in the presence of simpler population structure – Europeans and Africans, with or without an environmental factor. The results are similar to those using the GAW18 data and are shown in Appendix C.

5.3.1 In the presence of only population structure

For the purpose of controlling Type I error and the inflation factor λ (Devlin and Roeder 2004), we used model (5.1) with $\beta_1 = 0$ to simulate Y's under the null hypothesis; to compare the power, we used model (1) with a genetic effect $\alpha \neq 0$. We set $\beta_0 = 5$, and the vector u was sampled from $N(0, \sigma_g^2 K)$, $\epsilon \sim N(0, \sigma^2 I)$. σ_g^2 was the polygenic variance, and K was the IBS matrix estimated from the GAW18 data. We randomly selected 11133 pruned CVs to be tested. For the PCR and hybrid methods, we included the top 20 PCs unless specified otherwise. For all the tests the nominal significance level was 0.05.

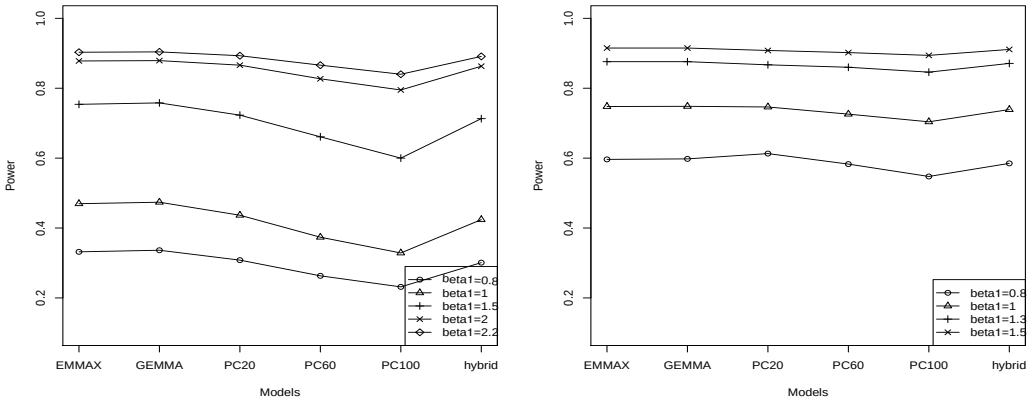
Under the null hypothesis (Table 5.1), as σ_g^2 increased, PCR gradually failed to control the Type I error rate and λ while EMMAX and GEMMA behaved well. For example, for $\sigma_g^2 = 90$ and σ^2=10, PCR had a severely inflated Type I error rate of 0.110 (and inflated $\lambda = 1.493 > 1$), while EMMAX and GEMMA had their type I errors around 0.05 (and $\lambda \approx 1$). If we gradually increased the number of PCs for the scenario $\sigma_g^2 = 90, \sigma^2 = 10$, both the Type I error and λ were reduced notably; however, even with the top 100 PCs used, the Type I error was still around 0.070 and λ around 1.18.

Figure C.3 shows the power comparison with different genetic effect β_1. GEMMA usually had the highest power, but as σ_g^2 increased, the power difference between GEMMA, PCR and the hybrid model decreased. The power of the hybrid method was slightly lower than EMMAX and GEMMA, and was close to that of PCR with 20 PCs. For example, when $\sigma_g^2 = 10$ with $\beta_1 = 1.5$, the power was 0.754 for EMMAX, 0.758 for GEMMA, 0.723 for PCR with 20 PCs and 0.600 with 100 PCs, and 0.713 for the hybrid method; for $\sigma_g^2 = 90$, the power for the methods was 0.915, 0.915, 0.908, 0.894 and 0.911, respectively.

Table 5.1: Association testing under H_0 in the presence of only population structure.

set-up		EMMAX	GEMMA	PCR(20)	hybrid	OLS
$\sigma_g^2 = 10$, $\sigma^2 = 90$	Type I error	0.050	0.051	0.053	0.050	0.055
	λ	0.992	1.004	1.025	0.980	1.025
$\sigma_g^2 = 60$, $\sigma^2 = 40$	Type I error	0.051	0.052	0.073	0.053	0.095
	λ	1.041	1.044	1.185	1.029	1.363
$\sigma_g^2 = 90$, $\sigma^2 = 10$	Type I error	0.050	0.051	0.109	0.050	0.153
	λ	1.000	1.003	1.465	1.011	1.828

Figure 5.1: Power of the association tests based on a simulated trait and the GAW 18 genotype data.



(a) $\sigma_g^2 = 10$, $\sigma^2 = 90$

(b) $\sigma_g^2 = 90$, $\sigma^2 = 10$

5.3.2 In the presence of an environmental confounder

We considered a scenario with both population structure and an environmental confounder. We assumed that 496 samples in 8 families were from the same spatial area thus sharing the same environmental risk while the remaining in another cluster. For illustration, we selected 10000 SNPs that were significantly associated with the clustering assignment and were to be tested in association analysis. We saw that EMMAX and GEMMA had gradually increasing Type I errors as the environmental effect $|\theta_j|$ became larger, due to the inadequacy of the genetic similarity matrix to capture the environmental confounder. The hybrid method was consistently the best performer. For example, when $|\theta_j| = 4$, while GEMMA had a Type I error rate of 0.109 (λ=1.645), PCR of 0.067 (λ=1.151) with 100 PCs, the hybrid method with 20 PCs could reduce it to 0.061 (λ=1.158), and further if more PCs were used. Here PCR also worked fine with 100 PCs mainly because σ_g^2 was not so big; when we increased σ_g^2 to 90, PCR lost its efficacy even with 100 PCs (Type I error =0.081 and λ=1.269) while the hybrid method could still control the Type I error rate to be 0.0618 (and λ=1.102) with only top 20 PCs.

Table 5.2: Association testing with a sample structure and environmental factor, and $\sigma_g^2 = 60$, $\sigma^2 = 40$ for the GAW18 data. 959 samples are artificially assigned into 2 clusters.

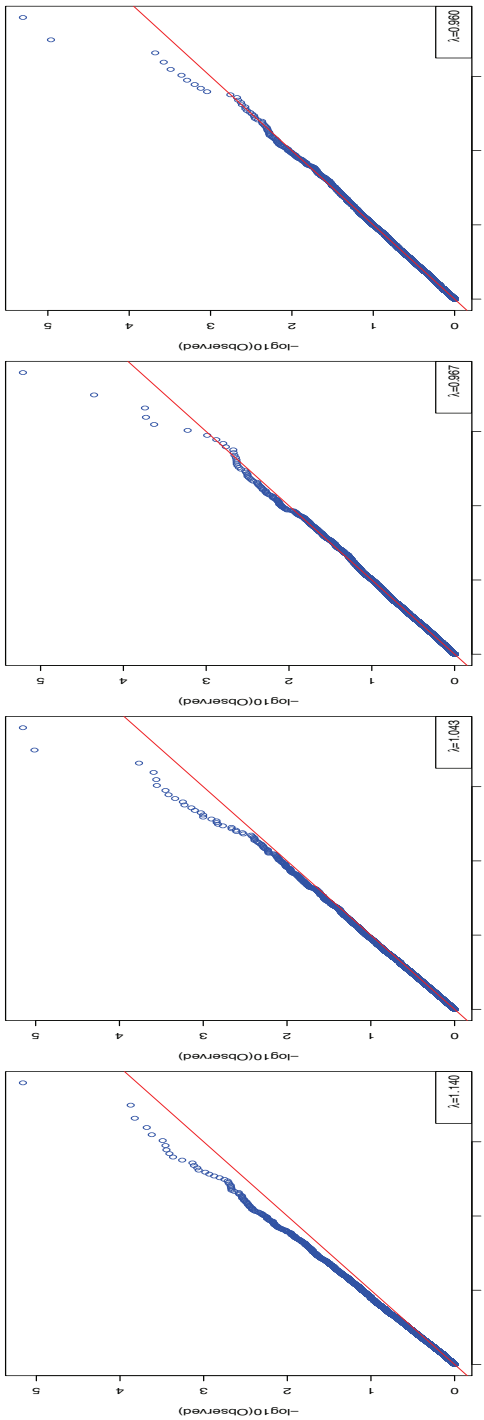
		EMMAX	GEMMA	PCR					hybrid	OLS
				20	40	60	80	100		
$\theta_1 = -1, \theta_2 = 1$	Type I	0.057	0.059	0.081	0.067	0.065	0.061	0.059	0.052	0.152
	λ	1.133	1.146	1.302	1.218	1.152	1.125	1.097	1.076	1.947
$\theta_1 = -2, \theta_2 = 2$	Type I	0.073	0.075	0.092	0.070	0.068	0.062	0.061	0.053	0.288
	λ	1.279	1.299	1.381	1.237	1.175	1.139	1.113	1.093	3.823
$\theta_1 = -4, \theta_2 = 4$	Type I	0.107	0.109	0.136	0.083	0.077	0.070	0.067	0.061[a]	0.626
	λ	1.629	1.645	1.769	1.334	1.275	1.208	1.151	1.158[a]	11.976

[a]For PC#=40, 60, 80, 100, Type I errors=0.056, 0.055, 0.053, 0.051 and λ=1.111, 1.060, 1.062, 1.063 respectively.

5.4 Example

We conducted an association analysis with the systolic blood pressure (SBP) at the baseline as the quantitative phenotype in the GAW 18 data. The genotype data used were the same as in simulations. In particular, we used the IBS matrix as the similarity matrix K; the use of the covariance matrix as K performed worse (not shown). All the methods included subject gender, smoking status and age as covariates. As shown in Figure 5.2, no adjustment for population structure (i.e. OLS) led to severely inflated false positives with an inflation factor λ of 1.14, since it failed to correct for within-family correlations in the data. In contrast, PCR with the top 20 PCs could largely control the inflated false positives, while the GEMMA implementation of LMM had a slight advantage over PCR with a λ closer to 1; furthermore, there were fewer more significant p-values (< 0.001) resulting from the LMM than those from PCR. The good performance of LMM also suggested the non-existence or negligible effects of environmental confounders (that could not be adjusted by genetic similarity matrix K), possibly due to all study subjects were from the San Antonio area and thus there was a lack of environmental heterogeneity. It is reassuring to see that the hybrid method controlled the false positives as well as LMM, and at the same time, gave a few more p-values < 0.001.

Figure 5.2: Q-Q plots of the p-values in the association tests for the SBP in the GAW 18 data.



(a) Unadjusted (b) PCR (c) LMM (d) Hybrid

5.5 Discussion

We have discussed a close connection between PCR and LMM in association analysis of structured samples. This connection suggests both theoretical and practical advantages of the LMM method over PCR. In particular, the choice of how many PCs to use in PCR is difficult; too few PCs may cause inflated Type I errors while too many lead to power loss. It appears increasingly acceptable to take the LMM as a general model for population structure, from which the connection between the two methods also offers an explanation on why PCR often performs well if the population structure is not sufficiently complex or subtle. For example, when we used the European and African samples from the 1000 Genomes project data, PCR with 20 PCs performed as well as the LMM method (Appendix C). A challenge however is how to tell whether a PCR model is adequate or not when compared to a LMM. In this regard, it seems that one should always use LMM over PCR. However, we have also pointed out a weakness of the LMM method in the presence of unknown environmental confounders that can arise from GWAS. Accordingly we have proposed a hybrid method, which performed consistently well across all scenarios in our study. In particular, the hybrid method can be easily implemented in any existing framework of fitting LMMs, such as in the EMMAX or GEMMA package. Therefore, we recommend the use of the hybrid method.

Chapter 6

Multivariate Trait Analysis with a New Adaptive Sum of Powered Score Test in Generalized Estimation Equations

6.1 Introduction

In previous three chapters, we endeavored to study how to model population structures in genetic association analysis. Being aware of the fact that it is almost inevitable for subjects to be genetically related, population stratification has become one of the major concerns in genome-wide association studies. Insufficient attention to this problem can cause spurious findings. We examined how the PCR model [63] performed in the presence of population stratification when testing CVs or RVs. In particular, we studied what types and set of variants and what dimension reduction method could extract more effective principal components (PCs) to represent (sub)groups and control population stratification in a fine scale. By relating the PCR with LMM, we explained why the LMM was a better performer for addressing population structure alone, and proposed a hybrid method targeting both genetic relatedness and environmental confounders.

Now we will pursue the other side of association testing – power boosting. In the study of complex diseases, multiple correlated phenotypes are usually collected. For example, in the AIDS study, outcomes like mortality, serious AIDS symptoms, CD4+ cell counts, HIV viral load, etc., are usually recorded to measure disease progression. Most genome-wide association studies (GWASs) only analyze each phenotype separately and find it underpowered to identify many disease markers. Thus, multivariate trait analyses are frequently encountered. On the other hand, there may or may not be some common genetic causes to all the traits under investigation. This fact can affect the performance of the multivariate trait analysis depending on the model assumption.

Various methods have been used for multivariate trait analysis. Broadly, those proposed for pedigree or longitudinal data are also suited. One commonly used method is to combine p-values from univariate trait tests by adjusting for multiple testing, like taking the minimum p-values with the Bonferroni correction. This adjustment is known to be conservative. Another way is dimension reduction on related traits, such as extracting a few top principal components (PCs) to be tested, or canonical correlation analysis (CCA) to seek a linear combination of traits to yield the greatest association [69, 70, 71]. Linear mixed models (LMM) or generalized linear mixed models (GLMM) can be readily applied to multivariate association

studies [72, 73, 74]. However, the likelihood-based approaches can be computationally demanding, and only be able to model phenotypes of the same distribution (e.g. all continuous or binary). After the generalized estimation equations (GEE) methodology is proposed [75], it has been widely adopted for analyzing correlated responses [76, 77, 78].

Most multivariate trait methods are mainly proposed for testing common variants (CVs) with minor allele frequencies (MAFs) greater than 0.05. However, it is believed that low frequency variants (LFVs with $0.01 \leq \text{MAF} < 0.05$) and rare variants (RVs with MAF$<$0.01) also play very important roles in complex diseases [3, 4]. Multivariate trait analysis with RVs can be a challenging yet interesting topic. Due to the sparsity and small effect of each RV, global testing, which is to test a group of RVs instead of single one, are more practical. Among all the tests available for RVs, including pooled association tests [48, 17], the sequencing kernel association test (SKAT) [13], the kernel-based adaptive clustering (KBAC) test [19], the estimated regression coefficient (EREC) test [20], we choose to use the sum of powered score tests (SPU) and the data-adaptive SPU (aSPU) test lately proposed by Pan et al. (2013) [40].

The SPU tests, including a class of tests of different power parameters, are extensions of the SSU test [16], C-alpha test [15] and SKAT [13]. The SPU tests are versatile in the sense that, for a given scenario, at least one SPU test with a certain parameter value performs relatively well. The SPU tests are targeted for global testing, and shows an exceptional performance when many non-associated variants are present in the analysis. This property is especially favorable in our multivariate analysis as we may "incidentally" generate many null variables. This point will be illustrated in Methods section. However, this method is first proposed for independent samples with a single trait. In this chapter, we will extend the SPU tests to the GEE framework with covariates so that it is usable for multivariate trait analysis.

In this chapter, we compared the SPU tests with the score test and UminP test via simulated case-control studies under various scenarios, for both CVs and RVs. P-values were evaluated by simulation-based as well as permutation-based methods. We then applied our method to two real datasets: one is a clinical trial data from the NvR HIV/AIDS study [79], to investigate which HIV viral mutations

are associated with drug resistance in advanced HIV patients; the other one is the Genetic Analysis Workshop (GAW) 18 data with the DNA sequencing data and periodical hypertension measurements.

6.2 Methods

6.2.1 Generalized Estimating Equations

In this chapter, we mainly consider binary traits, although for quantitative traits it follows similarly. Suppose for each subject i, we have k binary traits, y_{i1}, y_{i2}, ..., y_{ik} with y_{im}=1 for disease or $y_{im} = 0$ otherwise for m^{th} trait, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is a row vector for p single nucleotides variants (SNVs) of interest with x_{id} as 0, 1, 2 for count of the minor allele at locus d, $d = 1, ..., p$, and $z_i = (z_{i1}, z_{i2}, ..., z_{iq})$ is a row vector for q covariates. Univariate trait analysis is to regress each trait on x_i and z_i:

$$\begin{aligned} logitPr(y_{i1} = 1) &= a_1 + x_i b_1 + z_i c_1, \\ logitPr(y_{i2} = 1) &= a_2 + x_i b_2 + z_i c_2, \\ &\vdots \\ logitPr(y_{ik} = 1) &= a_k + x_i b_k + z_i c_k, \end{aligned}$$

where $logitPr(y_{im} = 1) = log(Pr(y_{im} = 1)/Pr(y_{im} = 0))$, $i = 1, ..., n$, a_m, b_m, c_m are vectors of coefficients for trait $m = 1, 2, ..., k$. Multivariate trait analysis is to combine and summarize the results of univariate trait analysis by considering the relation of the multiple traits.

The Generalized Estimating Equations (GEE) is a natural way to achieve that. It has been widely used to model correlated data, especially for binary traits. Suppose $Y_i = (y_{i1}, y_{i2}, ...y_{ik})'$ is the vector of phenotypes for each individual i, and we assume that the effect sizes of SNVs and covariates can be different on the marginal mean of

each trait. Thus, we construct for SNVs and covariates two block-diagonal matrices:

$$X_i = \begin{pmatrix} x_i & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & x_i & \cdots & \mathbf{0} \\ \vdots & & & \\ \mathbf{0} & \cdots & \mathbf{0} & x_i \end{pmatrix}, \qquad Z_i = \begin{pmatrix} 1 & z_i & \mathbf{0} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 & z_i & ... & \mathbf{0} \\ & & & \vdots & & \\ \mathbf{0} & \mathbf{0} & ... & \mathbf{0} & 1 & z_i \end{pmatrix},$$

where $\mathbf{0}$ is a row vector of all 0's, X_i is a $k \times kp$ matrix, Z_i is a $k \times k(q+1)$ matrix. Let us define the coefficients as $\beta = (\beta_{11}, ..., \beta_{1p}, ..., \beta_{k1}, ..., \beta_{kp})^{'}$, $\kappa = (\kappa_{11}, ..., \kappa_{1(q+1)}, ..., \kappa_{k1}, ..., \kappa_{k(q+1)})^{'}$. The marginal mean, $E(y_{im}|x_i, z_i) = \mu_{im}$, and covariates are related through:

$$g(\mu_i) = \eta_i = Z_i\kappa + X_i\beta = H_i\theta,$$

where $g(.)$ is a link function, $H_i = (Z_i,\ X_i)$ is a matrix of $k \times k(p+q+1)$, $\theta = (\kappa^{'},\ \beta^{'})^{'}$ is a vector of $k(p+q+1)$ entries, and η_i is a column vector of length k.

For binary traits, we assume y_{ij} is from a binomial distribution and use the logit link function $g(\mu_{im}) = logitPr(y_{im} = 1) = log(Pr(y_{im} = 1)/Pr(y_{im} = 0))$ for trait m.

The estimates of β and κ can be obtained by solving the GEE [75]:

$$U(\kappa, \beta) \quad = \quad \sum_{i=1}^{n} U_i(\kappa, \beta) = \sum_{i=1}^{n} \bigtriangledown {\mu_i}^{'} V_i^{-1}(Y_i - \mu_i) = 0,$$

where U_i is the score vector for subject i, $\bigtriangledown \mu_i = \partial \mu_i / \partial \theta^{'}$, $V_i = \phi A_i^{1/2} R_w(\alpha) A_i^{1/2}$, ϕ is the dispersion parameter, $v(\mu_i)$ is the variance function and $A_i = \text{diag}\{v(\mu_{i1})$, $v(\mu_{i2})$,...,$v(\mu_{ik})\}$. The marginal variance of y_{im} is $\phi v(\mu_{im})$ and for binary traits, we have $v(\mu_{im}) = \mu_{im}(1 - \mu_{im})$. $R_w(\alpha)$ is a working correlation matrix, which may depend on some unknown parameters α. Below we will derive the score vector and the variance of the score vector.

The derivative of μ_i with respect to the vector of coefficients, θ, is a matrix consisted of the partial derivative of μ_{im} with respect to each coefficient, indexed

by $l = 1, ..., (p+q+1)k$:

$$\left(\frac{\partial \mu_i}{\partial \theta'}\right)_{ml} = \frac{\partial \mu_{im}}{\partial \theta_l} = H_{iml} \times \mu_{im} \times (1 - \mu_{im}), \text{ where } \mu_{im} = \frac{1}{1 + \exp(-\eta_{im})}.$$

$\left(\frac{\partial \mu_i}{\partial \theta'}\right)_{ml}$ is the $(m,\ l)^{th}$ entry of the derivative matrix, H_{iml} is the $(m,\ l)^{th}$ element of matrix H_i, corresponding to the l^{th} variable for the m^{th} trait. Therefore we have

$$\frac{\partial \mu_i}{\partial \theta'} = \begin{pmatrix} \mu_{i1}(1-\mu_{i1}) \\ \vdots \\ \mu_{ik}(1-\mu_{ik}) \end{pmatrix} \otimes \begin{pmatrix} Z_i, & X_i \end{pmatrix},$$

and with a canonical link function and an independent working correlation matrix, it is not difficult to obtain the closed form of the score vector:

$$U(\kappa, \beta) = \sum_{i=1}^{n} \bigtriangledown {\mu_i}' A_i (Y_i - \mu_i) = \sum_{i=1}^{n} \begin{pmatrix} Z_i, & X_i \end{pmatrix}' (Y_i - \mu_i),$$

which is the equation we use for our analysis in this chapter.

6.2.2 Hypothesis testing

Our goal is to detect whether there is an association between any of the traits and a group of SNVs. To test H_0: $\beta = (\beta_{11}, \beta_{12}, ..., \beta_{kp})' = 0$ with covariates Z_i by score-based tests, we first need to fit the GEE model under the null hypothesis with $logitPr(Y_i = 1) = Z_i \varphi$ to obtain $\hat{\varphi}$ and $\hat{\mu} = 1/(1 + \exp(-Z\hat{\varphi}))$. For each subject i, if we denote for subject i, U_{i1} as the score vector corresponding to covariates Z_i and U_{i2} as the score vector for SNVs, then the score statistic under the null hypothesis, with an assumed independent working correlation for the traits, is:

$$U(\beta = 0, \varphi(\beta)) = (U_1, U_2)' = \sum_{i=1}^{n} (U_{i1}, U_{i2})',\ U_{i1} = Z_i'(Y_i - \hat{\mu_i}),\ U_{i2} = X_i'(Y_i - \hat{\mu_i}),$$

where n is the number of individuals. As $E(U(\beta = 0, \varphi(\beta)))$=0 under H_0, the estimated variance of $U(\beta = 0, \varphi(\beta))$ is:

$$\hat{\Sigma} = \sum_{i=1}^{n} U_{i.}U_{i.}^{'}/n = \sum_{i=1}^{n} \begin{pmatrix} U_{i1} \\ U_{i2} \end{pmatrix} \begin{pmatrix} U_{i1}^{'} & U_{i2}^{'} \end{pmatrix} /n$$

where $U_{i.} = (U_{i1}, U_{i2})^{'}$. In our implementation, we use a more robust estimator of Σ [80]

$$\begin{aligned} \tilde{\Sigma} &= \sum_{i=1}^{n} (Z_i,\ X_i)^{'} \hat{var}(Y)(Z_i,\ X_i) \\ &= \sum_{i=1}^{n} (Z_i,\ X_i)^{'} (\sum_{i} (Y_i - \hat{\mu_i})(Y_i - \hat{\mu_i})^{'}/n)(Z_i,\ X_i) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}. \end{aligned}$$

Because asymptotically we have

$$U_{2.1} = U_2 | \{\beta = 0, \kappa = \hat{\kappa}\} \sim N(0, \Sigma_{2.1}),\ \Sigma_{2.1} = V_{22} - V_{21}V_{11}^{-1}V_{12},$$

the score test statistic is:

$$T = U_{2.1}^{T} \Sigma_{2.1}^{-1} U_{2.1} \sim \chi_{pk}^{2}.$$

We will compare several tests:

The Wald Test: $T = \hat{\beta}^{'}(var(\hat{\beta}))^{-1}\hat{\beta}$. $\hat{\beta}$ is the estimate of β after fitting the full GEE model with $logitPr(Y_{im} = 1) = Z_{im}\kappa + X_{im}\beta$ for i^{th} subject and m^{th} trait. Under H_0, we have $T \sim \chi_{pk}^{2}$. When there are many SNVs to test, we may fail to fit the full model because of the complete separation or quasi-complete separation issue [81]. Namely, the linear combination of covariates and SNVs can perfectly separate the cases and controls so that the estimated odds ratio is infinite.

Score test: $T = U_{2.1}^{'}\Sigma_{2.1}^{-1}U_{2.1}$, where $U_{2.1}$ and $\Sigma_{2.1}$ are discussed above. Since we only need to fit the model with the covariates, it is computationally easier and less likely to have (quasi-)complete separation.

The UminP test: $T = \max_{j} U_{2.1(j)}^{2}/\Sigma_{2.1(j)}$ for $j \in 1, 2, ..., kp$, where $\Sigma_{2.1(j)}$ is the j^{th} entry on the diagonal of $\Sigma_{2.1}$. This method is as intuitive as taking the minimum p-value of testing each trait on each SNV with covariates. A Bonferroni

adjustment or a simulation-based method has to be used to calculate its p-value.

