Inferring, Estimating, and Accounting for Population and Pedigree Structure in Genetic Analyses

Matthew P Conomos

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Timothy Thornton, Chair

Bruce Weir, Chair

Sharon Browning

Program Authorized to Offer Degree:
Public Health: Biostatistics

University of Washington

Abstract

Inferring, Estimating, and Accounting for Population and Pedigree Structure in Genetic Analyses

Matthew P Conomos

Co-Chairs of the Supervisory Committee:
Timothy Thornton
Biostatistics

Bruce Weir
Biostatistics

Genetic studies of admixed individuals with ancestry derived from multiple previously isolated populations have become more common in recent years. Statistical methodology for analyzing genetic data in the presence of complex population structure and relatedness is currently of need. In this dissertation, we thoroughly investigate the performance of existing analysis methods and explore improvements where limitations exist. Methods for accurate ancestry inference and relatedness estimation when both structures are simultaneously present are developed and then utilized to achieve improved performance of genome-wide genetic association testing for complex quantitative traits. The performance of novel methodology is assessed through the use of extensive simulation studies as well as in applications to real data sets from admixed populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Timothy Thornton and Dr. Bruce Weir for all of their support and guidance in the process of conducting this research and writing this dissertation. I would also like to acknowledge all of my fellow students whom I have worked with throughout my graduate studies; their encouragement and collaboration greatly helped me to achieve this goal.

DEDICATION

In dedication to my father, Phillip Conomos,

for all of the love and support,

and for always pushing me to succeed.

Chapter 1

INTRODUCTION

As the cost of genotyping has substantially decreased in recent years, samples for the identification of genetic variants that influence complex traits and disease susceptibility have grown in both size and diversity. It is now common for large-scale genetic studies to include tens of thousands of individuals. Relatedness among sample individuals is common due to the inclusion of large pedigrees of known structure or due to cryptic relatedness between individuals assumed to be unrelated but actually sharing a common ancestor. Ancestral diversity of samples has also increased as populations from around the world have become the subject of genetic studies, with the hope of replicating previous associations and identifying novel variants that underlie differences in phenotypes across populations. A particularly challenging area of interest is that of analyzing genetic data from admixed populations, defined as populations with ancestry derived from two or more progenitor groups that were previously reproductively isolated. Admixture events have occurred between human populations due to historical events such as colonization and slave trade, and the two largest minority groups in the United States, African Americans and Hispanic Americans, are known to have admixed ancestry. Furthermore, the high connectivity of the globe significantly contributes to the admixing of populations today.

Statistical methods for inferring and estimating complex genetic structure from genome-screen data often make simplifying assumptions such as independence and homogeneity among sample individuals. These assumptions, however, often do not hold. Accurate inference and estimation of the genetic structure of sample individuals in the presence of both population and pedigree structure, including admixture and

cryptic relatedness, is a difficult problem that has yet to be adequately addressed. Improved statistical methodology for this area is of high interest, as both of these structures play important roles in many areas of genetic research. Accurate ancestry inference is essential to population genetics and can provide interesting insight into historical population events such as migration and colonization. Reliable estimation of familial relatedness enables confirmation of reported pedigree relationships and error detection of misspecified relationships, as well as provides valuable information for forensic genetics.

Accurate modeling of population and pedigree structure also provides utility for genetic association testing. At the start of the large-scale genome-wide association study (GWAS) era, the focus was primarily on identifying genetic variants associated with complex traits in samples of unrelated individuals with predominantly European ancestry. Simple regression models are suitable for the analysis of genome-wide data to identify associated markers in this setting with independent samples from a homogeneous population. However, in the presence of population structure and familial relatedness, more complex analysis methods are required. Appropriately and efficiently accounting for the genetic structure of these components is necessary to prevent spurious association and can increase statistical power to detect true associations. An improved understanding of population and pedigree structure and their roles in phenotypic variability could potentially help facilitate improved treatment on a personalized level in the medical and clinical setting.

The specific aims for this dissertation are:

- Develop a method for population structure inference that is robust to the presence of either known or cryptic relatedness among sample individuals and does not require genotype data from additional reference samples of known ancestry.

- Develop a method for accurately estimating measures of genetic relatedness in admixed populations with unspecified structure without requiring external

reference population samples of known ancestry.

- Improve on existing linear mixed model methods for genetic association testing in samples with relatedness and population structure by utilizing our improved relatedness and ancestry estimation methods.

- Extend linear mixed model methods for genetic association testing to allow for shared environmental effects and investigate the impact on type I error and power.

1.1 Population Structure Inference

Several approaches have been proposed for the identification of genetic ancestry differences in samples where study participants are assumed to be unrelated, including principal components analysis (PCA) [40, 41], multi-dimensional scaling (MDS) [44], and model-based methods for proportional ancestry estimation [3, 43, 54]. Many genetic studies, however, include individuals with some degree of relatedness, and existing methods for inferring genetic ancestry fail in related samples [42, 57]. In Chapter 3 of this dissertation, we look in depth at the problem of ancestry and population structure inference in samples that contain familial relatives. We present a method, PC-AiR, for robust population structure inference in the presence of known or cryptic relatedness. PC-AiR utilizes genome-screen data and an efficient algorithm to identify a diverse subset of unrelated individuals that is representative of all ancestries in the sample. The PC-AiR method directly performs principal components analysis (PCA) on the identified ancestry representative subset and then predicts components of variation for all remaining individuals based on genetic similarities. In simulation studies and in applications to real data from Phase III of the HapMap Project [18], we demonstrate that PC-AiR provides a substantial improvement over existing approaches for population structure inference in related samples. We also demonstrate significant

efficiency gains, where a single axis of variation from PC-AiR provides better prediction of ancestry in a variety of structure settings than using ten (or more) components of variation from widely used PCA and MDS approaches. PC-AiR provides accurate population structure inference without making any assumptions about or putting any constraints on the sample population and pedigree structures, and it does not require any external information such as additional reference samples of known ancestry or the known genealogy of the sample individuals. The ancestry inference from PC-AiR can be used for many genetic applications including prediction of individual ancestry and population stratification correction in genetic association studies.

1.2 Relatedness Estimation

A number of approaches have been proposed for relatedness inference in samples from homogeneous populations, including maximum likelihood [36, 55] and method of moments estimators such as those implemented in the widely used PLINK software [44] and the standard genetic relationship matrix (GRM) [64]. Many genetic studies, however, include individuals with genetic ancestry differences, and estimators that assume population homogeneity can fail in the presence of population structure [31, 38, 56]. Recent attempts at relatedness estimation in structured populations have been made, but these methods have limitations. The KING-robust [31] estimator relies on allele sharing counts, rather than allele frequency estimates, and provides consistent estimates of kinship coefficients for pairs of individuals with the same ancestry, but systematically biased estimates for pairs with different ancestry. Two other recent methods, REAP [56] and RelateAdmix [37], have been proposed for estimation of relatedness in populations with admixed ancestry; however, these methods require (1) prior knowledge about the ancestries that are present in the sample, (2) appropriate external reference panels for each of the ancestral subpopulations, and (3) reliable estimates of subpopulation-specific allele frequencies and individual admixture proportions. In Chapter 4 we develop PC-Relate, a novel principal component analysis

(PCA) based method for genetic relatedness inference in samples with unspecified population structure. PC-Relate provides accurate estimates of kinship coefficients and identity by descent (IBD) sharing probabilities in structured samples without requiring additional reference population panels or prior specification of the number of ancestral subpopulations. In simulation studies with population structure, including admixture, we demonstrate that PC-Relate provides a substantial improvement over commonly used existing methods for relatedness inference. Finally, we further demonstrate the utility of PC-Relate in an application to the Hispanic cohort of the Women's Health Initiative study.

1.3 Accounting for Structure in Genetic Association Testing

Linear mixed models (LMMs) have become the prevalent analysis approach for genetic association studies of complex quantitative traits in population based samples with known or cryptic structure. In recent years, many LMM methods have been proposed that all fit similar models [21, 22, 27, 28, 53, 65, 68, 72]. These methods treat sample structure as a random effect, simultaneously accounting for both population stratification and relatedness by modeling the trait covariance structure with a genetic relationship matrix (GRM). The variation between methods mostly lies in computational approaches and approximations to make genome-wide analysis feasible, as well as which genetic markers are used to construct the GRM. It has previously been shown that these LMM methods control genomic inflation well at the median of all test statistics genome-wide, but they may not be well-calibrated at markers with atypical structure [42, 63]. In admixed populations, however, the difference of allele frequencies between the underlying ancestral populations varies greatly across the genome. In Chapter 6 we demonstrate through simulation that existing LMM methods provide inflated test statistics at markers with large allele frequency differences and deflated test statistics at markers with small allele frequency differences. To address this issue, we develop MMAAPS, a new LMM approach for genetic as-

sociation testing in admixed populations, where sample structure is partitioned into fixed and random effects. Population structure is accounted for by including ancestry representative principal components as fixed effect covariates in the mean model, and relatedness is accounted for by modeling the trait covariance structure with an ancestry-adjusted kinship matrix. MMAAPS provides well-calibrated test statistics at all markers, regardless of the differentiation of allele frequency between ancestral populations, without suffering a loss of statistical power. To provide an application to real data, we compare the performance of MMAAPS to existing LMM methods for hematology phenotypes measured on the Hispanic cohort of the Women's Health Initiative study, where we see that MMAAPS provides stronger statistical evidence of association at previously identified hits.

In Chapter 6 we also investigate the importance of accounting for shared environmental effects in LMMs for genetic association testing. Existing LMM methods for GWAS only account for phenotypic correlation of sample individuals due to shared polygenic effects. Shared environmental effects have not been studied in depth in the context of genetic association studies, but individuals who share a common environment, such as household or geographic region, can have increased correlation for certain trait values. MMAAPS allows for the inclusion of multiple random effects, enabling analysis that accounts for additional covariance structure of the phenotype. Through simulation studies, we demonstrate that accounting for shared environmental effects does not appear to be critical for maintaining nominal type I error rates, provided that a variance component for polygenic effects is included in the LMM; however, we also demonstrate that accurately modeling the environmental covariance structure provides more efficient tests and can drastically improve statistical power to detect genetic associations in certain settings.

Chapter 2

BACKGROUND

In this chapter, we define some basic genetic concepts including cryptic relatedness and ancestry admixture, and we introduce notation that will be used throughout this dissertation. We also present a general model for genetic population structure. This model makes weak assumptions that are satisfied by commonly used models of population structure, such as the Balding-Nichols model [4]. We also introduce the standard genetic relationship matrix (GRM) that is widely seen throughout the statistical genetics literature [15,23,29,64,65]. We present a commonly used estimator of the GRM here, and briefly discuss its properties, as it is essential to much of the work in the following chapters.

2.1 Data and Notation

We begin by introducing some basic concepts for human genetics, notation, and data structure. Let the set $\mathcal{N}$ be a sample of individuals to be genotyped in a genome-screen. Each of these individuals has 23 pairs of chromosomes, 22 of which are autosomes, and 1 of which is a pair of sex chromosomes. One set of an individual's chromosomes is inherited from the mother of the individual, and the other set is inherited from the father. The DNA along each of these chromosomes consists of nucleotides, totaling approximately 3 billion guanine-cytosine or adenine-thymine base pairs genome-wide. Of these 3 billion base pairs, over 99% are identical among all humans. Positions along the genome where the base pairs are not fixed and have some variation across individuals in a population are referred to as single nucleotide polymorphisms (SNPs), and a small proportion of these SNPs are expected to drive

phenotype variation in the population. Each of the two possible base pairs at a SNP are referred to as an allele, often denoted as the "A" and "a" alleles or the "A" and "B" alleles. When genome-screen data are obtained via high-density SNP arrays, genotype values are measured on the set of sample individuals at some pre-specified set of SNPs, $\mathcal{S}$. One of the two allele types at a SNP is designated to be the "reference" allele, and the random variable g_{is} is a count of the copies of the reference allele that individual $i \in \mathcal{N}$ has at SNP $s \in \mathcal{S}$. Each individual has two alleles at each locus, i.e. position along the genome, one from each chromosome, and thus g_{is} takes values of 0, 1, or 2. A diagram of this data structure is presented in Figure 2.1.

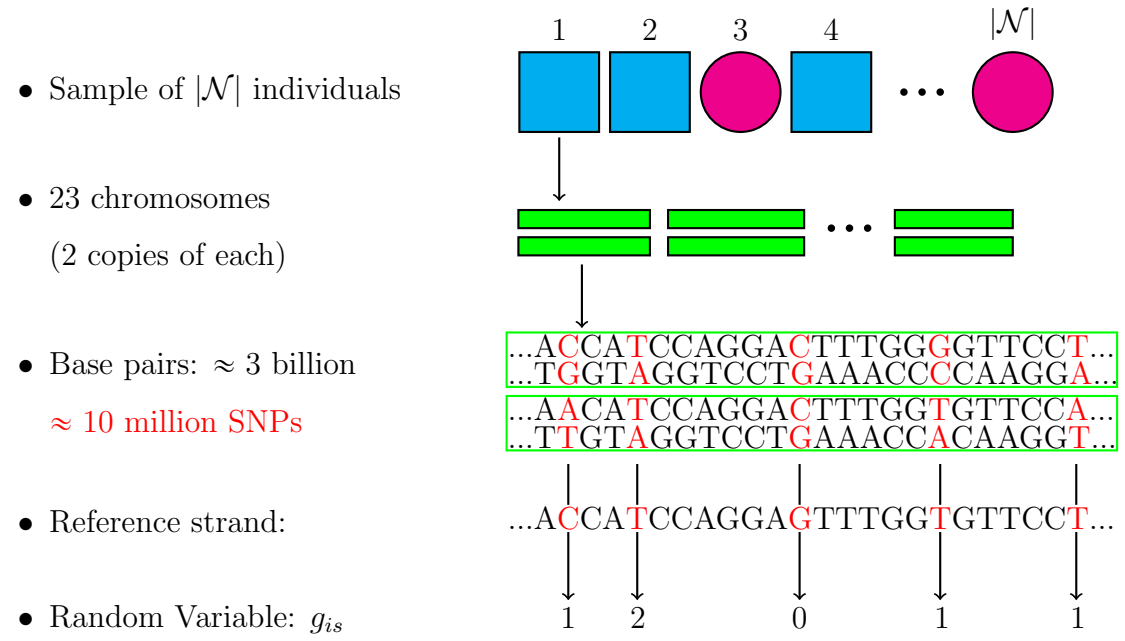


Figure 2.1: Basic Genetic Data Structure

Typical pedigree notation denotes each individual with a shape; squares for males and circles for females. An example of a short stretch of base pairs along a pair of chromosomes is shown, with SNPs highlighted in red. The reference strand for each chromosome is also shown, and g_{is} is constructed for this individual at these SNPs to be the number of copies of the reference allele in the top strand.

2.2 Measures of Genetic Relatedness

Commonly used genetic relatedness measures for a pair of individuals are often based on probabilities of sharing alleles that are identical-by-descent (IBD); i.e. copies of the same allele inherited from a recent common ancestor. For a pair of outbred individuals i and j, the set of IBD sharing probabilities $k_{ij}^{(0)}$, $k_{ij}^{(1)}$, and $k_{ij}^{(2)}$ are defined to be the proportion of loci for which i and j share 0, 1, or 2 alleles IBD, respectively. The kinship coefficient for the pair is defined to be the probability that a random allele selected from i and a random allele selected from j at a locus are IBD, and this quantity can be written as a function of the IBD sharing probabilities: $\phi_{ij} = \frac{1}{2}k_{ij}^{(2)} + \frac{1}{4}k_{ij}^{(1)}$.

In a population with inbreeding, there are no longer only three possible IBD states that a pair of individuals can have at a locus, but rather nine condensed IBD states as given by Jacquard [19]. We define $\Delta_{ij}^{(l)}$ to be the probability that individuals i and j are in Jacquard's l^{th} condensed IBD state at a locus, and the kinship coefficient can now be expressed as $\phi_{ij} = \Delta_{ij}^{(1)} + \frac{1}{2}\left(\Delta_{ij}^{(3)} + \Delta_{ij}^{(5)} + \Delta_{ij}^{(7)}\right) + \frac{1}{4}\Delta_{ij}^{(8)}$. Additionally, an individual may have two alleles at a locus that are IBD to each other in this setting, and the inbreeding coefficient, f_i, is defined to be the probability of this event for individual i. This quantity can also be expressed as the self-kinship coefficient, $\phi_{ii} = (1 + f_i)/2$.

2.3 Cryptic Relatedness

For family-based genetic studies, information on pedigree relationships among sample individuals is generally collected. However, genealogical information is often not available for population-based studies, and it is common for sample individuals who are assumed to be unrelated to actually share a common ancestor. Throughout this dissertation, we refer to these relationships that are unknown a priori as cryptic relatedness. Cryptic relatedness often arises between distant relatives, but may also

occur between close relatives that were not reported at the time of data collection. An example of cryptic relatedness among individuals that share a common ancestor a few generations back is illustrated in Figure 2.2.

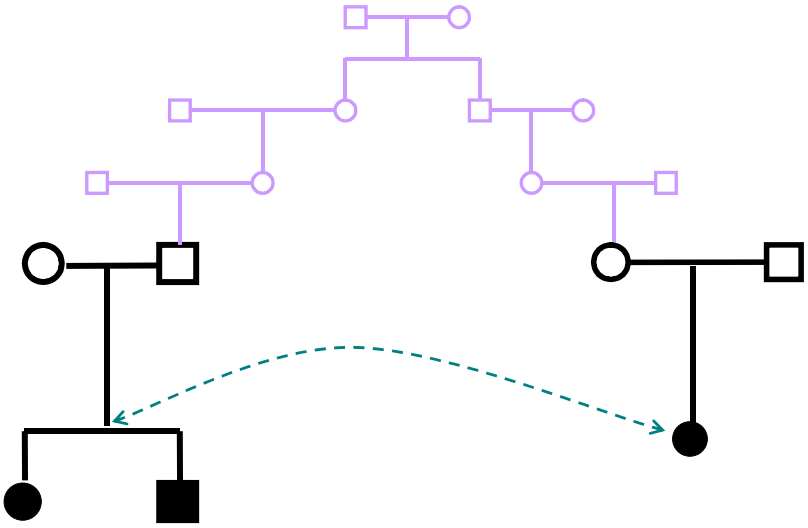


Figure 2.2: Cryptic Relatedness

A diagram of a pedigree; individuals further down the diagram are descendants of those individuals further up the diagram. The individuals in black are sampled with known relationships and believed to be two small nuclear families. The individuals in purple are unobserved, but they connect the sampled individuals into one large family. The individuals at the bottom of the pedigree joined by the green arrow are cryptic relatives that share a common ancestor 4 generations back.

2.4 Admixed Populations

An admixed population is a population with ancestry derived from two or more progenitor groups that were previously reproductively isolated, but have since interbred.

Admixture events in human populations have occurred throughout history due to events such as the colonization of the Americas and the trans-Atlantic slave trade. Admixed populations exist across the world, and the two largest minority populations in the United States, African Americans and Hispanic Americans, both have admixed ancestry.

Admixed populations present additional challenges in genetic analyses. Individuals within the same admixed population may have very different ancestry, both globally and locally. For example, our analysis of 86 individuals from the Mexican Americans in Los Angeles, California (MXL) population sample from release 3 of phase III of the International Haplotype Map Project (HapMap) [18] shows that global European ancestry ranges from 18.0% to 91.0%, and global Native American ancestry ranges from 4.2% to 80.4% in this sample of Mexican Americans (see Section 3.3.4 for further details). Additionally, admixed individuals who have similar global ancestry proportions, even close relatives, generally will have different ancestry locally at a particular position in the genome. For an illustration of this, see Figure 2.3.

2.5 Population Genetic Modeling Assumptions

Let the set $\mathcal{N}$ of sample individuals genotyped in the genome-screen belong to a structured population with ancestry derived from K subpopulations. We allow for these individuals to have admixed ancestry from the K subpopulations, and we denote $\mathbf{a}_i = (a_i^1, \ldots, a_i^K)^T$ to be the ancestry vector for individual $i \in \mathcal{N}$, where a_i^k is the proportion of ancestry across the autosomal chromosomes from subpopulation k for individual i, with $a_i^k \geqslant 0$ for all k, and $\sum_{k=1}^K a_i^k = 1$. Let $\mathcal{S}$ be the set of autosomal SNPs in the genome-screen, and for SNP $s \in \mathcal{S}$, denote $\mathbf{p}_s = (p_s^1, \ldots, p_s^K)^T$ to be the vector of subpopulation-specific allele frequencies, where p_s^k is the reference allele frequency at SNP s in subpopulation $k \in \{1, \ldots, K\}$. We assume that the p_s^k are random variables, independent across s but with possible dependence across the k's, with

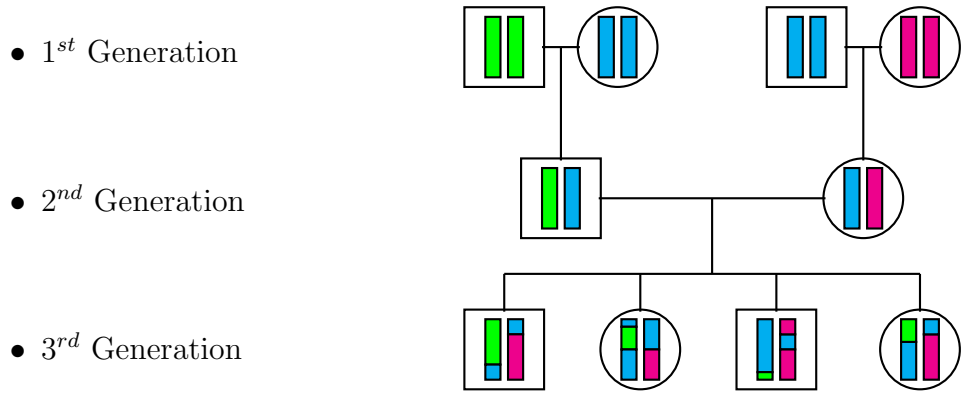


Figure 2.3: Ancestry Admixture in a Pedigree at One Chromosome

Each vertical bar within an individual represents one copy of a chromosome, and the color represents the ancestry at that position along the chromosome. The first admixture event occurs in the 2^{nd} generation, so whole copies of chromosomes within individuals have the same ancestry. In the 3^{rd} generation and beyond, recombination results in a mixture of segments of chromosomes with the same ancestry, and the length of these segments shrinks on average as more generations pass. As a consequence, while each of the siblings in the 3^{rd} generation has the same expected global ancestry proportions, the realization of global ancestry proportions as well as the ancestry at any one location may be different.

mean $\mathbb{E}[\mathbf{p}_s] = p_s\mathbf{1}$ and covariance $\text{Cov}[\mathbf{p}_s] = p_s(1-p_s)\mathbf{\Sigma}_K$ for every s, where $\mathbf{1}$ is a length K column vector of 1's, and $\mathbf{\Sigma}_K$ is a $K \times K$ matrix. In genetic models incorporating population structure, the allele frequency parameter p_s is typically interpreted as an "ancestral" reference allele frequency, or some average of reference allele frequencies across subpopulations. A schematic of this model is presented in Figure 2.4. Although we allow $\mathbf{\Sigma}_K$ to be completely general, including allowing for non-zero covariances across subpopulations, a special case is the Balding-Nichols model [4], where $\mathbf{\Sigma}_K$ is a diagonal matrix with (k,k)-th element equal to $F_k \geqslant 0$, and F_k is Wright's standardized measure of variation [62] for subpopulation k. In most contexts, the parameters K, $\mathbf{\Sigma}_K$, $\mathbf{p}_s$ and p_s for all $s \in \mathcal{S}$, and $\mathbf{a}_i$ for all $i \in \mathcal{N}$ are unknown. The goal of PC-AiR (Chapter 3) is to obtain inference on ancestry, i.e. the $\mathbf{a}_i$'s, for all sample individuals $i \in \mathcal{N}$ in the presence of known or cryptic relatedness. The goal of PC-Relate (Chapter 4) is to provide accurate relatedness estimation without making assumptions about any of these population structure parameters.

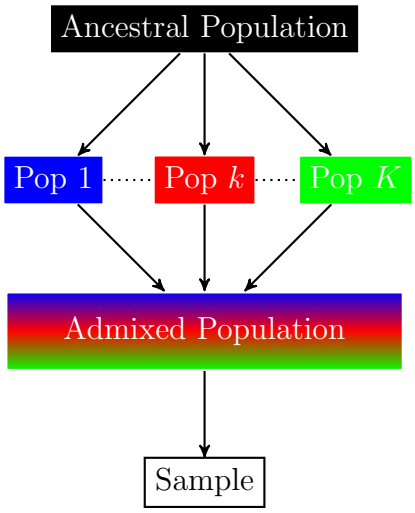
2.6 The Genetic Relationship Matrix (GRM)

The genetic relationship matrix (GRM) provides a measure of the genetic similarity between all pairs of individuals in the sample $\mathcal{N}$. For this sample of size $|\mathcal{N}|$, the empirical GRM is an $(|\mathcal{N}|\text{x}|\mathcal{N}|)$ symmetric matrix, $\widehat{\mathbf{\Psi}}$. The $[i,j]^{th}$ entry of $\widehat{\mathbf{\Psi}}$ is a scaled measure of the genetic covariance between individuals i and j, found under the assumption that the genetic correlation structure is the same across all SNPs using the formula

$$\widehat{\psi}_{ij} = \frac{1}{|\mathcal{S}_{ij}|} \sum_{s \in \mathcal{S}_{ij}} \frac{(g_{is} - 2\hat{p}_s)(g_{js} - 2\hat{p}_s)}{2\hat{p}_s(1-\hat{p}_s)}, \tag{2.1}$$

where $\mathcal{S}_{ij}$ is the subset of SNPs for which individuals i and j have non-missing genotype data, $|\mathcal{S}_{ij}|$ is the number of SNPs in this subset, and $\widehat{p}_s$ is an estimate of the reference allele frequency in the population at SNP s.

The GRM is used for many applications in statistical genetics including population



- p_s: ancestral allele freq. at SNP s
- $\mathbf{p}_s = (p_s^1, \ldots, p_s^K)^T$: subpopulation-specific allele frequencies at SNP s
- $\mathbb{E}[\mathbf{p}_s] = p_s\mathbf{1}$
- $\text{Cov}[\mathbf{p}_s] = p_s(1-p_s)\mathbf{\Sigma}_K$
- $\mathbf{a}_i = (a_i^1, \ldots, a_i^K)^T$: individual i admixture proportion from each pop
- $\sum_{k=1}^K a_i^k = 1$

Figure 2.4: Population Structure Model

A schematic of our basic population structure model. The ancestral population split many thousands of years ago into K discrete subpopulations. More recently, these previously isolated subpopulations came together reproductively to create an admixed population. Individuals in the admixed population may have highly varied ancestry proportions from each of the subpopulations. We consider sampling individuals from this admixed population.

structure inference, kinship coefficient estimation, heritability estimation, and adjustment for structure in genetic association studies. One of the most popular methods for population structure inference, principal components analysis (PCA), relies on an empirical GRM to measure ancestral similarity among sample individuals in order to find principal components (PCs) that reflect population structure. The elements of the GRM are also commonly used as estimates of pairwise kinship coefficients, as individuals that share recent common ancestors are more genetically similar. While the GRM can provide accurate population structure inference or accurate kinship coefficient estimation in certain situations, when both structures are simultaneously present, inference on either one will be confounded by the other.

To understand this confounding better, consider the properties of the empirical GRM in more detail. The GRM can be used for estimation of both population and pedigree structure because both manifest as increased genetic correlation. However, when the sample $\mathcal{N}$ comes from a structured population following the general population genetic modeling assumptions discussed in Section 2.5, it can be shown that under some weak convergence assumptions

$$\frac{1}{2}\widehat{\psi}_{ij} \rightarrow \phi_{ij}[1 - b_{\psi_1}(i,j)] + \mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_j, \tag{2.2}$$

where $b_{\psi_1}(i,j)$ is a function of the underlying population structure and the ancestry of individuals i and j. The details of this derivation are provided in Appendix A.2.1. This result shows that each entry of $\widehat{\boldsymbol{\Psi}}$ is a function of the ancestral similarity (population structure) and the kinship (pedigree structure) of that pair of individuals. When all of the sample individuals are unrelated, the contribution from kinship is 0, and the limiting values of the empirical GRM entries are $\mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_j$. In this setting, $\widehat{\boldsymbol{\Psi}}$ can be used for population structure inference with PCA, and the resulting PCs reflect ancestry. When all of the sample individuals have the same ancestry, i.e. are from a homogeneous population, the contribution from population structure is 0, and the limiting values of the empirical GRM entries are ϕ_{ij}. Therefore, in a homogeneous

population, the entries of $\widehat{\mathbf{\Psi}}$ can be used for kinship coefficient estimation.

Deconvoluting the population and pedigree structure when they are both present is challenging, and is a major focus of the work presented in this dissertation. The properties of this estimator are explored in depth in Chapters 3 and 4, where modifications are presented that provide accurate population structure inference and relatedness estimation, even when both structures are present and unknown a priori.

Chapter 3

ROBUST INFERENCE OF POPULATION STRUCTURE IN THE PRESENCE OF RELATEDNESS

3.1 Introduction

Ancestry inference with genetic data is an essential component for a variety of applications in genetic association studies, population genetics, and both personalized and medical genomics. Advances in high-throughput genotyping technology have allowed for an improved understanding of continental-level and fine-scale genetic structure of human populations, as well as other organisms. Principal components analysis (PCA) [40, 41] has been the prevailing approach in recent years for both population structure inference and correction of population stratification in genome-wide association studies (GWAS) with high-density single nucleotide polymorphism (SNP) genotyping data. Other widely used methods for inference on genetic ancestry include multi-dimensional scaling (MDS) [44], a dimension reduction method similar to PCA, and model-based methods, such as STRUCTURE [43], FRAPPE [54], and ADMIXTURE [3], for proportional ancestry estimation in samples from admixed populations.

Genetic studies often include related individuals; however, most existing population structure inference methods fail in the presence of relatedness. For example, the top principal components from PCA, as well as the top dimensions from MDS, can reflect family relatedness rather than population structure when applied to samples that include relatives [42]. Model-based ancestry estimation methods similarly fail in the presence of relatedness as they are not able to distinguish between ancestral groups and clusters of relatives [57]. For certain family-based study designs with known pedigrees, the population structure inference method proposed by Zhu

et al. [73], where SNP loadings calculated from a PCA on pedigree founders are used to obtain principal components values for genotyped offspring, can be used. However, this approach, which we refer to as "FamPCA," fails in the presence of cryptic or misspecified relatedness and is not applicable to most GWAS where genealogical information on sample individuals is often incomplete or unavailable. The FamPCA method requires genotype data to be available for pedigree founders, which can be prohibitive for many genetic studies. In addition, inference on population structure is limited to the ancestries in the subset of genotyped founders, which may lack sufficient diversity to be representative of the ancestries in the entire sample [6].

We address the problem of population structure inference and correction in samples with related individuals. We do not put constraints on how the individuals might be related, and we allow for the possibility that genealogical information on sample individuals could be partially or completely missing. We propose a method, which we call PC-AiR (principal components analysis in related samples), for inference on population structure from SNP genotype data in general samples with related individuals. The PC-AiR method implements a fast and efficient algorithm for the identification of a diverse subset of mutually unrelated individuals who are representative of the ancestries in the entire sample. Axes of variation are inferred using this ancestry representative subset, and coordinates along the axes are predicted for all remaining sample individuals based on genetic similarities with individuals in the ancestry representative subset. The top axes of variation (principal components) from PC-AiR are constructed to be both representative of ancestry and robust to both known or cryptic relatedness in the sample. A remarkable feature of PC-AiR is the method's ability to identify a diverse and representative subset of individuals for ancestry inference using only genome-screen data from the sample, without requiring additional samples from external reference population panels or genealogical information on the study individuals.

We assess the robustness and accuracy of PC-AiR for inference on genetic an-

cestry in simulation studies with both related and unrelated individuals under various types of population structure settings, including admixture. We also directly compare PC-AiR to existing population structure inference methods using both simulated data and real genotype data collected from the Mexican Americans in Los Angeles, California (MXL) and African American individuals in the southwestern USA (ASW) population samples of release 3 of phase III of the International Haplotype Map Project (HapMap) [18]. The population structure inference methods to which we compare PC-AiR are: (1) PCA with the EIGENSOFT [41] software, (2) MDS with the PLINK [44] software, (3) the model-based ancestry estimation methods FRAPPE [54] and ADMIXTURE [3], and (4) FamPCA [73] as implemented in the KING [31] software. Through these analyses we are able to demonstrate that PC-AiR provides more accurate and more efficient ancestry inference than existing methods.

3.2 Methods

3.2.1 Overview of the PC-AiR Method

Let the set $\mathcal{N}$ be a sample of individuals who have been genotyped in a genome-screen. An essential component of the PC-AiR method for population structure inference in the presence of relatedness is to use genome-screen data to partition $\mathcal{N}$ into two non-overlapping subsets, $\mathcal{U}$ and $\mathcal{R}$, i.e. $\mathcal{N} = \mathcal{U} \cup \mathcal{R}$ with $\mathcal{U} \cap \mathcal{R} = \varnothing$, where $\mathcal{U}$ is a subset of mutually unrelated individuals who are representative of the ancestries of all individuals in $\mathcal{N}$, and $\mathcal{R}$ is a "related subset" of individuals who have at least one relative in $\mathcal{U}$. We allow for individuals in $\mathcal{R}$ to be related to each other in addition to having relatives in $\mathcal{U}$. PC-AiR uses measures of pairwise relatedness and ancestry divergence calculated from autosomal SNP genotype data for the identification of $\mathcal{U}$, without requiring external reference panels or genealogical information. Population structure inference on the entire set of sample individuals, $\mathcal{N}$, is then obtained by first directly performing PCA on the selected ancestry representative subset, $\mathcal{U}$, and then

predicting values along the components of variation for all individuals in the related subset, $\mathcal{R}$, based on genetic similarities with the individuals in $\mathcal{U}$. In the following subsections, we describe the PC-AiR method in detail.

3.2.2 Relatedness Inference in Structured Populations

Estimating genetic relatedness in the presence of population structure is the subject of Chapter 4, but we provide an abreviated discussion here to facilitate understanding of the PC-AiR method. PC-AiR uses kinship coefficients to measure relatedness between all pairs of individuals in $\mathcal{N}$, where the kinship coefficient for individuals i and j, which we denote as ϕ_{ij}, is defined to be the probability that a random allele selected from i and a random allele selected from j at a locus are identical-by-descent (IBD). When the genealogy of the sample individuals is known, PC-AiR can use theoretical or pedigree-based kinship coefficients, and a number of software packages [1, 70] are available for calculating these according to a specified genealogy. However, genealogical information on sample individual is often unknown, incomplete, or misspecified, and PC-AiR can also use empirical kinship coefficients estimated from genome-screen data for samples with cryptic relatedness that must be genetically inferred. It is important to note that relatedness estimators that assume population homogeneity, such as those implemented in the widely used PLINK software [44] or obtained via a standard genetic relationship matrix (GRM) [64], are biased in samples from structured populations. Therefore, we do not recommended using these estimators with PC-AiR as it has been demonstrated that they give inflated kinship estimates in the presence of population structure [31, 56], where (1) unrelated pairs of individuals with similar ancestry can have kinship-coefficient estimates corresponding to values that are expected for close relatives, and (2) related individuals can have a systematic inflation in their estimated degree of relatedness.

To use the PC-AiR method when pedigree relationships are unknown or incomplete, we recommend using empirical kinship coefficient estimates from methods that

have been developed for samples from structured populations. One such estimator is KING (kinship-based inference for GWASs)-robust [31]. Rather than using estimated allele frequencies, which leads to biased relatedness estimates in the presence of population structure, KING-robust relies on shared genotype counts across the SNPs in the genome-screen to measure the genetic distance between individuals. KING-robust was developed for relatedness inference in samples from populations with discrete substructure without admixture, and it is a consistent estimator of the kinship coefficient for a pair of outbred individuals from the same subpopulation. The estimator, however, will generally be negatively biased for pairs of individuals that have different ancestries. Despite this bias, the KING-robust estimator is typically able to separate close relatives with similar ancestry from unrelated individuals, which is often sufficient for the PC-AiR method. Additionally, the PC-AiR method exploits the negative bias of the KING-robust estimator to gain insight on ancestry differences among individuals, as discussed in more detail in the following subsection.

Estimated kinship coefficients from the recently proposed REAP [56] and RelateAdmix [37] methods can also be used by PC-AiR. Both of these methods offer improved relatedness inference over KING-robust in samples with admixed ancestry by using external reference panels. REAP and RelateAdmix, however, may not be suitable for some studies as they require (1) some prior knowledge about the ancestries that are likely present in the sample, and (2) appropriate reference panels with suitable surrogates for the ancestral subpopulations. KING-robust does not require external reference panels and can be used with PC-AiR for admixed samples with cryptic relatedness when the REAP and RelateAdmix methods may not be practical.

3.2.3 Measuring Ancestry Divergence with Genome-Screen Data

Pairwise measures of relatedness, such as kinship coefficients, among individuals in a sample can be used for selecting a subset of mutually unrelated individuals [52]. In structured samples, however, identifying a subset of unrelated individuals based

solely on relatedness measures can result in a subset that lacks sufficient diversity for population structure inference on the entire sample, as it may not be representative of the ancestries of all individuals. For the identification of an ancestry representative subset of mutually unrelated individuals, PC-AiR incorporates measures of ancestry divergence in addition to the kinship coefficients used as measures of relatedness.

Consider a pair of individuals $i, j \in \mathcal{N}$ who have non-missing genotype data at the set $\mathcal{S}_{ij} \subset \mathcal{S}$ of autosomal SNPs in a genome-screen, and let $|\mathcal{S}_{ij}|$ denote the total number of SNPs in this set. Additionally, let the random variables g_{is} and g_{js} be the number of copies of the reference allele that individuals i and j each have, respectively, at SNP $s \in \mathcal{S}_{ij}$; thus, g_{is} and g_{js} take values of 0, 1, or 2. To measure ancestry divergence between a pair of unrelated individuals i and j, we use the estimator

$$\widehat{\kappa}_{ij} = \frac{1}{2}\left(1 - \frac{\sum_{s\in\mathcal{S}_{ij}}(g_{is} - g_{js})^2}{\sum_{s\in\mathcal{S}_{ij}}\left(\mathbb{1}_{[g_{is}=1]} + \mathbb{1}_{[g_{js}=1]}\right)}\right), \tag{3.1}$$

where $\mathbb{1}_{[g_{is}=1]}$ is an indicator for individual i being heterozygous at SNP s, i.e. $\mathbb{1}_{[g_{is}=1]}$ is 1 if $g_{is} = 1$ and is 0 otherwise, and $\mathbb{1}_{[g_{js}=1]}$ is similarly defined for individual j. Equation (3.1) is equivalent to the KING-robust estimator [31] that has been proposed for estimating kinship coefficients of related individuals in samples from discrete subpopulations. We consider the KING-robust estimator under the general population genetic modeling assumptions presented in Section 2.5. The limiting behavior of this estimator is derived in Appendix A.2.3 under the assumption that genotypes at different SNPs are independent and with $|\mathcal{S}_{ij}| \to \infty$. For unrelated individuals i and j with the same ancestry, $\widehat{\kappa}_{ij} \to 0$. However, when i and j have different ancestral backgrounds, $\widehat{\kappa}_{ij}$ tends to be a negatively biased estimator of kinship, and this bias provides a useful measure of the ancestry divergence between the pair of individuals. The magnitude of the negative bias depends on how divergent the ancestries for the pair are. The $\widehat{\kappa}_{ij}$ estimator will generally have more extreme negative values when (1) the F_k values are large, (2) i and j have large ancestry proportion differences, or (3) either i or j has an ancestry proportion that is close to 1 from one of the K

subpopulations. For the special case when i and j are non-admixed and have ancestry from different subpopulations k and k', the limiting value of the estimator reaches an extreme negative value and

$$\widehat{\kappa}_{ij} \to \frac{-\frac{1}{2}(F_k + F_{k'})}{1 - \frac{1}{2}(F_k + F_{k'})}. \tag{3.2}$$

PC-AiR uses the $\widehat{\kappa}_{ij}$ estimator given by Equation (3.1) for inference on ancestry divergence for all pairs of individuals $i, j \in \mathcal{N}$ who are not inferred to be related based on the kinship coefficient measures discussed in the previous subsection.

3.2.4 Identification of an Ancestry Representative Subset

We now provide details on how PC-AiR uses both the relatedness and ancestry divergence measures discussed in the previous two subsections for the identification of $\mathcal{U}$, a mutually unrelated subset of individuals that is representative of the ancestries of all individuals in the sample $\mathcal{N}$. Let $\widehat{\phi}_{ij}$ be the kinship coefficient measure that is chosen for relatedness inference on a pair of individuals $i, j \in \mathcal{N}$. When the genealogy of the sample individuals is known, $\widehat{\phi}_{ij}$ could be a pedigree-based kinship coefficient, and when the genealogy is partially or completely unknown, $\widehat{\phi}_{ij}$ would be an empirical kinship coefficient estimate from a relatedness estimation method that allows for population structure, e.g., the KING-robust estimator of Equation (3.1) , REAP, or RelateAdmix. In order to identify all pairs of relatives in $\mathcal{N}$, a relatedness threshold, τ_ϕ, is chosen such that i and j are designated to be related by the PC-AiR method if $\widehat{\phi}_{ij} > \tau_\phi$. When pedigree-based kinship coefficients are used with PC-AiR, all unrelated pairs will have $\widehat{\phi}_{ij} = 0$, and τ_ϕ should be set to 0. When empirical kinship coefficient estimates are used, there will be some noise in the estimation, and τ_ϕ can be set to an approximate upper bound that is expected for the chosen kinship coefficient estimator for an unrelated pair. For example, when using KING-robust for relatedness inference, i.e. using $\widehat{\phi}_{ij} = \widehat{\kappa}_{ij}$, we have found that 0.025 is an approximate upper bound with dense SNP genotyping data for unrelated pairs with the same

ancestry, and setting $\tau_\phi = 0.025$ works well in practice for identifying relatives with similar ancestry up to third-degree (and some fourth-degree) in a variety of population structure settings with ancestry admixture. For all sample individuals $i \in \mathcal{N}$, we calculate $\gamma_i = \sum_{j \neq i} \widehat{\phi}_{ij} \mathbb{1}_{[\widehat{\phi}_{ij} > \tau_\phi]}$ as a measure of the total kinship individual i has with inferred relatives in the sample, where $\mathbb{1}_{[\widehat{\phi}_{ij} > \tau_\phi]}$ is the indicator that individual j is inferred to be related to i.

PC-AiR uses $\widehat{\kappa}_{ij}$ to infer ancestry divergence for all pairs of individuals $i, j \in \mathcal{N}$ who are not inferred to be relatives. We showed that $\widehat{\kappa}_{ij}$ is a consistent estimator of 0 for unrelated pairs with the same ancestry, while unrelated pairs with different ancestry will have $\widehat{\kappa}_{ij}$ values that are systematically negative. We define a pair of individuals i and j to be "divergent" if they have different ancestral backgrounds, i.e. $\widehat{\kappa}_{ij} < -\tau_\kappa$, where $-\tau_\kappa$ is the expected lower bound of $\widehat{\kappa}_{ij}$ for a pair of unrelated individuals with the same ancestry. Since the distribution of $\widehat{\kappa}_{ij}$ for unrelated pairs with the same ancestry will be symmetric around 0, we expect that the vast majority of these pairs will satisfy $\widehat{\kappa}_{ij} \in [-0.025, 0.025]$ when $|\mathcal{S}_{ij}|$ is large, where 0.025 is the previously mentioned approximate upper bound for unrelated pairs. We have found that setting $-\tau_\kappa = -0.025$ works well in practice for identifying unrelated pairs of individuals with different admixed ancestries. For all sample individuals $i \in \mathcal{N}$, we calculate $\delta_i = \sum_{j \neq i} \mathbb{1}_{[\widehat{\phi}_{ij} < \tau_\phi,\ \widehat{\kappa}_{ij} < -\tau_\kappa]}$, the number of divergent ancestry pairs that individual i is a member of. Small δ_i values generally correspond to individuals with ancestry that is similar to the ancestries of many other individuals in $\mathcal{N}$, while the highest δ_i values generally correspond to individuals with unique ancestry and/or individuals with an ancestry proportion close to 1 from some subpopulation.

The algorithm used by PC-AiR for partitioning the set $\mathcal{N}$ based on measures of ancestry divergence and kinship is presented in Algorithm 3.1. It is both fast and efficient, and the two subsets returned from the algorithm are the ancestry representative and mutually unrelated subset, $\mathcal{U}$, and the related subset, $\mathcal{R}$, where each individual in $\mathcal{R}$ has at least one relative in $\mathcal{U}$. The algorithm is constructed in such a way that one

individual from any set of mutually related individuals in $\mathcal{N}$ will be included in $\mathcal{U}$, with priority given to the individual who is a member of the most divergent ancestry pairs (large δ_i). This helps to ensure that every ancestry in $\mathcal{N}$ is represented by some individual(s) in $\mathcal{U}$, while simultaneously satisfying the requirement that individuals in $\mathcal{U}$ are also mutually unrelated. It also favors choosing the individuals with the highest ancestry proportions from each of the K subpopulations for $\mathcal{U}$. These individuals will be at the extremes of the $K-1$ dimensional space spanned by the axes of variation representing the ancestries in $\mathcal{N}$, and selecting them for $\mathcal{U}$ helps to avoid shrinkage in prediction of principal component values for individuals in $\mathcal{R}$. Secondary priority for inclusion in $\mathcal{U}$ is given to individuals that share the most genetic information with their collection of relatives in $\mathcal{N}$ (large γ_i), also allowing for better prediction of principal component values for relatives in $\mathcal{R}$.

3.2.5 Genetic Relationship Matrix for PC-AiR

The traditional PCA approach for population structure inference with genetic data, e.g., the EIGENSOFT method, performs PCA on standardized genotypes, where the standardized genotype value for individual i at SNP s is given by

$$z_{is} = \frac{g_{is} - 2\hat{p}_s}{\sqrt{2\hat{p}_s(1-\hat{p}_s)}}, \tag{3.3}$$

and $\hat{p}_s$ will typically be an allele frequency estimate for SNP s calculated using all sample individuals. The PC-AiR method also uses standardized genotypes, but the allele frequencies used for the standardization are calculated using only the unrelated individuals selected for $\mathcal{U}$. The standardized genotype values for PC-AiR are calculated from Equation (3.3) by setting $\hat{p}_s = \hat{p}_s^u$, where

$$\hat{p}_s^u = \frac{1}{2|\mathcal{U}_s|}\sum_{i \in \mathcal{U}_s} g_{is}, \tag{3.4}$$

$\mathcal{U}_s$ is the subset of individuals in $\mathcal{U}$ who have non-missing genotype data at SNP s, and $|\mathcal{U}_s|$ is the number of individuals in $\mathcal{U}_s$. In samples with related individuals

Algorithm 3.1 PC-AiR Algorithm for Partitioning $\mathcal{N}$ into $\mathcal{U}$ and $\mathcal{R}$

(1) Compute: $\gamma_i = \sum_{\substack{j \neq i \\ j \in \mathcal{N}}} \widehat{\phi}_{ij} \mathbb{1}_{[\widehat{\phi}_{ij} > \tau_\phi]}$ for all $i \in \mathcal{N}$.

(2) Compute: $\delta_i = \sum_{\substack{j \neq i \\ j \in \mathcal{N}}} \mathbb{1}_{[\widehat{\phi}_{ij} < \tau_\phi,\ \widehat{\kappa}_{ij} < -\tau_\kappa]}$ for all $i \in \mathcal{N}$.

(3) Initialize the two subsets to be $\mathcal{U} = \mathcal{N}$ and $\mathcal{R} = \varnothing$, where $\varnothing$ is the empty set.

(4) Compute: $\eta_i = \begin{cases} \sum_{\substack{j \neq i \\ j \in \mathcal{U}}} \mathbb{1}_{[\widehat{\phi}_{ij} > \tau_\phi]} & \forall i \in \mathcal{U} \\ 0 & \forall i \in \mathcal{R} \end{cases}$.

If $\max_i(\eta_i) > 0$, continue to step (5), otherwise go to step (11).

(5) Identify $\mathcal{T}_1 = \{i | \eta_i = \max_j(\eta_j)\}$, the subset of individuals in $\mathcal{U}$ with the most relatives in $\mathcal{U}$.

If $|\mathcal{T}_1| > 1$, where $|\mathcal{T}_1|$ is the number of elements in $\mathcal{T}_1$, go to step (6). Otherwise set $\mathcal{T}^* = \mathcal{T}_1$ and go to step (9).

(6) Identify $\mathcal{T}_2 = \{i | \delta_i = \min_{j \in \mathcal{T}_1}(\delta_j)\}$, the subset of individuals in $\mathcal{T}_1$ that are members of the least divergent ancestry pairs.

If $|\mathcal{T}_2| > 1$, go to step (7). Otherwise set $\mathcal{T}^* = \mathcal{T}_2$ and go to step (9).

(7) Identify $\mathcal{T}_3 = \{i | \gamma_i = \min_{j \in \mathcal{T}_2}(\gamma_j)\}$, the subset of individuals in $\mathcal{T}_2$ that have the minimum total kinship with their inferred relatives.

If $|\mathcal{T}_3| > 1$, go to step (8). Otherwise set $\mathcal{T}^* = \mathcal{T}_3$ and go to step (9).

(8) Randomly select one element from $\mathcal{T}_3$ and define this element to be the set $\mathcal{T}^*$.

(9) Define the sets: $\mathcal{U}^* = \mathcal{U} \backslash \mathcal{T}^*$ and $\mathcal{R}^* = \mathcal{R} \cup \mathcal{T}^*$.

(10) Update $\mathcal{U} = \mathcal{U}^*$ and $\mathcal{R} = \mathcal{R}^*$ and return to step (4).

(11) The algorithm has completed.

and population structure, we have found that using the estimator $\hat{p}_s^u$ provides better ancestry inference with PC-AiR than using allele frequency estimates calculated from the entire sample, which can be heavily influenced by the correlated genotypes among relatives. For any individual $i \in \mathcal{N}$ with a missing genotype value at SNP s, z_{is} is set to 0, i.e. g_{is} is set equal to $2\hat{p}_s^u$, an estimate of its expected value. Provided that individuals with high levels of missingness are excluded from analysis as a result of standard quality control [24], a small percentage of mean imputed genotypes should not bias results.

Similar minor allele frequency filtering and LD pruning of SNPs that have been recommended for standard PCA [40, 41] should also be used for PC-AiR. Let $|\mathcal{S}^*|$ be the number of SNPs in the pruned and filtered set $\mathcal{S}^*$, and let n, n_u, and n_r be the number of individuals in set $\mathcal{N}$ and subsets $\mathcal{U}$ and $\mathcal{R}$, respectively, with $n = n_u + n_r$. We construct $\mathbf{Z}$, an $n \times |\mathcal{S}^*|$ standardized genotype matrix for $\mathcal{N}$, with (i, s)-th entry equal to z_{is}, ordered such that the first n_u rows correspond to individuals in $\mathcal{U}$, and the remaining n_r rows correspond to individuals in $\mathcal{R}$. The standardized genotype matrix for $\mathcal{U}$ is the $n_u \times |\mathcal{S}^*|$ submatrix $\mathbf{Z}_u$ corresponding to the first n_u rows of $\mathbf{Z}$. Similarly, the $n_r \times |\mathcal{S}^*|$ submatrix $\mathbf{Z}_r$ is the standardized genotype matrix for $\mathcal{R}$ corresponding to the last n_r rows of $\mathbf{Z}$.

