ABSTRACT

MICLAUS, KELCI JO. Addressing Sources of Bias in Genetic Association Studies. (Under the direction of Dr. Russ Wolfinger and Dr. Jason Osborne).

Genome-wide association studies (GWAS) have become a popular method for the discovery of genetic variants associated with complex diseases or traits. As the size and scope of these studies increase in order to obtain higher power for determining significant associations, careful consideration of population structure becomes paramount. If individuals in a study come from different ethnic or ancestral backgrounds, variation in allele frequencies and disproportionate ancestry representation in cases and controls can lead to inflated Type I error rates. Over the years, several methods for controlling population stratification have been introduced, many of which rely on the use of multivariate dimension reduction methods. An important aspect of population stratification is to determine which loci exhibit evidence of population allele frequency differences. We introduce a method based on Hardy-Weinberg Disequilibrium to find substructure-informative markers coupled with the use of nonmetric Multidimensional Scaling (NMDS) in order to visualize population structure in a sample. We extend the use of NMDS in conjunction with nonparametric clustering to develop a test for association that corrects for population stratification. We show that NMDS is a preferable visualization technique for detecting multiple levels of relatedness within a set of individuals and that the subsequent test correction model is a more powerful test under realistic scenarios. Recent research has shown that technical bias due to differential genotyping errors between cases and controls can also inflate the Type I error rate, possibly an even more severe source of bias in GWAS. Current genotype calling algorithms rely on processing samples in batches due to computational constraints as well as concerns of differences in DNA collection, lab preparation and heterogeneous samples that can skew results of genotype calls. This thesis also addresses possible bias caused by differential genotyping due to batch size and composition effects for the widely used BRLMM algorithm recommended for the Affymetrix GeneChip Human Mapping 500 K array set. Samples obtained from the Wellcome Trust Case Control Consortium are utilized to determine differential results due to genotype calling batch differences.

Addressing Sources of Bias in Genetic Association Studies

by
Kelci Jo Miclaus

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2009

APPROVED BY:

Dr. Dahlia Nielsen	Dr. Lexin Li
Dr. Russ Wolfinger Chair of Advisory Committee	Dr. Jason Osborne Co-chair of Advisory Committee

DEDICATION

To my mother, Michael MacLaughlin, and my husband, Radu Miclaus. Without your love, patience and encouragement I never would have reached my goals.

BIOGRAPHY

Kelci Jo Miclaus was born on May 5^{th}, 1982 in Gunnison, Colorado to Charles and Michael MacLaughlin. The MacLaughlin family moved to Clarksville, Missouri when she was 10 years old to live and work on a farm. Among summers of hauling hay and seasons of watching calves being born, Kelci attended Clopton High School where she played basketball and softball and suffered through boring math classes. She then went on to attend college at the University of Missouri-Columbia where she tried out majors in physical therapy, nuclear medicine, psychology, and communications. When she was forced to take a statisics class, she realized that this kind of math made sense to her and she graduated Summa Cum Laude with a bachelor's degree in Statistics in May 2004. Kelci was married to Radu Miclaus in the summer of 2004 and happily remains so. She accepted a fellowship to attend graduate school at North Carolina State University in Raleigh, North Carolina. She received her Master's degree in Statistics in May, 2006 from NCSU. This thesis is part of her work for completing the Ph.D. program in Statistics. While in graduate school, Kelci has gained experience working in the field of statistics with internships at Merck Laboratories, GlaxoSmithKline, and SAS Institute. She currently lives in Durham, NC with her husband and two unruly bulldog mutts.

ACKNOWLEDGMENTS

I would like to thank my co-advisor Russ Wolfinger for all the time, consideration and advice he has provided. I have learned so much about all aspects of statistics and genetics under his direction. I would also like to thank my co-advisor Jason Obsorne and the rest of my committee, Dahlia Nielsen and Lexin Li, for their time and contributions that enhanced the quality of my research. I thank the collaborators from the MicroArray Quality Control GWA working group and Wendy Czika from SAS Institute for their contributions as well. A big thanks goes out to the faculty and staff in the NCSU statistics department. Finally I thank my friends and family for their support and faith in me throughout this process.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

1 Understanding and Controlling Bias in Genetic Association Studies 1
1.1 Introduction 1
1.2 Population Stratification Bias 5
1.2.1 Population Structure Among Individuals 5
1.2.2 Adjustment Methods for Stratification 11
1.3 Differential Bias due to Genotype Calling 15

2 SNP Selection and Multidimensional Scaling to Quantify Population Structure 18
2.1 Introduction 18
2.2 SNP Selection Materials and Methods 20
2.2.1 Hardy-Weinberg Disequilibrium SNP Selection 20
2.2.2 Controlling the Confounding Effect of LD 21
2.2.3 SNP Selection Using the Test for HWE 22
2.2.4 Distance Metrics 23
2.2.5 Non-metric MDS Algorithm 25
2.3 Results 26
2.3.1 Simulation Details 26
2.3.2 Simulation Results 27
2.3.3 Empirical Relationship of F_{ST} and HWD SNP Selection 31
2.3.4 Application to HapMap Samples 31
2.4 Discussion 33

3 Population Stratification Correction in Association Studies Via Nonmetric Multidimensional Scaling 36
3.1 Introduction 36
3.2 Methods 39
3.2.1 Nonmetric Multidimensional Scaling 39
3.2.2 Nonparametric Clustering 40
3.2.3 Statistical Model for PS Adjusted Association Analysis 42
3.3 Results 43
3.3.1 Disease Association Simulation Results 43
3.3.2 Varying Levels of Relatedness and Visualization 47
3.3.3 Clustering Comparisons 48
3.3.4 SNP-Ethnicity Interaction Association Results 50

3.4 Discussion 52
3.5 Supplemental Information 52

4 Evaluating the Effects of Batch Composition and Size on BRLMM Genotype Calls for the Affymetrix 500K Array 54
4.1 Background 54
4.2 Results 57
4.2.1 Quality Control 57
4.2.2 Association Testing Results 60
4.2.3 Stringent Call Rate Threshold Evaluation 64
4.2.4 Statistical Tests for Batch Effects 65
4.3 Discussion 69
4.4 Materials and Methods 71
4.4.1 WTCCC Data 71
4.4.2 BRLMM Genotype Calling 71
4.4.3 Batch Effect Schema 72
4.4.4 Quality Control and Association Analysis Methods 73
4.4.5 Generalized Linear Mixed Models to Test Batch Effects 74

5 Conclusions and Avenues for Future Research 75
5.1 Bias Due to Population Structure 75
5.2 Bias Due to Genotype Calling 77

Appendices 86

A Chapter 2 Supplementary Details 87
A.1 MDS Algorithm 87
A.2 Supplemental Material 88
A.2.1 Rogers' Distance Manipulation 88
A.2.2 Plots of Fit Statistics for NMDS 89
A.2.3 Confounding Effect of Systematic LD 90

B Chapter 3 Supplementary Tables 93

LIST OF TABLES

Table 1.1 Allelic Contigency Table For Case-Control Association Tests 11

Table 2.1 LD Confounding SNP Proportions for HWD Selection 29

Table 2.2 LD Confounding SNP Proportions and F_{ST} for Various SNP Selections 29

Table 3.1 Chi-Squared Test with no PS Correction: Type I Error and Power 45

Table 3.2 Nonparametric and K-medoids Cluster Performance 48

Table 4.1 List of Significant SNPS for BRLMM Batch Sets with Stringent QC 64

Table 4.2 BRLMM Batch Effect Parameters for GLMMs 65

Table 4.3 Hypotheses and Estimates for Testing Batch Size and Composition Differences 67

Table B.1 Type I Error Estimates for EIGENSTRAT, NMDS, and CMDS, R=1.3 93

Table B.2 Power Estimates for EIGENSTRAT, NMDS, and CMDS, R=1.3 94

Table B.3 Type I Error Estimates for EIGENSTRAT, NMDS, and CMDS, R=1.5 95

Table B.4 Power Estimates for EIGENSTRAT, NMDS, and CMDS, R=1.5 95

Table B.5 Power and Type I Error Estimates for the SNP*Cluster Interaction Test 96

LIST OF FIGURES

Figure 2.1 HWE Test Quantile Plots for Confounding and Substructure Informative SNPs 28

Figure 2.2 Nonmetric MDS Coordinate Results under SNP Selection Criteria 30

Figure 2.3 Empirical Relationship between F_{ST} and HWD 32

Figure 2.4 Coordinates via PCA and NMDS for HapMap Populations 33

Figure 2.5 Coordinates via PCA and NMDS for HapMap CEPH Population 34

Figure 3.1 Nonparametric Density Estimation Histogram 42

Figure 3.2 Type I Error and Power Rates for PS Correction with Relative Risk=1.3 . . 46

Figure 3.3 Type I Error and Power Rates for PS Correction with Relative Risk=1.5 . . 47

Figure 3.4 NMDS and CMDS Display of Populations with Varying Levels of Relatedness 48

Figure 3.5 Nonparametric and K-medoids Cluster Visualization Example 49

Figure 3.6 Type I Error and Power for SNP*Cluster Interaction Effect 51

Figure 4.1 Quality Control Exclusion Results for BRLMM Datasets 58

Figure 4.2 Counts of SNPs with Differential QC Results for BRLMM Datasets 59

Figure 4.3 Genome-Wide Scatter Plots of Batch Set p-values 61

Figure 4.4 Venn Diagrams and Discordant Counts of Significant SNPs for BRLMM Batch Sets 62

Figure 4.5 Quantile Plots of Observed vs. Expected p-values for BRLMM Batch Sets 66

Figure A.1 Nonmetric MDS Fit Statistics Plots 91

Figure A.2 Systematic Linkage Disequilibrium: PCA vs. NMDS 92

1

Understanding and Controlling Bias in Genetic Association Studies

1.1 Introduction

The field of statistical genetics has grown significantly over the last few decades, particularly in the area of genetic disease association studies. Analysis of variability in Single Nucleotide Polymorphisms (SNPs) has led to many significant findings of genes associated with disease etiology. With technology making it affordable to carry out whole-genome association scans, many studies genotype thousands of individuals and hundreds of thousands of markers for analysis. Yet with the benefits of these large scale studies comes drawbacks that include loss of power (due to the sheer amount of tests carried out) and spurious association that may be caused by population stratification (resulting in an inflated false positive rate). In order to increase power, it is becoming popular to do collaborative association studies that combine samples from different studies. This must be done carefully to avoid unwanted associations due to heterogeneity in the individuals as they may belong to separate populations with varying allele frequency patterns. Thus is it crucial for researchers to understand and explore the structure of genetic data.

Much research has been devoted to unraveling the complex nature of genetic variation, in particular with population differences and individual classification. Findings show that human genetic variation can in part be attributed to geographic structure and migration patterns where individuals within a close region will tend to be more genetically similar

[1]; yet increase in the frequency of intercontinental travel over the last few decades invariably complicates this relationship. Differences in allelic frequencies due to racial and ethnic backgrounds have been used for population classification in research studies [2]. Clearly population variation has a large impact on the structure of genetic data, but it is not a complete picture. Findings show that 85-90% of the total genetic variability can be found between individuals within one of the major (Old World) continents (Africa, Asia, Europe); while only 10-15% is accounted for by differences between such populations [1]. Therefore it is necessary to not only quantify major population differences but more subtle individual structure within populations.

Genetic distance measures such as Reynolds' Distance (based on the coancestry coefficient) [3] and Rogers' Distance [4] are commonly used to compare allele frequencies between known populations to obtain measures of how genetically different two or more populations may be; yet many times a study is conducted where population membership or possible relatedness of individuals is unknown. Genetic distance measures have not been widely exploited for applications on individuals in order to visualize genetic structure. Currently, Principal Components Analysis (PCA) is a popular application for detecting genetic structure in populations, yet the impact of the theoretical drivers of the method has not been assessed. Applications of multivariate dimension reduction methods for quantifying individual level genetic structure will be explored in Chapter 2, particularly Nonmetric Multidimensional Scaling (NMDS). Further application of NMDS combined with nonparametric clustering analysis is used to propose a model for testing for genetic disease association while controlling for the confounding effect of population stratification. Chapter 3 introduces the new method and compares performance with established methods that are based on PCA as well as Classical Multidimensional Scaling (CMDS).

Chapter 2 is comprised of the article 'SNP Selection and Multidimensional Scaling to Quantify Population Structure' recently published in the journal *Genetic Epidemiology* [5]. The article briefly summarizes common genetic distance measures used for measuring population and individual genetic dissimilarity in order to obtain information about relatedness and clustering of individuals, and proposes the use of NMDS as a useful visualization tool to better interpret genetic structure (and how it can introduce bias in association studies). Certain characteristics of NMDS suggest that it may give a more accurate visualization of genetic patterns (on the individual level) compared to PCA. Applications of NMDS using a dissimilarity matrix of individual genetic distances is evaluated using simulated and real data. Additionally, a method for selecting ancestry informative markers based on the test for Hardy-Weinberg Equilibrium (HWE) is proposed and developed in Chapter 2 through theoretical and empirical study. This method is formulated in conjunction with the use of NMDS as a visualization tool with supporting evidence as to why PCA may fail to provide an adequate graphical summary when using SNPs based on HWE test results.

Chapter 3 contains an article where we extend the use of NMDS and propose a formal model that controls population stratification bias in the case-control association test. The contents of Chapter 3 are currently under review for journal publication. The NMDS coordinates for each individual in a study are used as continuous covariates in a logistic model to capture the effect of varying ethnic groups. Additionally, the coordinates are used as input to perform nonparametric clustering in order to assign individuals to discrete clusters, which are also used as covariates in the logistic model. The inclusion of the continuous NMDS dimension coordinates aids in capturing more subtle genetic variation (both within and between subpopulation groups) among individuals, while the discretized classification covariate enhances the correction of PS bias on markers that show more dramatic genetic differentiation than may have been captured by the markers used to form the NMDS dimensions. The model builds upon research done by Price et al. [6] in the development of EIGENSTRAT (a correction method employing PCA) and the more recent work done by Li and Yu [7] that makes use of Classical Multidimensional Scaling (CMDS) in a model to adjust for population structure in association studies [7].

Chapter 4 presents research into understanding another source of bias in genome-wide association studies (GWAS): differential bias due to inaccuracies in genotype calling. The implementation of high-throughput microarray technologies for SNP discovery and analysis has been a catalyst for successes with GWAS, which required the development of multi-chip clustering algorithms for genotype calling in order to estimate probe effects and allele signals for SNPs simultaneously [8]. As the size of the arrays used for genotyping increases from 10,000 to 100,000 to 500,000 to 1,000,000 SNPs, genotype calling algorithms continue to improve and advance [9, 10, 11, 12, 13]. Despite continuous improvements in accuracy and SNP call rate, genotyping errors have been shown to cause spurious results in association studies [14, 15, 16, 17]. Clayton et al. [14] reported that bias due to differential genotyping error between cases and controls can equal or even outweigh the effects of population stratification in terms of inflation to the test statistic in association studies.

One of the most common platforms used for GWAS currently is the Affymetrix GeneChip Human Mapping 500K array set, which was recently used by the Wellcome Trust Case Control Consortium (WTCCC) GWA study of several complex diseases [18]. Affymetrix recommends the use of the Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM) [11] genotype calling algorithm with the Affymetrix 500K array set, thus making BRLMM one of the most widely applied calling algorithms. Due to computing restrictions and the scale of current GWAS that contain thousands of individuals, calling algorithms (BRLMM included) must typically process individuals for genotype calls in batches. Using data obtained from the WTCCC for coronary artery disease (CAD) patients and a set of control individuals, we explore the effects that genotyping differences due to batch size (the number of individuals processed at a time) and batch composition (the distribution of cases and controls within a batch) have on the results of quality control and association testing in GWAS for the Affymetrix 500K array set using the BRLMM calling algorithm. We find that batch size and composition differences used for BRLMM processing can result in $2-3\%$ discordance in association results and so must be considered for reliable, comparable results in GWAS analyses. This work is done in collaboration with the Genome-Wide Association Working Group (GWAWG) of the MicroArray Quality Control Consortium (MAQC) [19] in a continuing effort to better understand how technical bias due to genotyping platforms and genotype calling algorithms can affect the results of large-scale association studies.

The research in Chapters 1-4 provides novel approaches to visualizing genetic data, proposes promising new methods for understanding, quantifying, and controlling (the potential bias caused by) population structure, and addresses bias due to more subtle genetic patterns that are a technical result of genotype calling differences within the BRLMM algorithm for the common Affymetrix 500K array platform. Section 1.2 and Section 1.3 outline some existing methods and issues found in current research that pertain to genetic data structure and sources of bias due to population stratification or introduced by genotype calling. This overview provides a better understanding of the area and enhances the motivation for the research presented in this thesis.

1.2 Population Stratification Bias

This section gives a summary of past research pertaining to genetic heterogeneity among populations and individuals. Methodologies used to control stratification effects in disease association studies are also reviewed and discussed in order to give insight into the use and interpretation of pattern visualization methods and subsequent population stratification (PS) controlling association test methods.

1.2.1 Population Structure Among Individuals

Results from selected studies have highlighted genetic homogeneity of humans globally, pointing out that much of the variation is found within populations and that the structure among populations follows geographic continuity [1, 20]. This may be seen as undermining the usefulness of distinct clusters in general; yet a power study done by Bamshad et al. [20] showed that despite the overall homogeneity, individuals tend to cluster according to their ancestral origins (most clearly when the population is comprised of highly differentiated subpopulations). Moreover, the authors concluded that clustering analysis becomes a powerful tool for detecting human population structure when certain criteria are optimized. The results of their power study showed that the degree to which population structure can be distinguished is dependent on the amount of variation between the populations studied and their respective sample sizes; along with the number of markers used and how much information pertaining to ancestry the chosen markers contain.

A common criterion of an informative marker for the detection of structure is F_{ST}, which measures the amount of differentiation between populations. This measure is one of the F statistics proposed for studying population genetics by Wright [21] and will be discussed in more detail in Chapter 2. Bamshad et al. [20] corroborate the use of F_{ST} with the conclusion that the power of a marker (to measure population structure) is a function of F_{ST}. Thus with careful selection (in addition to a sufficient number) of loci, clustering methods can yield positive results for detecting population structure in individuals.

There are two main clustering approaches utilized in the detection and quantification of population structure among individuals. These two approaches consist of model-based and distance-based clustering methods. The pros and cons of both methods have been discussed in detail over the years. The most well-known and often cited model-based method is *Structure*, which was proposed by Pritchard et al. [22]. *Structure* uses a Bayesian approach to estimate the number of populations present in the data (K) and calculate probabilities of ancestry for each individual (hence allowing for admixture). Based on the posterior probabilities of membership; an individual is assigned to a cluster. The Bayesian clustering model assumes that while the population sample may contain some markers in Linkage Disequilibrium (LD) and Hardy-Weinberg Disequilibrium (HWD), there is linkage equilibrium and HWE within subpopulations. The goal then is to group individuals into subpopulations characterized by the allelic frequency patterns in order to minimize within cluster HWD among alleles at a marker and LD between markers.

Despite the fact that distance-based methods are much simpler and easier to visualize, Pritchard et al. [22] advocated the use of a statistically sophisticated model-based method under the reasoning that results of distance-based methods rely heavily on the chosen distance measure and graphical method and hence are not suited for fine-scale analysis. *Structure* has been shown to be an extremely useful tool and has been widely applied in many studies pertaining to population genetics and case-control association [2, 20, 1]. The downfall of *Structure* is computation cost and parametric assumptions. The method employs a Markov chain Monte Carlo (MCMC) method to obtain sample estimates of each individual's population membership and the allele frequencies for each marker in all populations. Due to the complexity of computation, the method is not feasible for studies that employ a large amount of markers - the method cannot be performed with typical GWAS. Additionally, the computation of allele frequencies are subject to the assumptions specified previously concerning HWE and linkage equilibrium[1] which may be violated in many situations.

Due to the computational inefficiencies and parametric assumptions necessary in *Structure*, recent research has turned back to nonparametric and distance-based methods in order to address the issue of population structure in genetic studies (particularly for GWAS). Gao and Starmer [23] proposed a clustering method "AW-clust" that uses the Allele Sharing Distance (ASD) given in Equation 1.1 in a hierarchical clustering framework to cluster individuals to subpopulations within a study. Following the authors' notation,

[1] The updated version of *Structure* can now account for LD within a population due to admixture, yet still does not perform optimally with background LD

using markers $l = 1, 2, ..., L$, the allele sharing distance for individuals 1 and 2 is defined as

$$D_{ASD} = \frac{1}{L}\sum_{l=1}^{L} d_l \tag{1.1}$$

where

$$d_l = \begin{cases} 0 & \text{if two alleles are the shared at marker } l \\ 1 & \text{if one allele is shared at marker } l \\ 2 & \text{if no alleles are shared at marker } l. \end{cases}$$

The method results in a hierarchical tree that the authors find to be flexible for inference on K depending on the amount of resolution needed to address the particular study involved. Using a random selection of 20,000 SNPs across the genome, the method was successful at differentiating Chinese (CHB) and Japanese (JPT) individuals using HapMap data with close to 100% accuracy. The method can also accommodate a study which may for certain purposes wish to group CHB and JPT as one population of Asian individuals by using a cutoff for K at a higher level of the tree. Thus the authors emphasize K to be a variable instead of a fixed parameter that must be determined. The "AW-clust" method is exemplary of the success of distance-based methods for population structure detection; yet the variability of K may be seen as a downfall as it may be difficult to determine the cutoff level within the hierarchical framework[2]. In some cases, other nonparametric and nonhierarchical methods may be preferred.

Principal Components Analysis (PCA) has recently become a popular nonparametric approach to quantify and visualize population heterogeneity. While PCA is by no means a new method to genetic data analysis, it has experienced a resurgence in the population genetics realm much due to the development of the method EIGENSTRAT [6]. EIGENSTRAT is a method designed for GWAS to test for association while controlling for population stratification (which will be discussed in more detail in Section 1.2.2) and is built upon theory of eigenanalysis applied on the individual level by Patterson et al. [25] and employed in the method SMARTPCA. The rational for the use of PCA to identify

[2] Note: Gao and Starmer [23] propose using the Gap Statistic [24] to provide an objective verification of K.

population structure is based on the distributional theory of eigenvalues. The method normalizes the genetic data and performs a singular value decomposition upon the covariance matrix to obtain “continuous axes of genetic variation” [6] that contain the coordinates for each individual based upon their genetic information in reduced dimensionality.

Patterson et al. [25] postulate that the first few axes of variation (principal components) based on the largest eigenvalues will quantify subpopulations that (may) exist among a population, which is supported with the following brief argument summarized from the article. To define our data matrix G, let $c_{i,l}$ denote the count of (variant) alleles for the i^{th} individual, l^{th} locus ($i = 1, 2, ..., N$, $l = 1, 2, ..., L$). The (i^{th}, l^{th}) element G in Equation 1.2 represents the centered, normalized genetic information for the study, where $\mu_l = \frac{\sum_i^N c_{i,l}}{N}$ is the individual mean for the l^{th} marker and $\hat{p}_l = \mu_l/2$ is an estimate of the corresponding allele frequency,

$$G_{i,l} = \frac{c_{i,l} - \mu_l}{\sqrt{\hat{p}_l(1 - \hat{p}_i)}}. \tag{1.2}$$

The N-1 eigenvalues obtained from $GG^{'}$ are ordered and Patterson et al. [25] describes theory that states that the largest eigenvalue λ_1 follows a Tracy-Widom (TW) distribution. The motivation for the use of this distributional aspect springs from the fact that in PCA, the principal (corresponding to the largest eigenvalue) eigenvector ($e^{[1]}$) from $GG^{'}$ will maximize the sum of squares,

$$S = \sum_{l=1}^{L} \left(\sum_{i=1}^{N} e_i^{[1]} G_{i,l} \right)^2$$

and describe the largest amount of variation. Likewise, the successive eigenvectors will also maximize S conditional on orthogonality among the previous vectors. When substructure is present in the data, alleles within a subpopulation will be more similar, with large differences between populations, and will drive the coordinates of the principal components. This will cause the sum of squares S to be large; hence the eigenvectors that maximize S are representative of distinguishing groups of similar allele frequencies in the data (subpopulations).

Therefore, the development of a formal statistical test for the largest eigenvalues (using the TW distributional theory) is also formally testing for population structure and the significant number of axes needed to quantify population heterogeneity. Following the logic that each vector that maximizes S will represent the variability between subpopulations, Price et al. [6] suggest that (if there are K subpopulations present) $K-1$ axes will be meaningful and significant for quantifying between subpopulation variance. Furthermore, following principal components $(e^{[K]}, e^{[K+1]}, ...)$ will exhibit within-subpopulation variance if present (or simply sampling error if subpopulations are homogeneous). The results presented by Patterson et al. [25] show highly accurate agreement between an unsupervised PCA method and a supervised ANOVA means comparison using known population labels in simulated sample data sets. They also provide evidence that not only can the method display discrete clusters but also more gradual population heterogeneity that follows a cline pattern. For smaller data sets where *Structure* was feasible, the results between SMARTPCA and *Structure* were synonymous.

Another contribution from their work involved the 'phase change phenomenon', where the ability to detect population differentiation can be quantified and determined dependent on the number of individuals N, the number of markers L, and F_{ST}. Namely, for two subpopulations with an equal number of samples, there exists a threshold:

$$F_{ST} = \frac{1}{\sqrt{NL}}$$

above which ($F_{ST} > 1/\sqrt{NL}$) evidence of structure in the data is easily detectable, while below this threshold population differentiation is not successful. This result was found to be generalizable to other methods (*Structure* and even supervised ANOVA) in addition to the proposed PCA method. Thus one can conclude that no matter how efficient the method is, studies of population structure must consist of many markers and individuals, along with well chosen markers to maximize F_{ST} in the data.

The recent work in developing PCA for use with population genetics has resulted in SMARTPCA (and EIGENSTRAT for PS controlled association testing) becoming favored over *Structure* as an easy-to-use, nonparametric tool that requires no prior knowledge or assumptions for the data or the origins of individuals within a study. Early successes with the method have prompted continued research into using dimension reduction techniques to infer structure. Paschou et al. [26] have added to the research with the development of a SNP selection tool based on PCA, again emphasizing the importance of using a well chosen subset of SNPs with which to infer population heterogeneity (a fact that was downplayed in the proposal of SMARTPCA as the authors intended the method for genome-wide application[3]). With reduction of costs for genotyping along with insight into loci that exhibit selection forces in mind, Paschou et al. [26] developed a method that acts as a variable selection tool to extract a subset of SNPs that produce principal components that are highly correlated with the eigenvectors derived from the complete data set.