The SPU tests [40]: $T_\varrho = \sum_{j=1}^{kp} U_{2.1(j)}^{\varrho}$, where ϱ is the power parameter and $\varrho \in \Gamma$, Γ is a pre-selected set of values for ϱ, e.g. $\{1, 2, ..., 8, \infty\}$. As the power parameter, ϱ, increases, the SPU(ϱ) test puts more weights on the associated SNVs while gradually reducing the weights on the null SNVs. In particular when $\varrho \to \infty$ we have:

$$T_\infty \propto \max_j |U_{2.1(j)}|,\ j \in \{1, 2, ..., pk\}.$$

The aSPU test is a data-adaptive version of the SPU tests. It takes the minimum p-value among all SPU(ϱ) test, that is:

$$T_{aspu} = \min_\varrho P_{SPU(\varrho)},\ \varrho \in \Gamma,$$

where $P_{SPU(\varrho)}$ is the p-value of the SPU(ϱ) test. The p-value of aSPU test can be obtained based on permutation [40], and usually cannot be smaller than the smallest p-value of the SPU(ϱ) tests.

Besides these tests for multivariate traits, we also use the Score, SSU, SSUw, Sum test and UminP test for single trait analysis [16, 46]:

$$\begin{aligned} \text{SSU} &= \tilde{U}'\tilde{U}, \\ \text{SSUw} &= \tilde{U}' diag(\tilde{V})^{-1}\tilde{U}, \\ \text{Sum} &= 1'\tilde{U}, \\ \text{UminP} &= \max_s \tilde{U}_{(s)}^2/\tilde{V}_{(ss)}, \end{aligned}$$

where $\tilde{U}$ is the score vector for a single trait after adjusting for covariates in logistic regression, and $\tilde{U}_{(s)}$ is the component of the score vector for the s^{th} SNV, $\tilde{V} = cov(\tilde{U})$ and $\tilde{V}_{(ss)}$ is the s^{th} entry on the diagonal of $\tilde{V}$. The asymptotic distribution under the null hypothesis for these tests are derived in [16].

6.2.3 Simulation-based and permutation-based methods

It is not easy to directly derive the distribution of the SPU(ϱ) statistic under H_0, so we use simulation-based method to calculate p-values.

Suppose the sample size is large enough for the asymptotic null distribution of the score vector to hold, we can draw B samples of the score vector from its null distribution: $U_{2.1}^{(b)} \sim N(0, \hat{\Sigma}_{2.1})$, $b = 1, ..., B$, and obtain the statistic $T^{(b)} = \sum_{j=1}^{kp} U_{2.1(j)}^{(b)\varrho}$. $\hat{\Sigma}_{2.1}$ can be calculated as discussed before. We then calculate p-value=$\sum_{b=1}^{B} I(|T^{(b)}| > |T^{obs}|)/B$ [40].

When the sample size is too small, the permutation method may be needed. We permute residuals after fitting the trait on the covariates, but preserve the regression residuals of the SNVs on the covariates [82]. Here as our phenotypes are correlated, we proceed in three steps. First, we fit the model under H_0 by the GEE using $logit(E(Y_i|Z_i)) = Z_i\varphi$, and keep the residuals $e_i = Y_i - \hat{\mu}_i$, where $\hat{\mu}_i = 1/(1 + \exp(-Z_i\hat{\varphi}))$. Secondly we fit each SNV against the covariates, i.e. $E(X_{i(j)}|Z_i) = Z_i\zeta_j$, and keep the residuals $w_{i(j)} = X_{i(j)} - Z_i\hat{\zeta}_j$, where j indexes the j^{th} SNV, and $X_{i(j)}$ is the jth column of X_i, $w_i = (w_{i(1)}, ...w_{i(kp)})$ is a $k \times kp$ matrix. Thirdly we can estimate β, the coefficients of SNVs, by fitting a GEE on e_i against w_i with the identity link function $g(e_i) = E(e_i|w_i) = w_i\beta$. The score vector of β is

$$U = \sum_{i=1}^{n} \bigtriangledown (E(e_i|w_i))' V_i^{-1} e_i = \sum_{i=1}^{n} w_i' V_i^{-1} e_i.$$

Here we assume $v(.) = 1$ for e_i so $V_i = R_w(\alpha)$. The working correlation matrix, $R_w(\alpha)$, is assumed to be the same as in the first step: it can be either specified as any known structure or to be estimated as claimed "unstructured" in the GEE. So the observed score vector and its covariance are

$$\begin{aligned} U &= \sum_{i=1}^{n} w_i' R_w^{-1} e_i, \\ V &= \sum_{i=1}^{n} U_i U_i' = \sum_{i=1}^{n} w_i' R_w^{-1} \hat{var}(e_i) R_w^{-1} w_i, \end{aligned} \tag{6.1}$$

where $\hat{var}(e_i) = \sum_{i=1}^{n} e_i e_i'/n$.

Lastly, we can permute e_i's for B times while keeping w_i's in the original order. In permutation b we obtain $U_{2.1}^{(b)}$ and $T^{(b)}$ by Equation 6.1, and p-value=$\sum_{b=1}^{B} I(|T^b| > |T^{obs}|)/B$.

6.3 Simulations

6.3.1 Set-ups

We tested a group of CVs and RVs in simulated case-control studies using multiple traits, comparing the score test, UminP test and SPU(ϱ) test based on the GEE under various scenarios. To fully understand the advantages of multivariate trait analysis, we also carried out the comparison with univariate trait analysis, for which we used the Score, Sum, SSU, SSUw and UminP tests based on the generalized linear model (GLM).

We generated genotypes following [8]. Specifically, we had two independent blocks of SNVs, with each block in linkage disequilibrium (LD). The first block had 15 SNVs including the causal SNVs; the second block had 5 SNVs. Within each block for each subject, a latent vector $G_i = (G_{i1}, ...G_{ip})^{'}$ were first generated from a multivariate normal distribution with a first-order auto-regression (AR(1)) covariance structure: there was a correlation $Corr(G_{if}, G_{ig}) = \rho^{|f-g|}$ between any two latent components, G_{if} and G_{ig} for $f \neq g$. In our simulations we set $\rho = 0.7$. Secondly, the latent vector was dichotomized to yield a haplotype with MAFs randomly drawn from a uniform distribution. For CVs, the MAFs of the causal SNVs were between 0.3 and 0.4 while the MAFs of the null SNVs were from Unif(0.1, 0.5). For RVs, the MAFs of both the causal and null SNVs were drawn from Unif(0.005, 0.01). Thirdly, we combined two independent haplotypes to form the genotype $X_i = (X_{i1}, ...X_{ip})^T$. To mimic the real data situation, for CVs the causal SNVs were not included in our analysis, while for RVs the causal ones were in presence.

Our phenotype simulation followed [83]. We considered 3 binary traits of 1000 observations. We first generated continuous traits, $\dot{Y}_m$, m=1, 2, 3, by the following model:

$$\begin{aligned} \dot{Y}_1 &= \beta_{01} + X\beta_1 + \epsilon_1, \\ \dot{Y}_2 &= \beta_{02} + X\beta_2 + \gamma_{12}Y_1 + \epsilon_2, \\ \dot{Y}_3 &= \beta_{03} + X\beta_3 + \gamma_{13}Y_2 + \gamma_{23}Y_1 + \epsilon_3, \end{aligned} \tag{6.2}$$

$\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3)$ was a vector of extraneous variables, drawn from $N(0, W)$, where W was a compound symmetry matrix with diagonal elements as 1 and non-diagonal as

ν=0.3. Based on this model, the dependencies between traits came not only from the dependencies of the traits (γ's), but also from extraneous variables (ϵ's), such as environmental factors. $\dot{Y}_m$ was dichotomized for trait m: $Y_m = 1$ if $\dot{Y}_m \geq c_m$ and $Y_m = 0$ if $\dot{Y}_m < c_m$, where $c_m = \Phi^{-1}(1-r)$, and r was the prevalence. We set r_1=0.1, $r_2 = 0.2$, $r_3 = 0.3$, and we sampled 300 controls and 700 cases. The controls were defined as the subjects having none of the 3 diseases (Y_{im}=0 for m=1, 2, 3), while the cases were defined as the subjects having at least 1 disease (i.e. at least for one m Y_{im}=1). We chose 9 scenarios from [83] (Table 6.1). β_1, β_2, β_3 were the effect of causal SNVs on Y_1, Y_2, Y_3 respectively. $\gamma = (\gamma_{12}, \gamma_{13}, \gamma_{23})$ was a vector of the correlation between trait 1 and trait 2, trait 1 and trait 3, trait 2 and trait 3, as shown in Equation (6.2).

Briefly, in S2 and S3 there was no genetic effect, i.e. $\beta = 0$ so the proportion of simulations with p-values< 0.05 is actually the estimated Type I error. S9 and S10 had the same correlation structure among the traits but in S9, the causal SNVs were associated with the first trait while in S10 it was with the second trait. In S22 and S24 the causal SNPs directly affected 2 out of 3 traits. In S35 and S38 all traits were associated with the causal SNPs, although the correlation structure among the 3 traits was different. Notice that, although in S35 γ's=0, the 3 traits were not independent because of the extraneous correlation. 1000 simulations were conducted for each scenario. For both the simulation-based and permutation-based method, B=1000 was used.

Table 6.1: Simulation scenarios.

	for CVs			for RVs			
	β_1	β_2	β_3	β_1	β_2	β_3	γ
S2	0	0	0	0	0	0	(0.3,0,0)
S3	0	0	0	0	0	0	(0.3,0,0.3)
S9	(0.2,0.2,0.1)	0	0	(0.4,0.6,0.8)	0	0	(0.3,0,0)
S10	0	(0.2,0.2,0.1)	0				(0.3,0,0)
S16	0	0	(0.2,0.2,0.1)				(0,0.3,0.3)
S22	(0.2,0.2,0.1)	(0.2,0.2,0.1)	0	(0.4, 0.6, 0.8)	(0.4, 0.6, 0.8)	0	(0.3,0,0)
S24	0	(0.2,0.2,0.1)	(0.2,0.2,0.1)	0	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0,0.3,0)
S35	(0.2,0.2,0.1)	(0.2,0.2,0.1)	(0.2,0.2,0.1)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0,0,0)
S38	(0.2,0.2,0.1)	(0.2,0.2,0.1)	(0.2,0.2,0.1)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0.4,0.6,0.8)	(0,0.3,0.3)

6.3.2 Results

Table 6.2 shows the results of univariate trait analysis by using the Score test, Sum test, SSU/SSUw test and UminP test. S stands for simulation-based results and

B for permutation-based results. We notice that the 5 tests all had descent power to detect the causal SNVs. For example, in S9, provided that only trait 1 was associated with the causal SNP, SSU could reject the null hypothesis 75% of the time. For the other traits, even though they were related with the affected trait, the power was much lower.

Table 6.3 shows the Type I error rates and power when testing three traits together on CVs. The Type I error (S2 and S3) were all around 0.05 nominal level for both simulation-based and permutation-based methods. With more traits associated with the causal SNVs, we could see a generally increasing power. For example, based on asymptotic p-values, in S9, the score test had a power of 0.298, the aSPU had a power of 0.407; in S35, the score test had a power of 0.637 and the aSPU test had a power of 0.946. In more detail, we see that the highest power of SPU tests were all obtained when $\varrho = 2$ or 3. This was probably because we only included 5 null CVs in the analysis. As shown in Pan et al. (2013), with more null SNVs, the benefit of SPU tests with larger ϱ would be more obvious.

Comparing Table 6.2 with Table 6.3, we could see that when there was only 1 trait associated with the disease loci, multivariate trait analysis was not improved over univariate trait analysis, or even worse. For example, in S9, the highest power in univariate analysis was obtained by SSU when testing trait 1, 0.750, while the highest power testing all 3 traits was obtained by UminP, 0.465. However, as the number of associated traits increased, multivariate trait analysis presented a growing advantage. For example, in S35, while the highest power in univariate analysis was about 0.68 by the SSU test for trait 2, the aSPU test yielded a power of 0.937 in multivariate trait analysis.

Table 6.2: Simulations of univariate trait analysis with CVs

scenario	method	Y1					Y2					Y3				
		score	SSU	SSUw	SUM	UminP	score	SSU	SSUw	SUM	UminP	score	SSU	SSUw	SUM	UminP
S2	S	0.041	0.052	0.053	0.053	0.057	0.050	0.050	0.047	0.053	0.048	0.052	0.042	0.039	0.052	0.056
	B	0.041	0.058	0.058	0.052	0.057	0.052	0.053	0.051	0.054	0.048	0.051	0.044	0.040	0.053	0.054
S3	S	0.054	0.053	0.050	0.049	0.059	0.043	0.055	0.048	0.047	0.057	0.051	0.055	0.056	0.059	0.048
	B	0.058	0.053	0.054	0.048	0.058	0.045	0.054	0.047	0.048	0.054	0.054	0.058	0.056	0.060	0.051
S9	S	0.534	0.750	0.748	0.662	0.627	0.059	0.074	0.071	0.086	0.075	0.037	0.055	0.046	0.051	0.059
	B	0.539	0.756	0.754	0.659	0.622	0.058	0.078	0.076	0.087	0.074	0.038	0.057	0.050	0.052	0.058
S24	S	0.062	0.058	0.056	0.056	0.053	0.530	0.720	0.727	0.640	0.590	0.375	0.577	0.580	0.537	0.455
	B	0.063	0.059	0.054	0.053	0.054	0.535	0.727	0.729	0.643	0.594	0.379	0.584	0.580	0.537	0.459
S35	S	0.395	0.609	0.609	0.549	0.493	0.445	0.676	0.686	0.564	0.552	0.450	0.662	0.662	0.594	0.545
	B	0.395	0.617	0.613	0.547	0.494	0.443	0.677	0.686	0.559	0.548	0.459	0.670	0.666	0.591	0.542

Table 6.3: Simulations of multivariate trait analysis with CVs: S stands for simulation-based and B stands for permutation

scenario	method	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(6)	SPU(7)	SPU(8)	SPU(∞)	aSPU
S2	S	0.041	0.043	0.056	0.047	0.053	0.048	0.053	0.047	0.051	0.046	0.042	0.050
	B	0.043	0.046	0.056	0.051	0.055	0.052	0.056	0.050	0.057	0.047	0.046	0.051
S3	S	0.049	0.048	0.058	0.048	0.053	0.053	0.059	0.053	0.056	0.052	0.058	0.053
	B	0.046	0.048	0.058	0.049	0.054	0.054	0.059	0.055	0.059	0.055	0.057	0.054
S9	S	0.298	0.465	0.262	0.441	0.403	0.429	0.385	0.372	0.350	0.343	0.291	0.407
	B	0.306	0.474	0.254	0.450	0.405	0.430	0.391	0.386	0.354	0.353	0.286	0.420
S22	S	0.573	0.698	0.624	0.863	0.840	0.845	0.817	0.808	0.784	0.775	0.680	0.853
	B	0.576	0.703	0.630	0.866	0.840	0.845	0.823	0.809	0.785	0.782	0.683	0.859
S24	S	0.513	0.568	0.644	0.812	0.833	0.789	0.788	0.735	0.729	0.697	0.597	0.811
	B	0.515	0.575	0.652	0.812	0.832	0.797	0.788	0.738	0.730	0.699	0.609	0.811
S35	S	0.637	0.688	0.898	0.922	0.937	0.876	0.877	0.805	0.806	0.763	0.641	0.921
	B	0.639	0.689	0.887	0.917	0.933	0.878	0.882	0.812	0.814	0.768	0.649	0.922
S38	S	0.722	0.830	0.922	0.962	0.961	0.943	0.944	0.916	0.902	0.892	0.813	0.959
	B	0.722	0.829	0.924	0.960	0.960	0.945	0.941	0.917	0.908	0.891	0.811	0.962

For testing RVs, the comparison conclusion regarding univariate trait and multivariate trait analyses applied here, that as more phenotypes were associated with the disease loci, multivariate trait analysis was more powerful (Table 6.4 and Table 6.5). For example, in S9, the highest power in univariate trait analysis (0.738 by the SSUw test) for testing trait 1, was greater than the highest power in multivariate trait analysis (0.657 by the UminP test); in S35, multivariate trait analysis had a greater power (0.866 by the aSPU test) than univariate trait analysis (0.639 by the SSUw test). We also noticed that in S9, where only 1 out of 3 traits was affected by the causal RVs, the SPU(ϱ) test was not as well-performed as the UminP test and score test. The enhanced power of UminP was most probably due to the inclusion of causal RVs in the global test. However, as the number of associated traits increased, in S24 and S35, the SPU(ϱ) test had greater power than the UminP and score tests. Moreover, even now we had RVs with very low MAFs, the permutation-based method still generated a similar Type I error and power as the simulation-based method, possibly because the sample size was big enough for the asymptotic theory to hold.

Table 6.4: Simulations of univariate trait analysis with RVs: S stands for simulation-based method and B stands for permutation-based method

		Y1					Y2					Y3				
scenario	method	score	SSU	SSUw	SUM	UminP	score	SSU	SSUw	SUM	UminP	score	SSU	SSUw	SUM	UminP
S2	S	0.062	0.052	0.057	0.049	0.083	0.044	0.037	0.037	0.043	0.025	0.051	0.045	0.045	0.056	0.037
	B	0.051	0.046	0.048	0.053	0.045	0.056	0.046	0.048	0.043	0.045	0.058	0.055	0.059	0.060	0.063
S3	S	0.063	0.049	0.052	0.043	0.095	0.042	0.054	0.053	0.057	0.033	0.030	0.044	0.037	0.051	0.026
	B	0.048	0.041	0.039	0.047	0.046	0.058	0.060	0.058	0.059	0.056	0.037	0.052	0.045	0.050	0.048
S9	S	0.711	0.731	0.738	0.512	0.750	0.054	0.065	0.062	0.061	0.047	0.036	0.033	0.036	0.058	0.023
	B	0.692	0.721	0.723	0.513	0.678	0.064	0.074	0.073	0.062	0.077	0.048	0.039	0.044	0.056	0.040
S24	S	0.043	0.047	0.040	0.044	0.075	0.549	0.632	0.605	0.448	0.600	0.177	0.275	0.253	0.252	0.128
	B	0.034	0.042	0.035	0.049	0.040	0.562	0.650	0.621	0.454	0.610	0.208	0.295	0.272	0.255	0.183
S35	S	0.571	0.639	0.604	0.443	0.601	0.466	0.607	0.545	0.416	0.506	0.314	0.454	0.390	0.342	0.281
	B	0.563	0.634	0.591	0.442	0.533	0.486	0.624	0.566	0.412	0.541	0.348	0.475	0.412	0.338	0.365

Table 6.5: Simulations of multivariate trait analysis with RVs

scenario	method	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(6)	SPU(7)	SPU(8)	SPU(∞)	aSPU
S3	S	0.045	0.061	0.037	0.043	0.038	0.035	0.046	0.036	0.045	0.039	0.041	0.041
	B	0.048	0.045	0.036	0.053	0.046	0.046	0.051	0.046	0.055	0.052	0.055	0.051
S9	S	0.514	0.657	0.166	0.388	0.345	0.378	0.353	0.352	0.338	0.342	0.317	0.367
	B	0.528	0.645	0.162	0.400	0.361	0.422	0.382	0.388	0.363	0.371	0.338	0.403
S22	S	0.679	0.731	0.405	0.774	0.762	0.764	0.744	0.709	0.706	0.692	0.645	0.761
	B	0.688	0.709	0.399	0.784	0.766	0.787	0.767	0.752	0.738	0.723	0.679	0.775
S24	S	0.461	0.495	0.389	0.631	0.692	0.644	0.648	0.594	0.591	0.556	0.508	0.638
	B	0.459	0.420	0.388	0.643	0.703	0.662	0.658	0.613	0.613	0.578	0.530	0.665
S35	S	0.761	0.681	0.760	0.843	0.879	0.810	0.813	0.758	0.758	0.712	0.637	0.866
	B	0.769	0.652	0.763	0.855	0.878	0.826	0.826	0.785	0.771	0.741	0.662	0.871
S38	S	0.702	0.722	0.692	0.816	0.838	0.774	0.778	0.710	0.688	0.675	0.600	0.825
	B	0.705	0.630	0.694	0.826	0.844	0.789	0.778	0.718	0.706	0.677	0.599	0.829

Lastly, as the GEE model does not assume the distribution of the traits, we felt it would be possible to analyze traits of different distributions. We simulated two binary traits following the procedure before, and one quantitative trait, which was not discretized during data generation. The cases and controls were defined based on the two binary traits. The Type I error rates were under control (S3) (Table 6.6). In most cases the power was greater than using all 3 binary traits. For instance in S9, the aSPU test had a power of 0.367 when testing 3 binary traits with RVs (Table 6.5) while here it had a power of 0.465 (Table 6.6). The power gain could be due to the inclusion of the quantitative trait. However, it was surprising that, when testing CVs in S9, while the score test had a power 0.400, the SPU tests all had a very low power below 0.2. We suspected this was because we used the identity link for all three traits, including the binary traits. Thus, the variance was over-estimated. As the score and UminP tests were all divided by the covariance matrix or variance of the score statistic, the power was maintained, while the SPU tests were heavily affected.

Table 6.6: Simulation results with 2 binary traits and 1 quantitative traits. An identity link is used.

		score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(6)	SPU(7)	SPU(8)	SPU(∞)	aSPU
CVs													
S3	S	0.052	0.054	0.060	0.060	0.061	0.056	0.050	0.056	0.054	0.054	0.052	0.055
	B	0.053	0.052	0.058	0.062	0.060	0.057	0.054	0.054	0.051	0.050	0.051	0.058
S9	S	0.400	0.539	0.152	0.178	0.088	0.082	0.073	0.073	0.065	0.063	0.064	0.150
	B	0.409	0.543	0.152	0.188	0.092	0.091	0.073	0.078	0.069	0.066	0.063	0.158
S24	S	0.673	0.697	0.726	0.862	0.841	0.789	0.768	0.744	0.738	0.712	0.650	0.833
	B	0.685	0.708	0.726	0.866	0.841	0.793	0.775	0.742	0.731	0.716	0.659	0.841
RVs													
S3	S	0.049	0.058	0.048	0.053	0.053	0.059	0.053	0.056	0.052	0.058	0.053	0.048
	B	0.046	0.048	0.058	0.049	0.054	0.054	0.059	0.055	0.059	0.055	0.057	0.054
S9	S	0.298	0.262	0.441	0.403	0.429	0.385	0.372	0.350	0.343	0.291	0.407	0.465
	B	0.306	0.474	0.254	0.450	0.405	0.430	0.391	0.386	0.354	0.353	0.286	0.420
S24	S	0.513	0.644	0.812	0.833	0.789	0.788	0.735	0.729	0.697	0.597	0.811	0.568
	B	0.515	0.575	0.652	0.812	0.832	0.797	0.788	0.738	0.730	0.699	0.609	0.811

6.4 Real Data Analysis

6.4.1 Application to the NvR data

We first applied the score test, UminP test and SPU tests to the NvR study [43] to investigate the viral RNA mutations associated with HIV RNA level, CD4 counts and AIDs/death. The data and variables used in the analysis were explained in Chapter 2. We were first interested in how the mutations occurring in last visit $(t-1)$ affected the drug resistance response at visit t. For a cross-sectional analysis at each visit t, we allowed different odds ratios (ORs) of mutations and covariates for different traits.

The relation between the marginal mean, mutations and covariates were modeled as:

$$\begin{aligned} & (logitPr(\text{V(t)}=1),\ logitPr(\text{CD4(t)}=1),\ logitPr(\text{SYM(t)}=1))_i^T \\ = \ & \mathbf{I}\kappa_0 + \text{diag}\{\text{basePI}\}_i\kappa_1 + \text{diag}\{\text{NAVEBL}\}_i\kappa_2 + \text{diag}\{\text{AGE}\}_i\kappa_3 \\ & + \text{diag}\{\text{FEMALE}\}_i\kappa_4 + \text{diag}\{\text{WHITE}\}_i\kappa_5 + \text{diag}\{\text{PODBL}\}_i\kappa_6 \\ & + \text{IND\_RNA(t-1)}_i\kappa_7 + \text{geno(t-1)}_i\kappa_8 + \text{diag}\{\text{mute(t-1)}\}_i\beta, \end{aligned}$$

where $\mathbf{I}$ is an identity matrix, diag is defined with non-zero diagonal entries and zeros otherwise, κ_h is a vector of 3 coefficients for 3 traits, $h = 1, 2, ..., 8$.

We first tested on each single common mutation. The mutations having a p-value < 0.05 for any test in cross-sectional multivariate trait analysis are shown in Figure 6.1. The permutation-based p-values (-B) were obtained based on 1000 resamples, which were very close to simulation-based p-values (-S). For most of the mutations, like X83K in month 8, the p-values of all tests, especially of the score and SPU tests, were similar. However, there were some mutations found to be very significant by the Wald test while not significant at all by other tests, like X118I in month 8. We doubted this was owing to some convergence issue in the Wald test. There was no mutations being significant across all periods, though several mutations appeared to be significant in multiple periods and the summarized results are shown in Table 6.7. For example, the mutation X214L was significant for all periods except month 12. In fact, the mutation X214 and X211 are both identified to be associated with drug Azido-thymidine (AZT), which is widely used in the trial.

As a comparison, we did a cross-sectional univariate analysis only with the change of HIV RNA level. Comparing Figure 6.1 with Figure 6.2, there were less significant mutations found in the univariate trait analysis, and the significant mutations were less overlapped across different time points. Overall, most of the significant mutations found in the univariate trait analysis were also significant in multivariate analysis. The ones a p-value< 0.05 in both analyses are marked in yellow in Figure 6.2.