Similar to the traditional PCA approach, PC-AiR obtains a genetic relationship matrix (GRM) for population structure inference from standardized genotypes. It is important to note that PCA applied to a GRM that includes all individuals in $\mathcal{N}$, as in the traditional PCA approaches, leads to artifactual principal components for ancestry due to confounding from correlated genotypes among relatives, i.e. genetic similarities are reflecting alleles shared IBD among relatives. To protect against confounding caused by sample relatedness, PC-AiR instead calculates a GRM using only the mutually unrelated sample individuals who were selected to be included in the ancestry representative subset, $\mathcal{U}$. The empirical n_u x n_u GRM for $\mathcal{U}$ calculated with

the standardized genotype matrix $\mathbf{Z}_u$ is

$$\widehat{\mathbf{\Psi}}_u = \frac{1}{|\mathcal{S}^*|}\mathbf{Z}_u\mathbf{Z}_u^T, \tag{3.5}$$

and the (i, j)-th entry of $\widehat{\mathbf{\Psi}}_u$ provides a measure of the average genetic similarity across the autosomes for individuals $i, j \in \mathcal{U}$. For further discussion and explanation regarding the properties of the GRM, see Appendix A.2.1.

3.2.6 Population Structure Inference in Related Samples with PC-AiR

To obtain principal components that are ancestry representative on a set $\mathcal{N}$ containing related individuals, the PC-AiR method first performs a PCA using genome-screen data from only those individuals selected to be in the mutually unrelated ancestry representative subset, $\mathcal{U}$. PCA is performed by obtaining the eigendecomposition of the GRM $\widehat{\mathbf{\Psi}}_u$ from Equation (3.5). This procedure sequentially identifies orthogonal axes of variation, i.e. linear combinations of SNPs, that best explain the genotypic variability amongst the individuals in $\mathcal{U}$, where each axis of variation reflects the structure that leads to the greatest variability after accounting for the structure explained by all previously defined axes. The eigendecomposition of $\widehat{\mathbf{\Psi}}_u$ results in an $n_u \times n_u$ matrix, $\mathbf{V}_u = [\mathbf{V}_1^u, \mathbf{V}_2^u, \ldots, \mathbf{V}_{n_u}^u]$, with orthogonal, length n_u, column vectors, and a corresponding length n_u vector, $\mathbf{\Lambda}_u = (\lambda_1^u, \lambda_2^u, \ldots, \lambda_{n_u}^u)$, with the property of $\lambda_1^u > \lambda_2^u > \ldots > \lambda_{n_u}^u$. For $d \in \{1, \ldots, n_u\}$, $\mathbf{V}_d^u$ and λ_d^u are the corresponding d^{th} principal component (eigenvector) and eigenvalue of $\widehat{\mathbf{\Psi}}_u$, where λ_d^u is proportional to the percentage of variability in the genome-screen data for $\mathcal{U}$ that is explained by $\mathbf{V}_d^u$. By construction, individuals in $\mathcal{U}$ are mutually unrelated and have diverse ancestry, so the top principal components of $\widehat{\mathbf{\Psi}}_u$ are expected to be representative of ancestry.

Once PCA has been performed on $\mathcal{U}$, principal components values for individuals in the related subset, $\mathcal{R}$, can be obtained via prediction. Let $\mathbf{L}_u$ be a diagonal matrix created from the vector of eigenvalues, i.e. $\mathbf{L}_u = diag(\mathbf{\Lambda}_u)$. An $|\mathcal{S}^*| \times n_u$ SNP weight matrix giving the relative influence of each SNP on each of the n_u eigenvectors can

be obtained as $\mathbf{W}_u = \mathbf{Z}_u^T \mathbf{V}_u$, and from the form of the eigendecomposition of the real symmetric matrix $\widehat{\mathbf{\Psi}}_u = \mathbf{V}_u \mathbf{L}_u \mathbf{V}_u^{-1}$, it can be shown [16] that the principal components for the ancestry representative subset, $\mathcal{U}$, can alternatively be written as:

$$\mathbf{V}_u = \widehat{\mathbf{\Psi}}_u \mathbf{V}_u \mathbf{L}_u^{-1} = \left(\frac{1}{|\mathcal{S}^*|} \mathbf{Z}_u \mathbf{Z}_u^T \right) \mathbf{V}_u \mathbf{L}_u^{-1} = \frac{1}{|\mathcal{S}^*|} \mathbf{Z}_u \mathbf{W}_u \mathbf{L}_u^{-1}. \tag{3.6}$$

For the related subset, $\mathcal{R}$, the PC-AiR method predicts principal components values from Equation (3.6) by replacing $\mathbf{Z}_u$, the standardized genotype matrix for individuals in $\mathcal{U}$, with $\mathbf{Z}_r$, the standardized genotype matrix for individuals in $\mathcal{R}$. The n_r x n_u matrix of predicted eigenvectors for $\mathcal{R}$, which we denote as $\mathbf{Q}_r$, is thus given by:

$$\mathbf{Q}_r = \frac{1}{|\mathcal{S}^*|} \mathbf{Z}_r \mathbf{W}_u \mathbf{L}_u^{-1}. \tag{3.7}$$

The d^{th} column in the matrix $\mathbf{Q}_r$ corresponds to PC-AiR's predicted coordinates along the d^{th} principal component for the individuals in $\mathcal{R}$. We define $\mathbf{\Gamma}$ to be the $n \times n_u$ matrix of the combined principal components for $\mathcal{U}$ and $\mathcal{R}$, where:

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{V}_u \\ \hdashline \mathbf{Q}_r \end{bmatrix} = \left[\begin{array}{c:c:c:c} \mathbf{V}_1^u & \mathbf{V}_2^u & \cdots & \mathbf{V}_{n_u}^u \\ \hdashline \mathbf{Q}_1^r & \mathbf{Q}_2^r & \cdots & \mathbf{Q}_{n_u}^r \end{array}\right]. \tag{3.8}$$

The column vectors of $\mathbf{\Gamma}$ are the principal components (axes of variation) of the set $\mathcal{N} = \mathcal{U} \cup \mathcal{R}$ obtained from the PC-AiR method. The genetic structure that is reflected by all of the principal components for PC-AiR are found using only the ancestry representative subset, $\mathcal{U}$, and thus the top principal components from $\mathbf{\Gamma}$ are designed to be representative of ancestry in $\mathcal{N}$, even in the presence of known or cryptic relatedness.

3.2.7 Simulation Studies

We perform simulation studies in which both population and pedigree structure are simultaneously present in order to assess the accuracy and robustness of the PC-AiR method for population structure inference in the presence of relatedness, and to

compare the performance of PC-AiR to existing methods. We simulate a variety of population structure settings, including admixture and ancestry-related assortative mating, with differentiation between populations ranging from subtle to large. We evaluate population structure inference for four different relationship configurations, where each configuration corresponds to a specific setting of genealogical relationships among the sample individuals. In all simulation studies considered, pedigree information on the sample individuals is hidden and genetic relatedness is inferred from the genotype data with the PC-AiR method using the KING-robust kinship estimator in Equation (3.1).

Population Structure Settings

The population structure settings we consider are similar to the settings in Price et al. [41], where PCA was performed with the EIGENSOFT software in unrelated samples for inference on and adjustment for population structure in GWAS, except that our simulation studies include related individuals. We consider population structure settings where individuals have ancestry derived from two populations, and the allele frequencies at 100,000 SNPs for each of these two populations are generated using the Balding-Nichols model [4]. More precisely, for each SNP s, the allele frequency p_s in the ancestral population is drawn from a uniform distribution on $[0.1, 0.9]$, and the allele frequency in population $k \in \{1, 2\}$ is drawn from a beta distribution with parameters $p_s(1-F_k)/F_k$ and $(1-p_s)(1-F_k)/F_k$, where the quantity F_k is equivalent to Wright's measure of population differentiation [62] from the ancestral population. In all simulations, we set F_1 and F_2 equal to a common value, F_{ST}. To generate allele frequencies derived from populations ranging from closely related to highly divergent, we consider F_{ST} values from 0.01 to 0.2.

For each F_{ST} value considered, we simulate three population structure settings. Population structures I and II both consist of individuals sampled from an admixed population formed from populations 1 and 2. For population structure I, all unre-

lated individuals and pedigree founders have ancestry proportions a from population 1 and $(1 - a)$ from population 2, with the parameter a for each individual drawn from a uniform distribution on $[0, 1]$. Population structure II is similar to population structure I, but with the ancestry parameter, a, drawn from a beta distribution with mean 0.4 and standard deviation 0.1 for 50% of the unrelated individuals and pedigree founders, and with mean 0.6 and standard deviation 0.1 for the other 50%. All founders within the same pedigree have a drawn from the same beta distribution for population structure II. Population structure III consists of non-admixed individuals, where 50% of the unrelated individuals and pedigrees are sampled from population 1, and the other 50% are sampled from population 2. Both population structure settings II and III have ancestry-related assortative mating, i.e., the mating of founder individuals in every pedigree occurs with individuals who have either the same (population structure III) or similar (population structure II) ancestry, while population structure I has random mating that is independent of ancestry.

Relationship Configurations

Three of the four relationship configurations simulated include both related and unrelated individuals. Relationship configuration I consists of 200 unrelated individuals and 200 individuals from 10 four-generation pedigrees, where each pedigree has a total of 20 individuals (Figure 3.1). Relationship configuration II is comprised of 280 unrelated individuals with 20 parent-offspring trios, and relationship configuration III includes 260 unrelated individuals with 20 sibling pairs. To sample pedigree relationships within a given setting of population structure, we simulate genotypes for pedigree founders under Hardy-Weinberg equilibrium (HWE) according to the chosen population structure setting and then drop alleles down the pedigree. Relationship configuration IV consists of 320 unrelated individuals without any family structure. We include the unrelated sample setting in our simulation studies in order to evaluate any loss in population structure inference with the PC-AiR method compared to stan-

dard PCA in a setting where standard PCA is appropriate and has been previously demonstrated to perform well.

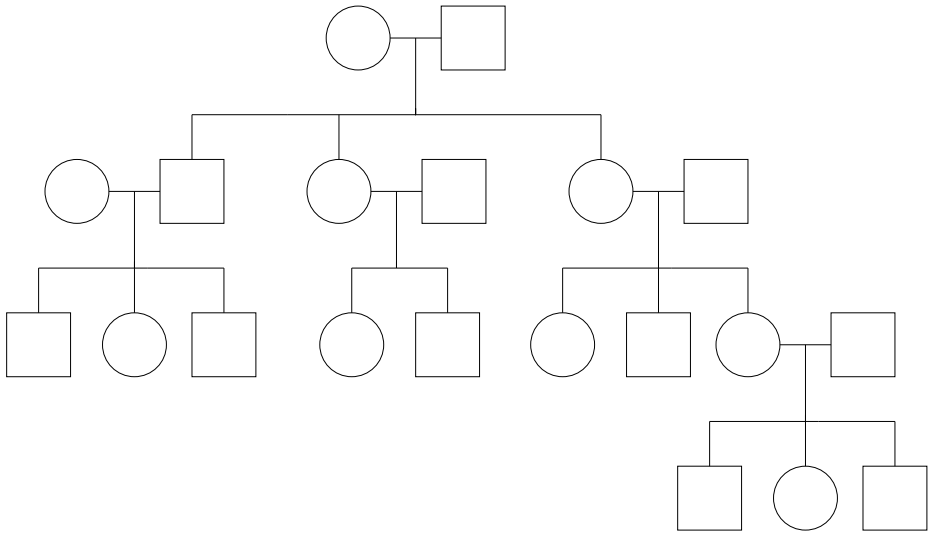


Figure 3.1: Extended Pedigree Configuration for Simulation Studies
The pedigree configuration for each of the 10 outbred, four-generation pedigrees included in Relationship Configuration I of the simulation studies, where the overall structure of each pedigree is as depicted, but the pattern of ancestry admixture varies according to the specified population structure setting.

3.3 Results

3.3.1 Subtle Population and Pedigree Structure

We first considered samples with subtle population structure, where the ancestry of the sample individuals was derived from two closely related populations. We set F_{ST} to 0.01 (a typical value for divergent European populations) and generated genotype data under population structure I for each of the four relationship configurations. Population structure inference with PC-AiR was compared to that of standard PCA with the EIGENSOFT software. To assess the performance of the two methods, we

included the top principal components (axes of variation) from each method as predictors for the true simulated ancestry of the sample individuals in a linear regression model, and the proportion of ancestry explained, as measured by R^2, was used to evaluate prediction accuracy. We also compared the efficiency of PC-AiR to EIGENSOFT by assessing the number of top axes of variation required to attain an R^2 of at least 0.99 for ancestry. It should be noted that since the data in the simulation studies contained only one added dimension of population structure, an optimal method would require only a single axis of variation for complete ancestry inference. Both PC-AiR and EIGENSOFT were provided only genotype data without any additional pedigree information on the sample individuals.

Figure 3.2 displays the population structure inference results for relationship configuration I from both PC-AiR and EIGENSOFT. Figure 3.2B displays the top two axes of variation obtained by EIGENSOFT, which almost entirely reflected pedigree structure in the sample. The ten spikes of points radiating from the center cluster in the figure correspond to the individuals who are members of the ten pedigrees, and the cluster of points in the center of the plot corresponds to the 200 individuals who do not have any relatives in the sample. In contrast, the top two axes of variation from PC-AiR were not confounded by family structure, as illustrated in Figure 3.2A, and the top axis explained ancestry in the sample nearly perfectly, with an R^2 of 0.993 (Figure 3.2C). Figure 3.2D shows that the top axis of variation from EIGENSOFT did not reflect population structure and did not adequately capture the ancestry of the sample individuals, with an R^2 of only 0.133. The efficiency for population structure inference of both methods is illustrated in Figure 3.2E, where the proportion of ancestry explained (R^2 values) for each of the top axes of variation is displayed. EIGENSOFT required the top 51 axes to be included as predictors in a linear regression model to achieve an R^2 of at least 0.99 for ancestry. In contrast, a single axis of variation from PC-AiR had an R^2 greater than 0.99, thus demonstrating a substantial improvement in efficiency with PC-AiR over EIGENSOFT in this setting with both

subtle population structure and relatedness.

Population structure inference results with PC-AiR and EIGENSOFT for relationship configurations II and III are presented in Figures 3.3 and 3.4. The top axes of variation from EIGENSOFT were influenced by relatedness, as expected; however, since relationship configurations II and III have substantially less pedigree structure than relationship configuration I, there was some improvement in ancestry prediction with the top axis in each of these two settings, with R^2 values of 0.870 and 0.933, respectively. For both relationship configurations II and III, the top 21 axes of variation from EIGENSOFT were required to attain an R^2 of at least 0.99 for predicting ancestry. In comparison, the PC-AiR analysis was robust to the relatedness in the sample, and the single top axis of variation for both relationship configurations II and III attained an R^2 value greater than 0.99 for predicting ancestry. For relationship configuration IV, PC-AiR accurately identified all sample individuals to be unrelated, i.e. the ancestry informative subset, $\mathcal{U}$, was the entire sample, $\mathcal{N}$, so the PC-AiR method reduced to standard PCA, and inference with either PC-AiR or EIGENSOFT was essentially identical. The R^2 for ancestry with the top axis of variation from both methods was greater than 0.99, illustrating that there is no loss in accuracy or efficiency compared to standard PCA when using PC-AiR for population structure inference in samples where all individuals are unrelated.

We also evaluated the performance of PC-AiR and EIGENSOFT under population structures II and III with F_{ST} set to 0.01 for each of the relationship configurations. The results are given in Table 3.1, and the conclusions drawn from these population structure settings are the same as those for population structure I. For the three relationship configurations that included related samples, a single axis of variation from PC-AiR fully explained the ancestry in the sample and provided better prediction of ancestry than using ten (or more) axes from EIGENSOFT. For relationship configuration IV, where all sample individuals were unrelated, PC-AiR and EIGENSOFT gave essentially identical results, with the top axis from both methods fully explaining

the true ancestry.

Table 3.1: Proportion of Ancestry Explained (R^2) by PC-AiR and EIGENSOFT in Simulation Studies

Relationship Configuration	Population Structure	F_{ST}	PC-AiR	EIGENSOFT			
			$d^a = 1$	$d = 1$	$d = 4$	$d = 10$	d^{*b}
I	I	0.01	0.993	0.133	0.145	0.165	51
		0.1	0.999	0.977	0.977	0.993	10
		0.2	0.999	0.995	0.995	0.999	1
	II	0.01	0.949	0.302	0.360	0.402	51
		0.1	0.998	0.741	0.741	0.755	41
		0.2	0.999	0.882	0.882	0.914	21
	III	0.01	0.999	0.832	0.832	0.832	22
		0.1	0.999	0.998	0.998	0.998	1
		0.2	0.999	0.999	0.999	0.999	1
II	I	0.01	0.994	0.870	0.871	0.872	21
		0.1	0.999	0.999	0.999	0.999	1
		0.2	0.999	0.999	0.999	0.999	1
	II	0.01	0.942	0.259	0.313	0.320	21
		0.1	0.998	0.983	0.983	0.984	21
		0.2	0.999	0.996	0.996	0.996	1
	III	0.01	0.999	0.990	0.990	0.990	1
		0.1	0.999	0.999	0.999	0.999	1
		0.2	0.999	0.999	0.999	0.999	1
III	I	0.01	0.990	0.933	0.933	0.936	21
		0.1	0.999	0.999	0.999	0.999	1
		0.2	0.999	0.999	0.999	0.999	1
	II	0.01	0.922	0.220	0.230	0.250	21
		0.1	0.997	0.992	0.992	0.993	1
		0.2	0.999	0.998	0.998	0.998	1
	III	0.01	0.998	0.995	0.995	0.995	1
		0.1	0.999	0.999	0.999	0.999	1
		0.2	0.999	0.999	0.999	0.999	1

[a] d denotes the number of axes of variation included as predictors in the linear regression model to determine the R^2 value for either PC-AiR or EIGENSOFT.

[b] d^* is the number of axes of variation from EIGENSOFT that are required to either match the R^2 value of the first axis of variation from PC-AiR or achieve an R^2 of 0.99, whichever is smaller.

Figure 3.2: Comparison of PC-AiR and EIGENSOFT for Relationship Configuration I and Population Structure I with $F_{ST} = 0.01$

(A and B) Scatter plots of principal components 1 and 2 from PC-AiR (A) and EIGENSOFT (B), respectively. (C and D) Scatter plots of the simulated population 1 ancestry proportions vs. coordinates along principal component 1 for each individual from PC-AiR (C) and EIGENSOFT (D), respectively. (A-D) The color of each point represents that individual's true ancestry; red for population 1, blue for population 2, and an intermediate color for an admixed individual. (A and C) A dot represents an individual in the mutually unrelated ancestry representative set, and a plus represents an individual in the related set. (B and D) A circle represents an individual not in a pedigree, and a triangle represents an individual who is a member of a pedigree. (E) Barplot of the efficiency of PC-AiR and EIGENSOFT. Each bar represents the proportion of ancestry explained (R^2 value) by each principal component from PC-AiR (gold) and EIGENSOFT (black), until a cumulative R^2 of 0.99 is achieved.

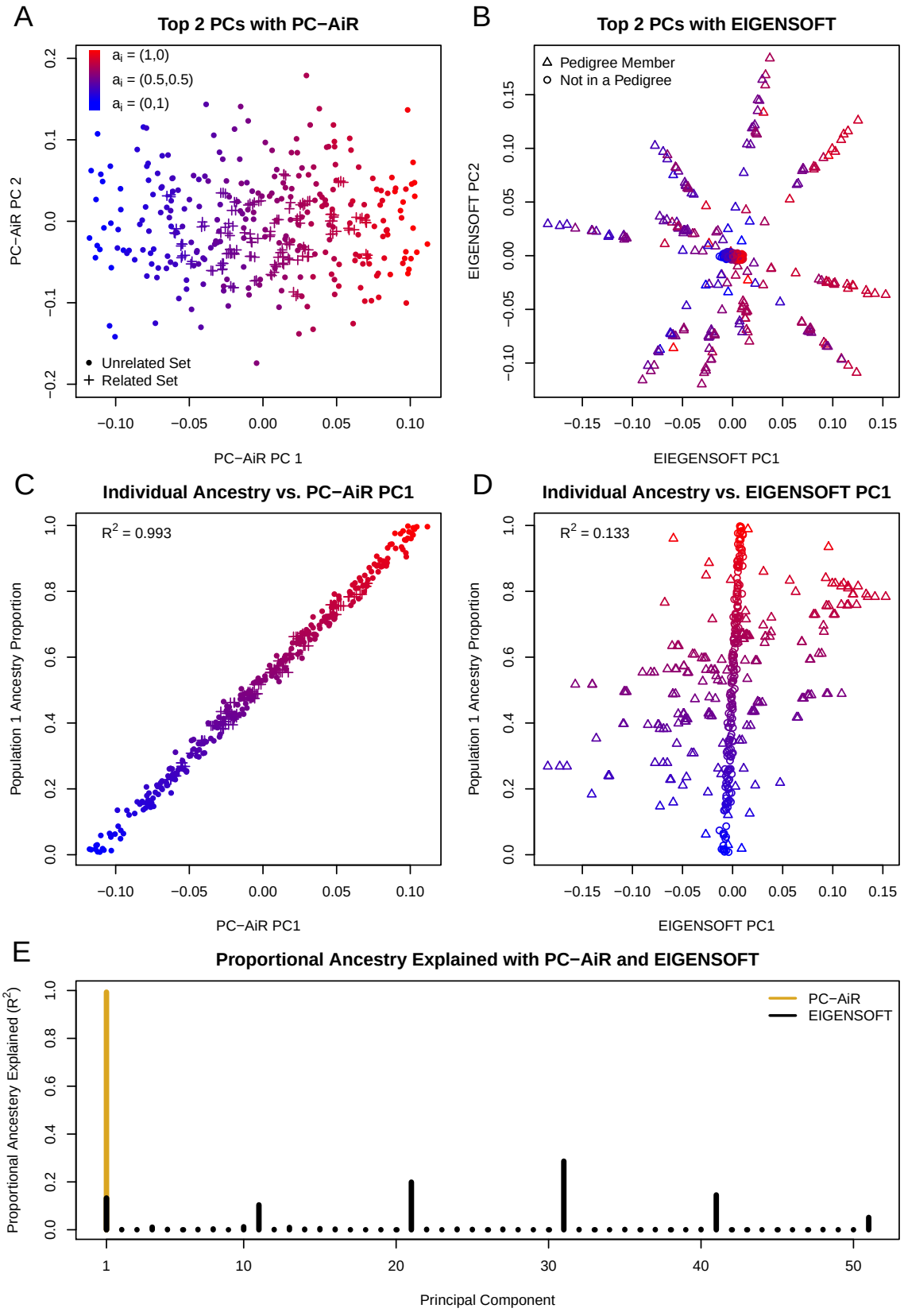

A
Top 2 PCs with PC−AiR
a_i = (1,0)
a_i = (0.5,0.5)
a_i = (0,1)
Unrelated Set
Related Set
PC−AiR PC 2
PC−AiR PC 1
B
Top 2 PCs with EIGENSOFT
Pedigree Member
Not in a Pedigree
EIGENSOFT PC2
EIEGENSOFT PC1
C
Individual Ancestry vs. PC−AiR PC1
R^2 = 0.993
Population 1 Ancestry Proportion
PC−AiR PC1
D
Individual Ancestry vs. EIGENSOFT PC1
R^2 = 0.133
Population 1 Ancestry Proportion
EIGENSOFT PC1
E
Proportional Ancestry Explained with PC−AiR and EIGENSOFT
PC−AiR
EIGENSOFT
Proportional Ancestery Explained (R^2)
Principal Component

Figure 3.3: Comparison of PC-AiR and EIGENSOFT for Relationship Configuration II and Population Structure I with $F_{ST} = 0.01$

(A and B) Scatter plots of principal components 1 and 2 from PC-AiR (A) and EIGENSOFT (B), respectively. (C and D) Scatter plots of the simulated population 1 ancestry proportions vs. coordinates along principal component 1 for each individual from PC-AiR (C) and EIGENSOFT (D), respectively. (A-D) The color of each point represents that individual's true ancestry; red for population 1, blue for population 2, and an intermediate color for an admixed individual. (A and C) A dot represents an individual in the ancestry representative, mutually unrelated set, and a plus represents an individual in the related set. (B and D) A diamond represents an individual who is a member of a parent/offspring trio, and a circle represents the remaining individuals. (E) Barplot of the efficiency of PC-AiR and EIGENSOFT. Each bar represents the proportion of ancestry explained (R^2 value) by each principal component from PC-AiR (gold) and EIGENSOFT (black), until a cumulative R^2 of 0.99 is achieved.

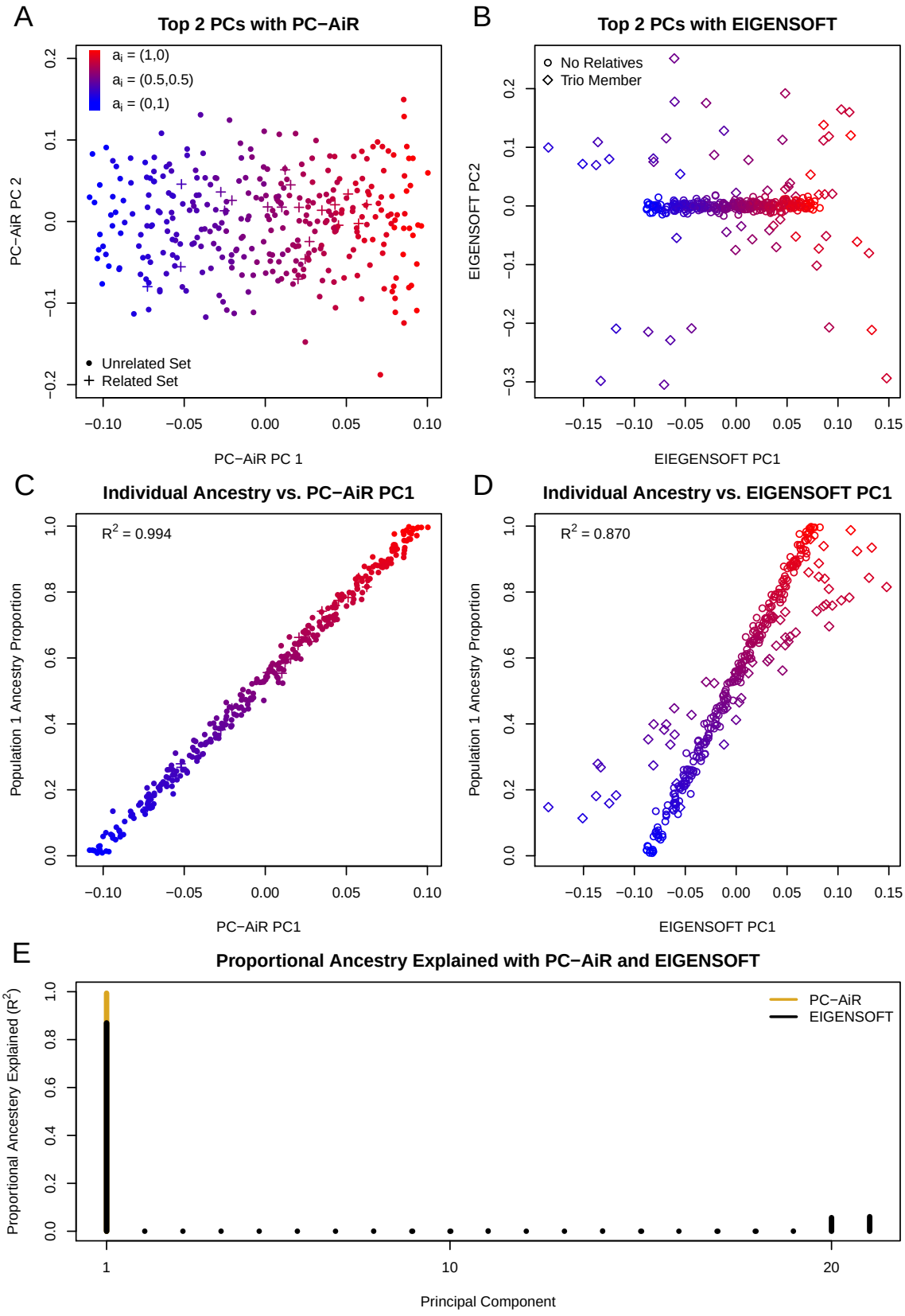
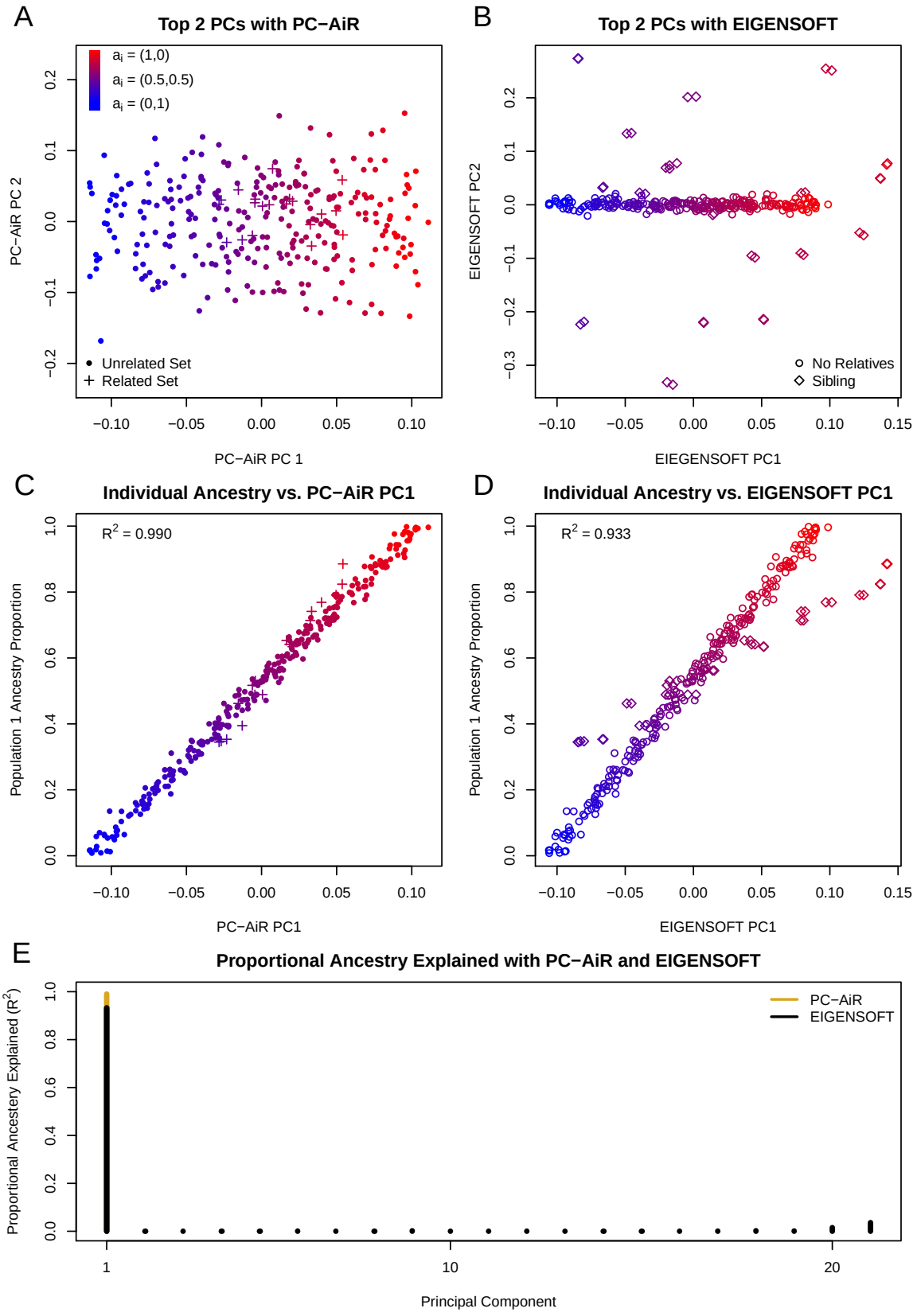

A
Top 2 PCs with PC−AiR
a_i = (1,0)
a_i = (0.5,0.5)
a_i = (0,1)
Unrelated Set
Related Set
PC−AiR PC 2
PC−AiR PC 1
B
Top 2 PCs with EIGENSOFT
No Relatives
Trio Member
EIGENSOFT PC2
EIEGENSOFT PC1
C
Individual Ancestry vs. PC−AiR PC1
R^2 = 0.994
Population 1 Ancestry Proportion
PC−AiR PC1
D
Individual Ancestry vs. EIGENSOFT PC1
R^2 = 0.870
Population 1 Ancestry Proportion
EIGENSOFT PC1
E
Proportional Ancestry Explained with PC−AiR and EIGENSOFT
PC−AiR
EIGENSOFT
Proportional Ancestery Explained (R^2)
Principal Component

Figure 3.4: Comparison of PC-AiR and EIGENSOFT for Relationship Configuration III and Population Structure I with $F_{ST} = 0.01$

(A and B) Scatter plots of principal components 1 and 2 from PC-AiR (A) and EIGENSOFT (B), respectively. (C and D) Scatter plots of the simulated population 1 ancestry proportions vs. coordinates along principal component 1 for each individual from PC-AiR (C) and EIGENSOFT (D), respectively. (A-D) The color of each point represents that individual's true ancestry; red for population 1, blue for population 2, and an intermediate color for an admixed individual. (A and C) A dot represents an individual in the ancestry representative, mutually unrelated set, and a plus represents an individual in the related set. (B and D) A diamond represents an individual who is a member of a sibling pair, and a circle represents the remaining individuals. (E) Barplot of the efficiency of PC-AiR and EIGENSOFT. Each bar represents the proportion of ancestry explained (R^2 value) by each principal component from PC-AiR (gold) and EIGENSOFT (black), until a cumulative R^2 of 0.99 is achieved.


A
Top 2 PCs with PC−AiR
a_i = (1,0)
a_i = (0.5,0.5)
a_i = (0,1)
Unrelated Set
Related Set
PC−AiR PC 2
PC−AiR PC 1
B
Top 2 PCs with EIGENSOFT
No Relatives
Sibling
EIGENSOFT PC2
EIEGENSOFT PC1
C
Individual Ancestry vs. PC−AiR PC1
R^2 = 0.990
Population 1 Ancestry Proportion
PC−AiR PC1
D
Individual Ancestry vs. EIGENSOFT PC1
R^2 = 0.933
Population 1 Ancestry Proportion
EIGENSOFT PC1
E
Proportional Ancestry Explained with PC−AiR and EIGENSOFT
PC−AiR
EIGENSOFT
Proportional Ancestery Explained (R^2)
Principal Component

3.3.2 Relatedness and Admixture from Divergent Populations

We also conducted simulation studies with relatedness and admixture from divergent populations. We considered relationship configuration I and population structure II, where we set F_{ST} to 0.1 (a value representative of continental-level ancestry differences) in the Balding-Nichols model to simulate allele frequencies at SNPs derived from two divergent populations. We evaluated and compared the performance of PC-AiR to PCA with the EIGENSOFT software, MDS with the PLINK software, and the two model-based methods ADMIXTURE and FRAPPE for proportional ancestry estimation. As in the previous subsection, no genealogical information on the sample individuals was provided to any of the analysis methods, so the FamPCA method could not be used as it is restricted to settings with known pedigrees. The ADMIXTURE and FRAPPE software analyses were conducted with the correct number of populations specified.

The population structure inference results for each method considered are shown in Figure 3.5, where each panel is a plot of the simulated population 1 ancestry proportions against the inferred ancestry from one of the methods. The top axis of variation from PC-AiR had an R^2 of 0.998 and provided nearly perfect inference on ancestry for the sample individuals (Figure 3.5A). Similar to the EIGENSOFT results for the simulations with subtle population structure and relatedness, the top axis of variation did not adequately reflect the ancestry in this related sample with admixture from divergent populations, attaining an R^2 of only 0.741 (Figure 3.5B). ADMIXTURE and FRAPPE gave identical ancestry proportion estimates for all individuals in the simulation, and Figure 3.5D shows estimated proportional ancestry plotted against the simulated ancestry proportions from population 1. These model-based ancestry estimation methods were confounded by the pedigree structure in the sample and performed similarly to PCA, with an R^2 of only 0.730. While the top dimension of MDS achieved an R^2 of 0.785 and provided some improvement in predicting ances-

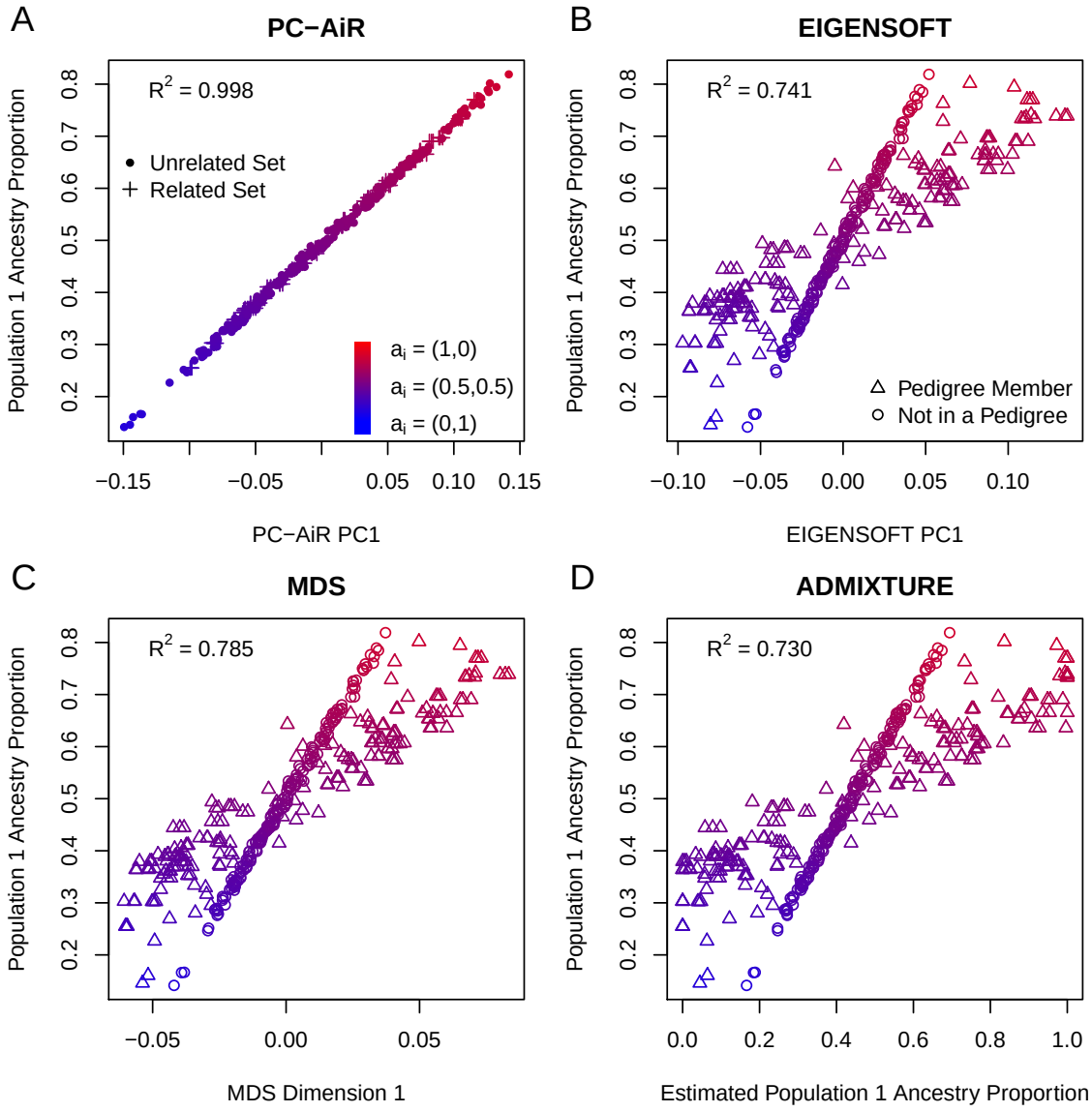
try over both of the model-based methods as well as the top axis of variation from EIGENSOFT, it was also confounded by sample relatedness, as shown in Figure 3.5C.

We also evaluated the performance of PC-AiR and EIGENSOFT for all combinations of relationship configurations and population structure settings with F_{ST} set to 0.1 and 0.2 (Table 3.1). For all settings considered, the top axis of variation from PC-AiR gave nearly perfect ancestry inference, attaining an $R^2 > 0.99$. The extent to which EIGENSOFT's PCA was confounded by the relatedness depended on how divergent the populations were, i.e. the F_{ST} values, and how complex the pedigree structure was; however, a single axis of variation from PC-AiR always performed as well as or better than using ten axes of variation from EIGENSOFT for ancestry prediction.

3.3.3 Ancestry Inference and Prediction in Related Samples with Reference Panels

Reference population panels are commonly used for improved ancestry inference in unrelated samples from admixed populations, such as African Americans and Hispanics. We conducted a simulation study evaluating population structure inference with reference panels in admixed samples with relatedness. We considered the same simulation study discussed in detail in the previous subsection, but we now included reference panels consisting of 50 unrelated individuals randomly sampled from each of the two underlying populations. The same population structure methods from the previous subsection were used, and the results are displayed in Figure 3.6. Ancestry inference with EIGENSOFT, MDS, ADMIXTURE, and FRAPPE was substantially improved by including the reference panels as compared to the analyses without them, but PC-AiR still outperformed all methods, with the top axis of variation achieving an R^2 of 0.999 with ancestry. The analyses with ADMIXTURE and FRAPPE, which were run supervised by specifying which samples were from the reference populations, once again gave identical results to each other, and the estimated ancestry proportions had an R^2 of 0.973 with the simulated ancestries. Similarly, the top axis of

Figure 3.5: Population Structure Inference Results for Relationship Configuration I and Population Structure II with $\mathbf{F_{ST}} = 0.1$
Simulated population 1 ancestry proportions for each individual are plotted against: (A) coordinates along principal component 1 from PC-AiR, (B) coordinates along principal component 1 from EIGENSOFT, (C) coordinates along dimension 1 from MDS, and (D) the estimated ancestry proportions from ADMIXTURE for the inferred population with the highest R^2. The color of each point represents that individual's true ancestry; red for population 1, blue for population 2, and an intermediate color for an admixed individual. (A) A dot represents an individual in the mutually unrelated ancestry representative set, and a plus represents an individual in the related set. (B-D) A circle represents an individual not in a pedigree, and a triangle represents an individual who is a member of a pedigree.


A
PC−AiR
R2 = 0.998
Unrelated Set
Related Set
ai = (1,0)
ai = (0.5,0.5)
ai = (0,1)
Population 1 Ancestry Proportion
PC−AiR PC1
B
EIGENSOFT
R2 = 0.741
Pedigree Member
Not in a Pedigree
EIGENSOFT PC1
C
MDS
R2 = 0.785
MDS Dimension 1
D
ADMIXTURE
R2 = 0.730
Estimated Population 1 Ancestry Proportion

variation from each of EIGENSOFT and MDS reached R^2 values of 0.970 and 0.979 respectively.

Interestingly, the top axis of variation from PC-AiR without additional reference population samples had an R^2 of 0.998 and provided better ancestry inference than all of the competing methods with the reference panels. Even with the inclusion of reference panels, there remains some bias in ancestry inference for all methods, except for PC-AiR, that is induced by the presence of related individuals in the sample. This can be seen in Figure 3.6, where the inferred ancestries for individuals with relatives in the sample were systematically biased for each of the competing methods. We have found that using ADMIXTURE (or FRAPPE) to conduct separate individual ancestry analyses for each of the admixed samples, i.e. analyses with genotype data from a single admixed sample individual and all individuals in the reference population panels, can remove the bias caused by sample relatedness, known or cryptic, as long as the reference panel samples are appropriate surrogates for the underlying populations. We performed these analyses with ADMIXTURE for each sample individual, and the estimated ancestries attained an R^2 of 0.999, the same as PC-AiR.

Recent work [6,30] has shown that principal components that are calculated with reference panels can be used to predict ancestry proportions in unrelated samples. We used the top principal component from PC-AiR and the methodology described in Chen et al. [6] to predict ancestry proportions for the simulated admixed sample individuals in the presence of relatedness. Figure 3.7 illustrates that the predicted ancestries with PC-AiR were nearly identical to the true simulated ancestries, with an R^2 of 0.999. In fact, the predicted ancestry proportions calculated with PC-AiR were more accurate than those from the supervised ADMIXTURE analysis ($R^2 = 0.973$) that directly estimated proportional ancestry for all sample individuals. Despite the fact that PC-AiR can provide accurate ancestry inference without reference population samples, we only recommend ancestry proportion prediction from principal components when they are used. We make this recommendation because the ancestry

proportion estimates found from principal components calculated without reference population samples will be biased if there are not individuals in the sample with 100% ancestry from each of the underlying populations.

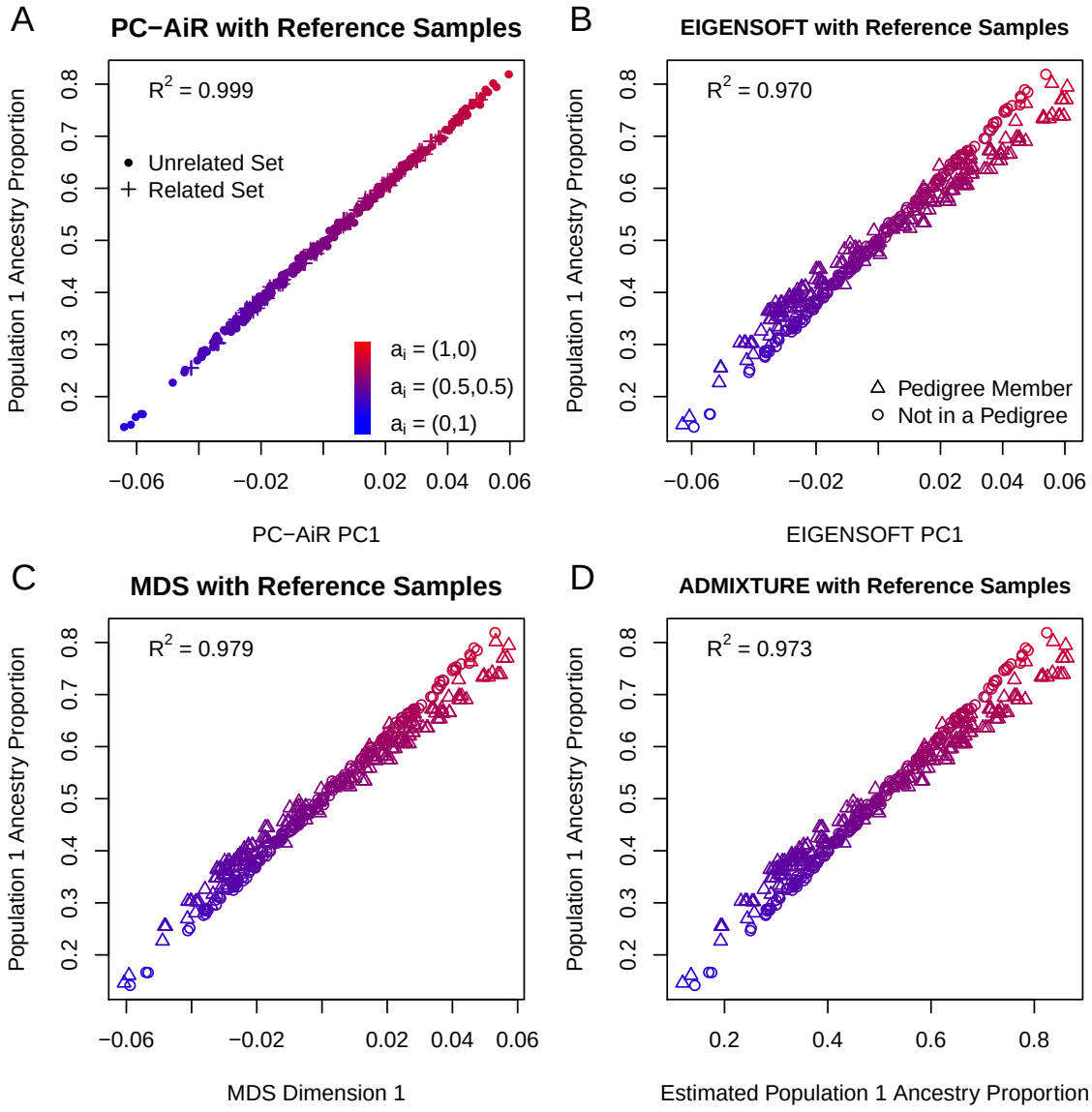
3.3.4 Population Structure Inference in Admixed HapMap Samples

HapMap MXL Data

We analyzed high-density genotype data from the Mexican Americans in Los Angeles, California (MXL) population sample of HapMap 3 for population structure inference. We applied PC-AiR, EIGENSOFT, MDS, ADMIXTURE, and FamPCA to the 86 genotyped individuals, and we compared the population structure inference results of these methods to a supervised individual ancestry estimation analysis with ADMIXTURE that included continental reference population panels. For the supervised analysis with ADMIXTURE, the number of ancestral populations was set to 3, for which the HapMap CEU (Utah residents with ancestry from northern and western Europe from the Centre d'Etude du Polymorphisme Human collection) and YRI (Yoruba in Ibadan, Nigeria) samples were included as the reference population panels for European and African ancestry, respectively, and for which the Human Genome Diversity Project (HGDP) [26] samples from the Americas were included for Native American ancestry. The analyses were based on 150,872 autosomal SNPs that were genotyped in both the HapMap and HGDP datasets. To protect against potential confounding due to relatedness in the supervised ancestry analysis, a separate ADMIXTURE analysis was conducted for each of the HapMap MXL individuals, where each analysis included a single HapMap MXL individual and the reference population panels. All methods, except for FamPCA, were provided only the SNP genotype data on the sample individuals for population structure inference, without any additional information on the pedigree relationships. The FamPCA method was also provided the documented pedigrees in the HapMap MXL which includes 24 genotyped trios, 5 families with two genotyped individuals, and 4 families with a single genotyped

Figure 3.6: Population Structure Inference Results including Reference Panels for Relationship Configuration I and Population Structure II with $\mathbf{F_{ST}} = \mathbf{0.1}$

Simulated population 1 ancestry proportions for each individual are plotted against: (A) coordinates along principal component 1 from PC-AiR, (B) coordinates along principal component 1 from EIGENSOFT, (C) coordinates along dimension 1 from MDS, and (D) the estimated ancestry proportions from ADMIXTURE for the inferred population with the highest R^2. The color of each point represents that individual's true ancestry; red for population 1, blue for population 2, and an intermediate color for an admixed individual. (A) A dot represents an individual in the ancestry representative, mutually unrelated set, and a plus represents an individual in the related set. (B-D) A circle represents an individual not in a pedigree, and a triangle represents an individual who is a member of a pedigree.