In PCA dimension reduction, the resulting vectors that span the reduced subspace are linear combinations of the marker columns. The SNP selection tool PCASNPS (the name of the implemented software provided by the authors) works by ordering scores that consist of the squared sum of coefficients (from the eigenvectors) for the top significant principal components at each SNP column. The top few scores then correspond to the columns of markers that most drive the coordinates that span the reduced subspace. To determine the significant number of dimensions (vectors), the authors do not employ the Tracy-Widom distribution as in SMARTPCA. Instead they propose the use of random matrix theory to test iteratively if each component (and subsequent ones as ordered by the eigenvalues) forms a matrix that is significantly distinct from a randomly generated matrix.

In the development of PCASNPS, it was shown that coupling PCA (using the entire dataset) with a clustering tool (k-means was utilized in their work) results in almost 100% classification accuracy in cases of both between continental groups and within continents. It was also shown that using the subset of PCA correlated SNPs could also achieve the same results. One study done on a worldwide data set showed that the use of just 30 PCA correlated markers produced the same results (near perfect clustering) as using the over 9,000 original SNPs. Additionally, these selected SNPs performed equally well as SNP selection based on an F_{ST} correlated measure, with a surprising finding of high amounts of overlap between markers chosen by the two methods (many markers that did differ between the methods were still found to be in LD with markers from the alternative method).

The results of the past research for population genetics outlined above indicates the trend of turning to easy, nonparametric methods such as distance-based clustering and PCA for understanding population structure in genetic data. The success of these methods is also influencing tools for correcting for population stratification as outlined in the following section; yet these methods do have downsides and drawbacks which will be discussed.

[3] When using this method in association studies, the authors do suggest (in the documentation of their program) using a set of substructure informative SNPs known to be not linked to disease if the study consists of less than 100,000 markers

Table 1.1: Allelic Contigency Table For Case-Control Association Tests

	Alleles at Marker M		
	A_1	A_2	Total
Case	$r_1 + 2r_2$	$r_1 + 2r_0$	$2R$
Control	$s_1 + 2s_2$	$s_1 + 2s_0$	$2S$
Total	$n_1 + 2n_2$	$n_1 + 2n_0$	$2N$

1.2.2 Adjustment Methods for Stratification

Multiple methods have been introduced that not only quantify, but develop tests to correct, adjust and control for population heterogeneity in association studies. Two of the earliest methods are Genomic Control (GC) [27] and Structured Association (SA) [28]. These two methods are representative of the two general approaches used to correct for population substructure. The first approach is to scale the usual test statistic for association by a factor that adjusts for the presence of population structure (GC is exemplary). The second approach aims to estimate the number of existing subgroups and probability of individuals belonging to a subgroup of the population, then carry out a test of association contingent on subpopulation stratification (SA is exemplary).

The premise of adjusting for population stratification to reduce the false positive rate in association studies was laid out by Devlin and Roeder [29]. Using the allelic contingency table given in Table 1.1 with notation taken from Sasieni [30], the Armitage Trend test [31] is expressed as

$$\boldsymbol{T} = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{R(N - R)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}. \tag{1.3}$$

If a population is truly homogeneous, under the null hypothesis that a marker is not associated with disease, then $\boldsymbol{T} \sim \boldsymbol{\chi}_1(0)$. In order to quantify bias and overdispersion due to population structure, the variance of the trend test statistic was derived. The following is an outline of their notation. Let $a_1, a_2, ..., a_K$ and $b_1, b_2, ..., b_K$ represent the number of cases and controls respectively for K subgroups existing within the population. If a population exhibits structure (due to relatedness specifically), then subgroups show allelic correlation within its members. This quantity is denoted F and is related to the fact that alleles within related individuals may be identical by descent (IBD). Finally let p and q denote the probability of the A_1 and A_2 allele respectively at marker M. Devlin and Roeder [29] showed that the variance of $\boldsymbol{T}$ (assuming $R = S$) under the null hypothesis that a marker is not linked to disease is given by Equation 1.4,

$$Var(\boldsymbol{T}) = 4Rpq(1+F) + 4Fpq\sum_{k=1}^{K}\{a_k(a_k-1) + b_k(b_k-1) - 2a_kb_k\}. \tag{1.4}$$

The derivation makes it clear that inflation of the variance is sensitive to correlation within subgroups F as well as unequal representation of a subpopulation among the cases and controls ($a_k \neq b_k$). Genomic Control sought to estimate the inflation to the test statistic for association and adjust the original test statistic for it; in contrast, Structured Association incorporated the inferred population information obtained by the aforementioned software program *Structure* into a likelihood-based test carried out by the software program STRAT [28]. Advances to both approaches have been made over the last several years with the majority focused on the latter approach.

While Genomic Control is a simple, easily applied solution to PS bias; a critical assumption is a constant effect of population structure on the inflation of the test statistic across the genome. Specifically, the amount of correlation of alleles (the value of F) must be similar over all markers and populations. Weir et al. [32] found that there exists substantial heterogeneity in the sample variance in allele frequencies among sampled populations with large differences across chromosomes. Similar results over the years (on the complexity of population differentiation across the genome) have led to more research devoted to estimating population membership similar to the *Structure* method and subsequent test corrections. Yu et al. [33] extended the use of *Structure* in a mixed model approach to test disease association with the genotypes ($\boldsymbol{X}$) to disease (y) of the form

$$y = \boldsymbol{X}\alpha + \boldsymbol{Q}\beta + \boldsymbol{I}u + \epsilon,$$

where $Var(u) = 2\boldsymbol{K}\sigma_g^2$ and $Var(\epsilon) = \sigma_e^2$. The model uses the matrix of estimated genetic information (population membership) as a fixed effect ($\boldsymbol{Q}$, given by *Structure*) with an additional random effect (u) quantifying genome-wide relatedness. The random effect adds a matrix of the kinship coefficients ($\boldsymbol{K}$) into the phenotypic covariance matrix to incorporate multiple levels of relatedness (between and within subpopulations) into the PS correction model. The complexity of the *Structure* algorithm does not allow for application with large GWAS and limits the mixed model approach. Zhao et al. [34] explored modifications to the mixed model approach by replacing the $\boldsymbol{Q}$ matrix (originally the vectors containing probabilities of ancestry from *Structure*) with principal component vectors to indicate ancestry, similar to the EIGENSTRAT correction method.

The motivation behind EIGENSTRAT was discussed in Section 1.2.1. Using principal components analysis to estimate 'axes of variation' corresponding to ancestry, the subsequent test correction adjusts the genotype for the i^{th} individual at marker M_l using coordinates a_i for each eigenvector to obtain a weighted allele count

$$c_{i,l}^{adj} = c_{i,l} - \gamma_l a_i, \tag{1.5}$$

where

$$\gamma_l = \frac{\sum_i a_i c_{i,l}}{\sum_i a_i^2}$$

is a weight derived from the regression of genotype on ancestry [6]. The adjustment is similarly made on the phenotype or trait of interest (i.e. $0, 1$ disease indication). These adjustments to genotype and phenotype are iteratively performed over all the eigenvectors obtained from PCA used to infer ancestry. A generalized trend test of the form given in Equation 1.6 is used for testing association while controlling for population stratification, where $t_{i,l}^{adj}$ represents the adjusted trait value of interest and $K^{'}$ is the number of eigenvectors used for adjustment:

$$\boldsymbol{\chi}^{\mathbf{2}}_{eigenstrat} = (N - K^{'} - 1) \times Corr(c_{i,l}^{adj}, t_{i,l}^{adj}) \sim \boldsymbol{\chi}^{\mathbf{2}}_{1}(0). \tag{1.6}$$

Price et al. [6] note that their correction method is analogous to creating matched case-control data to control inflation to the Type I error rate. EIGENSTRAT has become a popular PS correction method and has been applied in many studies including the Wellcome Trust Case Control Consortium [18]. It has been shown to be especially useful for discerning continuous patterns of subtle variation in European American samples where little PS was previously believed to exist [35]. One concern with EIGENSTRAT (and other PCA-based dimension reduction methods) is distortion to the PC coordinates due to LD structure among markers used to infer population structure, where variability from individuals with genotypic information in common across LD blocks may distort PCA results. PCA-based approaches are also highly influenced by outliers and require either an outlier removal process or a subset of representative samples to avoid principal components that only explain such variability. Research presented in Chapter 2 offers a SNP selection tool based on HWD that is more appropriately coupled with NMDS to select a set of markers better suited for elucidating accurate patterns population structure.

Li and Yu [7] employ Classical Multidimensional Scaling (CMDS) in a method that builds on EIGENSTRAT to better quantify not only subtle continuous patterns of ancestry but discrete ethnic/population groups. CMDS is a dimension reduction method that performs spectral decomposition to obtain eigenvectors of a given similarity or dissimilarity matrix. This can result in better quantification of population structure because a more appropriate choice of distance (such as genetic distance measures) can be employed to obtain individuals' coordinates along axes that better represent relatedness. Li and Yu [7] show that CMDS coordinates coupled with cluster indicators in a logistic regression setting provide improved control of Type I error rates. When using an inner product similarity metric, CMDS coordinates are analogous to PCA. The notable feature in the CMDS method is to use k-medoids clustering (where the number of clusters is chosen using the Gap statistic [24]) to form a covariate in the model for capturing discrete population group variability. The characterization of both continuous patterns (CMDS coordinates) and discrete patterns (cluster membership from k-medoids) of population structure controls Type I error at the nominal level in cases where EIGENSTRAT fails to do so, specifically in the case of discrete populations and when testing highly differentiated SNPs. Li and Yu [7] explore the use of other more genetic-oriented metrics, including a genotype-matching-based metric and an allele sharing metric, which we show in Section 2.2.4 to be synonymous with Identity By State (IBS) distance [36] as well as Gower's [37] distance with a range standardization.

We advance the research for methods used to control PS in association studies with the proposal of using Nonmetric Multidimensional Scaling (NMDS) coordinates calculated from the IBS distance matrix, which offers improvements on both EIGENSTRAT and CMDS for association testing in the presence of population stratification. The NMDS method relies on obtaining Euclidean coordinates in a reduced dimension space by an iterative optimization routine as opposed to singular value decomposition (SVD). This routine aims to minimize differences in the reduced dimension with the full distance matrix and tends to produce a more representative picture of the true data instead of simply producing coordinates that find the directions of maximal variance explained; this can allow for better estimation of multiple levels of relatedness (PS) in the data. The downside of NMDS comes in computational time in comparison with PCA as the optimization scheme can become cumbersome when the dimension of the matrix exceeds more than a few thousand individuals. More details on the NMDS algorithm are found in Chapters 2 and 3 as well as in Appendix A.1.

Population structure is a multifaceted, complex problem. The inflation of false positives is most simply attributed to differing allele frequencies that are inherent to subgroups in the population, along with variable disease susceptibility. Confounding effects are not only seen when the population sampled is heterogeneous due to different ethnic groups present; confounding may also be attributed to *cryptic relatedness*, where there is correlation in the cases and/or controls (individual relatedness) that is unbeknownst to the researcher. Spurious association can also arise due to population admixture, where individuals have differing proportions of ancestry from original subgroups; however a recent study that modeled spurious association for both admixture and discrete subgroups found admixture to be a less severe cause of spurious associations [38]. Many of the methods proposed in the literature may be less desirable under different causes of population structure, which has led to a heated debate over the last years as to which methods perform best. In this dissertation we employ NMDS and nonparametric clustering to develop a method for controlling PS bias in association studies due mainly to discrete populations or ethnic groups, although NMDS can also be applied to admixed populations.

1.3 Differential Bias due to Genotype Calling

Population stratification has been a dominant concern for bias in genetic studies, although recently other concerns have come to light with the use of high throughput microarray technologies for SNP data. For example, Price et al. [6] found that sometimes principal components computed in the EIGENSTRAT method were highly correlated with assay effect and genotype call rate. If these effects are confounded with case/control status and the researcher unknowingly includes those principal components in the model, then a loss of power will result. This highlights a source of bias in association studies that is due to technical errors in genotype calls, as mentioned in Section 1.1. Clayton et al. [14] found that differential errors in genotype calls for cases and controls were responsible for more than half of the 11.2% inflation to the test statistics (under the assumed null distribution for approximately 6,000 nonsynonymous SNPs) in a Type 1 diabetes association study. Other studies have reported similar results of bias due to genotyping errors [16, 39, 15, 17]. The Affymetrix GeneChip Human Mapping 500K array set is currently one of the most commonly used high throughput genotyping SNP platforms and Affymetrix recommends the use of the Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM) algorithm [11] for genotype determination.

The BRLMM algorithm begins by first normalizing and summarizing probe level intensities for the two alleles at each marker ('A' allele and 'B' allele) without a background correction in a SNP Robust Multiarray Average (RMA) process. BRLMM is an advancement to RLMM [10] with the addition of a Bayesian step that employs a Dynamic Model (DM) algorithm [9], an approach to call genotypes one chip and one SNP at a time, on a sample of SNPs to obtain estimates of the 'AA', 'AB', and 'BB' genotype cluster centers and variances. After quantile normalization and median polish summarization, the resulting S_A and S_B signals for the 'A' and 'B' alleles are transformed using the Cluster Center Stretch (CCS) transformation to allow for optimal clustering without heterozygous dropout. The CCS transformation is given as

$$CCS = \frac{asinh\left(T \times \frac{S_A - S_B}{S_A + S_B}\right)}{asinh(T)} \tag{1.7}$$

where T is a tuning parameter found to be useful for balancing homozygote and heterozygote clustering performance. The CCS transformed contrast is plotted against a strength measure that quantifies the overall brightness of the probe intensities given as

$$Strength = \log(S_A + S_B).$$

The Mahalanobis distances to clusters are then computed from each point x_i (the SNP intensity contrast for each sample) calculated as

$$\sqrt{(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)} \tag{1.8}$$

where μ_j is the cluster center for the j^{th} genotype ($j = 1, 2, 3$) cluster with covariance Σ_j, assumed to follow a multivariate normal distribution. Each point x_i is assigned a genotype by the minimum distance to a genotype cluster. The confidence of the genotype call for this point (i.e. individual at the particular SNP) is based on the ratio of the smallest distance to a cluster center (indicating the genotype assigned) and the next smallest distance. If the confidence does not meet a specified threshold, the call for the individual at that SNP is set to missing.

The simultaneous use of samples (i.e. the multi-chip aspect) of BRLMM is integral to the algorithm as the number and composition of samples used for forming clusters can produce differential results in terms of both genotype call and confidence assigned. This multi-chip clustering aspect is common with other genotype calling algorithms currently in use [18, 12, 13]. For this reason, it is imperative to understand the effect that batch size and composition has on downstream association analysis. The final portion of our work addresses the discordance that batch effects can introduce into large-scale association studies using samples obtained from the WTCCC [18].

This thesis outlines and addresses two major sources of potential bias, and hence discordance and lack of reproducibility, in genetic association studies: population stratification and batch effects in genotype calling algorithms. The development of methods for quantification, visualization, and appropriate test correction for population structure in case-control association studies presented in Chapters 2 and 3 constitute an important contribution to the set of analysis tools used for understanding the heritability of complex genetic diseases. The evaluation of discordance and bias introduced by genotype calls from the BRLMM algorithm in Chapter 4 highlights the necessity for proper handling and quality control of genetic data in order to obtain reliable, reproducible results in GWAS.

2

SNP Selection and Multidimensional Scaling to Quantify Population Structure

2.1 Introduction

Analysis of Single Nucleotide Polymorphism (SNP) variability has led to many significant findings of genes associated with disease etiology. Technology now enables whole-genome association scans for analysis; yet these large scale studies come with drawbacks such as loss of power and inflated false positive rates due to population stratification. Collaborative studies are becoming common to increase power, a solution that may cause spurious associations due to heterogeneity in individuals that may belong to separate populations with structurally different allele frequency patterns. It is crucial for researchers to understand, explore, and visualize the structure of the genetic data.

A plethora of methods [22, 27, 6, 23, 7] have been developed to control the confounding effect of population structure (PS) on disease association tests. An often sidelined part of such methods is selection of substructure-informative SNPs for quantifying the effect of PS. Devlin et al. [27] emphasized the use of "Null SNPs"[1] in Genomic Control to avoid bias in the inflation factor, although in exploratory studies it is difficult to obtain such a set. The current gold standard EIGENSTRAT [6], based on Principal Components Analysis

[1] "Null SNPs" refer to markers known to be unassociated to the disease of interest.

(PCA), advocates independent SNPs to avoid linkage disequilibrium (LD) that may distort the axes of variation. Patterson et al. [25] outline a correction method for EIGENSTRAT to minimize distortion in PCA due to high correlation in adjacent SNPs based on multivariate regression analysis for each marker in the SNP matrix; although recommend filtering out loci with high LD in large data. A recent method using Classical Multidimensional Scaling [7] corroborates the use of independent SNPs; yet in GWAS, obtaining independent markers or running multiple regression analyses to correct for high adjacent LD is computationally difficult. Other methods use randomly selected markers to infer structure [23].

Independent SNP sets (or an LD corrected SNP matrix) avoid distortion but do not select differentiated SNPs. A common differentiation measure is F_{ST} [21]. Bamshad et al. [20] concludes that the power of a marker (to exhibit PS) is a function of F_{ST}. Patterson et al. [25] explain the "phase change phenomenon", where the ability to detect PS is dependent on the number of individuals N, markers L, and F_{ST}. Namely for two populations, there is a threshold: $F_{ST} = \frac{1}{\sqrt{NL}}$, above which ($F_{ST} > 1/\sqrt{NL}$) evidence of structure in the data is detectable; otherwise differentiation is unsuccessful. Hence PS studies must consist of well chosen markers to maximize F_{ST} in the data and the selection of optimally informative markers must be addressed to quantify genetic heterogeneity among individuals.

We propose SNP selection based on deviation from Hardy-Weinberg Equilibrium (HWE) as a solution to confounding LD as well as choosing an informative SNP set. Markers in HWD yield a set of differentiated SNPs as evidenced by F_{ST} calculation. Additionally by partitioning the sources of LD, we show not all markers in LD confound PS inference and that by selecting markers based on the test for HWE, we only exclude confounding LD SNPs (thereby we are able to include markers in LD that actually signify stratification). The motivation for using HWD as a selection tool lies in ease-of-use through the incorporation of a standard test commonly performed in a study.

We also apply Non-metric Multidimensional Scaling (NMDS) [40] as a visualization tool for evaluating population stratification. The use of NMDS builds upon work by Li and Yu [7], which used Classical MDS (performing eigen-decomposition on a similarity matrix) to adjust the association test by the MDS eigenvectors to encapsulate the effect of PS. Unlike Classical MDS and PCA, NMDS does not return eigenvectors to form dimensions; instead the method finds an optimal configuration of points in a reduced dimension by iteratively minimizing a stress function. Minimization of the function results in the optimal coordination in the reduced dimension relative to the true distances between objects in a given distance matrix. Simulation as well as examples with HapMap samples [41] show the proposed SNP selection and NMDS visualization work well to differentiate populations in a stratified sample.

2.2 SNP Selection Materials and Methods

SNPs with high F_{ST} values equate to markers that exhibit large allele frequency differences as F_{ST} estimates are proportional to the sample variance for allele frequencies among populations [32]. Values of F_{ST} vary depending on the amount of existing population differentiation. Yet the use of F_{ST} for SNP selection to define structure between groups in a population is limited because calculation requires knowledge of group membership. Alternatively, population structure and variable allele frequencies have a large impact on deviation from Hardy-Weinberg proportions; which lends itself as a tool for selection of informative SNPs by testing for HWD in the combined case/control sample.

2.2.1 Hardy-Weinberg Disequilibrium SNP Selection

Differing allele frequencies between groups result in a loss of heterozygosity (in the population as a whole) termed the Wahlund effect [42]. An extreme exhibition of the effect is with two groups where within one group, all individuals have the genotype 'AA' at a locus; while in the other all samples carry 'aa'. If the data are analyzed without regard to group membership, there will be no heterozygosity in the population (a gross violation of HWE). Now suppose K subpopulations have allele frequencies p_k (probability for allele 'A') for $k = 1, 2, ..., K$ at a locus. Assume within each subpopulation, the locus exhibits HWE with genotype frequencies of p_k^2, $2p_kq_k$, and q_k^2, where $q_k = 1 - p_k$. For equal sample sizes n, if these subpopulations were treated as one population; the expectation μ_p and variance σ_p^2 of the population allele frequency p are given as

$$\mu_p = \overline{p} = \frac{\sum_{k=1}^{K} p_k}{K} \quad \text{and} \quad \sigma_p^2 = \frac{\sum_{k=1}^{K}(p_k - \overline{p})^2}{K} = \frac{\sum_{k=1}^{K} p_k^2}{K} - \overline{p}^2.$$

The expected genotype frequencies for homozygote 'AA' can then be calculated by (2.1):

$$E(AA) = P_{AA} = \frac{\sum_{k=1}^{K} p_k^2}{K} \quad = \quad \overline{p}^2 + \sigma_p^2. \tag{2.1}$$

Similarly, for homozygote 'aa', $E(aa) = (1 - \overline{p})^2 + \sigma_p^2$. The expected frequency for a heterozygous individual is then calculated as:

$$\begin{aligned} E(Aa) = P_{Aa} &= \frac{\sum_{k=1}^{K} 2p_k(1-p_k)}{K} = 2\left(\frac{\sum_{k=1}^{K} p_k}{K} - \frac{\sum_{k=1}^{K} p_k^2}{K}\right) \\ &= 2(\overline{p} - (\overline{p}^2 + \sigma_p^2)) \\ &= 2\overline{p}(1 - \overline{p}) - 2\sigma_p^2. \end{aligned} \tag{2.2}$$

These estimators show that difference in allele frequencies between subgroups directly impact Hardy-Weinberg proportions. Heterozygosity at a marker is reduced by $2\sigma_p^2$ (while the homozygous frequencies are inflated); thus loci with group variation in allele frequencies exhibit deviation from HWE.

2.2.2 Controlling the Confounding Effect of LD

Patterson et al. [25] discuss how LD can distort the PCA axes of variation used in EIGENSTRAT; yet Price et al. [6] recommend using all markers for studies of more than 100,000 SNPs[2] based upon HapMap analysis of the Asian populations that shows the effect of LD to be negligible for sets of over three million loci. Their analysis of the effect of LD in practice is limited (they examined results for removing only adjacent markers at a maximum of a ten locus range) as the behavior of LD is widely variable both among populations and within different gene regions. Smaller scale studies that include areas of the genome where LD extends much further than expected may still suffer from this distortion as PCA will indiscriminately elucidate factors that contribute to genetic variation (whether this be from LD/haplotype variation or PS). When markers from LD blocks are chosen, dissimilarity between individuals may be driven by genotype patterns due to LD as opposed to PS. The solution is an independent or LD corrected SNP set; yet LD may be created by PS and so this excludes potentially informative markers. Peterson et al. [43] note that total LD

[2] Found in the online Supplemental Material of their work.

(LD_T) can be partitioned as

$$LD_T = LD_W + LD_{SC} + LD_{CT},$$

representing components within populations (LD_W), between local populations (LD_{SC}), and among continents (LD_{CT}); where LD_{SC} and LD_{CT} are due only to allele frequency variation among populations. Only LD_W (blocks of LD common in populations with low variation in allele frequencies) distorts the effect of PS. Markers in LD that exhibit high allelic differences due to a subdivided population will also show evidence of HWD (Wahlund Effect); by selecting markers based on the test of Hardy-Weinberg Equilibrium, we not only can select independent substructure-informative markers but also markers in LD due to population differentiation. Markers in confounding LD regions will not display allele frequency differences (between groups) and will tend to be excluded from selection (assuming other sources of HWD are minimal). The simulation study and examples from HapMap data corroborate this behavior.

2.2.3 SNP Selection Using the Test for HWE

The goodness-of-fit test for HWE (HWT) is commonly performed in association studies. The test may be done using control individuals as a quality control measure, using cases (for evidence of disease-susceptible loci [44]), and/or using the combined set of cases and controls. The HWT as a formal test for population admixture has had limited success, though [45] showed that power of the HWT for PS detection is directly related to the degree of population differentiation. Recently Lee [46] proposed the use of the HWT for a combined panel of SNPs to better formally detect PS[3]. Here we do not require the HWT to formally indicate PS, but use it simply as a selection tool of informative SNPs for further use with methods suited for detection of PS. The HWD coefficient for allele 'A' used in the HWT,

$$\hat{D}_A = \tilde{P}_{AA} - \tilde{p}_A^2$$

($\tilde{P}_{AA}$ and $\tilde{p}_A$ are the respective estimated genotype frequency of 'AA' and allele frequency 'A'), will be large due to the Wahlund effect at informative SNPs; making the p-value small

[3] This promising research for the use of HWT supports our application of HWD for SNP selection.

for testing $H_0 : \ D_A = 0$ [47]. We use the p-value from the HWT for combined cases and controls[4] to provide mounting evidence of informativeness of a marker for PS detection. We refer to this use of the HWT as “HWD SNP selection” and explore cutoff values for the p-value as well as performance through simulation.

2.2.4 Distance Metrics

Li and Yu [7] give a good outline of metrics for Classical MDS which can also be employed with NMDS. We focus on general genetic distance metrics and show their relationship in response to criticism that distance-based methods are sensitive to the choice of distance metric. General metrics such as Gower's [37] and Rogers' [4] distances (whose properties apply to individual level calculation) can be related linearly to genetic distances such as the coancestry coefficient [48]. Allele Sharing Distance (ASD) [23] and Identical By State (IBS) [36] are also applied on the individual level for genetic distance. We show that these measures are synonymous under a range standardization and that distance-based methods in this context are not hindered by metric choice as critics claim.

Rogers' distance compares two populations based on allele frequencies. Let L be the number of loci/markers, a_l is the number of alleles at the l^{th} locus, p_{lj} and q_{lj} are the allele frequencies for the j^{th} allele, l^{th} marker for population 1 and 2 respectively. Roger's distance is given by:

$$D_R = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{2}\sum_{j=1}^{a_l}(p_{lj}-q_{lj})^2}\ . \tag{2.3}$$

Identity By State (IBS) is a commonly used measure to compute genetic similarity (S_{IBS}) between two individuals in a population by evaluating the number of allele copies two individuals have in common:

$$S_{IBS} = \frac{1}{2L}\sum_{l=1}^{L} s_l(x_l, y_l), \tag{2.4}$$

where $l = 1, 2, ..L$ is the number of loci over which to compute the sharing measure; x_l and y_l is the genotype for individuals 1 and 2 at the l^{th} marker; and s_l is a sharing function that takes values of 0 if no alleles are shared, 1 if individuals have one allele in common, and 2 if the individuals share both alleles [36]. This formulation can be recoded with $x_l = (0, 1, 2)$

[4] Preferably with a set of “Null SNPs” if possible

in place of $x_l = (aa, Aa, AA)$ and likewise for y_l. The sharing function is then expressed as $s_l(x_l, y_l) = 2 - |x_l - y_l|$. We wish to use the measure of IBS as a genetic *distance*; achieved by subtracting S_{IBS} from one to yield:

$$D_{IBS} = 1 - \frac{1}{2L}\sum_{l=1}^{L}(2 - |x_l - y_l|) = \frac{1}{2L}\sum_{l=1}^{L}|x_l - y_l|. \tag{2.5}$$

By treating individuals as populations, Rogers' distance yields an equivalent equation such that $D_R = D_{IBS}$ (See Equation A.2 in Appendix A).