We also did cross-sectional multivariate trait analysis on sliding windows of 5 neighboring rare mutations with moving step 2 (Table 6.8). There was a somewhat big difference between the simulations-based method and permutation-based method for the UminP test and score test. This could be due to the small-sample size issue, as the relative frequencies of some mutations could be extremely low. Based on the permutation-based method, for month 8, 12, and 20, the aSPU test had a greater rejection rate than the score and UminP tests.

Figure 6.1: Cross-sectional multivariate trait analysis on common mutations for month 4, 8, 12, 16, 20, 24. -B stands for permutation-based methods.

				SPU(ϱ)									SPU.B(ϱ)						
	mute	Wald	Score	$\varrho=1$	2	3	4	5	∞	aSPU	minU	Score.B	minU.B	$\varrho=1$	2	3	4	5	∞	aSPU.B
4 month																				
1	X10I	0.030	0.061	0.583	0.071	0.183	0.075	0.115	0.093	0.131	0.097	0.065	0.104	0.641	0.093	0.227	0.103	0.149	0.118	0.152
4	X15V	0.074	0.076	0.167	0.058	0.035	0.044	0.038	0.038	0.067	0.038	0.081	0.043	0.200	0.084	0.048	0.049	0.043	0.040	0.080
8	X36I	0.115	0.024	0.025	0.073	0.039	0.047	0.039	0.038	0.057	0.010	0.025	0.014	0.046	0.080	0.057	0.063	0.055	0.054	0.086
18	X67N	0.114	0.104	0.149	0.089	0.039	0.044	0.039	0.037	0.070	0.039	0.097	0.045	0.197	0.107	0.057	0.058	0.051	0.046	0.081
21	X83K	0.007	0.024	0.396	0.012	0.129	0.021	0.059	0.038	0.027	0.065	0.026	0.067	0.383	0.017	0.145	0.034	0.081	0.046	0.035
24	X123E	0.066	0.051	0.416	0.027	0.104	0.041	0.062	0.054	0.050	0.088	0.059	0.080	0.454	0.041	0.144	0.052	0.086	0.064	0.073
29	X184V	0.009	0.016	0.048	0.043	0.037	0.061	0.064	0.103	0.072	0.089	0.012	0.086	0.068	0.062	0.064	0.097	0.095	0.139	0.106
37	X214L	0.000	0.007	0.005	0.007	0.006	0.007	0.006	0.008	0.009	0.010	0.008	0.012	0.009	0.005	0.004	0.004	0.004	0.004	0.006
8 months																				
11	X41K	0.001	0.003	0.037	0.015	0.004	0.004	0.004	0.003	0.007	0.000	0.002	0.000	0.054	0.016	0.006	0.008	0.007	0.008	0.013
21	X83K	0.002	0.013	0.006	0.006	0.003	0.004	0.004	0.003	0.005	0.011	0.014	0.007	0.007	0.009	0.004	0.005	0.004	0.003	0.006
22	X93L	0.000	0.001	0.152	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.003	0.171	0.003	0.001	0.001	0.001	0.001	0.002
29	X184V	0.009	0.045	0.766	0.084	0.309	0.107	0.197	0.131	0.145	0.070	0.049	0.058	0.761	0.104	0.350	0.122	0.205	0.149	0.183
36	X211K	0.064	0.063	0.040	0.030	0.015	0.017	0.015	0.015	0.030	0.028	0.068	0.031	0.045	0.042	0.024	0.025	0.018	0.018	0.035
37	X214L	0.002	0.042	0.006	0.030	0.031	0.044	0.050	0.060	0.014	0.101	0.048	0.106	0.021	0.048	0.048	0.065	0.071	0.078	0.040
42	X90M	0.026	0.105	0.054	0.179	0.138	0.202	0.190	0.258	0.104	0.143	0.102	0.146	0.091	0.224	0.180	0.241	0.231	0.289	0.169
43	X118I	0.006	0.165	0.614	0.149	0.115	0.119	0.100	0.098	0.176	0.126	0.158	0.081	0.621	0.183	0.154	0.149	0.131	0.122	0.206
12 months																				
4	X15V	0.038	0.039	0.030	0.031	0.021	0.022	0.020	0.019	0.035	0.033	0.033	0.041	0.038	0.043	0.032	0.033	0.033	0.032	0.051
10	X37S	0.066	0.054	0.114	0.018	0.006	0.006	0.006	0.006	0.011	0.028	0.057	0.022	0.159	0.039	0.019	0.020	0.018	0.018	0.031
21	X83K	0.158	0.167	0.042	0.085	0.077	0.096	0.091	0.099	0.073	0.174	0.176	0.190	0.051	0.116	0.106	0.124	0.114	0.126	0.091
22	X93L	0.035	0.028	0.067	0.019	0.018	0.020	0.019	0.019	0.034	0.043	0.029	0.048	0.106	0.055	0.043	0.043	0.036	0.036	0.066
32	X202V	0.000	0.048	0.348	0.142	0.122	0.120	0.113	0.108	0.174	0.016	0.047	0.022	0.401	0.189	0.148	0.143	0.130	0.126	0.207
34	X207Q	0.000	0.028	0.443	0.012	0.097	0.026	0.045	0.035	0.027	0.070	0.020	0.050	0.481	0.034	0.161	0.057	0.098	0.079	0.067
44	X219Q	0.026	0.179	0.264	0.188	0.116	0.130	0.112	0.118	0.186	0.087	0.186	0.108	0.313	0.234	0.167	0.173	0.155	0.149	0.228
45	X46I	0.001	0.379	0.573	0.262	0.211	0.211	0.193	0.189	0.302	0.325	0.399	0.334	0.625	0.336	0.279	0.294	0.259	0.255	0.384
46	X71V	0.000	0.164	0.561	0.095	0.080	0.079	0.070	0.066	0.119	0.144	0.171	0.145	0.581	0.123	0.097	0.097	0.091	0.089	0.151
47	X82A	0.000	0.102	0.075	0.120	0.065	0.078	0.069	0.067	0.108	0.052	0.092	0.054	0.074	0.153	0.103	0.120	0.105	0.108	0.135
16 months																				
21	X83K	0.078	0.048	0.026	0.166	0.173	0.234	0.236	0.267	0.046	0.104	0.046	0.100	0.032	0.183	0.206	0.275	0.282	0.320	0.055
24	X123E	0.119	0.105	0.026	0.028	0.023	0.023	0.024	0.023	0.040	0.059	0.110	0.056	0.048	0.049	0.040	0.040	0.039	0.039	0.062
26	X162C	0.043	0.182	0.188	0.225	0.200	0.239	0.214	0.221	0.293	0.341	0.178	0.320	0.224	0.273	0.226	0.268	0.243	0.250	0.337
27	X177E	0.061	0.065	0.019	0.022	0.014	0.014	0.014	0.014	0.026	0.037	0.075	0.036	0.040	0.042	0.026	0.026	0.024	0.022	0.041
36	X211K	0.029	0.026	0.644	0.015	0.044	0.026	0.033	0.030	0.027	0.064	0.023	0.071	0.685	0.037	0.078	0.046	0.053	0.048	0.061
37	X214L	0.000	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.001	0.000	0.001	0.002	0.000	0.001	0.001	0.001	0.002	0.002	0.000
40	X245M	0.017	0.065	0.173	0.046	0.050	0.045	0.046	0.044	0.077	0.126	0.060	0.107	0.182	0.075	0.076	0.078	0.076	0.074	0.127
42	X90M	0.030	0.083	0.971	0.082	0.287	0.116	0.186	0.147	0.128	0.242	0.075	0.242	0.960	0.104	0.359	0.142	0.229	0.182	0.166
44	X219Q	0.000	0.016	0.085	0.029	0.019	0.020	0.018	0.018	0.033	0.011	0.019	0.014	0.110	0.062	0.043	0.048	0.044	0.046	0.072
45	X46I	0.000	0.130	0.050	0.036	0.016	0.017	0.013	0.011	0.021	0.052	0.123	0.051	0.075	0.062	0.031	0.035	0.033	0.034	0.057
46	X71V	0.001	0.156	0.456	0.150	0.164	0.144	0.136	0.133	0.216	0.219	0.162	0.233	0.460	0.178	0.167	0.161	0.156	0.152	0.235
47	X82A	0.000	0.132	0.328	0.233	0.280	0.292	0.296	0.304	0.352	0.299	0.136	0.309	0.366	0.303	0.351	0.365	0.357	0.368	0.437
48	X44D	0.000	0.306	0.329	0.203	0.177	0.178	0.167	0.164	0.265	0.327	0.286	0.310	0.380	0.267	0.246	0.243	0.222	0.226	0.343
49	X54V	0.000	0.274	0.252	0.392	0.330	0.386	0.356	0.364	0.371	0.399	0.242	0.349	0.268	0.453	0.371	0.432	0.409	0.416	0.387
50	X71T	0.116	0.139	0.016	0.077	0.069	0.096	0.098	0.111	0.028	0.238	0.133	0.197	0.030	0.099	0.096	0.124	0.124	0.147	0.053
54	X208Y	0.000	0.057	0.041	0.030	0.028	0.033	0.033	0.033	0.048	0.071	0.062	0.084	0.057	0.044	0.050	0.056	0.056	0.059	0.073
20 months																				
10	X37S	0.022	0.060	0.300	0.041	0.034	0.032	0.030	0.029	0.058	0.046	0.062	0.055	0.380	0.064	0.056	0.047	0.047	0.045	0.078
23	X122E	0.052	0.042	0.836	0.068	0.395	0.103	0.227	0.151	0.121	0.026	0.032	0.024	0.877	0.123	0.482	0.163	0.288	0.209	0.196
27	X177E	0.055	0.056	0.035	0.022	0.010	0.009	0.008	0.008	0.014	0.019	0.058	0.019	0.064	0.032	0.024	0.022	0.022	0.021	0.036
29	X184V	0.063	0.130	0.461	0.306	0.214	0.238	0.206	0.214	0.335	0.047	0.146	0.057	0.555	0.420	0.315	0.330	0.298	0.303	0.443
36	X211K	0.165	0.126	0.195	0.079	0.040	0.045	0.040	0.038	0.068	0.092	0.133	0.095	0.280	0.126	0.086	0.086	0.080	0.078	0.119
37	X214L	0.004	0.040	0.002	0.061	0.082	0.140	0.147	0.233	0.004	0.219	0.047	0.240	0.031	0.118	0.142	0.196	0.212	0.297	0.055
43	X118I	0.012	0.224	0.279	0.346	0.323	0.403	0.380	0.435	0.417	0.212	0.206	0.207	0.340	0.450	0.424	0.499	0.475	0.510	0.493
44	X219Q	0.000	0.041	0.030	0.060	0.031	0.036	0.030	0.030	0.057	0.022	0.034	0.028	0.059	0.096	0.068	0.067	0.062	0.061	0.097
45	X46I	0.001	0.251	0.125	0.151	0.074	0.079	0.067	0.062	0.113	0.145	0.259	0.123	0.189	0.179	0.101	0.110	0.097	0.096	0.162
49	X54V	0.000	0.211	0.866	0.351	0.337	0.347	0.319	0.314	0.469	0.132	0.209	0.100	0.881	0.408	0.436	0.413	0.396	0.384	0.532
50	X71T	0.048	0.160	0.308	0.189	0.142	0.153	0.136	0.129	0.214	0.131	0.146	0.103	0.354	0.237	0.200	0.197	0.188	0.185	0.285
55	X10V	0.030	0.385	0.295	0.425	0.295	0.338	0.298	0.300	0.439	0.289	0.375	0.268	0.320	0.484	0.348	0.385	0.349	0.343	0.470
56	X74S	0.000	0.112	0.489	0.159	0.579	0.249	0.626	0.384	0.258	0.312	0.109	0.267	0.560	0.200	0.631	0.290	0.693	0.451	0.312
57	X103N	0.005	0.439	0.314	0.281	0.227	0.252	0.221	0.217	0.345	0.387	0.413	0.341	0.398	0.335	0.279	0.311	0.282	0.277	0.401
58	X135V	0.000	0.072	0.731	0.055	0.346	0.102	0.187	0.136	0.108	0.261	0.069	0.233	0.788	0.121	0.422	0.176	0.247	0.209	0.199
59	X181C	0.000	0.152	0.571	0.154	0.758	0.252	0.634	0.356	0.253	0.139	0.142	0.125	0.620	0.215	0.789	0.321	0.689	0.420	0.334
24 months																				
25	X135T	0.099	0.125	0.134	0.057	0.048	0.049	0.043	0.041	0.065	0.102	0.144	0.120	0.168	0.097	0.085	0.083	0.077	0.075	0.122
27	X177E	0.009	0.013	0.010	0.004	0.003	0.003	0.003	0.003	0.005	0.000	0.015	0.006	0.020	0.010	0.005	0.005	0.005	0.005	0.010
34	X207Q	0.286	0.020	0.319	0.295	0.204	0.218	0.206	0.203	0.308	0.007	0.025	0.009	0.352	0.376	0.245	0.269	0.243	0.244	0.364
36	X211K	0.110	0.084	0.160	0.045	0.028	0.031	0.027	0.028	0.052	0.060	0.076	0.056	0.174	0.068	0.044	0.043	0.039	0.037	0.060
37	X214L	0.040	0.042	0.052	0.022	0.033	0.039	0.038	0.039	0.040	0.064	0.039	0.059	0.089	0.034	0.049	0.050	0.053	0.054	0.061
44	X219Q	0.009	0.138	0.123	0.131	0.099	0.113	0.102	0.100	0.164	0.077	0.130	0.060	0.130	0.146	0.109	0.122	0.114	0.113	0.182
54	X208Y	0.000	0.396	0.618	0.575	0.704	0.666	0.726	0.758	0.729	0.418	0.387	0.438	0.675	0.647	0.754	0.729	0.768	0.799	0.805
57	X103N	0.728	0.079	0.232	0.406	0.314	0.337	0.311	0.318	0.342	0.036	0.068	0.027	0.281	0.494	0.368	0.412	0.376	0.379	0.413
60	X72V	0.000	0.149	0.346	0.201	0.219	0.210	0.201	0.198	0.302	0.174	0.142	0.168	0.387	0.257	0.275	0.270	0.260	0.257	0.388
61	X179I	0.022	0.664	0.824	0.638	0.604	0.612	0.580	0.581	0.737	0.721	0.663	0.719	0.826	0.685	0.637	0.662	0.626	0.633	0.800

Table 6.7: Significant viral mutations across different visits

	4 mon	8 mon	12 mon	16 mon	20 mon	24 mon
X83K	+	+	+	+		
X184V	+	+			+	
X214L	+	+		+	+	+
X211K		+		+	+	+
X207Q			+			+
X219Q			+	+	+	+
X46I			+	+	+	
X71V			+	+	+	
X82A			+	+		
X54V				+	+	
X71T				+	+	
X208				+		
X177E				+	+	+
X103N					+	+

Figure 6.2: Univariate analysis on RVs with change of HIV RNA level

				SPU(ϱ)								SPU.B(ϱ)							
	mute	Wald	Score	$\varrho=1$	2	3	4	5	∞	aSPU	minU	Score.B	minU.B	$\varrho=1$	2	3	4	5	∞	aSPU.B
04 month																				
1	X10I	0.014	0.018	0.016	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018
12	X41L	0.274	0.039	0.044	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039
18	X67N	0.013	0.004	0.004	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
21	X83K	0.064	0.050	0.053	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061
29	X184V	0.025	0.003	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
31	X200A	0.125	0.022	0.025	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026	0.026
38	X215Y	0.208	0.015	0.010	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
08 month																				
1	X10I	0.027	0.021	0.017	0.021	0.021	0.021	0.021	0.021	0.021	0.021	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
41	X30N	0.008	0.020	0.014	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
43	X118I	0.000	0.009	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
44	X219Q	0.046	0.299	0.274	0.298	0.298	0.298	0.298	0.298	0.298	0.298	0.324	0.324	0.324	0.324	0.324	0.324	0.324	0.324	0.324
12 month																				
4	X15V	0.083	0.021	0.017	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025
16	X63P	0.345	0.039	0.029	0.044	0.044	0.044	0.044	0.044	0.044	0.044	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037
23	X122E	0.133	0.019	0.021	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018
26	X162C	0.027	0.022	0.021	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014
29	X184V	0.279	0.054	0.045	0.065	0.065	0.065	0.065	0.065	0.065	0.065	0.052	0.052	0.052	0.052	0.052	0.052	0.052	0.052	0.052
32	X202V	0.000	0.002	0.001	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
34	X207Q	0.000	0.002	0.000	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
41	X30N	0.030	0.071	0.105	0.051	0.051	0.051	0.051	0.051	0.051	0.051	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058
16 month																				
	mute	w	sco	minU	spu.1	spu.2	spu.3	spu.4	spu.5	spu+	spu.a	sco.p	minU.p	spu.1p	spu.2p	spu.3p	spu.4p	spu.5p	spu+p	spu.ap
10	X37S	0.028	0.085	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.081	0.080	0.080	0.080	0.080	0.080	0.080	0.080	0.080	0.080
16	X63P	0.134	0.006	0.012	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
26	X162C	0.007	0.015	0.017	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
28	X178M	0.024	0.042	0.044	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045
46	X71V	0.000	0.076	0.079	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.070	0.070	0.070	0.070	0.070	0.070	0.070	0.070	0.070
47	X82A	0.000	0.110	0.109	0.102	0.102	0.102	0.102	0.102	0.102	0.102	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.076
48	X44D	0.000	0.212	0.231	0.210	0.210	0.210	0.210	0.210	0.210	0.210	0.190	0.190	0.190	0.190	0.190	0.190	0.190	0.190	0.190
49	X54V	0.000	0.217	0.216	0.198	0.198	0.198	0.198	0.198	0.198	0.198	0.193	0.193	0.193	0.193	0.193	0.193	0.193	0.193	0.193
54	X208Y	0.000	0.242	0.229	0.232	0.232	0.232	0.232	0.232	0.232	0.232	0.216	0.216	0.216	0.216	0.216	0.216	0.216	0.216	0.216
20 month																				
6	X35D	0.048	0.033	0.031	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025
14	X60I	0.003	0.021	0.018	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018	0.018
23	X122E	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
29	X184V	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
31	X200A	0.128	0.042	0.052	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
33	X207E	0.122	0.036	0.025	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034
37	X214L	0.006	0.066	0.063	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.068	0.068	0.068	0.068	0.068	0.068	0.068	0.068	0.068
43	X118I	0.009	0.066	0.069	0.068	0.068	0.068	0.068	0.068	0.068	0.068	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059	0.059
49	X54V	0.000	0.031	0.033	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
56	X74S	0.000	0.135	0.135	0.130	0.130	0.130	0.130	0.130	0.130	0.130	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.112
58	X135V	0.000	0.178	0.171	0.170	0.170	0.170	0.170	0.170	0.170	0.170	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
59	X181C	0.000	0.022	0.017	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022
24 month																				
16	X63P	0.165	0.010	0.014	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016
27	X177E	0.406	0.049	0.051	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.065	0.065	0.065	0.065	0.065	0.065	0.065	0.065	0.065
34	X207Q	0.035	0.008	0.012	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
36	X211K	0.415	0.050	0.052	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
38	X215Y	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
54	X208Y	0.000	0.114	0.127	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.095	0.095	0.095	0.095	0.095	0.095	0.095	0.095	0.095
55	X10V	0.001	0.096	0.093	0.099	0.099	0.099	0.099	0.099	0.099	0.099	0.082	0.082	0.082	0.082	0.082	0.082	0.082	0.082	0.082
57	X103N	0.199	0.010	0.009	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016
60	X72V	0.000	0.071	0.058	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066
61	X179I	0.000	0.516	0.516	0.490	0.490	0.490	0.490	0.490	0.490	0.490	0.546	0.546	0.546	0.546	0.546	0.546	0.546	0.546	0.546

Table 6.8: Multivariate cross-sectional analysis on sliding windows of 5 RVs with moving step 2. S stands for simulation-based method and B stands for permutation-based method.

				SPU(ϱ)						
month	method	score	UminP	ϱ =1	2	3	4	5	∞	aSPU
4	S	0.074	0.162	0.037	0.037	0.037	0.037	0.037	0.037	0.030
	B	0.048	0.048	0.037	0.037	0.026	0.030	0.030	0.030	0.022
8	S	0.037	0.108	0.054	0.041	0.068	0.047	0.054	0.044	0.058
	B	0.031	0.037	0.051	0.047	0.068	0.058	0.054	0.047	0.051
12	S	0.086	0.128	0.045	0.048	0.045	0.061	0.048	0.051	0.054
	B	0.058	0.035	0.058	0.048	0.048	0.051	0.045	0.042	0.058
16	S	0.113	0.141	0.046	0.071	0.037	0.064	0.040	0.058	0.037
	B	0.064	0.071	0.052	0.064	0.031	0.052	0.049	0.046	0.040
20	S	0.045	0.126	0.030	0.054	0.063	0.072	0.069	0.072	0.054
	B	0.030	0.018	0.030	0.051	0.069	0.066	0.069	0.063	0.048
24	S	0.120	0.163	0.047	0.061	0.035	0.047	0.041	0.047	0.038
	B	0.061	0.067	0.050	0.047	0.035	0.047	0.041	0.047	0.050

We were also curious about if any mutations had effects on the change of HIV RNA level over time, so we carried out a longitudinal analysis. The relation between the mean, mutations and covariates were modeled as:

$$\begin{aligned} & (logitPr(V(1)=1), logitPr(V(4)=1), ..., logitPr(V(t)=1))_i^T \\ = \; & \eta_0 + \text{basePI}_i\eta_1 + \text{NAVEBL}_i\eta_2 + \text{AGE}_i\eta_3 + \text{FEMALE}_i\eta_4 + \text{WHITE}_i\eta_5 \\ & + \text{PODBL}_i\eta_6 + \text{IND\_RNA}_i\eta_7 + \text{CD4BL}_i\eta_8 + \text{mute}_i\beta, \end{aligned}$$

where η_h is a scalar coefficient for $h = 1, 2, ..., 8$. For the convergence issue in longitudinal analysis, we assumed the covariates and mutations had the same effect at all time points. IND_RNA_i is a vector containing the indicators of whether the viral RNA load was beyond 2000 copies/ml at each time point from baseline up to month $t-1$ for subject i; mute_i is a vector containing the mutations happened at baseline, month 1, 4, etc., up to month $t-1$ for each subject i.

Table 6.9: Longitudinal analysis on each single CV and sliding windows of 3 CVs with moving step 2. For testing sliding windows, the first mutation in the window is shown.

	Index	mutations	Wald	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
Year 1												
single variant	8	X35I	0.049	0.046	0.036	0.040	0.040	0.040	0.040	0.040	0.040	0.040
	30	X123E	0.031	0.033	0.040	0.031	0.031	0.031	0.031	0.031	0.031	0.031
	47	X245M	0.032	0.066	0.074	0.077	0.077	0.077	0.077	0.077	0.077	0.077
3 variants	6	X30N	0.060	0.145	0.154	0.020	0.077	0.063	0.084	0.075	0.078	0.038
	43	X214L	0.311	0.271	0.495	0.037	0.244	0.214	0.275	0.251	0.264	0.070
	45	X219Q	0.049	0.071	0.176	0.070	0.096	0.096	0.123	0.118	0.128	0.121
Year 2												
single variant	3	X13I	0.017	0.012	0.009	0.008	0.008	0.008	0.008	0.008	0.008	0.008
	17	X54V	0.002	0.075	0.067	0.079	0.079	0.079	0.079	0.079	0.079	0.079
	27	X72V	0.000	0.324	0.319	0.320	0.320	0.320	0.320	0.320	0.320	0.320
	46	X179I	0.000	0.293	0.297	0.297	0.297	0.297	0.297	0.297	0.297	0.297
	53	X207Q	0.020	0.013	0.012	0.016	0.016	0.016	0.016	0.016	0.016	0.016
3 variants	1	X10I	0.017	0.027	0.040	0.763	0.028	0.117	0.044	0.072	0.056	0.048
	2	X10V	0.065	0.063	0.036	0.087	0.110	0.074	0.082	0.077	0.078	0.121
	3	X13I	0.077	0.070	0.029	0.257	0.235	0.146	0.171	0.148	0.149	0.238
	7	X30N	0.076	0.116	0.189	0.017	0.066	0.079	0.103	0.105	0.134	0.032
	52	X207E	0.041	0.049	0.052	0.894	0.040	0.212	0.078	0.136	0.101	0.066
	53	X207Q	0.073	0.070	0.045	0.187	0.126	0.081	0.085	0.081	0.081	0.131

In the longitudinal analysis with single common mutation, most mutations barely show any significant association with the change of HIV viral load and the ones with a p-value < 0.05 in any tests are shown in Table 6.9. The incapability of finding significant mutations might be due to the dilution effect, as it was assumed mutations had the same odds ratios (ORs) at several time points. We also tested on sliding windows of size 3 on common mutations with moving step 1. As we considered more mutations together and more time points, the number of significant mutations increased slightly.

6.4.2 Application to the GAW18 data

We carried out a genome-wide association scan for CVs and RVs with multivariate traits in the Genetic Analysis Workshop (GAW) 18 data, which was introduced in Chapter 2. In simulations we did not see a big difference between the simulation-based and permutation-based methods, so we mainly used simulation-based method to obtain p-values. We first focused on a subset of 157 independent samples. For CVs, we tested each single SNP, while for RVs, we tested sliding windows of size 20 with moving step 10. We fit the GEE model for a longitudinal analysis on

hypertension, HTN, over 4 exam points. The model was:

$$\begin{aligned} & (LogitPr(\text{HTN}(1)), LogitPr(\text{HTN}(2)), LogitPr(\text{HTN}(3)), LogitPr(\text{HTN}(4)))_i^T \\ = \quad & \mathbf{I}\psi_0 + \text{diag}\{\text{age(t)}\}_i\psi_1 + \text{diag}\{\text{gender}\}_i\psi_2 + \text{diag}\{\text{smoke(t)}\}_i\psi_3 + \text{diag}\{SNVs\}_i\beta, \end{aligned}$$

where ψ_h, h=1, 2, 3, is a vector of coefficients for covariates and γ is a vector of SNV effects at 4 exam time points. There was a suggested covariate, MEDBP, but because HTN was based on that, it usually caused convergence issues in longitudinal analysis. Thus we were not including it.