A
PC−AiR with Reference Samples
R² = 0.999
Unrelated Set
Related Set
a_i = (1,0)
a_i = (0.5,0.5)
a_i = (0,1)
Population 1 Ancestry Proportion
PC−AiR PC1
B
EIGENSOFT with Reference Samples
R² = 0.970
Pedigree Member
Not in a Pedigree
Population 1 Ancestry Proportion
EIGENSOFT PC1
C
MDS with Reference Samples
R² = 0.979
Population 1 Ancestry Proportion
MDS Dimension 1
D
ADMIXTURE with Reference Samples
R² = 0.973
Population 1 Ancestry Proportion
Estimated Population 1 Ancestry Proportion

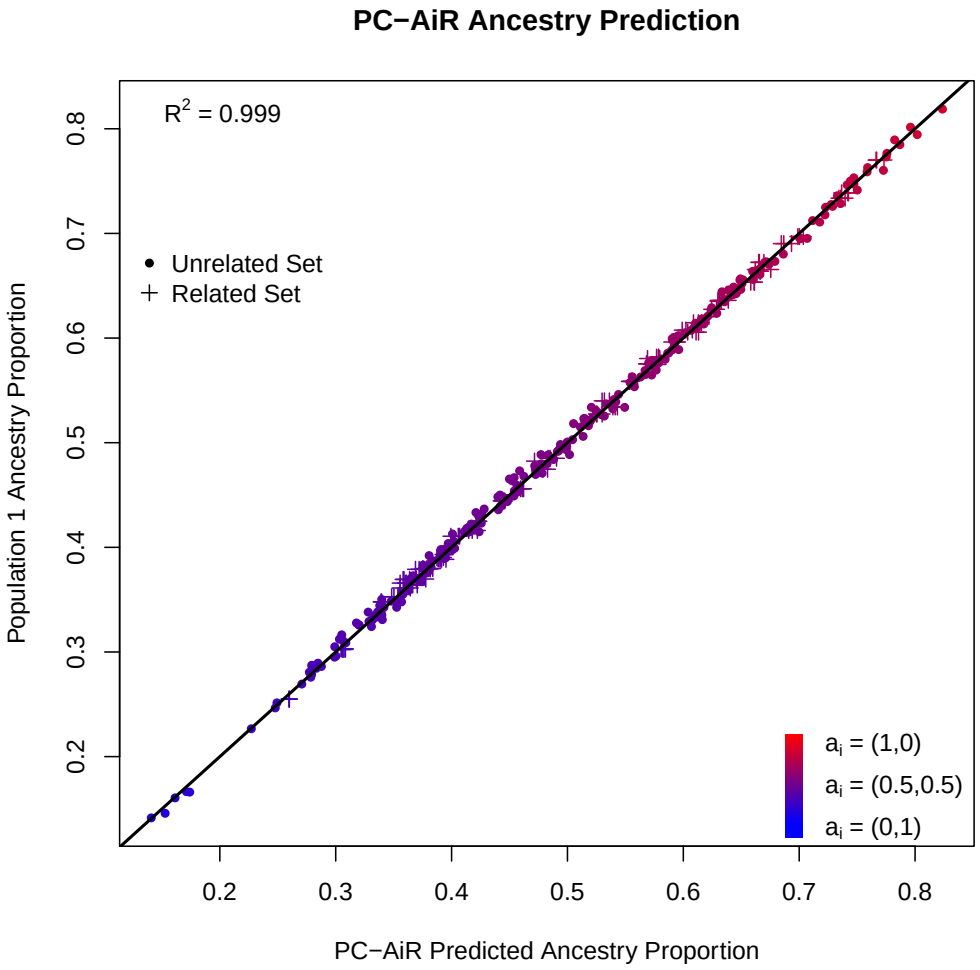


Figure 3.7: Ancestry Proportion Prediction using PC-AiR with Reference Panels for Relationship Configuration I and Population Structure II with $F_{ST} = 0.1$

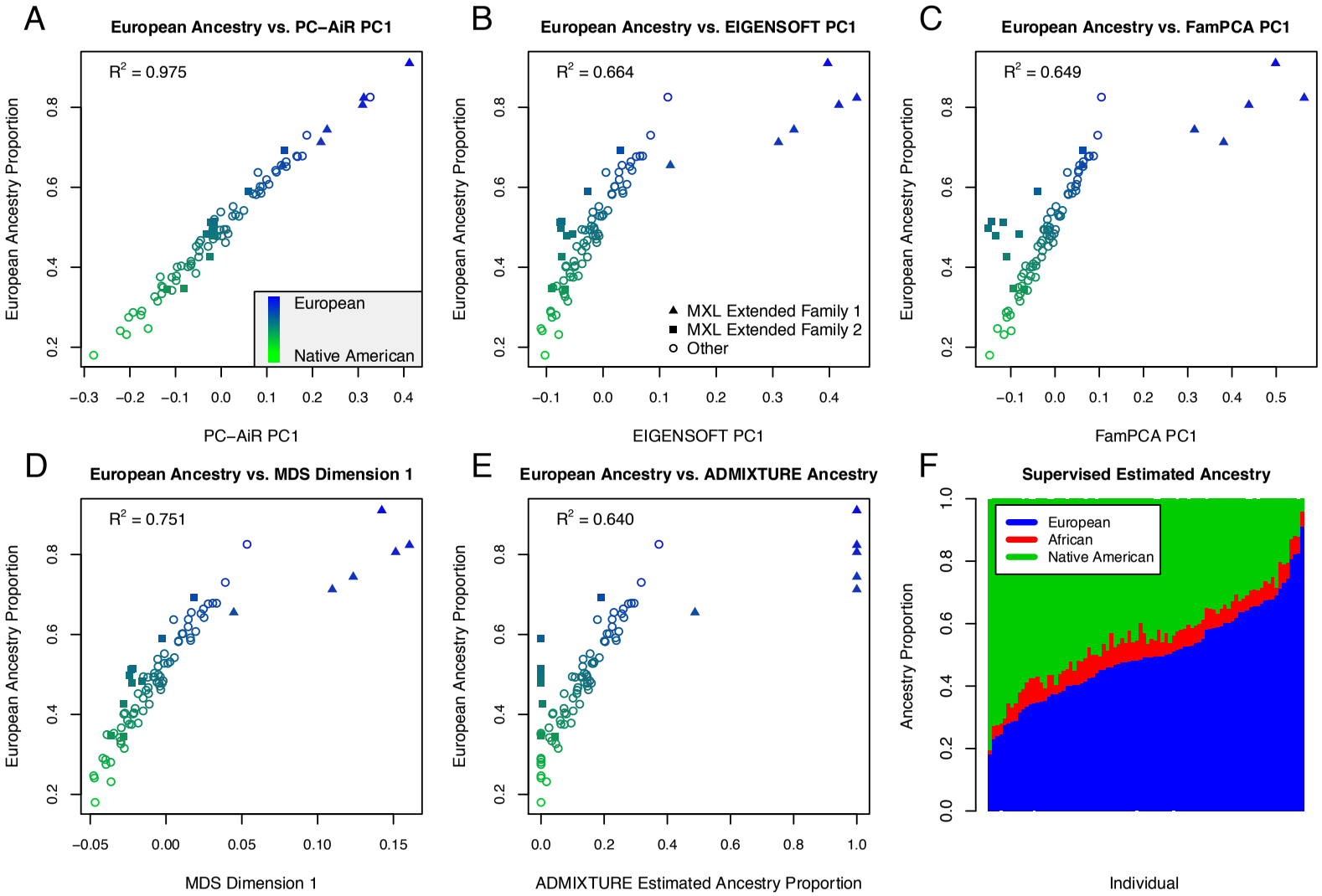
Scatter plot of the simulated population 1 ancestry proportions for each individual against predicted population 1 ancestry proportions using the top principal component from PC-AiR and the methodology given by Chen et al [10]. The color of each point represents that individual's true ancestry; red for population 1, blue for population 2, and an intermediate color for an admixed individual. A dot represents an individual in the ancestry representative, mutually unrelated set, and a plus represents an individual in the related set.

individual. The PC-AiR method used the KING-robust kinship coefficient estimator in Equation (3.1) and the relatedness threshold $\tau_\phi = 0.025$ to infer genetic relatedness in the sample, and a MAF filter of 5% was used on SNPs for population structure inference.

Figure 3.8F presents a bar plot of the results from the supervised individual ADMIXTURE ancestry analysis. In the bar plot of ancestry proportion estimates, individuals (vertical bars) are arranged in increasing order (left to right) of genome-wide European ancestry proportion. Our proportional ancestry estimates were similar to the results from a previous supervised analysis of this data [14, 56]. HapMap MXL individuals have modest African ancestry with little variation, with a mean of 6% and a standard deviation (SD) of 1.8%. The sample individuals are largely derived from European and Native American ancestry, with means of 49.9% (SD=14.8%) and 44.1% (SD=14.8%) respectively. Since the European and Native American ancestry proportions are predominant, nearly perfectly negatively correlated (with a correlation of -0.99), and quite variable, ranging from 18.0% to 91.0% and from 4.2% to 80.4% respectively, we expected that an optimal population structure inference method would require only a single axis of variation to explain these two ancestries in the HapMap MXL.

The population structure inference results for European and Native American ancestry in the HapMap MXL are given in Table 3.2. PC-AiR's top axis of variation was nearly perfectly correlated with European (and Native American) ancestry, as estimated from the supervised individual ADMIXTURE ancestry analysis, with an R^2 of 0.98 (Figure 3.8A). In contrast, the top axis of variation from each of EIGENSOFT, FamPCA, and MDS had an R^2 for European ancestry of only 0.66, 0.65, and 0.75 respectively. For the unsupervised ADMIXTURE analysis that did not include reference panels, the highest R^2 for either European or Native American ancestry with any estimated ancestry component was only 0.64. Figure 3.8B, 3.8C, 3.8D, and 3.8E illustrate that ancestry inference in the HapMap MXL for each of these compet-

Figure 3.8: Comparison of Population Structure Inference for the HapMap MXL Sample
Scatter plots of the European ancestry proportions estimated from a supervised individual ancestry analysis with ADMIXTURE for each individual are plotted against: (A) coordinates along principal component 1 from PC-AiR, (B) coordinates along principal component 1 from EIGENSOFT, (C) coordinates along principal component 1 from FamPCA, (D) coordinates along dimension 1 from MDS, and (E) the estimated ancestry proportions from an unsupervised analysis with ADMIXTURE for the inferred population with the highest R^2. The color of each point represents that individual's ancestry as estimated from the supervised individual ancestry analysis with ADMIXTURE; blue for European, green for Native American, and an intermediate color for an admixed individual. Individuals who are members of MXL Extended Family 1 or 2 are plotted as triangles or squares, respectively, and remaining individuals are plotted as circles. (F) Individual ancestry estimates for 86 HapMap MXL samples from a supervised individual ancestry analysis with ADMIXTURE. Each individual is represented by a vertical bar; estimated European (HapMap CEU), African (HapMap YRI), and Native American (HGDP samples from the Americas) ancestry proportions are shown in blue, red, and green, respectively.


A
European Ancestry vs. PC−AiR PC1
R^2 = 0.975
European Ancestry Proportion
PC−AiR PC1
European
Native American
B
European Ancestry vs. EIGENSOFT PC1
R^2 = 0.664
EIGENSOFT PC1
MXL Extended Family 1
MXL Extended Family 2
Other
C
European Ancestry vs. FamPCA PC1
R^2 = 0.649
FamPCA PC1
D
European Ancestry vs. MDS Dimension 1
R^2 = 0.751
MDS Dimension 1
E
European Ancestry vs. ADMIXTURE Ancestry
R^2 = 0.640
ADMIXTURE Estimated Ancestry Proportion
F
Supervised Estimated Ancestry
Ancestry Proportion
Individual
European
African
Native American

ing methods was confounded by relatedness, including the FamPCA method, which was provided the documented pedigree relationships. Ancestry inference with FamPCA was confounded by cryptic relatedness present in the HapMap MXL including a previously reported [56] extended pedigree consisting of two smaller documented pedigrees, which we have labeled in Figure 3.8 as MXL Extended Family 1. Without being provided any pedigree information, a single axis of variation from PC-AiR gave better prediction of both European and Native American ancestry than the top ten axes from EIGENSOFT, MDS, and FamPCA, as shown in Table 3.2. Remarkably, the top axis of variation from PC-AiR without using any reference population samples gave comparable ancestry inference on European and Native American ancestry to a supervised ancestry analysis that included reference panels, similar to the results from the simulation studies.

Combined HapMap ASW and MXL Data

To evaluate the performance of the population structure inference methods in an admixed population structure setting with three predominant continental ancestries and relatedness, we considered an analysis of the combined HapMap ASW (African American individuals in the southwestern USA) and MXL samples. Similar to our ancestry estimation analysis of the HapMap MXL, we also conducted a supervised individual ADMIXTURE analysis for the 87 genotyped individuals in the HapMap ASW with reference population panels included for European, Native American, and African ancestries. Figure 3.9A shows a barplot of the results from the supervised individual ADMIXTURE ancestry analysis of the HapMap MXL and ASW samples, which illustrates that these populations have very different ancestral backgrounds. Most of the HapMap ASW ancestry is African, with a mean of 77.5% (SD=8.4%). There is also a large European ancestry component, with a mean of 20.5% (SD=7.9%); however, unlike the HapMap MXL, there is very little Native American ancestry in the HapMap ASW, with a mean of only 1.9% (SD=3.5%). Since there are three

Table 3.2: Population Structure Inference Results for HapMap MXL and ASW

		R^2 Values				
Ancestry	d^a	PC-AiR	EIGENSOFT^b	MDS^c	FamPCA^d	ADMIXTURE^e
MXL Sample						
European	1	0.975	0.664	0.751	0.649	0.640
	4	-	0.914	0.935	0.969	-
	10	-	0.924	0.943	0.970	-
Native American	1	0.977	0.661	0.748	0.651	0.633
	4	-	0.908	0.929	0.968	-
	10	-	0.911	0.932	0.969	-
MXL + ASW Sample						
European	2	0.988	0.858	0.892	0.862	0.615
	4	-	0.868	0.899	0.878	-
	10	-	0.963	0.970	0.987	-
Native American	2	0.995	0.953	0.962	0.951	0.866
	4	-	0.958	0.967	0.961	-
	10	-	0.987	0.989	0.996	-
African	2	0.999	0.996	0.997	0.997	0.990
	4	-	0.996	0.997	0.997	-
	10	-	0.999	0.999	0.999	-

Population structure inference results from each method were compared to ancestry estimates from a supervised individual ancestry analysis with ADMIXTURE including reference population panels.

[a] d denotes the number of axes of variation included as predictors in the linear regression model to determine the R^2 value for each of the methods.

[b] PCA was performed with the EIGENSOFT software.

[c] MDS was implemented in the PLINK software.

[d] FamPCA is the Zhu et al. [73] method as implemented in the KING [31] software.

[e] An unsupervised ADMIXTURE analysis was conducted without including reference population panels. FRAPPE results were identical to ADMIXTURE.

predominant continental ancestries in the combined HapMap ASW and MXL samples, we expected that an optimal method would require two axes of variation to fully explain continental population structure.

We applied each of the dimension reduction methods (i.e. PC-AiR, EIGENSOFT, MDS, and FamPCA) to the combined HapMap ASW and MXL samples and compared the results to the supervised individual ancestry analysis with ADMIXTURE that included the reference population panels; results are shown in Table 3.2. All of the methods were able to fully explain the African ancestry with two axes of variation, achieving R^2 values greater than 0.99. For European ancestry, PC-AiR's top two axes of variation achieved an R^2 value of 0.99, while the top two axes from each of the competing population structure methods had R^2 values less than 0.90. With an R^2 value greater than 0.99, PC-AiR's top two axes of variation also explained Native American ancestry better than the top two axes from EIGENSOFT, MDS, and FamPCA, with corresponding R^2 values of 0.95, 0.96, and 0.95, respectively. These results are illustrated in Figure 3.10, where we can see that the top two axes of variation from each of these methods, except PC-AiR, were confounded by relatedness. In fact, the top ten axes of variation from EIGENSOFT, MDS, and FamPCA were highly confounded by pedigree structure, whereas axes beyond the top two from PC-AiR did not represent any identifiable structure and appear to be noise (Figures 3.11 - 3.15). As a consequence, the top ten axes of variation from both EIGENSOFT and MDS were not able to explain European and Native American ancestry as well as the top two axes from PC-AiR. Intersetingly, FamPCA required ten axes of variation to match PC-AiR's top two, despite FamPCA being provided the documented pedigree information for both the HapMap MXL and ASW samples (Table 3.2). PC-AiR appropriately accounted for both the known and cryptic relatedness in the sample for optimal and efficient inference on ancestry with only two axes of variation.

We also performed an unsupervised ancestry analysis with ADMIXTURE and FRAPPE without including reference panel samples and we compared the results

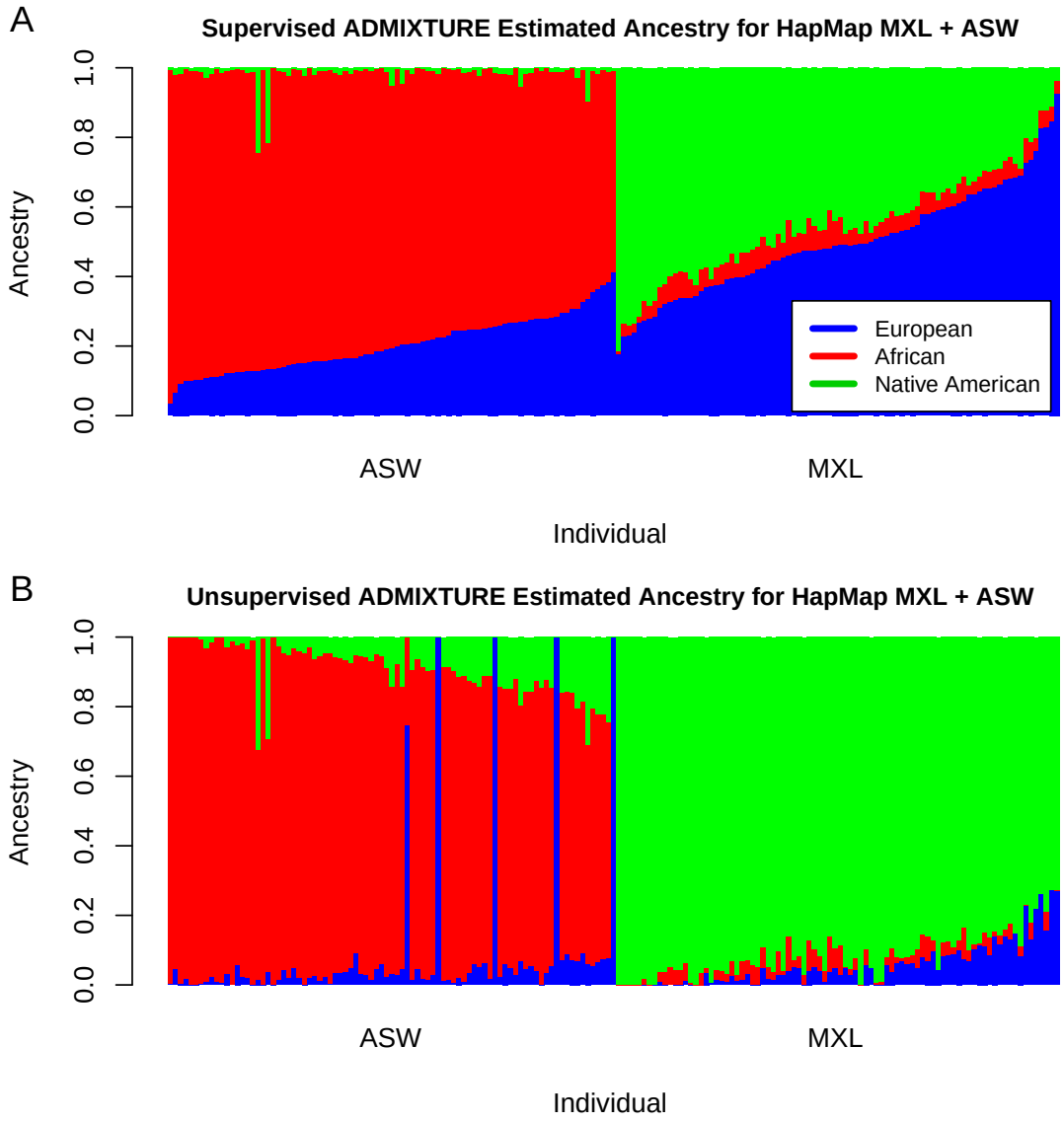


Figure 3.9: HapMap MXL and ASW Individual Ancestry Bar Plots
Individual ancestry estimates for 87 HapMap ASW and 86 HapMap MXL samples. (A) Supervised individual ancestry analysis with ADMIXTURE including reference population panels. Each individual is represented by a vertical bar; estimated European (HapMap CEU), African (HapMap YRI), and Native American (HGDP samples from the Americas) ancestry proportions are shown in blue, red, and green, respectively. (B) Unsupervised ancestry analysis with ADMXITURE without reference poplulation panels. The three colors represent the three inferred ancestral populations.

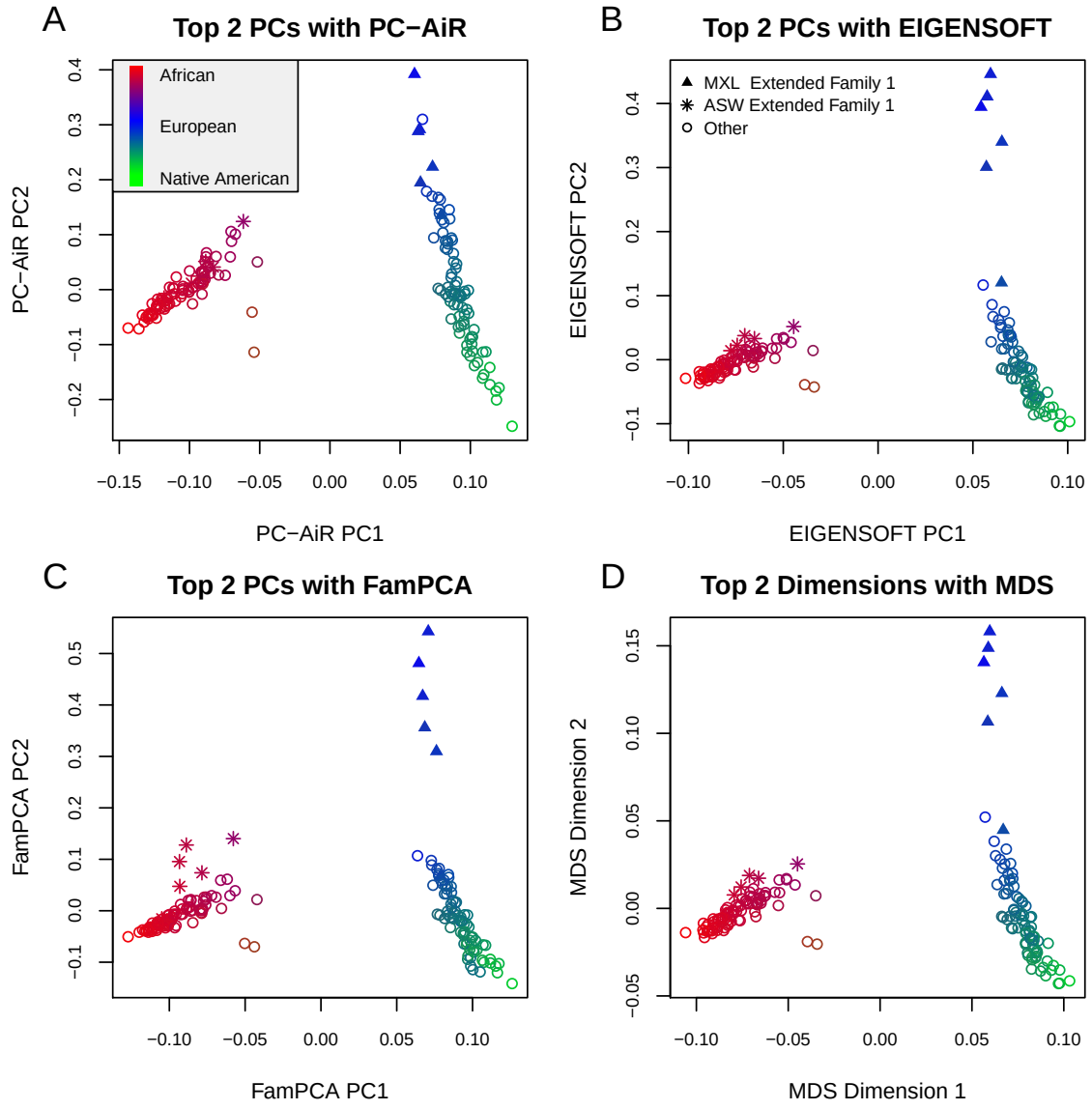


Figure 3.10: Comparison of Population Structure Inference for the HapMap MXL and ASW Combined Sample

Scatter plots of the top two axes of variation from PC-AiR (A), EIGENSOFT (B), FamPCA (C), and MDS (D). The color of each point represents that individual's ancestry as estimated from a supervised individual ancestry analysis with ADMIXTURE; blue for European (HapMap CEU), red for African (HapMap YRI), green for Native American (HGDP samples from the Americas), and an intermediated color for an admixed individual. Individuals who are members of MXL Extended Family 1 or ASW Extended Family 1 are plotted as triangles or stars, respectively, and remaining individuals are plotted as circles.

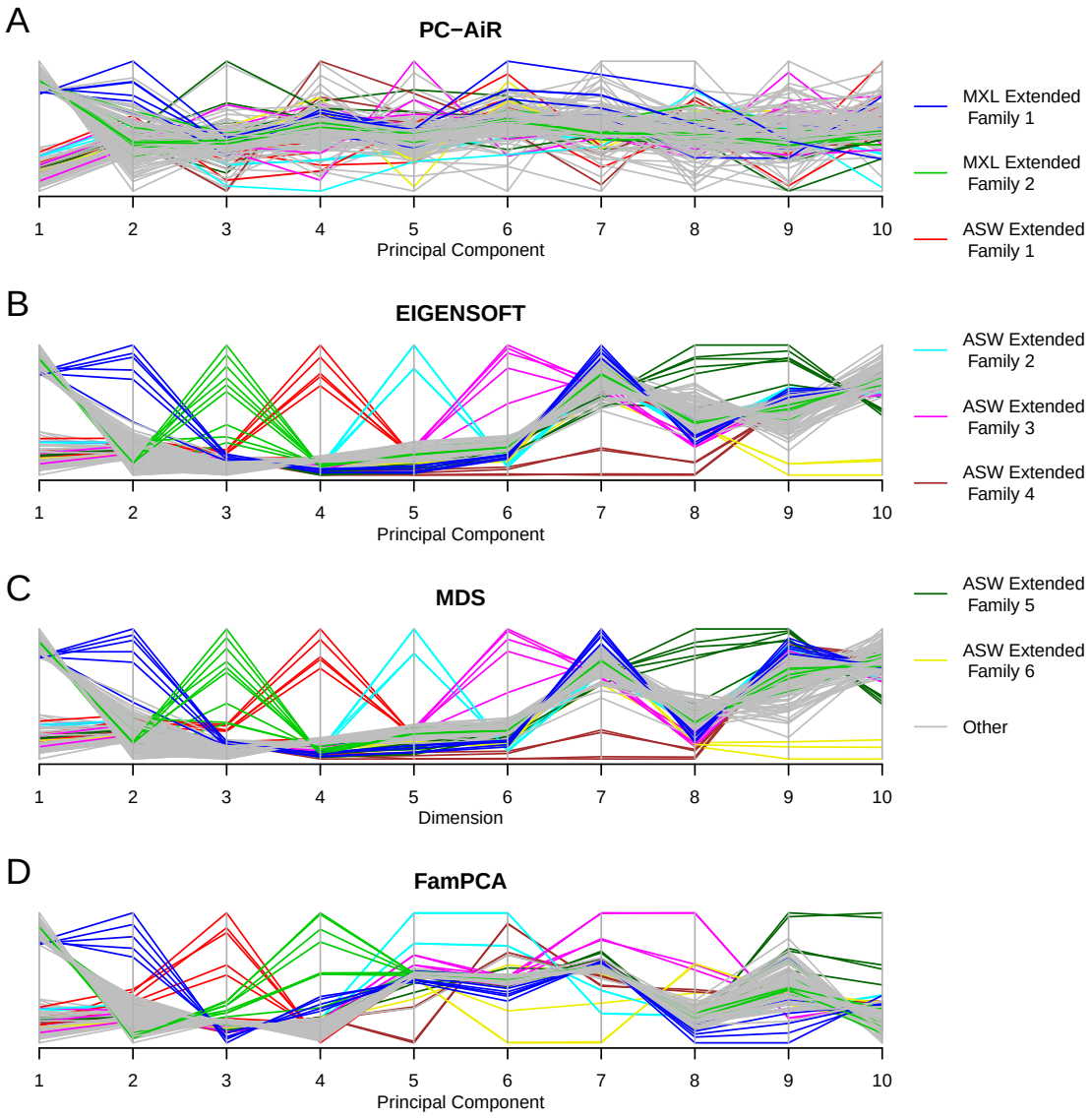


Figure 3.11: Parallel Coordinates Plots for the HapMap MXL and ASW Combined Sample

Parallel coordinates plots of the top ten axes of variation from PC-AiR (A), EIGENSOFT (B), MDS (C), and FamPCA (D). Each vertical bar represents one of the axes of variation, and each line traces out the coordinates for an indivdiual across all ten axes of variation. Colors are used to show individuals belonging to the same extended family. Many of the top ten axes of variation from all methods except PC-AiR are driven by the correlation of members in extended families.

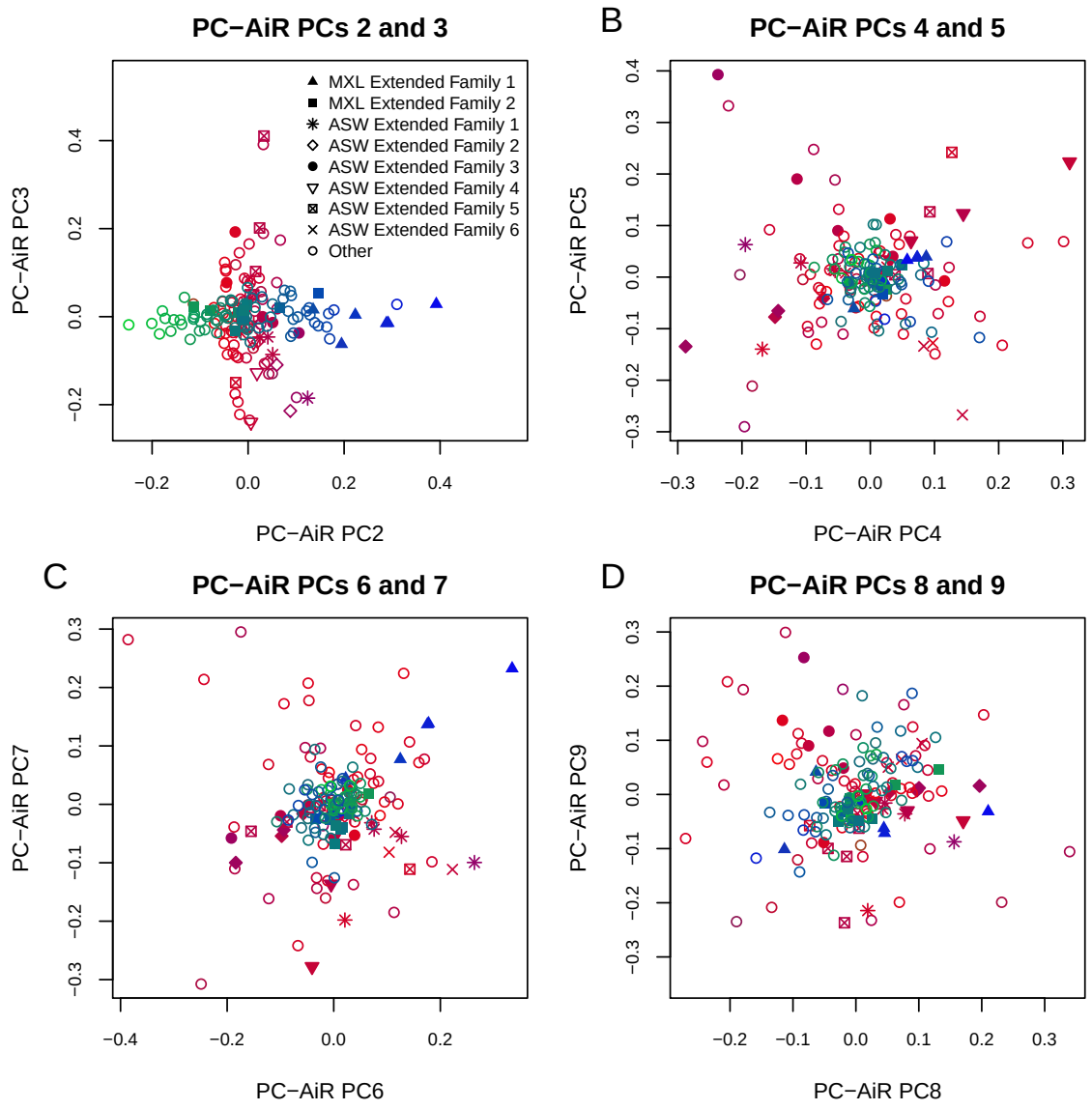


Figure 3.12: PC-AiR PCs 2-9 from HapMap MXL and ASW Combined Sample

Scatterplots of principal components 2-9 from PC-AiR. The color of each point represents that individual's ancestry as estimated from a supervised individual ancestry analysis with ADMIXTURE; blue for European (HapMap CEU), red for African (HapMap YRI), green for Native American (HGDP samples from the Americas), and an intermediated color for an admixed individual. Different plotting characters represent individuals belonging to different extended pedigrees. PC 2 reflects European vs. Native American ancestry, but PCs 3-9 do not reflect any apparent structure.

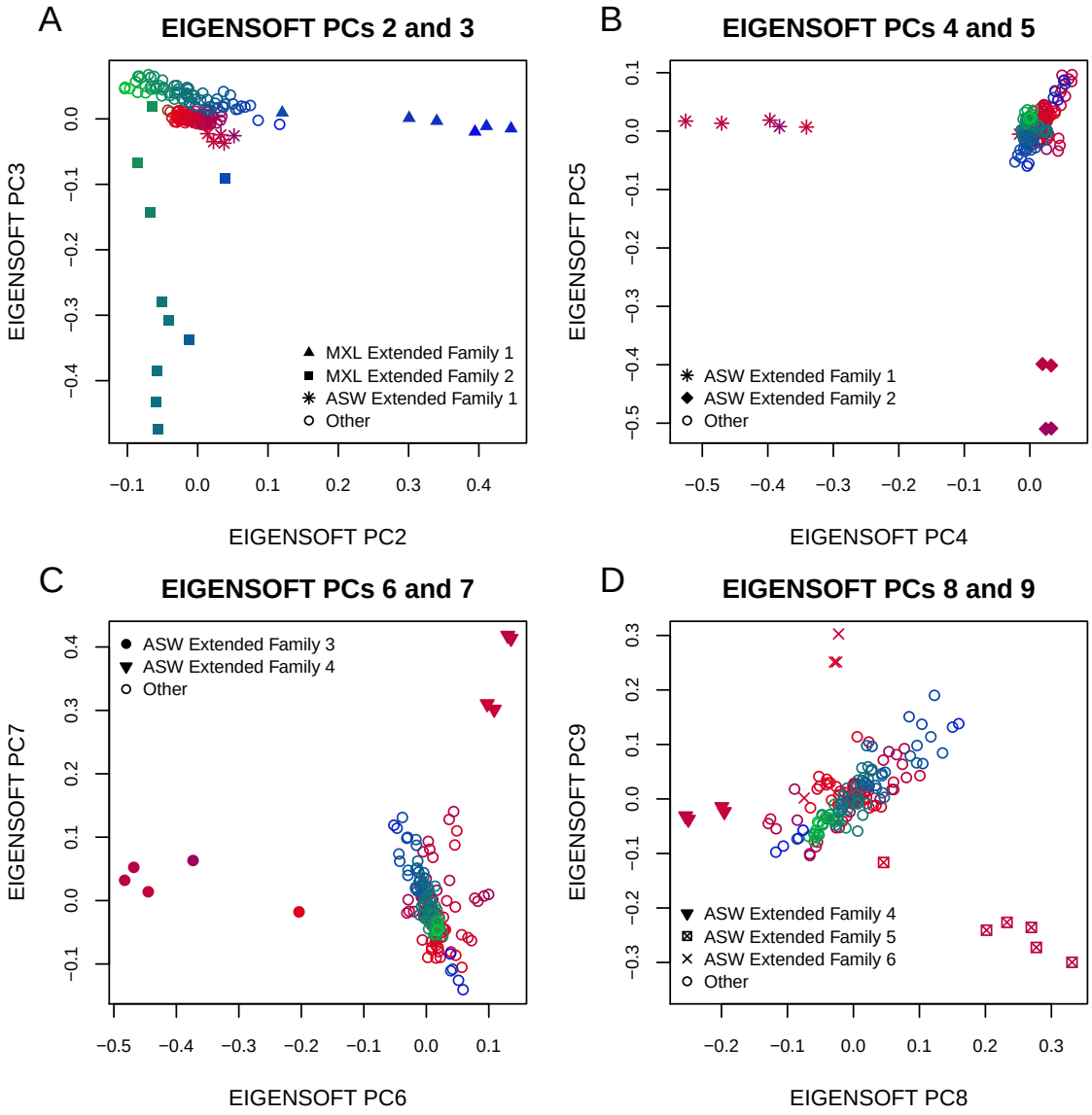


Figure 3.13: EIGENSOFT PCs 2-9 from HapMap MXL and ASW Combined Sample

Scatterplots of principal components 2-9 from EIGENSOFT. The color of each point represents that individual's ancestry as estimated from a supervised individual ancestry analysis with ADMIXTURE; blue for European (HapMap CEU), red for African (HapMap YRI), green for Native American (HGDP samples from the Americas), and an intermediated color for an admixed individual. Different plotting characters represent individuals belonging to different extended pedigrees. All axes of variation are strongly influenced by the correlation structure of groups of relatives belonging to extended families.

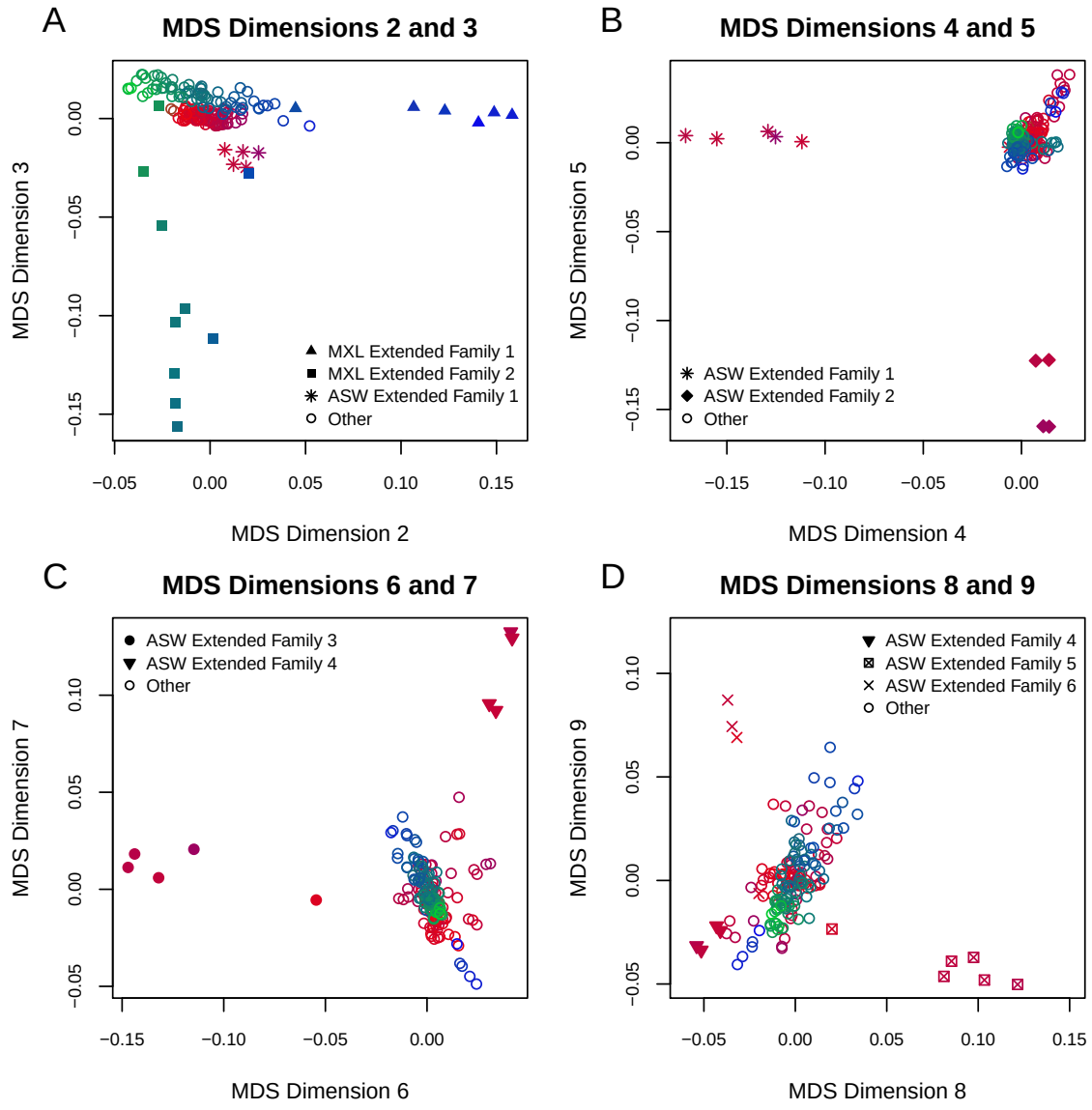


Figure 3.14: MDS Dimensions 2-9 from HapMap MXL and ASW Combined Sample

Scatterplots of dimensions 2-9 from MDS. The color of each point represents that individual's ancestry as estimated from a supervised individual ancestry analysis with ADMIXTURE; blue for European (HapMap CEU), red for African (HapMap YRI), green for Native American (HGDP samples from the Americas), and an intermediated color for an admixed individual. Different plotting characters represent individuals belonging to different extended pedigrees. All axes of variation are strongly influenced by the correlation structure of groups of relatives belonging to extended families.

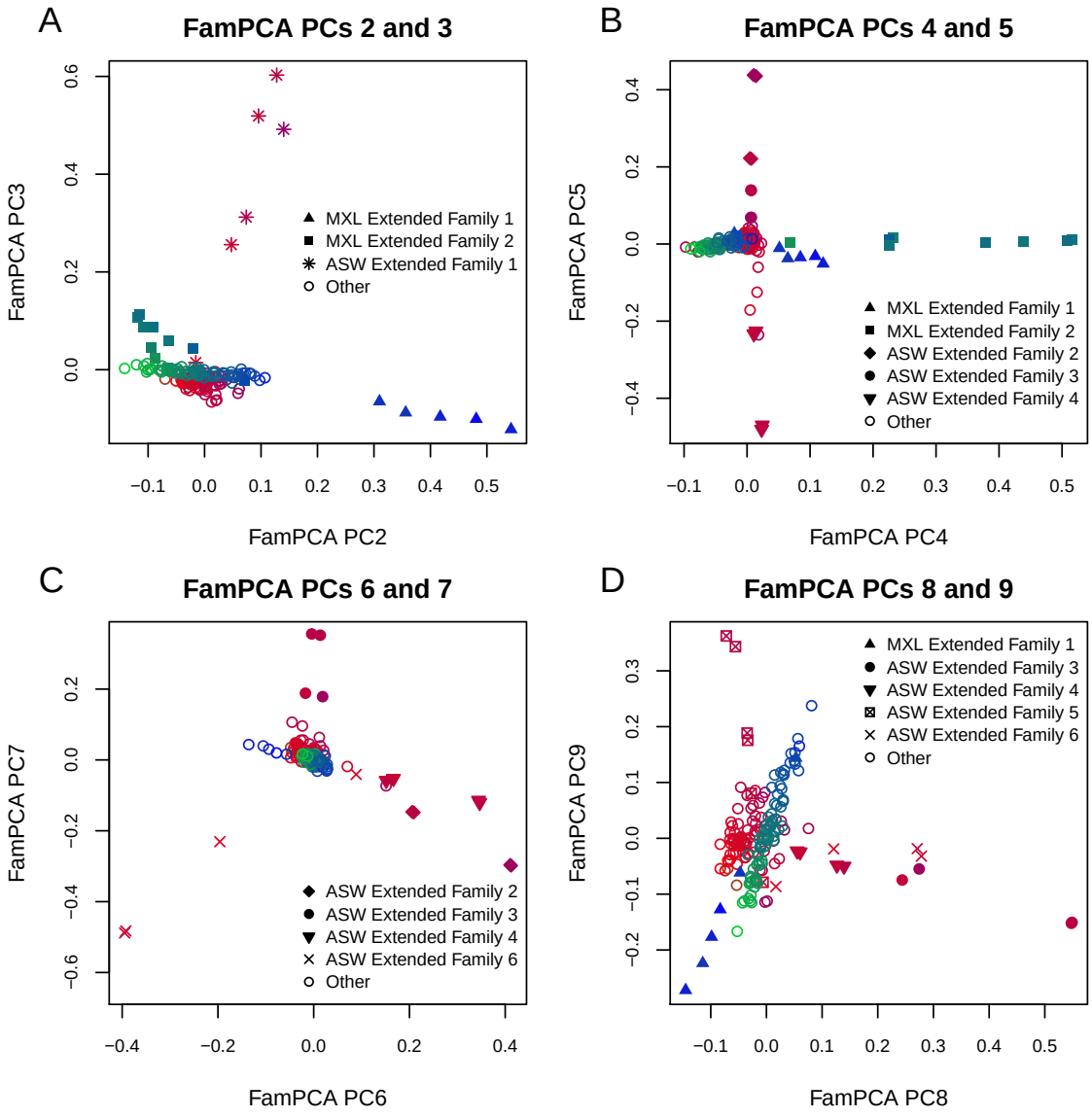


Figure 3.15: FamPCA PCs 2-9 from HapMap MXL and ASW Combined Sample

Scatterplots of PCs 2-9 from FamPCA. The color of each point represents that individual's ancestry as estimated from a supervised individual ancestry analysis with ADMIXTURE; blue for European (HapMap CEU), red for African (HapMap YRI), green for Native American (HGDP samples from the Americas), and an intermediated color for an admixed individual. Different plotting characters represent individuals belonging to different extended pedigrees. All axes of variation are strongly influenced by the correlation structure of groups of relatives belonging to extended families.

to the supervised ADMIXTURE analysis. ADMIXTURE and FRAPPE performed identically to each other, as expected, and a barplot of the estimated ancestry proportions from the unsupervised ancestry analysis is given in Figure 3.9B. Two of the three components of ancestry essentially distinguish the ASW from the MXL samples, while the third was completely confounded by pedigree structure. This estimated ancestry components were able to attain an R^2 value of 0.99 for African ancestry, but the R^2 values were only 0.87 for Native American ancestry and 0.62 for European ancestry, thus performing the worst of all the methods for ancestry inference in the combined HapMap MXL and ASW samples.

3.3.5 Assessment of Computation Time

The computation time for PC-AiR depends on both the sample size and the number of markers being analyzed. To analyze a simulated sample of 800 individuals, where 400 individuals are from 20 pedigrees and the remaining 400 individuals are unrelated, with 100K, 50K, and 20K SNPs required 28.5s, 14.8s, and 6.3s, respectively, on a 2.5 GHz laptop with 8 GB memory. The PC-AiR analysis of the HapMap data with 150,872 SNPs required 1.8s for the MXL sample with 86 individuals and 3.9s for the combined ASW and MXL sample with 173 individuals. All computation times refer to the time to run the PC-AiR algorithm, and do not include the time to estimate the measure of relatedness and divergence. The KING software implements a highly efficient algorithm for obtaining relatedness/divergence estimates, and evaluating millions of pairs of individuals in a sample can be conducted in a matter of minutes.

3.4 Discussion

Genetic ancestry inference has been motivated by a variety of applications in population genetics, genetic association studies, and other genomic research areas. Advancements in array-based genotyping technologies have largely facilitated the investigation

of genetic diversity at remarkably high levels of detail, and a variety of methods have been proposed for the identification of genetic ancestry differences among unrelated sample individuals using high-density genome-screen data. It is common, however, for genetic studies to have sample structure that is due to both population stratification and relatedness, and existing population structure inference methods can fail in related samples. We develop PC-AiR, a method for robust population structure inference in the presence of known or cryptic relatedness. PC-AiR applies a computationally efficient algorithm that uses pairwise measures of kinship and ancestry divergence from genome-screen data for the identification of a diverse subset of mutually unrelated individuals that is representative of the ancestries in the entire sample. Principal components that are representative of ancestry are obtained by performing PCA directly on genotype data from individuals selected for the ancestry representative subset, while coordinates along the axes of variation for the remaining individuals in the sample are predicted based on genetic similarities with the diverse subset. The PC-AiR method does not require the genealogy of the sampled individuals to be known, and it can be used across a variety of study designs, ranging from population based studies where individuals are assumed to be unrelated to family based studies with partially or completely unknown pedigrees.

In simulation studies with a broad range of population structure settings, including ancestry admixture, and with sample individuals related according to a variety of genealogical configurations, we demonstrated that the top axes of variation from PC-AiR were nearly perfectly correlated with ancestry. In contrast, widely used methods for population structure inference performed poorly in the presence of relatedness, including the PCA method implemented in the EIGENSOFT software, MDS as implemented in PLINK software, and model-based ancestry estimation methods ADMIXTURE and FRAPPE. We also applied PC-AiR and competing methods to the admixed HapMap MXL and ASW population samples. Without using any reference population panels or pedigree information on the sample individuals, the top two axes

of variation from PC-AiR nearly perfectly explained proportional European, Native American, and African ancestry in the HapMap MXL and ASW samples as compared to a supervised individual ancestry analysis with ADMIXTURE that included reference population panels. In contrast, all other population structure inference methods were confounded by relatedness, including the FamPCA method which was provided the documented pedigree relationships.

Performing PCA with genome-wide SNP weights that are calculated from external reference panels has recently been proposed [6] for certain admixed populations. This approach requires prior knowledge about the ancestries of the individuals in the sample, which may be partially or completely unknown, as well as having available reference panels that are adequate surrogates for ancestry. Nevertheless, the PC-AiR method can also easily incorporate SNP-weights from external reference panels for population structure inference. For example, by designating population samples from external reference panels to be the ancestry representative subset in the PC-AiR algorithm, principal components for individuals in the target sample for population structure inference will be calculated based solely on SNP weights from the reference panels. A potential limitation of using SNP weights from external reference panels, however, is that inference on population structure will be limited to the ancestries of individuals selected from the panels, which may not be representative of the ancestries of all individuals in the sample. An attractive alternative approach would be to perform a PC-AiR analysis on the study sample combined with the external reference panels, where genome-screen data would be used by the algorithm implemented in PC-AiR for the identification of an ancestry representative subset from the combined set of individuals, and where ancestries from both the reference panels and the sample will be allowed to contribute to the SNP weights.

The challenges of inferring genetic ancestry in related samples have been well documented [40, 42]. To our knowledge, PC-AiR is the first method to provide robust population structure inference in the presence of known or cryptic relatedness

without requiring reference population panels, external SNP loadings, or genealogical information on the sample individuals.

In Chapter 4 we utilize the ancestry representative PCs from PC-AiR to develop a set of method of moments estimators that accurately estimate genetic relatedness measures for all pairs of sample individuals in presence of unspecified population structure. In Chapter 6 we demonstrate the limitations of existing linear mixed model (LMM) methods for genetic association mapping in admixed populations, and we show that including ancestry representative principal components from PC-AiR as fixed effects in LMMs can address these limitations and provide well-calibrated association test statistics at all SNPs.

Chapter 4

PCA-BASED RELATEDNESS ESTIMATION FOR ADMIXED POPULATIONS WITH UNSPECIFIED STRUCTURE

4.1 Introduction

Genetic relatedness between pairs of individuals is important to many areas of genetic research including forensics, population genetics, medical genomics, and genetic association studies. Relatedness among sample individuals in genetic studies is common due to the inclusion of pedigrees of known structure or due to cryptic relatedness between individuals who are assumed to be unrelated but actually share a common ancestor. Reliable relatedness inference from genome-screen data allows for confirmation and error detection of reported pedigree relationships, as well as the identification of unreported cryptic relationships in both population and pedigree based studies. Furthermore, accurate characterization of the correlation structure of sample individuals due to genetic relatedness is essential for valid and optimal association testing.

Until recently, existing methods for estimating genetic relatedness measures focused on homogeneous populations with no underlying ancestral diversity. These methods, which include both maximum likelihood estimators [36, 55] and method of moments estimators such as PLINK [44] and the standard genetic relationship matrix (GRM) [64], work well for relatively homogeneous populations, but they have been found to be severely biased in the presence of population structure [31, 38, 56]. In recent years, the study of more diverse populations has created the need for accurate relatedness estimation in structured populations and has led to the development of new methods including KING-robust [31], REAP [56], and RelateAdmix [37]. Each of these methods provides a substantial improvement over methods that assume popula-

tion homogeneity, but each has its limitations. For example, KING-robust is designed for populations with discrete substructure and has been shown to provide biased estimates for pairs of individuals with different ancestry [56]; a problem for inference in admixed populations. REAP and RelateAdmix both provide accurate estimates in admixed populations, but they require prior knowledge of the ancestries present in the sample as well as available reference population panels with individuals that are appropriate proxies for these ancestral populations.