Another general measure is Gower's distance [37], which is similar to genotype matching. Using the notation of x_l and y_l, the genetic data are treated as interval variables and Gower's distance is expressed by:

$$D_G = 1 - \left(\frac{\sum_{l=1}^{L}(1 - |x_l - y_l|)}{L}\right) = \frac{1}{L}\sum_{l=1}^{L}|x_l - y_l|. \tag{2.6}$$

Gower's distance is equivalent to Allele Sharing Distance (ASD) given in (2.7) where d_l takes values of $(0, 1, 2)$ if two, one or no alleles are shared at a marker l. With the x_l and y_l notation, ASD becomes equal to Gower's distance:

$$D_{ASD} = \frac{1}{L}\sum_{l=1}^{L}d_l = \frac{1}{L}\sum_{l=1}^{L}|x_l - y_l| = D_G. \tag{2.7}$$

If we use a range standardization to Gower's and ASD[5], the range correction will be two. Thus the relationship of the distances is given by (2.8).

$$\frac{D_{ASD}}{2} = \frac{D_G}{2} = \frac{1}{2L}\sum_{l=1}^{L}|x_l - y_l| = D_{IBS} = D_R \tag{2.8}$$

This relationship is often not recognized due to different formulations; our notation reconciles the imposed differences in the metrics and supports the use of distance-based methods for quantifying genetic relatedness.

[5] Assuming all markers have at least one individual homozygous for the variant allele and for the common allele

2.2.5 Non-metric MDS Algorithm

The method by Li and Yu [7] computes a similarity matrix and uses the top eigenvectors as covariates to control the effect of PS in association testing. The method is a generalization of EIGENSTRAT where the decomposed matrix may no longer be a correlation or covariance matrix but formed from a chosen similarity. This method, coined "MDS", is also known as Classical MDS or Principal Coordinates Analysis (PCoA). NMDS, in contrast to both EIGENSTRAT and Classical MDS, does not return eigenvectors to describe variability. Instead, NMDS fits an iterative measurement model of the form

$$T(D_{rs}) = \hat{D}_{rs} + \epsilon$$

to estimate optimal dimension coordinates for objects based on the observed distance matrix for a specified number of dimensions (where $T(D_{r,s})$ is the optimally transformed dissimilarity for objects r and s and $\hat{D}_{rs}$ is the Euclidean distance between the estimated NMDS coordinates for objects r and s). The coordinates produced are the best fit possible to the reduced dimension space (as measured by Euclidean distance) compared to the full distance matrix by minimization of a Stress (here we use Stress formula 1 from Kruskal and Wish [49]) function given in (2.9). At each iteration of the NMDS algorithm, coordinates are estimated and the Stress or "Badness-Of-Fit" measure is evaluated.

$$Stress1 = \sqrt{\sum_{r,s} \left(\frac{(T(D_{rs}) - \hat{D}_{rs})^2}{\sum_{r,s} T(D_{rs})^2} \right)} . \tag{2.9}$$

The algorithm converges and iteration ends when the change in BOF becomes sufficiently small. Appendix A.1 gives more details on the NMDS algorithm.

NMDS can often visualize structure with the first two or three dimensions that may not be captured with the first few eigenvectors from PCA or PCoA, thereby giving a more "truthful" representation of the data in reduced dimension space as opposed to simply pulling out the most variable factors. These qualities make NMDS an appealing method for use in population structure visualization; particularly with the proposed HWD SNP selection due to quality control concerns and other sources of deviation in HWD.

2.3 Results

Simulation was used to study the empirical effects of SNP selection and NMDS visualization. Subgroups were simulated with relatedness among individuals (F_k) and varying allele frequencies at both null markers and a disease locus. While our methods do not attempt to adjust the test for association (current unpublished research), case/control status was generated using the disease locus and used in sampling to create imbalanced proportions of subgroups in cases and controls to assess how our SNP selection methods would perform in a disease association context. Additional (unpublished) simulations were carried out without generating a disease and produced synonymous results. F_k describes the correlation among individuals where high values correspond to low diversity. Creating relatedness among groups of individuals stratifies those groups from other individuals so that higher relatedness within groups yields more distinction between groups [6].

2.3.1 Simulation Details

Data were generated using Balding and Nichols [50] model similar to past studies [6, 7]. Populations consisted of individuals from three subgroups ($k = 1, 2, 3$) with (F_k) equal to 0.025, 0.015, 0.005 respectively. Disease status was assigned using a causal locus with additive penetrance rates and allele frequencies (q_k) with probabilities 30%, 20%, and 25% for the respective subgroups. Penetrance of disease (T_k) within a subgroup took values of 0.2, 0.15, and 0.1 respectively. Following the model used by Tzeng et al. [51], the probability of disease for an individual with no copies of the causal allele is $f_0 = T_k/(1 - 2q_k - 2q_k r)$ with r, the additive risk, chosen as $r = 1.5$. A heterozygous individual has risk of $f_1 = rf_0$, while an individual homozygous for the causal allele has risk $f_2 = 2rf_0 - f_0$.

Ten thousand loci were generated where 5,000 exhibit LD within subgroups to assess the robustness of HWD SNP selection (to the LD effect). Allele frequencies were generated as

$$Beta\left(\frac{(1-F_k)p_{lk}}{F_k}, \frac{(1-F_k)(1-p_{lk})}{F_k}\right),$$

where p_{lk} was generated as $Uniform(0.1, 0.55)$ for the 5,000 substructure-informative markers. To generate LD SNPs, p_{lk} was chosen from a $Uniform(0.35, 0.45)$ and $F_k = 0.001$ was

[6] If F_k is chosen to be equal in all subgroups, this becomes an estimate of F_{ST}.

set for all subgroups[7]. Two separate independent blocks of LD were created using the disequilibrium coefficient $D = \pi_{AB} - \pi_A \pi_B$ [47] , where given two markers with alleles 'A' and 'a' at locus 1 and 'B', 'b' at locus 2, π_{AB} represents the frequency of observing alleles 'A' and 'B' together with individual allele frequencies represented by π_A and π_B for the respective alleles. For each block of LD, a single SNP was generated and used to conditionally generate the alleles for LD markers. For example, if the allele of the SNP for LD Block 1 is 'A' then allele 'B' for an LD Marker is drawn from a

$$Binomial\left(1, p_{lk} + \frac{D}{p_{lk}}\right),$$

otherwise it is drawn from

$$Binomial\left(1, p_{lk} - \frac{D}{1 - p_{lk}}\right).$$

Using our values for p_{lk} and $D = 0.08$, the blocks generated within the subgroups had an average median p-value=0.037 for testing pairwise Linkage Disequilibrium[8] effect .

The samples exhibit PS due to variation in allele frequencies, correlation structure, and disease penetrance rates. While uncorrelated markers in the population were generated in HWE and linkage equilibrium, samples do not exhibit this due to sampling regardless of subgroup membership. LD generated by sampling in the 5,000 substructure-informative markers will not have a confounding effect as it is created due to PS. One hundred populations were generated to evaluate SNP selection and visualization where 500 cases and 500 controls were randomly selected in each population. For controls, an average percentage of 32.06%, 33.25%, and 34.69% belonged to subgroups 1, 2, and 3 respectively. The case group had averages of 42.17%, 34.74%, and 23.09% of individuals belonging to the respective subgroups.

2.3.2 Simulation Results

Figure 2.1 shows the p-values for the HWT at each marker over all 100 iterations for substructure-informative SNPs and confounding LD SNPs plotted against the expected

[7] This allows for more realistic variation in confounding LD SNPs through random draws from the Beta distribution

[8] A more realistic simulation of LD compared to that done by Patterson et al. [25] which generated perfect LD.

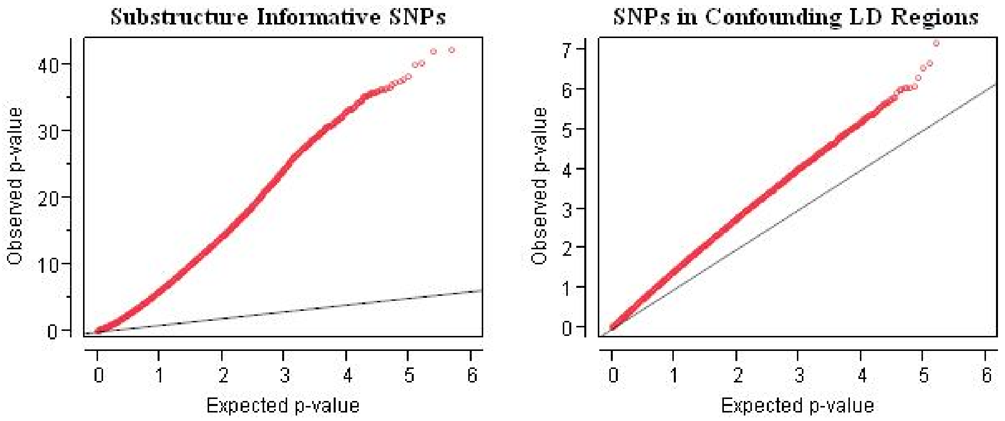

Figure 2.1: Quantile plots of observed -log10(p-values) from the test for Hardy-Weinberg Equilibrium against expected -log10(p-values) for SNPs generated in regions of confounding LD (right plot) and substructure-informative SNPs generated without LD (left plot). Under the assumption of no HWD, the p-values should follow a uniform distribution (black line).

distribution of p-values. There is marked deviation for the substructure-informative SNPs; supporting the use of HWD for informative SNP selection. Confounding LD markers show a small deviation, which is expected given that F_k and random assignment of allele frequencies influence the subgroups. The average median p-value for the substructure-informative SNP set was $0.062(\sigma = 0.003)$ compared to $0.373(\sigma = 0.036)$ for the confounding LD SNP set.

To explore how well HWD SNP selection eliminates confounding SNPs, sets of 1,000 SNPs were chosen from markers that met a selection cutoff based on significance levels of HWT $(0.001, 0.01, 0.05, 0.10, 0.20)$. The proportions of markers from confounding LD regions were calculated for the 100 populations. Table 2.1 shows that a cutoff of HWT $p < 0.01$ allows only 9.1% of markers to be confounding on average[9]. To further explore the performance of HWD SNP selection, 4 sets of 200 markers were randomly chosen (for each population) that met the following criteria: SNPs that were in HWD with $p < 0.01$, SNPs in HWE with $p > 0.01$, Randomly chosen SNPs (no criteria imposed), and Non-Confounding Random SNPs (chosen from regions where no LD was generated in original populations). Table 2.2 displays the average proportion of chosen confounding LD SNPs as well as the average F_{ST} value for each SNP set. Selection of markers in HWD yields a significantly more differentiated set than the contrasting selection methods including non-confounding

[9] Sample out of a panel of SNPs ten times larger than the set, half of which were confounding due to LD.

Table 2.1: Average proportion ($\mu_{P_{LD}}$)and standard deviation ($\sigma_{P_{LD}}$) of LD confounding SNPs chosen under HWD selection criteria

Selection Criteria	$\mu_{P_{LD}}$	$\sigma_{P_{LD}}$
HWT $p < .001$	0.023	0.008
HWT $p < .01$	0.091	0.019
HWT $p < .05$	0.199	0.029
HWT $p < .10$	0.264	0.028
HWT $p < .20$	0.336	0.025

Table 2.2: Average proportion ($\mu_{P_{LD}}$) of LD confounding SNPs chosen and average F_{ST} ($\mu_{F_{ST}}$) under various SNP selection criteria with associated standard deviations $\sigma_{P_{LD}}$ and $\sigma_{F_{ST}}$.

Selection Criteria	$\mu_{P_{LD}}$	$\sigma_{P_{LD}}$	$\mu_{F_{ST}}$	$\sigma_{F_{ST}}$
SNPs in HWD ($p < .01$)	0.089	0.025	0.083	0.006
SNPs in HWE ($p \geq .01$)	0.605	0.037	0.033	0.002
Random SNPs	0.505	0.037	0.043	0.003
Random Non-Confounding SNPs	0	0	0.051	0.005

random selection ($F_{ST} = 0.083$ vs. $F_{ST} = 0.051$). The effect of confounding LD present in the SNP selection is evident from the significantly smaller F_{ST} for random markers (0.043) as well.

To evaluate SNP selection methods in conjunction with NMDS, one population (out of the 100 iterations) was randomly chosen for visualization. Using JMP® Genomics Software, distance matrices using Gower's distance with a range standardization (equivalent to IBS and others as previously shown) for 1,000 individuals were calculated with each of the aforementioned SNP sets. The distance matrices were then used as input for NMDS. For each distance matrix, dimensions 1 through 5 were fit to the data. BOF and distance correlation plots were evaluated (Figure A.1 in Appendix A) and a two-dimensional fit was chosen[10]. Figure 2.2 shows 95% density ellipses (for the three subgroups from the structured population) for individual coordinates formed from 2-D NMDS using the four different SNP set distance matrices. Clearly, HWD SNPs outperform the other SNP sets for quantifying structure. The effect of confounding LD can be seen by contrasting randomly chosen SNPs

[10] Data with clear structure will yield a BOF plot similar to the Scree plot, with the "elbow" of the plot at the optimal dimension fit.

and non-confounding randomly chosen SNPs.

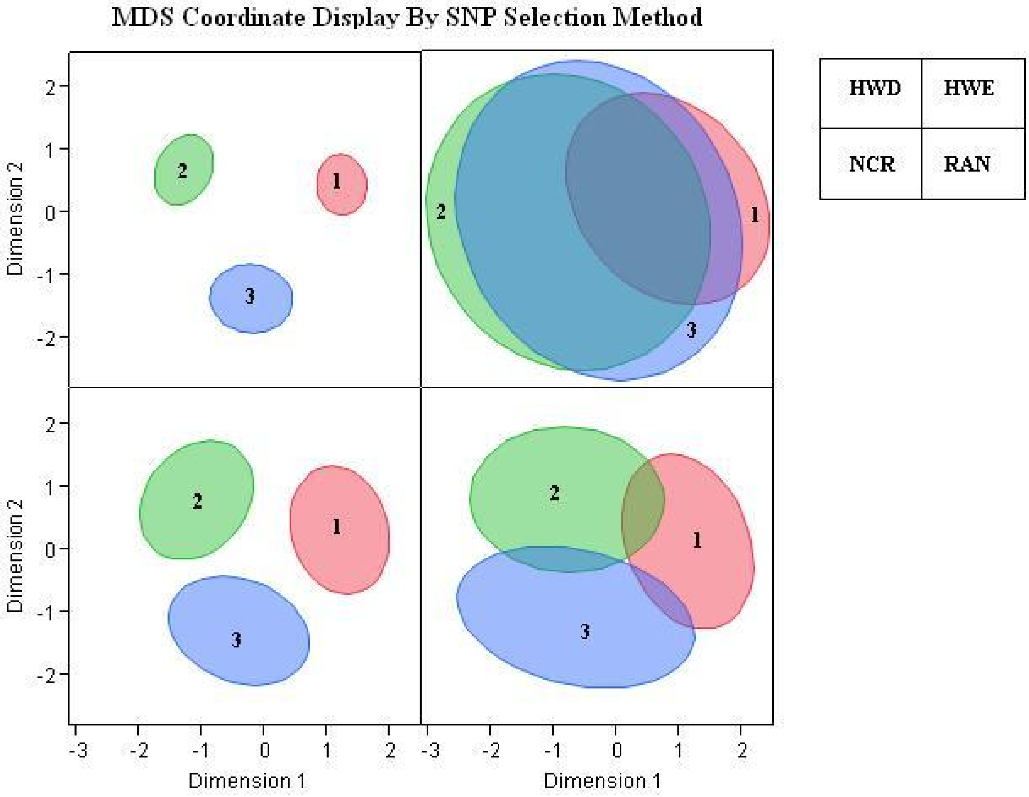

Figure 2.2: 95% density ellipses for NMDS coordinates of one simulated population using 200 markers to infer PS by the four SNP selection criteria. Distance matrices were calculated using Gower's Distance and NMDS coordinates were computed for each of the SNP selection methods (HWD SNPs, HWE SNPs, random SNPs (RAN), and random, non-confounding SNPS (NCR)). Simulated subpopulations 1, 2, and 3 are marked in the corresponding density ellipse.

An important result evident from the NMDS plot using HWD SNP selection is the visualization of variation not only between subgroups, but within as well. Allelic variation comes not only from differences between individuals from separate subgroups; it is also present among individuals within a subgroup. Recall the simulation of these three subgroups: the first subgroup (1) was generated to have most relatedness among individuals ($F_k = 0.025$), while subgroup 2 and 3 have progressively less relatedness among individuals ($F_k = 0.015, 0.005$ respectively). HWD SNP selection is able to quantify this visually while picking non-confounding random markers distinguishes the groups (not to the distinction of HWD SNP selection), but does not reflect the amount of relatedness/variation present within groups.

2.3.3 Empirical Relationship of F_{ST} and HWD SNP Selection

An additional simulation was carried out where F_k was set equal in all subgroups in order to explore the empirical relationship between F_{ST} and the number of HWD SNPs. Sets of 2,000 SNPs were generated for values of F_{ST} ranging from 0.005 to 0.20 and the proportion of HWD SNPs with $p < 0.01$ was plotted in Figure 2.3. For smaller values of F_{ST} (indicating very little differentiation), the HWT for deviation due to PS is not powerful (corroborating [45]); although the power increases dramatically as the sample size and differentiation increases. Additionally, for common GWAS, selecting 2% (the average proportion of HWD SNPs for 1,000 individuals for F_{ST} values between 0.005 and 0.02 in this simulation[11]) from 500,000 SNPs yields 10,000 substructure-informative SNPs. Further simulation of smaller F_{ST} values from 0.0005 to 0.005 showed the proportion of HWD SNPs to level off (not pictured) with an average proportion for 1,000 individuals in this range to be 1.59%.

2.3.4 Application to HapMap Samples

To assess HWD SNP selection and NMDS visualization on real data, unrelated individuals from the HapMap populations[12] were combined. Proposed methods were applied to SNPs from Chromosome 21. The HWD SNP set for visualization met the following criteria; Test for HWE: $p < 0.001$, Minor Allele Frequency: $maf > 0.10$, Proportion of missing data: $propmiss < 0.05$. These criteria yielded a set of 2918 SNPs; an additional random set of 2918 SNPs with Minor Allele Frequency: $maf > 0.10$, and Proportion of missing data: $propmiss < 0.05$ was used as well. Both PCA (the coordinates used in EIGENSTRAT correction) and NMDS were implemented to visualize the resulting pattern of variation (Figure 2.4). Both methods (for Random and HWD SNP selection) distinguish CEU and YRI from the Asian populations, and HWD Selection gives more defined clusters than random SNP selection for both methods. Weir et al. [32] notes that on Chromosome 21, the YRI population shows slightly more heterogeneity within the population (evidenced

[11] Increasing the HWT p-value cutoff to 0.05 increased this average to 7.4%.

[12] HapMap populations are made up of 208 Yoruban (YRI), Japanese (JPT), Han Chinese (CHB), and CEPH (CEU) individuals. Data downloaded was from the HapMap NCBI Build 35.

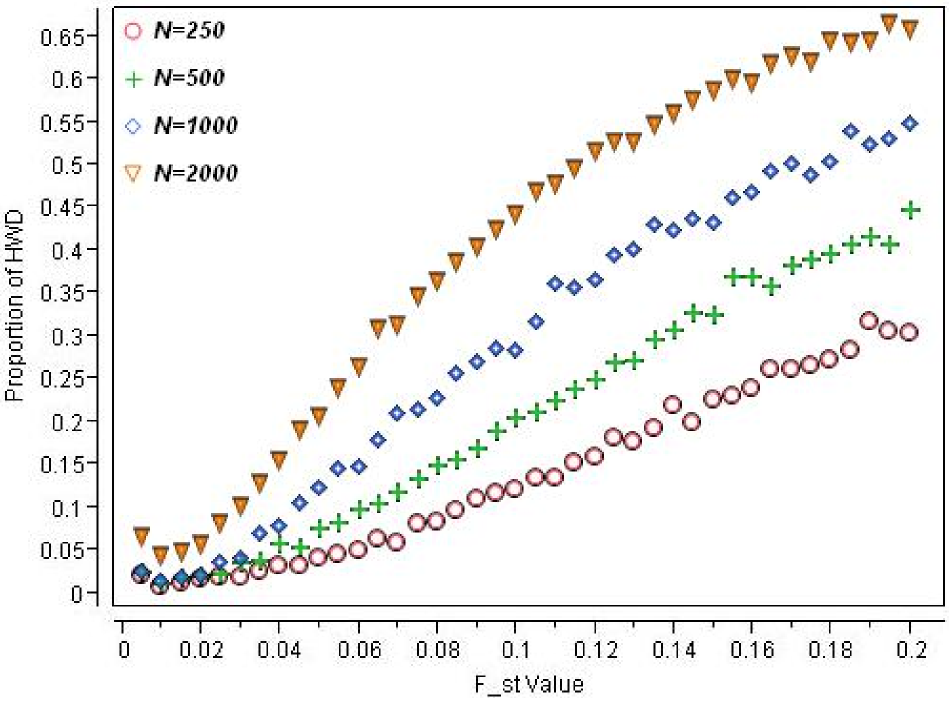

Figure 2.3: Plot of the empirical relationship between F_{ST} values and proportion of SNPs in HWD ($p < 0.01$). Two thousand SNPs were generated under the simulation scheme for values of F_{ST} ranging 0.005 to 0.200 and evaluated for sample sizes of 250, 500, 1000, and 2000 individuals.

by within population F_{ST} estimates) compared to the other populations. This coincides with both plots of individual coordinates from NMDS (right) in Figure 2.4, which shows more variation existing within the YRI population (in contrast to PCA). This supports the rationale behind the NMDS algorithm, which yields a more harmonious picture compared to the full genetic distance matrix (while PCA only displays the variation that is captured in the first two eigenvectors). It is of interest to see the effect of choosing HWD SNPs from a homogeneous population to evaluate if such selection imposes heterogeneity. Unrelated CEU individuals were utilized and SNPs from Chromosome 21 were selected by the criteria: Test for HWE: $p < 0.01$, Minor Allele Frequency: $maf > 0.01$, proportion of missing data: $propmiss < 0.10$. The PCA (EIGENSTRAT coordinates) and NMDS plots in Figure 2.5 show that HWD SNP selection with PCA visualization distorts the CEPH population, while NMDS is robust to the low quality of the data and accurately depicts the CEPH population.

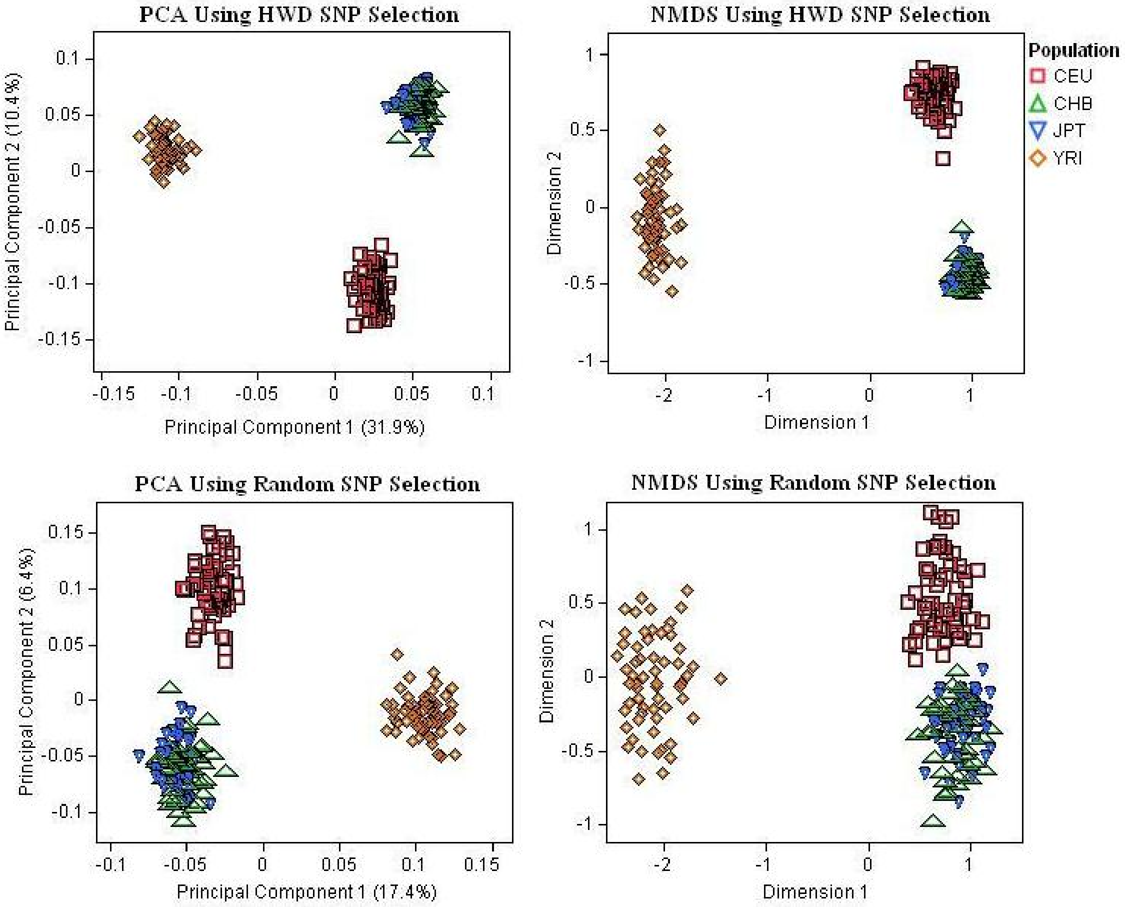

Figure 2.4: PCA (left) and NMDS (right) display of the HapMap populations for Chromosome 21. HWD SNP set used for visualization met the following criteria: Test for HWE: $p < 0.001$, Minor Allele Frequency: $maf > 0.10$, Proportion of missing data: $propmiss < 0.05$. Random SNP set met the criteria for minor allele frequency and proportion of missing data only.