We first used the 200 replicates of the simulated HTN to assess our methods. To evaluate the Type I error, we tested on each of 1000 CVs from chromosome 15, which was the chromosome not used for simulating the phenotype. We obtained 1000 Type I errors and the summary is shown in Table 6.10. By all methods, the 90% confidence interval was roughly (0.02, 0.09) with means around 0.05 and variance 0.001. We also tested on 400 non-overlapping windows with each window containing 20 RVs. The means of the Type I error rates by the score test (0.014) and UminP test (0.019) were a little conservative, probably due to the small sample size issue, while those of the SPU tests were around 0.05. Overall, the Type I error seemed under control. To evaluate power, we tested 13 causal variants in MAP4 gene. We tested on 1) each of the 7 common variants with MAF$\geq$ 0.01, 2) 6 rare variants with 0<MAF< 0.01 combined, 3) all 13 variants combined. While the aSPU test was generally more powerful than the score test for testing CVs, it was the opposite for testing RVs. And for testing all 13 variants, SPU(1) had a power of 0.670 while the aSPU had a power of 0.575, much better than the score test (0.114) (Table 6.12). In fact it was interesting to notice that in this analysis power decreased with ϱ. This could be explained: the major benefit of the SPU tests is that, with a greater ϱ the tests put more weights on the associated variants while down-weighting the null ones. Here as all the variants tested were actually associated variants, which was an ideal situation, SPU(ϱ) test with a larger ϱ no longer had an obvious advantage.

Table 6.10: Summary of the Type I error rates of testing 1000 CVs on chromosome 15.

	Score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
0%	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10%	0.025	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020
20%	0.030	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025
30%	0.040	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
40%	0.040	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035
50%	0.045	0.040	0.040	0.040	0.040	0.040	0.040	0.045	0.040
60%	0.052	0.050	0.045	0.045	0.050	0.050	0.050	0.050	0.050
70%	0.060	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055
80%	0.070	0.065	0.070	0.065	0.065	0.065	0.065	0.065	0.066
90%	0.080	0.090	0.090	0.090	0.085	0.090	0.090	0.086	0.090
100%	0.175	0.220	0.260	0.265	0.275	0.260	0.250	0.235	0.280
mean	0.051	0.049	0.049	0.049	0.050	0.050	0.050	0.050	0.050
var	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 6.11: Summary of the Type I error rates of testing 400 windows of RVs on chromosome 15.

	Score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
0%	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000
10%	0.000	0.000	0.020	0.010	0.010	0.005	0.010	0.010	0.010
20%	0.000	0.005	0.025	0.020	0.020	0.015	0.015	0.015	0.020
30%	0.000	0.010	0.034	0.025	0.025	0.020	0.020	0.020	0.025
40%	0.000	0.010	0.040	0.030	0.030	0.025	0.026	0.025	0.030
50%	0.005	0.015	0.045	0.040	0.035	0.030	0.035	0.034	0.039
60%	0.010	0.020	0.053	0.045	0.045	0.040	0.040	0.040	0.045
70%	0.015	0.025	0.063	0.055	0.054	0.049	0.052	0.050	0.054
80%	0.020	0.030	0.075	0.069	0.070	0.060	0.064	0.060	0.065
90%	0.040	0.040	0.101	0.090	0.091	0.080	0.085	0.080	0.090
100%	0.190	0.083	0.317	0.210	0.330	0.235	0.250	0.220	0.215
mean	0.014	0.019	0.056	0.046	0.047	0.040	0.043	0.040	0.046
var	0.001	0.000	0.002	0.001	0.002	0.001	0.001	0.001	0.001

Table 6.12: The proportion of 200 replicates that detect the causal variants with a p-value$<$ 0.05 in MAP4.

	SNP	MAF	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
1	3_47956424	0.359	0.155	0.235	0.275	0.290	0.285	0.280	0.265	0.275	0.270
2	3_47957996	0.021	0.005	0.100	0.290	0.225	0.205	0.195	0.180	0.170	0.130
3	3_47958037	0.317	0.255	0.375	0.450	0.450	0.440	0.430	0.410	0.370	0.435
4	3_47973345	0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	3_48040283	0.025	0.070	0.530	0.740	0.700	0.675	0.635	0.625	0.605	0.630
6	3_48040284	0.021	0.015	0.120	0.290	0.245	0.230	0.225	0.220	0.200	0.175
7	3_48054461	0.074	0.050	0.040	0.025	0.025	0.025	0.025	0.025	0.030	0.025
8	8 RVs	$\leq$ 0.01	0.225	0.025	0.035	0.250	0.030	0.095	0.025	0.020	0.130
9	All 13 variants		0.114	0.115	0.670	0.420	0.360	0.350	0.355	0.340	0.570

For analyzing the real HTN traits of the 157 unrelated samples, we carried out GWAS on each single CV and sliding windows of RVs with size 20 and moving step 10. For the aSPU test whose p-value was based on simulation, we gradually increased the simulation number in order to get a more precise p-value, i.e. p-value> 0. We first used B=1000 for an initial run. And for those with a p-value ≤ 0.05, we re-ran the tests with B increased to 10^4. From the second run, the cut-of value was set to $5/B$ and for those with a p-value$\leq 5/B$, we re-ran the test with B increased tenfold ($B = 10^5, 10^6, 10^7$). When B reached 10^7, all the p-values were greater than 5×10^{-7}, and the iteration stopped.

The Manhattan plots for p-values of testing CVs and RVs by the score test and aSPU test are shown in Figure 6.3. For CVs, based on the score test, several variants passed the first suggested cut-off value, $\log_{10}(10^{-5})$ (Panel (a)), and two of them were close to the genome-wide suggested cut-off value, $\log_{10}(5 \times 10^{-8})$. The aSPU test seemed to be less powerful than the score test in the sense that less CVs achieved significant p-values (Panel (b)). The variants found to be significant by the score test and aSPU test were also very different. For example, the most significant regions found by the score test were approximately located at the end of chromosome 7 while the most significant ones by the aSPU test were at the beginning of chromosome 5. This inconsistency was also found in Table 6.15, which displays the p-values of the SPU and aSPU tests for the top 10 CVs with the most significant p-values of the score test. While these 10 CVs had p-values of around 10^{-7} by the score test, their p-values by the aSPU test were all greater than 10^{-3}. For RVs, the score test detected several significant variants while no variants were shown to reach the suggestive line by the aSPU test.

To summarize, one immediate conclusion was that the aSPU test was less powerful than the score test in the genome-wide scan with the GAW18 data. Although disappointing, this did not completely contradict to our previous results. In the simulation studies with CVs, the aSPU test was always powerful than the score test. However, when testing on CVs in the MAP4 gene with 200 replicates of HTN, we observed that for SNP 3_48054461 with MAF 0.074, the score test had a power of 0.05 while the aSPU test had a power of 0.025. In the simulation studies of testing RVs (Table 6.5), when only one out of three traits was associated with the causal RVs (S9), the score test and UminP test were more powerful than the SPU tests.

Also, with 200 replicates of HTN to test on 6 RVs in the MAP4 gene, the score test had a power of 0.225 while the aSPU test only yielded a power of 0.130. As a reflection, we realized that in our simulations for testing CVs, we had 17 SNPs to be tested, which could be the situation that the score test lost its power for large degree of freedom. To test one SNP at a time, the SPU test might not outperform the score test. To confirm our suspect, we carried out another simulation with S9 but testing only on 1 SNP, and the 3 causal SNPs had opposite effect size (0.6, -0.5, 0.4) (Table 6.13). In this situation the score test had a greater power (0.736) than the aSPU test (0.690). This could serve as an illustration to our results in the GAW18 data analysis. This result suggests that the SPU tests may not win the score test in power for single SNP analysis. On contrary, the SPU tests may be more powerful in testing SNP-sets or genes.

The proportion of CVs and RVs having a p-value$< \alpha$ are shown in Table 6.14 for different α values. Overall for CVs there was no obvious inflation observed with $\alpha = 0.05$, approximately 5% of all CVs were rejected for all methods. However, as α decreased to 0.0005 (5E-05), the score test had a higher rejection rate (0.0093%) than expected. For RVs, the score test was very conservative, with only 0.34% out of them having p-values< 0.05. The SPU tests could maintain the rejection rates around 0.05.

Figure 6.3: Manhattan plots (a) for CV by the score test, (b) for CV by the aSPU test (c) for RV by the score test, (d) for RV by the aSPU test

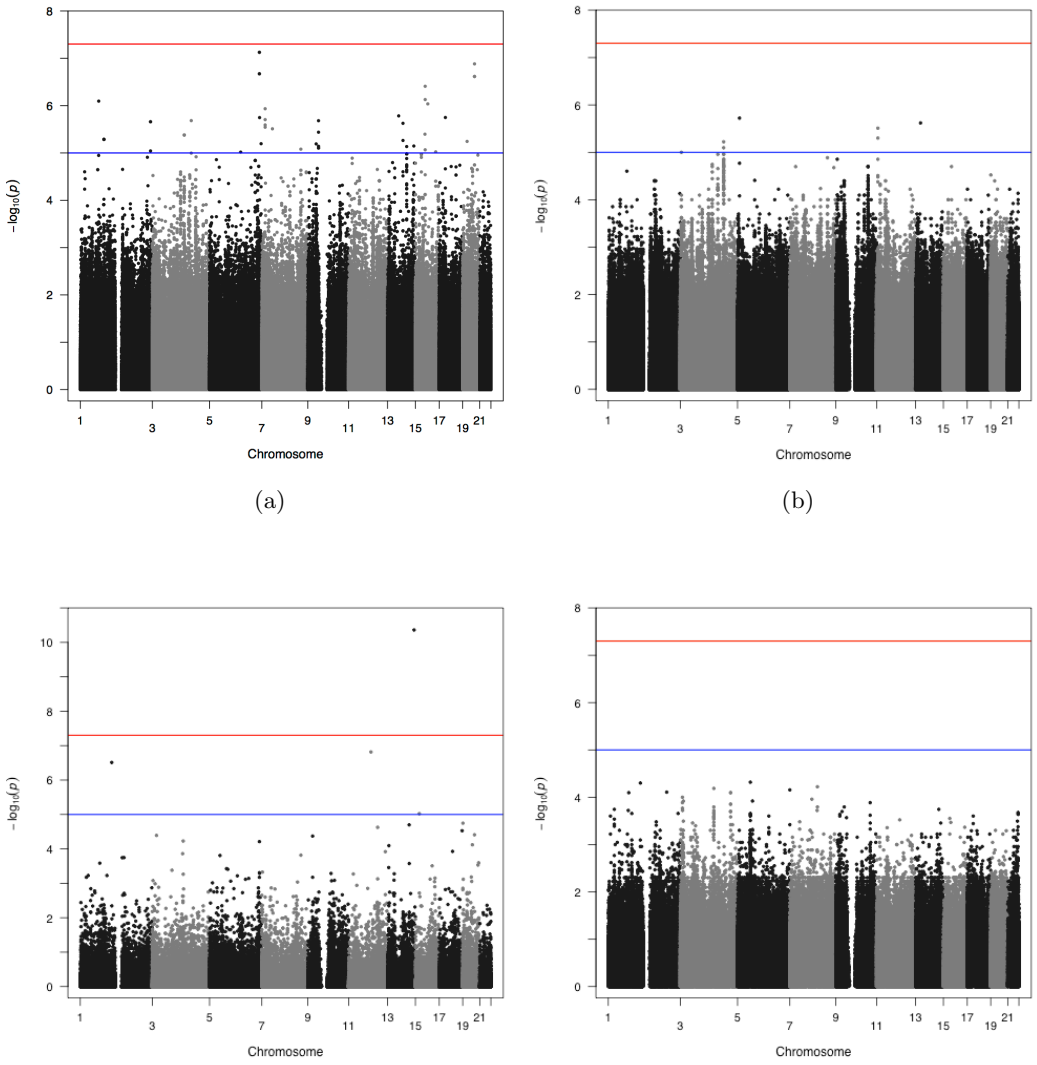


Table 6.13: An illustration simulation with S9 testing on 1 SNP. The effective size of 3 causal SNPs is (0.6, -0.5, 0.4) and are not included in the analysis.

	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(6)	SPU(7)	SPU(8)	SPU(∞)	aSPU
S	0.715	0.729	0.386	0.658	0.627	0.666	0.649	0.660	0.651	0.658	0.649	0.634
B	0.713	0.730	0.380	0.659	0.625	0.668	0.647	0.665	0.649	0.660	0.652	0.633

Table 6.14: The proportion of CVs and RVs having a p-value$< \alpha$

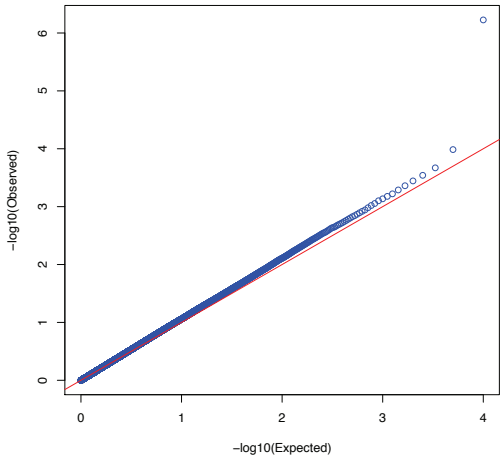
	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
					CVs				
$\alpha = 5E-2$	0.052209	0.051258	0.049379	0.050199	0.050382	0.050516	0.050645	0.049954	0.050931
$\alpha = 5E-3$	0.005321	0.004648	0.004389	0.004266	0.004237	0.004243	0.004247	0.004435	0.004337
$\alpha = 5E-4$	0.000625	0.000399	0.000291	0.000277	0.000283	0.000287	0.000296	0.000359	0.000352
$\alpha = 5E-5$	0.000093	0.000061	0.000015	0.000021	0.000023	0.000024	0.000025	0.000034	0.000029
$\alpha = 5E-6$	0.000011	0.000033	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001
					RVs				
$\alpha = 5E-2$	0.003444	0.011292	0.050995	0.043624	0.043460	0.041838	0.042695	0.042069	0.043496
$\alpha = 5E-3$	0.000535	0.000562	0.004176	0.002612	0.002641	0.002506	0.002578	0.002626	0.002157
$\alpha = 5E-4$	0.000203	0.000090	0.000322	0.000185	0.000175	0.000130	0.000122	0.000126	0.000157
$\alpha = 5E-5$	0.000164	0.000065	0.000018	0.000007	0.000005	0.000005	0.000004	0.000002	0.000004
$\alpha = 5E-6$	0.000149	0.000063	0.000002	0.000002	0.000002	0.000002	0.000002	0.000002	0.000002

Table 6.15: P-values for 10 variants with the most significant p-values

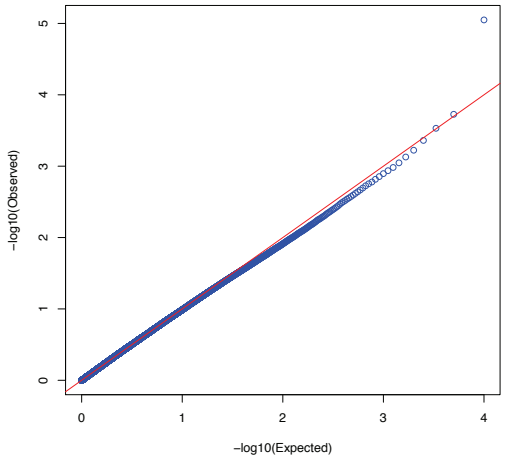
	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
19_41041164	1.312E-07	9.897E-02	2.116E-01	8.330E-02	9.809E-02	9.018E-02	9.339E-02	8.821E-02	1.075E-01
5_172381091	2.134E-07	1.496E-03	1.457E-01	3.840E-02	1.333E-02	6.488E-03	3.986E-03	1.162E-03	1.770E-03
5_172382836	7.477E-08	1.235E-03	1.656E-01	3.754E-02	1.214E-02	5.726E-03	3.523E-03	1.029E-03	1.564E-03
5_172383219	7.477E-08	1.283E-03	1.659E-01	3.766E-02	1.262E-02	6.046E-03	3.696E-03	1.044E-03	1.584E-03
5_172384088	2.116E-07	1.223E-03	1.428E-01	3.237E-02	1.161E-02	5.611E-03	3.491E-03	9.720E-04	1.480E-03
15_53946467	3.902E-07	6.162E-02	1.578E-01	7.447E-02	6.176E-02	5.545E-02	5.284E-02	5.071E-02	6.714E-02
15_53958050	7.452E-07	6.084E-02	1.426E-01	7.083E-02	5.844E-02	5.289E-02	5.043E-02	5.059E-02	6.667E-02
19_41034724	2.422E-07	7.619E-02	1.630E-01	6.980E-02	7.310E-02	6.973E-02	7.052E-02	6.722E-02	8.767E-02
19_41038081	2.422E-07	7.617E-02	1.628E-01	6.950E-02	7.282E-02	6.944E-02	7.035E-02	6.714E-02	8.738E-02
19_41040399	2.422E-07	7.665E-02	1.637E-01	7.026E-02	7.374E-02	7.014E-02	7.103E-02	6.779E-02	8.808E-02

To use a larger sample size, we have to deal with familial correlations in the GAW18 data. Based on our previous studies of single quantitative trait, we found that with a weak heritability, adding principal components (PCs) in regression model, abbreviated as PCR, could effectively account for familial correlations. Here we were curious about whether adding PCs in the GEE model could adequately adjust for the familial correlatedness in multivariate trait analysis. We first carried out an analysis of HTN at each time point. The top 10 PCs of the IBS matrix, which was estimated based on 31544 pruned CVs, were added to covariates (Figure 6.4). We can see that the p-values were not inflated, as most of the points are aligned along the 45 degree line and λ's are around 1.

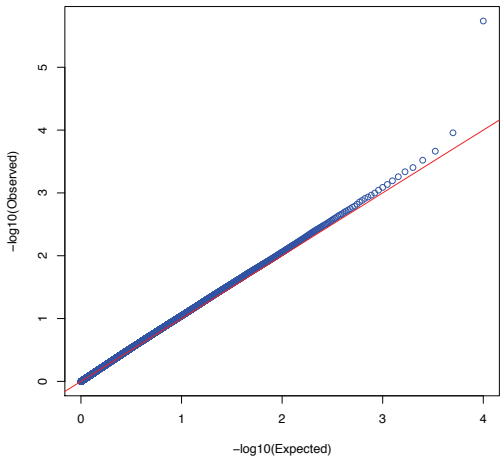
Figure 6.4: Q-Q plots of p-values from GWAS by PLINK



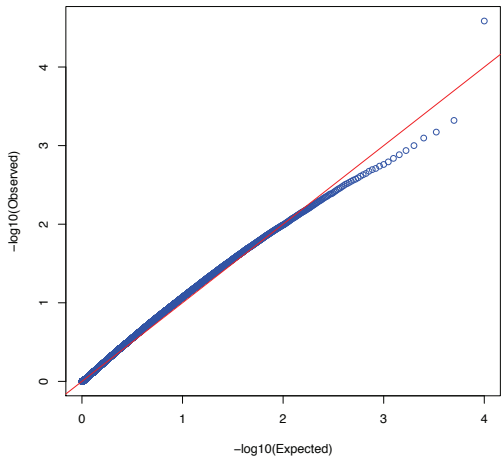
(a) Exam 1



(b) Exam 2



(c) Exame 3



(d) Exam 4

Further, with all 855 samples included, we tested on 1000 CVs from chromosome 15 with 200 replicates of the simulated HTN (Table 6.16). The means of the Type I error rates were between 0.07 to 0.085, still showing some inflation. The Type I error rates of testing RVs were even more inflated for each method. For example, by the score test, the mean of the Type I error was 0.086 while by the aSPU test it was

0.117 (Table 6.17). This was highly suspected to be caused by the correlations among family members. Further investigations are needed to understand why PCs failed to adjust for familial correlations. It might be because now we used the GEE model, and the covariance among samples were not modeled as LMMs or generalized linear model (GLM). Hence the linear combination of PCs can not capture the covariance structure among family members. It would be helpful to understand this result by fitting a LMM for the multivariate analysis using all samples.

Table 6.16: Summary of the Type I error rates for testing 10000 CVs with all samples

	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
0%	0.010	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10%	0.030	0.025	0.020	0.020	0.025	0.025	0.025	0.025	0.025
20%	0.035	0.030	0.030	0.030	0.030	0.030	0.030	0.035	0.030
30%	0.045	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040
40%	0.050	0.045	0.050	0.050	0.050	0.050	0.050	0.050	0.050
50%	0.057	0.055	0.060	0.060	0.060	0.060	0.060	0.060	0.060
60%	0.065	0.070	0.075	0.075	0.070	0.072	0.070	0.070	0.075
70%	0.080	0.090	0.095	0.095	0.095	0.095	0.095	0.095	0.095
80%	0.095	0.120	0.130	0.130	0.130	0.130	0.126	0.125	0.130
90%	0.140	0.180	0.210	0.200	0.191	0.185	0.185	0.175	0.190
100%	0.570	0.660	0.750	0.730	0.725	0.725	0.715	0.685	0.725
mean	0.072	0.083	0.089	0.089	0.087	0.087	0.086	0.084	0.088
var	0.003	0.006	0.008	0.007	0.007	0.007	0.006	0.006	0.007

Table 6.17: Summary of the Type I error rates of testing 150 windows of RVs with all samples

	score	UminP	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(∞)	aSPU
0%	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10%	0.011	0.052	0.016	0.040	0.027	0.029	0.027	0.026	0.032
20%	0.028	0.074	0.032	0.055	0.039	0.040	0.040	0.040	0.045
30%	0.040	0.088	0.042	0.065	0.045	0.050	0.050	0.048	0.055
40%	0.053	0.110	0.050	0.076	0.058	0.065	0.059	0.059	0.068
50%	0.070	0.131	0.056	0.091	0.070	0.076	0.073	0.072	0.084
60%	0.084	0.158	0.070	0.118	0.091	0.099	0.092	0.085	0.100
70%	0.107	0.186	0.095	0.155	0.121	0.126	0.117	0.113	0.125
80%	0.125	0.211	0.136	0.200	0.176	0.176	0.159	0.151	0.169
90%	0.188	0.261	0.189	0.301	0.241	0.260	0.245	0.218	0.265
100%	0.388	0.571	0.540	0.512	0.512	0.495	0.490	0.465	0.528
mean	0.086	0.149	0.092	0.134	0.111	0.114	0.107	0.100	0.117
var	0.006	0.009	0.009	0.012	0.011	0.010	0.009	0.007	0.010

6.5 Discussion

In this chapter we have extended the SPU tests to multivariate trait analysis under the GEE framework. There were several reasons we advocated the SPU tests and GEE model. As briefed in Pan et al. (2013), the SPU tests allowed many possible and informative choices of weights, $U^{\varrho-1}$, by varying the power parameter ϱ. With a larger ϱ, it gave larger weights to associated variants and smaller weights to the null ones. Thus, its performance deteriorated more slowly than many other methods when a lot of null SNVs were present in the test. Moreover, depending on the minimum p-value of the SPU(ϱ) test, the aSPU test was a convenient data-adaptive test to summarize all the SPU tests. Besides, as the GEE model is actually a marginal model, it is a much easier to obtain the score vector than by using GLMM.

We carried out simulation studies on global testing of CVs and RVs with multiple binary traits in different scenarios, and compared the results with univariate trait analysis. The simulation results were consistent with our intuition, that when there was only one trait associated with the causal SNVs, multivariate trait analysis did not necessarily win over univariate trait analysis, but as the number of associated traits increased, multivariate trait analysis became more powerful. We also noticed that when testing a group of CVs, the SPU tests generally outperformed the score and UminP tests. For RVs, when the causal RVs only contributed to one out of three traits, the SPU tests were not as powerful as the score and UminP tests; but for other situations, the SPU tests were still the winners. More, we implemented the score test, UminP test, and SPU tests on cross-sectional multivariate trait analysis and longitudinal analysis with the NvR study. We found several significant common and rare viral mutations significantly associated with disease progression in multiple visits. In future, it is more of interest to perform prognostic analysis to study how genotypic drug resistance is associated with clinical outcome in the following visits, such as whether the HIV RNA level is above 400 copies/ml, whether CD4 counts are below 200 copies/ml, etc. We also carried out a genome-wide association testing with the GAW18 data. For CVs, we tested on each single variant; for RVs, we tested on sliding windows of 20 variants. We observed a power loss of the aSPU test compared to the score test for both CVs and RVs. It was possibly because we were only testing one SNP at a time, while the advantage of the SPU tests lied in

testing with a large number of null variants. For single SNP analysis, the SPU tests might not be as powerful as the score test and the UminP test. Further analysis for SNP-sets or genes is of interest and necessary.