Here we present a new principal components analysis (PCA) based method, PC-Relate, for estimating measures of genetic relatedness in the presence of unspecified population structure, including ancestry admixture. PC-Relate uses ancestry representative principal components to estimate expected individual-specific allele frequencies at each SNP for every individual, conditional on their autsome-wide ancestry. These individual-specific allele frequency estimates are then used in a set of method of moments estimators to obtain relatedness estimates. We demonstrate through simulation that PC-Relate provides accurate estimation in the presence of population structure, including discrete substructure and varying patterns of ancestry admixture, while other commonly used estimators, such as the standard GRM, PLINK, and KING-robust, provide systematically biased estimates when their assumptions are violated. Additionally, through analysis of real genotype data from the Hispanic cohort of the Women's Health Initiative SNP Health Association Resource (WHI-SHARe) study, we show that PC-Relate provides comparable estimates to REAP and RelateAdmix, but without any prior assumptions about the underlying population structure and without requiring any additional reference panel samples. Furthermore, PC-Relate is computationally efficient and is applicable to large studies with genome-wide SNP data and many thousands of individuals.

4.2 Methods

4.2.1 Expectation of the Genetic Relationship Matrix

For each $i \in \mathcal{N}$ and $s \in \mathcal{S}$, let the random variable g_{is} be the number of copies of the reference allele that individuals i has at SNP s; thus, g_{is} takes values of 0, 1, or 2 and has unconditional expectation $\mathbb{E}[g_{is}] = 2p_s$. For a pair of individuals i and j, we define the quantity ψ_{ij} to be the unconditional correlation between g_{is} and g_{js} under the assumption that the population from which $\mathcal{N}$ is sampled is in Hardy-Weinberg Equilibrium (HWE), i.e. when the variance of g_{is} is $\text{Var}[g_{is}] = 2p_s(1-p_s)$. The correlation structure is assumed to be the same across all SNPs, so a standard estimator of ψ_{ij} is

$$\widehat{\psi}_{ij} = \frac{1}{|\mathcal{S}_{ij}|} \sum_{s \in \mathcal{S}_{ij}} \frac{(g_{is} - 2\hat{p}_s)(g_{js} - 2\hat{p}_s)}{2\hat{p}_s(1-\hat{p}_s)}, \tag{4.1}$$

where $\mathcal{S}_{ij}$ is the subset of SNPs for which both i and j have non-missing genotype data, $|\mathcal{S}_{ij}|$ is the number of SNPs in this subset, and $\hat{p}_s$ is some estimator of the population allele frequency at SNP s. The $\widehat{\psi}_{ij}$ estimates given by Equation (4.1) are the elements of what is often referred to as the genetic relationship matrix (GRM) [64]. Note that even if the HWE assumption does not hold, ψ_{ij} is still a scaled measure of the genetic covariance between individuals i and j. This matrix and the elements within it are often used for inference on genetic relatedness and/or population structure.

Consider a sample from a structured population following the general population genetic modeling assumptions discussed in Section 2.5. Under the assumption that genotypes at different SNPs are independent, and with $|\mathcal{S}_{ij}| \to \infty$, when the true p_s is known for each SNP $s \in \mathcal{S}_{ij}$, it can be shown that (see Appendix A.2.1)

$$\frac{1}{2}\widehat{\psi}_{ij} \to \phi_{ij}[1 - b_{\psi_1}(i,j)] + \mathbf{a}_i^T \mathbf{\Sigma}_K \mathbf{a}_j, \tag{4.2}$$

where $b_{\psi_1}(i,j)$ is a function of the ancestry vectors for i and j's set of most recent common ancestors and the covariance structure, $\mathbf{\Sigma}_K$, of the subpopulation-specific allele frequencies. In the special case of discrete population substructure, $b_{\psi_1}(i,j) =$

$\mathbf{a}_i^T \mathbf{\Sigma}_K \mathbf{a}_j$ when i and j are from the same subpopulation (see Appendix A.2.1 for details). What we see from Equation (4.2) is that $\widehat{\psi}_{ij}$ is consistent for ϕ_{ij} only in a homogeneous population, where $K = 1$ and the random vector $\mathbf{p}_s = p_s$ is a degenerate random variable with $\mathbf{\Sigma}_K = 0$. This estimator measures the genetic covariance between individuals i and j relative to the underlying ancestral population, which includes covariance due to both recent family structure (kinship) and distant population structure. This realization suggests an estimator that conditions on each individual's ancestry to remove the contribution of population structure. This is the motivation behind the REAP method, where the genotype values in Equation (4.1) are centered and standardized by individual-specific allele frequencies, calculated conditionally on each individual's autosome-wide ancestral background, rather than population average allele frequencies.

4.2.2 Estimating Individual-Specific Allele Frequencies with Principal Components

As in Thornton et al. (2012) [56], we consider the expectation of g_{is} conditional on individual i's ancestry, $\mathbf{a}_i$, and the vector of subpopulation-specific allele frequencies at SNP s, $\mathbf{p}_s$, and we define the quantity

$$\mu_{is} \equiv \frac{1}{2}\mathbb{E}[g_{is}|\mathbf{a}_i, \mathbf{p}_s] = \mathbf{a}_i^T \mathbf{p}_s = \sum_{k=1}^{K} a_i^k p_s^k. \tag{4.3}$$

This μ_{is} can be interpreted as the individual-specific allele frequency for individual i at SNP s, as it is a linear combination of the subpopulation-specific allele frequencies weighted by individual i's autosomal ancestry proportions from each of the ancestral subpopulations. In Thornton et al. (2012) [56], estimates of individual-specific allele frequencies are found directly with Equation (4.3) by plugging in estimates of $\mathbf{p}_s$ and $\mathbf{a}_i$ obtained from model-based ancestry estimation methods such as STRUCTURE [43], FRAPPE [54], or ADMIXTURE [3]. However, these model-based methods may not be suitable for some studies as they require (1) some prior knowledge about the ancestries that are likely present in the sample and (2) appropriate reference panels

with suitable surrogates for each of the ancestral subpopulations. If the number of ancestral subpopulations, K, is mis-specified, or if suitable reference panels are not used, then the estimates obtained may be severely biased.

Here we propose an alternative method for estimating individual-specific allele frequencies that does not require any modeling assumptions or reference population samples. We accomplish this through the use of ancestry representative principal components, such as those calculated from the genome-screen data with PC-AiR (Chapter 3). Let $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_D]$ be a matrix with column vectors that are the top D principal components required to fully explain the population structure present in $\mathcal{N}$, and let $\mathbf{g}_s$ be the vector of genotype values for each individual at SNP s. As long as $\mathbf{V}$ captures all of the population structure in the sample, the expectation of $\mathbf{g}_s$ conditional on this set of principal components is equivalent to its expectation conditional on the set of true ancestry vectors. Therefore, we propose estimating individual-specific allele frequencies at each SNP $s \in \mathcal{S}$ with the fitted values from the linear regression model $\mathbb{E}[\mathbf{g}_s|\mathbf{V}] = \mathbf{1}\beta_0 + \mathbf{V}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a length D vector of regression coefficients for each of the D principal components. The estimate of μ_{is} for each individual $i \in \mathcal{N}$ at SNP s is calculated as

$$\widehat{\mu}_{is} = \frac{1}{2}\widehat{\mathbb{E}}[g_{is}|V_{i1}, \ldots, V_{iD}] = \frac{1}{2}\left(\widehat{\beta}_0 + \sum_{d=1}^{D} \widehat{\beta}_d V_{id}\right), \tag{4.4}$$

where V_{id} is the coordinate for individual i along principal component d. Since each of the principal components has mean 0, $\widehat{\beta}_0$ is twice the sample average allele frequency, an estimator of $2p_s$, and each of the estimators $\widehat{\beta}_d$ for $d \in \{1, \ldots, D\}$ can be interpreted as a measure of the change in allele frequency due to having a shift in ancestry away from the sample average in the direction represented by $\mathbf{V}_d$.

4.2.3 Estimating Kinship in a Structured Population

The PC-Relate kinship coefficient estimator,

$$\widehat{\phi}_{ij}^{PC} = \frac{\sum_{s \in \mathcal{S}_{ij}} (g_{is} - 2\widehat{\mu}_{is})(g_{js} - 2\widehat{\mu}_{js})}{4\sum_{s \in \mathcal{S}_{ij}} \sqrt{\widehat{\mu}_{is}(1 - \widehat{\mu}_{is})\widehat{\mu}_{js}(1 - \widehat{\mu}_{js})}}, \tag{4.5}$$

is a modification of the standard GRM. This estimator uses genotype values centered and scaled by estimates of individual-specific allele frequencies calculated as in Equation (4.4) to remove genetic covariance due to the population structure represented by these principal components. A further modification of the estimator is taking the ratio of the averages of the numerator and the denominator terms, essentially a weighted average over loci, as opposed to the average of the ratios of these terms as in Equation (4.1), an unweighted average over loci [46, 61]. The ratio of averages estimator is better behaved as it increases stability and reduces sampling variability, especially when SNPs with low minor allele frequencies are included [5]. Under the assumption that genotypes at different SNPs are independent, and with $|\mathcal{S}_{ij}| \to \infty$, when the true values of μ_{is} and μ_{js} are known, it can be shown that

$$\widehat{\phi}_{ij}^{PC} \to \phi_{ij}[1 - b_{\phi_1}(i,j)], \tag{4.6}$$

where $b_{\phi_1}(i,j)$ is a function of $\mathbf{\Sigma}_K$ and the ancestry vectors for individuals i, j, and their set of most recent common ancestors. What we can see from this result is that $\widehat{\phi}_{ij}^{PC}$ provides consistent estimates of 0 for unrelated pairs of individuals, regardless of the underlying population structure. Furthermore, in the presence of discrete population substructure, $b_{\phi_1}(i,j) = 0$, and $\widehat{\phi}_{ij}^{PC}$ also provides consistent estimates of kinship coefficients for relatives (see Appendix A.2.2 for mathematical details). While we can not claim consistency for relative pairs in all population structure settings, we demonstrate in our simulation studies (Section 4.2.8) that the bias is very small and that the method allows for accurate inference of pedigree relationships, even for distant relatives in settings with ancestry admixture from highly divergent populations.

4.2.4 An Alternative Genotype Coding

For each $i \in \mathcal{N}$ and $s \in \mathcal{S}$, let the random variable g_{is}^{D} be an alternative genotype coding that is constructed to be orthogonal to the traditional additive genotype coding,

g_{is}, under HWE; i.e. $\text{Cov}[g_{is}, g_{is}^D] = 0$. We refer to g_{is}^D as the dominance genotype coding, and it is presented in Table 4.1. For now, consider an outbred homogeneous population where $\mu_{is} = p_s$, the population allele frequency, for every $i \in \mathcal{N}$. In this scenario, the coding of g_{is}^D is the same as one presented in Vitezica et al. [58] up to a shift and re-scaling. Under the HWE assumption, g_{is}^D has expectation $\mathbb{E}[g_{is}^D] = p_s(1-p_s)$ and variance $\text{Var}[g_{is}^D] = [p_s(1-p_s)]^2$. Analogous to ψ_{ij}, we define the quantity δ_{ij} to be the unconditional correlation between g_{is}^D and g_{js}^D, and a reasonable estimator is given by

$$\widehat{\delta}_{ij} = \frac{1}{|\mathcal{S}_{ij}|} \sum_{s \in \mathcal{S}_{ij}} \frac{[g_{is}^D - \hat{p}_s(1-\hat{p}_s)][g_{js}^D - \hat{p}_s(1-\hat{p}_s)]}{[\hat{p}_s(1-\hat{p}_s)]^2}. \tag{4.7}$$

The $\widehat{\delta}_{ij}$ estimates are the elements of what can be interpreted as to as a dominance genetic relationship matrix, and they have been shown to be good estimates of the coefficient of fraternity, $k_{ij}^{(2)}$ [11]. In fact, under the same convergence assumptions as presented for $\widehat{\psi}_{ij}$, when individuals have genotype counts in expected HW proportions, it can be shown that $\widehat{\delta}_{ij}$ is a consistent estimator for $k_{ij}^{(2)}$ (see Appendix A.5.1).

Table 4.1: Genotype Codings for Individual i at SNP s

Genotype	Genotype Coding	
	Additive: g_{is}	Dominance: g_{is}^D
AA	2	$(1-\widehat{\mu}_{is})$
Aa	1	0
aa	0	$\widehat{\mu}_{is}$

μ_{is} is the frequency of the A allele for individual i at SNP s. In a homogeneous population, $\mu_{is} = p_s$ for every $i \in \mathcal{N}$.

4.2.5 Estimating IBD Sharing Probabilities in a Structured Population

Similar to the properties of $\widehat{\psi}_{ij}$, when the sample $\mathcal{N}$ is from a structured population, $\widehat{\delta}_{ij}$ also reflects additional covariance structure due to the underlying population structure. To remove the contribution of population structure, we again propose conditioning on each individual's ancestry with the use of individual-specific allele frequencies. The adjusted PC-Relate estimator is

$$\widehat{\delta}_{ij}^{PC} = \frac{\sum_{s \in \mathcal{S}_{ij}} [g_{is}^{D} - \widehat{\mu}_{is}(1 - \widehat{\mu}_{is})(1 + \widehat{f}_{i}^{PC})][g_{js}^{D} - \widehat{\mu}_{js}(1 - \widehat{\mu}_{js})(1 + \widehat{f}_{j}^{PC})]}{\sum_{s \in \mathcal{S}_{ij}} \widehat{\mu}_{is}(1 - \widehat{\mu}_{is})\widehat{\mu}_{js}(1 - \widehat{\mu}_{js})}, \tag{4.8}$$

where $\widehat{f}_{i}^{PC}$ is the inbreeding coefficient estimate for individual i from the PC-Relate estimator presented in Equation (4.12) of the following subsection, but is used here to account for departures from HWE. The improvement in stability and decrease in sampling variance from taking the ratio of the averages is particularly noticeable for this estimator, especially when including SNPs with low minor allele frequency, because the denominator consists of terms with allele frequencies raised to the fourth power. The convergence results for $\widehat{\delta}_{ij}^{PC}$ are very similar to those presented for $\widehat{\phi}_{ij}^{PC}$. This estimator provides consistent estimates of 0 for unrelated pairs of individuals in all population structure settings, and it provides consistent estimates of $k_{ij}^{(2)}$ for outbred relatives in the presence of discrete population substructure (see Appendix A.5.2 for mathematical details). Once again, while we can not claim consistency of this estimator in all admixed populations, simulations show that the bias tends to be small, and we therefore define $\widehat{k}_{ij}^{(2)^{PC}} \equiv \widehat{\delta}_{ij}^{PC}$.

PC-Relate also implements estimators for the remaining IBD sharing probabilities in structured populations. Two estimators of $k_{ij}^{(0)}$ are used in combination as we have found in practice that each one has lower sampling variability for particular relationship types. For pairs of individuals with estimated kinship coefficients consistent with values for first degree relatives, an estimator that is a function of the number of opposite homozygotes and individual-specific allele frequencies is used [31, 56]. For

pairs of individuals with estimated kinship coefficients less than those of first degree relatives, an estimate is calculated as a function of the estimators $\widehat{\phi}_{ij}^{PC}$ and $\widehat{k}_{ij}^{(2)^{PC}}$ using the identities $k_{ij}^{(0)} + k_{ij}^{(1)} + k_{ij}^{(2)} = 1$ and $\phi_{ij} = \frac{1}{2}k_{ij}^{(2)} + \frac{1}{4}k_{ij}^{(1)}$. The PC-Relate estimator is therefore

$$\widehat{k}_{ij}^{(0)^{PC}} = \begin{cases} \frac{\sum_{s\in\mathcal{S}_{ij}} \mathbb{1}_{[|g_{is}-g_{js}|=2]}}{\sum_{s\in\mathcal{S}_{ij}}\left[\hat{\mu}_{is}^2(1-\hat{\mu}_{js})^2+(1-\hat{\mu}_{is})^2\hat{\mu}_{js}^2\right]} & \text{if} \quad \widehat{\phi}_{ij}^{PC} > 2^{-5/2} \approx 0.177 \\ 1 - 4\widehat{\phi}_{ij}^{PC} + \widehat{k}_{ij}^{(2)^{PC}} & \text{if} \quad \widehat{\phi}_{ij}^{PC} < 2^{-5/2} \approx 0.177 \end{cases}. \tag{4.9}$$

The final IBD sharing probablity, $k_{ij}^{(1)}$, is also found from the identities above and is simply estimated as $\widehat{k}_{ij}^{(1)^{PC}} = 1 - \widehat{k}_{ij}^{(0)^{PC}} - \widehat{k}_{ij}^{(2)^{PC}}$.

4.2.6 Estimating Inbreeding Coefficients in the Presence of Population Structure

We now focus on individuals sampled from a population with inbreeding. Numerous inbreeding coefficient estimators have been proposed for homogeneous populations, and the estimator used by GCTA has been shown [65] to have smaller sampling variance than both the estimator of self kinship, $\widehat{\psi}_{ii}$, and the estimator based on excess homozygosity implemented in PLINK. It is interesting to notice that the GCTA estimator, which is a complicated function of the additive genotype coding, g_{is}, can be more simply expressed in terms of our dominance genotype coding as

$$\widehat{f}_i = \frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}\left[\frac{g_{is}^D}{\hat{p}_s(1-\hat{p}_s)}\right] - 1, \tag{4.10}$$

where $|\mathcal{S}_i|$ is the number of SNPs in $\mathcal{S}_i$, the subset of SNPs for which individual i has non-missing genotype data. When population homogeneity is incorrectly assumed, under the assumption that genotypes at different SNPs are independent, and when the true values of p_s are known, it can be shown that

$$\widehat{f}_i \to f_i[1 - \mathbf{a}_{M(i)}^T\mathbf{\Sigma}_K\mathbf{a}_{P(i)}] + \mathbf{a}_{M(i)}^T\mathbf{\Sigma}_K\mathbf{a}_{P(i)} \tag{4.11}$$

as $|\mathcal{S}_i| \to \infty$, where the indices $M(i)$ and $P(i)$ represent the mother and father of individual i, respectively. This estimator is consistent for f_i only in a homogeneous

population ($\mathbf{\Sigma}_K = 0$). PC-Relate provides an extension of this estimator to structured populations by conditioning on individual i's ancestry, once again through the use of individual-specific allele frequencies. When an individual is inbred, the conditional expectation of the dominance genotype coding is $\mathbb{E}[g_{is}^D|\mathbf{a}_i, \mathbf{p}_s] = \mu_{is}(1-\mu_{is})(1+f_i)$, so the PC-Relate inbreeding coefficient estimator is

$$\widehat{f}_i^{PC} = \frac{\sum_{s\in\mathcal{S}_i} g_{is}^D}{\sum_{s\in\mathcal{S}_i} \widehat{\mu}_{is}(1-\widehat{\mu}_{is})} - 1. \tag{4.12}$$

Similar to the other PC-Relate estimators, this estimator provides consistent estimates in the presence of discrete population substructure, and it typically has small bias in more general population structure settings (see Appendix A.4 for further details).

The parameter f_i can also be viewed as a measure of the departure of individual i's genotype counts from the expected HW proportions. A positive value of f_i indicates more homozygous genotypes than expected, and a negative value of f_i indicates more heterozygous genotypes than expected. Estimators that assume population homogeneity, such as the one presented in Equation (4.10), compare an individual's counts to expected counts based on population average allele frequencies, typically leading to inflated estimates of inbreeding. The PC-Relate estimator, on the other hand, compares an individual's counts to expected counts based on individual-specific allele frequencies, providing accurate estimates of inbreeding coefficients even in the presence of population structure. It is worth noting, however, that an individual who is the offspring of parents with different ancestries may have more heterozygous genotypes than expected under HW proportions based on their individual-specific allele frequencies, resulting in a negative f_i value. In fact, for an outbred individual, under our usual convergence assumptions,

$$\widehat{f}_i^{PC} \to \frac{-\frac{1}{4}(\mathbf{a}_{M(i)} - \mathbf{a}_{P(i)})^T \mathbf{\Sigma}_K (\mathbf{a}_{M(i)} - \mathbf{a}_{P(i)})}{1 - \mathbf{a}_i^T \mathbf{\Sigma}_K \mathbf{a}_i} \tag{4.13}$$

as $|\mathcal{S}_i| \to \infty$. This results shows a bias that is systematically negative for inbreeding coefficients, but is an accurate representation of excess heterozygosity from recent

admixture events. The magnitude of this value tends to be small unless $M(i)$ and $P(i)$ have very different ancestries, and while this bias is not ideal for estimating inbreeding coefficients, it does allow for the possible identification of individuals recently admixed from highly divergent populations.

4.2.7 Estimating Pairwise Relatedness Measures in Inbred Populations

In a population with inbreeding, careful consideration must be taken in estimating pairwise relatedness measures. There are no longer only three IBD states that a pair of individuals can take at a locus, but rather nine condensed IBD states as given by Jacquard [19]. The estimator $\widehat{\phi}_{ij}^{PC}$ is still an accurate estimator of the kinship coefficient, which can now be written as $\phi_{ij} = \Delta_{ij}^{(1)} + \frac{1}{2}\left(\Delta_{ij}^{(3)} + \Delta_{ij}^{(5)} + \Delta_{ij}^{(7)}\right) + \frac{1}{4}\Delta_{ij}^{(8)}$, where $\Delta_{ij}^{(l)}$ is the probability that individuals i and j are in Jacquard's l^{th} condensed IBD state at a locus. On the other hand, $\widehat{\delta}_{ij}^{PC}$ is no longer necessarily an estimator of the probability of sharing two alleles IBD and may be difficult to interpret (see Appendix A.5.3 for further details). Additionally, it has been shown [11] that the set of nine condensed IBD sharing probabilities are not estimable with biallelic SNP markers. However, as long as the expectation $\widehat{\mu}_{is}(1-\widehat{\mu}_{is})(1+\widehat{f}_i^{PC})$ is used to center genotype values in Equation (4.8), possible departures from HWE are accounted for, and $\widehat{\delta}_{ij}^{PC} \to 0$ for unrelated pairs, regardless of the underlying population structure or true inbreeding.

4.2.8 Simulation Studies

We perform simulation studies with related individuals in the presence of population structure in order to (1) assess the accuracy of the PC-Relate estimators for kinship coefficients, IBD sharing probabilities, and inbreeding coefficients in structured populations, and to (2) compare the performance of PC-Relate to existing methods including the PLINK method of moments and KING-robust estimators. We also consider the PC-Relate estimators under the assumption of population homogeneity,

where individual-specific allele frequencies are replaced by sample average allele frequencies. We refer to these unadjusted versions of the PC-Relate estimators as the homogeneous estimators. The homogeneous kinship coefficient estimator is a slight modification of the standard GRM given by Equation (4.1), the homogeneous $k^{(2)}$ estimator is a slight modification of the Garcia-Cortes et al. [11] estimator presented in Equation (4.7), and the homogeneous inbreeding coefficient estimator is a slight modification of the GCTA estimator in Equation (4.10). These homogeneous estimators and PLINK both assume population homogeneity, while KING-robust is designed for populations with discrete substructure. Note that we can only compare the performance of KING-robust for kinship coefficients as it does not provide IBD sharing probability estimates for structured populations. In each simulation, two principal components from PC-AiR are used to adjust the PC-Relate estimators, and SNPs with an estimated minor allele frequency below 5% are filtered.

Relationship Configuration

We consider two different relationship configurations in our simulation studies. For relationship configuration I, we generate 1000 individuals from 40 non-inbred four-generation pedigrees, where each pedigree consists of 25 individuals, including first- through fifth-degree relatives. For relationship configuration II, we generate 1000 individuals from 50 inbred pedigrees, where each pedigree consists of 20 individuals and includes a first cousin mating as well as a first cousin once removed mating, both with two offspring. The exact pedigree configurations are shown in Figures 4.1 and 4.2. For pedigree founders, ancestry vectors are generated under one of many population structure settings, described in the following subsection, and genotypes are generated independently at each SNP. For a given SNP, the population from which each allele is drawn is selected with probabilities equal to that founder's ancestry proportions, and the type of each allele is drawn from a Bernoulli distribution with probability given by the allele frequency at that SNP for the chosen population. Alleles are then

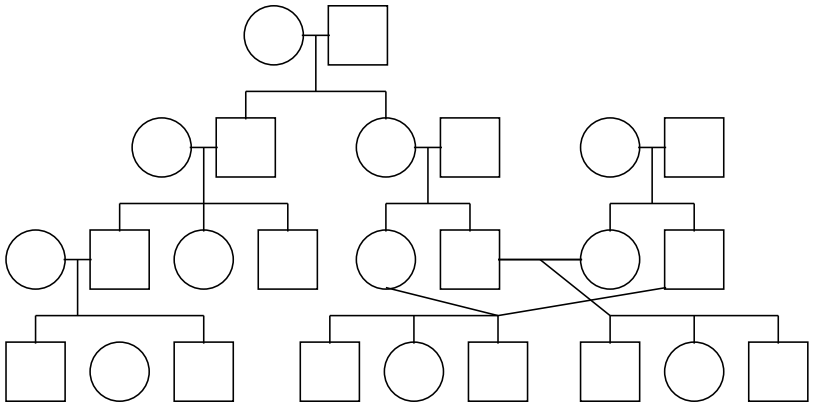


Figure 4.1: Pedigree Configuration for Simulations with Double First Cousins
The pedigree configuration for each of the 40 non-inbred four-generation pedigrees included in Relationship Configuration I of the simulation studies, where the overall structure of each pedigree is as depicted, but the pattern of ancestry admixture varies according to the specified population structure.

dropped down the pedigree to generate genotypes for descendants.

4.2.9 Population Structure Settings

We consider population structure settings where individuals have ancestry derived from three populations. The allele frequencies at 100,000 independent SNPs for each of these populations are generated using the Balding-Nichols model [4]. More precisely, for each SNP s, the ancestral allele frequency p_s is drawn from a uniform distribution on $[0.1, 0.9]$, and the allele frequency in population $k \in \{1, 2, 3\}$ is drawn from a beta distribution with parameters $p_s(1-F_k)/F_k$ and $(1-p_s)(1-F_k)/F_k$, where the quantity F_k is equivalent to Wright's measure of population differentiation [62] from the ancestral population. For each population structure setting, we simulate highly divergent populations representing continental differences, where the values of F_k are $F_1 = 0.05$, $F_2 = 0.15$, and $F_3 = 0.25$, leading to pairwise F_{ST} values of roughly

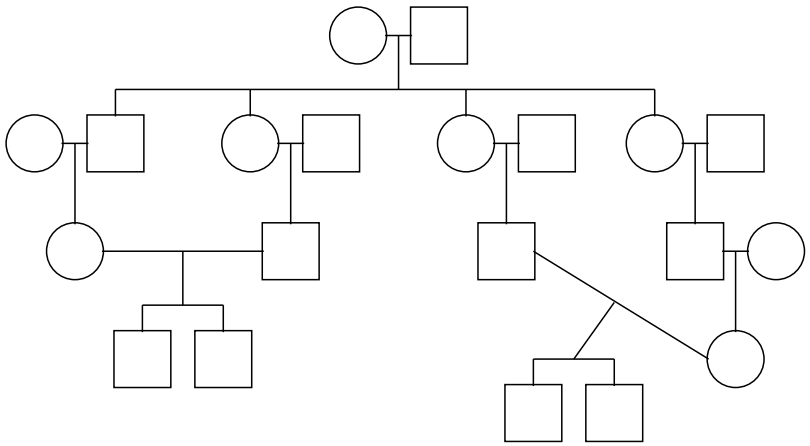


Figure 4.2: Pedigree Configuration for Simulations with Inbreeding
The pedigree configuration for each of the 50 inbred pedigrees included in Relationship Configuration II of the simulation studies, where the overall structure of each pedigree is as depicted, but the pattern of ancestry admixture varies according to the specified population structure.

0.10 between populations 1 and 2, 0.15 between populations 1 and 3, and 0.20 between populations 2 and 3. Population structures I and II both consist of individuals with admixed ancestry from the three populations. For population structure I, pedigree founders have ancestry vectors drawn from a Dirichlet(1,1,1) distribution, resulting in equal contributions of ancestry from each population on average. For population structure II, founders of half of the pedigrees have ancestry vectors drawn from a Dirichlet(6,2,0.25) distribution, resulting in mean ancestry proportions of 0.73, 0.24, and 0.03 from populations 1, 2, and 3 respectively. The roles of populations 1 and 2 are reversed for founders in the other half of the pedigrees, who have ancestry vectors drawn from a Dirichlet(2,6,0.25) distribution. Population structure III consists of non-admixed individuals, where an equal number of pedigrees are sampled from each of populations 1 and 2, and slightly fewer pedigrees are sampled from population 3. Population structure settings II and III result in ancestry assortative mating, as

the founder individuals in every pedigree have either the same (population structure III) or similar (population structure II) ancestry, while population structure I has completely random mating that may lead to relatives with very different ancestry.

4.3 Results

4.3.1 Admixed Populations

The relatedness estimates from each method were used to infer relationship types for all pairs of individuals according to criteria similar to that from Manichaikul et al. (2010) [31]. A pair of individuals was inferred to have a d^{th} degree relationship if $2^{-(d+3/2)} < \widehat{\phi}_{ij} < 2^{-(d+1/2)}$, where monozygotic twins were considered $d = 0$. Within first degree relationships, parent/offspring pairs were distinguished from full siblings for methods that provided IBD sharing probability estimates when $\widehat{k}_{ij}^{(0)} < 0.05$ and for KING-robust when the proportion of loci for which the pair shared zero alleles identical by state (IBS0) was less than 0.005. Double first cousins have an expected $k^{(2)}$ of 0.0625, so these were identified by methods that provided IBD sharing probability estimates as second degree relatives with $\widehat{k}_{ij}^{(2)} > 2^{(-9/2)} \approx 0.044$.

The relationship estimation results for relationship configuration I under population structure II for each of the methods considered are shown in Figure 4.3. The PC-Relate estimators provided accurate estimates with low variability and were able to correctly infer relationships for all pairs of individuals, even up to 5^{th} degree relatives. All other methods had biases that made it difficult to distinguish relationships at or below 3^{rd} degree. Figure 4.4A shows that the PC-Relate kinship coefficient estimates were accurate regardless of the ancestry of the pair of individuals. On the other hand, Figure 4.4B shows that KING-robust only provided accurate estimates for pairs of individuals with the same ancestry and had a negative bias that increased as the ancestry difference for the pair of individuals increased. This was most noticeable in distant relatives, where multiple generations of admixture led to quite different ancestry proportions, and in unrelated pairs with different ancestries, for which negative

kinship coefficient estimates were obtained. The homogeneous estimators and PLINK both provided inflated kinship coefficient estimates for pairs of individuals with similar ancestry, due to the additional genetic similarity, resulting in many unrelated pairs of individuals being identified as relatives (Figures 4.4C and 4.4D). Similar to KING-robust, these estimators were also deflated for pairs of individuals with very different ancestry, although PLINK not as badly due to the truncations it imposed. Figure 4.5 shows that PC-Relate also provided accurate, low variability, estimates of IBD sharing probabilities, while the estimators that assume population homogeneity tended to give inflated $k^{(2)}$ estimates (especially PLINK) and deflated $k^{(0)}$ estimates, both with much higher variability.

The relationship estimation results under population structure I are shown in Figures 4.6, 4.7, and 4.8. In this setting, mating was entirely random, possibly between individuals with very different ancestries, resulting in close relatives with very different ancestry proportions. Interestingly, the KING-robust kinship coefficient estimate for a pair of individuals may be either negatively or positively biased in this setting. In short, when the individuals have different ancestries, this contributes a negative bias, as seen in the results from population structure II. However, when either of the individuals are the offspring of recent admixture events, descending from parents with different ancestry from each other, this contributes a positive bias (see Appendix A.2.3 for further details). As a result, many of the KING-robust kinship estimates were negatively biased, similar to population structure II but more extreme, which resulted in 6 2^{nd} degree and 68 3^{rd} degree relative pairs being inferred as unrelated. Additionally, however, the estimates for many other pairs of individuals were positively biased. In fact, 51 unrelated pairs and 59 4^{th} degree relative pairs were inferred to be 3^{rd} degree relatives, while 6 additional 3^{rd} degree relative pairs were identified as 2^{nd} degree relatives.

In comparison, all of the PC-Relate estimators still performed well under population structure I, though the variability was slightly increased, and a slight negative

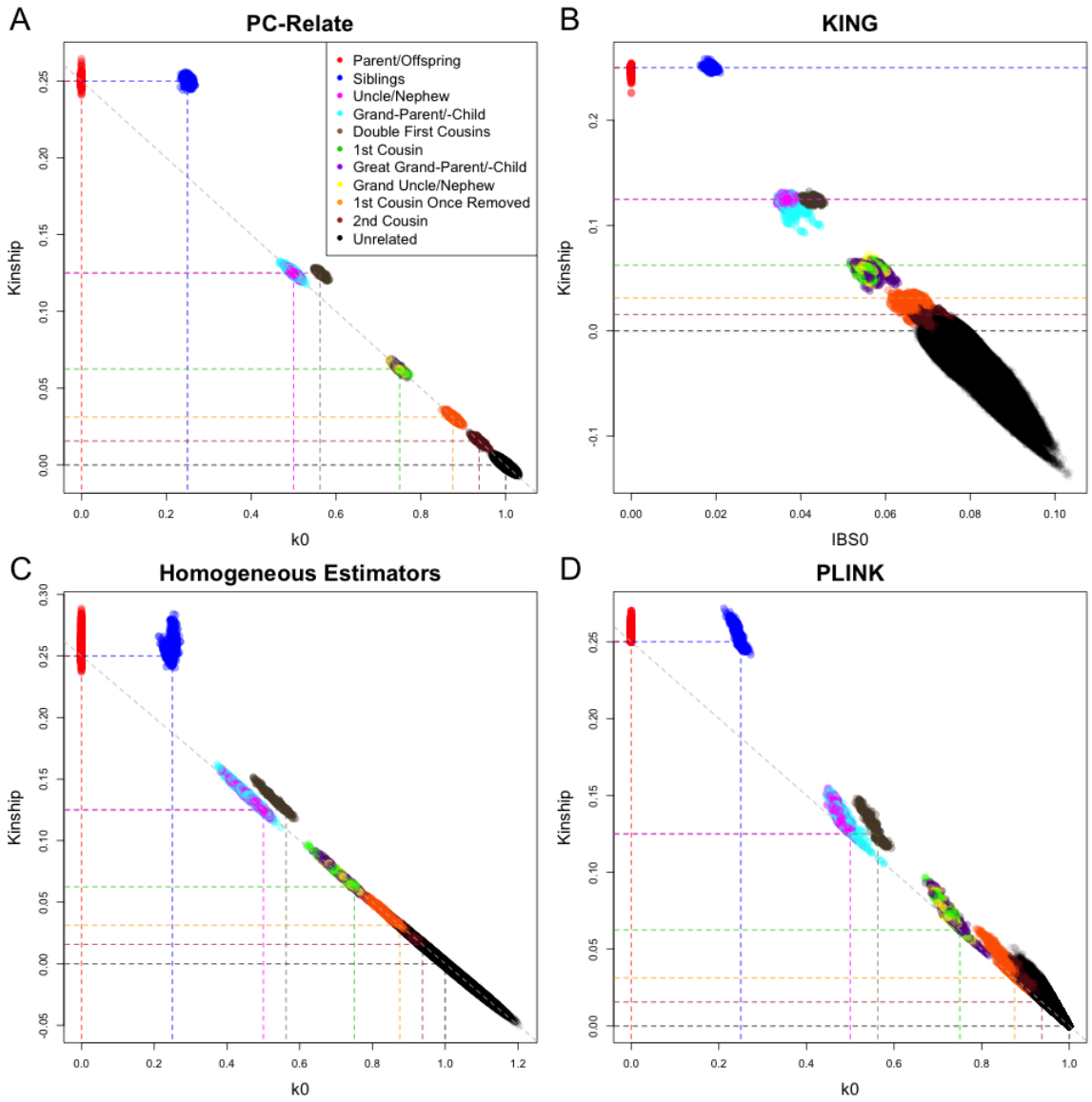


Figure 4.3: Relationship Estimation for Relationship Configuration I under Population Structure II

Scatter plots of the estimated kinship coefficient against the estimated probability of sharing zero alleles IBD, $k^{(0)}$, from (A) PC-Relate, (C) the Homogeneous Estimators, and (D) PLINK. (B) KING-robust does not provide IBD sharing probability estimates for structured populations, so the estimated kinship coefficients are plotted against the proportion of SNPs where the pair of individuals are opposite homozygotes; i.e. share zero alleles identical by state (IBS). Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical values for the corresponding relationship type.

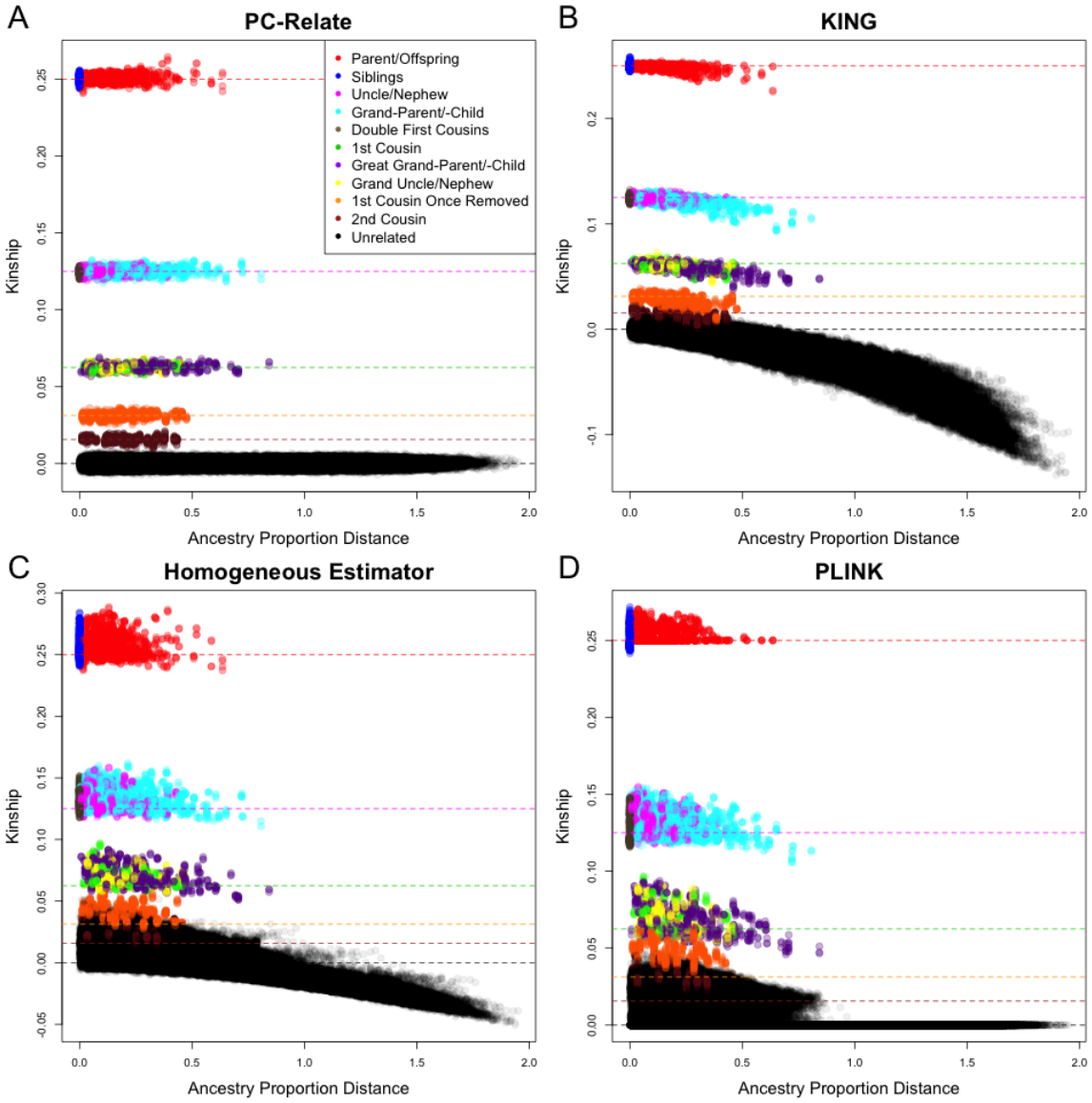


Figure 4.4: Kinship Estimation as a Function of Ancestry Difference for Relationship Configuration I under Population Structure II

Scatter plots of the estimated kinship coefficient against the ancestry proportion distance, defined as $\sum_{k=1}^{K} |a_i^k - a_j^k|$, for (A) PC-Relate, (B) KING-robust, (C) the Homogeneous Estimators, and (D) PLINK. Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical value for the corresponding relationship type.

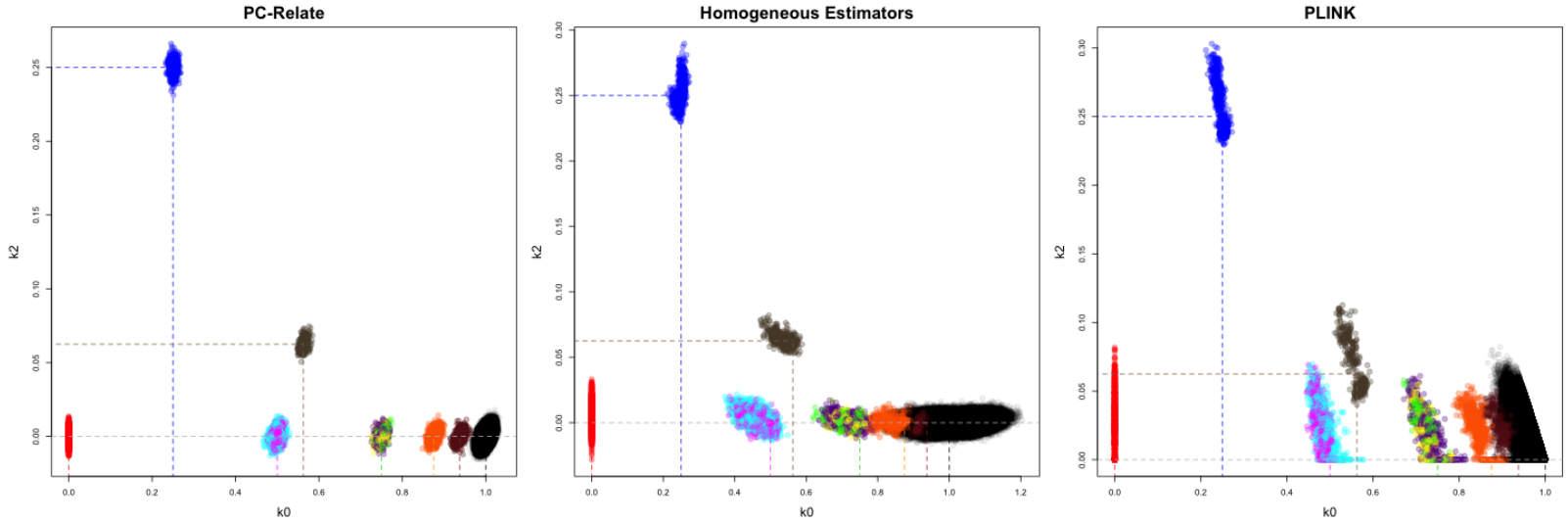


Figure 4.5: IBD Sharing Results for Relationship Configuration I under Population Structure II
Scatter plots of the estimated probabilities of sharing two alleles IBD, $k^{(2)}$, against sharing zero alleles IBD, $k^{(0)}$, from (A) PC-Relate, (B) the Homogeneous Estimators, and (C) PLINK. Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical values for the corresponding relationship type.

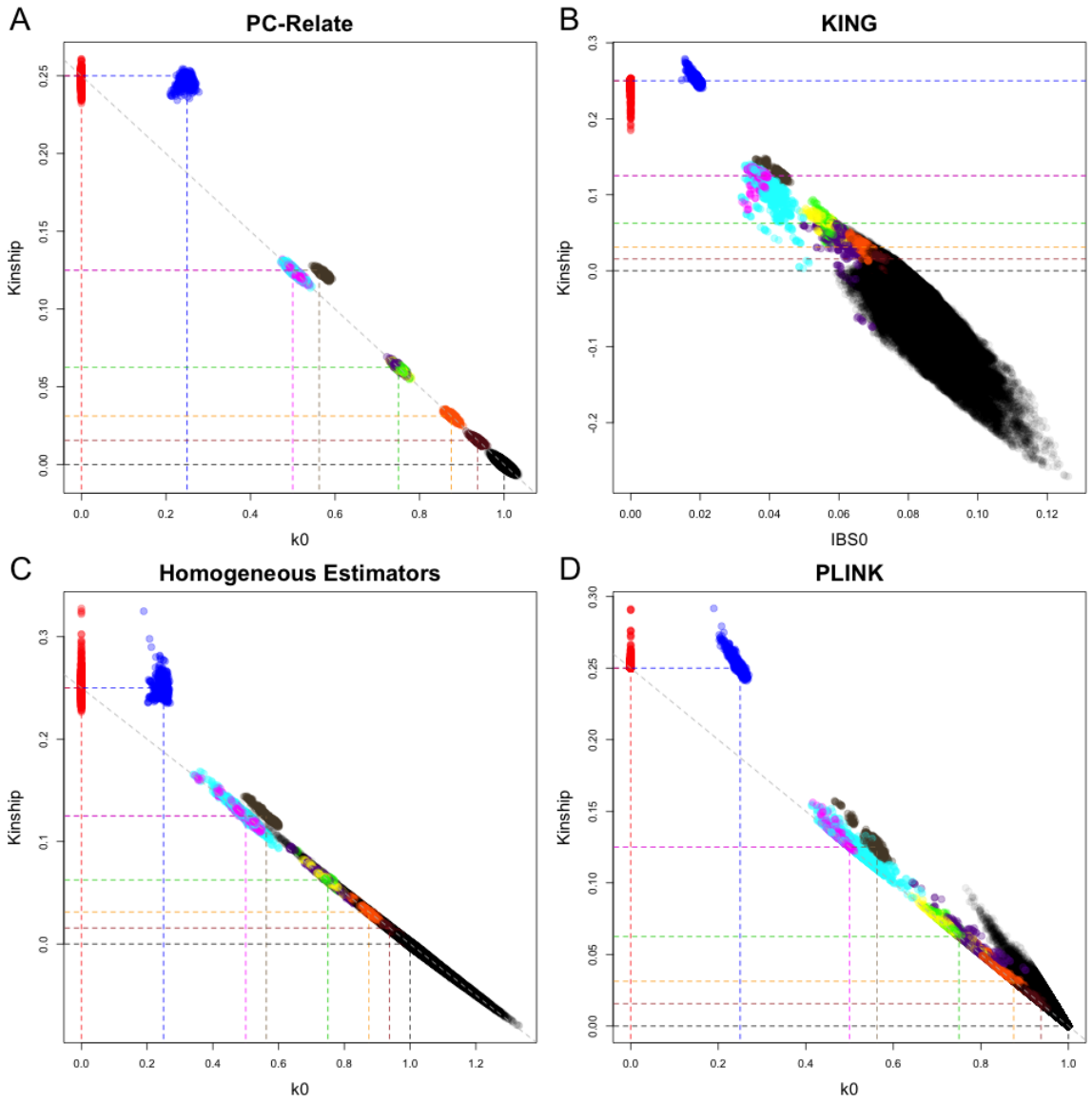


Figure 4.6: Relationship Estimation for Relationship Configuration I under Population Structure I

Scatter plots of the estimated kinship coefficient against the estimated probability of sharing zero alleles IBD, $k^{(0)}$, from (A) PC-Relate, (C) the Homogeneous Estimators, and (D) PLINK. (B) KING-robust does not provide IBD sharing probability estimates for structured populations, so the estimated kinship coefficients are plotted against the proportion of SNPs where the pair of individuals are opposite homozygotes; i.e. share zero alleles identical by state (IBS). Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical values for the corresponding relationship type.

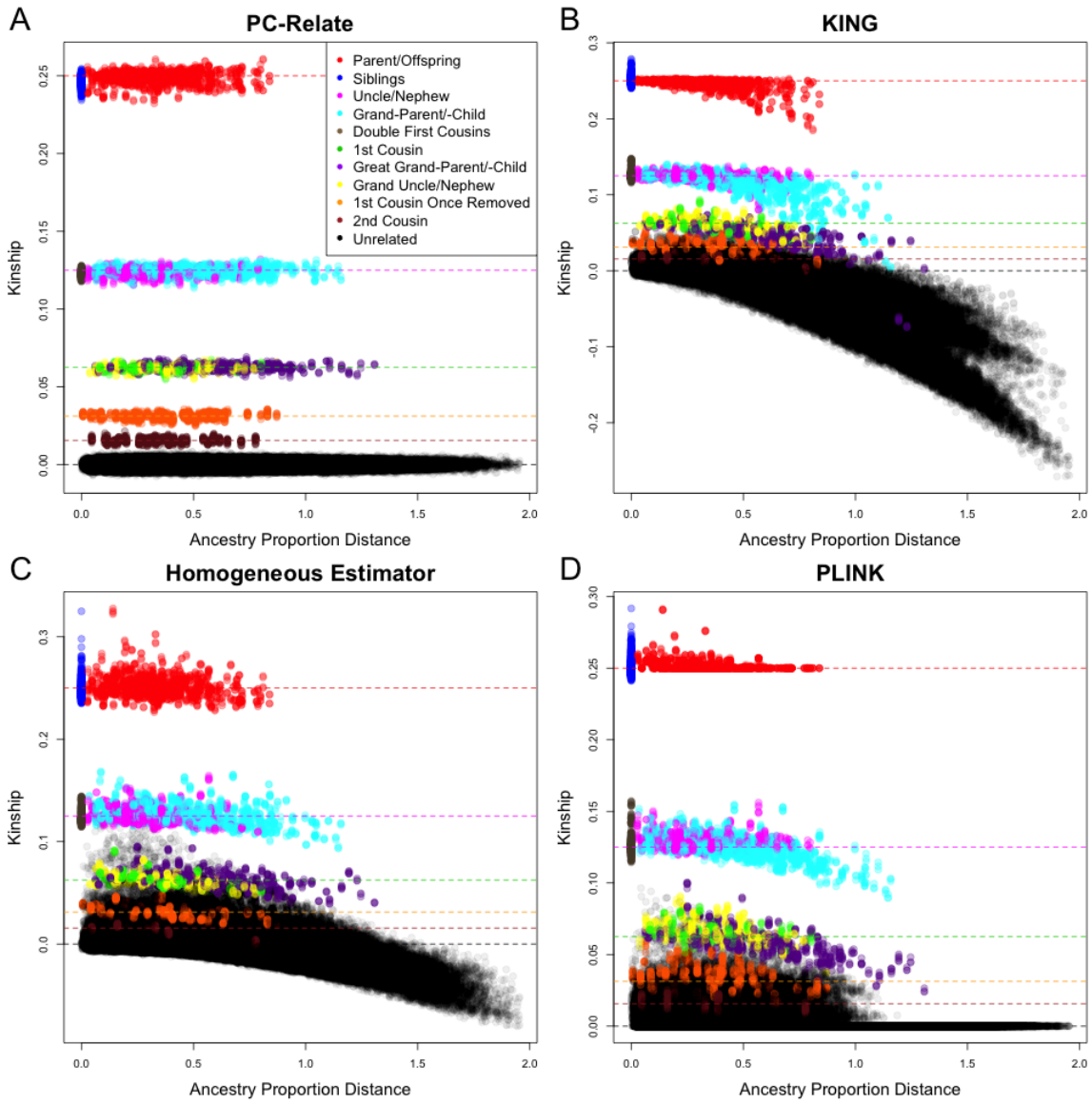


Figure 4.7: Kinship Estimation as a Function of Ancestry Difference for Relationship Configuration I under Population Structure I

Scatter plots of the estimated kinship coefficient against the ancestry proportion distance, defined as $\sum_{k=1}^{K} |a_i^k - a_j^k|$, for (A) PC-Relate, (B) KING-robust, (C) the Homogeneous Estimators, and (D) PLINK. Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical value for the corresponding relationship type.