2.4 Discussion

HWD SNP selection yields an informative SNP set that reveals more differentiation than random or non-confounding random SNPs. Simulations show that HWD SNP selection is robust against SNPs that confound the effect of population structure. This is preferable to independent SNP selection which suffers from computational constraints and restricts selection of SNPs in LD due to PS. HWD SNP selection promises to be a more powerful set for population differentiation as evidenced by F_{ST} calculations. Visualization

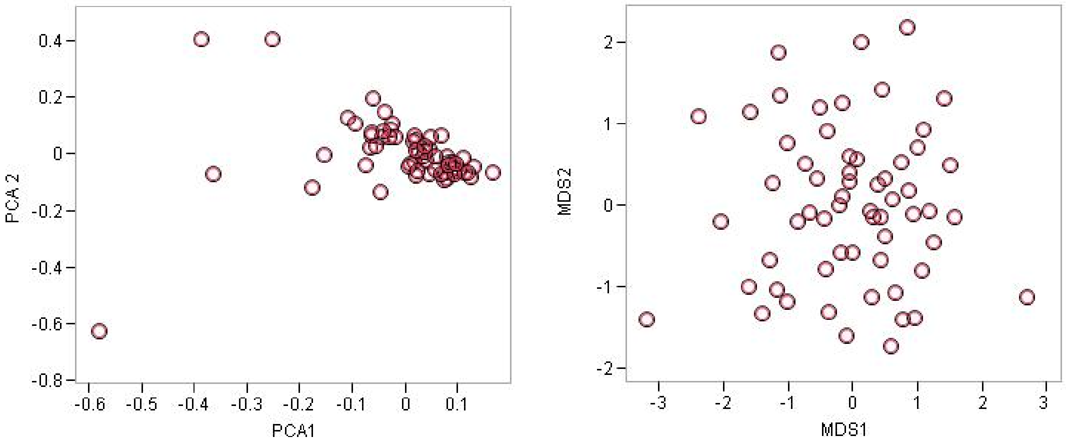

Figure 2.5: PCA (left) and NMDS (right) display of the HapMap CEPH population for Chromosome 21. SNP set used for visualization met the following criteria. Test for HWE: $p < 0.01$, Minor Allele Frequency: $maf > 0.01$, Proportion of missing data: $propmiss < 0.10$.

using NMDS works well with HWD SNP selection by yielding a more faithful representation of observed distances among individuals than PCA, which is a more "extreme" projection. NMDS minimizes differences between Euclidean distance in the reduced dimension space and the observed dissimilarities from the full distance matrix, whereas PCA coordinates maximize explained variance in the reduced dimension (and thus may be more prone to outlier influence and finding projections that inordinately pull apart the data). This distinction helps to assuage worries about Type I error in the context of finding differentiation in a homogeneous population, which in turn would relate to an over-conservative correction for PS in association testing. Additionally, this property of NMDS makes the method more robust against other sources of HWD unrelated to population stratification, although data quality is a large concern when using the HWT in this manner.

The influence of markers in HWD not attributable to PS will be negligible if standards of data quality are upheld. In practice it is still recommended to first use the HWT for quality control purposes as well as elimination of SNPs with inadequate call rates, missing data, and rare alleles (excluded for substructure quantification). Such steps help ensure HWD is reflective of allele frequency patterns among possible groups in the data. Once data quality has been addressed, the HWT p-values from the combined case/control set can be used for SNP selection to quantify PS. A cutoff based on the HWT p-value < 0.01

works well with our simulated data and in HapMap samples; although depending on the size and quality of the data, a more or less stringent cutoff may be used. If data quality remains an issue, a less stringent cutoff and random sampling from the markers that meet this cutoff may be used to obtain a manageable SNP set. An alternative cutoff may be decided based on effect size or ranks of the test statistic to help avoid bias/influence due to minor allele frequencies and p-value threshold concerns [52].

The data quality issues plague all methods for quantifying population structure (in both SNP selection and association test correction). The proposed SNP selection method helps to answer confounding problems encountered such as blocks of LD, as well as obtaining a differentiated set of SNPs. The incorporation of nonmetric MDS complements the method as a visualization technique. Additionally, NMDS shows promise for future use by building upon research done for current association correction tests such as EIGENSTRAT and the recently proposed (classical) MDS method.

3

Population Stratification Correction in Association Studies Via Nonmetric Multidimensional Scaling

3.1 Introduction

Techniques for correcting bias in genetic association studies due to population stratification (PS) have been an active area of research in recent years. PS is the increase in spurious associations in case-control genetic studies due to ancestral allelic variation. When individuals from a common ethnic background are over-represented for the trait of interest, Type I errors can occur at loci that exhibit allele frequency differences for that ethnic group (in comparison to others). In order to achieve sufficient power for testing associations of Single Nucleotide Polymorphisms (SNPs) in large scale studies, increasingly large sample sizes are required; often resulting in a heterogeneous population sample. Genomic Control (GC) [29] and *Structure* [28] are two of the earliest PS correction methods, since which numerous studies have highlighted the complexity and impact of PS [53, 54, 55, 32, 56]. GC, which corrects the association test statistic by a fixed inflation factor across the genome, has been shown to be too simplistic in many cases; whereas *Structure* is computationally difficult with large numbers of loci.

EIGENSTRAT is a recently popular approach that employs principal components analysis (PCA) to capture and correct for genetic variation in samples [6], although it has been shown to be insufficient in certain cases [57, 7]. A method that builds on EIGENSTRAT was introduced by Li and Yu [7] and uses classical multidimensional scaling (CMDS) coupled with clustering to improve PS correction due to admixed and/or discrete population structure. While PCA performs eigen-decomposition on a correlation or covariance matrix, CMDS is a generalization that allows eigen-decomposition of a chosen similarity/dissimilarity matrix[1] such that more appropriate genetic distance measures can be used for dimension reduction to capture population structure. The correction model is the defining distinction between EIGENSTRAT and CMDS. EIGENSTRAT uses principal components as axes of variation to iteratively weight both phenotype and genotype with adjustments obtained from simple linear regression on the principal components, while CMDS uses the coordinates as continuous covariates as well as an additional categorical covariate (of individual classifications obtained from clustering) in a logistic model.

A drawback of EIGENSTRAT arises when select loci exhibit more allelic variation between population groups than captured by the axes of variation [7]; evidence that EIGENSTRAT may be an insufficient correction in cases of data with discrete population groups. Additionally, a recent study found data with discrete population groups to have more severe spurious associations in contrast to data exhibiting admixture [38]. For this reason, CMDS performs k-medoid clustering based on the principal coordinates and uses the cluster assignments as an additional covariate in the correction model. Li and Yu [7] show that their model corrects Type I error to the nominal level for highly differentiated SNPs while EIGENSTRAT fails to do so under the discrete populations simulation framework. The use of classical multidimensional scaling with hierarchical clustering has been independently proposed to work very well for population structure inference by Gao and Starmer [23].

Previous research has shown that nonmetric multidimensional scaling (NMDS) yields a more faithful representation of genetic variation patterns across populations in comparison to PCA (and consequently CMDS) [5], which chases directions of maximal variability. We propose a modification of the CMDS model that instead uses coordinates

[1] Li and Yu [7] show that with an inner product similarity, CMDS and PCA are synonymous.

obtained by the NMDS dimension reduction method as well as nonparametric clustering techniques. Simulation results show that our method (referenced as 'NMDS') controls Type I error as well as CMDS. Moreover NMDS is uniformly more powerful, with highly significant differences in power for scenarios where a portion of the loci used to infer coordinates for population variation are associated with the trait of interest. It has been a concern with PS correction that the set of markers used to infer population structure may also capture variability in the trait of interest, resulting in a loss of power and an increase in Type II errors. Through empirical analysis, NMDS is shown to be more robust for capturing population structure in comparison to EIGENSTRAT and CMDS when a number of disease-associated markers exist in the dataset.

The concern about over-correction and loss of power due to the influence of many positively associated markers has largely been ignored under belief in the common variant-common disease (CVCD) hypothesis. The debate that a handful of commonly found SNP variants explain heritability of complex common disease (CVCD) as opposed to many slightly deleterious rare variants (rare variant-common disease, RVCD)[58] has been shifting as more and more studies highlight the importance of rare (MAF $< 1-5\%$) SNPs in disease etiology [59, 60, for example]. Associations of common SNPs have thus far only explained a very small proportion of variance in traits and new sequencing technology can now accurately capture low frequency SNPs that show promise for explaining more heritability [60]. Moreover, initial successes with common variant associations are shadowed by modest effect sizes and low reproducibility. It is typical for studies to result in many highly significant 'hits' whereas the 'true associations' are often much further down the list or ranked p-values [52, 61]; thus a good recommendation for reproducibility studies is to follow up with thousands of possibly associated markers [61]. Many loci may also exhibit moderate evidence of trait association due to differential call rates for cases and controls due to genotyping batch effects or laboratory effects when cases and controls are collected separately; Price et al. [6] found evidence of this in a real data application where a top principal component was highly correlated with plate assay effects. In light of these findings, it is critical to assume there may be a large number of common and/or rare variants showing possible evidence of association with complex diseases and methods employed to correct for PS must result in a minimal loss of power for this scenario as well as appropriately control Type I error.

3.2 Methods

The NMDS model proposed (adopted from the model framework in Li and Yu [7]) for PS correction uses the coordinates computed for each individual in a study by nonmetric MDS to account for continuous genetic variation as well as cluster assignments using nonparametric density estimation to account for discrete subgroup membership. The following sections outline the dimension reduction method, clustering details and subsequent logistic modeling for testing association in the presence of confounding population structure.

3.2.1 Nonmetric Multidimensional Scaling

NMDS is an optimization routine that produces coordinates for N objects in a reduced dimension space that most closely represents (in Euclidean space) an $N \times N$ similarity or dissimilarity matrix. In this study we use Gower's distance [37] as a genotype matching dissimilarity metric computed for each pair of individuals across their SNP data. Both Miclaus et al. [5] and Li and Yu [7] provide details of other metrics for quantifying genetic relatedness that can be used with the NMDS algorithm. Gower's distance (using a range standardization) for a pair of individuals is calculated as

$$D_G = 1 - \left(\frac{\sum_{l=1}^{L} (1 - |x_l - y_l|)}{2L} \right) = \frac{1}{2L} \sum_{l=1}^{L} |x_l - y_l|, \tag{3.1}$$

where $l = 1, 2, ..L$ is the number of loci over which to compute the measure and x_l and y_l is the genotype for two respective individuals at the l^{th} marker. Optimized coordinates representing population structure are obtained through minimization of a Stress (here we use Stress formula 1 from Kruskal and Wish [49]) function given in (3.2). The iterative NMDS process first finds an optimal transformation function $T(\cdot)$ through isotonic regression (on the estimated Euclidean coordinates which are initialized on the first iteration) and then minimizes the Stress function with respect to $\hat{D}_{rs}$ using the Gauss-Newton algorithm. $T(D_{r,s})$ is the transformed dissimilarity for objects r and s and $\hat{D}_{rs}$ is the Euclidean distance between the estimated NMDS coordinates for objects r and s in (3.2),

$$Stress1 = \sqrt{\sum_{r,s} \left(\frac{(T(D_{rs}) - \hat{D}_{rs})^2}{\sum_{r,s} T(D_{rs})^2} \right)}\ . \tag{3.2}$$

The algorithm converges and iteration ends when the change in the Stress function becomes sufficiently small. More details of NMDS specific to genetic population structure can be found in Miclaus et al. [5] as well as more general multidimensional scaling techniques (including Classical MDS) in Kruskal and Wish [49] and Cox and Cox [40]. The number of adequate dimensions can be chosen by analyzing the change in the badness-of-fit (Stress1 values) as well as the correlation between the estimated pairwise Euclidean distances and true distances in the full matrix (similar to scree plots) [5]. If desired, the eigenvalues of the dissimilarity matrix can be produced upon initialization and a significance test using the Tracy-Widom distribution can be performed similar to EIGENSTRAT [25]. The most significant distinction in optimization between NMDS and eigen-based methods like CMDS and PCA is that NMDS seeks the coordinates that are the best fit possible in the specified number of dimensions to relationships found in the full dimensional data, whereas CMDS and PCA solely perform rotation and scaling of the matrix to extract axes in the direction of the maximum variance explained. Principal components will only explain interesting facets of the data when there is more signal in comparison to noise such that large variances correspond with the signal (i.e. population structure) and may be unable to elucidate more subtle patterns in noisy genetic data. Additionally, PCA-based methods are extremely sensitive to outliers and EIGENSTRAT requires an outlier removal process to account for this, which may exclude informative samples.

3.2.2 Nonparametric Clustering

Nonparametric density estimation is used for cluster assignment based on the NMDS coordinates due to the flexibility and generality of the method [62, 63]. Many clustering methods show bias toward certain specifications of clusters groups. Optimization methods based on least-squares (such as k-means) are biased to finding clusters containing the same number of observations, yet other methods may tend to find clusters of equal size. Cluster methods based on nonparametric density estimation are the least biased in general and more applicable for irregular data with varying size and shapes [62, 64] while still providing good results for data with clusters of equal size and variance. Nonparametric density estimation allows for the density function to be determined from the sample data and thus imposes no fixed assumptions on the number or values of parameters. Nonparametric estimation is done by estimating the underlying kernel density of the data given as

$$\hat{f} = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \tag{3.3}$$

where K is the kernel (weight function) and h is a smoothing parameter that elucidates key patterns in the data $x_1, ..., x_N \sim f$ (kernel density formula taken from Scott [63, pg. 125]).

We used the MODECLUS procedure in SAS/STAT software which implements nonparametric clustering to obtain density estimates at each x_i given as $\hat{f}_i = \frac{n_i}{N\nu_i}$ where n_i is the number of points within a calculated neighborhood of x_i and ν_i is the volume of the neighborhood. Neighborhood or 'sphere' size is determined using hyper-spherical uniform kernels with fixed or variable radii decided by the smoothing parameter specified [64]. Neighborhoods are formed by a valley-finding method which joins observations into clusters based on the nearest point with a higher estimated density. The number of adequate clusters is chosen based on the smoothing parameter choice. Silverman [62] provides recommendations for choice of smoothing parameter that are implemented in the MODECLUS procedure. We use the recommendation of a reference to a Gaussian distribution, weighted by the number of variables v used for clustering (number of NMDS dimensions) to prevent over smoothing, given as

$$\left[\frac{2^{v+2}(v+2)\mathbf{\Gamma}(0.5v+1)}{Nv^2}\right]^{1/(v+4)} \sqrt{\sum_{j=1}^{v} s_j^2}\ , \tag{3.4}$$

where s_j^2 is the variance estimate for each variable v_j (representing each NMDS dimension vector used). A simpler approach is to specify an array of smoothing values and one can visually inspect the fit through histogram plots. Figure 3.1 is an example of nonparametric (kernel) density estimation for a univariate case. Further detail for nonparametric density estimation and clustering can be found in Silverman [62], Scott [63] as well as the online SAS documentation for the MODECLUS procedure.

The choice of clustering method is a flexible aspect of the proposed NMDS method. The CMDS method uses a k-medoids algorithm (similar to k-means) and estimates the number of clusters via the Gap statistic [24], which evaluates the within-cluster dispersion for a range of estimates for cluster number. For our simulations of populations with differentia-

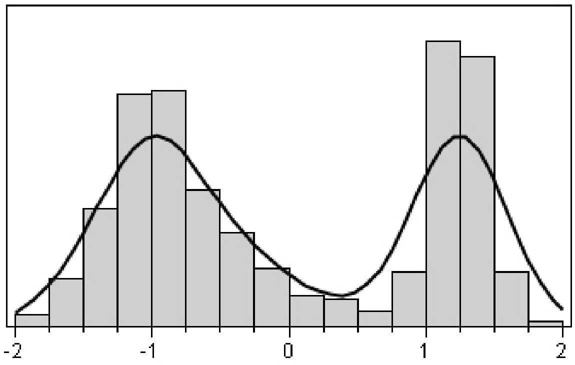

Figure 3.1: Histogram of a univariate distribution overlaid with a smoothed kernel density fit.

tion measured as $F_{ST} = 0.01$, both methods performed equally well. Simulations with less distinct populations ($F_{ST} < 0.005$) showed the nonparametric clustering with our smoothing estimate to more accurately select the correct number of populations in contrast to k-medoids. Further analysis (not covered in this work) may be necessary under varying population sizes and unequal population membership counts to better hone the choice of clustering method specific to genetic data.

3.2.3 Statistical Model for PS Adjusted Association Analysis

The proposed correction model uses the computed NMDS coordinates and cluster membership for all individuals as covariates in a logistic model. The logistic model can be expressed as

$$logit(p) = \boldsymbol{X\beta} + \boldsymbol{M\psi} + \boldsymbol{C\eta}, \tag{3.5}$$

where p is the probability of having the trait of interest. $\boldsymbol{X}$, $\boldsymbol{M}$, and $\boldsymbol{C}$ represent the $N \times 1$ SNP genotype vector, the $N \times K^{'}$ matrix of NMDS coordinates (where $K^{'}$ is the number of dimensions selected), and the $N \times 1$ vector containing cluster membership information respectively. $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\eta}$ are the respective vectors of parameters in the corrective association model. The inclusion of the cluster membership covariate has been shown by Li and Yu [7] as a key advantage over EIGENSTRAT for correcting for PS in a discrete population framework. The simplicity and ease of implementation of the correction model is an additional advantage; it can be easily expanded to include further phenotypic or environmental effects that may contribute to the trait as well as interactions. In Section 3.3.4, we extend the model to include an interaction effect between genotype and cluster membership. This interaction tests for differences in disease susceptibility at a SNP between different ethnic populations and can provide insight of how genetic disease risk may vary among different populations of individuals.

3.3 Results

To evaluate the performance of the NMDS method, simulations were done to obtain Type I error and power estimates, with emphasis on the scenario when many SNPs may show evidence of association with disease. NMDS is compared with CMDS and EIGENSTRAT. For all simulation scenarios, two dimensional coordinates were calculated for all three methods based on scree plots and visualized best fit of the data. We do not address the use of formal testing of eigenvalues for determining dimension, nor do we use the recommendation by EIGENSTRAT to incorporate the first 10 principal components, as this would bias the EIGENSTRAT method to return poor results in our scenario (extreme over-correction due to the inclusion of SNPs associated with case/control status).

The Balding and Nichols [50] model was used to generate SNP data under the discrete subpopulation framework similar to studies in Price et al. [6] and Li and Yu [7]. Additional simulations explore the effect of varying levels of relatedness among populations, contrast nonparametric clustering with the k-medoids cluster method employed by CMDS, and test the power in the NMDS method for finding interactions between SNP and ethnicity associated with disease risk. Gower's distance metrics and nonmetric MDS as well as the EIGENSTRAT method are implemented using JMP Genomics®Version 4.0 software (SAS Institute Inc.) and nonparametric clustering is performed with the MODECLUS procedure in the SAS/STAT package of SAS®Version 9.2 software. The CMDS method along with Gap statistic calculation is computed using R software packages.

3.3.1 Disease Association Simulation Results

We generated independent sets of 10,000 SNPs for 500 cases and 500 controls from two separate populations (with 80% cases and 40% of controls for the 1^{st} population and the rest taken from population 2)[2] to use for calculating coordinates via NMDS, CMDS, and Eigenstrat. Population allele frequencies were randomly drawn as

$$p_{lk} \sim Beta\left(\frac{(1-F_{ST})p_l}{F_{ST}}, \frac{(1-F_{ST})(1-p_l)}{F_{ST}}\right),$$

where F_{ST} measures the degree of population differentiation [21]. We use a value of $F_{ST} = 0.01$ representing divergence indicative of European populations. The ancestral allele frequencies at each loci, p_l, were sampled from the $Uniform(0.1, 0.9)$ distribution. For each set, 100 *random null* SNPs (unassociated to case-control status under the Balding-Nichols model with $F_{ST} = 0.01$) as well as 100 *differentiated null* SNPs (allele frequencies of 0.90 in population 1 and 0.10 in population 2) were generated for testing Type I error. Disease *associated* SNPs for testing power were generated using the multiplicative risk model adopted from Price et al. [6] where controls from population k were assigned 0, 1, or 2 copies of an allele under the $Binomial(2, p_{lk})$ distribution while genotypes for cases were drawn from a

$$Binomial\left(2, \frac{Rp_{lk}(Rp_{lk} + (1-p_{lk}))}{(1-p_{lk})^2 + 2Rp_{lk}(1-p_{lk}) + R^2p_{lk}^2}\right),$$

where R is the relative disease risk for the causal allele and p_{lk} is the allele frequency at each loci for populations $k = 1, 2$.

To address power issues when a portion of the data used to infer coordinates show evidence of disease association, we analyzed groups of 10 generated SNP sets where the 10,000 SNPs contained 0%-10% associated SNPs under relative risk $R = 1.3$ (typical low to moderate effect). Additional analysis was done on sets of 10,000 SNPs with 0%-5% disease-associated SNPs when the relative risk was set to 1.5. At every percentage change of SNP composition, 10 sets of 100 disease-associated, 100 random unassociated and 100 differentiated SNPs were tested under NMDS, CMDS, and EIGENSTRAT models. For SNP sets with 1% or more disease-associated SNPs included in coordinate computation, an additional 100 of those associated SNPs were tested to infer power as well. Significance was determined on the p-values using a threshold of $\alpha = 0.05$. The traditional Chi-squared test with no PS correction was also computed and the Type I error and power rates for the three categories of markers is given in Table 3.1.

[2] Other case-control proportions were also examined and yielded similar results to those found in Price et al. [6] and Li and Yu [7] (results not shown).

Table 3.1: Type I error and power rates (and associated standard errors) for the traditional Chi-squared test with no PS correction for random null, differentiated null, and disease-associated SNPs for risk models with values of 1.3 and 1.5 relative risk. Significant associations determined by $p < 0.05$.

SNP Type	χ^2 for R=1.3	χ^2 for R=1.5
Random Null (Type I Error)	0.219 (0.013)	0.195 (0.013)
Differentiated Null (Type I Error)	1.000 (0.000)	1.000 (0.000)
Associated SNPS (Power)	0.652 (0.015)	0.884 (0.010)

Figure 3.2 and Figure 3.3 show the Type I error and power comparisons between the three methods under the risk model for disease associated SNPs using $R = 1.3$ and $R = 1.5$ respectively. The proposed NMDS method as well as the CMDS method adequately control Type I error for both random and differentiated (both appear to be slightly conservative, yet not significantly so in this case) SNPs in all scenarios. EIGENSTRAT does not sufficiently control Type I error at the nominal level for differentiated SNPs (agreeing with findings of Li and Yu [7]), yet as the percent of disease associated SNPs increases, that error rate decreases for both relative risk rates. This interesting finding appears to be indicative of an overall loss of power for both true and false positive findings when associated SNPs confound the estimation of population structure in the axes of variation.

Power is similar among the three methods when there is little to no SNPs that show a positive association in the SNP set used to form the coordinates. For relative risk equal to 1.3, CMDS and EIGENSTRAT have significantly lower power compared to NMDS when around 6-8% markers of the SNP set included show evidence of disease association. For SNPs generated with a relative risk of 1.5, the results are more dramatic; NMDS is significantly more powerful with just 3% or more of associated markers used to form NMDS, CMDS, and PCA coordinates. When testing the power of SNPs that are associated to the trait as well as included in coordinate estimation (plots B of Figure 3.2 and Figure 3.3), the decrease in power for CMDS and EIGENSTRAT is much more dramatic (a much steeper negative slope) resulting in power estimates that show NMDS to have around a 20% power increase over the other methods (42.4% for EIGENSTRAT, 37.1% for CMDS, and 61.3% for NMDS) when 5% of SNPs included in dimension reduction calculation are show disease association.

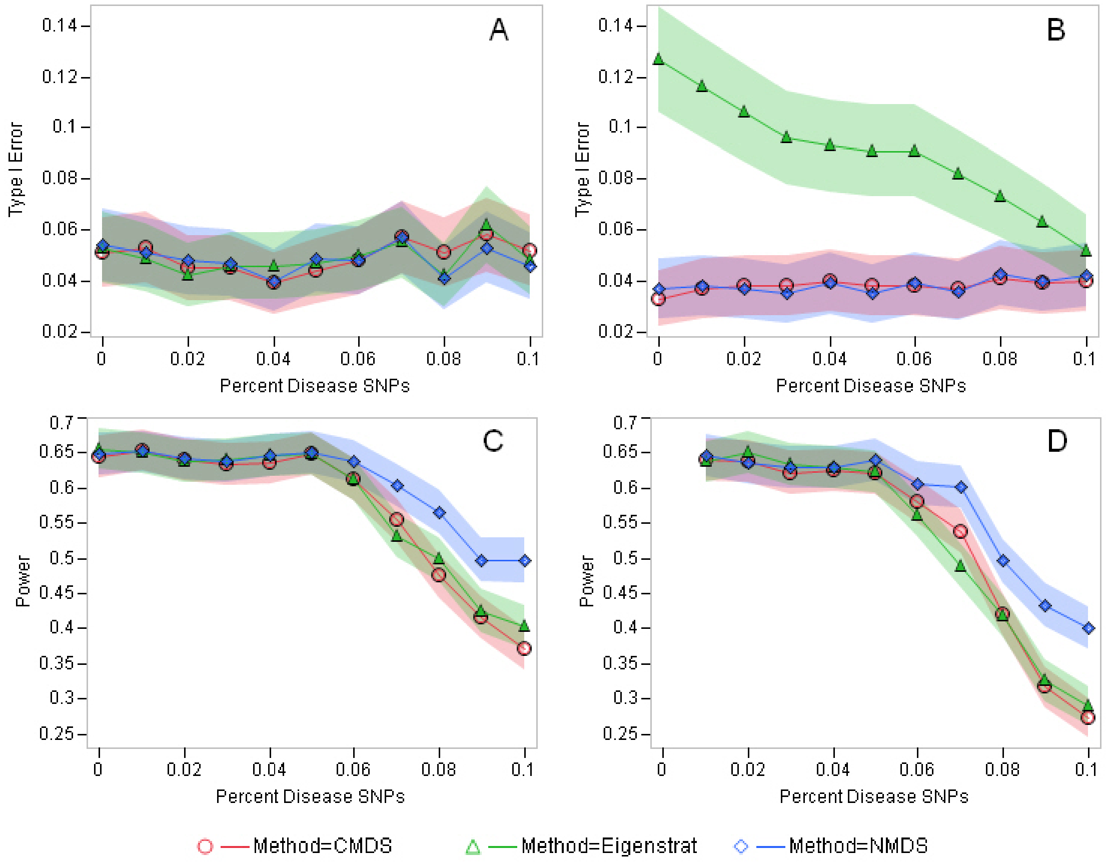

Figure 3.2: Type I error and power results for NMDS (diamonds), CMDS (circles), and EIGENSTRAT (triangles) for $R = 1.3$ under an increasing percentage of associated markers used to form coordinates. Plots A and B show respective Type I error rates for the sets of random unassociated SNPs and differentiated SNPs. Power comparisons for associated SNPs not used to form dimension coordinates (plot C) as well as for those used in forming the NMDS, CMDS, and PCA axes (plot D) are given. Shading around lines represent the 95% confidence intervals for each estimate. Significant associations determined by $p < 0.05$.