In this chapter, besides the simulation-based method, we also proposed the use of a permutation-based method to estimate p-values. Since we did not need to re-fit the GEE in each permutation, this method was relatively fast. However, sometimes it may be preferred to use a "non-parametric" permutation method, especially when the sample size is extremely small, i.e. n$<$ 10. [84]. We have adopted the following strategy: we can estimate the event rate p_i for each subject i as well as the correlation matrix, $R_w(\alpha)$, by fitting Y_i on covariates (Z_i), and calculate the statistic (T). Then based on $\hat{p}_i$ and $R_w(\hat{\alpha})$, we can draw $Y_i^{(b)}$ as a vector of correlated binary samples [85], and re-fit $Y_i^{(b)}$ on X_i and Z_i to obtain $T^{(b)}$. The p-value is calculated as the proportion of $I(|T^{(b)}| > |T|)$, where I is an indicator function. The problem with this method is that sometimes $\hat{p}_i$ is not compatible with the correlation matrix $R_w(\hat{\alpha})$ for constructing a joint probability matrix, thus unable to be used for drawing the samples.

This study is worth to be explored further. Firstly, with some more simulations studies with multiple quantitative traits, we confirmed that the advantage of the aSPU tests is obvious when there are a large number of null SNPs present in the tests. However, the aSPU test is not necessarily more powerful than the score test when the number of traits increased with small number of null SNPs. To improve the performance of the aSPU test in multivariate trait analysis, the test statistic needs to adapt to traits in addition to the genetic variants. Namely, the improved SPU and aSPU test statistics are

$$\begin{aligned} T_{SPU} &= \sum_{m=1}^{k} (\sum_{j=1}^{p} U_{mj}^{\varrho_1})^{\varrho_2}, \ \varrho_1 = 1, 2, ..., \infty, \ \varrho_2 = 1, 2, ..., \infty, \\ T_{aSPU} &= \min_{\varrho_1, \varrho_2} P_{SPU(\varrho_1, \varrho_2)}, \end{aligned}$$

where m denotes the index of traits and j denotes the index of SNVs. Further studies are needed for the new SPU and aSPU tests . Secondly, more simulations with a much smaller sample size may be needed, to compare the performance of simulation-based and permutation-based methods. Another compelling topic is to extend the

SPU tests to handle data with various sources of correlations. For example in the GAW18 data, there are both familial correlations as well as between-trait correlations. A direct way may be modeling multivariate traits for all members of one family as one cluster under the GEE model. In spite of that, more computational issues may be encountered, especially for binary traits. Lastly but challengingly, since the GEE does not require specification of traits distribution, it is potential to analyze mixed traits. So far, the available packages unnecessarily force to use the same link function on all traits, possibly resulting in a reduced power of the SPU tests. Further works are needed to develop programs to estimate the parameters in the correlation matrix, $R_w(\alpha)$, and the marginal variance, $v(\mu_{im})$, with different link function for different responses.

Chapter 7

Conclusions and Future Work

In this thesis we targeted on two major issues in genetic association testing with next-generation sequencing data: controlling false positives and enhancing true positives. Regarding the first issue, we compared the strategies for adjusting for known and unknown relatedness among samples, including population structures, familial correlatedness, cryptic relatedness and environmental factors. For power boosting, we developed a class of competent tests, the SPU tests under the GEE model, for testing multiple traits on a group of variants.

As pointed out by Yu at al. (2005), it is almost impossible to collect samples completely independent. Population structure is one of the primary sources for genetic correlation, which spoils the association testing as a confounder. The principal component regression (PCR), which includes the top principal components (PCs) of dense genetic data in the regression model [63, 25], is one of the most appealing methods for adjusting for population stratification. In Chapter 3 we applied the PCR on three selected regions of low frequency variants (LFVs) and rare variants (RVs) using the 1000 Genomes Project data with European and African samples. We showed that using the top 10 PCs constructed by principal components analysis (PCA) based on 10000 randomly selected CVs/LFVs was largely effective in controlling the Type I error under the nominal level, while PCs of 10000 RVs were less effective. The power was largely unaffected when the associated variants were independent of the population structure. Nevertheless, we also observed some power loss possibly caused by over-adjustment.

To tackle some tightly related but unclear questions, in Chapter 4, we compared two dimension reduction methods, namely PCA and SDR, for constructing PCs to adjust for population stratification in a fine scale. Variants of different MAFs and with or without pruning were also studied for constructing a similarity matrix. We showed that, in general, SDR had a better performance than PCA. Its benefit was pronounced both in association testing, especially on RVs, and uncovering subgroups. PCs based on all variants and all CVs constructed by SDR could consistently control the Type I error around the nominal level 0.05. Moreover, in the presence of a local non-genetic risk, while it was relatively easy to control the inflation when testing CVs, it was difficult to achieve the same effectiveness when testing RVs. Lastly, we also validated the reasons for why RVs performed worse than expected in the use of adjusting for population structure.

Another method commonly used for addressing correlated samples and of arising interest is Linear mixed models (LMM). It was recently reported to outperform PCR as a more powerful and general method to adjust for various population structures. In Chapter 5, we studied PCR and LMM in handling cryptic relatedness, familiar correlations and environmental factors. By putting PCR under a probabilistic PCA framework, we showed that the PCR was an approximation to LMM. The LMM modeled the correlation by a random effect, yet PCR captured the correlation structure by estimating the fixed effects of the included PCs. While the LMM was mostly efficacious in adjusting for a population structure and more powerful, the PCR could be advantageous to accommodate environmental confounders. Based on these inspections, we proposed a hybrid model combining these two models. Simulation results validated the benefits of the proposed method.

Beside population structures, another topic discussed here was how to improve power in detecting disease loci in association testing with multiple traits. In Chapter 6, by extending the SPU tests proposed by Pan et al. (2013) for independent samples, we proposed a multivariate trait analysis using the SPU tests under the generalized estimation equations (GEE). With simulated binary traits, we showed that as the number of associated traits increased, the SPU tests had a more impressive performance than the score test and the UminP test. With these methods, we analyzed the association between the viral RNA mutations with drug resistance in the NvR study, where we found several interesting mutations across different visits. We also carried out the genetic association analysis on hypertension in the GAW18 data. However, we did not observe any genome-wide significant SNVs, and the aSPU test lost its strength in the single SNP analysis. To combine the first and second theme in our thesis, our original plan was to use PCs as covariates to take into account the familial correlations in the GAW18 data. Yet we found that the PCs were insufficient to adjust for the correlatedness in the multivariate trait analysis, so we only used 157 unrelated samples.

As discussed in Chapter 3 to 5, the determination of the number of PCs to capture the population structure is a hard yet important problem in dimension reduction on massive genetic data. While using too few PCs may be insufficient to capture the population structure, using too many can cause over-adjustment problem. In our thesis, we used the Tracy-Widom test [63] for PCA and the heuristic

method [86] for SDR. The scree plots of eigenvalues, or an ANOVA test may also help to yield general idea. However, their performances are not always reliable, and how to determine the number of PCs is still an exciting topic for future work. For multivariate trait analysis, the class of SPU tests is a prospective method because its performance is less impaired by increasing number of null variants than other methods. This is especially beneficial in analyzing sequencing data. Based on our current findings, each SNP only bears a small increased disease risk, thus testing on a set of SNVs or a whole gene attracts growing attention. In our implementation we only did single SNP analysis for CVs, from which we did not see a remarkable advantage of the SPU tests. A SNP-set or gene-set analysis is definitely of interest later.

References

[1] J. Asimit and E. Zeggini. Rare variant association analysis methods for complex traits. *Annual Review of Genetics*, 44:293–308, 2010.

[2] V. Bansal, O. Libiger, A. Torkamani, and N.J. Schork. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785, 2010.

[3] W. Bodmer and C. Bonilla. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, 40(6):695–701, 2008.

[4] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, and T.A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[5] B.M. Henn, S. Gravel, A. Moreno-Estrada, S. Acevedo-Acevedo, and C.D. Bustamante. Fine-scale population structure and the era of next-generation sequencing. *Human Molecular Genetics*, 19(R2):R221–R226, 2010.

[6] T.M. Baye, H. He, L. Ding, B.G. Kurowski, X. Zhang, and L.J. Martin. Population structure analysis using rare and common functional variants. In *BMC proceedings*, volume 5, page S8. BioMed Central Ltd, 2011.

[7] H. Siu, L. Jin, and M. Xiong. Manifold learning for human population structure studies. *PloS One*, 7(1):e29901, 2012.

[8] S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619, 2011.

[9] W. Pan. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic Epidemiology*, 35(4):211–216, 2011.

[10] J.J. Goeman, S.A. Van De Geer, and H.C. Van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

[11] L.C. Kwee, D. Liu, X. Lin, D. Ghosh, and M.P. Epstein. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397, 2008.

[12] M.C. Wu, P. Kraft, M.P. Epstein, D.M. Taylor, S.J. Chanock, D.J. Hunter, and X. Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

[13] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[14] J. Wessel and N.J. Schork. Generalized genomic distance–based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79(5):792–806, 2006.

[15] B.M. Neale, M.A. Rivas, B.F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S.M. Purcell, K. Roeder, and M.J. Daly. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322, 2011.

[16] W. Pan. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497–507, 2009.

[17] B.E. Madsen and S.R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, 2009.

[18] A.L. Price, G.V. Kryukov, P.I.W. de Bakker, S.M. Purcell, J. Staples, L.J. Wei, and S.R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics*, 86(6):832, 2010.

[19] D.J. Liu and S.M. Leal. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS genetics*, 6(10):e1001156, 2010.

[20] D.Y. Lin and Z.Z. Tang. A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367, 2011.

[21] G. Sirugo, B.J. Hennig, A.A. Adeyemo, A. Matimba, M.J. Newport, M.E. Ibrahim, K.K. Ryckman, A. Tacconelli, R. Mariani-Costantini, G. Novelli, et al. Genetic studies of African populations: an overview on disease susceptibility and response to vaccines and therapeutics. *Human Genetics*, 123(6):557–598, 2008.

[22] K. Bryc, A. Auton, M.R. Nelson, J.R. Oksenberg, S.L. Hauser, S. Williams, A. Froment, J.M. Bodo, C. Wambebe, S.A. Tishkoff, et al. Genome-wide patterns of population structure and admixture in west Africans and African Americans. *Proceedings of the National Academy of Sciences*, 107(2):786–791, 2010.

[23] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 2004.

[24] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

[25] C. Wu, A. DeWan, J. Hoh, and Z. Wang. A comparison of association methods correcting for population stratification in case–control studies. *Annals of Human Genetics*, 75(3):418–427, 2011.

[26] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[27] W. Guan, L. Liang, M. Boehnke, and G.R. Abecasis. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genetic Epidemiology*, 33(6):508–517, 2009.

[28] S. Lee, F.A. Wright, and F. Zou. Control of population stratification by correlation-selected principal components. *Biometrics*, 67(3):967–974, 2011.

[29] J. Yu, G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2005.

[30] K. Zhao, M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, et al. An arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1):e4, 2007.

[31] H.M. Kang, N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[32] H.M. Kang, J.H. Sul, N.A. Zaitlen, S. Kong, N.B. Freimer, C. Sabatti, E. Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

[33] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, 2012.

[34] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[35] Greg Gibson. The environmental contribution to gene expression profiles. *Nature Reviews Genetics*, 9(8):575–581, 2008.

[36] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

[37] Jennifer Listgarten, Carl Kadie, Eric E Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 107(38):16465–16470, 2010.

[38] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44(3):243–246, 2012.

[39] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[40] Wei Pan, Peng Wei, and Xiaotong Shen. An adaptive test on a high-dimensional parameter with application to detect disease association with rare variants. *Submmitted*, 2013.

[41] D.M. Altshuler, E.S. Lander, L. Ambrogio, T. Bloom, K. Cibulskis, T.J. Fennell, S.B. Gabriel, D.B. Jaffe, E. Shefler, C.L. Sougnez, et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[42] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. De Bakker, M.J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[43] G. Perez, S. MacArthur, R.D.and Walmsley, J.A. Baxter, C. Mullin, and J.D. Neaton. A randomized clinical trial comparing nelfinavir and ritonavir in patients with advanced HIV disease. *HIV Clinical Trial*, 5(1):7–18, 2004.

[44] A.L. Price, N.A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[45] A.B. Lee, D. Luca, L. Klei, B. Devlin, and K. Roeder. Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology*, 34(1):51–59, 2009.

[46] F. Han and W. Pan. A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1):42–54, 2010.

[47] S. Morgenthaler and W.G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums

test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56, 2007.

[48] B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[49] K.N. Conneely and M. Boehnke. So many correlated tests, so little time! rapid adjustment of p-values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168, 2007.

[50] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 1994.

[51] W. Pan and X. Shen. Adaptive tests for association analysis of rare variants. *Genetic Epidemiology*, 35(5):381–388, 2011.

[52] N. Tintle, H. Aschard, I. Hu, N. Nock, H. Wang, and E. Pugh. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 genomes project exon sequencing data in unrelated individuals: summary results from group 7 at genetic analysis workshop 17. *Genetic Epidemiology*, 35(S1):S56–S60, 2011.

[53] C. Dering, C. Hemmelmann, E. Pugh, and A. Ziegler. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genetic epidemiology*, 35(S1):S12–S17, 2011.

[54] J.A. Tennessen, A.W. Bigham, T.D. OConnor, W. Fu, E.E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, 2012.

[55] S.C. Heath, I.G. Gut, P. Brennan, J.D. McKay, V. Bencko, E. Fabianova, L. Foretova, M. Georges, V. Janout, M. Kabesch, et al. Investigation of the fine structure of european populations with applications to disease association studies. *European Journal of Human Genetics*, 16(12):1413–1429, 2008.

[56] M.C. Babron, M. de Tayrac, D.N. Rutledge, E. Zeggini, and E. Génin. Rare and low frequency variant stratification in the UK population: Description and impact on association tests. *PloS One*, 7(10):e46519, 2012.

[57] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[58] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[59] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[60] D. Luca, S. Ringquist, L. Klei, A.B. Lee, C. Gieger, H. Wichmann, S. Schreiber, M. Krawczak, Y. Lu, A. Styche, et al. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *The American Journal of Human Genetics*, 82(2):453–463, 2008.

[61] Carrie B Moore, John R Wallace, Alex T Frase, Sarah A Pendergrass, Marylyn D Ritchie, et al. Using biobin to explore rare variant population stratification. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 332–343, 2012.

[62] M. Bouaziz, C. Ambroise, and M. Guedj. Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide association studies. *PLoS One*, 6(12):e28845, 2011.

[63] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[64] Michael P Epstein, Andrew S Allen, and Glen A Satten. A simple and improved correction for population stratification in case-control studies. *The American Journal of Human Genetics*, 80(5):921–930, 2007.

[65] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, 2008.

[66] Chaolong Wang, Sebastian Zöllner, and Noah A Rosenberg. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genetics*, 8(8):e1002886, 2012.

[67] Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, 2008.

[68] Kai Wang and Yingwei Peng. An analytical comparison of the principal component method and the mixed effects model for genetic association studies. *Statistics in Medicine*, 00:1–9, 2012.

[69] H. Lan, J.P. Stoehr, S.T. Nadler, K.L. Schueler, B.S. Yandell, and A.D. Attie. Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics*, 164(4):1607–1614, 2003.

[70] Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

[71] Paul F OReilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

[72] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[73] G.M. Fitzmaurice and N.M. Laird. A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1):141–151, 1993.

[74] Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9):1066–1071, 2012.

[75] K.Y. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[76] J. Liu, Y. Pei, C.J. Papasian, and H.W. Deng. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genetic Epidemiology*, 33(3):217–227, 2009.

[77] Ming-Huei Chen, Xuan Liu, Fengrong Wei, Martin G Larson, Caroline S Fox, Ramachandran S Vasan, and Qiong Yang. A comparison of strategies for analyzing dichotomous outcomes in genome-wide association studies with general pedigrees. *Genetic Epidemiology*, 35(7):650–657, 2011.

[78] C. Lange, E.K. Silverman, X. Xu, S.T. Weiss, and N.M. Laird. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics*, 4(2):195–206, 2003.

[79] D. Podzamczer, E. Ferrer, E. Consiglio, J.M. Gatell, P. Perez, J.L. Perez, E. Luna, A. González, E. Pedrol, L. Lozano, et al. A randomized clinical trial comparing nelfinavir or nevirapine associated to zidovudine/lamivudine in HIV-infected naive patients (the combine study). *Antiviral Therapy*, 7(2):81–90, 2002.

[80] Wei Pan. On the robust variance estimator in generalised estimating equations. *Biometrika*, 88(3):901–906, 2001.

[81] A. Albert and JA Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[82] Charles Kooperberg and Michael LeBlanc. Increasing the power of identifying gene× gene interactions in genome-wide association studies. *Genetic Epidemiology*, 32(3):255–263, 2008.

[83] W. Zhu and H. Zhang. Why do we test multiple traits in genetic association studies? *Journal of the Korean Statistical Society*, 38(1):1–10, 2009.

[84] Marti J Anderson. Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(3):626–639, 2001.

[85] Friedrich Leisch, Andreas Weingessel, and Kurt Hornik. On the generation of correlated artificial binary data. *SFB Adaptive Information Systems and Modelling in Economics and Management Science*, 1998.

[86] Ann B Lee, Diana Luca, and Kathryn Roeder. A spectral graph approach to discovering genetic ancestry. *The Annals of Applied Statistics*, 4(1):179, 2010.

[87] DY Lin and D Zeng. Correcting for population stratification in genomewide association studies. *Journal of the American Statistical Association*, 106(495):997–1008, 2011.

Appendix A

Supplimentary Results for Chapter 4

Table A.1: Fst statistics calculated between every pair of subgroups based on all pruned CVs from chromosomes 1-22.

	GBR	FIN	PUR	CEU	MXL	TSI	PUR2	YRI	LWK	ASW
GBR	0.000	0.007	0.011	0.004	0.022	0.008	0.028	0.163	0.149	0.107
FIN	0.007	0.000	0.015	0.010	0.026	0.017	0.032	0.168	0.154	0.113
PUR	0.011	0.015	0.000	0.015	0.016	0.015	-0.084	0.133	0.119	0.083
CEU	0.004	0.010	0.015	0.000	0.024	0.004	0.031	0.159	0.150	0.107
MXL	0.022	0.026	0.016	0.024	0.000	0.025	0.029	0.147	0.134	0.092
TSI	0.008	0.017	0.015	0.004	0.025	0.000	0.031	0.155	0.145	0.104
PUR2	0.028	0.032	-0.084	0.031	0.029	0.031	0.000	0.113	0.100	0.067
YRI	0.163	0.168	0.133	0.159	0.147	0.155	0.113	0.000	0.013	0.014
LWK	0.149	0.154	0.119	0.150	0.134	0.145	0.100	0.013	0.000	0.012
ASW	0.107	0.113	0.083	0.107	0.092	0.104	0.067	0.014	0.012	0.000

Table A.2: Fst statistics calculated between every pair of subgroups based on all pruned RVs from chromosomes 1-22.

	GBR	FIN	PUR	CEU	MXL	TSI	PUR2	YRI	LWK	ASW
GBR	0.000	0.003	0.003	0.001	0.020	0.002	0.004	0.009	0.009	0.003
FIN	0.003	0.000	0.005	0.003	0.021	0.004	0.006	0.011	0.011	0.005
PUR	0.003	0.005	0.000	0.003	0.003	0.003	-0.101	0.008	0.008	0.005
CEU	0.001	0.003	0.003	0.000	0.020	0.001	0.004	0.008	0.009	0.003
MXL	0.020	0.021	0.003	0.020	0.000	0.020	0.005	0.018	0.018	0.009
TSI	0.002	0.004	0.003	0.001	0.020	0.000	0.003	0.008	0.008	0.003
PUR2	0.004	0.006	-0.101	0.004	0.005	0.003	0.000	0.008	0.008	0.004
YRI	0.009	0.011	0.008	0.008	0.018	0.008	0.008	0.000	0.005	0.002
LWK	0.009	0.011	0.008	0.009	0.018	0.008	0.008	0.005	0.000	0.004
ASW	0.003	0.005	0.005	0.003	0.009	0.003	0.004	0.002	0.004	0.000

Table A.3: Clustering results without AMR samples

		w/o pruning				w/ pruning				10000 pruned			
		all	CVs	LFVs	RVs	all	CV s	LFVs	RVs	all	CVs	LFVs	RVs
SDR	best cluster #	10	12	19	8	9	17	14	10	10	27	19	10
	best RI	0.963	0.959	0.925	0.895	0.900	0.873	0.894	0.940	0.935	0.866	0.905	0.895
	best aRI	0.865	0.842	0.672	0.600	0.615	0.410	0.515	0.765	0.757	0.324	0.585	0.594
	RI at 10 clusters	0.929	0.950	0.912	0.895	0.848	0.821	0.837	0.940	0.809	0.812	0.848	0.870
	aRI at 10 clusters	0.754	0.833	0.672	0.600	0.472	0.426	0.443	0.786	0.437	0.450	0.518	0.559
PCA	best cluster #	9	11	12	12	10	10	10	18	9	12	18	12
	best RI	0.921	0.927	0.893	0.787	0.878	0.908	0.887	0.831	0.868	0.897	0.870	0.803
	best aRI	0.718	0.692	0.536	0.295	0.496	0.618	0.542	0.347	0.509	0.546	0.384	0.322
	RI at 10 clusters	0.884	0.894	0.699	0.707	0.819	0.790	0.790	0.738	0.699	0.837	0.721	0.784
	aRI at 10 clusters	0.623	0.654	0.300	0.316	0.440	0.358	0.411	0.337	0.291	0.485	0.313	0.311

Table A.4: Clustering result based on PCs of SDR with all pruned CVs.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
CEU	3	11	18	42	0	0	8	8	0	0	0	0	0	0	0	0	0
FIN	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0
GBR	23	11	4	3	2	0	0	0	0	0	0	0	0	0	0	0	0
TSI	0	0	2	24	0	0	48	0	18	0	0	0	0	0	0	0	0
MXL	0	0	0	1	16	0	0	0	0	0	0	0	0	0	0	0	0
PUR	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
YRI	0	0	0	0	0	0	0	0	0	19	13	25	4	17	0	0	0
LWK	0	0	0	0	0	0	0	0	0	0	10	0	10	0	22	25	0
ASW	0	0	0	0	4	0	0	0	0	0	11	0	0	0	0	0	9
PUR2	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0

Table A.5: Clustering result based on PCs of SDR with all pruned RVs.