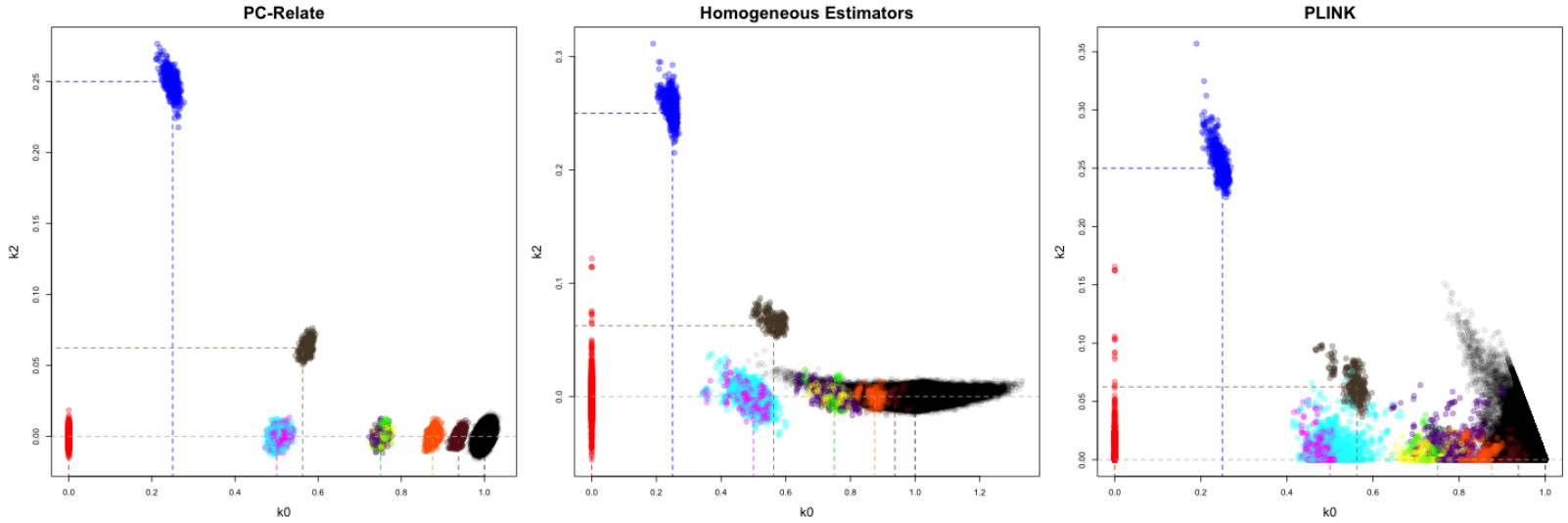


Figure 4.8: IBD Sharing Results for Relationship Configuration I under Population Structure I
Scatter plots of the estimated probabilities of sharing two alleles IBD, $k^{(2)}$, against sharing zero alleles IBD, $k^{(0)}$, from (A) PC-Relate, (B) the Homogeneous Estimators, and (C) PLINK. Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical values for the corresponding relationship type.

bias was present in some kinship coefficients for relatives with very different ancestry from their shared ancestors. Despite this, however, PC-Relate was still able to accurately infer relationship types for all pairs of individuals up to 5^{th} degree relatives. The biases in the estimators that assume population homogeneity followed the same patterns as under population structure II, but were even larger to the point that it was difficult to distinguish relationship types other than 1^{st} degree relatives. For example, the homogeneous kinship coefficient estimates were so inflated that 134 unrelated pairs were inferred to be 2^{nd} degree relatives, and the PLINK $k^{(2)}$ estimates were so inflated that 39 additional 2^{nd} degree relative pairs were inferred to be double first cousins.

4.3.2 Discrete Subpopulations

We also considered relationship configuration I under population structure III, a setting with outbred individuals and discrete population substructure. In this setting, PC-Relate provided consistent estimates of kinship coefficients and IBD sharing probabilities with low variability. The KING-robust estimator was designed for this setting and also performed very well, providing consistent kinship coefficient estimates for relatives and unrelated pairs with the same ancestry. However, the KING-robust estimates were negative for unrelated pairs of individuals from different subpopulations, and the magnitude of this negative value was a function of the differentiation, as measured by F_{ST}, between these two subpopulations. The estimators that assume population homogeneity performed terribly in this setting, with huge positive and negative biases both within and across subpopulations (Figures 4.9, 4.10, and 4.11).

4.3.3 Inbred Populations

Kinship coefficients were also estimated for relationship configuration II under each of the three population structure settings. IBD sharing probabilities were not estimated because relationship configuration II consists of pedigrees with some inbreeding, and,

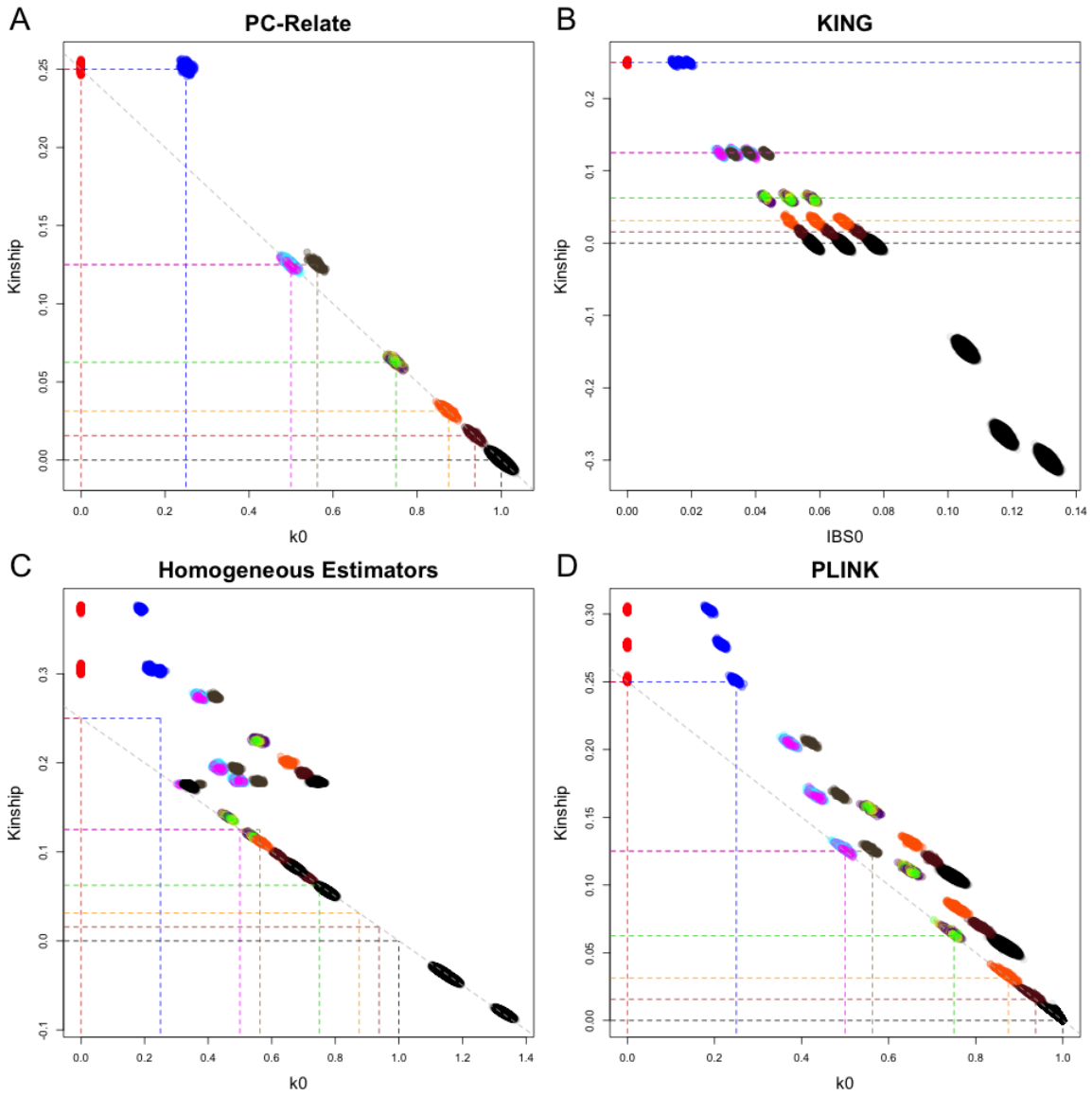


Figure 4.9: Relationship Estimation for Relationship Configuration I under Population Structure III

Scatter plots of the estimated kinship coefficient against the estimated probability of sharing zero alleles IBD, $k^{(0)}$, from (A) PC-Relate, (C) the Homogeneous Estimators, and (D) PLINK. (B) KING-robust does not provide IBD sharing probability estimates for structured populations, so the estimated kinship coefficients are plotted against the proportion of SNPs where the pair of individuals are opposite homozygotes; i.e. share zero alleles identical by state (IBS). Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical values for the corresponding relationship type.

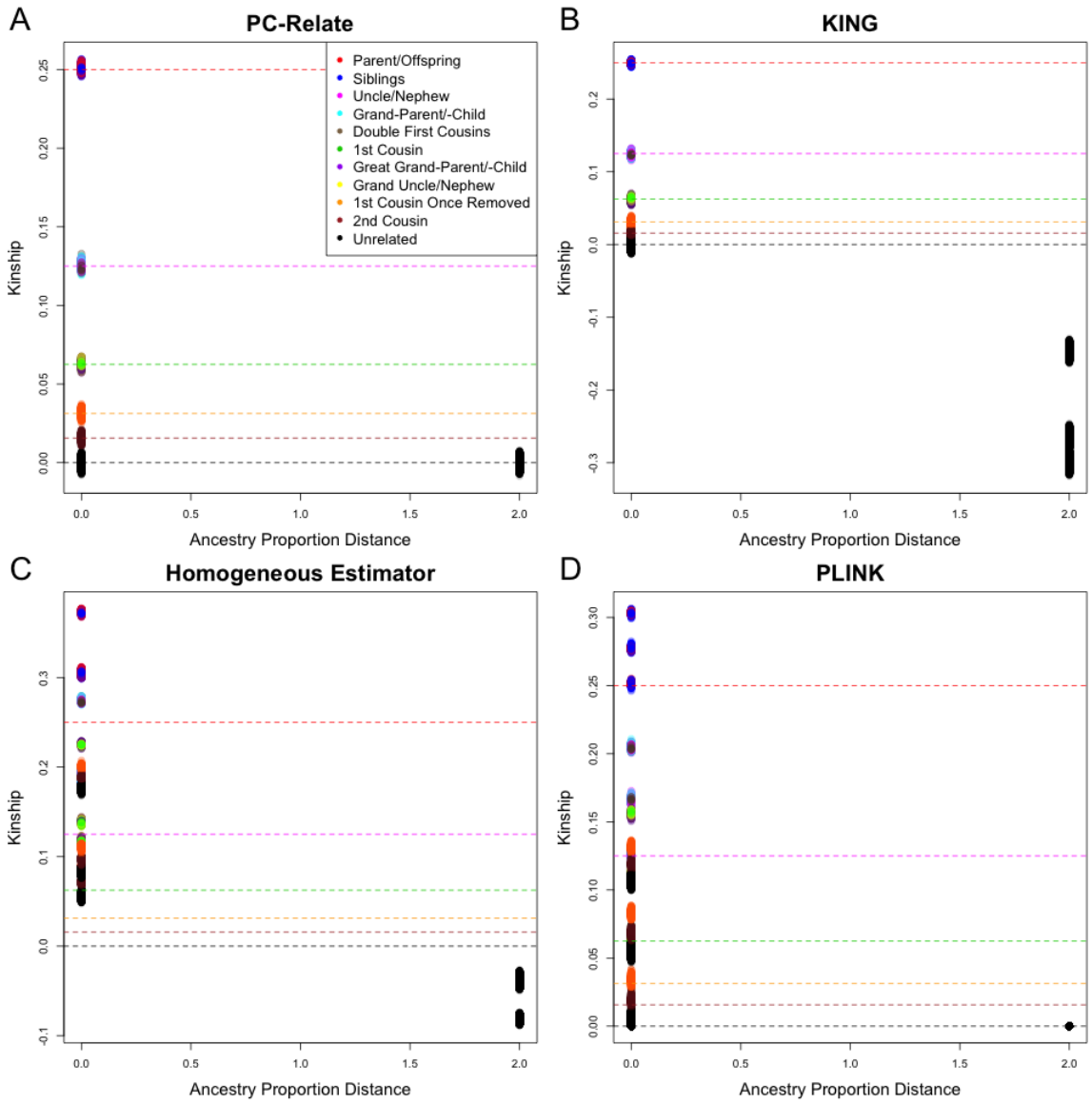


Figure 4.10: Kinship Estimation as a Function of Ancestry Difference for Relationship Configuration I under Population Structure III

Scatter plots of the estimated kinship coefficient against the ancestry proportion distance, defined as $\sum_{k=1}^{K} |a_i^k - a_j^k|$, for (A) PC-Relate, (B) KING-robust, (C) the Homogeneous Estimators, and (D) PLINK. Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical value for the corresponding relationship type.

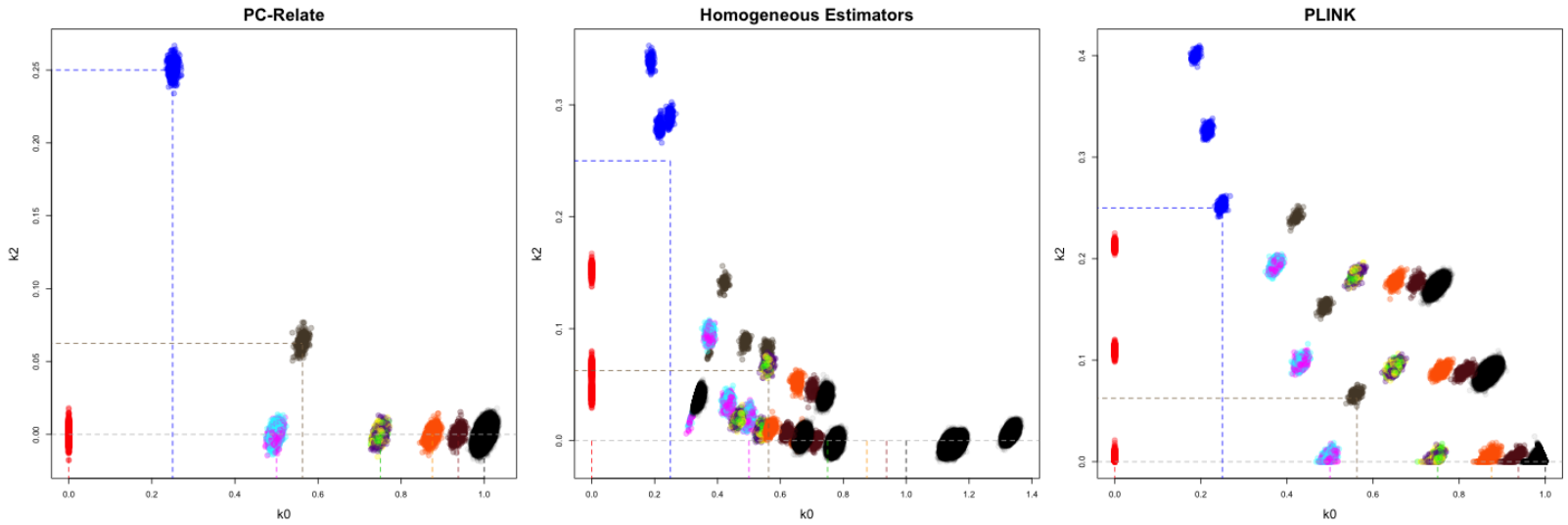


Figure 4.11: IBD Sharing Results for Relationship Configuration I under Population Structure III
Scatter plots of the estimated probabilities of sharing two alleles IBD, $k^{(2)}$, against sharing zero alleles IBD, $k^{(0)}$, from (A) PC-Relate, (B) the Homogeneous Estimators, and (C) PLINK. Each point is color coded by the true relationship type of the pair of individuals, and the colored dashed lines show the theoretical values for the corresponding relationship type.

as discussed previously, there are no longer only three IBD states. PC-Relate once again provided accurate kinship coefficient estimates for all settings, KING-robust was still biased in the presence of ancestry admixture, and the estimators that assume population homogeneity struggled in the presence of any population structure. However, the point of interest from these simulations was the effect of inbreeding on the estimators. We present the results from population structure III for PC-Relate and KING-robust in Figure 4.12. We selected this setting because both PC-Relate and KING-robust provided consistent kinship coefficient estimates for outbred samples with discrete population substructure. Figure 4.12A shows that PC-Relate still provided consistent kinship coefficient estimates with low variability, even in the presence of inbreeding. On the other hand, Figure 4.12B shows that KING-robust only provided consistent estimates for pairs of individuals that were both outbred and were from the same subpopulation. If either individual in the pair was inbred, then the KING-robust kinship coefficient estimates were negatively biased. The magnitude of this negative bias grew larger with higher levels of inbreeding for each individual, and the exact relationship is given in Appendix A.2.3.

We also estimated inbreeding coefficients using PC-Relate and the corresponding estimator that assumes population homogeneity for each setting. Histograms of the estimates for relationship configuration II under population structure II are presented in Figure 4.13. The PC-Relate estimates were accurate, and individuals could easily be identified as outbred, the offspring of a first cousin once removed mating, or the offspring of a first cousin mating. In comparison, the estimates from the homogeneous estimator were inflated, as expected from Equation (4.11), and many outbred individuals had estimates consistent with first cousin once removed or first cousin inbreeding. The results from population structure III were similar, but it is interesting to look at the results from population structure I in more detail. In this setting, some individuals were the offspring of parents with very different ancestry, resulting in excess heterozygosity. The histogram in Figure 4.14 shows that PC-Relate gave

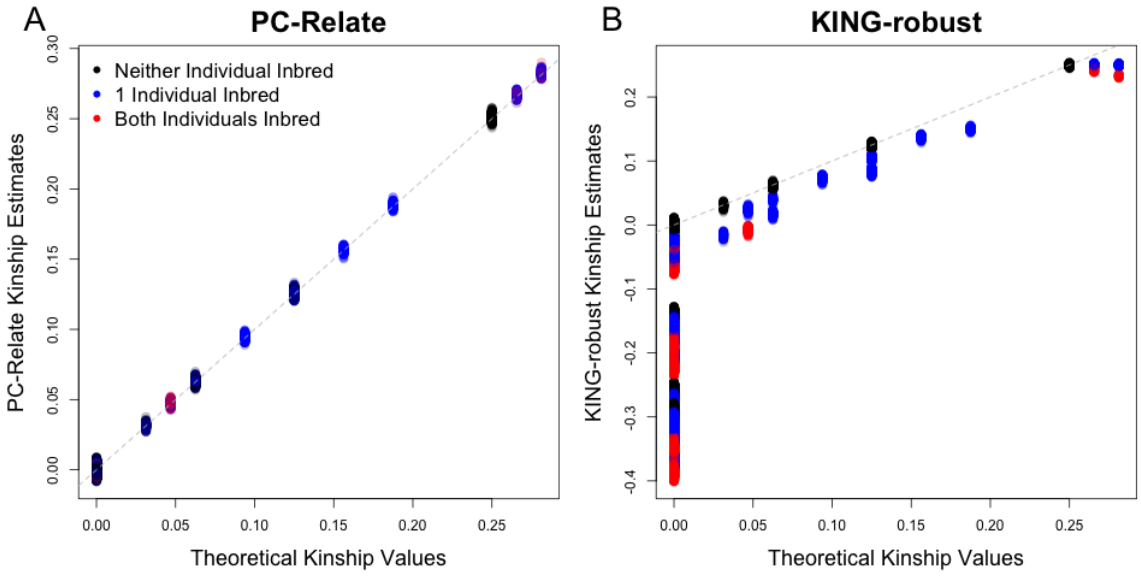


Figure 4.12: Kinship Coefficient Estimation for Relationship Configuration II under Population Structure III

Scatter plots of estimated kinship coefficients against theoretical kinship coefficients for (A) PC-Relate and (B) KING-robust. Theoretical kinship coefficients were calculated from the pedigrees using the QTLRel package [7] for R. Each point is color coded by the number of inbred individuals in the pair.

negative estimates for these individuals, as expected from Equation (4.13), providing insight into these recent admixture events. Despite this, Figure 4.14 also shows that PC-Relate was still able to accurately identify the inbred individuals in the sample. In comparison, the homogeneous estimator provided biased and extremely variable estimates, and many outbred individuals appeared to be highly inbred.

4.3.4 WHI-SHARe Hispanic Cohort

We analyzed the high-density genotype data collected from 3,587 women of self-identified Hispanic descent in the Women's Health Initiative SNP Health Association Resource (WHI-SHARe) study with PC-Relate, KING-robust, and PLINK. Hispanic populations are well known to be admixed with three major continental ancestries; European, Native American, and African, and a recent study [32] has shown that there is further subcontinental population structure within these populations. We filtered SNPs with a sample minor allele frequency (MAF) < 0.05 and LD pruned using an r^2 threshold of 0.10 to obtain a set of 87,180 nearly independent markers for the purpose of ancestry inference and relatedness estimation. The PC-Relate estimates were found in two iterations of both: (1) inferring ancestry representative principal components using the PC-AiR method (Chapter 3), and (2) estimating relatedness measures using PC-Relate. For the first iteration, PC-AiR used the KING-robust kinship coefficient estimates to identify pairs of relatives, and PC-Relate used the resulting PC-AiR PCs to estimate individual-specific allele frequencies and measures of relatedness. However, the KING-robust estimates were badly biased due to ancestry admixture and departures from HWE, and far too many pairs of relatives were identified. To improve both the ancestry inference with PC-AiR and relatedness estimation with PC-Relate, a second iteration was performed. For the second iteration, PC-AiR used the PC-Relate kinship coefficient estimates from the first iteration to identify pairs of relatives, and PC-Relate used the new PC-AiR PCs for estimating individual-specific allele frequencies and measures of relatedness. A further descrip-

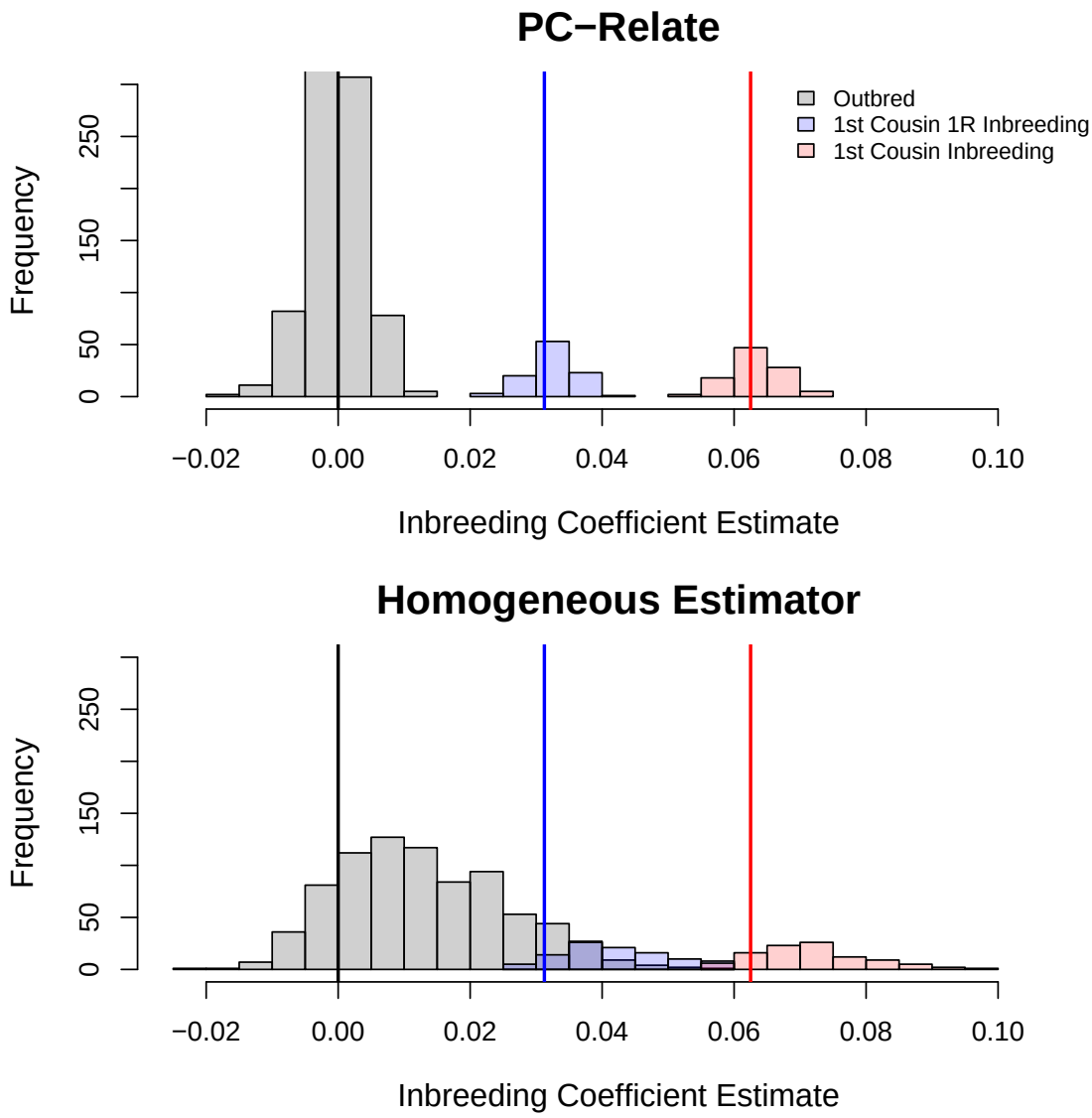


Figure 4.13: Histograms of Inbreeding Coefficient Estimates for Relationship Configuration II under Population Structure II

Histograms of estimated inbreeding coefficients for (A) PC-Relate, (B) the Homogeneous Estimator. Values are color coded by whether the individual is outbred, the offspring of a first cousin once removed mating, or the offspring of a first cousin mating. Theoretical values are shown with color corresponding vertical lines.

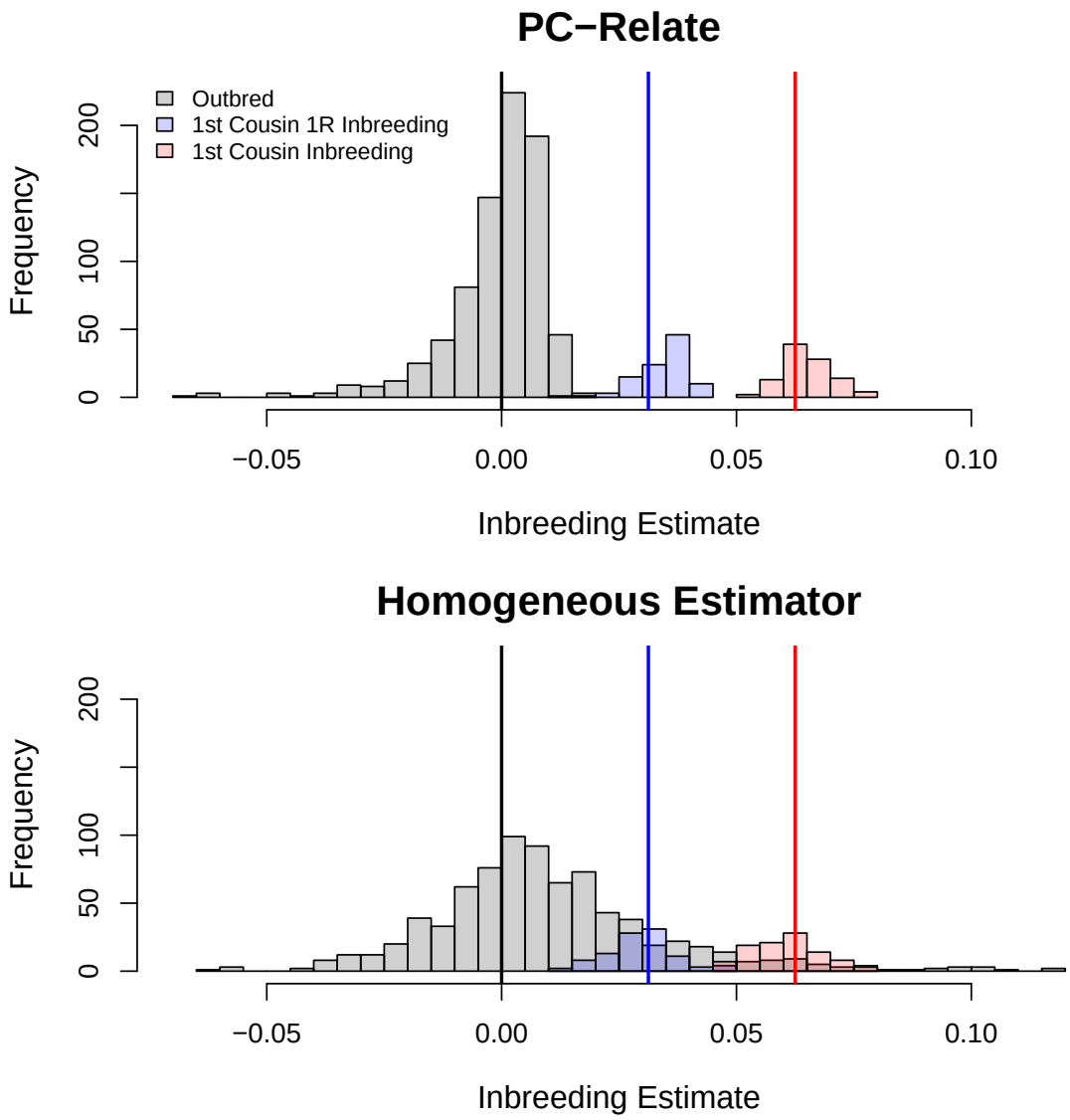


Figure 4.14: Histograms of Inbreeding Coefficient Estimates for Relationship Configuration II under Population Structure I

Histograms of estimated inbreeding coefficients for (A) PC-Relate, (B) the Homogeneous Estimator. Values are color coded by whether the individual is outbred, the offspring of a first cousin once removed mating, or the offspring of a first cousin mating. Theoretical values are shown with color corresponding vertical lines.

tion of the iterative procedure is given in Chapter 5. In both iterations of the analysis, PC-Relate was adjusted for the top 8 principal components from PC-AiR, which all appeared to reflect population structure (see Figure 4.15).

Relationship types were assigned to all pairs of individuals according to the estimates from each method, using the same criteria as described previously. Due to Mendelian sampling and linkage, variation in the actual proportion of alleles shared IBD by a pair of individuals of a specific relationship type results in no distinct clusters of relatives at or below 3^{rd} degree [17]. For this reason, relationship types were only inferred up to 3^{rd} degree, and a pair of individuals with $\widehat{\phi}_{ij} < 2^{(-9/2)} \approx 0.044$ was classified as unrelated. A comparison of the relationship assignment from PC-Relate and KING-robust is presented in Table 4.2. We see perfect concordance between PC-Relate and KING-robust for 1^{st} degree relatives. The majority of more distant relatives identified by PC-Relate were also identified by KING-robust, but KING-robust identified tens of additional 2^{nd} degree and thousands of additional 3^{rd} degree relatives that PC-Relate estimated to be unrelated. We believe that KING-robust is overestimating relatedness for these pairs of individuals, as this is consistent with the results we saw in our simulations under population structure I, where recent admixture for the offspring of parents with different ancestries led to inflated estimates. Further support of this hypothesis is provided by the fact that PC-Relate estimated negative inbreeding coefficients ($\widehat{f}_i^{PC} < -2^{(-11/2)} \approx -0.022$) for 74 individuals in the sample, also indicating recent admixture and excess heterozygosity. Further, 77 of 78 additional 2^{nd} degree relative pairs and 1,546 of 2,395 additional 3^{rd} degree relative pairs identified by KING-robust involve at least one of these individuals with negative inbreeding estimates. Relationship inference with PLINK is also perfectly concordant for 1^{st} degree relatives, but PLINK identified 20 additional 2^{nd} degree and 36,351 additional 3^{rd} degree relatives as compared to PC-Relate. These results were not surprising, as we expected inflated estimates from PLINK for pairs of individuals with similar ancestry. The results described above can also be seen in Figure 4.16,

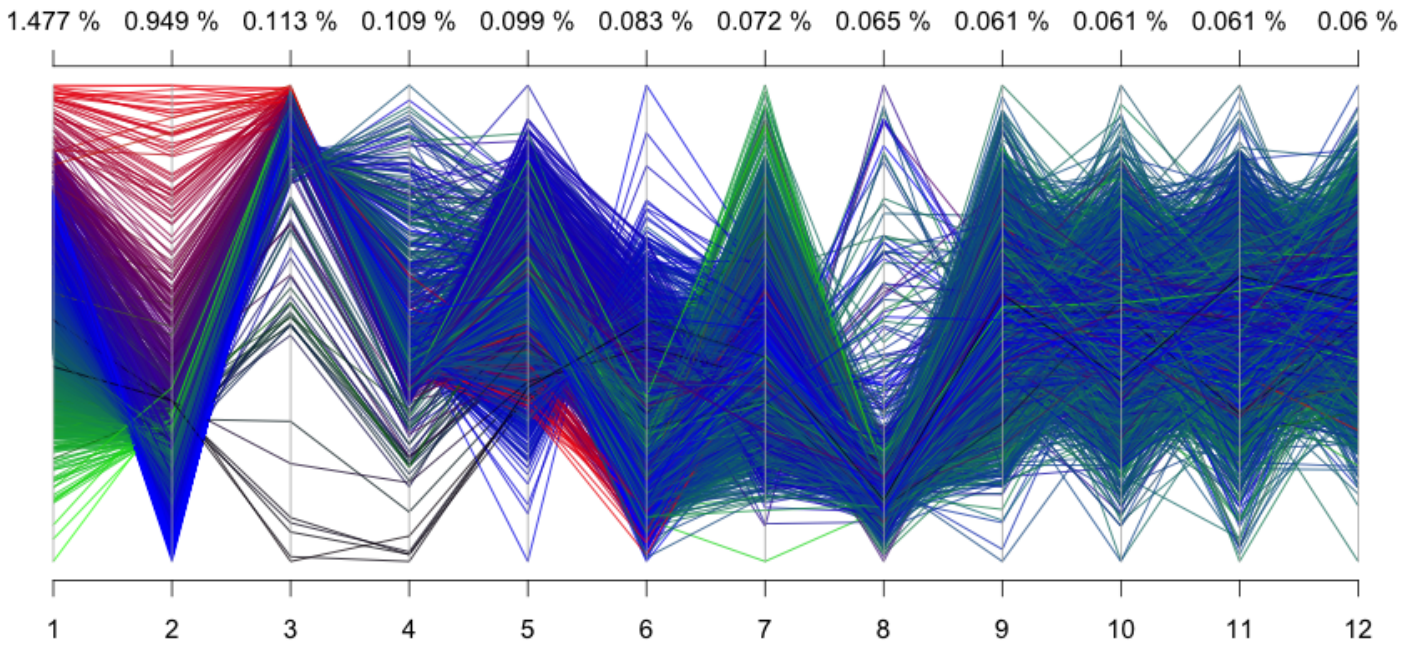


Figure 4.15: PC-AiR Ancestry Inference for the WHI SHARe Hispanic Cohort
Parallel coordinates plot showing the first 12 PCs from PC-AiR. Each vertical bar represents one of the principal components, and each line traces out the coordinates for an individual across all 12 PCs. The color represents the ancestry inferred from the supervised ADMIXTURE analysis; red for African, blue for European, green for Native American, and black for East Asian, with gradients for admixed individuals. The percentage of variation of the genetic data in the unrelated subset that is accounted for by each principal component is shown at the top of the plot. Only the first 8 PCs appear to show any structure.

where the kinship coefficient estimates from each method are directly compared. Finally, we examined the number of inferred relatives for each individual according to each method, and the results for PC-Relate (maximum = 3, mean = 0.089) were much more reasonable than for KING-robust (maximum = 118, mean = 1.463) or for PLINK (maximum = 2897, mean = 20.360, lending more support to the argument that PC-Relate outperformed the other methods.

Table 4.2: Pairwise Relationship Assignment from PC-Relate and KING-robust

		PC-Relate						
		MZ	FS	PO	2^{nd}	3^{rd}	Unrel	Total
KING-robust	MZ	**1**	0	0	0	0	0	1
	FS	0	**71**	0	0	0	0	71
	PO	0	0	**8**	0	0	0	8
	2^{nd}	0	0	0	**17**	5	73	95
	3^{rd}	0	0	0	1	**53**	2395	2449
	Unrel	0	0	0	0	3	**6428864**	6428867
	Total	1	71	8	18	61	6431332	6431491

Relationship Types: MZ: monozygotic twins; FS: full siblings; PO: parent/offspring; 2^{nd}: second degree relatives; 3^{rd}: third degree relatives; Unrel: unrelated pair.

4.3.5 Comparison with Methods that Use Reference Population Panels

To compare PC-Relate with relationship estimation methods that are specifically designed for admixed populations, we also analyzed the WHI SHARe Hispanic cohort data with REAP and RelateAdmix, both of which require individual ancestry proportion and subpopulation-specific allele frequency estimates as input. To obtain these estimates, a supervised ancestry analysis was performed using the ADMIXTURE [3] software, with the number of ancestral populations set to four, for which the HapMap [18] CEU (Utah residents with ancestry from northern and west-

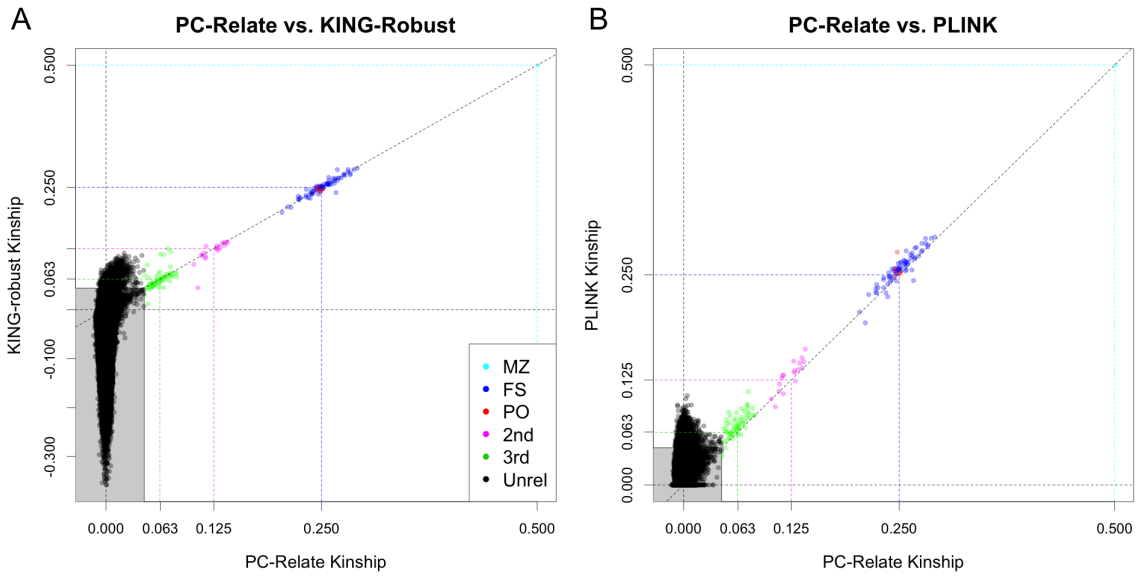


Figure 4.16: Comparison of Kinship Coefficient Estimates for WHI SHARe Hispanic Cohort

Scatter plot of the estimated kinship coefficients from PC-Relate vs. (A) KING-robust and (B) PLINK. The shaded gray box indicates estimates where both methods infer unrelated pairs. Each point is color coded by the relationship type of the pair of individuals, as inferred from PC-Relate, and the colored dashed lines show the theoretical kinship values for the corresponding relationship type. The same relationship type abbreviations are used as in Table 4.2.

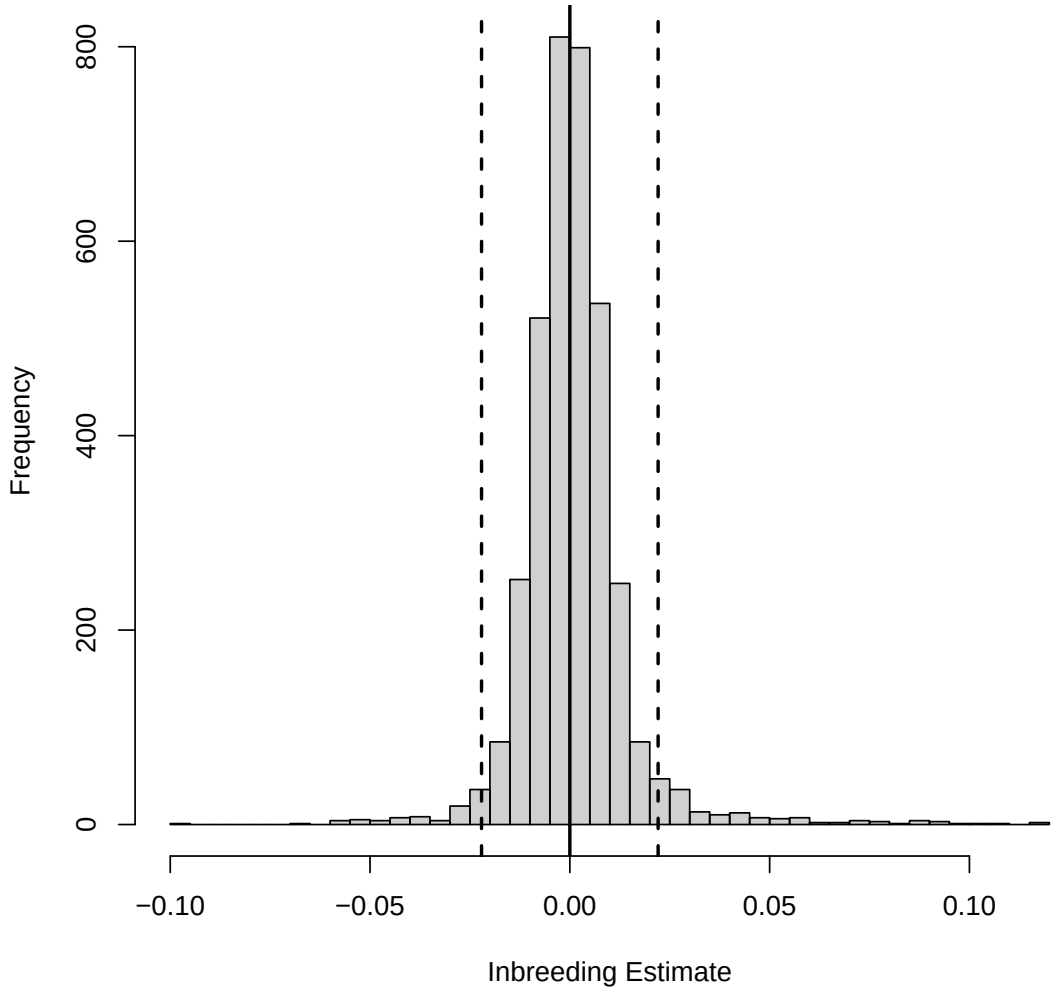


Figure 4.17: Histogram of PC-Relate Inbreeding Coefficient Estimates in WHI SHARe Hispanic Cohort

Histogram of estimated inbreeding coefficients from PC-Relate. The dashed vertical lines are at $\pm 2^{(-11/2)} \approx \pm 0.022$, representing the expected upper and lower bounds of estimates for outbred individuals.

ern Europe from the Centre dEtude du Polymorphisme Human collection) and YRI (Yoruba in Ibadan, Nigeria) samples were included as the reference population panels for European and African ancestry, respectively, the HapMap CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan) samples were included jointly as the reference population panel for East Asian ancestry, and the Human Genome Diversity Project (HGDP) [26] samples from the Americas were included as the reference population panel for Native American ancestry. A set of SNPs common to the WHI-SHARe, HapMap, and HGDP data sets was found, filtered with a sample MAF < 0.05, and LD pruned using an r^2 threshold of 0.10, resulting in a set of 59,969 nearly independent markers for the purpose of ancestry and relatedness estimation.

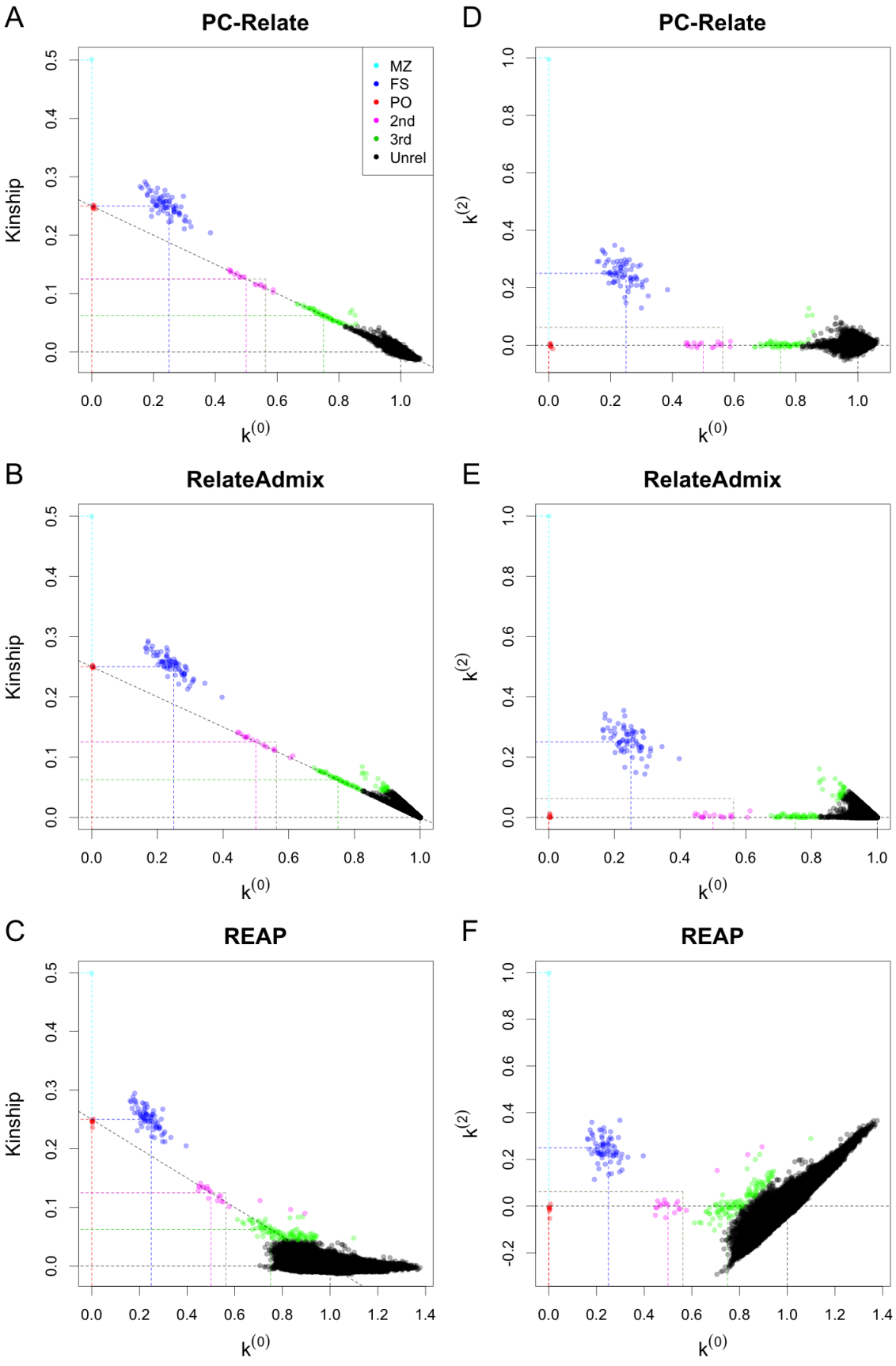
The relationship estimation results for each of the three methods are presented in Figure 4.18. RelateAdmix is the only method that provided truncated estimates, restricting $\phi_{ij} > 0$, $k^{(0)} < 1$, and $k^{(2)} > 0$. We can see that the kinship coefficient estimates from all three methods were very similar, with only slight variation. Additionally, PC-Relate and RelateAdmix provided similar estimates for IBD sharing probabilities, while, compared to the other methods, the REAP $k^{(0)}$ estimates showed considerably more noise for distant relatives and unrelated pairs. The REAP $k^{(0)}$ estimator for all pairs of individuals takes the same form as the PC-Relate estimator for first degree relatives, which we have found to have high sampling variability in distant relatives and unrelated pairs. Due to this increased noise in the $k^{(0)}$ estimator, REAP also provided poor $k^{(2)}$ estimates for more distant relatives and unrelated pairs.

Relationship types were also assigned from the RelateAdmix and REAP estimates using the same criteria as before. For 1^{st} and 2^{nd} degree relatives, there was perfect concordance between PC-Relate and RelateAdmix and nearly perfect concordance between PC-Relate and REAP, where REAP identified two additional 2^{nd} degree relative pairs that PC-Relate and RelateAdmix both estimated to be 3^{rd} degree relatives. There was also high concordance among all three methods for 3^{rd} degree relationships, although, amongst pairs of individuals that PC-Relate identi-

fied as unrelated, RelateAdmix and REAP identified 26 and 59 additional 3^{rd} degree relationships, respectively. In comparison, PC-Relate only identified two 3^{rd} degree relationships that RelateAdmix inferred to be unrelated, and one 3^{rd} degree relationship that REAP inferred to be unrelated. Kinship coefficient estimates from all three methods are directly compared in Figure 4.19. Compared to PC-Relate, we can see slightly higher estimates from RelateAdmix and REAP for distant relative or unrelated pairs, in agreement with the additional identified relative pairs. However, pairwise correlations of estimated kinship coefficients, restricted to pairs of individuals that either method identified as related, were greater than 0.99 for all three methods. We also computed the number of inferred relatives for each individual according to these methods, and the results for RelateAdmix (maximum = 6, mean = 0.102) and REAP (maximum = 11, mean = 0.121) only showed slightly more relatedness than PC-Relate.

It is likely that RelateAdmix and REAP provide slightly larger kinship coefficient estimates than PC-Relate for some distant relative or unrelated pairs because they require prior specification of the number and types of ancestral populations contributing ancestry to the sample. For the pre-specified continental ancestry groups, there was high concordance between the PC-AiR analysis performed without reference panels and the supervised ADMIXTURE analysis. The top 3 PC-AiR PCs were able to jointly explain $> 99\%$ of the variability in each of the ADMIXTURE estimated European, African, and Native American ancestry proportions, while the top 4 PC-AiR PCs were able to jointly explain 95.5% of the variability in the estimated East Asian ancestry proportions. However, PC-AiR PCs 4-8 also appear to capture further subcontinental population structure that ADMIXTURE can not estimate. As a result, PC-Relate is able to adjust for this fine scale structure, while RelateAdmix and REAP can not account for it.

Figure 4.18: Relationship Estimation in WHI SHARe Hispanic Cohort
Scatter plots of the estimated kinship coefficients against the estimated probabilities of sharing zero alleles IBD, $k^{(0)}$, from (A) PC-Relate, (B) RelateAdmix, and (C) REAP, and of the estimated probabilities of sharing two alleles IBD, $k^{(2)}$, against $k^{(0)}$ from (D) PC-Relate, (E) RelateAdmix, and (F) REAP. Each point is color coded by the relationship type of the pair of individuals, as inferred from the respective method, and the colored dashed lines show the theoretical values of each measure for the corresponding relationship type.