While the power decreases in NMDS as well (as expected), NMDS is more robust to the influence of the disease associated SNPs while still capturing population structure. Under the 10,000 SNPs generated for inferring axes to capture PS, significant power differences occur when only a few hundred SNPs in the set show very modest effect sizes with disease association; a common situation in association studies before stringent testing thresholds are applied to declare a SNP significant (and possibly even more common under the RVCD hypothesis).

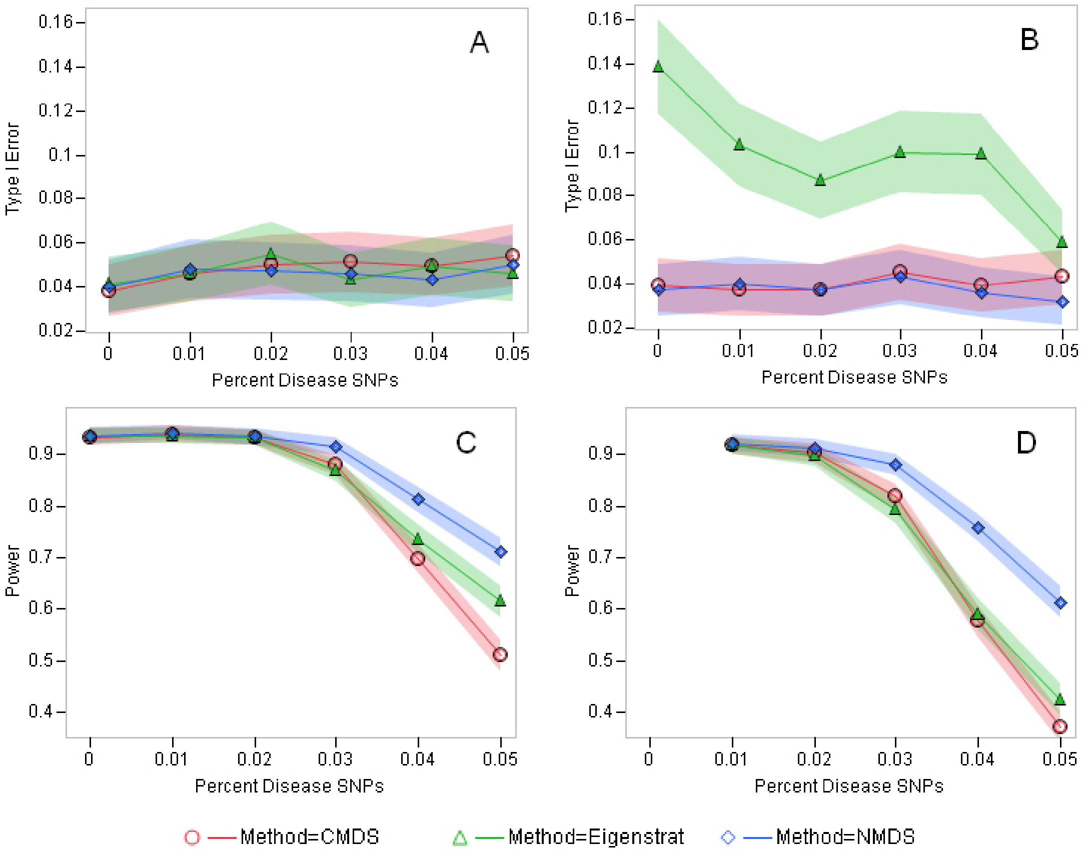

Figure 3.3: Type I error and power results for NMDS (diamonds), CMDS (circles), and EIGENSTRAT (triangles) for $R = 1.5$ under an increasing percentage of associated markers used to form coordinates. Plots A and B show respective Type I error rates for the sets of random unassociated SNPs and differentiated SNPs. Power comparisons for associated SNPs not used to form dimension coordinates (plot C) as well as for those used in forming the NMDS, CMDS, and PCA axes (plot D) are given. Shading around lines represent the 95% confidence intervals for each estimate. Significant associations determined by $p < 0.05$.

3.3.2 Varying Levels of Relatedness and Visualization

Further study was done where SNPs were generated for three populations under the Balding-Nichols model with varying levels of relatedness (F_k in place of F_{ST} for $k = 1, 2, 3$). Figure 3.4 shows the coordinates for nonmetric MDS and classical MDS (synonymous to PCA here) calculated from 10,000 random SNPs (no disease-associated markers included) for 500 individuals drawn from three populations with relatedness measures of

$F_k = \{0.03, 0.02, 0.005\}$. While CMDS simply finds the clustered populations, NMDS accurately reflect the relatedness within populations by the varying sizes of the clusters. It is quite common to find varying relatedness (common ancestry) of individuals within different ethnicities; which is evident even from the HapMap populations [32]. This lack of sensitivity to multiple levels of relatedness found in CMDS and PCA would have detrimental effects when a possible disease of interest has strong epidemiological interactions with regional or spatial influence.

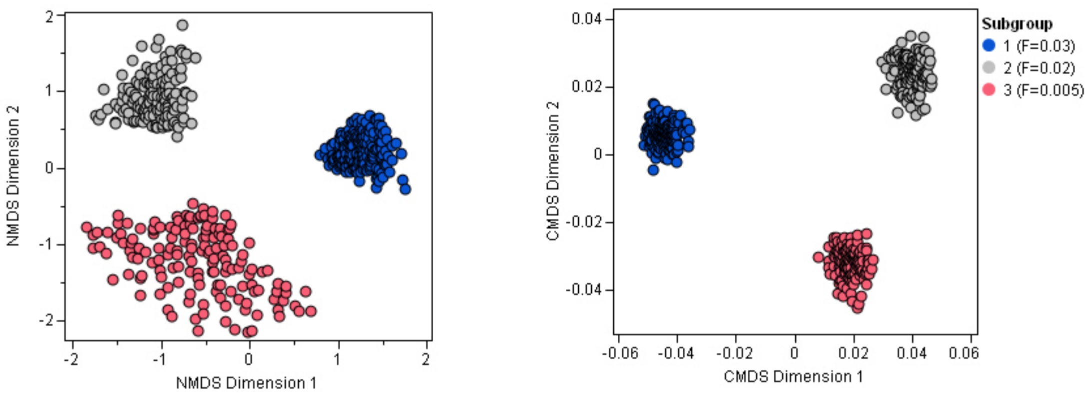

Figure 3.4: NMDS and CMDS display of 500 individuals sampled from 3 populations with varying levels of relatedness ($F_k = \{0.03, 0.02, 0.005\}$).

3.3.3 Clustering Comparisons

Table 3.2: Percentage of sets correctly determining the number of clusters by the Gap Statistic in conjunction with k-medoids vs. nonparametric clustering with the kernel radius determined by an estimated smoothing parameter. True data is composed of two populations with $F_{ST} = 0.004$

Clusters	K-medoids + Gap	Nonparametric Clustering
1	27%	0%
2	73%	92%
3	0%	6%
4	0%	2%

An additional 100 sets of 10,000 SNPs were generated for two populations under the Balding-Nichols model with $F_{ST} = 0.004$ to be indicative of population structure found in an European-American sample [6]. This simulation set is used to briefly compare the nonparametric clustering method we employ against k-medoids with the Gap statistic used in CMDS. For each set of SNPs, 1000 individuals were sampled and nonmetric MDS coordinates for two dimensions were computed on the individual distance matrix using Gower's dissimilarity metric. Both nonparametric clustering and k-medoids were performed on the NMDS coordinates for each set, with the number of clusters determined by our choice of smoothing parameter and the Gap statistic respectively. Estimation of the Gap statistic to choose the number of clusters for k-medoids followed the algorithm briefly described by Li and Yu [7] and derived from Tibshirani et al. [24] in more detail. Table 3.2 shows the results for the number of clusters determined over the 100 iterations.

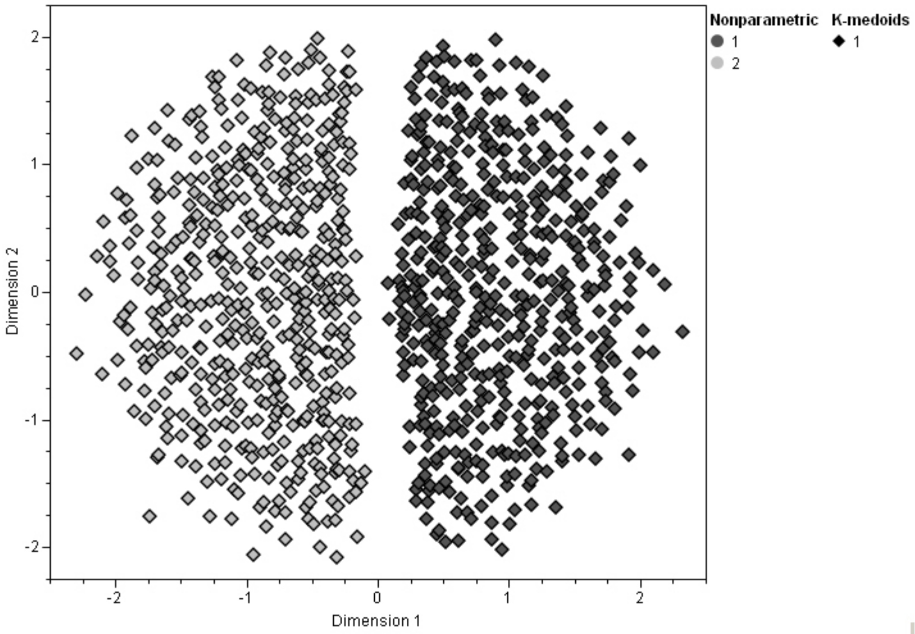

Figure 3.5: NMDS coordinate example of a dataset simulated with two discrete populations with relatedness/differentiation measured by $F_{ST} = 0.004$. Nonparametric clustering on the 1000 individuals is indicated by color (two clusters) while k-medoids clustering with the Gap statistic is notated by the symbol (only one cluster found).

The nonparametric clustering method found the correct number of clusters in 92% of the simulation runs in comparison to 73% for k-medoids with the Gap statistic computation. Figure 3.5 highlights a case where nonparametric found two clusters while the Gap statistic only determined one cluster with k-medoids[3]. The remaining proportion of datasets for k-medoids were determined to have only one cluster whereas nonparametric clustering appeared to be over-sensitive in estimating the number of clusters. Under the assumption that population stratification exists, over-estimation of the number of clusters would be likely to have less dire consequences.

3.3.4 SNP-Ethnicity Interaction Association Results

A final simulation scheme was performed to exemplify how the proposed NMDS correction model can be expanded to test for interactions of genotype with environment/ethnic backgrounds contributing to trait susceptibility. Two populations were generated with $F_{ST} = 0.01$ and again 10 sets of 10,000 SNPs were generated to infer NMDS coordinates as well as 100 random null SNPs to evaluate Type I error and 100 causal SNPs to test power. Two risk model scenarios were used to generate causal SNPs in the two populations where causal SNPs were assigned alleles 0, 1, or 2 drawn from a

$$Binomial\left(2, \frac{R_k p_{lk}(R_k p_{lk} + (1 - p_{lk}))}{(1 - p_{lk})^2 + 2R_k p_{lk}(1 - p_{lk}) + R_k^2 p_{lk}^2}\right)$$

for the case population. To emulate an interaction effect that would correspond to a relative risk of 1.3 with disease status, $R_k = (1.15, 1.5)$ values were used to generate alleles for the cases in population 1 and 2. This translates to a 30% increase is disease risk given the presence of the causal allele if an individual is from population 2 as opposed to population 1. An additional risk model of $R_k = (1.0, 1.5)$ was also tested to emulate an interaction relative risk of 1.5 for disease. These scenarios show how the PS correction model can test if disease susceptibility at a causal SNP changes between two discrete populations.

Under the proposed NMDS correction model, an additional parameter was added to model the interaction of SNP×Cluster. Figure 3.6 shows the power of testing that interaction effect over a range of sample sizes. Results show that a much larger sample size (around 3,000 to 4,000 individuals) is needed in order to obtain similar power for testing disease interaction with population membership and genotype information. In well powered studies with large sample sizes, this type of analysis is recommended when possible discrete

[3] Note that the resulting NMDS coordinates accurately represent two discrete populations with very little relatedness among them such that the only defining population structure is distinguishing between the two groups.

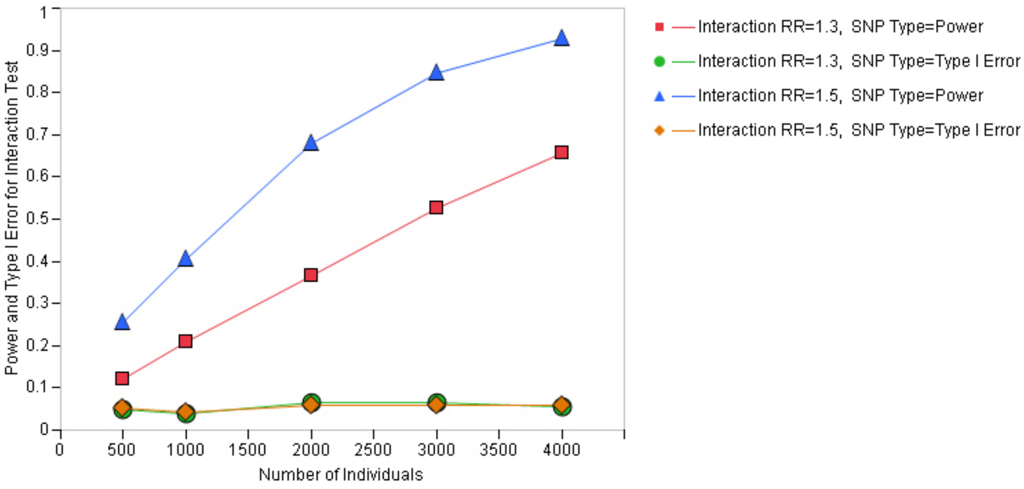

Figure 3.6: Type I error and power for testing interaction effect of SNP*Cluster on disease.

population structure exists in order to determine if disease association across loci tested differs among populations.

The value of the interaction test between population membership (such as ethnic groups or even regional groups) and genotype is that it allows the researcher to primarily determine if a disease exhibits varied regional susceptibility as well as lending more power to further studies when testing SNP association. For example, if a study contains a small number of individuals from a distinct ethnic background possessing loci that substantially increase disease risk while a another larger sample of individuals from a different population shows little genetic susceptibility; the main effect test averaged over both populations would be difficult to interpret and could diminish the chances of a significant finding. An interaction test would signify further study of disease association to be carried out separately for these populations. This is a plausible scenario where disease prevalence is negatively correlated with subpopulation size. Findings from Gorrochurn et al. [65] show that this scenario is one where population stratification is most damaging and give an example of a small Jewish founder population that had 10-50 times higher allele frequency for mutations related to breast cancer in comparison with the general population [66]. In such a case, the main effect test would likely be powerless to detect the causal loci if individuals from the founder population are sampled along with individuals for the general population.

3.4 Discussion

Dimension reduction methods such as principal components analysis and multidimensional scaling have become invaluable tools for quantifying genetic data, particularly for elucidating structural patterns among individuals that may lead to bias in association studies. Simulation results show the proposed NMDS method to accurately correct for inflation in false positive findings due to PS in case/control genetic association studies. Moreover, the NMDS method is more robust to markers that may confound the detection of population structure such as loci that show allelic variation that correlates with case/control status and yields significantly higher power rates in this scenario.

In our simulations, the SNP set used to infer coordinates via the three dimension reduction methods were solely proportions of trait-related SNPs to substructure-informative SNPs. Many loci in a study will not exhibit allele frequency variation among populations and so the proportion of associated markers in comparison to the full set of SNPs analyzed in GWAS can be much smaller than percentages studied in our simulation scheme. Inclusion of such (non-varying) SNPs can make effect of PS harder to detect and may allow the associated markers to drive coordinate estimation even more than shown in our study. While plots of components can be used to determine if the coordinates are over-correcting for case/control status, such visualization may not be possible when individuals in a study come from many populations.

The use of a covariate-based logistic model has been shown to work well for correcting population stratification in past studies [67, 7] and is easily expanded to model additional phenotypic covariates and interaction effects. Our research builds on this simple correction model by improving the estimation of the covariates used to infer population structure in terms of accuracy (to detect multiple levels of relatedness) as well as robustness to over correcting for case/control status. The use of nonparametric clustering methods enhances the model to be more sensitive to stratification caused by discrete populations.

3.5 Supplemental Information

Tables of Type I error and power estimates corresponding to the simulation results presented in Figure 3.2, Figure 3.3, and Figure 3.6 can be found in Appendix B.

URLS: SAS/STAT online documentation for MDS and MODECLUS procedures can be found at http://support.sas.com/documentation/onlinedoc/stat/index_proc.html.
JMP Genomics®Software information is found at http://jmp.com/software/genomics.

Table B.1: Type I Error Estimates and Standard Deviations Under Relative Risk=1.3 for CMDS, NMDS, and EIGENSTRAT corresponding to Figure 3.2.

Table B.2: Power Estimates and Standard Deviations Under Relative Risk=1.3 for CMDS, NMDS, and EIGENSTRAT corresponding to Figure 3.2.

Table B.3: Type I Error Estimates and Standard Deviations Under Relative Risk=1.5 for CMDS, NMDS, and EIGENSTRAT corresponding to Figure 3.3.

Table B.4: Power Estimates and Standard Deviations Under Relative Risk=1.3 for CMDS, NMDS, and EIGENSTRAT corresponding to Figure 3.3.

Table B.5: Type I Error and Power Estimates with Standard Deviations Under Relative Risk Values 1.3 and 1.5 for the NMDS Model Testing for Disease Interaction with Causal SNP and Population Membership corresponding to Figure 3.6.

4

Evaluating the Effects of Batch Composition and Size on BRLMM Genotype Calls for the Affymetrix 500K Array

4.1 Background

The advent of high-throughput genotyping technologies has facilitated many success stories for genome-wide association studies (GWAS) in finding genetic variants associated with common complex diseases. As we gain understanding about common complex disease heritability, the size and scope of GWAS continue to grow. Past results have shown that significant associations found in GWAS are typically with common SNPs showing very little to moderate effects with relative risks in the range of 1.1 to 1.3 [61]. One explanation for lack of reproducibility in candidate gene studies and early GWAS was an inadequate sample size to obtain the level of power necessary to capture a moderate effect [68]. GWAS have since begun to rectify this issue with studies now containing thousands of individuals, as well as adopting multi-tiered replication study designs. With such large-scale studies, minimal errors can introduce bias that result in inflation of both the Type I and Type II error rates, making adequate quality control essential to ensure reliability of GWAS results

[61]. Population stratification and genotype error rates are two main sources of bias in GWAS and have been shown to result in inflation of the test statistic by over 11%, where more than half of that overdispersion could be attributed to bias introduced by inaccurate genotype calls [14]. While clustering algorithms used to call genotypes continue to improve in accuracy [9, 12, 13], even miniscule errors can result in large biases due to a shifting paradigm that has researchers looking to rare variants (with minor allele frequencies less than 1-5%) to find associations with complex diseases, leading to new studies of now tens of thousands of individuals [59].

The Affymetrix GeneChip Human Mapping 500K array set is a common genotyping platform used in many GWAS [18, 69, 70, 71, 72] . Hence the genotype calling algorithm recommended by Affymetrix for use with the Affy 500K arrays, Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM) [11], one of the most widely used algorithms for producing genotype calls. BRLMM is a multi-chip clustering algorithm that assigns calls after a SNP RMA (Robust Multiarray Average) process of normalization, transformation, and summarization of probe level intensities for the 'A' and 'B' alleles. The ability to simultaneously call genotypes for a set of samples at multiple loci has been shown to improve accuracy due to better estimation of genotype (AA, AB, BB) cluster centers [11]. One issue with a multi-chip approach is the introduction of batch effects on the resulting genotype calls. Recent research has shown that discordant results can arise due to changes in both batch size (the number of samples processed simultaneously through BRLMM at a time) and batch composition (including or separating cases and controls and/or heterogeneous samples from different populations in batch sets) [14, 15, 17].

Using a sample of 816 cases for type 1 diabetes and 877 controls for 6,322 SNPs, Clayton et al. [14] found that the clustering clouds for the three genotypes differed positionally for cases and controls and can result in ambiguous genotype calls when samples are combined. Yet in their study, it is noted that calling cases and control separately may also lead to overdispersion of the test statistic in association testing. Moreover, using high thresholds in QC and only calling genotypes with high confidence can lead to differential bias due to non-independence in genotypes set to missing; as well as possibly excluding valuable data. Miyagawa et al. [39] also found that the effects of differential bias due to genotype errors can be remedied with stringent QC (namely SNP call rate), but their QC steps resulted in excluding approximately 250,000 SNPs from the Affymetrix 500K array in their study. Anney et al. [16] reported Type I error rates in a simulated case-control association study up to 59% when up to 5% non-random missingness was generated in the set of case subjects. Such findings indicate that while it may be inadvisable to call cases and controls together (due to differences in data collection, DNA source and preparation etc..), calling cases and controls separately can lead to differential bias that inflate the test for association. Moreover, Plagnol et al. [15] state that allowing different *a priori* frequencies for genotype clusters in calling algorithms between cases and controls conflicts with the null hypothesis of no association with genetic markers and case-control status.

Hong et al. [17] used the 270 HapMap samples for the Affymetrix 500K array set to evaluate batch size and composition effects on call rates and genotype call concordance. Batch sizes of 30, 45, and 90 individuals were shown to have concordance rates for both homozygous and heterozygous calls above 99.9% yet also showed significant differences in sample and SNP call rate results. Similar findings were reported for batch composition where batches were comprised of combinations of Asian, African, or European HapMap samples. Hong et al. [17] also demonstrated that batch differences can be propagated to association testing results by treating one of the HapMap populations as a 'case' and the others as a 'control' to mimic a case-control study. A primary recommendation from the study was use of larger batch sizes and homogeneous (in terms of ancestral population background which are well known to have allele frequency differences) batch compositions. Such results highlight the fact that batch effects can significantly influence the outcome of a study and necessitate a better understanding of batch effect behavior in the BRLMM algorithm extended to GWAS with thousands of individuals. In our study, we extend the evaluations of BRLMM batch effects to better understand the implications on further downstream analysis of a case-control study.

Using 3491 individuals obtained from the Wellcome Trust Case Control Consortium (WTCCC) [18] for coronary artery disease, CAD (1991 individuals), and a set of controls (1500 individuals) for the Affymetrix 500K array set, we determined how batch size (on a larger scale) and composition[1] (using case and control sample combinations) affects the results of common quality control steps in GWAS and subsequent association testing. We show that both batch size and composition can greatly change the results of

[1] Note that here batch composition has a different interpretation from the study by Hong et al. [17] as we are looking at cases and controls assumed to be from a homogeneous population as opposed to different major world populations.

GWAS in terms of the set of SNPs that pass QC as well as the set of SNPs determined to be highly significant. Through evaluation of subject-specific differences in the probability for a marker to pass QC and subject-specific differences in the probability for a SNP to be deemed significant, our results show that batch size and composition and their interaction have both practical and statistically significant effects on GWAS.

4.2 Results

Five datasets of genotype calls were generated under different levels of batch size and composition for the WTCCC data (see Section 4.4): **C500**, **C2000**, **C3500**, **S500**, and **S2000**, where the dataset label conveys the batch composition (Combined or Separate) and size (500, 2000, 3500) used in the BRLMM algorithm. These datasets were then all subjected to identical quality control steps and subsequent association testing. Results of QC and testing were utilized in generalized linear mixed models to test if batch size and composition (and their interaction) significantly changed the result of both QC SNP exclusion and significant disease association.

4.2.1 Quality Control

Quality control to exclude both individuals and markers prior to association testing was carried out for all the datasets. Figure 4.1 shows work-flows of the number of individuals (top figure) and number of SNPs (bottom figure) excluded at each QC step. The resulting number of individuals and SNPs given in the far right boxes for each dataset was subsequently used for SNP association testing. For individual exclusion, there is a slight trend in more individuals being excluded with a call rate less than 97% as batch size increases for both combined and separate batch compositions. This trend was also followed for the QC exclusion step for SNPs with a minor allele frequency less than 1% or a call rate less than 95%. As batch size increased, many more SNPs were excluded due to low call rates (minor allele frequency differences were negligible, results not shown), indicating that the number of samples ran simultaneously through BRLMM seems to be negatively correlated with the confidence[2] of the genotype call (the threshold for setting a SNP either to a geno-

[2] We define confidence as 1 minus the ratio of the distance of the closest genotype cluster over the second closest cluster. See Section 4.4 for more detail.