	1	2	3	4	5	6	7
CEU	84	3	1	1	1	0	0
FIN	1	0	35	0	0	0	0
GBR	30	12	1	0	0	0	0
TSI	1	0	0	0	91	0	0
MXL	0	16	0	1	0	0	0
PUR	0	0	0	5	0	0	0
YRI	0	4	0	0	0	74	0
LWK	0	5	0	3	0	0	59
ASW	0	11	0	13	0	0	0
PUR2	0	0	0	5	0	0	0

Table A.6: Association testing results on CVs in simulation 2 with AMRs excluded

			w/o pruning				with pruning				10000 pruned			
		#PCs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
TypeI	SDR	0	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637	0.637
		2	0.074	0.074	0.074	0.080	0.074	0.074	0.074	0.079	0.074	0.074	0.074	0.097
		10	0.077	0.065	0.077	0.074	0.080	0.072	0.079	0.073	0.077	0.077	0.072	0.092
		15	0.061	0.066	0.076	0.075	0.080	0.062	0.063	0.078	0.081	0.078	0.075	0.091
		20	0.069	0.067	0.065	0.076	0.064	0.066	0.067	0.073	0.078	0.079	0.077	0.089
		25	0.070	0.069	0.069	0.078	0.064	0.072	0.071	0.070	0.082	0.082	0.078	0.091
	PCA	2	0.077	0.078	0.077	0.597	0.077	0.075	0.076	0.150	0.076	0.075	0.076	0.166
		10	0.078	0.078	0.070	0.072	0.079	0.056	0.075	0.071	0.074	0.067	0.072	0.083
		15	0.070	0.074	0.070	0.073	0.077	0.064	0.084	0.070	0.077	0.069	0.074	0.084
		20	0.074	0.069	0.077	0.075	0.078	0.065	0.077	0.073	0.077	0.070	0.076	0.083
		25	0.082	0.069	0.081	0.075	0.083	0.070	0.079	0.075	0.081	0.071	0.078	0.081
λ	SDR	0	18.425	18.425	18.425	18.425	18.425	18.425	18.425	18.425	18.425	18.425	18.425	18.425
		2	1.206	1.205	1.202	1.226	1.211	1.221	1.203	1.228	1.212	1.213	1.201	1.380
		10	1.194	1.129	1.205	1.192	1.285	1.182	1.271	1.184	1.259	1.271	1.207	1.315
		15	1.149	1.134	1.207	1.215	1.251	1.106	1.122	1.210	1.286	1.181	1.227	1.296
		20	1.167	1.135	1.145	1.222	1.111	1.170	1.189	1.197	1.256	1.213	1.230	1.272
		25	1.161	1.176	1.206	1.230	1.122	1.227	1.209	1.158	1.285	1.219	1.268	1.324
	PCA	2	1.200	1.189	1.195	14.386	1.223	1.206	1.209	1.832	1.228	1.209	1.208	1.941
		10	1.207	1.203	1.152	1.190	1.222	1.029	1.213	1.173	1.180	1.120	1.177	1.260
		15	1.177	1.166	1.190	1.179	1.234	1.120	1.265	1.150	1.219	1.155	1.211	1.204
		20	1.210	1.156	1.248	1.190	1.223	1.126	1.245	1.181	1.186	1.138	1.235	1.215
		25	1.230	1.174	1.272	1.204	1.249	1.188	1.295	1.187	1.209	1.181	1.263	1.235

Table A.7: Association testing results on CVs in simulation 1.

			w/o pruning				with pruning				10000 pruned			
		#PCs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
TypeI	SDR	0	0.134	0.134	0.134	0.134	0.134	0.134	0.134	0.134	0.134	0.134	0.134	0.134
		2	0.159	0.159	0.160	0.167	0.162	0.162	0.160	0.159	0.161	0.162	0.156	0.161
		10	0.077	0.083	0.102	0.094	0.118	0.063	0.091	0.092	0.139	0.089	0.071	0.104
		15	0.077	0.068	0.093	0.095	0.077	0.033	0.071	0.095	0.092	0.069	0.063	0.101
		20	0.076	0.066	0.089	0.092	0.067	0.067	0.069	0.097	0.085	0.060	0.067	0.104
		25		0.077	0.089	0.087	0.089	0.073	0.078	0.088	0.089	0.078	0.066	0.108
	PCA	2	0.109	0.069	0.129	0.135	0.153	0.128	0.114	0.134	0.155	0.125	0.113	0.135
		10	0.096	0.093	0.110	0.194	0.104	0.061	0.099	0.111	0.106	0.072	0.097	0.119
		15	0.100	0.093	0.111	0.161	0.100	0.071	0.097	0.114	0.103	0.068	0.096	0.131
		20	0.103	0.089	0.114	0.094	0.102	0.065	0.083	0.116	0.104	0.063	0.096	0.139
		25	0.103	0.087	0.115	0.100	0.105	0.081	0.079	0.121	0.106	0.071	0.093	0.138
λ	SDR	0	1.711	1.711	1.711	1.711	1.711	1.711	1.711	1.711	1.711	1.711	1.711	1.711
		2	1.950	1.961	1.944	1.996	1.984	1.994	1.931	1.963	1.988	1.991	1.921	1.946
		10	1.221	1.232	1.452	1.400	1.611	1.120	1.310	1.335	1.769	1.322	1.188	1.459
		15	1.246	1.160	1.370	1.397	1.213	0.872	1.191	1.371	1.387	1.189	1.151	1.430
		20	1.269	1.159	1.350	1.343	1.241	1.181	1.179	1.405	1.352	1.074	1.154	1.434
		25		1.234	1.331	1.326	1.363	1.291	1.271	1.317	1.357	1.241	1.136	1.468
	PCA	2	1.492	1.198	1.692	1.708	1.903	1.698	1.530	1.728	1.930	1.659	1.520	1.733
		10	1.375	1.377	1.484	2.301	1.482	1.146	1.427	1.493	1.498	1.232	1.424	1.608
		15	1.416	1.358	1.502	1.935	1.449	1.205	1.415	1.535	1.483	1.213	1.408	1.717
		20	1.436	1.361	1.533	1.380	1.468	1.213	1.333	1.558	1.490	1.133	1.428	1.779
		25	1.445	1.367	1.541	1.415	1.477	1.300	1.274	1.583	1.480	1.173	1.420	1.759

Table A.8: Association testing results on CVs in simulation 3.

			w/o pruning				with pruning				10000 pruned			
		#PCs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
TypeI	SDR	0	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625
		2	0.172	0.175	0.153	0.129	0.155	0.169	0.157	0.187	0.153	0.169	0.152	0.164
		10	0.073	0.073	0.089	0.133	0.064	0.050	0.079	0.142	0.077	0.072	0.088	0.102
		15	0.071	0.071	0.075	0.139	0.073	0.089	0.089	0.077	0.075	0.083	0.077	0.097
		20	0.072	0.077	0.084	0.139	0.129	0.105	0.149	0.094	0.067	0.099	0.094	0.095
		25	0.081	0.078	0.095	0.107	0.029	0.144	0.163	0.084	0.069	0.146	0.117	0.102
	PCA	2	0.095	0.074	0.135	0.622	0.123	0.134	0.109	0.620	0.137	0.129	0.108	0.623
		10	0.068	0.067	0.130	0.156	0.076	0.073	0.081	0.163	0.127	0.071	0.068	0.167
		15	0.068	0.068	0.137	0.171	0.074	0.077	0.077	0.148	0.067	0.077	0.067	0.165
		20	0.079	0.074	0.143	0.172	0.074	0.075	0.083	0.139	0.069	0.076	0.074	0.154
		25	0.087	0.083	0.136	0.166	0.078	0.083	0.092	0.129	0.070	0.083	0.078	0.157
λ	SDR	0	16.542	16.542	16.542	16.542	16.542	16.542	16.542	16.542	16.542	16.542	16.542	16.542
		2	2.005	2.052	1.828	1.614	1.841	1.974	1.855	2.184	1.807	1.974	1.824	1.960
		10	1.283	1.263	1.316	1.667	1.147	0.975	1.178	1.751	1.186	1.147	1.328	1.390
		15	1.185	1.203	1.266	1.721	1.253	1.371	1.396	1.176	1.175	1.308	1.288	1.413
		20	1.224	1.234	1.324	1.729	1.787	1.513	1.906	1.314	1.138	1.429	1.428	1.396
		25	1.264	1.278	1.427	1.367	0.825	1.899	2.023	1.288	1.175	1.826	1.618	1.432
	PCA	2	1.350	1.203	1.709	16.466	1.584	1.665	1.431	16.176	1.716	1.605	1.440	16.476
		10	1.183	1.161	1.713	1.873	1.221	1.212	1.255	1.965	1.611	1.193	1.187	1.945
		15	1.162	1.166	1.762	2.046	1.211	1.262	1.262	1.839	1.135	1.259	1.149	1.964
		20	1.256	1.225	1.771	2.040	1.251	1.280	1.304	1.742	1.156	1.260	1.228	1.896
		25	1.328	1.302	1.734	1.991	1.247	1.322	1.380	1.657	1.170	1.294	1.268	1.946

Table A.9: Association testing results on CVs in simulation 1 with AMR samples excluded.

			w/o pruning				with pruning				10000 pruned			
		#PCs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
TypeI	SDR	0	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154
		2	0.131	0.131	0.135	0.113	0.132	0.135	0.146	0.117	0.133	0.131	0.141	0.135
		10	0.088	0.063	0.088	0.089	0.084	0.056	0.074	0.079	0.079	0.077	0.080	0.089
		15	0.065	0.065	0.090	0.090	0.090	0.060	0.060	0.080	0.076	0.065	0.072	0.086
		20	0.069	0.069	0.079	0.083	0.062	0.065	0.064	0.075	0.082	0.077	0.075	0.084
		25	0.071	0.071	0.073	0.079	0.065	0.069	0.068	0.073	0.081	0.080	0.075	0.085
	PCA	2	0.138	0.130	0.136	0.131	0.135	0.137	0.127	0.154	0.134	0.135	0.127	0.143
		10	0.086	0.075	0.137	0.150	0.091	0.063	0.086	0.140	0.093	0.067	0.087	0.140
		15	0.087	0.076	0.138	0.151	0.091	0.060	0.081	0.136	0.092	0.066	0.085	0.143
		20	0.086	0.076	0.139	0.146	0.092	0.068	0.083	0.137	0.093	0.069	0.085	0.141
		25	0.090	0.078	0.132	0.143	0.092	0.071	0.082	0.143	0.097	0.072	0.087	0.143
λ	SDR	0	1.916	1.916	1.916	1.916	1.916	1.916	1.916	1.916	1.916	1.916	1.916	1.916
		2	1.744	1.741	1.754	1.578	1.731	1.772	1.842	1.604	1.735	1.748	1.793	1.727
		10	1.286	1.127	1.312	1.384	1.284	1.083	1.226	1.190	1.273	1.218	1.279	1.358
		15	1.099	1.113	1.324	1.389	1.301	1.094	1.061	1.211	1.265	1.145	1.202	1.313
		20	1.148	1.171	1.266	1.314	1.111	1.130	1.090	1.203	1.278	1.234	1.213	1.326
		25	1.142	1.177	1.207	1.248	1.136	1.161	1.118	1.229	1.263	1.285	1.223	1.340
	PCA	2	1.715	1.639	1.718	1.689	1.676	1.707	1.663	1.920	1.665	1.708	1.654	1.793
		10	1.255	1.191	1.810	1.843	1.396	1.088	1.344	1.777	1.422	1.197	1.347	1.766
		15	1.240	1.206	1.804	1.853	1.415	1.090	1.297	1.752	1.427	1.178	1.336	1.791
		20	1.263	1.192	1.809	1.824	1.411	1.145	1.298	1.763	1.438	1.164	1.365	1.787
		25	1.296	1.192	1.729	1.806	1.402	1.123	1.273	1.782	1.448	1.216	1.366	1.787

Table A.10: Association testing results on CVs in simulation 3 with AMR samples excluded.

			w/o pruning				with pruning				10000 pruned			
		#PCs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	0.571	0.571	0.571	0.571	0.571	0.571	0.571	0.571	0.571	0.571	0.571	0.571
		2	0.103	0.103	0.103	0.116	0.105	0.105	0.093	0.116	0.105	0.105	0.093	0.151
		10	0.077	0.066	0.071	0.108	0.075	0.074	0.073	0.088	0.067	0.074	0.071	0.098
		15	0.062	0.066	0.077	0.106	0.063	0.067	0.066	0.086	0.073	0.076	0.071	0.097
		20	0.068	0.069	0.060	0.102	0.067	0.070	0.070	0.068	0.075	0.072	0.077	0.095
		25	0.069	0.069	0.065	0.094	0.071	0.076	0.076	0.069	0.077	0.076	0.080	0.093
	PCA	2	0.115	0.115	0.114	0.547	0.103	0.102	0.103	0.144	0.102	0.098	0.103	0.160
		10	0.077	0.076	0.091	0.094	0.068	0.058	0.061	0.092	0.077	0.063	0.059	0.107
		15	0.072	0.070	0.069	0.095	0.069	0.064	0.067	0.092	0.073	0.067	0.060	0.111
		20	0.075	0.066	0.077	0.096	0.070	0.071	0.065	0.091	0.073	0.066	0.062	0.114
		25	0.078	0.067	0.082	0.097	0.072	0.071	0.068	0.093	0.077	0.067	0.065	0.116
λ	SDR	0	12.190	12.190	12.190	12.190	12.190	12.190	12.190	12.190	12.190	12.190	12.190	12.190
		2	1.450	1.446	1.433	1.549	1.461	1.449	1.341	1.558	1.453	1.446	1.356	1.853
		10	1.205	1.142	1.146	1.401	1.156	1.148	1.153	1.258	1.112	1.162	1.141	1.356
		15	1.150	1.138	1.210	1.397	1.102	1.154	1.152	1.251	1.207	1.169	1.187	1.337
		20	1.178	1.136	1.108	1.358	1.149	1.200	1.216	1.156	1.240	1.138	1.236	1.313
		25	1.174	1.158	1.181	1.323	1.196	1.225	1.226	1.124	1.265	1.161	1.283	1.328
	PCA	2	1.484	1.513	1.453	10.795	1.419	1.363	1.420	1.831	1.404	1.342	1.416	2.008
		10	1.162	1.203	1.331	1.319	1.152	1.036	1.076	1.259	1.221	1.079	1.083	1.471
		15	1.152	1.143	1.185	1.314	1.174	1.120	1.161	1.268	1.219	1.102	1.084	1.455
		20	1.175	1.153	1.221	1.320	1.166	1.162	1.160	1.274	1.217	1.110	1.095	1.475
		25	1.219	1.155	1.224	1.326	1.176	1.207	1.168	1.310	1.246	1.157	1.115	1.498

Table A.11: Association testing results on RVs in simulation 1.

				w/o pruning				with pruning				10000 pruned			
		#PCs	Tests	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T1	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136
			Fp	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136
		10	T1	0.062	0.056	0.076	0.073	0.114	0.058	0.105	0.081	0.140	0.084	0.079	0.095
			Fp	0.060	0.057	0.074	0.071	0.110	0.056	0.104	0.081	0.135	0.083	0.077	0.096
		25	T1		0.071	0.065	0.072	0.100	0.099	0.078	0.092	0.097	0.107	0.057	0.089
			Fp		0.071	0.065	0.072	0.098	0.100	0.075	0.093	0.098	0.106	0.057	0.089
	PCA	10	T1	0.090	0.098	0.083	0.112	0.088	0.096	0.127	0.097	0.086	0.101	0.112	0.095
			Fp	0.089	0.097	0.082	0.106	0.088	0.096	0.126	0.096	0.085	0.100	0.111	0.092
		25	T1	0.080	0.094	0.081	0.087	0.103	0.101	0.077	0.085	0.103	0.094	0.093	0.085
			Fp	0.079	0.092	0.081	0.086	0.101	0.103	0.075	0.084	0.100	0.094	0.094	0.082
λ	SDR	0	T1	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795
			Fp	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795	1.795
		10	T1	1.170	1.122	1.176	1.190	1.592	1.059	1.635	1.281	1.826	1.260	1.234	1.440
			Fp	1.165	1.135	1.191	1.179	1.551	1.056	1.618	1.299	1.827	1.243	1.264	1.409
		25	T1		1.271	1.208	1.274	1.428	1.353	1.295	1.398	1.453	1.463	1.017	1.334
			Fp		1.242	1.183	1.292	1.439	1.350	1.265	1.391	1.434	1.451	1.037	1.335
	PCA	10	T1	1.464	1.544	1.273	1.583	1.291	1.430	1.742	1.399	1.263	1.446	1.604	1.429
			Fp	1.497	1.587	1.244	1.552	1.306	1.453	1.776	1.368	1.268	1.456	1.655	1.369
		25	T1	1.274	1.427	1.278	1.326	1.416	1.441	1.258	1.278	1.443	1.370	1.353	1.274
			Fp	1.297	1.411	1.240	1.332	1.423	1.455	1.265	1.245	1.445	1.375	1.379	1.250

Table A.12: Association testing results on RVs in simulation 3.

				w/o pruning				with pruning				10000 pruned			
		#PCs	Tests	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T1	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277
			Fp	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249
		10	T1	0.069	0.066	0.074	0.100	0.062	0.037	0.065	0.140	0.071	0.049	0.075	0.129
			Fp	0.069	0.067	0.074	0.098	0.062	0.038	0.064	0.140	0.069	0.049	0.073	0.132
		25	T1	0.073	0.070	0.088	0.107		0.118		0.126	0.060	0.081	0.095	0.114
			Fp	0.073	0.073	0.088	0.105		0.119		0.129	0.061	0.084	0.094	0.113
	PCA	10	T1	0.187	0.160	0.234	0.197	0.080	0.140	0.109	0.139	0.229	0.136	0.122	0.185
			Fp	0.191	0.165	0.237	0.201	0.078	0.145	0.109	0.138	0.234	0.139	0.122	0.185
		25	T1	0.085	0.095	0.176	0.140	0.079	0.117	0.099	0.088	0.096	0.131	0.093	0.139
			Fp	0.084	0.093	0.177	0.137	0.081	0.122	0.099	0.085	0.097	0.136	0.093	0.137
λ	SDR	0	T1	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567
			Fp	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166
		10	T1	1.123	1.106	1.287	1.540	1.071	0.849	1.110	1.976	1.151	0.971	1.263	1.686
			Fp	1.143	1.151	1.289	1.530	1.075	0.845	1.119	1.979	1.160	0.970	1.232	1.699
		25	T1	1.215	1.186	1.363	1.487		1.566		1.747	1.052	1.276	1.413	1.509
			Fp	1.208	1.201	1.335	1.479		1.522		1.797	1.088	1.275	1.408	1.516
	PCA	10	T1	2.453	2.130	3.231	2.684	1.283	1.941	1.565	1.868	2.978	1.878	1.720	2.537
			Fp	2.481	2.194	3.254	2.705	1.297	1.986	1.573	1.848	3.058	1.906	1.753	2.556
		25	T1	1.328	1.433	2.445	1.911	1.233	1.653	1.378	1.364	1.434	1.785	1.375	1.936
			Fp	1.327	1.483	2.451	1.915	1.216	1.669	1.354	1.347	1.449	1.810	1.389	1.911

Table A.13: Association testing results on LFVs in sim 1.

				w/o pruning				with pruning				10000 pruned			
		#PCs	Tests	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T5	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.163
			Fp	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.164
		10	T1	0.086	0.084	0.089	0.080	0.149	**0.061**	0.100	0.087	0.163	0.088	0.064	0.098
			Fp	0.088	0.082	0.092	0.077	0.154	**0.063**	0.107	0.085	0.164	0.091	0.070	0.096
		25	T5		0.081	0.094	0.096	0.096	0.078	0.074	0.091	0.076	0.079	**0.057**	0.098
			Fp		0.084	0.091	0.094	0.095	0.080	0.075	0.092	0.079	0.079	**0.058**	0.093
	PCA	10	T5	0.097	0.093	0.103	0.175	0.086	**0.052**	0.094	0.109	0.081	**0.065**	0.080	0.097
			Fp	0.100	0.097	0.106	0.169	0.088	**0.056**	0.096	0.110	0.080	**0.066**	0.082	0.098
		25	T1	0.108	0.086	0.106	0.090	0.091	0.077	0.072	0.116	0.086	0.070	0.078	0.101
			Fp	0.110	0.085	0.109	0.092	0.091	0.083	0.070	0.116	0.087	0.071	0.075	0.098
λ	SDR	0	T5	1.816	1.816	1.816	1.816	1.816	1.816	1.816	1.816	1.816	1.816	1.816	1.823
			Fp	1.849	1.849	1.849	1.849	1.849	1.849	1.849	1.849	1.849	1.849	1.849	1.853
		10	T5	1.307	1.272	1.334	1.287	1.706	**1.164**	1.504	1.357	1.906	1.295	1.198	1.463
			Fp	1.317	1.270	1.348	1.240	1.703	**1.166**	1.539	1.332	1.890	1.370	1.211	1.454
		25	T5		1.265	1.362	1.350	1.379	1.199	1.228	1.397	1.304	1.236	**1.068**	1.435
			Fp		1.293	1.341	1.333	1.409	1.200	1.204	1.393	1.302	1.240	**1.058**	1.423
	PCA	10	T5	1.394	1.347	1.425	2.072	1.313	**1.046**	1.406	1.533	1.298	**1.140**	1.266	1.444
			Fp	1.434	1.412	1.461	1.964	1.306	**1.067**	1.437	1.540	1.305	**1.144**	1.297	1.456
		25	T5	1.444	1.317	1.460	1.356	1.351	1.278	1.221	1.546	1.370	1.163	1.254	1.468
			Fp	1.458	1.316	1.478	1.377	1.353	1.286	1.218	1.518	1.385	1.210	1.228	1.441

Table A.14: Association testing results on LFVs in sim 2.

				w/o pruning				with pruning				10000 pruned			
		#PCs	Tests	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T5	0.613	0.633	0.633	0.633	0.633	0.633	0.633	0.633	0.633	0.633	0.633	0.604
			Fp	0.578	0.596	0.596	0.596	0.596	0.596	0.596	0.596	0.596	0.596	0.596	0.566
		10	T5	0.062	0.065	0.070	0.096	0.063	0.062	0.092	0.106	0.070	0.078	0.081	0.108
			Fp	0.059	0.062	0.068	0.091	0.064	0.060	0.092	0.100	0.069	0.078	0.082	0.105
		25	T5	0.114	0.074	0.092	0.109	0.084	0.089	0.130	0.115	0.083	0.075	0.080	0.116
			Fp	0.119	0.078	0.092	0.105	0.085	0.093	0.131	0.110	0.083	0.077	0.082	0.113
	PCA	10	T5	0.087	0.086	0.093	0.092	0.083	0.087	0.074	0.095	0.104	0.084	0.073	0.096
			Fp	0.088	0.090	0.094	0.093	0.081	0.088	0.072	0.092	0.106	0.083	0.072	0.094
		25	T5	0.084	0.083	0.097	0.109	0.090	0.088	0.088	0.103	0.084	0.091	0.085	0.101
			Fp	0.083	0.086	0.096	0.106	0.089	0.093	0.087	0.098	0.083	0.091	0.083	0.097
λ	SDR	0	T5	14.816	16.658	16.658	16.658	16.658	16.658	16.658	16.658	16.658	16.658	16.658	13.876
			Fp	12.071	13.171	13.171	13.171	13.171	13.171	13.171	13.171	13.171	13.171	13.171	11.488
		10	T5	1.104	1.103	1.223	1.381	1.145	1.091	1.397	1.487	1.156	1.217	1.285	1.556
			Fp	1.101	1.098	1.191	1.351	1.148	1.103	1.378	1.448	1.164	1.207	1.293	1.508
		25	T5	1.598	1.257	1.367	1.516	1.300	1.343	1.756	1.606	1.280	1.283	1.287	1.630
			Fp	1.657	1.261	1.360	1.478	1.275	1.381	1.755	1.534	1.272	1.288	1.285	1.604
	PCA	10	T5	1.324	1.323	1.376	1.343	1.299	1.347	1.232	1.330	1.488	1.317	1.255	1.370
			Fp	1.331	1.339	1.392	1.351	1.283	1.335	1.241	1.326	1.495	1.301	1.243	1.366
		25	T5	1.322	1.279	1.384	1.424	1.385	1.324	1.362	1.395	1.326	1.318	1.359	1.360
			Fp	1.307	1.298	1.381	1.407	1.364	1.331	1.367	1.340	1.310	1.362	1.333	1.351

Table A.15: Association testing results on LFVs in sim 3.

				w/o pruning				w/ pruning				10000 pruned			
		#PCs	Tests	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T5	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277	0.277
			Fp	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249	0.249
		10	T5	0.069	0.066	0.074	0.100	0.062	0.037	0.065	0.140	0.071	0.049	0.075	0.129
			Fp	0.069	0.067	0.074	0.098	0.062	0.038	0.064	0.140	0.069	0.049	0.073	0.132
		25	T5	0.073	0.070	0.088	0.107		0.118		0.126	0.060	0.081	0.095	0.114
			Fp	0.073	0.073	0.088	0.105		0.119		0.129	0.061	0.084	0.094	0.113
	PCA	10	T5	0.187	0.160	0.234	0.197	0.080	0.140	0.109	0.139	0.229	0.136	0.122	0.185
			Fp	0.191	0.165	0.237	0.201	0.078	0.145	0.109	0.138	0.234	0.139	0.122	0.185
		25	T5	0.085	0.095	0.176	0.140	0.079	0.117	0.099	0.088	0.096	0.131	0.093	0.139
			Fp	0.084	0.093	0.177	0.137	0.081	0.122	0.099	0.085	0.097	0.136	0.093	0.137
λ	SDR	0	T5	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567	3.567
			Fp	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166	3.166
		10	T5	1.123	1.106	1.287	1.540	1.071	0.849	1.110	1.976	1.151	0.971	1.263	1.686
			Fp	1.143	1.151	1.289	1.530	1.075	0.845	1.119	1.979	1.160	0.970	1.232	1.699
		25	T5	1.215	1.186	1.363	1.487		1.566		1.747	1.052	1.276	1.413	1.509
			Fp	1.208	1.201	1.335	1.479		1.522		1.797	1.088	1.275	1.408	1.516
	PCA	10	T5	2.453	2.130	3.231	2.684	1.283	1.941	1.565	1.868	2.978	1.878	1.720	2.537
			Fp	2.481	2.194	3.254	2.705	1.297	1.986	1.573	1.848	3.058	1.906	1.753	2.556
		25	T5	1.328	1.433	2.445	1.911	1.233	1.653	1.378	1.364	1.434	1.785	1.375	1.936
			Fp	1.327	1.483	2.451	1.915	1.216	1.669	1.354	1.347	1.449	1.810	1.389	1.911

Table A.16: Association testing results on RVs with window size=20 in R1.

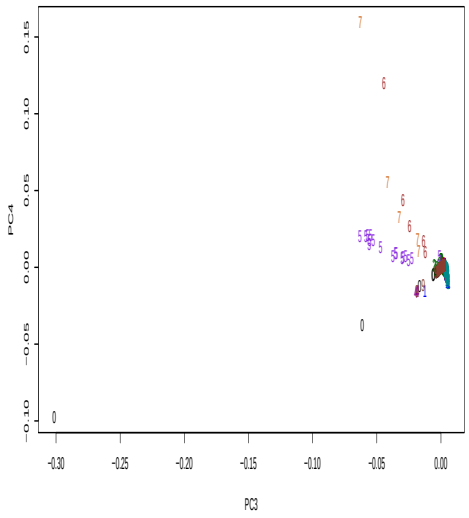
				w/o pruning				with pruning			
		#PCs	Tests	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T1	0.201	0.201	0.201	0.201	0.201	0.201	0.201	0.201
			Fp	0.184	0.184	0.184	0.184	0.184	0.184	0.184	0.184
		2	T1	0.136	0.134	0.147	0.176	0.148	0.141	0.147	0.141
			Fp	0.133	0.131	0.145	0.173	0.145	0.136	0.145	0.136
		10	T1	0.077	0.078	0.089	0.075	0.147	0.093	0.081	0.083
			Fp	0.075	0.076	0.087	0.074	0.143	0.092	0.079	0.081
		25	T1	0.082	0.084	0.084	0.080	0.080	0.077	0.079	0.084
			Fp	0.079	0.080	0.082	0.078	0.077	0.077	0.079	0.083
	PCA	2	T1	0.159	0.155	0.107	0.229	0.175	0.093	0.084	0.248
			Fp	0.158	0.153	0.108	0.211	0.173	0.090	0.083	0.229
		10	T1	0.115	0.104	0.123	0.159	0.095	0.119	0.093	0.145
			Fp	0.116	0.105	0.123	0.149	0.096	0.120	0.095	0.147
		25	T1	0.081	0.086	0.096	0.127	0.098	0.088	0.084	0.092
			Fp	0.079	0.084	0.094	0.123	0.096	0.086	0.082	0.091
λ	SDR	0	T1	2.324	2.324	2.324	2.324	2.324	2.324	2.324	2.324
			Fp	2.078	2.078	2.078	2.078	2.078	2.078	2.078	2.078
		2	T1	1.798	1.752	1.912	2.170	1.919	1.840	1.921	1.815
			Fp	1.764	1.748	1.861	2.125	1.888	1.821	1.877	1.759
		10	T1	1.201	1.209	1.293	1.197	1.879	1.401	1.244	1.246
			Fp	1.193	1.203	1.260	1.199	1.838	1.385	1.235	1.228
		25	T1	1.312	1.309	1.341	1.268	1.269	1.250	1.259	1.274
			Fp	1.284	1.289	1.305	1.278	1.295	1.289	1.250	1.263
	PCA	2	T1	2.027	2.000	1.447	2.575	2.253	1.359	1.279	2.798
			Fp	1.987	1.945	1.452	2.343	2.232	1.318	1.277	2.599
		10	T1	1.522	1.435	1.608	1.908	1.383	1.607	1.347	1.824
			Fp	1.560	1.449	1.640	1.790	1.403	1.627	1.365	1.848
		25	T1	1.320	1.338	1.349	1.697	1.457	1.366	1.279	1.383
			Fp	1.302	1.329	1.387	1.673	1.440	1.357	1.294	1.377

Table A.17: Association testing results on RVs with window size=20 in R2.