A
PC-Relate
MZ
FS
PO
2nd
3rd
Unrel
Kinship
k(0)
D
PC-Relate
k(2)
k(0)
B
RelateAdmix
Kinship
k(0)
E
RelateAdmix
k(2)
k(0)
C
REAP
Kinship
k(0)
F
REAP
k(2)
k(0)

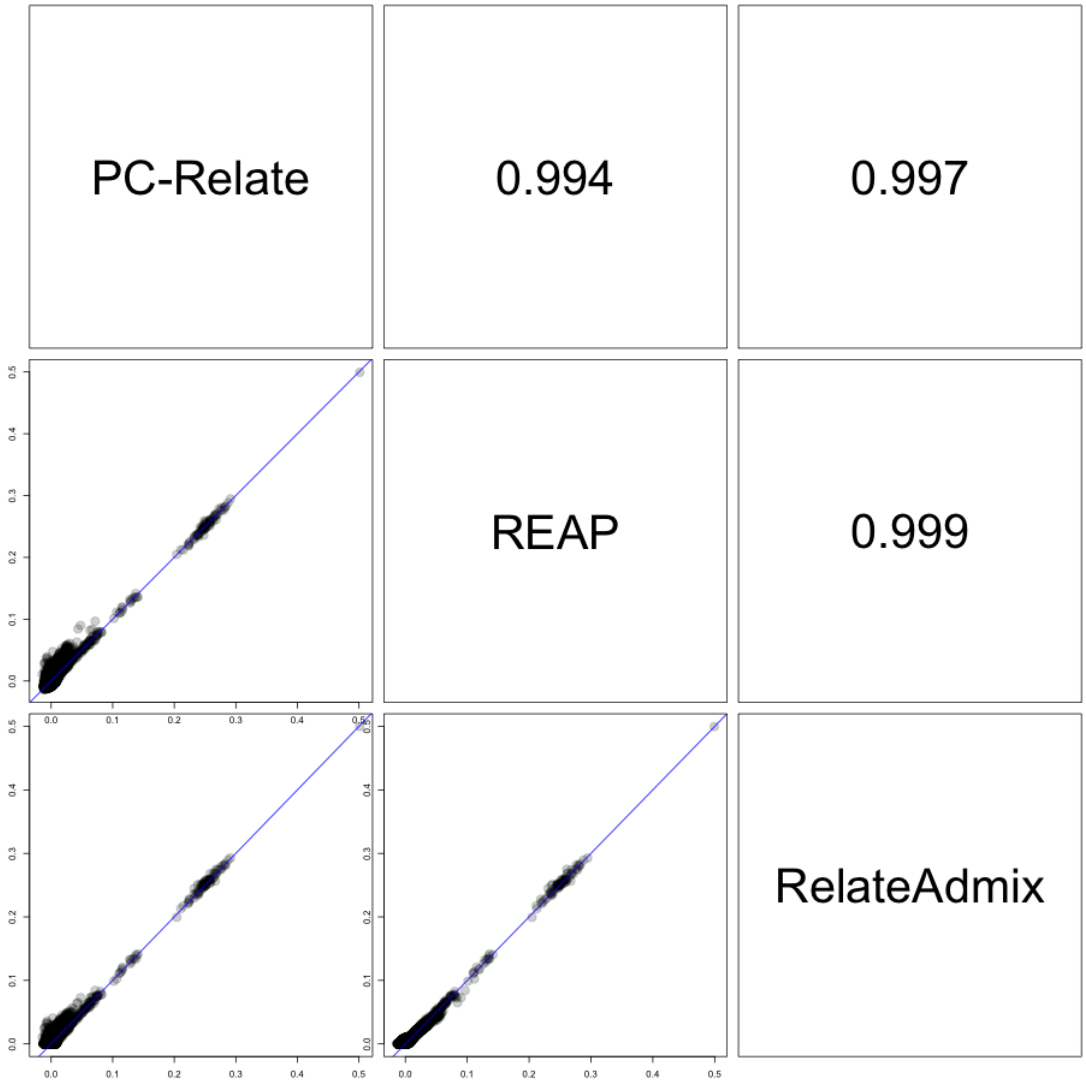


Figure 4.19: Comparison of Kinship Estimation with Methods that use Reference Panels in WHI SHARe Hispanic Cohort

The lower panels in the pairs plot show scatter plots directly comparing the PC-Relate, RelateAdmix, and REAP kinship coefficient estimates. The upper panels give the pairwise correlations of the estimated kinship coefficients restricted to pairs of individuals that either method identified as relatives.

4.3.6 Assessment of Computation Time

The computation time for each of these methods depends on both the sample size and the number of markers being analyzed. To analyze all 3,587 individuals in the WHI-SHARe Hispanic cohort with 87,180 markers took PC-Relate 12.1 minutes, KING-robust 5.0 minutes, and PLINK 282.8 minutes, respectively, on a 2.5GHz Intel Core i7 MacBook Pro with 8GB of 1333 MHz DDR3 RAM. To analyze all 3,587 individuals with 59,969 markers took REAP 73.2 minutes on the same laptop. RelateAdmix could not be run on the same system due to the increased computational time, and was instead parallelized on a 12 core 2.6GHz compute cluster with 128GB of RAM; the total computation time for RelateAdmix on the data set with 59,969 markers was 8.3 days. All of these computation times are only for relatedness estimation and do not include any prior analyses for ancestry inference such as PC-AiR or ADMIXTURE. While KING-robust is the fastest, it is important to note that the software does not provide IBD sharing probability or inbreeding coefficient estimates. PC-Relate can also be run without the computation of IBD sharing probabilities, which took only 5.9 minutes on the same laptop.

4.4 Discussion

Genetic relatedness estimation is essential to many areas of genetic research. Both known and cryptic relatives are common amongst sample individuals in genetic studies, and the collection of high-density genome-wide data allows for estimation of relatedness measures. Existing methods for relatedness estimation, however, all have some limitations. Many methods make assumptions such as population homogeneity or discrete population substructure, while others require prior knowledge of the underlying ancestral populations and the availability of suitable reference population panels. Here, we presented a new principal component based method, PC-Relate, for accurately estimating genetic relatedness measures from genome-screen data in

the presence of complex population structure, including ancestry admixture, without making any prior assumptions about the population structure or requiring additional reference panel samples.

In simulation studies with complex population structure, we demonstrated the accuracy of PC-Relate and its improvement over competing estimators that are biased as a result of making simplifying assumptions such as population homogeneity or discrete population substructure. We showed that PC-Relate can provide accurate estimates of kinship coefficients and IBD sharing probabilities, even in admixed populations. On the other hand, KING-robust, while accurate in populations with discrete substructure, is biased in admixed populations, providing pairwise kinship estimates that are too small for individuals with different ancestry, and too large in the presence of very recent admixture. Additionally, we simulated inbred samples to demonstrate that PC-Relate can provide accurate inbreeding and kinship coefficient estimates in this scenario, while KING-robust has difficulties with departures from HWE and provides a negatively biased kinship estimate when either individual in a pair is inbred.

We also performed relatedness estimation in the WHI SHARe Hispanic cohort, analyzing the data with PC-Relate, KING-robust, PLINK, REAP, and RelateAdmix. As expected from the simulation study, PC-Relate significantly outperformed KING-robust and PLINK due to the presence of complex ancestry admixture. REAP and RelateAdmix are both designed for admixed populations; however, they require prior knowledge of the underlying ancestral populations contributing to the ancestry of the target sample. Both methods use estimates of individual ancestry proportions and subpopulation-specific allele frequencies as input, and accurate estimation of these quantities requires reference population panels that are suitable surrogates for the ancestral populations. Remarkably, PC-Relate performed as well as REAP and RelateAdmix for kinship coefficient estimation, without requiring additional reference population samples. In fact, PC-Relate may have outperformed these methods when

considering distant relative pairs. Slight inflation of the kinship estimates from REAP and RelateAdmix is plausible for distant relatives or unrelated pairs, most likely due to more subtle sub-continental population structure that only PC-Relate can account for, as it makes no prior assumptions about the number of ancestral populations contributing to the underlying population structure, and it can account for an unrestricted number of dimensions of ancestry. For IBD sharing probability estimation, PC-Relate was able to perform as well as RelateAdmix, while we demonstrated that REAP may provide noisy estimates for more distant relatives and unrelated pairs. Further, while PC-Relate and RelateAdmix provide very similar results, PC-Relate is much more computationally efficient, and RelateAdmix is unfeasible for large studies with many thousand individuals.

A challenge of the PC-Relate method is selecting the appropriate number of principal components required to fully capture the population structure in the sample. Often, a reasonable guess can be made after performing the principal components analysis by examining plots of the top PCs to identify how many appear to reflect structure, and by examining the eigenvalues of these PCs to identify at what point they become relatively constant. Fortunately, including a few too many principal components in our estimating procedure does not seem to bias the results and leads to only a very minor increase in noise (Figure 4.20). We do not, however, recommend foregoing examination of the principal components and choosing an arbitrary number, as blindly including far too many PCs can result in a substantial increase in noise. It is, however, important to use ancestry informative principal components that are not confounded by family structure, such as those obtained from PC-AiR. If artefactual principal components that reflect family structure are used for PC-Relate, the adjustment procedure may remove correlation due to kinship and result in negatively biased estimates.

PC-Relate also provides further versatility in the analyses that it can perform. In the case where the underlying population structure is well known a priori and

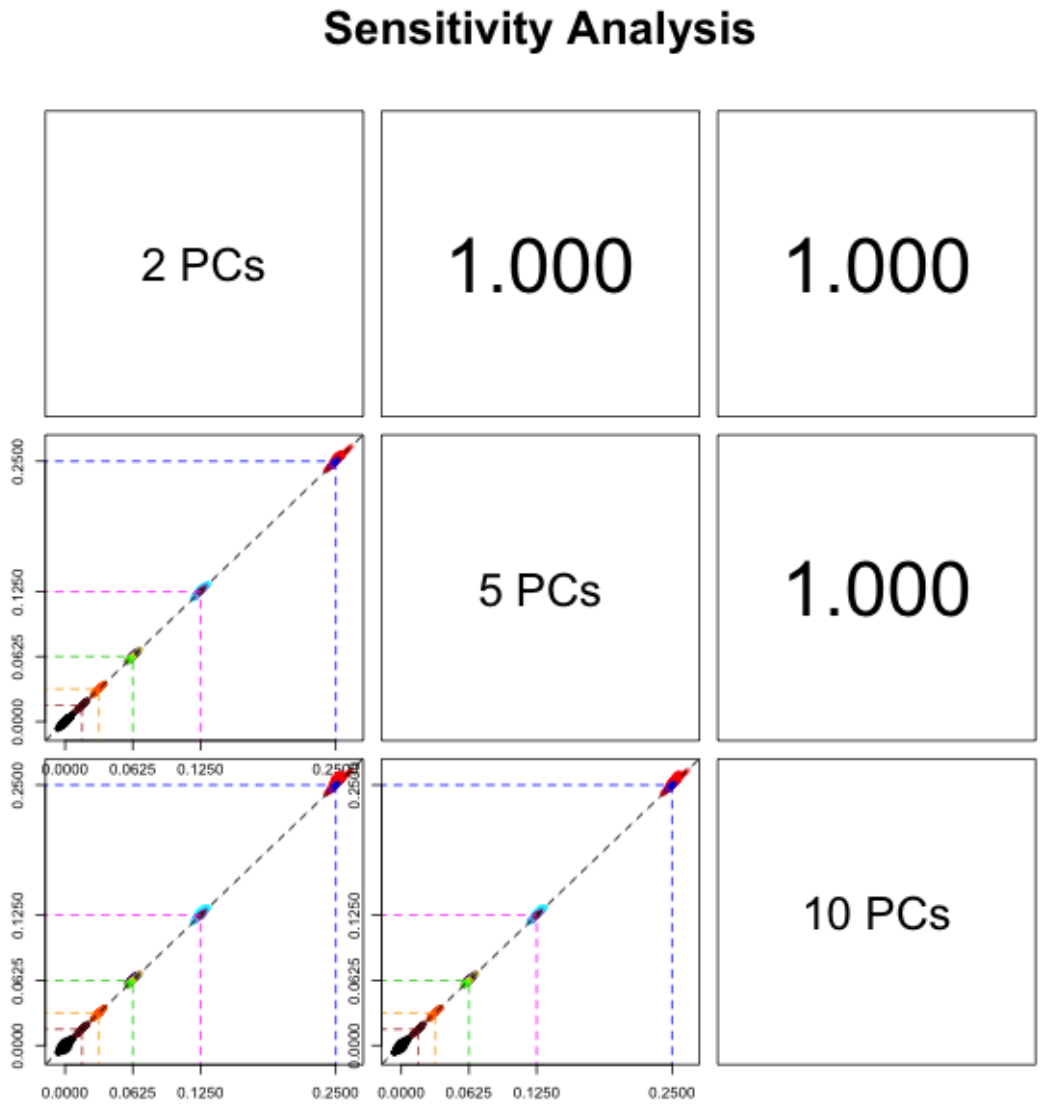


Figure 4.20: Sensitivity Analysis for Relationship Configuration I under Population Structure II

The lower panels in the pairs plot show scatter plots directly comparing the PC-Relate kinship coefficient estimates when using 2, 5, or 10 PCs to adjust for ancestry (only 2 truly reflect ancestry). The upper panels give the pairwise correlations of the estimated kinship coefficients restricted to pairs of individuals that either analysis identified as relatives.

suitable reference panels are available, accurate estimates of ancestry proportions can often be obtained from model based methods. PC-Relate can use vectors of ancestry proportions for all sample individuals in place of principal components for the computation of individual-specific allele frequencies and ancestry adjusted relatedness estimates. Additionally, PC-Relate should be suitable for relatedness inference with rare variants and sequence data. Each term of the estimator for the standard GRM involves division by $p_s(1-p_s)$, resulting in unstable estimates when rare variants are included. PC-Relate, however, uses estimators that take the ratio of the averages over loci of the numerator and denominator terms, preventing individual markers, even those with very low MAF, from having too much influence on the value of the estimate. The performance of PC-Relate with sequence data is an area of future research.

In Chapter 5 we demonstrate a procedure to combine ancestry inference with PC-AiR (Chapter 3) and relatedness estimation with PC-Relate (Chapter 4) for samples with known or cryptic relatedness and unspecified population structure. In Chapter 6 we show that when including ancestry representative principal components from PC-AiR as fixed effects in LMMs, the natural covariance structure of the polygenic random effect is the matrix of PC-Relate kinship coefficients.

Chapter 5

DECONVOLUTING ANCESTRY AND RELATEDNESS WITH PC-AIR AND PC-RELATE

In this brief chapter we tie together the work of the previous two chapters and present a procedure utilizing PC-AiR 3 and PC-Relate 4 for attaining accurate ancestry inference as well as accurate relatedness estimation in a sample with both known and cryptic relatives from a population with unspecified structure. It is probably apparent that ancestry inference with PC-AiR relies on kinship coefficient estimates, while relatedness estimation with PC-Relate relies on ancestry representative principal components. This begs the question of where to start when both structures are unknown, as is often the case. PC-AiR can use kinship coefficient estimates from methods such as KING-robust, but as shown with the WHI SHARe Hispanic cohort analysis in Section 4.3.4, these estimates may be quite biased. To address this issue, an iterative procedure for deconvoluting ancestry and relatedness is presented in Algorithm 5.1.

This iterative procedure is required when the initial kinship coefficient estimates are unreliable. Pedigrees often have mis-annotated relationships and can not account for cryptic relatives. As shown in Chapter 4, KING-robust provides biased estimates in admixed populations, and it may both fail to identify true relatives as well as incorrectly identify unrelated pairs as relatives. Both of these biases can be problematic, but the iterative procedure suggested can correct mis-specification in most scenarios.

We do not provide a specific criteria for how many iterations to run or when to declare that the estimates are stable. This is intentional, as the criteria should probably be considered on a study by study basis. Reasonable measures to examine include correlation between ancestry representative principal components from one

Algorithm 5.1 Iterative Procedure to Deconvolute Ancestry and Relatedness

(1) Estimate ancestry divergence for all pairs of individuals with KING-robust.

(2) Find initial estimates of kinship coefficients for all pairs of individuals.

- If it is strongly believed that pedigrees are reliable, pedigree values can be used.
- If reference panel samples are available and reliable, REAP or RelateAdmix can be used.
- If no good information is available a priori, KING-robust can be used.

(3) Perform PC-AiR using the ancestry divergence estimates from KING-robust and the initial kinship coefficient estimates.

(4) Perform PC-Relate using the initial principal components from PC-AiR.

(5) Perform PC-AiR using the PC-Relate estimates of kinship coefficients and the KING-robust estimates of ancestry divergence.

(6) Perform PC-Relate using the new principal components from PC-AiR.

(7) If necessary, iterate (5) and (6) until the estimates are stable.

iteration to the next as well as correlation between kinship coefficients for identified pairs of relatives from one iteration to the next. High correlations are indicative of stable estimates that have converged. For the WHI SHARe Hispanic cohort analysis in Section 4.3.4, no further iterations of the procedure were necessary (i.e. Step (7) of Algorithm 5.1 was not required). One additional iteration was performed, but it did not substantially alter results from the previous iteration. We have seen similar results in other analyses of real data.

Chapter 6

LINEAR MIXED MODEL TO PROPERLY AND EFFICIENTLY ACCOUNT FOR POPULATION STRUCTURE AND RELATEDNESS IN GENETIC ASSOCIATION STUDIES

6.1 Introduction

To date, hundreds of thousands of individuals have been included in genome wide association studies (GWAS), leading to the discovery of numerous associations of genetic polymorphisms with a variety of traits and diseases. However, few of these studies have focused on populations with multi-way admixture, including Hispanic Americans and African Americans, the two largest minority populations in the United States. Studying admixed populations presents a challenge, as there are typically two types of genetic structure that must be accounted for in genetic association studies: ancestry (population structure) and relatedness (family structure). Failure to appropriately and efficiently account for both of these genetic structures can result in spurious associations, as well as reduced power to detect true associations.

A popular approach for GWASs is to use a linear mixed model (LMM) that simultaneously accounts for both population structure and relatedness among sample individuals by including a genome-wide genetic relationship matrix (GRM) as part of the covariance structure of the phenotype. Several LMM methods with slightly different computational approaches have been developed recently, including EMMAX [21], GRAMMAR-Gamma [53], FaST-LMM [27,28], GEMMA [72], and GCTA-LOCO [65, 68]. These LMM methods control inflation of test statistics on average across the genome for samples with subtle population structure [42, 63] ; however, it has been shown that they may not adequately control for false positives at SNPs that

are unusually differentiated between ancestral populations [42]. As the differentiation in allele frequencies between ancestral populations varies greatly across the genome, proper calibration of test statistics from these methods should be a concern when studying admixed populations, even at SNPs that are not unusually differentiated. We have found that modeling the covariance structure of the phenotype using a GRM calculated genome-wide controls type I error well at SNPs with an a typical amount of differentiation, but it also leads to a systematic inflation or deflation of test statistics for SNPs that are more or less differentiated. Further, existing LMM approaches use GRMs calculated with allele frequencies estimated from the entire sample. In admixed populations, however, it has been shown that this can lead to inflated pairwise kinship estimates [56], and this may lead to reduced power to detect true associations.

Here we propose an LMM approach, MMAAPS, for genetic association testing that is well-calibrated, even in highly structured populations that include ancestry admixture. MMAAPS (mixed model association in admixed populations) appropriately and efficiently accounts for both population structure and relatedness in the sample by (1) including ancestry representative principal components (PCs) calculated with PC-AiR (Chapter 3) as fixed effects covariates in the mean model to account for ancestry differences, and (2) including an ancestry-adjusted GRM calculated with PC-Relate (Chapter 4) in the covariance structure to account for familial relationships. We demonstrate through simulation studies that including ancestry representative PCs in addition to modeling the phenotypic covariance structure with an ancestry-adjusted GRM accounts for varying ancestral allele frequency differences across SNPs, correcting the inflation/deflation issues in other methods, and obtaining well-calibrated test statistics genome-wide. Additionally, we show that MMAAPS can provide a modest increase in statistical power for detecting genetic associations with phenotypes that are correlated with ancestral background.

A further advantage of MMAAPS over existing methods is that it allows for the inclusion of multiple random effects. Current LMM approaches for GWAS only allow

for modeling one random effect - the background polygenic variance captured by the GRM. However, there may be further covariance structure of the phenotype among samples due to shared environment. Shared environmental effects may arise in large genetic studies where individuals from the same household are included, or when complex sampling procedures are implemented. We perform simulations to assess the importance of modeling these additional random effects in a variety of settings. We show that existing LMM methods that do not model correlation due to shared environment may still provide well-calibrated test statistics at null SNPs and can have type I error rates that are not significantly different from the nominal level. However, we also show that accurately modeling shared environment using MMAAPS can provide more efficient tests and a substantial increase in power to detect causal SNPs. To what extent the power is increased with MMAAPS depends on the proportion of variance accounted for by the shared environmental effects as well as the cluster sizes of individuals who share a common effect.

To compare the performance of MMAAPS to existing LMM methods in a real data example, we analyze traits from the Hispanic cohort of the Women's Health Initiative SNP Health Association Resource (WHI-SHARe) Study. This cohort consists of 3,587 women of self-identified Hispanic descent. It is well known that the Hispanic American population is admixed from multiple, previously isolated, continental progenitor groups including Native American, European, and African ancestries, with complex population substructure on both continental and sub-continental levels. Additionally, over 100 pairs of close relatives have been identified in this cohort [56], making this study particularly ideal for an association analysis with MMAAPS. Consistent with our simulation study results, MMAAPS achieves slightly more significance than existing LMM methods at SNPs known to be associated with white blood cell counts [45], providing smaller p-values and reaching genome-wide significance at one additional SNP in the region.

6.2 Methods

6.2.1 Polygenic Trait Model

The phenotypic value for an individual is the observed quantity for a trait measured on some scale. Consider a sample of individuals, $\mathcal{N}$, and let $\mathbf{Y}$ be a length $|\mathcal{N}|$ vector of their values for some quantitative trait of interest. In what follows, we consider the simplest form of Fisher's (1918) polygenic trait model [10]. Under the assumptions that there are no gene-environment interactions and that genetic loci act independently, the vector of trait values can be partitioned as

$$\mathbf{Y} = \sum_{l \in \mathcal{L}} \mathbf{G}_l + \mathbf{E}, \tag{6.1}$$

where $\mathcal{L}$ is the set of all loci in the genome, $\mathbf{G}_l$ is the genotypic contribution from locus l, and $\mathbf{E}$ is the environmental contribution. If we further assume that the effects at each locus act additively, then the covariance structure of the trait can be written as

$$\begin{aligned} \text{Cov}[\mathbf{Y}] &= \sum_{l \in \mathcal{L}} \sigma^2_{A_l} \mathbf{\Phi} + \sigma^2_E \mathbf{I}_{|\mathcal{N}|} \\ &= \sigma^2_A \mathbf{\Phi} + \sigma^2_E \mathbf{I}_{|\mathcal{N}|}, \end{aligned} \tag{6.2}$$

where $\sigma^2_{A_l}$ is the contribution to the additive genetic variance from locus l, $\sigma^2_A = \sum_{l \in \mathcal{L}} \sigma^2_{A_l}$ is the total additive genetic variance of the trait, σ^2_E is the environmental variance of the trait, $\mathbf{\Phi}$ is a relationship matrix with $[i, j]^{th}$ element equal to $2\phi_{ij}$, where ϕ_{ij} is the pairwise kinship coefficient for individuals $i, j \in \mathcal{N}$, and $\mathbf{I}_{|\mathcal{N}|}$ is an ($|\mathcal{N}|$x$|\mathcal{N}|$) identity matrix. The covariance structure in Equation 6.2 assumes that the environmental variances for each individual are independent, but this can easily be extended by including additional terms that represent shared environmental covariances.

6.2.2 Linear Mixed Model for GWAS

Now consider that the sample $\mathcal{N}$ has been genotyped at a set $\mathcal{S}$ of SNP markers. The polygenic model for the trait given in Equation 6.1 can be expressed in a linear regression framework as

$$\mathbf{Y} = \sum_{s \in \mathcal{S}} \mathbf{g}_s \beta_s + \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{6.3}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{|\mathcal{N}|}),$$

where $\mathbf{g}_s$ is the vector of genotypes for all individuals at SNP s, β_s is the effect of SNP s on the trait value, $\mathbf{X}$ is a matrix of environmental covariates including an intercept with corresponding effect sizes $\boldsymbol{\alpha}$, and $\boldsymbol{\epsilon}$ is a random variable that captures unaccounted for environmental effects. For now we assume that these random environmental effects ϵ_i are i.i.d. The goal of a GWAS is to determine if the genotype value at each of the SNPs is associated with the observed phenotypic value; i.e to determine which $\beta_s \neq 0$ for all $s \in \mathcal{S}$. However, the set $\mathcal{S}$ typically contains hundreds of thousands or millions of SNPs genotyped in the genome-screen, so $|\mathcal{S}| \gg |\mathcal{N}|$, and this multivariate regression can not be directly fit. Instead, SNPs are usually tested for association one at a time.

When testing for association at SNP $s' \in \mathcal{S}$, the multivariate regression equation can be rewritten as

$$\mathbf{Y} = \mathbf{g}_{s'} \beta_{s'} + \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}', \tag{6.4}$$

where the new error term, $\boldsymbol{\epsilon}'$, is a function of unaccounted for environmental effects as well as the polygenic effects from all of the other SNPs. We refer to this set of SNPs that excludes s' as the set $\mathcal{S}/s'$. Therefore, this new error term can be expressed as $\boldsymbol{\epsilon}' = \sum_{s \in \mathcal{S}/s'} \mathbf{g}_s \beta_s + \boldsymbol{\epsilon} = \sum_{s \in \mathcal{S}/s'} \mathbf{z}_s u_s + \boldsymbol{\epsilon}$, where $\mathbf{z}_s = (\mathbf{g}_s - 2p_s\mathbf{1})/[2p_s(1-p_s)]^{1/2}$ is the vector of standardized genotypes at SNP s, and u_s is the rescaled genotype effect on the trait. In order to account for this background polygenic effect, it is common practice to assume that the genotype effects for causal SNPs in the set $\mathcal{S}/s'$

are independent random effects drawn from a common distribution (non-causal SNPs all have effect size 0). We refer to this set of causal SNPs as $\mathcal{S}_c \subset \mathcal{S}/s'$. The linear regression model in Equation 6.4 can then be reformulated as the linear mixed effects model

$$\begin{aligned} \mathbf{Y} &= \mathbf{g}_{s'}\beta_{s'} + \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \\ \mathbf{u} &\sim \mathrm{N}(\mathbf{0}, \sigma_u^2 \mathbf{I}_{|\mathcal{S}_c|}) \\ \boldsymbol{\epsilon} &\sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{|\mathcal{N}|}), \end{aligned} \tag{6.5}$$

where $\mathbf{Z}$ is an $(|\mathcal{N}|\mathrm{x}|\mathcal{S}_c|)$ standardized genotype matrix with columns $\mathbf{z}_s$ for $s \in \mathcal{S}_c$, and $\mathbf{u}$ is a length $|\mathcal{S}_c|$ vector of the corresponding causal SNP effects. It can be seen from the linear mixed model equation that the covariance structure of the trait is

$$\begin{aligned} \boldsymbol{\Sigma} \equiv \mathrm{Cov}[\mathbf{Y}] &= \mathbf{Z}\mathrm{Cov}[\mathbf{u}]\mathbf{Z}^T + \mathrm{Cov}[\boldsymbol{\epsilon}] \\ &= \sigma_u^2 \mathbf{Z}\mathbf{Z}^T + \sigma_\epsilon^2 \mathbf{I}_{|\mathcal{N}|} \\ &= \sigma_{A_{\mathcal{S}/s'}}^2 \frac{1}{|\mathcal{S}_c|} \mathbf{Z}\mathbf{Z}^T + \sigma_\epsilon^2 \mathbf{I}_{|\mathcal{N}|}, \end{aligned} \tag{6.6}$$

where $\sigma_{A_{\mathcal{S}/s'}}^2 = |\mathcal{S}_c|\sigma_u^2 = \sum_{\mathcal{S}_c} \sigma_u^2$ is the total additive genetic variance of the trait, excluding the effect of SNP s', and the parameter σ_ϵ^2 is the residual variance.

Generalized least squares (GLS) can be used to fit this linear mixed model and test the null hypothesis that $\beta_{s'} = 0$. Since the covariance structure of the trait, $\boldsymbol{\Sigma}$, is unknown, it must first be estimated. In practice, the set $\mathcal{S}_c$ of causal SNPs is unknown a priori, so the covariance matrix $\frac{1}{|\mathcal{S}_c|}\mathbf{Z}\mathbf{Z}^T$ is estimated by the standard genetic relationship matrix (GRM), $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}/s'}$, constructed from the SNPs in the set $\mathcal{S}/s'$ [64]. Estimates of the variance components $\sigma_{A_{\mathcal{S}/s'}}^2$ and σ_ϵ^2 are obtained, typically using restricted maximum likelihood (REML) [65], and an estimate of $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\sigma}_{A_{\mathcal{S}/s'}}^2 \widehat{\boldsymbol{\Psi}}_{\mathcal{S}/s'} + \widehat{\sigma}_\epsilon^2 \mathbf{I}_{\mathcal{N}}. \tag{6.7}$$

Model fitting is then performed with GLS using this estimated covariance structure. Re-computing $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}/s'}$ and re-estimating $\widehat{\sigma}_{A_{\mathcal{S}/s'}}^2$ and $\widehat{\sigma}_\epsilon^2$ for every choice of $s' \in \mathcal{S}$ is

computationally expensive, and efficient approaches have been a major point of study in recent years. EMMAX [21] implements an approximation by assuming that most SNPs have small to no effect on a trait; a seemingly reasonable assumption for the majority of complex quantitative traits [33]. Under this assumption, $\sigma^2_{A_{\mathcal{S}/s'}} \approx \sigma^2_{A_{\mathcal{S}}}$ for every choice of s', so the variance components and $\boldsymbol{\Sigma}$ can be estimated once under the null model of no SNP effect using a matrix $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}}$ generated from all of the SNP data. GEMMA also uses the relationship matrix $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}}$, but it avoids this approximation and computes the exact test by using a clever optimization algorithm to efficiently recompute the variance components and $\widehat{\boldsymbol{\Sigma}}$ for each SNP tested. The results from both of these methods tend to be similar, with the possibility of a loss of power when using the approximation for testing SNPs with large effect sizes [72].

Other papers have demonstrated a potential loss in power from using $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}}$, as data is double counted by including the SNP that is being tested for association, or SNPs in close LD with that SNP, in the relationship matrix [27, 28, 68]. Since recomputing $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}/s'}$ for every choice of s' is infeasible genome-wide, different approaches have been taken for selecting a subset of SNPs, $\mathcal{S}^* \subset \mathcal{S}$, to use for constructing a suitable GRM, $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}^*}$. FaST-LMM [27] uses an equally spaced subset of random SNPs, FaST-LMM-Select [28] uses the subset of SNPs that are most highly correlated with the trait, and GCTA-LOCO [65, 68] uses a leave-one-chromosome-out approach, where $\mathcal{S}^*$ is the set of all SNPs not on the same chromosome as the SNP being tested. All of these approaches may improve statistical power in certain settings; however, if the set $\mathcal{S}^*$ is too small, $\widehat{\boldsymbol{\Psi}}_{\mathcal{S}^*}$ may not adequately adjust for sample structure, and test statistics at null SNPs may be inflated [68].

It is worth noting that the covariance structure of the trait specified by the linear mixed model and derived in Equation 6.6 takes a very similar form to the theoretical phenotype covariance structure presented in Equation 6.2. In fact, the standard GRM, $\widehat{\boldsymbol{\Psi}}$, generated from dense genome-wide SNP data is a consistent estimator of the relationship matrix, $\boldsymbol{\Phi}$, in a homogeneous population. On the other hand, in the

presence of population structure, the elements of $\widehat{\mathbf{\Psi}}$ are inflated for individuals with similar ancestry, and this matrix is no longer consistent for $\mathbf{\Phi}$ (see Appendix A.2.1).

6.2.3 LMM for GWAS with Population Structure

When there is population structure in the sample $\mathcal{N}$, ancestry may be a confounder for the SNP-phenotype relationship. Ancestry representative PCs should be included as fixed effects covariates in the mean model to remove this confounding effect and prevent spurious associations. The linear regression model including PCs is

$$\begin{aligned} \mathbf{Y} &= \sum_{s \in \mathcal{S}} \mathbf{g}_s \beta_s + \mathbf{X}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathrm{N}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_{|\mathcal{N}|}), \end{aligned} \tag{6.8}$$

where the columns of the matrix $\mathbf{V}$ are the set of ancestry representative PCs and the intercept (no longer in the $\mathbf{X}$ matrix), and $\boldsymbol{\gamma}$ is the corresponding vector of effect sizes. Equivalently, by the Frisch-Waugh Lovell theorem this model can be written as

$$\begin{aligned} \widetilde{\mathbf{Y}} &= \sum_{s \in \mathcal{S}} \widetilde{\mathbf{g}}_s \beta_s + \widetilde{\mathbf{X}}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathrm{N}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_{|\mathcal{N}|}), \end{aligned} \tag{6.9}$$

where $\mathbf{M}_\mathbf{V} = \left(\mathbf{I}_{|\mathcal{N}|} - \mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\right)$, and $\widetilde{\mathbf{Y}} = \mathbf{M}_\mathbf{V}\mathbf{Y}$, $\widetilde{\mathbf{g}}_s = \mathbf{M}_\mathbf{V}\mathbf{g}_s$, and $\widetilde{\mathbf{X}} = \mathbf{M}_\mathbf{V}\mathbf{X}$ are the residuals of the phenotypes, genotypes, and environmental covariates, respectively, adjusted for the PCs and intercept. Since testing is performed one SNP at a time in practice, the model for SNP s' is

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{g}}_{s'} \beta_{s'} + \widetilde{\mathbf{X}}\boldsymbol{\alpha} + \boldsymbol{\epsilon}', \tag{6.10}$$

where $\boldsymbol{\epsilon}' = \sum_{s \in \mathcal{S}/s'} \widetilde{\mathbf{g}}_s \beta_s + \boldsymbol{\epsilon} = \sum_{s \in \mathcal{S}/s'} \widetilde{\mathbf{z}}_s u_s + \boldsymbol{\epsilon}$. The residuals from the regression of the genotype vector on the PCs and intercept can also be written as $\widetilde{\mathbf{g}}_s = \left(\mathbf{g}_s - \widehat{\mathbb{E}}[\mathbf{g}_s|\mathbf{V}]\right)$. The quantity $\widehat{\mathbb{E}}[\mathbf{g}_s|\mathbf{V}]$ is an estimator of twice the vector of individual-specific allele frequencies at SNP s, as used in PC-Relate (Section 4.2.2), so the residualized genotype

vector can be rewritten as $\widetilde{\mathbf{g}}_s = (\mathbf{g}_s - 2\widehat{\boldsymbol{\mu}}_s)$, and a standardized vector of genotypes can be created as $\widetilde{\mathbf{z}}_s = (\mathbf{g}_s - 2\widehat{\boldsymbol{\mu}}_s)/[2\widehat{\boldsymbol{\mu}}_s(1-\widehat{\boldsymbol{\mu}}_s)]^{1/2}$. By the same argument as before, translating this into a linear mixed model gives

$$\begin{aligned}\widetilde{\mathbf{Y}} &= \widetilde{\mathbf{g}}_{s'}\beta_{s'} + \widetilde{\mathbf{X}}\boldsymbol{\alpha} + \widetilde{\mathbf{Z}}\mathbf{u} + \boldsymbol{\epsilon} \\ \mathbf{u} &\sim \mathrm{N}(\mathbf{0}, \sigma_u^2 \mathbf{I}_{|\mathcal{S}_c|}) \\ \boldsymbol{\epsilon} &\sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{|\mathcal{N}|}),\end{aligned} \tag{6.11}$$

where $\widetilde{\mathbf{Z}}$ is an ($|\mathcal{N}|$x$|\mathcal{S}_c|$) matrix of ancestry-adjusted standardized genotypes for $s \in \mathcal{S}_c$, and $\mathbf{u}$ is the length $|\mathcal{S}_c|$ vector of corresponding scaled causal SNP effects. The covariance structure of the phenotype under this model can be expressed as

$$\begin{aligned}\boldsymbol{\Sigma} \equiv \mathrm{Cov}[\widetilde{\mathbf{Y}}] &= \widetilde{\mathbf{Z}}\mathrm{Cov}[\mathbf{u}]\widetilde{\mathbf{Z}}^T + \mathrm{Cov}[\boldsymbol{\epsilon}] \\ &= \sigma_u^2 \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T + \sigma_\epsilon^2 \mathbf{I}_{|\mathcal{N}|} \\ &= \sigma^2_{A_{\mathcal{S}/s'}} \frac{1}{|\mathcal{S}_c|}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T + \sigma_\epsilon^2 \mathbf{I}_{|\mathcal{N}|},\end{aligned} \tag{6.12}$$

which is very similar to that in Equation 6.6, except with the matrix $\mathbf{Z}$ replaced by the matrix $\widetilde{\mathbf{Z}}$. The interpretation of the variance components is less clear than in a homogeneous population, but the parameter $\sigma^2_{A_{\mathcal{S}/s'}}$ is a measure of the variance accounted for by the sharing of alleles IBD among familial relatives, and σ^2_ϵ is still the residual variance. Once again, the set $\mathcal{S}_c$ of causal SNPs is unknown a priori, so the ancestry-adjusted covariance matrix $\frac{1}{|\mathcal{S}_c|}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T$ is estimated in practice from all of the SNPs in the set $\mathcal{S}/s'$. This estimated matrix turns out to be $\widehat{\boldsymbol{\Phi}}^{PC}_{\mathcal{S}/s'}$, the matrix of twice the PC-Relate kinship coefficient estimates found using SNPs in the set $\mathcal{S}/s'$.

This linear mixed model can still be fit with the same GLS procedure described for the model without PCs, but the result of the derivation in Equation 6.12 suggests that it is appropriate to use an ancestry adjusted relationship matrix when the mean model includes PCs. Fitting models that include PCs as fixed effects in addition to a standard GRM as a random effect has been proposed [65, 66], but a subsequent paper [20] demonstrated that adjusting for the same structure twice leads to erratic

variance component estimates due to double counting the data. Further, the covariance structure of the trait specified by the recommended linear mixed model once again reflects the theoretical covariance structure given in Equation 6.2. The same approximate and exact computational approaches for estimating the variance components and the covariance structure of the trait are still valid under the same assumptions, and the same considerations regarding what set of SNPs to use for constructing the relationship matrix to optimize power while preventing spurious association still apply.

6.2.4 Including Additional Random Effects

Until this point, we've assumed that the environmental variance for each individual in the sample is independent. However, this assumption often may not hold, for example, when individuals share the same household or are sampled from the same geographic region. In this situation, we may want to include additional random effects in the linear mixed model to capture these common environmental effects. For the purpose of fitting this model to test SNPs for association, this can be accomplished by including additional terms for shared environmental covariances in the estimate of the trait covariance structure:

$$\widehat{\mathbf{\Sigma}} = \widehat{\sigma}^2_{A_{\mathcal{S}*}} \widehat{\mathbf{\Phi}}^{PC}_{\mathcal{S}*} + \sum_{r=1}^{R} \widehat{\sigma}^2_{C_r} \mathbf{C}_r + \widehat{\sigma}^2_{\epsilon} \mathbf{I}_{|\mathcal{N}|}. \tag{6.13}$$

The parameter $\sigma^2_{C_r}$ in this model is the variance accounted for by environmental effect $r \in \{1 \dots R\}$, and $\mathbf{C}_r$ is a matrix indicating which pairs of individuals have a common value for this environmental effect. For example, when modeling a shared household effect, the $[i, j]^{th}$ entry of $\mathbf{C}_r$ is 1 if individuals i and j belong to the same household, and it is 0 if they do not. The parameter σ^2_{ϵ} can be interpreted as the residual variance unaccounted for by the polygenic and common environmental effects. Estimating these variance components is still performed via REML. However, the computational approach implemented in GEMMA to quickly recompute the variance

component estimates and $\widehat{\boldsymbol{\Sigma}}$ for each SNP tested is not applicable with multiple random effects, so the exact test can not be efficiently implemented genome-wide. As a result, it is necessary to use an approximation and estimate the variance components under the null model, either once for all SNPs genome-wide, or possibly once for each chromosome when using a leave-one-chromosome-out approach.

6.2.5 Simulated Data

We simulate genotype data at two sets of 100,000 independent SNPs, which we refer to as Set 1 and Set 2, for a sample of 3,000 individuals with ancestry derived from two divergent populations. Allele frequencies for the two populations at all SNPs are generated using the Balding Nichols model [4]. More precisely, for each SNP s, the allele frequency p_s in the ancestral population is drawn from a uniform distribution on $[0.1, 0.9]$, and the allele frequency in population $k \in \{1, 2\}$ is drawn from a beta distribution with parameters $p_s(1 - F_k)/F_k$ and $(1 - p_s)(1 - F_k)/F_k$, where the quantity F_k is equivalent to Wright's measure of population differentiation [62] from the ancestral population. We set F_1 and F_2 equal to a common value, $F_{ST} = 0.15$, representing highly divergent continental populations. The simulated sample includes both related and unrelated individuals; 30 four-generation pedigrees, where each pedigree has a total of 20 individuals (Figure 6.1), 120 cousin pairs, and 2,160 individuals unrelated to anyone else in the sample. The ancestry proportions for each individual can be represented as a and $(1 - a)$ from population 1 and 2 respectively, where the parameter a is drawn from a beta distribution with mean 0.3 and standard deviation 0.1 for one third of unrelated samples and pedigree founders, a beta distribution with mean 0.7 and standard deviation 0.1 for another third, and a uniform distribution on $[0, 1]$ for the remaining third. All founders within the same pedigree have a drawn from the same distribution. Pedigrees with admixture proportions drawn from a beta distribution have ancestry-related assortative mating, while those with admixture proportions drawn from a uniform distribution have random mating that

is independent of ancestry. Genotypes for unrelated samples and pedigree founders are drawn according to the the simulated population allele frequencies and individual admixture proportions. Genotypes for pedigree descendants and the cousin pairs are generated by passing alleles down the pedigree, and admixture proportions are calculated from the ancestry of their relatives.

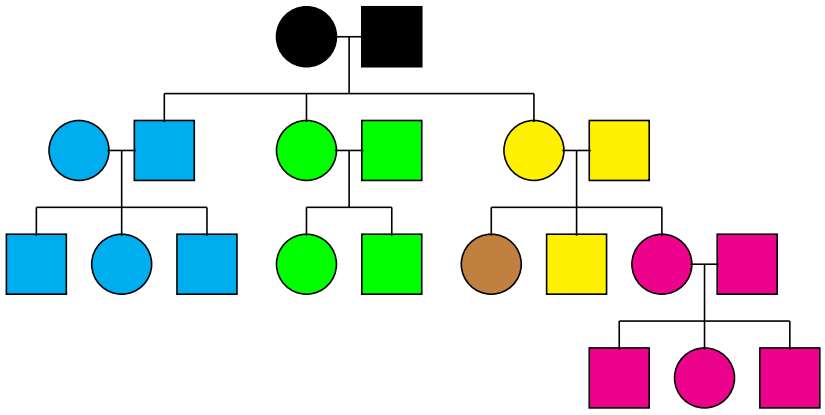


Figure 6.1: Pedigree Configuration for Simulated Data
The pedigree configuration for each of the 30 outbred, four-generation pedigrees included in the simulated data, where the overall structure of each pedigree is as depicted, but the pattern of ancestry admixture varies. For simulations that include household effects, the six different colors represent distinct households within the pedigree.

6.3 Results

6.3.1 Traits Correlated with Ancestry

We performed simulation studies to demonstrate that existing LMM methods that only utilize a GRM to account for sample structure have improper calibration at null SNPs for traits correlated with ancestry due to varying allele frequency differentiation between ancestral populations. We also demonstrate that including ancestry

representative PCs in the mean model, in addition to modeling the trait covariance structure with a relationship matrix, can correct this mis-calibration, providing appropriate test statistics and controlling Type-I error rate genome-wide.

This simulation study used simulated genotype data so that we could know details of the true underlying populations. Trait values were simulated for 1,000 heritable phenotypes correlated with genome-wide ancestry according to the model

$$\mathbf{Y} = \sum_{s \in \mathcal{S}_c} \mathbf{g}_s \beta_s + 2\mathbf{a}^1 + \boldsymbol{\epsilon}, \tag{6.14}$$

where $\mathcal{S}_c$ is a set of 100 causal SNPs from Set 2, $\mathbf{a}^1$ is the vector of ancestry proportions for all individuals from population 1, and $\epsilon_i \sim \mathrm{N}(0,1)$ is individual random noise. The 100 SNPs from Set 2 chosen for $\mathcal{S}_c$ were different for each of the 1,000 phenotypes generated, and the genotype effect sizes, β_s, were chosen so that each SNP explained 0.1% of the variability of the trait using the formula

$$\beta_s = \left[\frac{h_s^2 \sigma_\epsilon^2 / \left(1 - \sum_{t \in \mathcal{S}_c} h_t^2\right)}{2\hat{p}_s(1-\hat{p}_s)} \right]^{1/2}, \tag{6.15}$$

where $h_s^2 = 0.001$ for every $s \in \mathcal{S}_c$, $\sigma_\epsilon^2 = 1$, and $\hat{p}_s$ is the sample average allele frequency. The 100,000 SNPs in Set 1 have no direct association with the trait value, but may be correlated due to an association through genome-wide ancestry or alleles shared IBD between relatives. These null SNPs were tested for association with each of the 1,000 simulated phenotypes using MMAAPS, EMMAX, GEMMA, EIGENSTRAT [40] (linear regression including PCs), and unadjusted linear regression. MMAAPS used the top principal component from PC-AiR performed on Set 1 to adjust the mean model for ancestry, and it used a relationship matrix generated from the kinship estimates of PC-Relate from Set 1 to account for correlation due to relatedness. EMMAX used the Balding Nichols matrix generated by the software, and GEMMA used the standardized relatedness matrix generated by the software, both with default filters, to account for sample structure. The EIGENSTRAT approach was performed using the top PC from PC-AiR to ensure that ancestry was accurately captured.

We define the quantity $D_s = (p_s^1 - p_s^2)$ to be the allele frequency difference between the two ancestral populations. Additionally, we define three classes of SNPs based on the absolute allele frequency difference between the ancestral populations: highly ($|D_s| \geqslant 0.4$), moderately ($0.4 > |D_s| \geqslant 0.2$), and lowly ($0.2 > |D_s| \geqslant 0$) differentiated. Of the 100,00 null SNPs, 10.9%, 28.6%, and 60.5% belong to each of these three classes respectively. The performance of each method for null SNPs was examined in two ways. First, LOWESS (locally weighted scatterplot smoothing) curves showing the relationship between the local mean of the test statistics from each method and $|D_s|$ are shown in Figure 6.2. The mean test statistic at null SNPs for a well-calibrated test should be 1, regardless of the allele frequency difference. Second, genomic control inflation factors [9], λ_{GC}, were calculated genome-wide, as well as in each of the three classes of SNPs, for each of the 1,000 simulated phenotypes. The genomic control inflation factor is used to evaluate confounding due to unaccounted for sample structure, where $\lambda_{GC} \approx 1$ indicates appropriate correction for population and family structure, while $\lambda_{GC} > 1$ indicates elevated type-I error rate. The means and standard errors of the genomic inflation factors are presented in Table 6.1.

MMAAPS is the only method that is well-calibrated for all null SNPs, showing no important inflation or deflation and obtaining genomic inflation factors near 1 genome-wide as well as in all three classes of SNPs. The test statistics from EIGENSTRAT are equally inflated across all values of $|D_s|$, and the genomic inflation factor is nearly the same (≈ 1.026) genome-wide as in all three classes of SNPs. EIGENSTRAT adjusts for the allele frequency differences in the mean model, but it does not account for the correlation of phenotypes among relatives, resulting in uniformly inflated test statistics. On the other hand, EMMAX and GEMMA do not adjust for ancestry differences in the mean model but do account for correlation in the covariance structure. The result is a genomic inflation factor near 1 when considering all SNPs genome-wide, but test statistics that are substantially inflated when $|D_s|$ is large and deflated when $|D_s|$ is small. The test statistics from unadjusted linear

regression (omitted from Figure 6.2) are very highly inflated for all values of $|D_s|$, and increasingly so as $|D_s|$ increases.

Table 6.1: Genomic Control λ_{GC} for Association Testing Simulation Study

Method	Genome-Wide	Highly[a] Differentiated	Moderately[b] Differentiated	Lowly[c] Differentiated
MMAAPS	1.000 (0.006)	0.999 (0.022)	1.001 (0.013)	1.001 (0.009)
EMMAX	1.001 (0.006)	1.098 (0.034)	1.016 (0.014)	0.979 (0.010)
GEMMA	1.004 (0.006)	1.110 (0.036)	1.020 (0.014)	0.980 (0.010)
EIGENSTRAT	1.026 (0.019)	1.025 (0.030)	1.027 (0.022)	1.026 (0.020)
Linear Reg.	20.10 (4.114)	134.6 (28.94)	52.23 (11.21)	7.006 (1.371)

Values presented are mean (s.e.) across all 1000 phenotype replicates.
[a] Highly differentiated SNPs: $|D_s| \geqslant 0.4$ between the two ancestral populations.
[b] Moderately differentiated SNPs: $0.4 > |D_s| \geqslant 0.2$ between the two ancestral populations.
[c] Lowly differentiated SNPs: $|D_s| < 0.2$ between the two ancestral populations.

Simulation studies were also performed to assess the power of each of the LMM methods to detect causal SNPs. Causal SNPs were generated with the following procedure. One of the 1,000 previously simulated heritable phenotypes correlated with genome-wide ancestry was selected as a base phenotype, $\mathbf{Y}_{\text{old}}$, and one additional SNP from Set 2 was selected to act as the causal SNP of interest. This procedure was repeated 10,000 times, one at a time, to generate 10,000 new phenotypes, each equal to the base phenotype plus an effect due to one selected causal SNP's genotype: $\mathbf{Y}_{\text{new}} = \mathbf{Y}_{\text{old}} + \mathbf{g}_{s'}\beta_{s'}$. The effect size, $\beta_{s'}$, for the causal SNP was again chosen using Equation 6.15, where the set $\mathcal{S}_c$ now consisted of the 100 original causal SNPs, each with $h^2_s = 0.001$, and the one new causal SNP of interest to be tested, s', with some pre-specified heritability $h^2_{s'}$. Each of the 10,000 new phenotype-genotype pairs were tested for association with MMAAPS, EMMAX, and GEMMA, and the proportion of significant associations was used to assess the statistical power of each method. This entire procedure was repeated for varying proportions of phenotypic variance, $h^2_{s'} \in \{0.0075, 0.0100, 0.0125, 0.0150\}$, accounted for by the causal SNP to be tested.

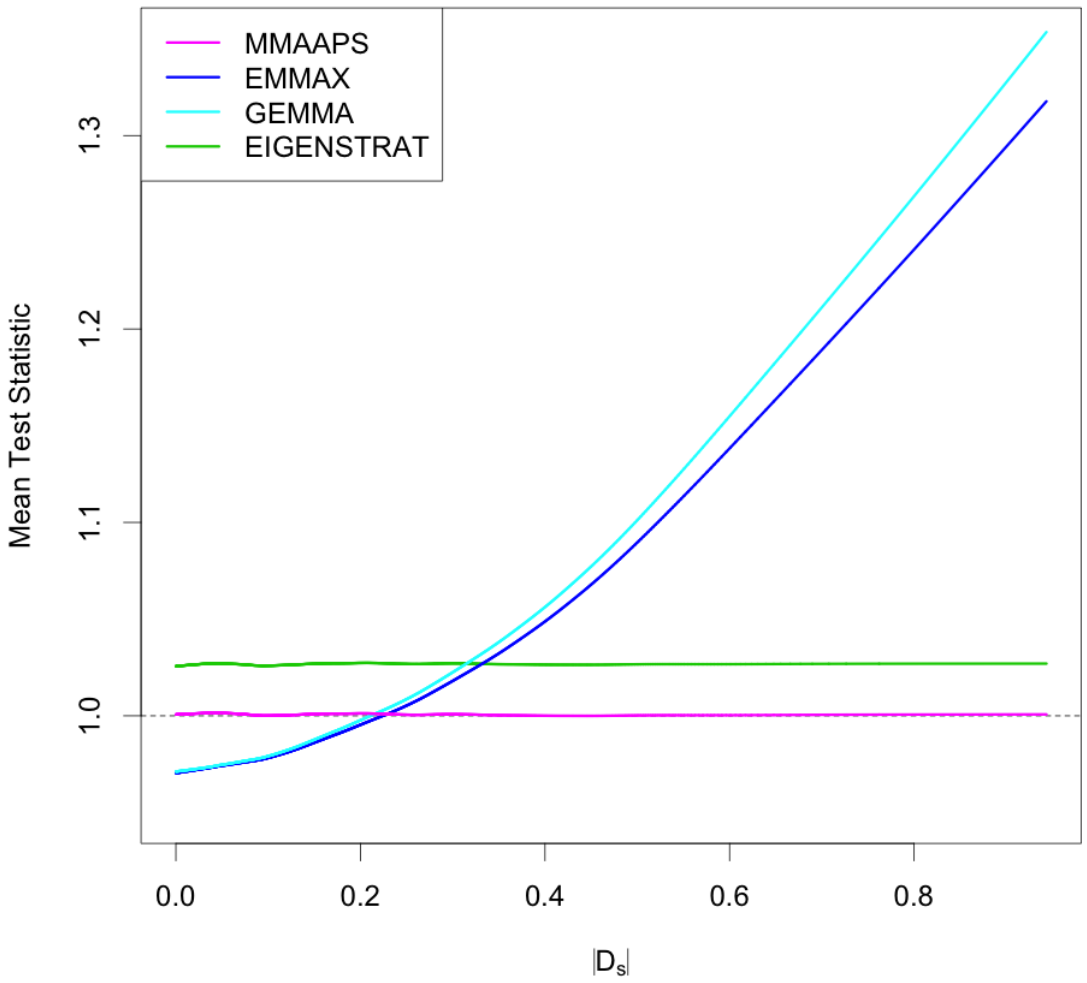


Figure 6.2: Mean Test Statistics by Ancestral Allele Frequency Difference LOWESS smooth curves showing the relationship between the local mean of the test statistics and the absolute difference in the allele frequencies of the two ancestral populations, $|D_s|$. The curves shown are the average relationship across all 1,000 simulated phenotypes. MMAAPS is the only method that is properly calibrated for all values of $|D_s|$.