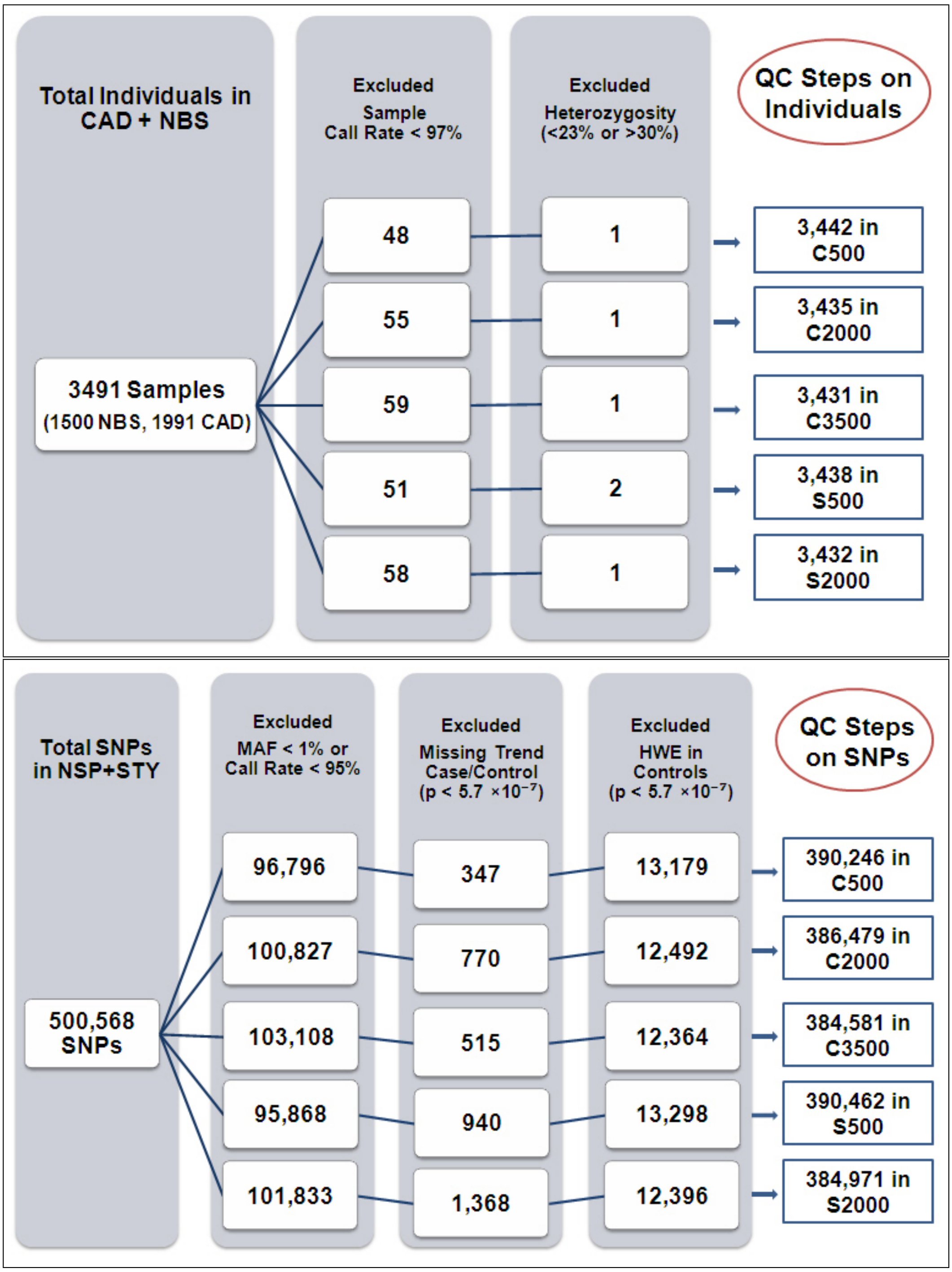

Figure 4.1: Results of quality control for excluding individuals (top figure) and SNPs (bottom figure) from association analysis for each of the five datasets.

type or missing). Quality control results for SNPs show that for batch composition, more SNPS tended to have significant differences in call rate between cases and controls when cases and controls were called in separate batches. From the bottom chart in Figure 4.1, a much larger proportion of SNPs were found to have highly significant proportions of missing data for cases vs. controls even after excluding markers with more than 5% missing calls overall. The number of SNPs that passed QC steps indicated that as batch size increases, our QC protocol excluded a larger amount of SNPs for both levels of batch composition.

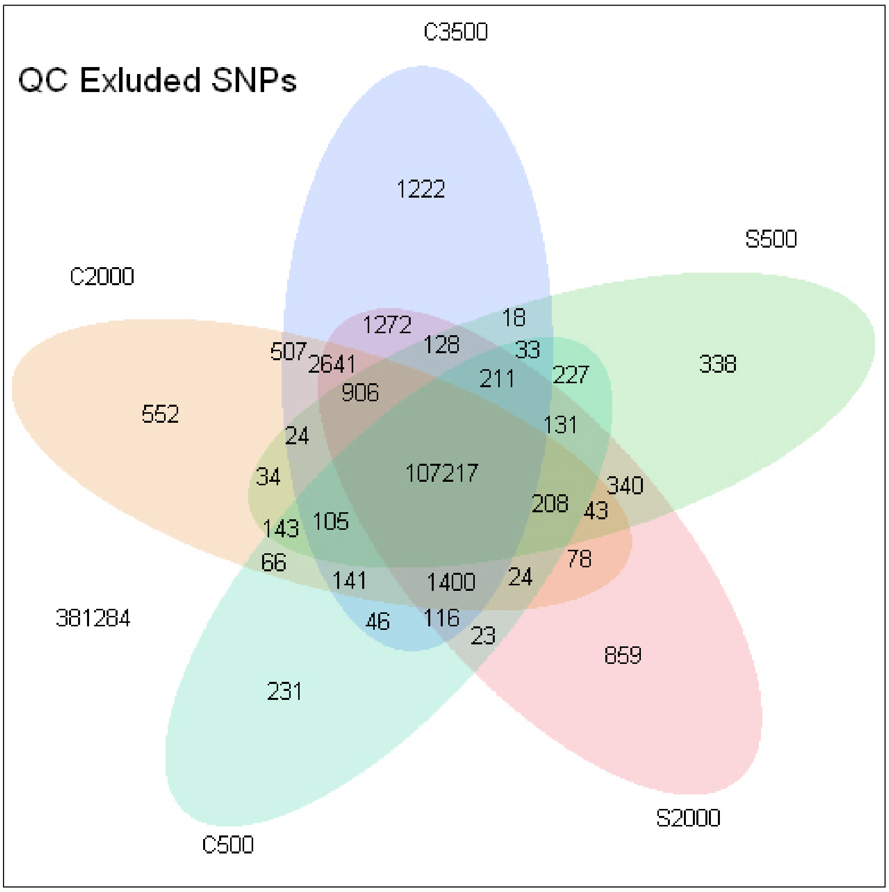

Figure 4.2: Five-way Venn diagram of the counts of SNPs excluded from each dataset as a result of the QC process.

Batch effects not only influenced the number of SNPs that were excluded by QC; the sets of SNPs excluded included thousands of discordant results. While the number of SNPs dropped from a given batch set was approximately 110, 000 to 116, 000; only 107, 217 loci in common were excluded in all 5 datasets. The proportion of discordant SNPs due to QC between two batch sets is no greater than 2%, yet this translates to several thousand markers that were dropped due to QC for genotype calls from one batch size and composition formulation but were analyzed for significant associations in another. The Venn diagram in Figure 4.2 shows the breakdown of all SNP counts that were either concordant or discordant across batch sets.

4.2.2 Association Testing Results

Markers that survived QC were deemed significant using an $\alpha = 5.0 \times 10^{-7}$ threshold for the trend test for association with CAD status in each of the 5 batch sets. Figure 4.3 displays the p-values for SNP calls from C500, C2000, C3500, S500, and S2000. The top plot in Figure 4.3 reports the $-\log 10(p\text{-value})$ across the genome sorted by Chromosome and Mb position and aids in understanding the overall trend in the behavior of the p-values among the different batch sets. In all batch sets, several SNPs in the chromosome 9 region were highly significant (similar to findings in the original WTCCC analysis). The results for S500 indicate a genome-wide trend of higher p-values that is constant across all chromosomes. C500 also shows similar results of slightly more inflated p-values in comparison to the batch sets of C2000, S2000, and C3500 (which appears to have the lowest trend in p-value magnitude). The bottom scatter-plot matrix in Figure 4.3 plots the pairwise p-values for each batch set on the $-\log 10$ scale for all SNPs that passed QC in both datasets. The solid gray indicates the 45^o angle and the dotted gray lines fall at the $\alpha = 5.0 \times 10^{-7}$ level [18]. Points that fall off the diagonal line indicate SNPs with discordant p-values and any SNPs in either of the two off-diagonal rectangles boxed in by the dotted lines are those that result in differential significance decisions. There was very little discordance found between the C2000 and C3500 batch sets while S500 has the most discordance in comparison with all other datasets. Overall, discordant p-values tended to be more significant in data with separated case-control batch composition and smaller batch sizes.

One of the main goals of this study was to evaluate how the list of SNPs that are deemed statistically significant is influenced by batch variations for genotype calling with the BRLMM algorithm. Figure 4.4 contains Venn diagrams representing the counts of SNPs with a $p < 5.0 \times 10^{-7}$ for one batch set compared with results from another batch set. Five pairwise comparisons were of interest to interrogate effects of batch size (C500

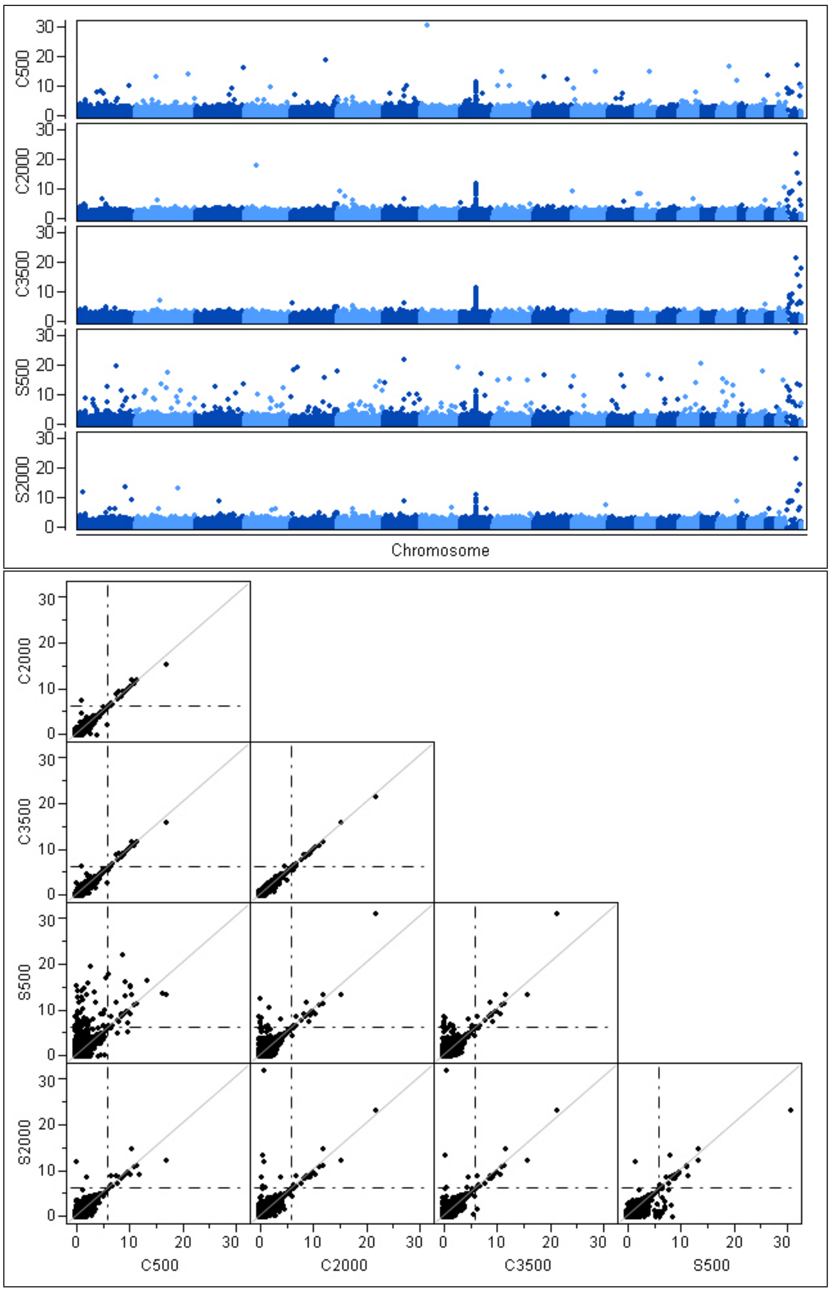

Figure 4.3: Genome-wide scatter plots of p-values for each dataset (top figure) on the -log10 scale. Chromosomes are alternatingly colored for distinction. The bottom figure shows the pairwise comparisons of the -log10(p-values) between C500, C2000, C3500, S500, and S2000 datasets. Points that fall off the diagonal line indicate discordant results, the dotted line is the 5.0×10^{-7} significance threshold. Only markers that passed QC in both datasets are plotted.

vs. C2000, C500 vs. C3500, and S500 vs. S2000) and composition (C500 vs. S500 and C2000 vs. S2000). The areas of the circles that do not overlap show that with the same data (allele intensities), different batch processing can produce highly discordant lists of significant findings.

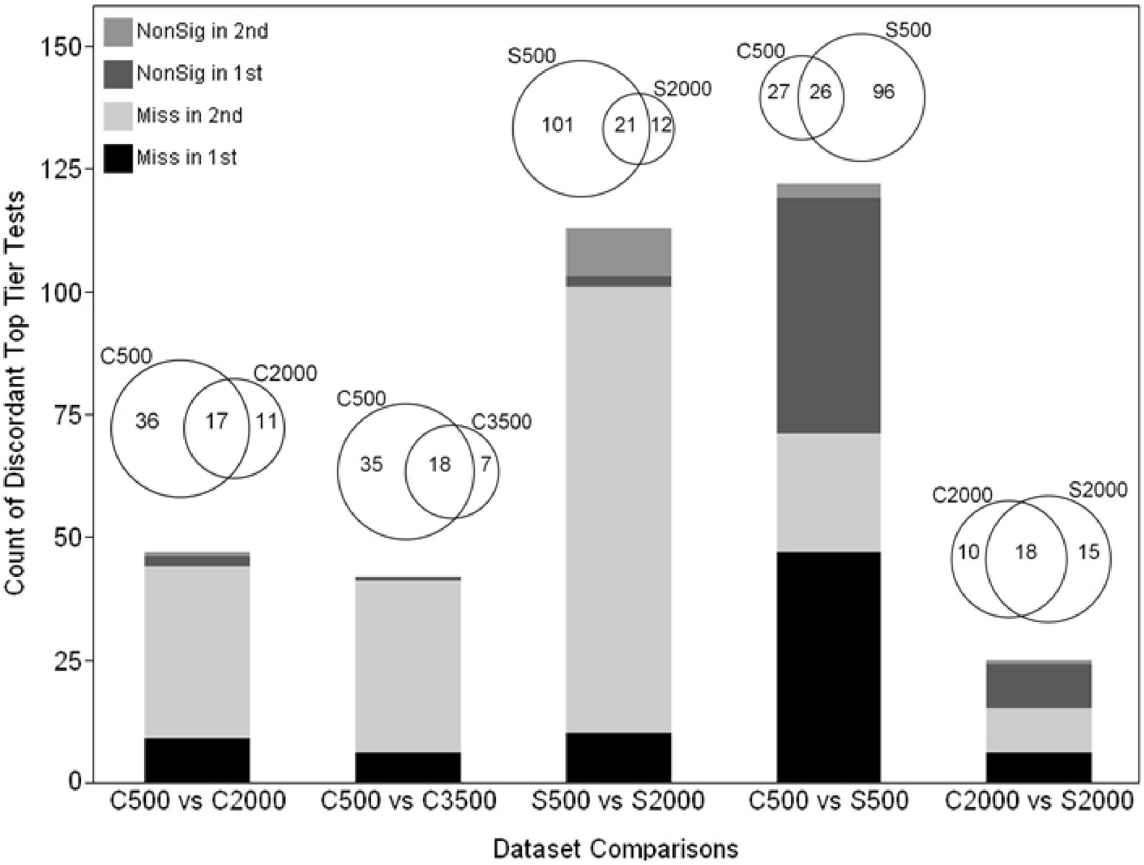

Figure 4.4: Counts of concordant/discordant significant SNPs ($p < 5.0 \times 10^{-7}$) among datasets. The Venn diagrams above the bars indicate the number of SNPs that were found significant in either or both data sources for pairwise comparisons of: C500 vs. C2000, C500 vs. C3500, S500 vs. S2000, C500 vs. S500, and C2000 vs. S2000. The bars below the Venn diagrams represent the count of discordant significant SNPs for that Venn (i.e. the area where the circles do not overlap), broken down by differences due to QC exclusion or association testing. Bar color indicates the source of differential results and the color legend corresponds to the order of the comparison labeled on the X-axis.

For each Venn diagram, the source of discordance (non-overlapping circles) can be categorized as due to quality control (a SNP was excluded and hence missing in one dataset but found significant in another) or due to differential test results (the SNP passed QC in both datasets but was only found significant in one batch set). The bars below the Venn diagrams represent the source of discordance for that pairwise batch set comparison and are broken down by color where black corresponds to the SNPs that were missing (excluded due to QC) in the first dataset in the comparison label but found significant in the second batch set. Light gray corresponds to discordant SNPs found significant in the first batch set but QC excluded in the second. Dark gray indicates the significant SNPs in the second batch set that were not significant in the first. Finally medium gray corresponds to SNPs found significant in the first batch set in the comparison but not in the second. Using the first bar on the left as an example, black (at the bottom of the bar) represents the count of SNPs that were found significant in C2000 but were excluded due to QC in C500 (i.e. missing in 1^{st} dataset listed on the X-axis label), light gray (second color from the bottom) is the count of discordant SNPs that are significant in C500 but QC excluded in C2000, the darker gray is the number of SNPs that were significant in C2000 but not in C500 (differential testing) and the medium gray is the count of SNPs that were significant in C500 but not in C2000. The medium and light gray portions of the bar correspond to the SNP count in the left non-overlapping Venn diagram portion, while the black and dark gray add up to the count found in the right non-overlapping portion of the Venn diagram circle.

The largest amount of overall discordance was found between C500 and S500 and also shows by far the largest amount of discordance due to association testing results. The comparison S500 vs. S2000 shows the largest amount of discordance due to QC exclusion where many SNPs that were found significant in S500 were excluded due to QC in S2000 (light gray). Comparisons of different batch sizes showed a much higher proportion of discordance due to QC exclusion as opposed to test results in contrast to batch composition comparisons. The discordance for C500 vs. C2000 and C500 vs. C3500 showed similar patterns, indicating that an increase in the magnitude of the batch size differences beyond 2,000 did not appear to influence results. While C500 vs. S500 showed the most discordance, C2000 vs. S2000 showed the least; hence the effect of batch composition was less severe (in terms of the significant SNP list results) for a larger batch size and suggests an interactive effect of batch size and composition. The take-away information obtained from Figure 4.4 is that BRLMM calls using larger batch sizes where cases and controls are combined within batches leads to more conservative, yet more concordant, results for significant association.

4.2.3 Stringent Call Rate Threshold Evaluation

Miyagawa et al. [39] found that stringent data cleaning (particularly SNP call rate thresholds) could reduce the false positive rate to nominal levels (which eliminated over 50% of SNPs in the study). We found that by increasing the SNP call rate necessary to pass QC to 99% we could eliminate nearly all differential significant SNP results in the 5 datasets. Table 4.1 gives a list of loci that were found to be significant at the chosen threshold of 5.0×10^{-7} for any of the datasets tested. All concordant significant SNPs are found on chromosome 9 in the region of SNP rs1333049, the locus also reported by WTCCC. Of the two SNPs with discordant significance results, rs7865618 was found to not be significant in S2000 by the slightest of margins while rs16846351 was excluded due to QC in all but S500. While these results may be encouraging that appropriate data cleansing will eliminate the effects of genotyping errors, using such a stringent SNP call rate threshold forces 40-50% of the data to be excluded from analysis; which is not optimal in association studies geared toward discovery. Under the additional QC constraint, a little under 100,000 additional SNPs did not pass quality control for each dataset leaving SNP counts of 305,984 for C500, 298,382 for C2000, 297,165 for C3500, 308,473 for S500, and 301,511 for S2000 that were used in association analysis.

Table 4.1: List of p-values for concordant and discordant SNPs that were found significant in at least one dataset with a p-value below 5.0×10^{-7} when using a SNP call rate threshold of 99% to pass quality control steps. 'QC' indicates that the SNP did not pass quality control and was not tested for association and '*' indicates a p-value that was not significant.

SNP	Chr.	Pos. (Mb)	C500	C2000	C3500	S500	S2000
Concordant							
rs10965219	9	22.04	1.2×10^{-7}	1.3×10^{-7}	8.5×10^{-8}	1.9×10^{-7}	1.5×10^{-7}
rs9632884	9	22.06	1.7×10^{-9}	2.6×10^{-9}	2.5×10^{-9}	1.9×10^{-9}	3.4×10^{-9}
rs6475606	9	22.07	6.3×10^{-10}	3.8×10^{-10}	7.6×10^{-10}	3.6×10^{-10}	7.3×10^{-10}
rs4977574	9	22.09	1.6×10^{-10}	8.6×10^{-11}	9.0×10^{-11}	1.2×10^{-10}	1.3×10^{-10}
rs2891168	9	22.09	2.0×10^{-10}	1.1×10^{-10}	1.5×10^{-10}	1.3×10^{-10}	2.0×10^{-10}
rs1333048	9	22.12	7.0×10^{-12}	4.7×10^{-12}	7.9×10^{-12}	4.7×10^{-12}	7.6×10^{-12}
rs1333049	9	22.12	2.1×10^{-12}	8.3×10^{-13}	2.1×10^{-12}	3.2×10^{-12}	4.4×10^{-12}
Discordant							
rs7865618	9	22.02	4.2×10^{-7}	$6.3 \times 10^{-7*}$	5.0×10^{-7}	3.8×10^{-7}	4.5×10^{-7}
rs16846351	1	224.91	QC	QC	QC	9.0×10^{-14}	QC

4.2.4 Statistical Tests for Batch Effects

A generalized linear mixed model (GLMM) approach was used to estimate the statistical effect that batch size and composition has on QC and association testing in GWAS. In striving for a more simple balanced design without losing information on batch size and composition, the C3500 batch set is not included in the models. Figures 4.1 and 4.2 clearly show differences in the counts of SNPs that pass QC among the batch sets. In order to test those differences, we built a GLMM to model the probability of a SNP passing QC (Y_1 coded as '0' if the SNP passes QC and '1' if the SNP is excluded in each dataset) dependent on different levels of batch size, composition, and size×composition. A SNP identifier is incorporated as a random effect to account for correlation due to the fact that the same SNP is genotyped in each of the batch sets. An additional GLMM was fit for modeling the probability of a SNP deemed significant ($Y_2 = 1$) using $\alpha = 1.0 \times 10^{-5}$, a threshold that was used in the WTCCC original analysis as a second-tier cutoff for moderate significance that could be recommended for follow-up replication studies. We also found this new threshold to be a statistically appropriate cut-off by visual inspection of the quantile plots given in Figure 4.5, which show that differential divergence from the expected distribution of p-values for the batch sets begins around the $\alpha = 1.0 \times 10^{-5}$ level (corresponding to $-\log 10(p) = 5$). We will refer to the GLMM for QC exclusion as Model I and the GLMM for SNP significance as Model II; see Section 4.4 for further details on the implementation of the logistic models with batch size and composition as fixed effects and SNP as a random effect.

A set of hypotheses were formed to test for subject-specific (i.e. SNP-specific) differences between levels of batch size and composition. Table 4.2 gives the format of the $\boldsymbol{X}\beta$ design matrix and parameter vector for the fixed batch effect levels: C500, S500, C2000, S2000. This table is useful for interpreting hypothesis tests given in Table 4.3.

Table 4.2: Format of the $\boldsymbol{X\beta}$ design matrix and parameter vector for fixed effects modeled in the GLMMS. This table serves as a guide for forming hypothesis tests.

Size	Composition	$\boldsymbol{X}\beta$
500	Combined	$\beta_0 + \beta_1$
500	Separated	$\beta_0 + \beta_2$
2000	Combined	$\beta_0 + \beta_3$
2000	Separated	$\beta_0 + \beta_4$

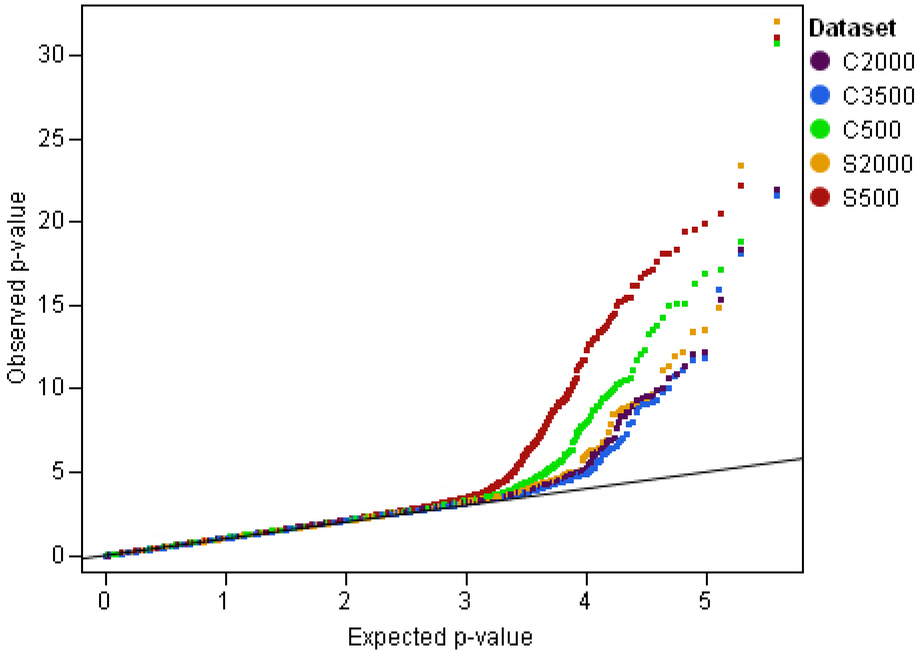

Figure 4.5: Overlaid quantile plots show the distribution of p-values for each of the five datasets on the $-\log 10$ scale. The X-axis plots the expected p-value under the Uniform distribution while the Y-axis is the observed p-values. The black solid line represents the 45^o angle.

Using the GLMM framework, we tested seven hypotheses for both Model I and Model II to estimate subject-specific differences (the difference in probability estimates between a SNP genotyped by one batch set vs. that same SNP genotyped in another batch set). Two overall main effect tests for $Size$ and $Composition$, and a third hypothesis test for the overall interaction of $Size \times Composition$ were performed. The remaining four tests are for simple interaction effects to determine subject-specific differences across levels of batch size at each level of batch composition and vice versa. The result of these tests for Model I and II are presented in Table 4.3.

Due to the extremely large sample size, the power to detect significant differences is extremely high; thus we do not try to use a significance threshold to evaluate p-values (although for Model II there were instances of p-values that are relatively large due to more variability in the estimates as there was a much smaller proportion of data where $Y_2 = 1$). Instead it is more informative to look at the direction and magnitude of the estimates; which

Table 4.3: Estimates and p-values for contrasts (and the associated hypothesis tested) from two GLMMs to test for significant subject-specific differences in QC exclusion and association testing significance results between batch size and composition levels. The estimates are in the form of the log odds ratios associated with the contrast. 'Exclusion' in the Modeling Probability column indicates the random effects logistic modeling of probability that a SNP was excluded due to QC from analysis (Model I) , while 'Significance' indicates the modeling of probability that a SNP was deemed significant at $p < 1.0 \times 10^{-5}$ (Model II).