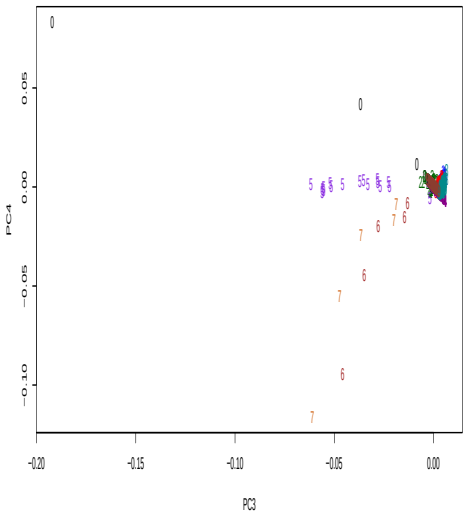
				w/o pruning				with pruning			
		#PCs	Tests	all	CVs	LFVs	RVs	all	CVs	LFVs	RVs
Type I	SDR	0	T1	0.121	0.121	0.121	0.121	0.121	0.121	0.121	0.121
			Fp	0.123	0.123	0.123	0.123	0.123	0.123	0.123	0.123
		2	T1	0.094	0.095	0.096	0.086	0.094	0.093	0.095	0.090
			Fp	0.093	0.093	0.099	0.085	0.096	0.093	0.095	0.087
		10	T1	0.092	0.090	0.097	0.087	0.093	0.089	0.099	0.089
			Fp	0.091	0.090	0.094	0.087	0.091	0.089	0.097	0.087
		25	T1	0.097	0.099	0.103	0.069	0.101	0.096	0.099	0.073
			Fp	0.098	0.101	0.104	0.069	0.100	0.100	0.099	0.075
	PCA	2	T1	0.111	0.110	0.111	0.153	0.110	0.100	0.111	0.157
			Fp	0.111	0.110	0.112	0.154	0.111	0.101	0.112	0.159
		10	T1	0.119	0.117	0.121	0.148	0.111	0.098	0.096	0.137
			Fp	0.121	0.118	0.122	0.143	0.112	0.099	0.098	0.138
		25	T1	0.095	0.096	0.106	0.161	0.090	0.100	0.096	0.110
			Fp	0.096	0.096	0.107	0.161	0.092	0.102	0.096	0.110
λ	SDR	0	T1	1.584	1.584	1.584	1.584	1.584	1.584	1.584	1.584
			Fp	1.584	1.584	1.584	1.584	1.584	1.584	1.584	1.584
		2	T1	1.392	1.396	1.412	1.301	1.411	1.403	1.408	1.338
			Fp	1.376	1.385	1.410	1.290	1.393	1.367	1.402	1.325
		10	T1	1.322	1.324	1.350	1.283	1.355	1.320	1.360	1.307
			Fp	1.358	1.337	1.337	1.291	1.364	1.331	1.373	1.305
		25	T1	1.429	1.423	1.482	1.169	1.441	1.415	1.429	1.192
			Fp	1.422	1.437	1.487	1.176	1.444	1.435	1.453	1.163
	PCA	2	T1	1.468	1.459	1.500	1.832	1.475	1.399	1.479	1.883
			Fp	1.491	1.477	1.477	1.881	1.489	1.412	1.498	1.913
		10	T1	1.534	1.533	1.567	1.741	1.465	1.351	1.349	1.691
			Fp	1.528	1.506	1.554	1.757	1.488	1.377	1.360	1.706
		25	T1	1.379	1.427	1.431	1.910	1.335	1.450	1.386	1.453
			Fp	1.410	1.443	1.461	1.902	1.340	1.448	1.415	1.478

Figure A.1: PC3 and PC4 of SDR. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), MXL=5(blueviolet), PUR=6(brown), PUR2=7(chocolate), TSI=8(cyan), YRI=9(coral)

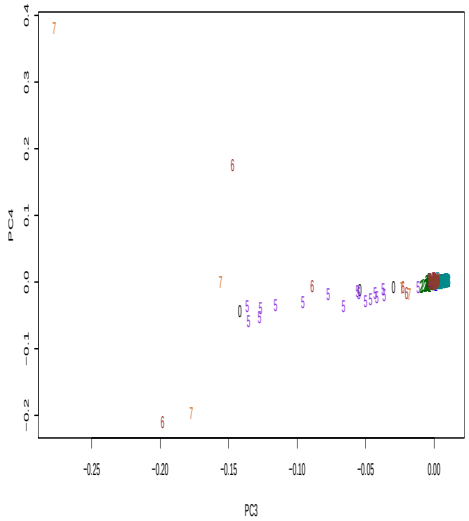
all variants



all CVs



all LFVs



all RVs

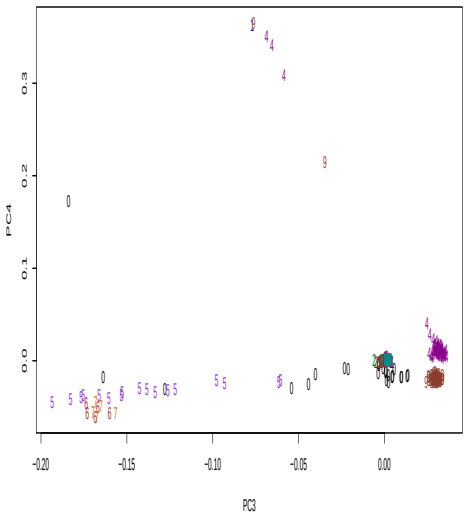
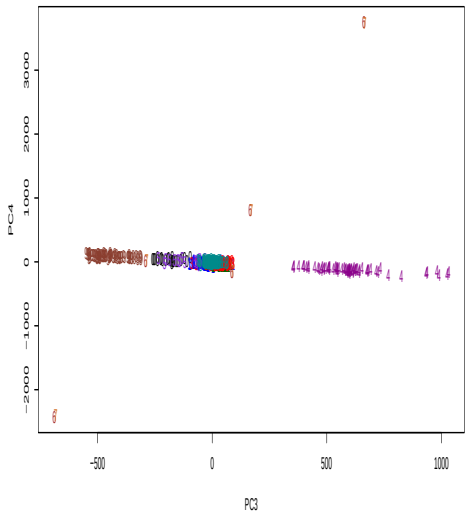
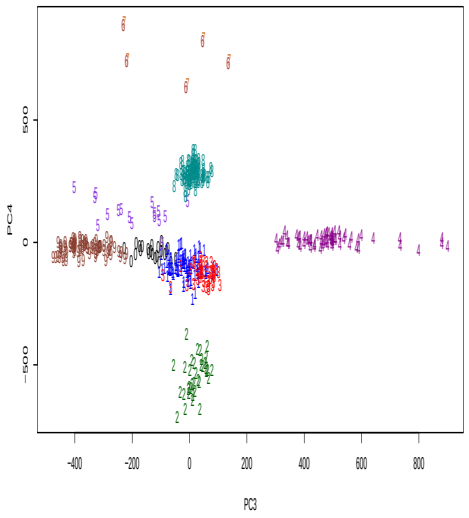


Figure A.2: PC3 and PC4 of PCA. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), MXL=5(blueviolet), PUR=6(brown), PUR2=7(chocolate), TSI=8(cyan), YRI=9(coral)

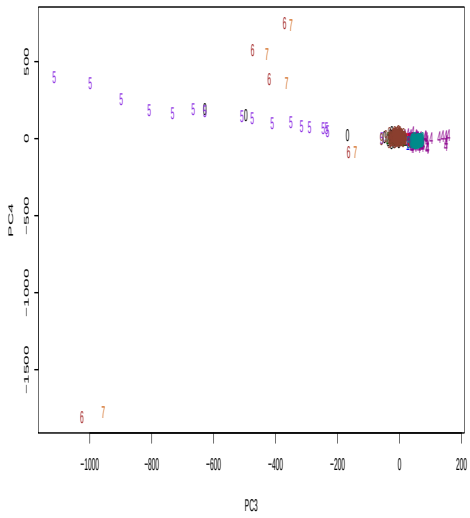
all variants



all CVs



all LFVs



all RVs

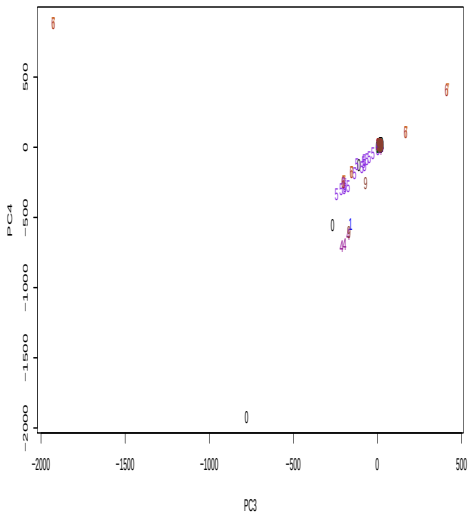
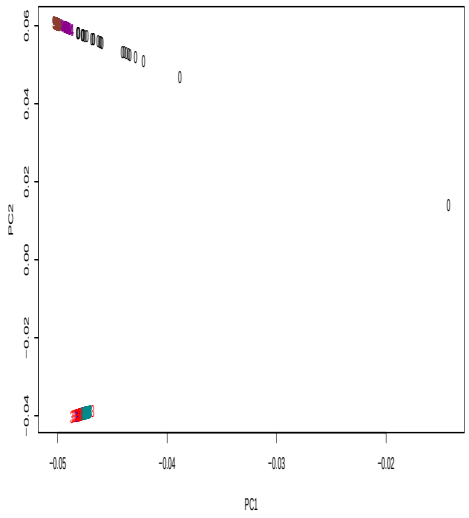
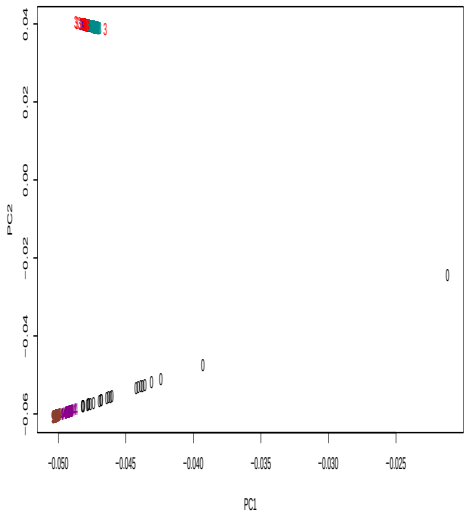


Figure A.3: PC1 and PC2 of SDR without AMRs. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), TSI=8(cyan), YRI=9(coral)

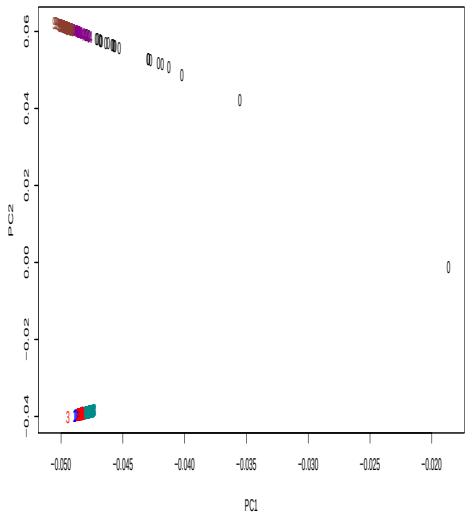
all variants



all CVs



all LFVs



all RVs

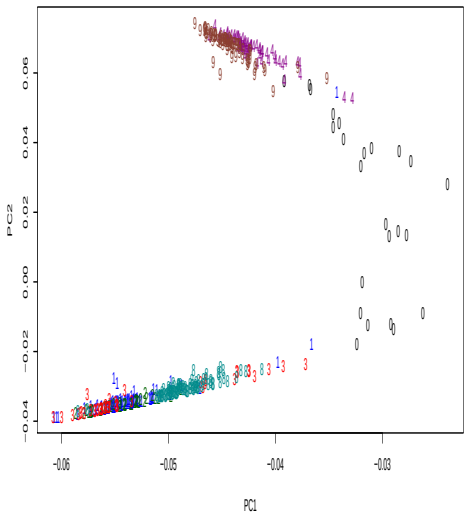
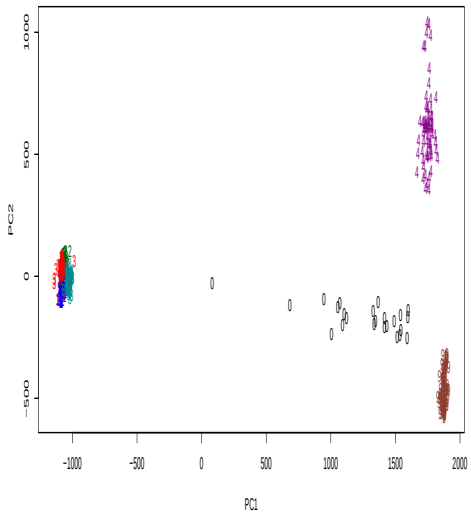
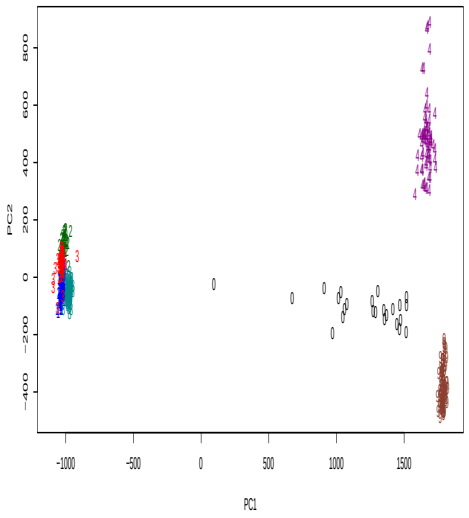


Figure A.4: PC1 and PC2 of PCA without AMRs. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), TSI=8(cyan), YRI=9(coral)

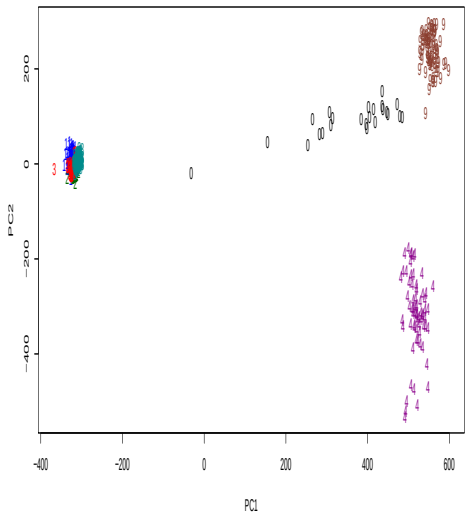
all variants



all CVs



all LFVs



all RVs

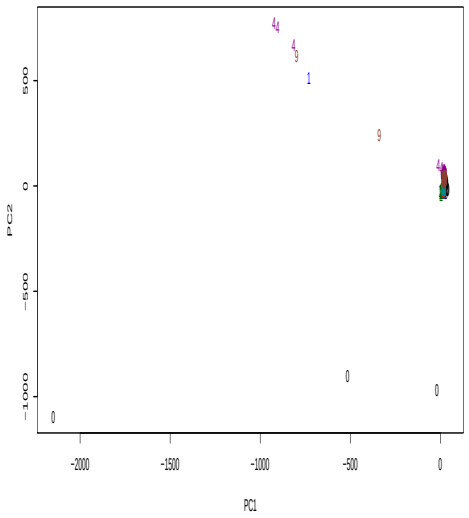
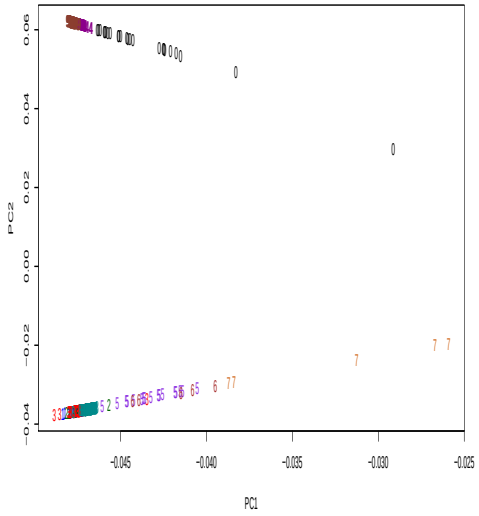
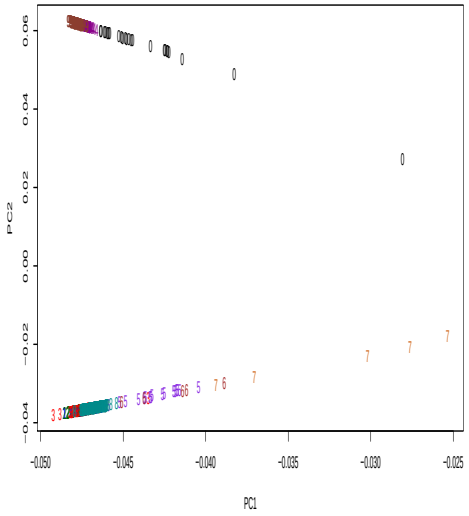


Figure A.5: PC1 and PC2 of SDR without AMRs of pruned variants and 10000 pruned variants. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), TSI=8(cyan), YRI=9(coral)

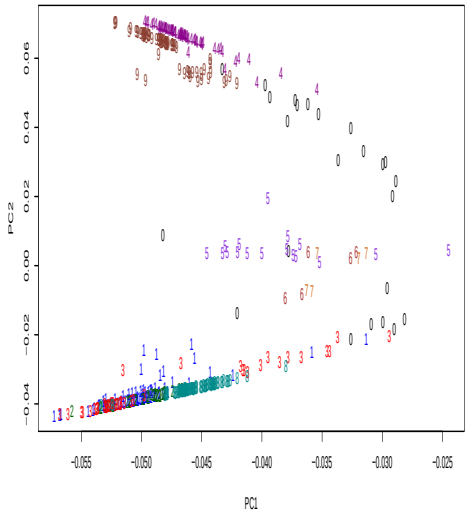
pruned CV



10000 pruned CVs



pruned RVs



10000 pruned RVs

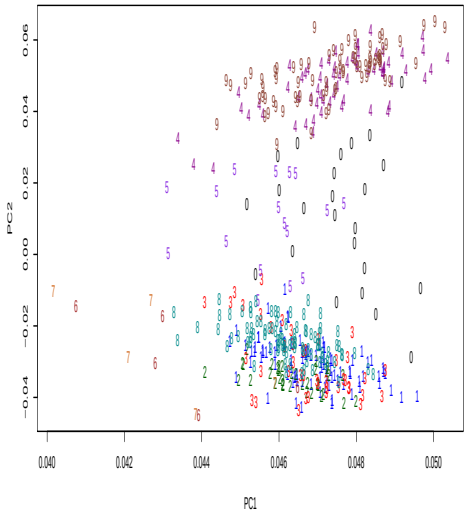
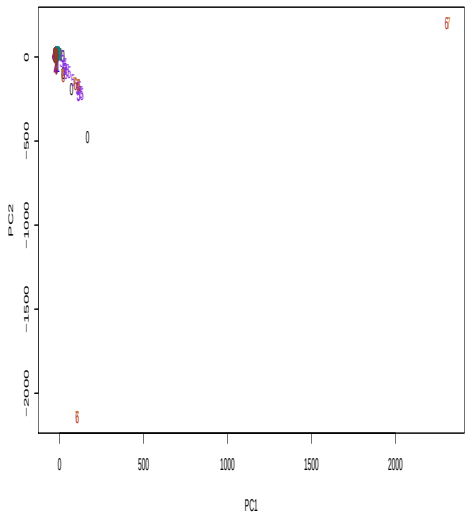
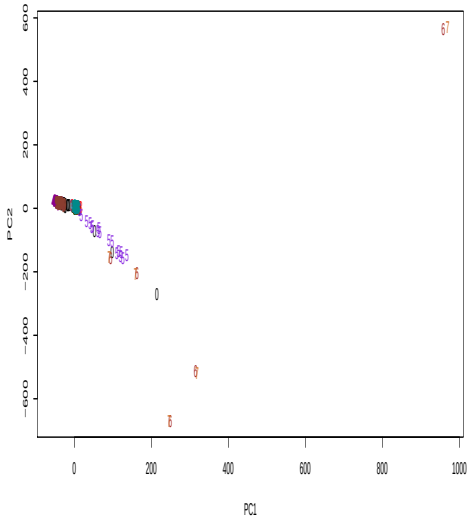


Figure A.6: PC1 and PC2 of SDR of all RVs, all pruned RVs and 10000 pruned RVs. ASW=0(black), CEU=1(blue), FIN=2(darkgreen), GBR=3(red), LWK=4(darkmagenta), MXL=5(blueviolet), PUR=6(brown), PUR2=7(chocolate), TSI=8(cyan), YRI=9(coral)

all RVs



all pruned RVs



10000 pruned RVs

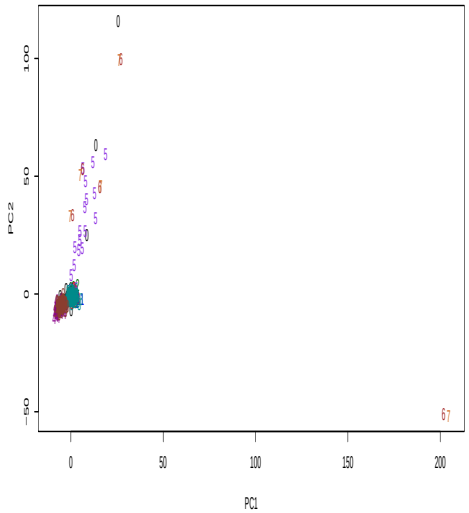
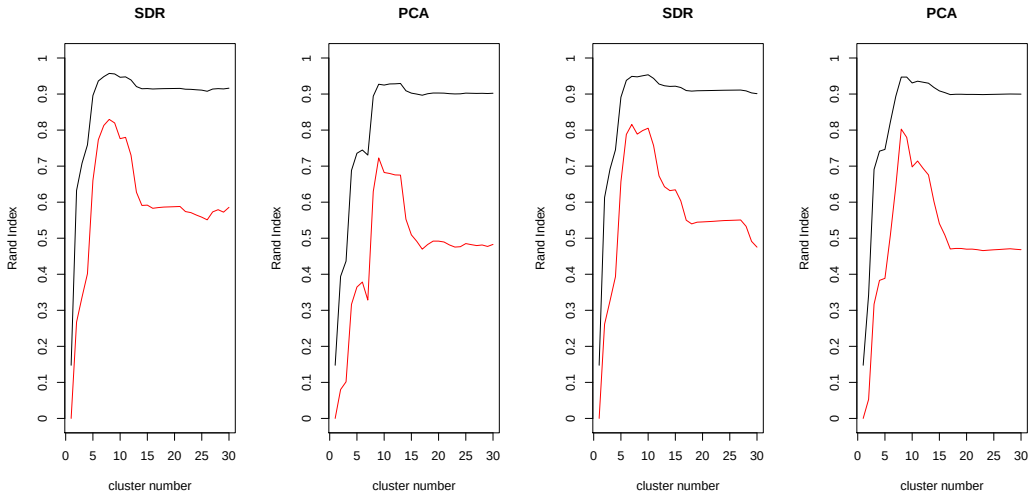
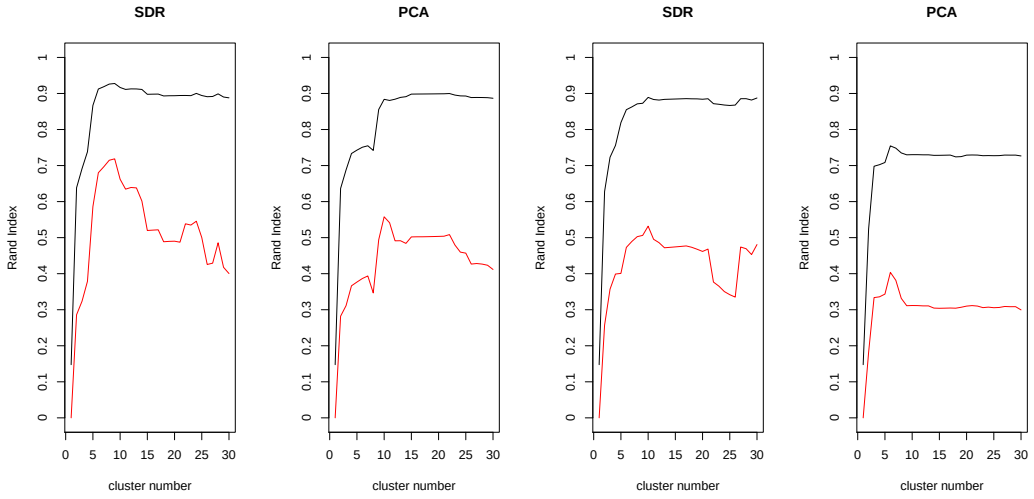


Figure A.7: Change of (a)RI with the number of clusters. The black line is for RI and the red line is for aRI.