The power of each of the LMM methods to detect a causal SNP in the simulated sample at significance level $\alpha = 5\text{x}10^{-8}$ for each choice of $h^2_{s'}$ is shown in Figure 6.3, and numeric values are presented for $h^2_{s'} = 0.0075$ at significance levels $\alpha = 5\text{x}10^{-8}$ and $\alpha = 5\text{x}10^{-6}$ in Table 6.2. The power for GEMMA is slightly greater than that for EMMAX, as expected, due to performing an exact test that estimates the variance components accounting for the causal SNP being tested. There is a modest increase in power of about $1-2\%$ genome-wide for MMAAPS as compared to the other LMM methods. The biggest gain in power for MMAAPS is in lowly differentiated SNPs, where we have shown that EMMAX and GEMMA provide over-corrected, deflated test statistics. The power gain for MMAAPS over GEMMA at highly differentiated SNPs is not as large; however, the increased number of true positive associations identified by GEMMA comes at the cost of many more false positive associations, as seen in Figure 6.4B. While GEMMA appears to have similar power for highly differentiated SNPs at the nominal level of the test, α, its true type I error rate is greater than α due to the inflation of its test statistics, and MMAAPS actually identifies many more true positives for a given honest false positive rate. A similar but less striking pattern is seen for moderately differentiated SNPs, where the inflation of test statistics for GEMMA and EMMAX is not as severe (Figure 6.4C). The comparison of true and false positive rates is nearly identical for all three methods at lowly differentiated SNPs (Figure 6.4D); GEMMA and EMMAX are less likely to identify true positives, but they are also conservative tests and less likely to identify false positives due to the deflation of their test statistics. While the potential power gain with MMAAPS is only modest in this setting, it is important to note that it suffers no appreciable loss in power while providing well-calibrated test statistics at all SNPs, providing better protection against false positives.

To investigate the relevance of these results in real populations, we examined select populations from release 3 of phase III of the International Haplotype Map Project (HapMap) [18]. Pairwise proportions of SNPs that are highly, moderately,

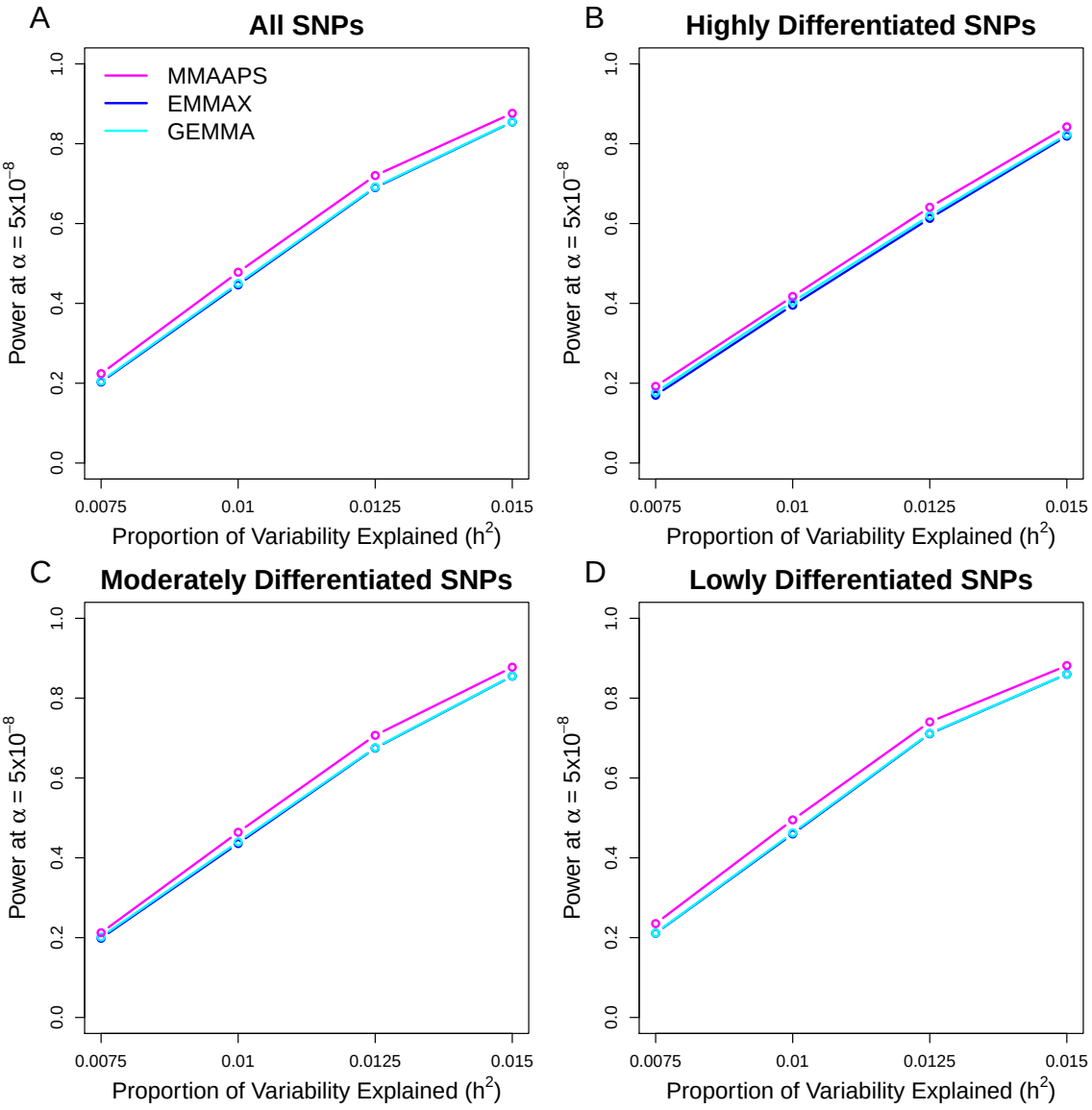


Figure 6.3: Power Curves for LMM Methods
The proportion of true positive associations identified (power) at a nominal significance level of $\alpha = 5\text{x}10^{-8}$ by MMAAPS, EMMAX, and GEMMA, both genome-wide and in all three classes of SNPs, is shown for each choice of h^2_s, for the causal SNP of interest.

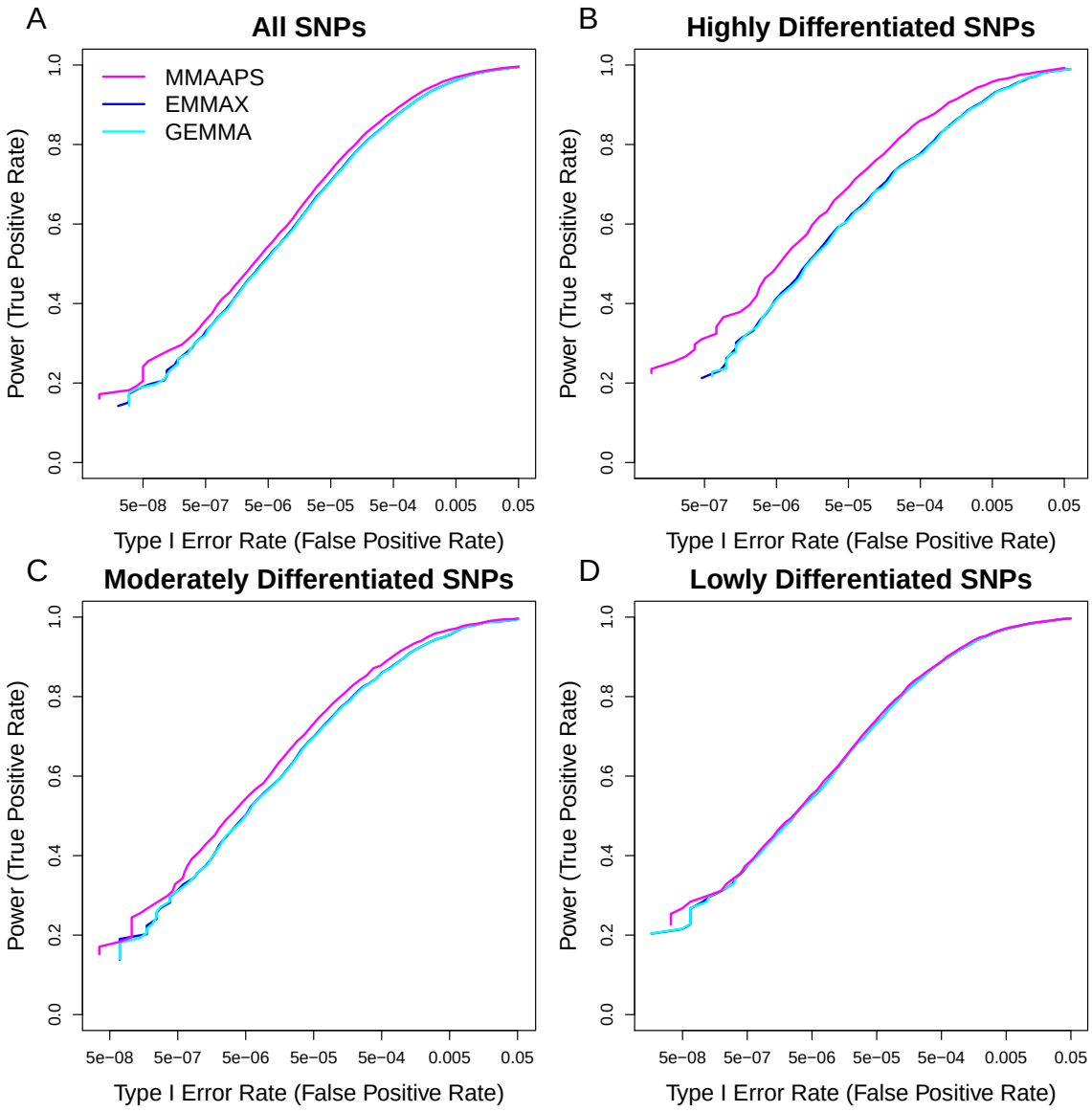


Figure 6.4: Comparison of True and False Positive Rates
The proportion of true positive associations identified (true positive rate) is compared to the proportion of null SNPs incorrectly identified as associations (false positive rate) by MMAAPS, EMMAX, and GEMMA, both genome-wide and in all three classes of SNPs. A higher curve indicates better performance. The honest false positive rate shown here may not match the nominal level of the test, α, for EMMAX and GEMMA due to mis-calibration of these tests at null SNPs.

Table 6.2: Power for LMM Methods with $h_s^2 = 0.0075$

Method	Genome-Wide	Highly[a] Differentiated	Moderately[b] Differentiated	Lowly[c] Differentiated
Power at Level $\alpha = 5\text{x}10^{-8}$				
MMAAPS	0.224	0.192	0.212	0.235
EMMAX	0.203	0.170	0.199	0.211
GEMMA	0.205	0.177	0.202	0.212
Power at Level $\alpha = 5\text{x}10^{-6}$				
MMAAPS	0.546	0.510	0.538	0.556
EMMAX	0.518	0.469	0.510	0.531
GEMMA	0.520	0.477	0.512	0.532

[a] Highly differentiated SNPs: $|D_s| \geqslant 0.4$ between the two ancestral populations.
[b] Moderately differentiated SNPs: $0.4 > |D_s| \geqslant 0.2$ between the two ancestral populations.
[c] Lowly differentiated SNPs: $|D_s| < 0.2$ between the two ancestral populations.

and lowly differentiated were estimated from 1,457,897 SNPs in the consensus data set. Not surprisingly, almost all SNPs are lowly differentiated between populations from the same continent. In contrast, a large proportion of SNPs have substantial allele frequency differences between populations from different continents, and the results are presented in Table 6.3. This result validates the concern of using existing LMM methods for association mapping in populations with cross continental admixture. African Americans, for example, have genetic contributions from European and African ancestral populations, and comparing the Northern European (CEU) and West African (YRI) populations, we see that approximately 9.5% of the SNPs are highly differentiated, and 26.2% are moderately differentiated.

It is also worth nothing that it seems likely that existing LMM methods may be poorly calibrated even in samples with ancestry admixture from less divergent populations. The ancestry correction provided by these methods is of appropriate size for SNPs with typical allele frequency differences. When the population divergence is less extreme, the typical allele frequency difference is smaller, resulting in a smaller ancestry correction. One would expect that this should still lead to deflation of test

statistics at the least differentiated SNPs and inflation of test statistics at the most highly differentiated SNPs. The amount of inflation is most likely not as severe, however, as the largest values of $|D_s|$ should presumably be smaller.

Table 6.3: Proportion of SNPs Highly and Moderately Differentiated Between HapMap Populations

	CEU	TSI	CHD	JPT	LWK	YRI
CEU	-	0.00	0.047	0.048	0.084	0.095
TSI	0.001	-	0.047	0.049	0.080	0.092
CHD	0.208	0.208	-	0.000	0.111	0.121
JPT	0.209	0.209	0.003	-	0.112	0.122
LWK	0.254	0.251	0.261	0.261	-	0.000
YRI	0.262	0.260	0.266	0.267	0.004	-

The upper half of the table gives the proportion of SNPs highly differentiated ($|D_s| \geqslant 0.4$) between the two populations. The lower half of the table gives the proportion of SNPs moderately differentiated ($0.4 > |D_s| \geqslant 0.2$) between the two populations.
CEU: Utah residents with Northern and Western European ancestry from the CEPH collection ($n = 165$)
TSI: Toscans in Italy ($n = 88$)
CHD: Chinese in Metropolitan Denver, Colorado ($n = 85$)
JPT: Japanese in Tokyo, Japan ($n = 86$)
LWK: Luhya in Webuye, Kenya ($n = 90$)
YRI: Yoruba in Ibadan, Nigeria ($n = 172$)

6.3.2 Traits with Shared Environmental Effects

We also performed simulations to explore the performance of each of the LMM methods for traits with shared environmental effects. MMAAPS is the only method that allows for additional random effects to account for the shared environmental covariances, and we demonstrate that modeling this structure can result in a significant increase in power to detect causal SNPs. This study used the same simulated genotype data as the previous subsection, and all individuals were assigned to both a household and a subgroup. Households within pedigrees were assigned according to Figure 6.1, and 120 additional two person households were created from unrelated

individuals. This resulted in 2,460 households with mean size 1.22 and median size 1. These households were randomly divided into ten subgroups, each containing 246 households. The households were therefore nested within subgroups, and the mean and median subgroup sizes were 300 and 301 individuals, respectively.

To assess the power of each LMM method, we considered 8 settings with a range of contributions to the trait variance from the background polygenic effect, the household effect, and the subgroup effect. For each setting, we simulated 1000 phenotypes, each with one main causal SNP to be tested for association. The trait value for individual i from household j in subgroup k was generated by the model

$$Y_{ijk} = g_{is'}\beta_{s'} + \sum_{s \in \mathcal{S}_c} g_{is}\beta_s + h_j + v_k + \epsilon_i, \tag{6.16}$$

where $g_{is'}$ is the genotype value for individual i at the causal SNP, $\beta_{s'}$ is the effect size of the causal SNP, $\mathcal{S}_c$ is a set of 100 SNPs contributing to the background polygenic effect with respective effect sizes β_s, $h_j \sim \mathrm{N}(0, \sigma_H^2)$ is the effect for household j, $v_k \sim \mathrm{N}(0, \sigma_V^2)$ is the effect for subgroup k, and $\epsilon_i \sim \mathrm{N}(0, 1)$ is a unique environmental effect for individual i. The effect sizes for the 100 SNPs in the set $\mathcal{S}_c$ were found using the formula

$$\beta_s = \left[\frac{\sigma_A^2/100}{2\hat{p}_s(1-\hat{p}_s)} \right]^{1/2}, \tag{6.17}$$

so they each contributed equally to the entire background polygenic variance given by σ_A^2. The effect size for the causal SNP was chosen to explain 1% of the total trait variance, and it was found from the formula

$$\beta_{s'} = \left[\frac{h_{s'}^2(\sigma_A^2 + \sigma_H^2 + \sigma_V^2 + \sigma_\epsilon^2)}{(1-h_{s'}^2)2\hat{p}_{s'}(1-\hat{p}_{s'})} \right]^{1/2}, \tag{6.18}$$

with $h_{s'}^2 = 0.01$. Each of the 1,000 causal SNP-phenotype pairs from each setting were tested for association with each of the LMM methods. MMAAPS again used the top PC from PC-AiR to adjust for ancestry in the mean model, but it now estimated the trait covariance structure as in Equation 6.13, with $R = 2$. The PC-Relate kinship matrix was used to account for correlation due to relatedness, and

two additional variance components were included to account for correlation due to shared household and shared subgroup. EMMAX and GEMMA do not allow for the inclusion of additional shared environmental random effects and relied on only a relationship matrix to account for all sample structure.

The power for each method to detect the causal SNP in each setting is presented in Table 6.4. In all 8 settings, EMMAX and GEMMA provided roughly equal power, and MMAAPS performed as well as or better than both. When there was no additional environmental correlation, all three methods provided nearly the same performance, as expected. Consider the settings with only one additional shared environmental effect. The household cluster sizes were small, mostly consisting of one individual, and when σ^2_H was small relative to σ^2_A, all three methods still attained similar power. However, as σ^2_H became larger relative to σ^2_A, MMAAPS had a modest increase in power over the other methods. On the other hand, the subgroup cluster sizes were larger than the households, and there was a more drastic change in the power across methods due to a shared subgroup effect. MMAAPS had much higher power than the other methods, even when σ^2_V was small relative to σ^2_A, and the increase in power grew as the relative size of σ^2_V grew. In all settings where both shared environmental effects were non-zero, MMAAPS had much higher power than EMMAX and GEMMA. As may be expected, the increase in power for MMAAPS was highest when σ^2_V was the dominating variance component and was lowest when σ^2_A was the dominating variance component. These simulations demonstrate that two main factors seem to contribute to how much power is gained by modeling the shared environmental variance components. The gain in power increases as (1) the proportion of trait variance accounted for by shared environmental effects increases, and (2) the cluster sizes of individuals that share the environmental effects grow larger.

Test statistics at the null SNPs were examined for the setting with $\sigma^2_A = 0.2$, $\sigma^2_H = 0.2$, and $\sigma^2_V = 0.4$ to assess the control of type I error rate for each of these methods in the presence of additional environmental covariates. This setting was

Table 6.4: Power of LMM Methods for a Trait with Additional Environmental Covariances

σ^2_A	σ^2_H	σ^2_V	MMAAPS	EMMAX	GEMMA
0.2	0.0	0.0	0.460	0.456	0.458
0.2	0.4	0.0	0.510	0.492	0.495
0.4	0.2	0.0	0.435	0.439	0.441
0.2	0.0	0.4	0.729	0.465	0.467
0.4	0.0	0.2	0.601	0.463	0.466
0.2	0.4	0.2	0.591	0.449	0.452
0.2	0.2	0.4	0.739	0.504	0.508
0.4	0.2	0.2	0.558	0.466	0.471

σ^2_A is the background polygenic variance
σ^2_H is the household effect variance
σ^2_V is the subgroup effect variance
$h^2_{s'} = 0.0075$ for the causal SNP and $\sigma^2_\epsilon = 1$ for all scenarios.

chosen because it had the greatest amount of shared environmental structure and showed a very large increase in power with MMAAPS. Perhaps surprisingly, all three methods were well calibrated, as can be seen from the QQ-plot in Figure 6.5 for one of the simulated phenotypes. The mean genomic control inflation factors across all of the 1,000 phenotype replicates for MMAAPS, EMMAX, and GEMMA were 0.997, 0.997, and 0.998 respectively. It may have been expected that EMMAX and GEMMA would provide inflated test statistics due to the unaccounted environmental covariance, similar to the inflation of test statistics from linear regression in the presence of unaccounted for genetic relatedness. To further investigate this, consider the variance component estimates from these methods. In the chosen setting, the true proportions of variance accounted for by genetic, household, and subgroup effects were 12%, 11%, and 22% respectively. MMAAPS provided quite accurate estimates of these quantities, on average assigning 12.3% (s.e. = 5.4%) to genetic effects, 10.2% (s.e. = 3.7%) to household effects, and 21.4% (s.e. = 7.7%) to subgroup effects. In comparison, EMMAX provided highly inflated estimates of the genetic variance, assigning 37.1%

(s.e. = 6.8%) on average. One possible explanation for the properly calibrated test statistics may be that the GRM used by EMMAX is able to account for a lot of the shared environmental variance in the estimation procedure because it is dense (i.e. not sparse). While these results indicate that unaccounted for environmental covariances may not affect the performance of these LMM methods at null SNPs, we also demonstrated that accurately modeling the phenotype covariance can improve efficiency and drastically improve power at causal SNPs.

6.3.3 WHI SHARe Hispanic Cohort

A GWAS was performed for log white blood cell count in the Hispanic cohort of the Women's Health Initiative SNP Health Association Research (WHI-SHARe) study using MMAAPS, EMMAX, and GEMMA. EMMAX and GEMMA were run with the default settings and filters, using the Balding-Nichols and standardized genetic relationship matrices respectively. Analysis with MMAAPS required principal components and an ancestry-adjusted relationship matrix. Ancestry representative principal components were found with PC-AiR as described in Section 4.3.4, and an ancestry-adjusted relationship matrix was constructed from PC-Relate kinship coefficient estimates found using all 656,852 SNPs and the first 8 PC-AiR PCs, which all appear to reflect population structure (Figure 4.15).

The genomic control inflation factors and p-values for the genome-wide significant hits from each of the analyses are presented in Table 6.5. The Manhattan plot from MMAAPS in Figure 6.6 shows a strong hit on chromosome 1. This hit is near the Duffy gene, which has previously been shown through admixture mapping of African Americans to be associated with white blood cell count [45]. It can be seen from the p-values of the top hits, as well as the QQ-plot in Figure 6.7, that MMAAPS provides the strongest evidence of significance, followed by GEMMA, and lastly EMMAX. GEMMA and EMMAX do not show inflation in the QQ-plot and obtain genomic control inflation factors very close to 1. MMAAPS has a slightly higher genomic

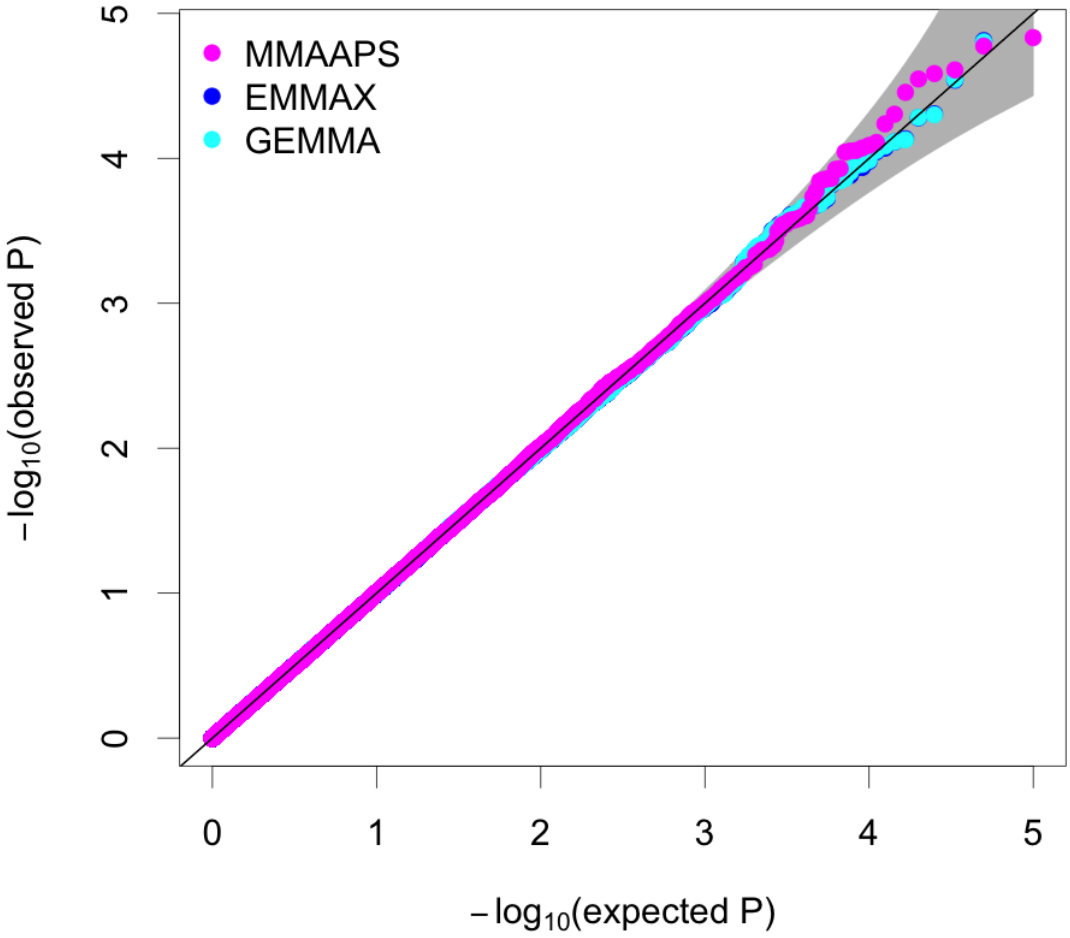


Figure 6.5: QQ-plot for Null SNPs in the Presence of Shared Environmental Effects
QQ-plot showing the distribution of $-\log_{10}(p)$-values at null SNPs from MMAAPS, EMMAX, and GEMMA for one of the phenotype replicates in the setting with $\sigma^2_A = 0.2$, $\sigma^2_H = 0.2$, and $\sigma^2_V = 0.4$. The gray shaded region shows the 95% confidence interval under the null hypothesis of no association.

inflation value of 1.012 and shows what could be some early inflation in the QQ-plot. The only SNPs approaching genome-wide significance were on chromosome 1, so we examined all autosomal SNPs not on chromosome 1 to investigate if MMAAPS was providing inflated test statistics at null SNPs or had slightly more power to detect small causal effects. The genomic control inflation factors excluding chromosome 1 SNPs are 1.005, 0.993, and 0.994 for MMAAPS, EMMAX, and GEMMA, respectively. A QQ-plot excluding chromosome 1 SNPs is presented in Figure 6.8, and MMAAPS no longer appears inflated. These results indicate that the higher genomic inflation for MMAAPS was likely due to increased power to detect a large number of associated SNPs on chromosome 1.

Table 6.5: Results for log White Blood Cell Count GWAS

	MMAAPS	EMMAX	GEMMA
λ_{GC}	1.012	0.997	0.999
SNP ID	p-value		
rs11265198	$1.59\text{x}10^{-12}$	$2.22\text{x}10^{-10}$	$7.36\text{x}10^{-11}$
rs2808666	$1.07\text{x}10^{-10}$	$2.30\text{x}10^{-9}$	$1.22\text{x}10^{-9}$
rs7534472	$2.81\text{x}10^{-10}$	$1.24\text{x}10^{-8}$	$6.01\text{x}10^{-9}$
rs857682	$8.47\text{x}10^{-10}$	$1.99\text{x}10^{-8}$	$1.23\text{x}10^{-8}$
rs856065	$1.44\text{x}10^{-9}$	$5.07\text{x}10^{-8}$	$2.95\text{x}10^{-8}$
rs6656586	$2.40\text{x}10^{-8}$	$3.36\text{x}10^{-7}$	$2.34\text{x}10^{-7}$

Each of the 6 genome-wide significant hits are in the same region of chromosome 1. MMAAPS provides the strongest evidence of significance, indicating an increase in power due to better modeling of ancestry and relatedness in the sample.

The proportion of phenotypic variance due to polygenic effects, $h^2_A = \sigma^2_A/(\sigma^2_A+\sigma^2_\epsilon)$, was also estimated from the variance component estimates given by these methods. As expected, the estimate of h^2_A from MMAAPS (0.351) was smaller than that from EMMAX (0.406), because MMAAPS adjusted for the ancestry effect in the mean model. GEMMA re-estimates the variance components at every SNP, so we do not

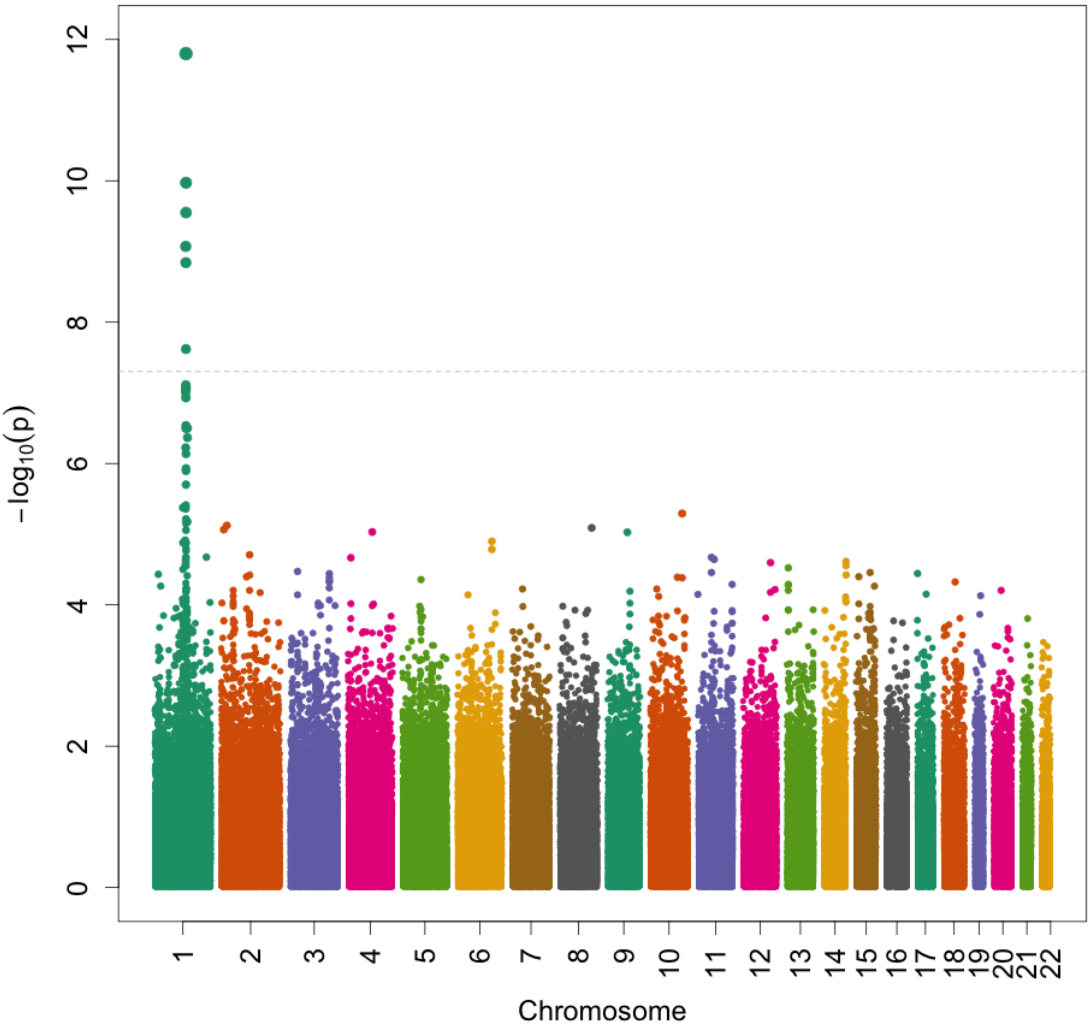


Figure 6.6: Manhattan Plot for log White Blood Cell Count
Manhattan plot of $-\log_{10}(p)$-values at all SNPs from MMAAPS for log white blood cell count in the Hispanic cohort of the WHI-SHARe study. The only SNPs approaching genome-wide significance are in a region of chromosome 1.

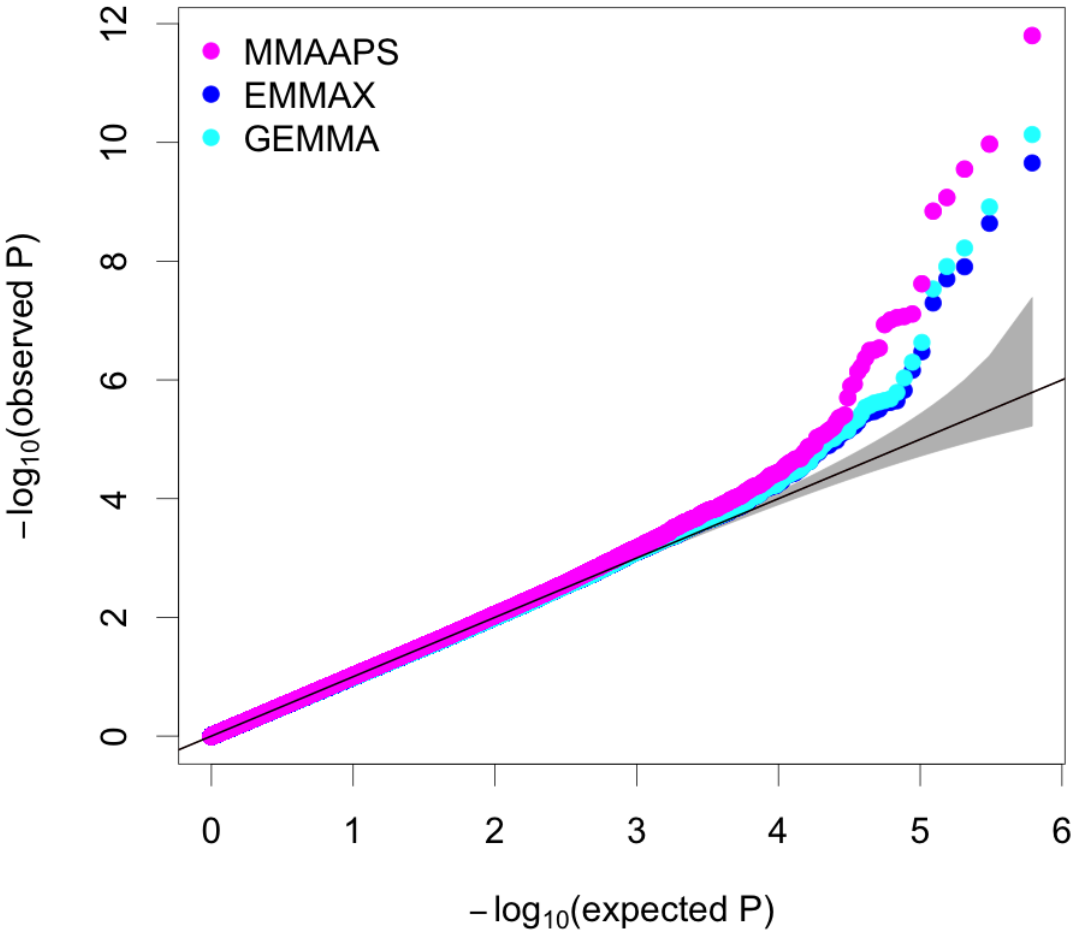


Figure 6.7: QQ-plot for log White Blood Cell Count
QQ-plot showing the distribution of $-\log_{10}(p)$-values at all SNPs from MMAAPS, EMMAX, and GEMMA for log white blood cell count in the Hispanic cohort of the WHI-SHARe study. The gray shaded region shows the 95% confidence interval under the null hypothesis of no association. MMAAPS shows the highest significance at associated SNPs.

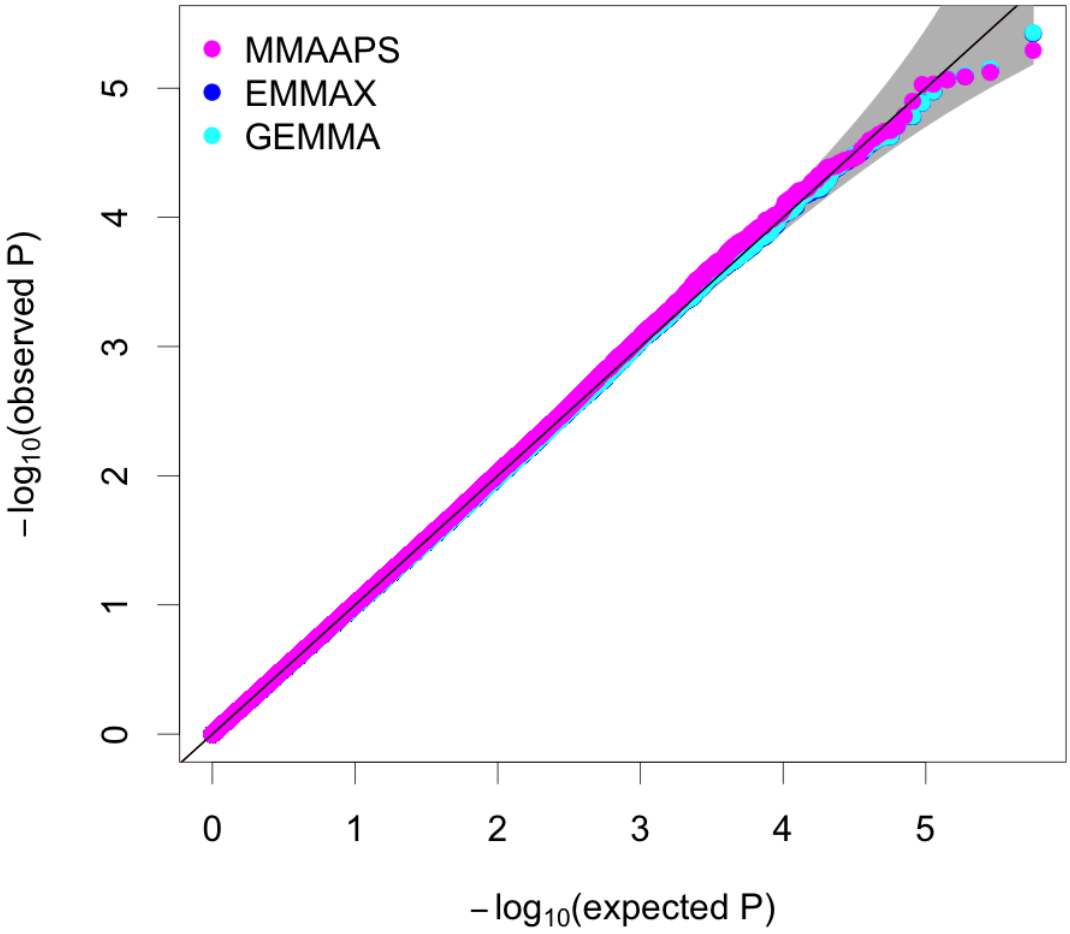


Figure 6.8: QQ-plot for log White Blood Cell Count Excluding Chromosome 1

QQ-plot showing the distribution of $-\log_{10}(p)$-values at SNPs not on chromosome 1 from MMAAPS, EMMAX, and GEMMA for log white blood cell count in the Hispanic cohort of the WHI-SHARe study. The gray shaded region shows the 95% confidence interval under the null hypothesis of no association.

report an estimate here. The quantity h^2_A is also referred to as the narrow-sense heritability of the trait in a homogeneous population, but this parameter is difficult to interpret in the presence of population structure. The estimate from MMAAPS represents the proportion of trait variability accounted for by genetic effects shared recently IBD, and the estimate from EMMAX represents the proportion accounted for by genetic effects due to both recent sharing IBD as well as distant sharing through ancestral similarity.

It has been suggested that including the SNP being tested, or other SNPs in LD with that SNP, in the relationship matrix results in proximal contamination and loss of power due to double counting data [27, 28, 68]. To investigate the effects of using different sets of SNPs to construct the relationship matrix for the log white blood cell count GWAS, we re-ran MMAAPS (1) using a kinship matrix constructed from the 87,180 LD pruned and MAF filtered SNPs used to find the PCs with PC-AiR, and (2) using a leave-one-chromosome-out approach, where all SNPs not on the same chromosome as the SNP being tested for association were used to construct the kinship matrix. A comparison of the results using the three different kinship matrices is presented in Table 6.6. The leave-one-chromosome-out approach gives the most significant p-values, followed by using a pruned set of SNPs, and lastly using all SNPs. These results are concordant with the idea of proximal contamination since the leave-one-chromosome-out approach includes no SNPs in the kinship matrix that are near the SNP being tested, the pruned SNPs approach may include few SNPs near the one being tested, and the all SNPs approach includes everything near the SNP being tested. However, the methods that used a subset of SNPs to construct the kinship matrix also gave inflated λ_{GC}, even when excluding chromosome 1, the only with a significant hit. It is unclear whether the inflated value is representative of insufficient sample structure correction or due to polygenic effects captured on the remaining chromosomes.

Table 6.6: Comparison of MMAAPS Results with Different Kinship Matrices

	All SNPs	Pruned SNPs	LOCO
λ_{GC} (All Chr)	1.012	1.032	1.044
λ_{GC} (No Chr 1)	1.005	1.026	1.036
SNP ID	p-value		
rs11265198	$1.59\text{x}10^{-12}$	$4.96\text{x}10^{-13}$	$1.12\text{x}10^{-13}$
rs2808666	$1.07\text{x}10^{-10}$	$3.22\text{x}10^{-11}$	$1.28\text{x}10^{-11}$
rs7534472	$2.81\text{x}10^{-10}$	$5.67\text{x}10^{-11}$	$3.73\text{x}10^{-11}$
rs857682	$8.47\text{x}10^{-10}$	$2.41\text{x}10^{-10}$	$1.29\text{x}10^{-10}$
rs856065	$1.44\text{x}10^{-9}$	$5.65\text{x}10^{-10}$	$2.66\text{x}10^{-10}$
rs6656586	$2.40\text{x}10^{-8}$	$8.52\text{x}10^{-9}$	$3.52\text{x}10^{-9}$

All SNPs uses a kinship matrix constructed from all 656,852 SNPs. Pruned SNPs uses a kinship matrix constructed from 87,180 LD pruned and MAF filtered SNPs. LOCO (leave-one-chromosome-out) uses a different kinship matrix when testing each chromosome; the kinship matrix is constructed from all SNPs not on the same chromosome as the SNP being tested.

6.4 Discussion

Linear mixed models (LMMs) have become an attractive and widely used approach for genetic association testing of quantitative traits on a genome-wide scale. They have been shown to be effective at controlling genomic inflation in samples with relatedness and subtle population structure [42, 63], and they have been shown to increase statistical power to detect true associations when implemented carefully [68]. Genetic studies involving ancestrally diverse populations from around the globe are becoming more common as interest grows in replicating associations across populations as well as identifying novel variants that are population specific and underly differences in phenotypes. Allele frequency differences between highly divergent populations can be quite large, and in populations with continental ancestry admixture, such as African Americans and Hispanic Americans, the pattern of this allele frequency differentiation varies greatly across the genome. Many LMM methods have been proposed for

genome-wide association studies (GWAS) [21,22,27,28,53,68,69,72], but their performance in highly admixed populations has not been explored thoroughly and should be of concern, as it has been shown that LMMs may not perform well at markers that are unusually differentiated between populations [42].

In a simulation study of an admixed population generated from two divergent ancestral populations, we demonstrated that existing implementations of LMMs for GWAS that only use a genome-wide genetic relationship matrix to account for both population structure and relatedness provide systematically biased test statistics in admixed populations. These methods provide inflated test statistics at SNPs with large allele frequency differences between ancestral populations, and test statistics that are deflated at SNPs with small allele frequency differences. This result is contrary to what is suggested in the literature [68], where it is stated that "markers with large allele frequency differences between populations receive a larger correction." Notably, the overcorrection for lowly differentiated SNPs and the undercorrection for highly differentiated SNPs balances out across the genome, leading to a genomic inflation factor near 1 when considering all SNPs, likely contributing to why this phenomenon has not been identified previously. This bias should be of concern when analyzing admixed populations, as we demonstrated through analysis of HapMap populations that over 30% of SNPs may be highly or moderately differentiated across continental populations.

To address this issue, we developed MMAAPS for linear mixed model association testing in admixed populations. MMAAPS partitions the genetic covariance structure among sample individuals into two pieces; ancestry representative principal components are included as fixed effects in the mean model, and ancestry-adjusted estimates of kinship coefficients are used to model relatedness as a random effect. In the same simulation study, we demonstrated that the testing procedure implemented in MMAAPS provides well-calibrated test statistics at all SNPs in admixed populations, regardless of the level of differentiation of allele frequencies between ancestral

populations. As a result, MMAAPS provides a better true positive to false positive ratio at highly differentiated SNPs when existing LMM methods identify many false positives due to mis-calibration. Additionally, MMAAPS does not suffer a loss in statistical power to detect true SNP-phenotype associations, and it may even provide a modest increase in power genome-wide. Some further evidence supporting this result was provided from analysis of log white blood cell count in the Hispanic cohort of the WHI SHARe study, where MMAAPS provided more significant p-values than EMMAX or GEMMA at SNPs near a previously identified hit [45].

We also explored the effects of using different sets of SNPs to construct the kinship matrix that MMAAPS utilizes to account for background polygenic effects due to relatedness. Many LMM methods for GWAS use a genetic relationship matrix constructed from all SNPs, but it has been reported in the literature that including the SNP that is being tested in the matrix leads to proximal contamination and loss of power due to double counting data [27, 28, 68]. We found that the significance of top hits in the WHI SHARe data was increased when using a leave-one-chromosome-out procedure or using an LD pruned set of SNPs to construct the kinship matrix. However, we also found that the genomic control inflation factor was larger with these analyses, even when excluding regions with significant hits. It is unclear whether the apparent inflation is due to insufficient sample structure correction, or if it is due to improved power to detect polygenic effects of small magnitude across the genome, as suggested by Yang et al. [67, 68].

Finally, we conducted simulations to investigate the consequences of modeling, or not modeling, additional random effects such as those due to shared environmental variances. Existing LMM software for GWAS does not allow for additional random effects beyond the polygenic effect to be included in the model, but MMAAPS has no restriction on the number of random effects that can be included. Perhaps surprisingly, we found that not modeling additional environmental correlation structure with methods such as EMMAX and GEMMA, even when these effects are strong,

did not lead to inflated test statistics at null SNPs. One possible explanation for this is that the genetic relationship matrix is fairly unstructured, having non-zero elements everywhere, so it is able to soak up a lot of the extra variability due to shared environmental effects, even though they do not follow the same covariance structure. This conjecture is something that needs to be explored in more detail. Regardless of the explanation, we also showed that modeling shared environmental random effects accurately with MMAAPS can lead to much greater power to detect true causal SNPs. A likely explanation is improved efficiency of the test from using the correct covariance structure. MMAAPS achieves the greatest power gains over existing methods when the proportion of variance that is due to shared environment is large or when the cluster sizes of individuals that share a common environmental effect are large.

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, we have discussed and explored in depth the properties and performance of current statistical methodology for inferring, estimating, and accounting for population structure and relatedness in genetic analyses. In the process, we have identified and demonstrated deficiencies with existing approaches, and we have developed new and improved methodology to overcome these limitations. The focus has been primarily on populations with ancestry admixture from multiple previously isolated populations. We have developed methodology that can provide accurate population structure inference in the presence of unreported cryptic relatedness, accurate relatedness estimation in the presence of population structure that is unspecified a priori, and improved association mapping in samples with ancestry admixture and partially or completely unknown genealogy. In this chapter, we briefly summarize the statistical methodology that we have developed and discuss possible extensions of these methods for future research expanding on what we have done.

In Chapter 3 we demonstrated that existing methodology for ancestry inference can have substantial bias in samples with family relatedness, either known or cryptic. Attempts to correct for this problem have been made, but have not been satisfactory up to this point. For example, simply choosing one individual from a set of relatives to be included in an ancestry analysis and throwing away the rest of the data results in a loss of efficiency due to a decreased sample size; the FamPCA [73] method requires prior knowledge of pedigree structure; and even model based methods that use reference population samples [3,43,54] provide biased proportional ancestry estimates in related samples unless an individual ancestry analysis is conducted separately for each sample individual, one at a time. To address the problem of population structure

inference in samples with pedigree relationships, we developed the PC-AiR method for principal components analysis in related samples. PC-AiR provides accurate population structure inference that is robust to both known and cryptic relatedness in the sample. PC-AiR is able to provide reliable population structure inference in the presence of complex relatedness without requiring external reference population samples by utilizing a fast and sophisticated algorithm to partition the entire sample into an ancestry representative mutually unrelated subset and a related subset. The only input required for the algorithm are pairwise measures based on the genome-wide SNP data from the sample individuals. While PC-AiR does not require reference population samples, its implementation is flexible enough to allow for them if desired.

One possible extension to the PC-AiR method is to allow for the use of sequence data. Sequence data contains many rare variants that are often excluded from inference on population stratification using standard minor allele frequency (MAF) thresholds. The current implementation of PC-AiR uses a genetic relationship matrix for the unrelated set constructed as the mean of ratios across all SNPs. This means that each term in the estimator involves division by $p_s(1 - p_s)$, resulting in unstable estimates when rare variants are included. To extend PC-AiR for ancestry inference in sequencing studies, a genetic relationship matrix constructed as the ratio of means of the numerator and denominator terms, similar to the PC-Relate kinship estimator, could be used (see Section 4.2.3). This estimator avoids giving too much influence to SNPs with very small MAF and remains stable even with rare variants. It has been suggested that population structure for rare variants could be substantially different from that for common variants [35]. Ancestry inference with a combination of common and rare variants, either jointly or separately, could provide interesting insight into different structure within and between populations.

Extending PC-AiR to improve the projection of principal component values for individuals in the related subset is currently of interest. It has been shown before that prediction of principal components values using SNP loadings (i.e. weights) in

high dimensional settings can lead to shrinkage towards zero [25]. In our experience, we have only witnessed this bias in very small samples, but it remains a valid concern in all situations. A recently published manuscript introduced the LASER method for ancestry estimation with sequence data [60]. This method uses a reference panel to perform PCA and then uses Procrustes analysis [48, 59] to accurately map sample individuals, one at a time, to the reference PCA space. A reasonable integration of these two methods would be to identify an ancestry representative mutually unrelated subset of individuals using the PC-AiR algorithm, run PCA directly on these individuals, and then predict principal components values for the excluded relatives in this PCA space using the technique implemented by LASER.

Chapter 4 thoroughly explored the properties of many estimators for measures of genetic relatedness and demonstrated that the simplifying assumptions made by existing methods can lead to poor performance in populations with ancestry admixture. Estimators that assume population homogeneity are systematically inflated for pairs of individuals with similar ancestry, and the KING-robust [32] kinship coefficient estimator can have substantial positive or negative bias in the presence of ancestry admixture and inbreeding. There are existing methods, such as REAP [56] and RelateAdmix [37], that can provide accurate relatedness estimates for admixed populations when the population structure can be accurately estimated using reference population samples. However, these methods will be biased if the assumptions regarding the underlying ancestry of the sample are incorrect, and they are not able to account for more subtle sub-continental population structure. Furthermore, analysis with RelateAdmix for large samples of many thousand individuals is not computationally feasible.