Contrasts	Hypothesis Tested	Modeling Probability	Estimate	P-value
500 vs. 2000	$H_0 : \frac{(\beta_1-\beta_3)+(\beta_2-\beta_4)}{2} = 0$	Exclusion Significance	-1.725 0.912	1.0×10^{-300} 1.8×10^{-13}
Combined vs. Separate	$H_0 : \frac{(\beta_1-\beta_2)+(\beta_3-\beta_4)}{2} = 0$	Exclusion Significance	-0.196 -0.415	2.1×10^{-23} 8.1×10^{-4}
Size × Composition	$H_0 : \frac{(\beta_1-\beta_3)-(\beta_2-\beta_4)}{2} = 0$	Exclusion Significance	0.292 -0.217	9.7×10^{-50} 8.0×10^{-2}
500 vs. 2000, Combined	$H_0 : \beta_1 - \beta_3 = 0$	Exclusion Significance	-1.434 0.695	1.0×10^{-300} 2.3×10^{-4}
500 vs. 2000, Separate	$H_0 : \beta_2 - \beta_4 = 0$	Exclusion Significance	-2.017 1.129	1.0×10^{-300} 2.2×10^{-12}
Combined vs. Separate, 500	$H_0 : \beta_1 - \beta_2 = 0$	Exclusion Significance	0.096 -0.632	1.3×10^{-3} 2.5×10^{-6}
Combined vs. Separate, 2000	$H_0 : \beta_3 - \beta_4 = 0$	Exclusion Significance	-0.488 -0.198	1.7×10^{-80} 3.4×10^{-1}

are the log odds ratio of the probabilities of $Y_i = 1$ for the contrasts specified in the table. For Model I, Table 4.3 reports a large negative estimate for the '500 vs. 2000' Contrast (-1.725 for the QC Exclusion modeling probability) which indicates there is a much higher probability for a SNP to be excluded from QC in the batch sets of size 2000 than for the same SNP to be excluded in the 500 batch size sets. The main effect contrast for 'Combined vs. Separate' is also negative, yet is of a much smaller magnitude. More importantly, the contrast *Size* × *Composition* shows strong evidence that there is an interactive effect of batch size and composition that influences the probability of a locus being excluded due to QC. The following simple interaction effect contrasts show that there is a large subject-specific difference in probabilities for batch sizes of '500 vs. 2000' when batch composition is separate (-2.017 compared to -1.434 estimate when '500 vs. 2000, Combined'). For simple interaction effect batch composition contrasts, there is very little evidence of subject-specific differences in QC results for the 'Combined vs. Separate' contrast at batch size of 500 (0.096 estimate with a p-value =.0013). The 'Combined vs. Separate' simple interaction effect for size 2000 reported more evidence of differences, indicating that separated batch composition at a batch size of 2000 has a higher probability of a SNP not passing QC; referring back to Figure 4.1, these differences are caused by more SNPs with a lower call rate and more SNPs with significant trends in missing data between cases and controls. The contrast estimates for Model I, modeling the probability of QC exclusion, show that batch size differences have a larger effect on quality control results than does batch composition; although the effect of batch size is slightly less severe for a combined batch composition.

Estimates are generally smaller in magnitude for Model II (Modeling Probability is notated as Significance in the table) compared to Model I, evidence that batch effects influence the results of QC much more strongly than association test results. For Model II, the batch 'Size×Composition' interaction is relatively insignificant at a $p = .08$. Batch size differences again result in the most subject-specific differences in the probability of a SNP being significant (0.912 log odds ratio for '500 vs. 2000' batch size contrast with $p = 1.8 \times 10^{-13}$) with more significant SNPs expected to be found at a batch size of 500. The batch composition main effect contrast indicates a higher probability of significant findings for the separated composition as compared to combined batch sets (-0.415 log odds ratio). The simple interaction effects corroborate these results.

It is important to note that while the estimates for the hypotheses tested in our models showed strong evidence of highly differential results, the subset of discordant outcomes for SNPs across the 4 batch sets was still very low. Overall concordance for the outcome of a SNP to pass QC or to be excluded from analysis in all 4 batch sets was $489,931/500,568 = 97.88\%$. Overall concordance for association testing (using the 1.0×10^{-5} threshold to assign a significant result) was $382,480/393,143 = 97.29\%$ when including differential QC results. For SNPs that passed QC in all 4 batch sets, overall concordance in association testing was $382,480/382,506 = 99.993\%$. In smaller scale studies, these concordance rates would be encouraging. Yet when over 500,000 SNPs are tested, discordance of up to 2-3% can dramatically change the outcome of association testing (largely due to differential QC results) as this translates to thousands of markers influenced by batch effects.

4.3 Discussion

We show that batch size and composition considerations can produce discordant results in terms of quality control decisions as well as the final list of significantly associated SNPs for GWAS data using the WTCCC CAD disease dataset as an example. Figure 4.1 and Figure 4.2 make it clear that increasing batch size from sets of 500 individuals called by the BRLMM algorithm to sets of 2,000 or 3,500 results in more SNPs being excluded from analysis under typical quality control thresholds, as well as differential sets of SNPs that pass QC. Additionally, batch composition differences results in more discordant results at a batch size of 500 as opposed to 2,000. Subsequent testing of association with disease status is also highly affected by batch changes, in large part due to the differential SNPs that passed QC but also due to different trends in the magnitude of p-values as seen in Figure 4.3. These discordant results in quality control and p-value results propagate to the list of SNPs that are deemed significant for the different batch sets as evidenced in Figure 4.4.

Using generalized linear mixed models, we show that batch size and composition effects have a highly significant impact on results of GWAS and while stringent quality control could eliminate much of the discordance, it would also result in a drastic loss of potentially useful data. These results indicate that careful consideration to the implementation of the BRLMM algorithm with the Affymetrix 500K array set should be taken to ensure more reproducible concordant results. The Wellcome Trust Case Control Consortium also found issue with the BRLMM algorithm and proposed a new calling algorithm CHIAMO in the course of their work [18]; which was used to process their samples in one single batch. Other recent research by Carvalho et al. [12] and Lin et al. [13] propose the use of modified normalization and summarization steps for better genotype calls with the method CRLMM. Another algorithm called Birdseed that is tailored for the Genome-Wide Human SNP Array 6.0 has also been introduced by Affymetrix (see their website for further details), although it is still recommended to use BRLMM for the 500K array set. More research and study is needed to evaluate how batch considerations affect the results of these alternative genotype calling algorithms. Alternatively, a promising focus is on better clustering methods adapted to allow for differences in genotype clusters in cases and controls as proposed by Plagnol et al. [15], as such methods could help alleviate bias that is introduced due to batch processing.

Differences in batch size and composition when using the BRLMM algorithm for genotype calling can drastically change the results of a GWA study and are a potential source for lack of reproducibility. Batches containing more individuals (a larger size) with cases and controls combined resulted in more conservative, concordant association testing results with the CAD case-control samples from the WTCCC, indicating there is a lower probability of making Type I Errors with such a batch schema. Alternatively, the exclusion of more loci from analysis due to QC could result in increases in Type II errors as a SNP that may have been associated is not even tested due to QC thresholds. The opposite is true for smaller batch sizes with cases and controls separated, where positive findings are more likely to be spurious associations (as evidenced by higher instances of discordant findings in our study). Imposing strict quality control measures can eliminate discordance in association results due to batch variation, but also eliminates a large portion of data that could have been informative. Thus there is a trade-off between high accuracy at the cost of losing potentially informative data and increased discovery of possible variants associated with disease that may contain a higher number of false positives.

Significance rules based on p-value cutoffs are not designed to be reproducible, but rather to control Type I and Type II errors, so some discordance evident in Figure 4.4 may be due simply to the fact that we are determining significance by a p-value-based criterion. A similar issue of discordance arose in the context of gene expression in the MAQC I project [19]. A primary conclusion of that work was that filtering on both p-value and fold change criterion can lead to more reproducible results. This suggests that adding a criterion such as the absolute difference of the numerator of the association test statistic (e.g. the Cochran-Armitage [31] χ^2) to current p-value filtering rules could enhance concordance and/or reproducibility of genome-wide association testing results.

Decisions on batch size and composition should be based on the researcher's goals, discovery or accurate reproducibility, as well as careful attention to DNA collection and preparation. It is recommended to use a combined batch composition in studies where cases and controls are prepared by similar labs and randomly assigned to plates for processing to avoid differential bias due to varying patterns of missing data in cases and controls (although a test for this as a quality control step alleviates some of the bias) or to employ methods such as those developed by Plagnol et al. [15]. In all follow-up, collaborative, and replication studies it is paramount to use a common genotype calling algorithm and batch schema to have comparable results. Another possible approach would be to analyze the allele probe intensities directly to avoid genotype calling errors. Future research into the behavior of batch effects in other popular calling algorithms such as Birdseed, CRLMM, and CHIAMO is necessary for researchers to make an informed choice for genotyping platform and genotype calling algorithm and is currently under study by the MAQC GWAWG.

4.4 Materials and Methods

4.4.1 WTCCC Data

The raw data was obtained from the Wellcome Trust Case Control Consortium, an organization of several research groups aimed at better understanding genetic variants and complex disease heritability through GWA studies. The CEL files of the 1500 individuals from the UK Blood Service Control Group and the 1991 CEL files of individuals that are cases for coronary artery disease (CAD) on the Affymetrix GeneChip Human Mapping 500K array were used to form batches for processing in the BRLMM genotype calling algorithm. The Affymetrix 500K array set consists of two arrays, Nsp and Sty, with approximately 262,000 and 238,000 SNPs on each, respectively.

4.4.2 BRLMM Genotype Calling

Algorithm

BRLMM is a multi-chip genotype calling algorithm that estimates cluster centers and variances for each genotype A/A, A/B, and B/B. The genotype call and associated confidence score is then decided using Mahalanobis distance to the cluster centers. The confidence score for the genotype call is derived by the ratio of d_1/d_2 where d_1 is the smallest distance to a genotype cluster (resulting call) and d_2 is the distance to the next closest cluster [11]. In this work, we refer to the confidence of the call as $1 - d_1/d_2$ so that higher confidences indicate that the genotype call is more likely to be accurate. If this confidence falls below a certain threshold, the genotype call is set to missing [11]. The 'B' in BRLMM is a Bayesian step that employs the use of DM [9], a single-chip algorithm that calls genotypes one SNP at a time on a sample of SNPs to estimate priors for the cluster centers and variances. The BRLMM algorithm has been shown to produce higher call rates and improve accuracy in comparison to DM, and has now become the recommended algorithm by Affymetrix for their 100K and 500K array sets.

Implementation

BRLMM calling was performed using the Affymetrix Power Tools (APT) version 1.10.2 that is available for download from http://www.affymetrix.com/. BRLMM is implemented in the 'apt-probeset-genotype' application of the package and all default parameters, except those to control batch size to match our experimental design, were used (which included the use of 0.5 as the confidence threshold for reporting a SNP call as missing). Chip description files (CDF) for Nsp and Sty arrays were used for genotype calling and were downloaded from the Affymetrix website as well. Genotype calls were done for the Nsp and Sty arrays separately and tab-delimited text files of calls were outputted for each array and batch.

4.4.3 Batch Effect Schema

Several runs of the BRLMM algorithm were performed under the experiment designed to estimate batch effects due to size and composition differences (i.e. the number of CEL files processed simultaneously and the case-control status of the samples processed within a batch). Three levels of batch size were used: 500, 2000, and 3500 samples (approximately) in each batch. Two levels of batch composition were employed: Separated (S) with cases and controls in different batches, and Combined (C) with a 1.25:1 ratio of cases to controls were randomly assigned to batches. For the final analysis, five different datasets of genotype calls for the 3491 samples were used and notated as follows: **C500** contained genotypes called from BRLMM in 7 batches consisting of 285 cases and 215 controls for a combined batch composition of size 500, **C2000** genotypes were called in two batches of 995 cases + 750 controls, **C3500** was created with one single batch of the 3491 samples,

S500 used 4 batches of 500 cases and 3 batches of 500 controls, and **S2000** consisted of one batch of 1991 cases and one batch of 1500 controls. Note that due to the unequal number of 1991 cases, the batches may have contained one more or less samples for that scenario. This design allows for the quantification of the effect of both batch size and composition as well as the interaction of the two effects. Due to the fact that our design is unbalanced with the inclusion of C3500, this dataset is not used in the formal statistical models we form to test significant changes in QC exclusion and significant test results; although it is utilized to elucidate trends in results due to increasing batch sizes.

4.4.4 Quality Control and Association Analysis Methods

Each text file produced by BRLMM was imported into JMP Genomics Statistical Discovery Software from SAS Institute for analysis. The genotype data for all batches over the Nsp and Sty arrays were merged and formatted into a wide dataset with rows corresponding to individuals and columns as markers. SNP annotation data was downloaded from the Affymetrix website. For each dataset the following QC steps were carried out. To eliminate low quality chips, individuals with a call rate less than 97% were excluded, further individuals were dropped if their average heterozygosity was below 25% or exceeding 30% (empirical threshold used in the original data analysis by the WTCCC [18]). After low quality individuals were dropped, SNP quality control consisted of three steps. First markers with either a minor allele frequency (MAF) less than 1% or with a call rate of less than 95% were excluded. Remaining markers were then filtered using a χ^2 one degree of freedom trend test of significant differences in the proportion of missing data between cases and controls. Finally SNPs that failed the test for Hardy-Weinberg Equilibrium (HWE) in the set of controls were also eliminated from further analysis. Significant results for the test of a trend in missing data between cases and controls and the HWE test were determined using $\alpha = 5.7 \times 10^{-7}$, an empirical threshold used by the original WTCCC study. See Figure 4.1 for all counts of individuals and loci excluded for C500, C2000, C3500, S500, and S2000 data.

The SNPs that passed QC were tested for significant associations with disease status, coronary artery disease, by the Cochran-Armitage trend test [31] for additive allele effects for each of the five datasets. SNPs were classified as significant if the p-value from the one degree of freedom χ^2 test was less than 5.0×10^{-7} (a commonly used threshold for uncorrected p-values [18, 68])and differential SNPs for batch sets were compared. To study the effect of stringent quality control measures, a second statistical association analysis was performed for SNPs that passed all previous QC steps as well as had a call rate no less than 99%. The set of SNPs that were significant at $\alpha = 5.0 \times 10^{-7}$ were again evaluated for differences among the five batch sets.

4.4.5 Generalized Linear Mixed Models to Test Batch Effects

Generalized Linear Mixed Models (GLMMs) were performed using the NLMIXED procedure in SAS/STAT version 9.2 in order to estimate subject-specific batch size composition and size×composition effects for modeling the probability of a SNP passing quality control and for modeling the probability of a SNP being deemed significant. The model is given as

$$\log\left(\frac{Pr(Y_i = 1|\boldsymbol{x})}{Pr(Y_i = 0|\boldsymbol{x})}\right) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}\ , \tag{4.1}$$

where $\boldsymbol{X}$ is the design matrix for fixed effects (batch size and composition) and $\boldsymbol{Z}$ is the design matrix for the random effect, SNP (to account for the 4 responses at each unique SNP that was genotyped in the 4 batch sets C500, S500, C2000, S2000). The SNP random effect is assumed to follow a $\gamma_i \sim Normal(0, \sigma_s^2)$ distribution, where σ_s^2 is the covariance within each unique SNP. A logit link function was used to model the probability that a marker passes QC dependent on the batch effects (Model I) where the response is coded as

$$Y_1 = \begin{cases} 0 & \text{for each SNP that passed QC steps} \\ 1 & \text{for each SNP that is excluded due to QC.} \end{cases}$$

Model II models the probability of a SNP being deemed significant given a batch size and composition using the response Y_2 coded as

$$Y_2 = \begin{cases} 0 & \text{for each SNP that passed QC but is not significant} \\ 1 & \text{for each SNP found signficant at p} < 1.0 \times 10^{-5}. \end{cases}$$

For Model II, SNPs that did not pass QC in a batch set were set to missing for Y_2 and if the SNP was excluded from all batch sets by QC it was not used in the model.

5

Conclusions and Avenues for Future Research

5.1 Bias Due to Population Structure

Population structure can result in bias in genetic association studies when differing allele frequency patterns exist at loci due to ancestry and there is disproportionate sampling of cases and controls across population groups. In Chapter 2 we show that SNPs with evidence of Hardy-Weinberg disequilibrium form a substructure-informative SNP set that better quantifies population structure than using randomly selected loci. HWD SNP selection is also shown to work well in avoiding the confounding influence that blocks of linkage disequilibrium may have on inferring structure among individuals. Visualization using nonmetric multidimensional scaling with HWD SNP selection provides a more faithful representation of population structure than PCA as evidenced by simulated applications as well as examples using HapMap data. PCA, which produces coordinates that maximize explained variance, is prone to the influence of outliers as well as producing projections that can inordinately separate the data; whereas NMDS finds coordinates that most closely represent the relationships between individuals in a full distance matrix (as measured using Identity-By-State distance in our studies). This important distinction also helps NMDS to avoid producing coordinates that over-estimate the amount of population differentiation that exists. NMDS is also shown to better quantify multiple levels of population structure that exist both between and within population groups.

In Chapter 3 we extend the use of NMDS coordinates to develop an association testing model that incorporates the coordinates as covariates to better control for population stratification in disease association studies. Nonparametric clustering was also employed to classify individuals to groups for better quantification and control of bias due to discrete population group membership. When loci exhibit more extreme population differentiation than evidenced by the coordinates produced from PCA or multidimensional scaling methods (CMDS and NMDS), the EIGENSTRAT method is unable to control Type I error at the nominal level; meanwhile the covariate models using CMDS and NMDS correct the bias appropriately. In a realistic simulation study, NMDS is shown to be significantly more powerful when multiple associated loci in the set of SNPs used to infer structure exist, providing evidence that NMDS coordinates are more robust to the contamination of markers that elucidate case-control structure as opposed to population structure. The use of nonparametric clustering in our NMDS method is also shown to be more sensitive to subtle structure in comparison with k-medoids clustering using the Gap statistic employed in the CMDS model.

The NMDS model we propose is a simple, easily applied model for testing for SNP association in the presence of population structure. The NMDS coordinates are shown to adjust for structure both between and within population groups. This provides promise for NMDS application when possible relatedness exists between individuals. Association studies, particularly those studying plants, that contain samples with complex pedigree structures as well as population structure (such as different species for crop studies) have favored the use of the '**Q+K**' mixed model approach proposed by Yu et al. [33] and modified by Zhao et al. [34]. Such a model, which uses coordinates from PCA as a fixed effect and an estimated kinship matrix to define a random effect, is favored in studies where multiple levels of relatedness exist and may influence disease behavior. Our research with NMDS suggests that substituting NMDS coordinates as a fixed effect in this mixed model may provide more accurate, less biased results and is an avenue for future research.

Another promising area of research is application of clustering tools and NMDS visualization to better quantify structure on the marker level (i.e. LD patterns across the genome). It is well known that patterns of linkage disequilibrium show high variation across the genome as well as across different populations of individuals. Understanding the structure that exists between markers and integrating that knowledge with association results can aid in the discovery of biological pathways and systems that explain the heritability of complex genetic diseases.

5.2 Bias Due to Genotype Calling

The analysis in Chapter 4 shows that another source of bias and unreliable results of association studies is batch processing differences in genotype calling algorithms. The novel approaches we employ to quantify batch effects in GWAS indicate that it is crucial to carefully consider batch size and composition when using the BRLMM algorithm for GWAS, particularly if the goals of the study include comparative analysis with previous findings for a certain disease. More research into differential errors produced by current genotype calling algorithms for the various SNP platforms is necessary to ensure high quality findings and reliable results. Research into alternative analysis techniques for GWAS such as tools to analyze the allele intensity signals directly (avoiding genotype calling algorithms) could potentially eliminate such errors. New technologies such as next-generation sequencing platforms may also help solve issues of genotype errors, although at this time such applications are not usually cost effective in comparison to current microarray technologies.

While bias in association studies is an ever-present concern, the research presented in this thesis contributes to better visualization and correction for bias due to population structure as well as addresses issues of technical bias due to genotype call algorithms. Such contributions help advance the use of GWAS for producing dependable, replicable results that aid in unraveling the complex nature of genetic diversity in organisms across the globe.

Bibliography

[1] Lynn B Jorde and Stephen P Wooding. Genetic variation, classification and 'race'. *Nature Genetics Supplement*, 36(11):S28–S33, 2004.

[2] Hua Tang, Tom Quertermous, Beatriz Rodriquez, Sharon L. R. Kardia, Xiaofeng Zhu, Andrew Brown, James S. Pankow, et al. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *The American Journal of Human Genetics*, 76:268–275, 2005.

[3] John Reynolds, B. S. Weir, and C. Clark Cockerham. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*, 105:767–779, 1983.

[4] J.S. Rogers. Measures of genetic similarity and genetic distance. *Studies in Genetics VII. University of Texas Publications 7213, Austin*, 1972.

[5] Kelci J. Miclaus, Russ Wolfinger, and Wendy Czika. Snp selection and multidimensional scaling to quantify population structure. *Genetic Epidemiology*, page In Press, 2009.

[6] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[7] Qizhai Li and Kai Yu. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genetic Epidemiology*, 32:215–226, 2008.

[8] G.C. Kennedy, H. Matsuzaki, S. Dong, W.M. Liu, J. Huang, G. Liu, X. Su, and Others. Large-scale genotyping of complex dna. *Nature Biotechnology*, 21(10):1233–1237, 2003.

[9] X. Di, H. Matsuzaki, T.A. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, and Others. Dynamic model based algorithms for screening and genotyping over 100k snps on oligonucleotide microarrrays. *Bioinformatics*, 21:1958–1963, 2005.

[10] Nusrat Rabbee and Terence P. Speed. A genotype calling algorithm for affymetrix snp arrays. *Bioinformatics*, 22(1):7–12, 2005.

[11] Affymetrix White Paper Publication. Brlmm: an improved genotype calling method for the genechip human mapping 500k array set. *http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf*.

[12] Benilton Carvalho, Henrik Bengtsson, Terence P. Speed, and Rafael A. Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. *Biostatistics*, 8(2):485–499, 2007.

[13] Shin Lin, Benilton Carvalho, David J. Cutler, Dan E. Arking, Aravinda Chakravarti, and Rafael A. Irizarry. Validation and extension of an empirical bayes method for snp calling on affymetrix microarrays. *Genome Biology*, 9:doi:10.1186/gb–2008–9–4–r63, 2008.

[14] David G. Clayton, Neil M. Walker, Deborah J. Smyth, Rebecca Pask, Jason D. Cooper, Lisa M. Maier, and Others. Population structure, differential bias and genomic control in large-scale, case-control association study. *Nature Genetics*, 37(11):1243 – 1246, 2008.

[15] Vincent Plagnol, Jason D. Cooper, John A. Todd, and David G. Clayton. A method to address differential bias in genotyping in large-scale association studies. *PLOS Genetics*, 3(5):759–767, 2007.

[16] Richard J.L. Anney, Elaine Kenny, Colm T. O'Dushlaine, Jessica Lasky-Su, Barbara Franke, Derek W. Morris, and Others. Non-random error in genotype calling procedures: Implications for family-based and case-control genome-wide association studies. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 147:1379–1386, 2008.

[17] Huixiao Hong, Zhenqiang Su, Weigong Ge, Leming Shi, Roger Perkins, Hong Fang, Joshua Xu, and Others. Assessing batch effect of genotype calling algorithm brlmm

for affymetrix genechip human mapping 500 k array set using 270 hapmap samples. *BMC Bioinformatics*, 9(9):S17doi:10.1186/1471–2105–9–S9–S17, 2008.

[18] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[19] MAQC Consortium. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006.

[20] Michael J. Bamshad, Stephen P. Wooding, W. Scott Watkins, Christopher T. Ostler, Mark A. Batzer, and Lynn B. Jorde. Human population genetic structure and inference of group membership. *American Journal of Human Genetics*, 72:578–589, 2003.

[21] S. Wright. *Evolution and the Genetics of Populations. Vol 2: The Theory of Gene Frequencies.* University of Chicago Press, 1969.

[22] Jonathan K. Pritchard, Matthew, and Stephens Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[23] Xiaoyi Gao and Joshua Starmer. Human population structure detection via multilocus genotype clustering. *BMC Genetics*, 8(34), 2007.

[24] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of Royal Statistical Society B*, 63: 411–423, 2001.

[25] Nick Patterson, Alkes L. Price, and David Reich. Population structure and eigenanalysis. *PLOS Genetics*, 2(12):2074–2093, 2006.

[26] Peristera Paschou, Elad Ziv, Estaband G. Burchard, Shweta Choudhry, William Rodriquez-Cintron, Michael W. Mahoney, and Petros Drineas. Pca-correlated snps for structure identification in worldwide human populations. *PLOS Genetics*, 3:1672–1686, 2007.

[27] B. Devlin, Kathryn Roeder, and Larry Wasserman. Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, 60:155–166, 2001.

[28] Jonathan K. Pritchard, Matthew Stephens, Noah A. Rosenberg, and Peter Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67:170–181, 2000.

[29] B. Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

[30] P. D. Sasieni. From genotypes to genes: Doubling the sample size. *Biometrics*, 53: 1253–1261, 1997.

[31] P Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3): 375–386, 1971.

[32] Bruce S. Weir, Lon r. Cardon, Amy D. Anderson, Dahlia M. Nielsen, and William G. Hill. Measures of human population structure show hetereogeneity among genomic regions. *Genome Research*, 15:1468–1476, 2005.

[33] Jianming Yu, Gael Pressoir, William H. Briggs, Irie Vroh Bi, Massanori Yamasaki, Joh F. Doebly, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38:203–208, 2006.

[34] Keyan Zhao, Maria Jose Aranzana, Sung Kim, et al. An aribidopsis example of association mapping in structured samples. *PLOS Genetics*, 3(1):71–82, 2007.

[35] Alkes L. Price, Johanna Butler, Nick Patterson, Cristian Capelli, Vencenzo L. Pascali, Frecesca Scarnicci, and Others. Discerning the ancestry of european americans in genetics association studies. *PLOS Genetics*, 4:e236. doi:10.1371/journal.pgen.0030236, 2008.

[36] Jennifer Wessel and Nicholas J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79:792–806, 2006.

[37] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27, 1971.

[38] Noah A. Rosenberg and Magnus Nordborg. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations

in discrete, admixed or spatially distributed populations. *Genetics*, 173(3):1665–1678, 2006.

[39] Taku Miyagawa, Nao Nishida, Jun Ohashi, Ryosuke Kimura, Akihiro Fujimoto, Minae Kawashima, and Others. Appropriate data cleaning methods for genome-wide association study. *Journal of Human Genetics*, 53:886–893, 2008.