(a) all variants

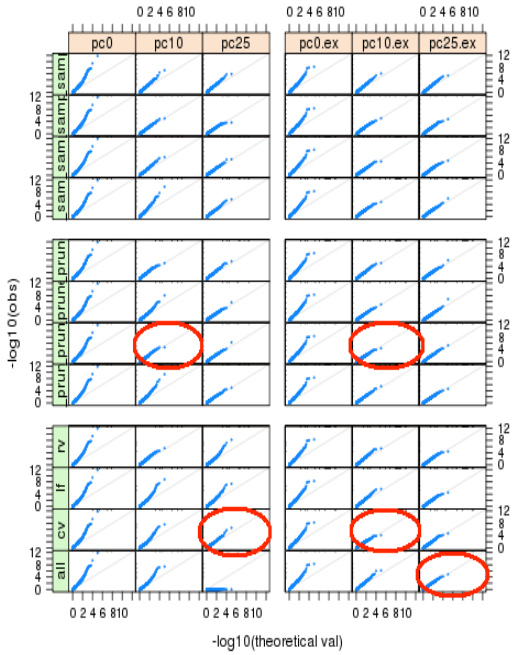
(b) all CVs



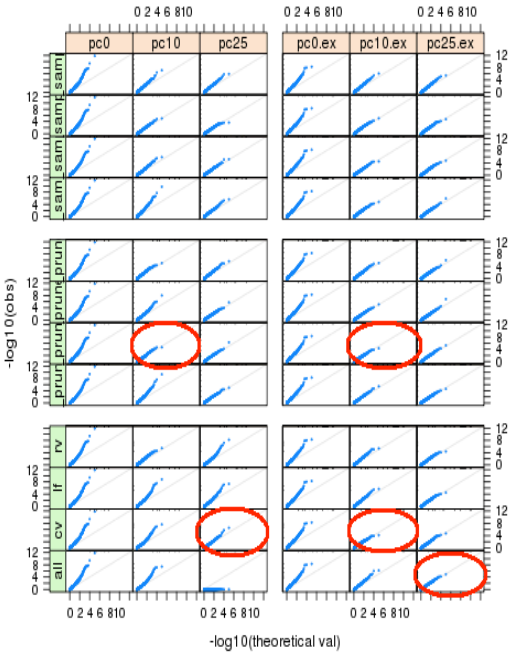
(c) all LFVs

(d) all RVs

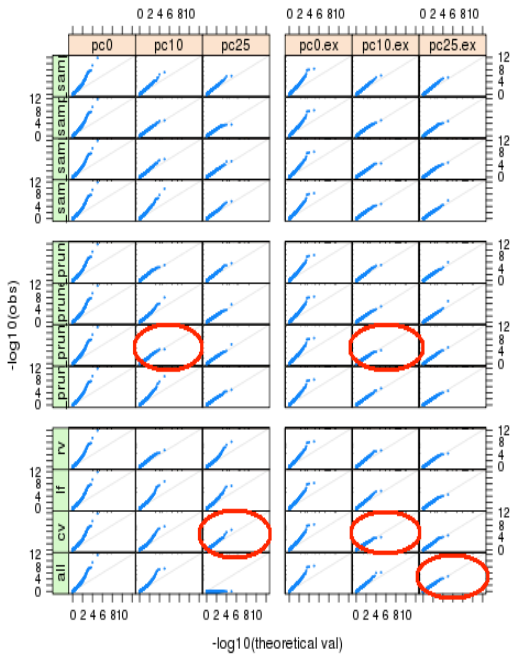
Figure A.8: Q-Q plots of association testing of CVs with adjustment of PCs of SDR in all 3 simulations.



(a) sim1

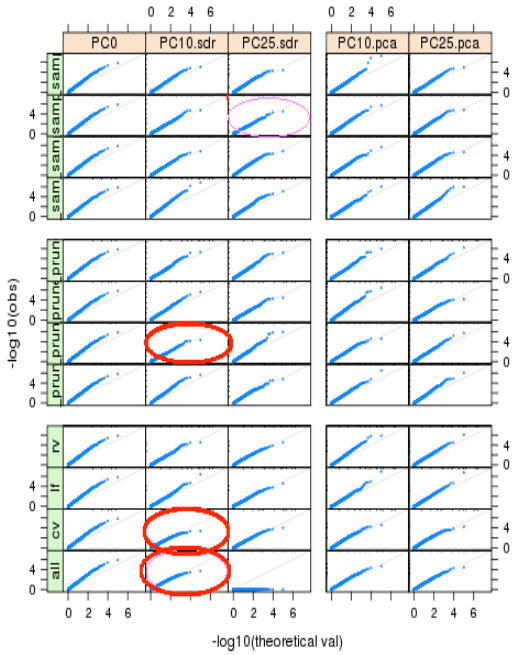


(b) sim2

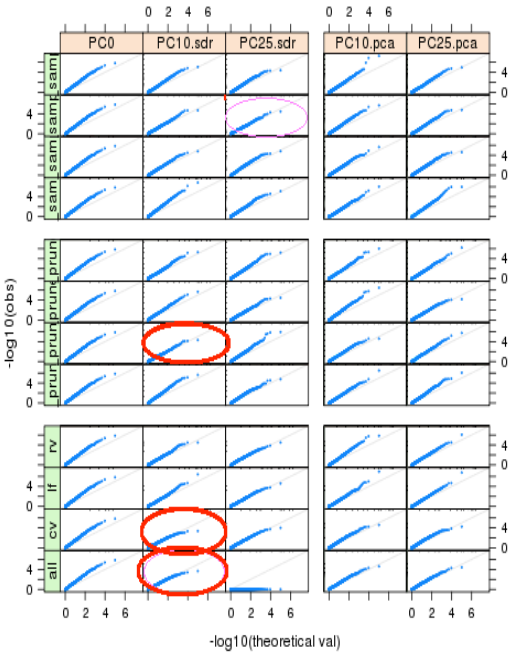


(c) sim3

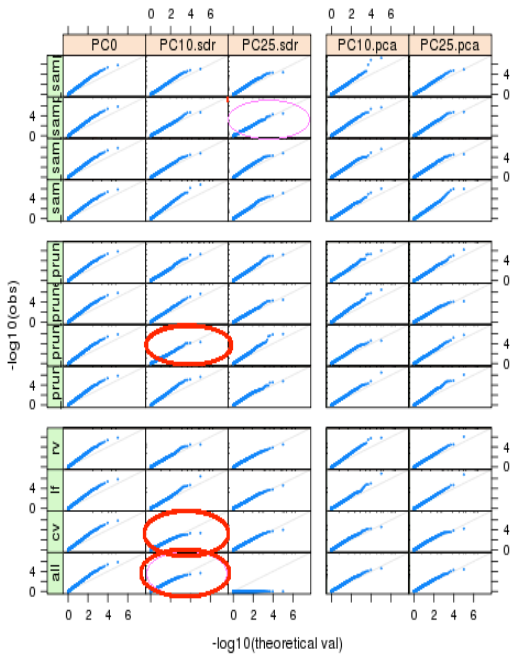
Figure A.9: Q-Q plots of association testing of RVs with adjustment of PCs of SDR in all 3 simulations.



(a) sim1



(b) sim2



(c) sim3

Appendix B

Proof of the Equivalence Between the Two Hybrid Models

The hybrid model contains two parts: the random term which aims to catch the sample structure and the PC terms which aim to catch the environmental effect. In the hybrid model, suppose Z is the matrix of the top k PCs that we want to include, and u is a random effect with $u \sim N(0, \sigma_g^2 S)$, where S is the similarity matrix. For convenience, we consider the hybrid model without an intercept here. We have 2 options of S:

1. $$Y = g\beta + Z\gamma + u + \epsilon, \tag{B.1}$$

   where $\epsilon \sim N(0, I)$, $u \sim N(0, \sigma_g^2 K)$, K is the IBS matrix estimated from high-density variants, σ_g^2 is the polygenic variance, and σ^2 is the individual variance. So we have $\text{var}(Y) = H = \sigma_g^2 K + \sigma^2 I$. We can write $\sigma_g^2 K = Q\Lambda Q^T = (Q_1\ Q_2)\text{diag}\{\Lambda_1\ \Lambda_2\}(Q_1\ Q_2)^T$, where Q_1 contains the top k PCs that we use as Z, $Z = Q_1$. diag$\{.\}$ is a diagonal matrix.

2. As Z is generated from K, some of the effects are modeled both by Z and K.

To avoid the redundancy, a direct modification to Model (B.1) is:

$$Y = g\beta + Z\gamma + u_2 + \epsilon, \tag{B.2}$$

where $\epsilon \sim N(0, I)$, $u_2 \sim N(0, \sigma_g^2 K_2)$, where $\sigma_g^2 K_2 = Q_2 \Lambda Q_2^T$, is the modified IBS matrix after excluding the variance explained by the top k PCs.

In the LMM, since we use the IBS matrix estimated from the high-density variants, K is already known and we only need to estimate σ_g^2, σ^2 and the fixed effects.

Actually, Model (B.1) and (B.2) are equivalent. Intuitively, the REML estimator and inference of β should be the same because REML estimator is defined as a maximum likelihood estimator based on a linear transformed set of data $Y^* = AY$ such that Y^* does not depend on the fixed effects any more. One way to achieve this is, suppose $X = (Z\ g) = (Q_1\ g)$, then $A = I - P_x = I - X(X^T X)^{-1} X^T$ which is the projection onto the orthogonal space of of X. After the projection, Model (B.1) becomes:

$$Y^* = (I - P_x)u + (I - P_x)\epsilon \tag{B.3}$$

and Model (B.2) becomes:

$$Y^* = (I - P_x)u_2 + (I - P_x)\epsilon. \tag{B.4}$$

The only difference between equation (B.3) and (B.4) is $(I - P_x)u$ and $(I - P_x)u_2$. And we have

$$\begin{aligned} \text{var}((I - P_x)u) &= (I - P_x)\sigma_g^2 K(I - P_x) = (I - P_x)Q\Lambda Q^T(I - P_x) \\ &= (I - P_x)(Q_1\ Q_2)\Lambda(Q_1\ Q_2)^T(I - P_x). \end{aligned}$$

Recall that $I - P_x$ is the projection onto the orthogonal space of X, so we have $(I - P_x)Q_1 = (I - P_x)X(I\ 0)^T = 0$, and

$$(I - P_x)(Q_1\ Q_2)\Lambda(Q_1\ Q_2)^T(I - P_x) = (I - P_x)Q_2\Lambda Q_2^T(I - P_x),$$

which is equal to

$$\text{var}((I - P_x)u_2) = (I - P_x)\sigma_g^2 K_2 (I - P_x) = (I - P_x)Q_2 \Lambda_2 Q_2^T (I - P_x).$$

.

Appendix C

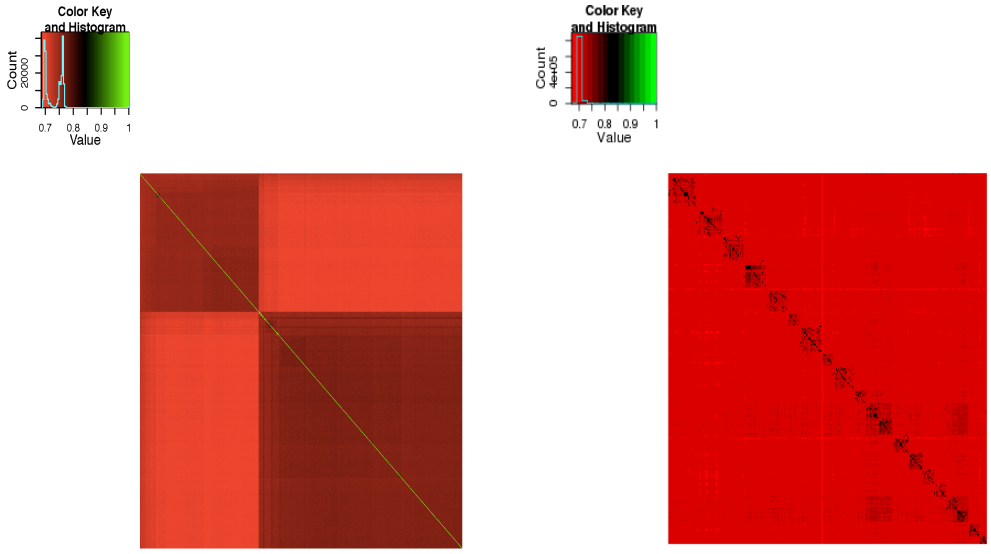
Simulations with the 1000 Genomes Project Data

In Chapter 5, we also used the 1000 Genomes Project data for comparing the PCR and LMM following the model described before. We first pruned all the common variants (CVs) by PLINK (Purcell et al., 2007) with a sliding window of size 50, a moving step 5 and $r^2 \leq 0.05$. We randomly selected 31292 pruned CVs with MAF $\geq$0.05 from all autosomes, including European (EUR) and African (AFR) ancestry groups, to estimate the similarity matrix. Wae selected 11464 unpruned CVs from chromosome 1, which were significantly associated with population structure, and carried out single variant analysis.

C.1 Visualization of population structure

Figure C.1 displays the heatmaps of two IBS matrices of the 1000 Genomes Projects data and the GAW18 data, which is used in the main text. The IBS matrix of the 1000 Genomes project data (panel (a)) shows two blocks for Europeans and Africans respectively, while the IBS matrix of the GAW18 data (panel (b)) shows about 20 small blocks for the families. The off-diagonal elements are all beyond 0.5, suggesting that people from different populations or families share a large proportion of common alleles as well.

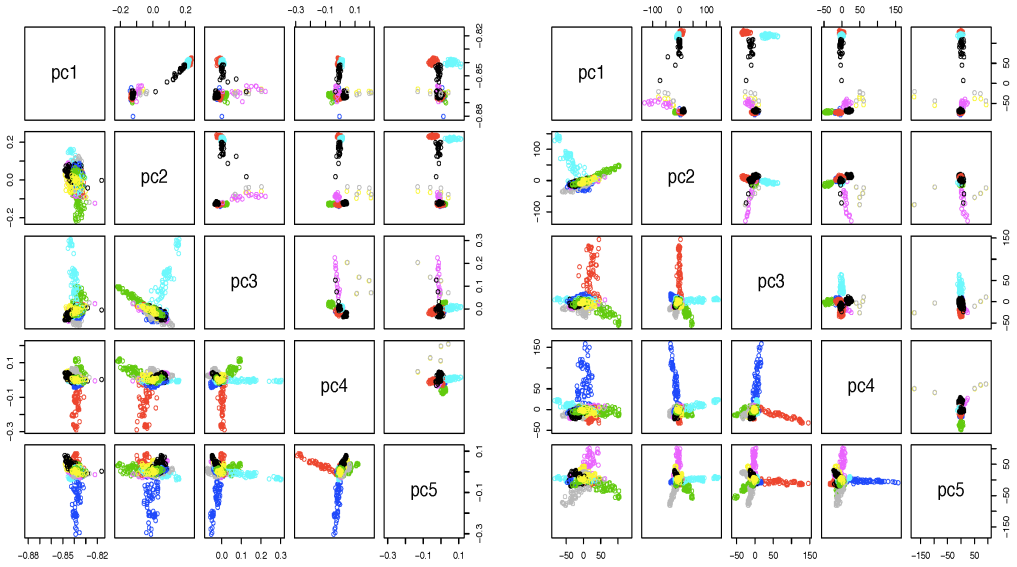
Figure C.1: The IBS matrix estimated with (a) 31292 pruned CVs in the 1000 Genomes project data and (b) 31544 pruned CVs in the GAW18 data.



(a) (b)

The top 5 PCs of the IBS matrix and the covariance matrix appear to be somewhat different (Figure C.2). For the 1000 Genomes project data (panel (a)), two continental groups are well separated but not all for the subgroups; for the GAW18 data (panel (b)), at most only 6 families are distinguishable, e.g. by PC3 and PC5 of both the IBS matrix and the covariance matrix, while the remaining families are mixed.

Figure C.2: PC-plots of the (a) IBS matrix (b) covariance matrix. The upper diagonal panel is for the 1000 Genomes project data and the lower diagonal panel is for the GAW18 data with different subgroups or families in different colors.



(a) (b)

C.2 In the presence of only population structure

The simulation model is the same as that for the GAW18 data. In general, under null hypothesis, PCR has the best control of the Type I error to be around 0.05 and λ around 1. GEMMA is still a little inflated for $\sigma_g^2 = 10$, possibly due to the large discrepancy of the IBS matrix and the true covariant matrix of y. EMMAX is conservative because the SNPs to be tested are all strongly associated with population structure, so estimating the variance under null hypothesis will cause the over-estimation of variance and thus larger p-values. For example, when $\sigma_g^2 = 10$ and $\sigma^2 = 90$, EMMAX had a Type I error of 0.040 (λ=0.898) and GEMMA 0.064 (λ=1.120). The means of the estimated genetic effect ($\hat{\alpha}$) were all near 0 without obvious bias.

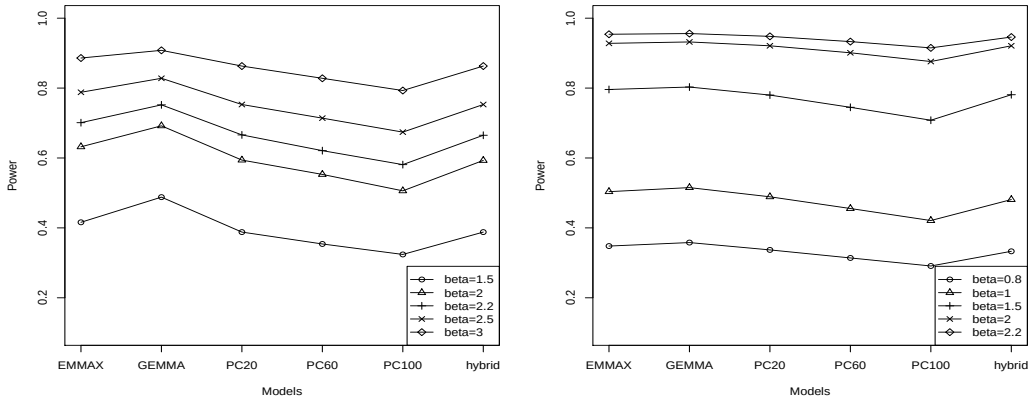
Figure C.3 shows the power comparison of different methods with different genetic effect α_1. The results are similar as using the GAW18 data. As the SNPs

tested here are strongly associated with population structure, we can see a substantial power loss of all methods compared to the results of GAW18 data. For example, for $\sigma_g^2 = 10$ with $\alpha_1 = 2.2$, the power of EMMAX is 0.701, GEMMA 0.752, PCR with 20 PCs is 0.666 and with 100 PCs is 0.581, the hybrid model is 0.665; for $\sigma_g^2 = 90$ with $\alpha_1 = 2.2$, EMMAX has a power of 0.954, GEMMA 0.948, PCR with 20 PCs 0.948 and the hybrid model is 0.954. Overall, all methods lose power evenly. This phenomenon has been pointed out by several papers that all methods would lose power to detect the signal SNPs which are related to population structure [30, 13, 87].

Table C.1: Association testing with population stratification of continental groups using the 1000 Genomes project data

set-up		EMMAX	GEMMA	PCR	hybrid	OLS
$\sigma_g^2 = 10$, $\sigma^2 = 90$	Type I	0.040	0.064	0.048	0.048	0.088
	λ	0.898	1.120	0.958	0.957	1.286
	$\hat{\mathrm{E}}(\hat{\alpha}_1)$(SD)	0.0020(0.833)	0.001(0.906)	-0.002(0.967)	0.001(0.969)	0.001(0.933)
	$\hat{\mathrm{E}}(\hat{SE})(SD)$(SD)		0.830(0.233)	0.952(0.245)	0.963(0.245)	0.789(0.226)
	$\hat{\sigma}_g^2$(SD), $\hat{\sigma}^2$(SD), $\hat{h}(SD)$		27.871(62.452), 85.549(15.980), 0.077(0.167)			
$\sigma_g^2 = 60$, $\sigma^2 = 40$	Type I	0.048	0.066	0.049	0.048	0.324
	λ	0.939	1.052	0.980	0.986	3.921
	$\hat{\mathrm{E}}(\hat{\alpha}_1)$(SD)	-0.003(0.690)	-0.003(0.725)	0.002(0.742)	-0.003(0.742)	0.009(1.263)
	$\hat{\mathrm{E}}(\hat{SE})(SD)$		0.677(0.182)	0.728(0.187)	0.729(0.188)	0.612(0.176)
	$\hat{\sigma}_g^2$(SD), $\hat{\sigma}^2$(SD), $\hat{h}$(SD)		63.508(57.691), 39.118(13.931), 0.295(0.258)			
$\sigma_g^2 = 90$, $\sigma^2 = 10$	Type I	0.047	0.051	0.050	0.048	0.500
	λ	0.963	0.999	0.990	0.964	8.427
	$\hat{\mathrm{E}}(\hat{\alpha}_1)$(SD)	-0.005(0.543)	-0.005(0.550)	0.005(0.567)	-0.005(0.564)	0.013(1.424)
	$\hat{\mathrm{E}}(\hat{SE})(SD)$		0.533(0.138)	0.552(0.142)	0.554(0.142)	0.475(0.138)
	$\hat{\sigma}_g^2$(SD), $\hat{\sigma}^2$(SD), $\hat{h}$(SD)		87.259(33.564), 10.620(7.740), 0.680(0.244)			

Figure C.3: Power of association testing using the 1000 Genomes Project data



(a) $\sigma_g^2 = 10$, $\sigma^2 = 90$

(b) $\sigma_g^2 = 90$, $\sigma^2 = 10$

C.3 In the presence of an environmental confounder

We first only consider an environmental risk ($\sigma_g^2 = 0$) without a structure. θ_1 and θ_2 are set different for EURs and AFRs and fixed throughout the scan of all CVs (Table C.2). We see that the results can be quite misleading without considering the environmental risk, as the Type I error rates of OLS are about 0.9 due to the confounding effect. More interestingly, we notice that as θ_1 and θ_2 become more divergent, EMMAX and GEMMA are less effective in correcting the underestimated p-values, ending up with inflated Type I error rates. For example, when $\theta_1 = -1$ and $\theta_2 = 1$, EMMAX has a Type I error=0.057, GEMMA has 0.061 and PCR has 0.046, but when $\theta_1 = -2$ and $\theta_2 = 2$, the Type I error rates for EMMAX and GEMMA both increase to about 0.1 while PCR maintains it to be 0.044.

Of course the setting above is no very realistic. Then we consider a scenario with both population structure, generated as $u \sim N(0, \sigma_g^2 K)$ with $\sigma_g^2 = 60$ and $\sigma^2 = 40$, and an environmental risk as different intercepts for EURs and AFRs. Table C.3 shows that when we only consider different environmental factors for 2 continental groups ($k = 2$), PCR is a little conservative (Type I error=0.041) while both EMMAX (Type I error =0.116) and GEMMA (Type I error=0.119) are unsuccessful in controlling inflations. The reason for this observation has been

discussed in Chapter 5, which is mainly due to the incapability of defining the true phenotype covariance by the IBS matrix alone when an extreme environmental effect is in presence. The hybrid model yields almost equal Type I error (0.041) and λ (0.887) as PCR with 20 PCs. As we consider different environmental effect for subgroups ($k = 10$), EMMAX (Type I error=0.061) and GEMMA (Type I error=0.061) could largely control the inflation while PCR was less conservative.

Table C.2: Association testing results with only an environmental risk using the 1000 Genomes project data

		EMMAX	GEMMA	PCR	hybrid	OLS
	TypeI	0.057	0.061	0.046	0.050	0.938
θ_1=-0.5, θ_2=0.5	$\hat{E}(\hat{\alpha}_1)$(SD)	-0.015 (0.103)	-0.015 (0.104)	0.005(0.101)	-0.005(0.104)	0.143 (0.336)
	TypeI	0.072	0.075	0.046	0.049	0.999
θ_1=-1, θ_2=1	$\hat{E}(\hat{\alpha}_1)$(SD)	-0.022 (0.113)	-0.023 (0.114)	0.005 (0.103)	-0.005 (0.106)	0.308 (0.706)
	TypeI	0.101	0.104	0.044	0.046	0.999
θ_1=-2, θ_2=2	$\hat{E}(\hat{\alpha}_1)$(SD)	-0.030 (0.146)	-0.039 (0.147)	0.005 (0.112)	-0.004 (0.114)	0.637 (1.454)

Table C.3: Simulations with $u \sim N(0, \sigma_g^2 K)$ and an environmental factor. k stands for the number of clusters, with each cluster sharing the same environmental factor.

		EMMAX	GEMMA	PCR					hybrid	OLS
				20	40	60	80	100		
k=2	Type I	0.116	0.119	0.041	0.0404	0.038	0.037	0.036	0.041	0.947
	λ	1.819	1.844	0.891	0.855	0.823	0.810	0.800	0.887	174.366
k=10	Type I	0.061	0.061	0.052	0.050	0.049	0.046	0.045	0.046	0.675
	λ	0.978	0.979	0.927	0.924	0.909	0.910	0.894	0.876	22.496