[40] Trevor F. Cox and Michael A.A. Cox. *Multidimensional Scaling.* Chapman & Hall, 1994.

[41] The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

[42] C. C. Li. *First Course in Population Genetics.* Boxwood Press, Pacific Grove, CA, 1976.

[43] Raymond J. Peterson, David Goldman, and Jeffrey C. Long. Effects of worldwide population subdivision on aldh2 linkage disequilibrium. *Genome Research*, 9:844–852, 1999.

[44] Dahlia M. Nielsen, Meg G. Ehm, and Bruce S. Weir. Detecting marker-disease association by testing for hardy-weinberg disequilibrium at a marker locus. *American Journal of Human Genetics*, 63(5):1531–1540, 1998.

[45] Hong-Wen Deng, Wei-Min Chen, and Robert R. Recker. Population admixture: Detection by hardy-weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. *Genetics*, 157:885–897, 2001.

[46] Wen-Chung Lee. Detecting population stratification using a panel of single nucleotide polymorphisms. *International Journal of Epidemiology*, 32:1120, 2003.

[47] Bruce S. Weir. *Genetic Data Analysis II.* Sinauer Associates, Inc., 1996.

[48] J. C. Reif, A. E. Melchinger, and M. Frisch. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science*, 45(5):1–7, 2005.

[49] J. B. Kruskal and M. Wish. *Multidimensional Scaling.* Sage University Paper series on Quantitative Application in the Social Sciences, 07-011. Beverly Hills and London: Sage Publications, 1978.

[50] David J. Balding and Richard A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3–12, 1995.

[51] Jung-Ying Tzeng, Chi-Hao Wang, Jau-Tzuen Kao, and Chuhsing Kate Hsiao. Regression-based association analysis with clustered haplotypes through use of genotypes. *The American Journal of Human Genetics*, 78:231–242, 2006.

[52] Dimitri V. Zaykin and Lev A. Zhivotovsky. Ranks of genuine associations in whole-genome scans. *Genetics*, 171:1813–823, 2005.

[53] Jonathan Marchini, Lon R Cardon, Michale S. Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, 2004.

[54] Matthew L. Freedman, David Reich, Kathry L. Penney, Gavin J. McDonald, et al. Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, 36(4):388–393, 2004.

[55] Catarina D. Campbell, Elizabeth L. Ogburn, Kathryn L. Lunetta, Helen N. Lyon, et al. Demonstrating stratification in a european american population. *Nature Genetics*, 37 (8):868–872, 2005.

[56] Mark D. Shriver, Rui Mei, Esteban J. Parra, Vibhor Sonpar, Indrani Halder, et al. Large-scale snp analysis reveals clustered and continuous patterns of human genetic variation. *Human Genomics*, 2(2):81–89, 2005.

[57] Gad Kimmel, Michael I. Jordan, Eran Halperin, Ron Shamir, and Richard M. Karp. A randomization test for controlling population stratification in whole-genome association studies. *The American Journal of Human Genetics*, 81(5):895–905, 2007.

[58] Jonathan K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 69:124–137, 2001.

[59] Ivan P. Gorlov, Olga Y. Gorlova, Shamil R. Sunyaev, Margaret R. Spitz, and Christopher I Amos. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 82:100–112, 2008.

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

[61] Peter Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 456(11):728–731, 2008.

[62] B. W. Silverman. *Density Estimation.* New York: Chapman & Hall, 1986.

[63] David W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization.* New York: John Wiley & Sons, 1985.

[64] David Yeo. *Applied Clustering Techniques.* SAS Institute Inc, Cary NC, 2005.

[65] Prakash Gorrochurn, Susan E. Hodge, Gary Heiman, and David A. Greenberg. Effect of population stratification on case-control association studies. *Human Heritability*, 58: 40–48, 2004.

[66] S. L. Neuhausen. Founder populations and their uses for breast cancer genetics. *Breast Cancer Research*, 2(2):77–81, 2000.

[67] Efrosini Setakis, Heide Stirnadel, and David J. Balding. Logistic regression protects against population structure in genetic association studies. *Genome Research*, 16:290–296, 2006.

[68] Stephen F. Kingsmore, Ingrid E. Lindquist, Joann Mudge, Damian D. Gessler, and William D. Beavis. Genome-wide assocation studies: progress and potential for drug discovery and development. *Nature Reviews: Drug Discovery*, 7:221–230, 2008.

[69] J. Winkelmann, B. Schormair, P. Lichtner, S. Ripke, L. Xiong, S. Jalilizadeh, and Others. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nature Genetics*, 39(8):1000–6, 2007.

[70] Christa Meisinger, Hogler Prokisch, Christian Gieger, Nicol Soranzo, Divya Mehta, Dieter Rosskopf, and Others. A genome-wide association study identifies three loci

associated with mean platlet volume. *The American Journal of Human Genetics*, 84 (1):66–71, 2008.

[71] Bert Gold, Tomas Kirchhoff, Stefan Stefanov, James Lautenberger, Agnex Viale, Judy Garber, and Others. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *PNAS*, 105(11):4340–4345, 2008.

[72] Andre Franke, Tobias Balschun, Tom H. Karlsen, Jurgita Sventoraityte, Susanna Nikolaus, Gabriele Mayr, and Others. Sequence variants in il10, arpc2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nature Genetics*, 40(5):1319 – 1323, 2007.

Appendices

A

Chapter 2 Supplementary Details

Here we provide details for Chapter 2: ‘SNP Selection and Multidimensional Scaling to Quantify Population Structure’ which correspond to the Appendix A and the Online Supplemental Material provided for the corresponding article published in *Genetic Epidemiology*

A.1 MDS Algorithm

There are many MDS algorithms for optimization of a stress function. We use Stress Formula 1 and associated algorithm by Kruskal and Wish [49] implemented in SAS® Software via PROC MDS. The distinction of Metric MDS versus Non-metric MDS is the transformation of the original dissimilarities. Metric MDS restricts the transformation to be reflective of the scale of the metric (i.e. linear or absolute) while NMDS finds an optimal monotone transformation; which allows a wider choice of dissimilarity metrics.

We employ a two-part alternating iterative algorithm to minimize the Stress function over both $T(D_{rs})$ and $\hat{D}_{rs}$ for all objects $r = 1, ..., N$ and $s = 1, ..., N$. The first component uses isotonic (monotonic) regression (Cox and Cox [40], pp. 47-51) with the monotonicity constraint of preserving the rank order of the original dissimilarities D_{rs}:

$$D_{rs} < D_{tu} \Rightarrow T(D_{rs}) \leq T(D_{tu})$$

in order to obtain an optimal $T(\cdot)$ for the model

$$T(D_{rs}) = \hat{D}_{rs} + \epsilon$$

where the Euclidean reduced dimension coordinates $\hat{D}_{rs}$ are initialized by a weighted average of the data matrix. The second component minimizes the Stress1 function (Equation 2.9) by the Gauss-Newton algorithm using the gradient of Stress1 with respect to $\hat{D}_{rs}$; which has been shown to exist and is continuous everywhere (Cox and Cox [40], pp. 58-60).

Other alternative (to the gradient method) algorithms include ASCAL (Alternative Least squares SCALing) and SMACOF (Scaling by Majorizing a COmplicated Function) (Cox and Cox [40], pp. 152-161) which may give slightly different coordination results due to different Stress functions and methods.

A.2 Supplemental Material

A.2.1 Rogers' Distance Manipulation

Recall the formula for computing Rogers' distance given by Equation 2.3.

$$D_R = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{2}\sum_{j=1}^{a_l}(p_{lj} - q_{lj})^2} \tag{A.1}$$

This equation can be simplified when applied to a bi-allelic locus with $a_l = 2$, using the fact that $p_{l2} = 1 - p_{l1}$ and $q_{l2} = 1 - q_{l1}$.

$$\begin{aligned} D_R &= \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{2}\sum_{j=1}^{2}(p_{lj} - q_{lj})^2} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{2}\left[(p_{l1} - q_{l1})^2 + (p_{l2} - q_{l2})^2\right]} \\ &= \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{2}\left[(p_{l1} - q_{l1})^2 + (1 - p_{l1} - 1 + q_{l1})^2\right]} \\ &= \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{2}\left[(p_{l1} - q_{l1})^2 + (q_{l1} - p_{l1})^2\right]} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{(p_{l1} - q_{l1})^2} \end{aligned}$$

Moreover, in the case of computing Roger's distance between individuals, p_{i1} and q_{i1} can be further simplified to $p_{l1} = x_l/2$ and $q_{l1} = y_l/2$, where x_l and y_l are once again the copies of allele ‘a’ that Individual 1 and Individual 2 have respectively. This is true because generally (at the l^{th} locus) $p_{l1} = n_{l1}/2n$, where n is the number of individuals in the population and n_{l1} is the number of copies of allele ‘a’ in said population. For a population with $n = 1$ individual, p_{l1} will take values of 0/2 , 1/2 , 2/2 corresponding to the number of copies of allele ‘a’ the individual has out of the total number possible at the l^{th} locus. With the transformation of the allele frequencies, it is clear that Roger’s distance is equivalent to IBS distance:

$$\begin{aligned} D_R &= \frac{1}{L}\sum_{l=1}^{L}\sqrt{(p_{l1}-q_{l1})^2} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\left(\frac{x_l}{2}-\frac{y_l}{2}\right)^2} \\ D_R &= \frac{1}{2L}\sum_{l=1}^{L}|x_l - y_l| = D_{IBS} \end{aligned} \qquad \text{(A.2)}$$

A.2.2 Plots of Fit Statistics for NMDS

The Badness-of-Fit (BOF) criterion as well as the distance correlation measure are useful statistics for evaluating the dimension choice for NMDS. The BOF measure is the value of Stress1 at the completion of the optimization process, while distance correlation measures how correlated the observed distances are with the fitted Euclidean distances from the NMDS coordinates for a given dimension. Many multivariate texts attempt to create a rule to determine an adequate fit based on values of BOF, yet this is difficult to standardize for all data and visual interpretation similar to the scree plot (as in PCA) is often more appropriate. The fit plots for our simulated data under the four SNP set criteria are given by Supplementary Figure A.1. If the data (SNP set) captures existing structure in the data, the BOF will be low as structured data can be better explained in reduced dimension as opposed to unstructured data. Hence, to pick dimensionality, evaluating the change in BOF across dimensions is key. When a clear “elbow” in the BOF plot is not discernible, the distance correlation plot helps in optimal dimension choice and is conceptually more meaningful.

With our simulation, HWD SNP selection has the most clear choice of 2D visualization as evidenced by the top two plots in Supplementary Figure A.1; which is also an indication of the structure in the data captured by the HWD SNPs chosen. The rule of thumb proposed with EIGENSTRAT postulates that if there exist K discrete populations,

$K-1$ axes will be required to capture significant PS variability. This can also be said of NMDS for optimal visualization, but no longer be *required* as NMDS will still attempt to capture the structure with fewer than $K-1$ dimensions. This distinction marks a favorable property of NMDS in comparison with methods based on eigen-analysis as NMDS will still incorporate meaningful variability that may not be captured by the requested number of principal components; hence NMDS is well suited for visualization (which becomes difficult past two or three dimensions).

A.2.3 Confounding Effect of Systematic LD

In the creation of LD blocks in our simulation study, the assumption was made that the blocks are generated independently (emulating blocks across chromosomes where it is assumed there is no LD structure between SNPs from separate blocks). In other words, LD was simulated assuming there was not systematic pattern of LD across the SNP set. Alternatively if there exists LD (common in all subpopulations) across many markers contribute to the same pattern of genotype frequencies for individuals, this will not only confound the identification of population structure but impose a structure in the data itself based on grouping individuals with similar genotypes across the markers in LD.

An alternative simulation was designed where two LD blocks were simulated in reference to a single locus so that not only were SNPs within each LD block correlated, but also between the two LD blocks. This would be similar to the scenario of an unknown disease locus in LD with two separate regions of a gene with the same correlation pattern. Under this simulation (with 5,000 SNPs that are in systematically structured LD and 5,000 substructure-informative loci) we again sampled 200 markers for 1000 individuals randomly as well as 200 markers using HWD SNP selection with a p-value cutoff of 0.01. The resulting NMDS as well and PCA (EIGENSTRAT correction) plots are given in S Figure A.2 and clearly show that HWD SNP selection still correctly distinguishes the three subpopulations while randomly chosen SNPs display structure due to LD generation instead of population stratification. The principal components plots show similar results to NMDS and appear to better distinguish Subgroup 3 from the other subgroups on the 2^{nd} component.

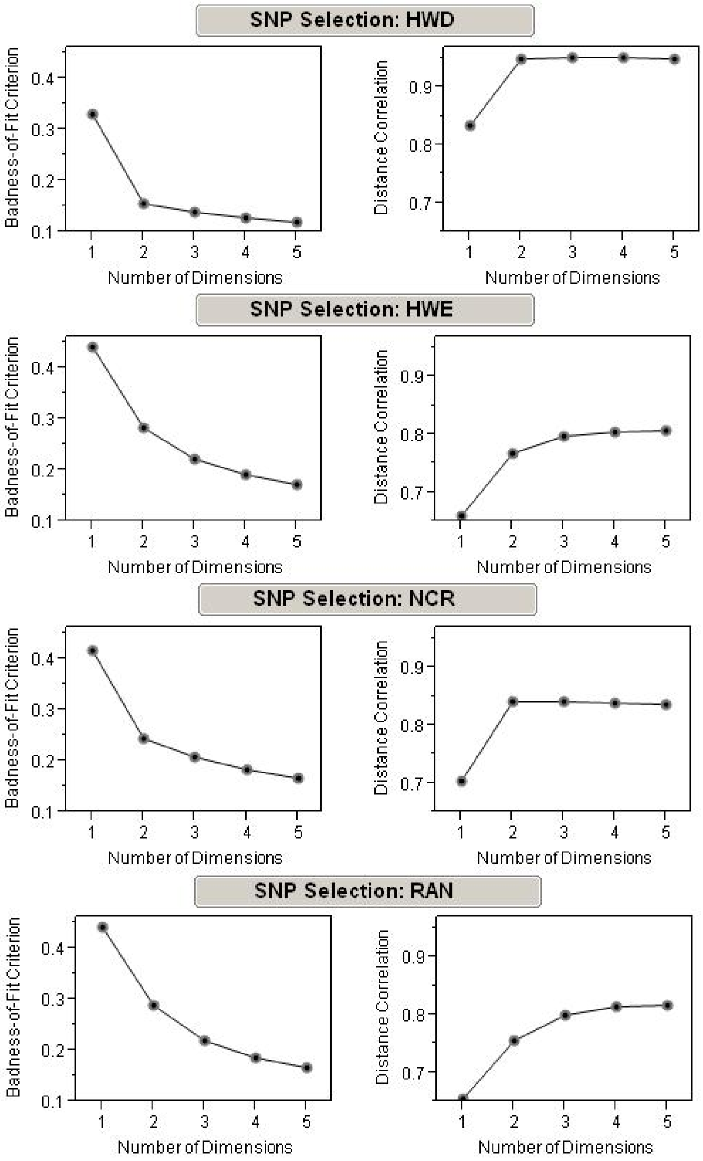

Figure A.1: Badness-of-Fit and distance correlation plots for calculations of NMDS coordinates of 1 randomly drawn simulated population consisting of 1000 individuals (500 cases, 500 controls) using 200 markers to infer substructure population by the four SNP selection criteria. Distance matrices were calculated using Gower's Distance and MDS coordinates were computed for each of the SNP selection methods (HWD SNPs, HWE SNPs, random SNPs (RAN), and random, non-confounding SNPS (NCR)).

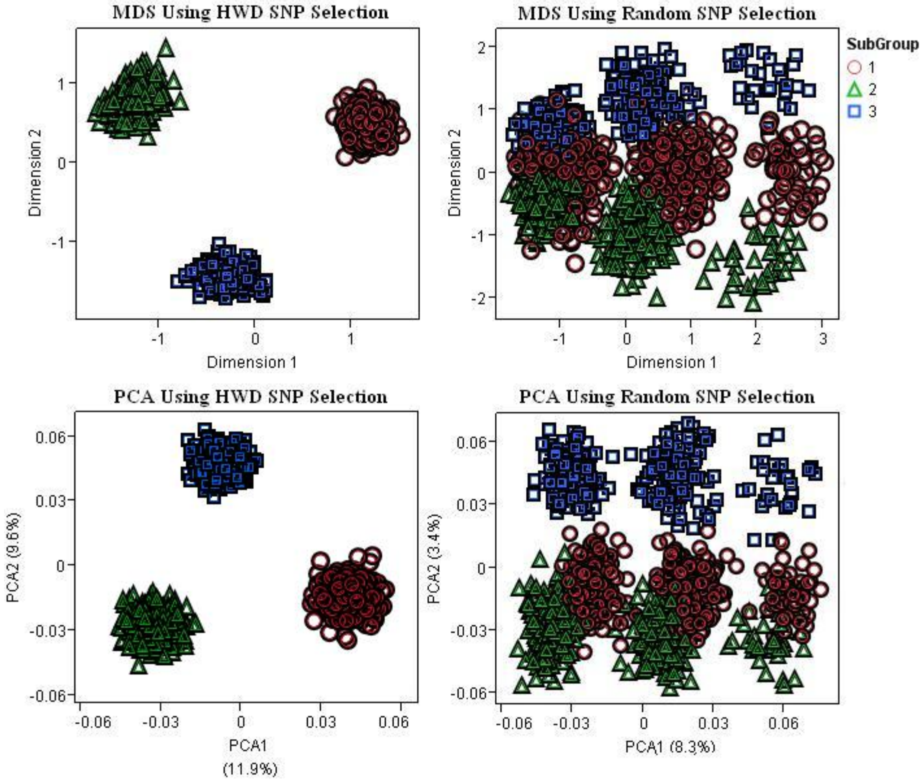

Figure A.2: NMDS (top) and PCA (bottom) Dimensions using 200 HWD ($p < 0.01$) chosen markers (left) and randomly chosen markers (right) for a simulated population of 1000 individuals (500 cases, 500 controls). Markers were chosen from a set of 10,000 generated SNPs, half of which were generated with two blocks of LD under a systematic pattern.

B

Chapter 3 Supplementary Tables

Table B.1: Type I Error estimates and associated standard errors for EIGENSTRAT, NMDS, and CMDS (corresponding to power estimates for risk model R=1.3). Values in this table correspond to plots C and D in Figure 3.2.

Disease SNP %	SNP Type	EIGENSTRAT	Classical MDS	Non-metric MDS
0%	Random Null	0.053 (0.007)	0.051 (0.007)	0.054 (0.007)
	Differentiated Null	0.127 (0.011)	0.033 (0.006)	0.037 (0.006)
1%	Random Null	0.049 (0.007)	0.053 (0.007)	0.051 (0.007)
	Differentiated Null	0.116 (0.01)	0.037 (0.006)	0.038 (0.006)
2%	Random Null	0.042 (0.006)	0.045 (0.007)	0.048 (0.007)
	Differentiated Null	0.106 (0.01)	0.038 (0.006)	0.037 (0.006)
3%	Random Null	0.046 (0.007)	0.045 (0.007)	0.047 (0.007)
	Differentiated Null	0.096 (0.009)	0.038 (0.006)	0.035 (0.006)
4%	Random Null	0.046 (0.007)	0.039 (0.006)	0.040 (0.006)
	Differentiated Null	0.093 (0.009)	0.040 (0.006)	0.039 (0.006)
5%	Random Null	0.047 (0.007)	0.044 (0.006)	0.049 (0.007)
	Differentiated Null	0.091 (0.009)	0.038 (0.006)	0.035 (0.006)
6%	Random Null	0.05 (0.007)	0.048 (0.007)	0.048 (0.007)
	Differentiated Null	0.091 (0.009)	0.038 (0.006)	0.039 (0.006)
7%	Random Null	0.055 (0.007)	0.057 (0.007)	0.057 (0.007)
	Differentiated Null	0.082 (0.009)	0.037 (0.006)	0.036 (0.006)
8%	Random Null	0.042 (0.006)	0.051 (0.007)	0.041 (0.006)
	Differentiated Null	0.073 (0.008)	0.041 (0.006)	0.043 (0.006)
9%	Random Null	0.062 (0.008)	0.058 (0.007)	0.053 (0.007)
	Differentiated Null	0.063 (0.008)	0.039 (0.006)	0.040 (0.006)
10%	Random Null	0.048 (0.007)	0.052 (0.007)	0.046 (0.007)
	Differentiated Null	0.052 (0.007)	0.040 (0.006)	0.042 (0.006)

Table B.2: Power estimates and associated standard errors for EIGENSTRAT, NMDS, and CMDS with relative risk, R=1.3. 'Power*' indicates the power estimates of associated SNPs that were also used in forming the coordinates in the respective dimension reduction techniques. 'Power' reflects the estimates for associated SNPs not used in the NMDS, CMDS, and PCA coordinate estimation. Values in this table correspond to plots A and B in Figure 3.2.

Disease SNP %	SNP Type	EIGENSTRAT	Classical MDS	Non-metric MDS
0%	Power	0.656 (0.015)	0.644 (0.015)	0.648 (0.015)
1%	Power	0.650 (0.015)	0.654 (0.015)	0.652 (0.015)
	Power*	0.637 (0.015)	0.640 (0.015)	0.646 (0.015)
2%	Power	0.638 (0.015)	0.640 (0.015)	0.642 (0.015)
	Power*	0.651 (0.015)	0.639 (0.015)	0.636 (0.015)
3%	Power	0.641 (0.015)	0.633 (0.015)	0.639 (0.015)
	Power*	0.633 (0.015)	0.622 (0.015)	0.630 (0.015)
4%	Power	0.647 (0.015)	0.635 (0.015)	0.647 (0.015)
	Power*	0.629 (0.015)	0.625 (0.015)	0.629 (0.015)
5%	Power	0.648 (0.015)	0.649 (0.015)	0.651 (0.015)
	Power*	0.623 (0.015)	0.621 (0.015)	0.640 (0.015)
6%	Power	0.613 (0.015)	0.613 (0.015)	0.638 (0.015)
	Power*	0.561 (0.016)	0.581 (0.016)	0.607 (0.015)
7%	Power	0.532 (0.016)	0.554 (0.016)	0.604 (0.015)
	Power*	0.489 (0.016)	0.538 (0.016)	0.602 (0.015)
8%	Power	0.499 (0.016)	0.475 (0.016)	0.566 (0.016)
	Power*	0.418 (0.016)	0.420 (0.016)	0.496 (0.016)
9%	Power	0.425 (0.016)	0.416 (0.016)	0.498 (0.016)
	Power*	0.326 (0.015)	0.317 (0.015)	0.434 (0.016)
10%	Power	0.402 (0.016)	0.370 (0.015)	0.497 (0.016)
	Power*	0.290 (0.014)	0.272 (0.014)	0.401 (0.016)

Table B.3: Type I Error estimates and associated standard errors for EIGENSTRAT, NMDS, and CMDS (corresponding to power estimates for risk model R=1.5). Values in this table correspond to plots C and D in Figure 3.3.

Disease SNP %	SNP Type	EIGENSTRAT	Classical MDS	Non-metric MDS
0%	Random Null	0.041 (0.006)	0.038 (0.006)	0.040 (0.006)
	Differentiated Null	0.139 (0.011)	0.039 (0.006)	0.037 (0.006)
1%	Random Null	0.046 (0.007)	0.046 (0.007)	0.048 (0.007)
	Differentiated Null	0.103 (0.01)	0.037 (0.006)	0.040 (0.006)
2%	Random Null	0.055 (0.007)	0.050 (0.007)	0.047 (0.007)
	Differentiated Null	0.087 (0.009)	0.037 (0.006)	0.037 (0.006)
3%	Random Null	0.043 (0.006)	0.051 (0.007)	0.046 (0.007)
	Differentiated Null	0.100 (0.009)	0.045 (0.007)	0.043 (0.006)
4%	Random Null	0.049 (0.007)	0.049 (0.007)	0.043 (0.006)
	Differentiated Null	0.099 (0.009)	0.039 (0.006)	0.036 (0.006)
5%	Random Null	0.046 (0.007)	0.054 (0.007)	0.050 (0.007)
	Differentiated Null	0.059 (0.007)	0.043 (0.006)	0.032 (0.006)

Table B.4: Power estimates and associated standard errors for EIGENSTRAT, NMDS, and CMDS with relative risk, R=1.5. 'Power*' indicates the power estimates of associated SNPs that were also used in forming the coordinates in the respective dimension reduction techniques. 'Power' reflects the estimates for associated SNPs not used in the NMDS, CMDS, and PCA coordinate estimation. Values in this table correspond to plots A and B in Figure 3.3.

Disease SNP %	SNP Type	EIGENSTRAT	Classical MDS	Non-metric MDS
0%	Power	0.936 (0.008)	0.934 (0.008)	0.937 (0.008)
1%	Power	0.936 (0.008)	0.940 (0.008)	0.943 (0.007)
	Power*	0.917 (0.009)	0.917 (0.009)	0.922 (0.008)
2%	Power	0.933 (0.008)	0.933 (0.008)	0.935 (0.008)
	Power*	0.897 (0.01)	0.903 (0.009)	0.913 (0.009)
3%	Power	0.869 (0.011)	0.879 (0.01)	0.916 (0.009)
	Power*	0.793 (0.013)	0.818 (0.012)	0.880 (0.01)
4%	Power	0.735 (0.014)	0.698 (0.015)	0.812 (0.012)
	Power*	0.590 (0.016)	0.576 (0.016)	0.757 (0.014)
5%	Power	0.614 (0.015)	0.509 (0.016)	0.710 (0.014)
	Power*	0.424 (0.016)	0.371 (0.015)	0.613 (0.015)

Table B.5: Power and Type I error estimates (and associated standard errors) for the SNP*Cluster interaction test under various sample sizes. Values correspond to Figure 3.6.

Number of Samples	SNP Type	Estimate (R=1.3)	Estimate (R=1.5)
500	Type I Error	0.048 (0.007)	0.053 (0.007)
	Power	0.122 (0.010)	0.254 (0.014)
1000	Type I Error	0.039 (0.006)	0.041 (0.006)
	Power	0.208 (0.013)	0.404 (0.016)
2000	Type I Error	0.065 (0.008)	0.058 (0.007)
	Power	0.367 (0.015)	0.679 (0.015)
3000	Type I Error	0.066 (0.008)	0.058 (0.007)
	Power	0.526 (0.016)	0.848 (0.011)
4000	Type I Error	0.054 (0.007)	0.06 (0.008)
	Power	0.657 (0.015)	0.927 (0.008)