We developed the PC-Relate method for accurate relatedness estimation in admixed populations with unspecified structure. Rather than relying on reference panels for estimates of sub-population specific allele frequencies and individual admixture proportions as REAP does, PC-Relate implicitly infers individual-specific allele fre-

quencies by considering the expectation of the observed sample genotype values conditional on a set of ancestry representative principal components. PC-Relate provides accurate pairwise kinship coefficient and IBD sharing probability estimates, even in the presence of complex and unknown population structure, including ancestry admixture. PC-Relate also does not make an assumption of HWE, and it can also be used in inbred populations for accurate kinship and inbreeding coefficient estimation.

A possible extension of PC-Relate, as with PC-AiR, is to relatedness inference with rare and common variants from sequence data. The PC-Relate estimators are already implemented as ratio of means estimators, so they should be stable with sequence data and rare variants. Minimal effort should be required for the PC-Relate method to be adapted for the analysis of sequence data, but its performance for rare variants has not yet been thoroughly evaluated. As sequence-based association studies become more common, PC-Relate may prove to be a very useful tool for relatedness estimation.

In Chapter 6 we closely examined the performance of existing linear mixed model (LMM) methods for genetic association testing. All of the existing approaches fit the same basic model, where both population structure and relatedness are accounted for as random effects by modeling the phenotype covariance structure with a genetic relationship matrix (GRM), and the contribution of polygenic variance is directly estimated from the observed data. The differences between existing methods are largely based on implementation of various algorithms optimized for computational speed and allowing for genome-wide analysis to be performed efficiently, as well as different strategies for selecting which SNPs are used to construct the empirical GRM to account for sample structure. While these methods perform well for samples with relatives and subtle population structure, we demonstrated that the association test statistics are not well-calibrated in the presence of continental ancestry admixture, resulting in inflated type I error rates at SNPs with large allele frequency differences between the underlying ancestral populations.

To address the issue of association testing in related admixed samples with ancestry derived from highly divergent populations, we developed the MMAAPS method for linear mixed model analysis in admixed populations. Rather than treating all of the sample structure as one random effect, MMAAPS partitions the sample structure into two components. The population stratification in the sample is adjusted for using ancestry representative principal components, such as those from PC-AiR, in the mean model, and the relatedness in the sample is accounted for by fitting an ancestry adjusted kinship matrix, generated from PC-Relate kinship estimates, as a random effect. MMAAPS provides well-calibrated test statistics at all SNPs, even in continentally admixed populations, regardless of the ancestral population allele frequency differentiation. Further, MMAAPS does not pay a penalty in power due to its improved control of type I error rates, and it provides a better true positive to false positive rate than existing LMM methods at highly differentiated SNPs.

Another advantage of MMAAPS is that it does not put a limitation on the number of random effects covariance structures that can be included in the model. This allows for the modeling of shared environmental effects, which we demonstrated can greatly increase power to detect associated SNPs in certain scenarios. Beyond accounting for shared environment, the flexibility of this model also opens up opportunities for future extensions to MMAAPS.

One possible extension for MMAAPS is accounting for X-chromosome structure. Since males only have one copy of the X-chromosome, the correlation structure for male-male or female-male pairs does not follow the same pattern as on the autosomes. As a consequence, polygenic effects on the X-chromosome can not be modeled using the same kinship matrix as for the autosomes, and the effects of X-chromosome SNPs are typically ignored in analyses. MMAAPS could allow for the inclusion of two kinship matrices, one representing autosomal structure, and one representing X-chromosome structure, in order to account for polygenic effects genome-wide rather than only autosome-wide. This may provide better control of sample structure when

testing autosomal SNPs, and it may allow for well-calibrated genetic association tests at X-chromosome SNPs.

A second possible extension of MMAAPS is adjustment for and testing of genetic dominance effects. There has been limited consideration of dominance effects in complex trait mapping, but increasing sample sizes of genetic studies may provide more power to attempt to detect them. At any given SNP, a joint test of both the additive and dominance genotype codings can be used to provide a more general test that does not assume additive genotype effects. Because they are orthogonal, the tests of the additive and dominance genotype codings could be run separately for computational efficiency, and test statistics could easily be combined. When fitting a model that tests for dominance genotype effects, it would be appropriate to also account for background polygenic dominance effects. Fortunately, the machinery to do this is already in place. The appropriate covariance structures for non-additive genetic effects, including dominance effects, have been specified previously [2, 8]. An estimated dominance genetic relationship matrix could be constructed with the values of $\widehat{\delta}^{PC}$ from PC-Relate on the off diagonals and $(1 - \widehat{f}^{PC})$ on the diagonals. By fitting a model that includes both a kinship matrix that measures additive genetic covariance and a matrix that measures dominance genetic covariance, estimates of both the additive and dominance variance components could be obtained, and both the narrow and broad sense heritability of a trait could be estimated.

A final extension to the MMAAPS method is to admixture mapping [39,49,51]. In admixture mapping, the number of alleles that each individual has from a particular population are tested for association rather than the number of alleles of a certain base pair type. Admixture mapping requires local ancestry estimation, and software is available for this [34, 47]. Admixture mapping can be extended to any number of underlying populations, as counts of alleles from all but one reference population can be tested jointly for association. The extension to MMAAPS is therefore relatively simple, but the performance of this method has not yet been evaluated. Whether or

not global ancestry should be adjusted for in the model, and what the appropriate way to make this adjustment is should be explored in detail.

Appendix A

MATHEMATICAL DERIVATIONS

A.1 Expectations of Individual-specific Allele Frequencies

Recall that g_{is} is the number of copies of the reference allele that individual i has at SNP s, and thus g_{is} can take the values 0, 1, or 2. Also recall that the definition of the individual-specific allele frequency, μ_{is}, is one half of the expectation of g_{is}, conditional on individual i's ancestry, $\mathbf{a}_i$, and the vector of subpopulation-specific allele frequencies, $\mathbf{p}_s$, at SNP s:

$$\mu_{is} \equiv \frac{1}{2}\mathbb{E}[g_{is}|\mathbf{a}_i, \mathbf{p}_s] = \mathbf{a}_i^T \mathbf{p}_s = \sum_{k=1}^{K} a_i^k p_s^k. \tag{A.1}$$

In Thornton et al. (2012) [56], both $\mathbf{a}_i$ and $\mathbf{p}_s$ are treated as fixed quantities. Here, we similarly treat the ancestry vectors as fixed, and we implicitly condition on $\mathbf{a}_i$ and $\mathbf{a}_j$ throughout what follows, but we allow $\mathbf{p}_s$ to be a random vector with the properties $\mathbb{E}[\mathbf{p}_s] = p_s\mathbf{1}$ and $\text{Cov}[\mathbf{p}_s] = p_s(1-p_s)\mathbf{\Sigma}_K$ for every $s \in \mathcal{S}$. Under these weak genetic

modeling assumptions, we calculate the following expectations:

$$\mathbb{E}[\mu_{is}] = \mathbb{E}[\mathbf{a}_i^T \mathbf{p}_s] = \mathbf{a}_i^T \mathbb{E}[\mathbf{p}_s] = (\mathbf{a}_i^T \mathbf{1}) p_s = p_s, \tag{A.2}$$

$$\begin{aligned}
\mathbb{E}[\mu_{is}\mu_{js}] &= \mathbb{E}\left[\sum_{k=1}^{K} a_i^k p_s^k \sum_{k'=1}^{K} a_j^{k'} p_s^{k'}\right] \\
&= \sum_{k=1}^{K}\sum_{k'=1}^{K} a_i^k a_j^{k'} \mathbb{E}[p_s^k p_s^{k'}] \\
&= \sum_{k=1}^{K}\sum_{k'=1}^{K} \left[a_i^k a_j^{k'} (p_s)^2 + a_i^k a_j^{k'} \left(\mathbb{E}[p_s^k p_s^{k'}] - (p_s)^2\right)\right] \\
&= (p_s)^2 \sum_{k=1}^{K} a_i^k \sum_{k'=1}^{K} a_j^{k'} + \sum_{k=1}^{K}\sum_{k'=1}^{K} \left(a_i^k a_j^{k'} \mathrm{Cov}[p_s^k, p_s^{k'}]\right) \\
&= (p_s)^2 + p_s(1-p_s)\mathbf{a}_i^T \mathbf{\Sigma}_K \mathbf{a}_j,
\end{aligned} \tag{A.3}$$

$$\text{and} \qquad \mathbb{E}[\mu_{is}(1-\mu_{is})] = p_s(1-p_s)[1 - \mathbf{a}_i^T \mathbf{\Sigma}_K \mathbf{a}_i]. \tag{A.4}$$

These expectations are used to derive the limiting behavior of the relatedness estimators in what follows.

A.2 Properties of Kinship Coefficient Estimators

In this section we examine the properties of three kinship coefficient estimators: (1) the elements of the standard genetic relationship matrix (GRM) as calculated in GCTA [65], (2) the PC-Relate kinship coefficient estimator $\widehat{\phi}_{ij}^{PC}$, and (3) the KING-robust kinship coefficient estimator. We derive the limiting behavior of each estimator under general population structure settings and also consider special situations such as unrelated pairs of individuals, discrete population substructure, homogeneous populations, and outbred populations.

A.2.1 Genetic Relationship Matrix (GRM)

Recall the form of the elements of the standard empirical GRM from Equation 4.1:

$$\widehat{\psi}_{ij} = \frac{1}{|\mathcal{S}_{ij}|} \sum_{s \in \mathcal{S}_{ij}} \frac{(g_{is} - 2\hat{p}_s)(g_{js} - 2\hat{p}_s)}{2\hat{p}_s(1-\hat{p}_s)}. \tag{A.5}$$

In order to derive the limiting behavior of this estimator, we begin by making a few assumptions that we later relax. First, assume that the true values of the ancestral allele frequencies are known so that $\hat{p}_s = p_s$. Secondly, assume that the p_s for every $s \in \mathcal{S}$ are independent and identically distributed (i.i.d.) random variables from some unspecified distribution on $[0, 1]$. Under this assumption, the unconditional expectation of each term in the summation in Equation (A.5) is the same for every choice of $s \in \mathcal{S}$. Finally, assume that genotypes at different SNPs are independent and condition on the known values of p_s. From the law of large numbers, it can be shown that

$$\frac{1}{2}\widehat{\psi}_{ij} \to \mathbb{E}\left[\frac{(g_{is} - 2p_s)(g_{js} - 2p_s)}{4p_s(1-p_s)}\right] = \frac{\mathbb{E}[g_{is}g_{js}] - 4(p_s)^2}{4p_s(1-p_s)} \tag{A.6}$$

as $|\mathcal{S}_{ij}| \to \infty$. However, as previously mentioned, the assumptions stated above can be relaxed and the convergence in Equation (A.6) will still hold. The independence of SNPs is not necessary, and a sufficient condition is that the effective number of independent SNPs tends to ∞. Additionally, the true values of p_s need not be known, as the convergence result will hold as long as the estimators $\hat{p}_s$ are consistent for p_s. Lastly, in what follows we derive $\mathbb{E}[g_{is}g_{js}]$ and show that the limiting value of $\frac{1}{2}\widehat{\psi}_{ij}$ does not depend on p_s, allowing for this convergence to hold for any random p_s on $[0, 1]$, and an assumption of i.i.d. p_s is not necessary.

Define the random variable x_{is_r} to be the indicator that individual i's allele $r \in \{1, 2\}$ at SNP s is the reference allele; by definition $g_{is} = x_{is_1} + x_{is_2}$. To calculate $\mathbb{E}[g_{is}g_{js}]$, we consider the quantity $\mathbb{E}[x_{is_r}x_{js_{r'}}]$, i.e. the probability that alleles r and r' at SNP s are each the reference allele for individuals i and j, respectively. Define the following sets and quantities. The set $\mathcal{M}_{ij}$ is the set of shared most recent common

ancestors of individuals i and j, possibly including individuals i or j. For example, if j is a direct descendant of i, then $\mathcal{M}_{ij} = \{i\}$; if i and j are siblings, then $\mathcal{M}_{ij}$ is their two parents; if i and j are cousins, then $\mathcal{M}_{ij}$ is their two shared grandparents, etc. The set $\mathcal{M}_i$ is a set of individual i's ancestors that account for all sources of individual i's genetic material. This set could be both parents, all four grandparents, one parent and two grandparents, etc., as long as none of the individuals in $\mathcal{M}_i$ are descendants of each other. The quantity n_{im} gives the length of the path in the pedigree from individual i to m, including both of these individuals. For example, if i and m are the same individual, $n_{im} = 1$; if i is the child of m, $n_{im} = 2$, etc. Through a path counting argument tracing back alleles to the individuals from which they descended, we can see that:

$$\begin{aligned}
\mathbb{E}[x_{is_r}x_{js_{r'}}|\mathbf{p}_s] &= \sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}-1)}\left(\frac{1}{2}\right)^{(n_{jm}-1)}\left\{(1-f_m)\left(\frac{1}{2}\mu_{ms}+\frac{1}{2}\mu_{ms}^2\right)\right.\right.\\
&\qquad\left.\left.+f_m\left(\frac{1}{2}\mu_{ms}+\frac{1}{2}\mu_{ms}\right)\right\}\right]\\
&\quad+\sum_{\substack{m_i\in\mathcal{M}_i\\ m_i\neq m_j}}\sum_{m_j\in\mathcal{M}_j}\left[\left(\frac{1}{2}\right)^{(n_{im_i}-1)}\mu_{(m_i)s}\left(\frac{1}{2}\right)^{(n_{jm_j}-1)}\mu_{(m_j)s}\right]\\
&=\sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}-1)}\left(\frac{1}{2}\right)^{(n_{jm}-1)}\left\{\frac{1}{2}(1+f_m)\mu_{ms}(1-\mu_{ms})\right\}\right]\\
&\quad+\left(\sum_{m_i\in\mathcal{M}_i}\left[\left(\frac{1}{2}\right)^{(n_{im_i}-1)}\mu_{(m_i)s}\right]\right)\left(\sum_{m_j\in\mathcal{M}_j}\left[\left(\frac{1}{2}\right)^{(n_{jm_j}-1)}\mu_{(m_j)s}\right]\right)\\
&=\sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}+n_{jm}-1)}(1+f_m)\mu_{ms}(1-\mu_{ms})\right]+\mu_{is}\mu_{js}. \qquad \text{(A.7)}
\end{aligned}$$

In the last line above, we get that $\mu_{is} = \sum_{m_i\in\mathcal{M}_i}\left[\left(\frac{1}{2}\right)^{(n_{im_i}-1)}\mu_{(m_i)s}\right]$ because $\left(\frac{1}{2}\right)^{(n_{im_i}-1)}$ represents the proportion of genetic material that individual i inherited from ancestor m_i. Since $\mathbb{E}[g_{is}g_{js}|\mathbf{p}_s] = 4\mathbb{E}[x_{is_r}x_{js_{r'}}|\mathbf{p}_s]$, by taking the expectation of this quantity

over the distribution of $\mathbf{p}_s$, we obtain the unconditional expectation to be

$$\begin{aligned}
\mathbb{E}[g_{is}g_{js}] &= \mathbb{E}[\mathbb{E}[g_{is}g_{js}|\mathbf{p}_s]] \\
&= 4\sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}+n_{jm}-1)}(1+f_m)\mathbb{E}[\mu_{ms}(1-\mu_{ms})]\right] + 4\mathbb{E}[\mu_{is}\mu_{js}] \\
&= 4p_s(1-p_s)\sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}+n_{jm}-1)}(1+f_m)[1-\mathbf{a}_m^T\boldsymbol{\Sigma}_K\mathbf{a}_m]\right] \\
&\quad + 4\left[(p_s)^2 + p_s(1-p_s)\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_j\right]. \qquad \text{(A.8)}
\end{aligned}$$

From Equation (A.6), we can see that

$$\frac{1}{2}\widehat{\psi}_{ij} \to \sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}+n_{jm}-1)}(1+f_m)[1-\mathbf{a}_m^T\boldsymbol{\Sigma}_K\mathbf{a}_m]\right] + \mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_j. \qquad \text{(A.9)}$$

The kinship coefficient can also be written in terms of path counting as

$$\phi_{ij} = \sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}+n_{jm}-1)}(1+f_m)\right] = \sum_{m\in\mathcal{M}_{ij}}\phi_{ij|m}, \qquad \text{(A.10)}$$

where $\phi_{ij|m} \equiv \left(\frac{1}{2}\right)^{(n_{im}+n_{jm}-1)}(1+f_m)$ is defined to be the contribution to the kinship for individuals i and j through alleles shared IBD from common ancestor m. Therefore, we can write

$$\frac{1}{2}\widehat{\psi}_{ij} \to \phi_{ij}[1-b_{\psi_1}(i,j)] + b_{\psi_2}(i,j), \qquad \text{(A.11)}$$

where the two bias terms are given by the functions

$$b_{\psi_1}(i,j) \equiv \frac{\sum_{m\in\mathcal{M}_{ij}}\phi_{ij|m}\mathbf{a}_m^T\boldsymbol{\Sigma}_K\mathbf{a}_m}{\sum_{m\in\mathcal{M}_{ij}}\phi_{ij|m}} \qquad \text{(A.12)}$$

$$\text{and} \qquad b_{\psi_2}(i,j) \equiv \mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_j. \qquad \text{(A.13)}$$

The first bias term, $b_{\psi_1}(i,j)$, results from the incorrect scaling of genotype values in $\widehat{\psi}_{ij}$, and the second bias term, $b_{\psi_2}(i,j)$, results from the incorrect centering of the genotype values. Note that for an unrelated pair of individuals, $\phi_{ij} = 0$, $\mathcal{M}_{ij} = \{\}$,

and $\frac{1}{2}\widehat{\psi}_{ij} \rightarrow \mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_j$, resulting in inflated estimates for pairs of individuals with similar ancestry.

In some particular population structure settings, the limiting value of the GRM given in Equation (A.11) simplifies considerably. With discrete population substructure, for pairs of individuals from the same subpopulation, $\mathbf{a}_i = \mathbf{a}_j = \mathbf{a}_m \equiv \mathbf{a}_*$ for every $m \in \mathcal{M}_{ij}$, the bias terms become $b_{\psi_1}(i,j) = b_{\psi_2}(i,j) = \mathbf{a}_*^T \boldsymbol{\Sigma}_K \mathbf{a}_*$, and

$$\frac{1}{2}\widehat{\psi}_{ij} \rightarrow \phi_{ij}[1 - \mathbf{a}_*^T \boldsymbol{\Sigma}_K \mathbf{a}_*] + \mathbf{a}_*^T \boldsymbol{\Sigma}_K \mathbf{a}_*. \tag{A.14}$$

Even in the presence of discrete population substructure, the GRM will provide inflated estimates for all pairs of individuals with similar ancestry. In a homogeneous population, $K = 1$, and the random vector of subpopulation-specific allele frequencies, $\mathbf{p}_s$, is just a single scalar value, p_s. In this scenario, $\mathbf{p}_s = p_s$ is a degenerate random variable, so $\boldsymbol{\Sigma}_K = 0$, the bias terms become $b_{\psi_1}(i,j) = b_{\psi_2}(i,j) = 0$, and

$$\frac{1}{2}\widehat{\psi}_{ij} \rightarrow \phi_{ij}. \tag{A.15}$$

This demonstrates that the standard GRM provides consistent kinship coefficient estimates in homogeneous populations.

A.2.2 PC-Relate

Now recall the form of the PC-Relate kinship coefficient estimator from Equation 4.5:

$$\widehat{\phi}_{ij}^{PC} = \frac{\sum_{s \in \mathcal{S}_{ij}} (g_{is} - 2\widehat{\mu}_{is})(g_{js} - 2\widehat{\mu}_{js})}{4 \sum_{s \in \mathcal{S}_{ij}} \sqrt{\widehat{\mu}_{is}(1 - \widehat{\mu}_{is})\widehat{\mu}_{js}(1 - \widehat{\mu}_{js})}}. \tag{A.16}$$

In order to show the limiting behavior of this estimator, we make the same set of assumptions as presented for $\widehat{\psi}_{ij}$, but we now assume that the true individual-specific allele frequencies are known so that $\widehat{\mu}_{is} = \mu_{is}$ and $\widehat{\mu}_{js} = \mu_{js}$. With these assumptions, by the law of large numbers,

$$\widehat{\phi}_{ij}^{PC} \rightarrow \frac{\mathbb{E}[g_{is}g_{js}] - 2\mathbb{E}[\mu_{is}g_{js}] - 2\mathbb{E}[\mu_{js}g_{is}] + 4\mathbb{E}[\mu_{is}\mu_{js}]}{4\mathbb{E}\left[\sqrt{\mu_{is}(1 - \mu_{is})\mu_{js}(1 - \mu_{js})}\right]} \tag{A.17}$$

as $|\mathcal{S}_{ij}| \to \infty$. Similar to our previous limiting result, each of the assumptions can be relaxed, and the convergence will still hold as long as the estimators $\widehat{\mu}_{is}$ and $\widehat{\mu}_{js}$ are consistent for their respective individual-specific allele frequencies. We previously calculated $\mathbb{E}[g_{is}g_{js}]$ in Equation (A.8) and $\mathbb{E}[\mu_{is}\mu_{js}]$ in Equation (A.3), and because μ_{js} is a fixed quantity conditional on $\mathbf{p}_s$, it can easily be seen that $\mathbb{E}[g_{is}\mu_{js}] = \mathbb{E}[\mathbb{E}[g_{is}\mu_{js}|\mathbf{p}_s]] = \mathbb{E}[\mu_{js}\mathbb{E}[g_{is}|\mathbf{p}_s]] = 2\mathbb{E}[\mu_{is}\mu_{js}]$. The expectation in the denominator of Equation (A.17) is not straightforward to calculate, but we can define it to be

$$\mathbb{E}\left[\sqrt{\mu_{is}(1-\mu_{is})\mu_{js}(1-\mu_{js})}\right] \equiv p_s(1-p_s)[1-d_\phi(i,j)], \tag{A.18}$$

where $d_\phi(i,j)$ is some function of the ancestry vectors for individuals i and j and the underlying subpopulation allele frequency covariance structure, $\mathbf{\Sigma}_K$. Plugging the appropriate values into Equation (A.17), it can be seen that

$$\begin{aligned}
\widehat{\phi}_{ij}^{PC} &\to \frac{\sum_{m\in\mathcal{M}_{ij}}\left[\left(\frac{1}{2}\right)^{(n_{im}+n_{jm}-1)}(1+f_m)[1-\mathbf{a}_m^T\mathbf{\Sigma}_K\mathbf{a}_m]\right]}{1-d_\phi(i,j)} \\
&= \sum_{m\in\mathcal{M}_{ij}}\left[\phi_{ij|m}\left(\frac{1-\mathbf{a}_m^T\mathbf{\Sigma}_K\mathbf{a}_m}{1-d_\phi(i,j)}\right)\right] \\
&= \sum_{m\in\mathcal{M}_{ij}}\left[\phi_{ij|m}\left(\frac{[1-d_\phi(i,j)]-[\mathbf{a}_m^T\mathbf{\Sigma}_K\mathbf{a}_m-d_\phi(i,j)]}{1-d_\phi(i,j)}\right)\right] \\
&= \sum_{m\in\mathcal{M}_{ij}}\left[\phi_{ij|m}\left(1-\frac{\mathbf{a}_m^T\mathbf{\Sigma}_K\mathbf{a}_m-d_\phi(i,j)}{1-d_\phi(i,j)}\right)\right] \\
&= \phi_{ij}[1-b_{\phi_1}(i,j)],
\end{aligned} \tag{A.19}$$

where the one bias term is given by the function

$$b_{\phi_1}(i,j) \equiv \frac{\sum_{m\in\mathcal{M}_{ij}}\phi_{ij|m}\left(\frac{\mathbf{a}_m^T\mathbf{\Sigma}_K\mathbf{a}_m-d_\phi(i,j)}{1-d_\phi(i,j)}\right)}{\sum_{m\in\mathcal{M}_{ij}}\phi_{ij|m}}. \tag{A.20}$$

Since this estimator uses adjusted genotype values that are centered according to individuals' specific ancestry, the second bias term that appears in the limiting value of the GRM is 0 for the ancestry adjusted estimator. As a result, $\widehat{\phi}_{ij}^{PC} \to 0$ for unrelated pairs of individuals, regardless of their ancestry and the underlying population

structure. However, the incorrect scaling of genotype values can not be fixed entirely without prior knowledge of the ancestries of all individuals in the set $\mathcal{M}_{ij}$. As a result, we can not claim consistency of $\widehat{\phi}_{ij}^{PC}$ for related pairs of individuals in all population structure scenarios.

We can show, however, that PC-Relate provides consistent estimates for relatives in the presence of discrete population substructure. If individuals i and j are relatives in this setting, they must belong to the same subpopulation, so $\mathbf{a}_i = \mathbf{a}_j = \mathbf{a}_m \equiv \mathbf{a}_*$ and $\mu_{is} = \mu_{js} = \mu_{ms} \equiv \mu_{*s}$ for every $m \in \mathcal{M}_{ij}$. As a result, Equation (A.18) simplifies to $\mathbb{E}[\mu_{*s}(1-\mu_{*s})] = p_s(1-p_s)[1 - \mathbf{a}_*^T \mathbf{\Sigma}_K \mathbf{a}_*]$, implying that $d_\phi(i,j) = \mathbf{a}_*^T \mathbf{\Sigma}_K \mathbf{a}_* = \mathbf{a}_m^T \mathbf{\Sigma}_K \mathbf{a}_m$ for every $m \in \mathcal{M}_{ij}$. Therefore, the bias term $b_{\phi_1}(i,j) = 0$, and

$$\widehat{\phi}_{ij}^{PC} \to \phi_{ij}. \tag{A.21}$$

Further, we have seen through simulations that the bias of the PC-Relate estimator tends to be very small, even in highly admixed populations. The bias only seems to become appreciable in populations with extremely disassortative mating, where it is still not large enough to lead to misclassification of relative types. Additionally, $b_{\phi_1}(i,j)$ is proportional to ϕ_{ij}, so the largest (though still small) bias arises for close relatives, which are easiest to identify.

A.2.3 KING-robust

We also derive the limiting behavior of the KING-robust kinship coefficient estimator under different sample structure settings. This estimator can be written as

$$\widehat{\kappa}_{ij} = \frac{1}{2}\left(1 - \frac{\sum_{s \in \mathcal{S}_{ij}} (g_{is} - g_{js})^2}{\sum_{s \in \mathcal{S}_{ij}} (\mathbb{1}_{[g_{is}=1]} + \mathbb{1}_{[g_{js}=1]})}\right). \tag{A.22}$$

Making the same assumption as for the previous two estimators, as the effective number of independent SNPs in $\mathcal{S}_{ij}$ tends to ∞,

$$\widehat{\kappa}_{ij} \to \frac{1}{2}\left(1 - \frac{\mathbb{E}[g_{is}^2] - 2\mathbb{E}[g_{is}g_{js}] + \mathbb{E}[g_{js}^2]}{\mathbb{E}[\mathbb{1}_{[g_{is}=1]}] + \mathbb{E}[\mathbb{1}_{[g_{js}=1]}]}\right). \tag{A.23}$$

We have already calculated $\mathbb{E}[g_{is}g_{js}]$ in Equation (A.8). The other expectations can be obtained directly from the observed genotype probabilities for individual i conditional on $\mathbf{p}_s$; however, it should be noted that these probabilities may not be what is expected under HWE based on individual i's individual specific allele frequencies, μ_{is}. The observed genotype probabilities are presented in Table A.1, and they take into account that individual i inherits one allele from its mother, $M(i)$, and one allele from its father, $P(i)$, at every locus.

Table A.1: Genotype Frequencies for Individual i at SNP s

Genotype	Individual-Specific Genotype Frequency	
	Expected (HWE)	Observed
AA	μ_{is}^2	$\mu_{M(i)s}\mu_{P(i)s}(1-f_i) + f_i\tilde{\mu}_{is}$
Aa	$2\mu_{is}(1-\mu_{is})$	$[\mu_{M(i)s}(1-\mu_{P(i)s}) + \mu_{P(i)s}(1-\mu_{M(i)s})](1-f_i)$
aa	$(1-\mu_{is})^2$	$(1-\mu_{M(i)s})(1-\mu_{P(i)s})(1-f_i) + f_i(1-\tilde{\mu}_{is})$

$\tilde{\mu}_{is}$ is the allele frequency given that $M(i)$ and $P(i)$ share the allele IBD

We see that departure from HWE in the observed genotype frequencies may result from two causes. The first is inbreeding, which leads to excess homozygosity. The second is recent admixture. If individual i's parents do not have the same ancestry, then individual i will have excess heterozygosity. In the case where individual i is outbred ($f_i = 0$) and has parents with the same ancestry ($\mu_{M(i)s} = \mu_{P(i)s} = \mu_{is}$), the observed genotype probabilities reduce to the same values as those expected under

HWE. From these observed genotype probabilities, we can calculate

$$\begin{aligned}
\mathbb{E}[g_{is}^2] &= \mathbb{E}[\mathbb{E}[g_{is}^2|\mathbf{p}_s]] \\
&= \mathbb{E}\left\{[\mu_{M(i)s}(1-\mu_{P(i)s}) + \mu_{P(i)s}(1-\mu_{M(i)s})](1-f_i)\right. \\
&\qquad \left. +4[\mu_{M(i)s}\mu_{P(i)s}(1-f_i) + f_i\tilde{\mu}_{is}]\right\} \\
&= 2p_s(1-p_s)[1 + \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}](1-f_i) + 4(p_s)^2(1-f_i) + 4f_ip_s \\
&= 2p_s(1-p_s)[1 + \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}] + 4(p_s)^2 + 2f_ip_s(1-p_s)[1 - \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}].
\end{aligned} \tag{A.24}$$

To obtain $\mathbb{E}[\mathbb{1}_{[g_{is}=1]}]$, we note that the expectation of an indicator function is just the probability of the event it indicates, so

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{[g_{is}=1]}] &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{[g_{is}=1]}|\mathbf{p}_s]] \\
&= \mathbb{E}[\mathbb{P}[g_{is}=1|\mathbf{p}_s]] \\
&= \mathbb{E}\left\{[\mu_{M(i)s}(1-\mu_{P(i)s}) + \mu_{P(i)s}(1-\mu_{M(i)s})](1-f_i)\right\} \\
&= 2p_s(1-p_s)[1 - \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}](1-f_i).
\end{aligned} \tag{A.25}$$

Plugging the appropriate expectations into Equation (A.23), through a lot of algebraic manipulation we get that

$$\widehat{\kappa}_{ij} \to \phi_{ij}[1 - b_{\kappa_1}(i,j)] + b_{\kappa_2}(i,j), \tag{A.26}$$

where we have defined the two bias terms for this estimator to be

$$b_{\kappa_1}(i,j) \equiv \frac{\sum_{m\in\mathcal{M}_{ij}} \phi_{ij|m}\left(\frac{\mathbf{a}_m^T\boldsymbol{\Sigma}_K\mathbf{a}_m - d_\kappa(i,j)}{1-d_\kappa(i,j)}\right)}{\sum_{m\in\mathcal{M}_{ij}}\phi_{ij|m}} \tag{A.27}$$

$$\text{and} \qquad b_{\kappa_2}(i,j) \equiv \frac{\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_j - d_\kappa(i,j)}{1-d_\kappa(i,j)}, \tag{A.28}$$

and we have also defined the function

$$\begin{aligned}
d_\kappa(i,j) \equiv \frac{1}{2}\left[\mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)} + f_i(1 - \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)})\right. \\
\left. +\mathbf{a}_{M(j)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(j)} + f_j(1 - \mathbf{a}_{M(j)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(j)})\right].
\end{aligned} \tag{A.29}$$

This first bias term, $b_{\kappa_1}(i,j)$, takes the same form as the bias term for the PC-Relate estimator, $b_{\phi_1}(i,j)$, but with the function $d_\kappa(i,j)$ in place of the function $d_\phi(i,j)$. The bias term $b_{\kappa_1}(i,j)$ is presumably small, as $b_{\phi_1}(i,j)$ is, although it is worth noting that $d_\kappa(i,j)$ is a function of the inbreeding coefficients for individuals i and j, whereas $d_\phi(i,j)$ is not. However, unlike the PC-Relate kinship coefficient estimator, the KING-robust estimator also has a second bias term, $b_{\kappa_2}(i,j)$. Similar to the GRM, this second bias term is the limiting value of the estimator for a pair of unrelated individuals; i.e $\widehat{\kappa}_{ij} \to b_{\kappa_2}(i,j)$ when $\phi_{ij} = 0$. We can look at the quantity $b_{\kappa_2}(i,j)$ in more detail to demonstrate that KING-robust may provide either a negatively or positively biased kinship coefficient estimate for a pair of individuals.

To start, consider a pair of outbred individuals, where $f_i = f_j = 0$, and the quantity in Equation (A.29) simplifies to $d_\kappa(i,j) = \frac{1}{2}\left[\mathbf{a}_{M(i)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(i)} + \mathbf{a}_{M(j)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(j)}\right]$. When the parents of individual i have the same ancestry, $\mathbf{a}_i = \mathbf{a}_{M(i)} = \mathbf{a}_{P(i)}$, and when the parents of individual j have the same ancestry, $\mathbf{a}_j = \mathbf{a}_{M(j)} = \mathbf{a}_{P(j)}$, the second bias term becomes

$$b_{\kappa_2}(i,j) = \frac{-\frac{1}{2}(\mathbf{a}_i - \mathbf{a}_j)^T \boldsymbol{\Sigma}_K (\mathbf{a}_i - \mathbf{a}_j)}{1 - \frac{1}{2}\left[\mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_i + \mathbf{a}_j^T \boldsymbol{\Sigma}_K \mathbf{a}_j\right]}. \tag{A.30}$$

If i and j also have the same ancestry, then $\mathbf{a}_i = \mathbf{a}_j$ and $b_{\kappa_2}(i,j) = 0$. However, if i and j have different ancestry, then $b_{\kappa_2}(i,j)$ is systematically negative, and the magnitude of this negative value is large when i and j have very different ancestry proportions. On the other hand, when the parents of an individual have very different ancestry, $b_{\kappa_2}(i,j)$ can become positive. To see this, consider a more extreme ancestry setting where the parents of individual i are from different populations, and the parents of individual j are from different populations. In this setting $\mathbf{a}_{M(i)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(i)} = \mathbf{a}_{M(j)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(j)} = 0$, so $d_\kappa(i,j) = 0$, and the second bias term becomes

$$b_{\kappa_2}(i,j) = \mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_j, \tag{A.31}$$

which is the same value as $b_{\psi_2}(i,j)$. If i and j have no common ancestry from any population, then $b_{\kappa_2}(i,j) = 0$, otherwise it is positive, and the magnitude is large if

i and j have similar ancestry proportions. In general, the quantity $b_{\kappa_2}(i,j)$ increases as $\mathbf{a}_i$ and $\mathbf{a}_j$ become more similar and it decreases as $\mathbf{a}_{M(i)}$ and $\mathbf{a}_{P(i)}$ or $\mathbf{a}_{M(j)}$ and $\mathbf{a}_{P(j)}$ become more similar. The kinship coefficient estimate for a pair of individuals with similar ancestry who are the offspring of recent admixture events, i.e. parents with different ancestry, will be positively biased. In contrast, the estimate for a pair of individuals with different ancestry who are each the offspring of parents with similar ancestry will be negatively biased. The magnitude of this bias term is also influenced by other factors. When the ancestral subpopulations are highly divergent, as measured by large values on the diagonal of $\mathbf{\Sigma}_K$, the absolute magnitude of the bias will increase. When either individual i or j is inbred, $d_\kappa(i,j)$ becomes larger, and $b_{\kappa_2}(i,j)$ becomes increasingly negative.

KING-robust was designed for discrete population substructure, and the limiting value in Equation (A.26) simplifies considerably in this setting. When individuals i and j are from the same subpopulation, $\mathbf{a}_i = \mathbf{a}_{M(i)} = \mathbf{a}_{P(i)} = \mathbf{a}_j = \mathbf{a}_{M(j)} = \mathbf{a}_{P(j)} = \mathbf{a}_m \equiv \mathbf{a}_*$ for every $m \in \mathcal{M}_{ij}$, and

$$\widehat{\kappa}_{ij} \to \frac{\phi_{ij} - \frac{1}{2}(f_i + f_j)}{1 - \frac{1}{2}(f_i + f_j)}, \tag{A.32}$$

once again illustrating that inbreeding leads to deflated kinship estimates. The only situation in which KING-robust provides consistent estimates is for a pair of outbred individuals from the same subpopulation, where

$$\widehat{\kappa}_{ij} \to \phi_{ij}. \tag{A.33}$$

Of course, as demonstrated from the scenarios above, when i and j are from different subpopulations, the KING-robust kinship estimates are biased and systematically negative.

A.3 Expectation of the Dominance Genotype Coding

Recall that g_{is}^D is an alternative genotype coding for individual i at SNP s, which we refer to as the dominance genotype coding. The values that g_{is}^D can take are a

function of $\widehat{\mu}_{is}$ and are given in Table 4.1 of the main text. In what follows, we derive the expectation of this genotype coding under different sets of assumptions, as the results will be used to explore the properties of the estimators $\widehat{f}_i^{PC}$ and $\widehat{\delta}_{ij}^{PC}$.

First, assume an outbred homogeneous population. In practice, we use $\widehat{\mu}_{is} = \hat{p}_s$ for every $i \in \mathcal{N}$ under this assumption, where $\hat{p}_s$ is an estimate of the population allele frequency. However, also assume that the true population allele frequency is known, so that $\widehat{\mu}_{is} = p_s$ can be used to construct g_{is}^D. Using the genotype probabilities for an outbred homogeneous population in HWE and this genotype coding, it can easily be shown that $\mathbb{E}[g_{is}^D] = p_s(1-p_s)$ and $\text{Var}[g_{is}^D] = [p_s(1-p_s)]^2$. If we allow that individuals in the population may be inbred, then it can also be shown that $\mathbb{E}[g_{is}^D] = p_s(1-p_s)(1+f_i)$.

Second, consider a structured population that is incorrectly assumed to be homogeneous. The coding for g_{is}^D still uses the population allele frequencies, but in the presence of population admixture, the observed genotype frequencies are those given in Table A.1. Computing the expectation of the dominance genotype coding in this setting, we derive

$$\begin{aligned}
\mathbb{E}[g_{is}^D] &= \mathbb{E}[\mathbb{E}[g_{is}^D|\mathbf{p}_s]] \\
&= \mathbb{E}\left[(1-p_s)[\mu_{M(i)s}\mu_{P(i)s}(1-f_i) + f_i\tilde{\mu}_{is}]\right. \\
&\qquad \left. + p_s[(1-\mu_{M(i)s})(1-\mu_{P(i)s})(1-f_i) + f_i(1-\tilde{\mu}_{is})]\right] \\
&= \left\{\mathbb{E}[\mu_{M(i)s}\mu_{P(i)s}] + p_s(1-\mathbb{E}[\mu_{M(i)s}] - \mathbb{E}[\mu_{P(i)s}])\right\}(1-f_i) \\
&\qquad + \{(1-p_s)\mathbb{E}[\tilde{\mu}_{is}] + p_s(1-\mathbb{E}[\tilde{\mu}_{is}])\}\, f_i \\
&= p_s(1-p_s)[1 + \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}](1-f_i) + 2p_s(1-p_s)f_i \\
&= p_s(1-p_s)\left\{1 + f_i[1 - \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}] + \mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}\right\}. \qquad \text{(A.34)}
\end{aligned}$$

On the other hand, if population homogeneity is not assumed, then the dominance genotype coding, g_{is}^D, is constructed using estimates of individual-specific allele frequencies. Assuming that these true values are known, so that $\widehat{\mu}_{is} = \mu_{is}$, the expecta-

tion of the dominance genotype coding is

$$\begin{aligned}
\mathbb{E}[g_{is}^D] &= \mathbb{E}[\mathbb{E}[g_{is}^D|\mathbf{p}_s]] \\
&= \mathbb{E}\left\{(1-\mu_{is})[\mu_{M(i)s}\mu_{P(i)s}(1-f_i)+f_i\tilde{\mu}_{is}]\right. \\
&\quad \left.+\mu_{is}[(1-\mu_{M(i)s})(1-\mu_{P(i)s})(1-f_i)+f_i(1-\tilde{\mu}_{is})]\right\} \\
&= \left\{\mathbb{E}[\mu_{M(i)s}\mu_{P(i)s}]-\mathbb{E}[\mu_{is}\mu_{M(i)s}]-\mathbb{E}[\mu_{is}\mu_{P(i)s}]+\mathbb{E}[\mu_{is}]\right\}(1-f_i) \\
&\quad +\left\{\mathbb{E}[\tilde{\mu}_{is}]+\mathbb{E}[\mu_{is}]-2\mathbb{E}[\mu_{is}\tilde{\mu}_{is}]\right\}f_i \\
&= p_s(1-p_s)[1+\mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}^T-\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}^T-\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_{M(i)}^T](1-f_i) \\
&\quad +2p_s(1-p_s)[1-\mathbf{a}_i^T\boldsymbol{\Sigma}_K\tilde{\mathbf{a}}_i]f_i \\
&= p_s(1-p_s)\left\{1+[\mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}^T-\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_{M(i)}^T-\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}^T]\right. \\
&\quad \left.+f_i[1-\mathbf{a}_{M(i)}^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}^T+\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_{M(i)}^T+\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_{P(i)}^T-2\mathbf{a}_i^T\boldsymbol{\Sigma}_K\tilde{\mathbf{a}}_i]\right\}, \quad \text{(A.35)}
\end{aligned}$$

where $\tilde{\mathbf{a}}_i$ gives the ancestry proportions for alleles shared IBD between $M(i)$ and $P(i)$. If we consider the setting where there is discrete population substructure, $\mathbf{a}_i = \mathbf{a}_{M(i)} = \mathbf{a}_{P(i)} = \tilde{\mathbf{a}}_i$, and we can see that

$$\mathbb{E}[g_{is}^D|\mathbf{p}_s] = \mu_{is}(1-\mu_{is})(1+f_i), \quad \text{(A.36)}$$

so Equation (A.35) simplifies considerably to

$$\mathbb{E}[g_{is}^D] = p_s(1-p_s)(1+f_i)[1-\mathbf{a}_i^T\boldsymbol{\Sigma}_K\mathbf{a}_i]. \quad \text{(A.37)}$$

A.4 Properties of the Inbreeding Coefficient Estimators

We first examine the limiting behavior of the GCTA inbreeding coefficient estimator, which can be written as

$$\widehat{f}_i = \frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}\left[\frac{g_{is}^D}{\hat{p}_s(1-\hat{p}_s)}\right]-1. \quad \text{(A.38)}$$

This estimator is often used under an assumption of population homogeneity, where $\widehat{\mu}_{is} = \hat{p}_s$ for every $i \in \mathcal{N}$ to construct g_{is}^D. Under the same set of relaxed convergence

assumptions as presented for $\widehat{\psi}_{ij}$, when the true p_s are known, we see by plugging in the expectation given in Equation (A.34) that

$$\widehat{f}_i \to \frac{\mathbb{E}[g_{is}^D]}{p_s(1-p_s)} - 1 = f_i[1 - \mathbf{a}_{M(i)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(i)}] + \mathbf{a}_{M(i)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(i)} \tag{A.39}$$

as $|\mathcal{S}_i| \to \infty$. Similar to the kinship coefficient estimates from the GRM, these inbreeding coefficient estimates are inflated due to the underlying population structure. For an outbred individual, $f_i = 0$, but $\widehat{f}_i \to \mathbf{a}_{M(i)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(i)}$. It is interesting to note that the amount of bias in the estimate for individual i is a function of how similar the ancestries of individual i's parents are.

In comparison, when using the PC-Relate inbreeding coefficient estimator

$$\widehat{f}_i^{PC} = \frac{\sum_{s \in \mathcal{S}_i} g_{is}^D}{\sum_{s \in \mathcal{S}_i} \widehat{\mu}_{is}(1-\widehat{\mu}_{is})} - 1, \tag{A.40}$$

population structure is accounted for by constructing g_{is}^D from estimates of individual-specific allele frequencies, $\widehat{\mu}_{is}$. Under the same set of relaxed convergence assumptions given above, when the true μ_{is} are known, we can see from the expectation given in Equation (A.35), a lot of algebraic manipulation, and the identity $\mathbf{a}_i = \frac{\mathbf{a}_{M(i)} + \mathbf{a}_{P(i)}}{2}$, that

$$\widehat{f}_i^{PC} \to \frac{\mathbb{E}[g_{is}^D]}{\mathbb{E}[\mu_{is}(1-\mu_{is})]} - 1 = f_i[1 - b_{f_1}(i)] + b_{f_2}(i) \tag{A.41}$$

as $|\mathcal{S}_i| \to \infty$, where we have defined the bias functions

$$b_{f_1}(i) \equiv \frac{\mathbf{a}_{M(i)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(i)} - \mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_i - 2\mathbf{a}_i^T \boldsymbol{\Sigma}_K (\mathbf{a}_i - \tilde{\mathbf{a}}_i)}{1 - \mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_i} \tag{A.42}$$

$$\text{and} \qquad b_{f_2}(i) \equiv \frac{\mathbf{a}_{M(i)}^T \boldsymbol{\Sigma}_K \mathbf{a}_{P(i)} - \mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_i}{1 - \mathbf{a}_i^T \boldsymbol{\Sigma}_K \mathbf{a}_i}. \tag{A.43}$$

Similar to the result for the PC-Relate kinship coefficient estimator, we can not claim consistency for the PC-Relate inbreeding coefficient estimator in all population structure scenarios; however, we can show that $\widehat{f}_i^{PC}$ is consistent in the presence of discrete population substructure. In this setting, $\mathbf{a}_i = \mathbf{a}_{M(i)} = \mathbf{a}_{P(i)} = \tilde{\mathbf{a}}_i$, so $b_{f_1}(i) = b_{f_2}(i) = 0$, and

$$\widehat{f}_i^{PC} \to f_i. \tag{A.44}$$

This result also implies that when the parents of individual i have similar ancestry, then the bias in the $\widehat{f}_i^{PC}$ estimator will be small, and this has been confirmed through simulation.

For an outbred individual, $f_i = 0$, and $\widehat{f}_i^{PC} \rightarrow b_{f_2}(i)$ under general population structure. This bias term can be rewritten as

$$b_{f_2}(i) = \frac{-\frac{1}{4}(\mathbf{a}_{M(i)} - \mathbf{a}_{P(i)})^T \mathbf{\Sigma}_K (\mathbf{a}_{M(i)} - \mathbf{a}_{P(i)})}{1 - \mathbf{a}_i^T \mathbf{\Sigma}_K \mathbf{a}_i}, \tag{A.45}$$

which shows that this bias is systematically negative and is a measure of the difference in individual i's parents' ancestries. When $M(i)$ and $P(i)$ have the same ancestry, the estimator will be consistent for 0, but as their ancestries grow further apart, this value will become increasingly negative. While it is not ideal that $\widehat{f}_i^{PC}$ provides negative estimates for individuals with parents of different ancestry, it can provide some insight into recent admixture events from highly divergent populations.

A.5 Properties of Dominance Genotype Estimators

To derive the limiting values of the estimators $\widehat{\delta}_{ij}$ and $\widehat{\delta}_{ij}^{PC}$, the expectation of the product of the dominance genotype values for two individuals must be examined. In an outbred homogeneous population, this can be calculated by considering the number of copies of independent alleles amongst the two individuals, conditional on

the possible IBD states:

$$\begin{aligned}
\mathbb{E}[g_{is}^D g_{js}^D] &= \sum_{\substack{\text{Geno}_{is}\in\{aa,Aa,AA\}\\ \text{Geno}_{js}\in\{aa,Aa,AA\}}} g_{is}^D g_{js}^D \mathbb{P}[\text{Geno}_{is}, \text{Geno}_{js}] \\
&= \sum_{\substack{\text{Geno}_{is}\in\{aa,Aa,AA\}\\ \text{Geno}_{js}\in\{aa,Aa,AA\}}} g_{is}^D g_{js}^D \sum_{\text{IBD}\in\{0,1,2\}} \mathbb{P}[\text{Geno}_{is}, \text{Geno}_{js}|\text{IBD}]\mathbb{P}[\text{IBD}] \\
&= (p_s)^2[(1-p_s)^2 k_{ij}^{(2)} + (1-p_s)^3 k_{ij}^{(1)} + (1-p_s)^4 k_{ij}^{(0)}] \\
&\quad + (1-p_s)^2[(p_s)^2 k_{ij}^{(2)} + (p_s)^3 k_{ij}^{(1)} + (p_s)^4 k_{ij}^{(0)}] \\
&\quad + 2p_s(1-p_s)[(p_s)^2(1-p_s)^2 k_{ij}^{(0)}] \\
&= [p_s(1-p_s)]^2(2k_{ij}^{(2)} + k_{ij}^{(1)} + k_{ij}^{(0)}) \\
&= [p_s(1-p_s)]^2(k_{ij}^{(2)} + 1). \qquad \text{(A.46)}
\end{aligned}$$

In the presence of inbreeding, an analogous derivation to the one given above, that instead conditions on the nine condensed IBD states given by Jacquard, gives us that

$$\begin{aligned}
\mathbb{E}[g_{is}^D g_{js}^D] = [p_s(1-p_s)]^2 \Bigg\{ &\frac{1-p_s(1-p_s)}{p_s(1-p_s)}\Delta_{ij}^{(1)} + 3\Delta_{ij}^{(2)} \\
&+\Delta_{ij}^{(3)} + \Delta_{ij}^{(4)} + \Delta_{ij}^{(5)} + \Delta_{ij}^{(6)} + \Delta_{ij}^{(7)} + 1 \Bigg\}. \qquad \text{(A.47)}
\end{aligned}$$

Unfortunately, the derivation of this quantity under general population structure, including ancestry admixture, becomes intractable; however, it is possible under the assumption of discrete population substructure. In this population structure setting, when individuals i and j are relatives they must come from the same subpopulation, which implies that $\mu_{is} = \mu_{js} \equiv \mu_{*s}$. From the same argument used to obtain the expectation in Equation (A.46), we can see that the conditional expectation of the product of dominance genotype values in this population structure setting is

$$\mathbb{E}[g_{is}^D g_{js}^D|\mathbf{p}_s] = [\mu_{*s}(1-\mu_{*s})]^2(k_{ij}^{(2)} + 1). \qquad \text{(A.48)}$$

A.5.1 Outbred Homogeneous Population

Recall the form of the $k_{ij}^{(2)}$ estimator given by Equation (4.7),

$$\widehat{\delta}_{ij} = \frac{1}{|\mathcal{S}_{ij}|} \sum_{s \in \mathcal{S}_{ij}} \frac{[g_{is}^D - \hat{p}_s(1-\hat{p}_s)][g_{js}^D - \hat{p}_s(1-\hat{p}_s)]}{[\hat{p}_s(1-\hat{p}_s)]^2}, \tag{A.49}$$

and the value of $\mathbb{E}[g_{is}^D g_{js}^D]$ calculated in Equation (A.46). With the same convergence assumptions that we previously made, when the true p_s are known, and with the law of large numbers, we obtain that

$$\begin{aligned}
\widehat{\delta}_{ij} &\to \frac{\mathbb{E}[g_{is}^D g_{js}^D] - p_s(1-p_s)\mathbb{E}[g_{is}^D] - p_s(1-p_s)\mathbb{E}[g_{js}^D] + [p_s(1-p_s)]^2}{[p_s(1-p_s)]^2} \\
&= \frac{[p_s(1-p_s)]^2(k_{ij}^{(2)}+1) - [p_s(1-p_s)]^2}{[p_s(1-p_s)]^2} \\
&= k_{ij}^{(2)}
\end{aligned} \tag{A.50}$$

as $|\mathcal{S}_{ij}| \to \infty$. As with $\widehat{\psi}_{ij}$, the convergence assumptions can be relaxed, and this result will still hold as long as the estimators $\hat{p}_s$ are consistent for p_s. From this derivation, we see that the estimator $\widehat{\delta}_{ij}$ provides consistent estimates of $k_{ij}^{(2)}$ in outbred homogeneous populations.

A.5.2 Structured Population

Now recall the form of the proposed PC-Relate $k_{ij}^{(2)}$ estimator for structured populations given by Equation (4.8):

$$\widehat{\delta}_{ij}^{PC} = \frac{\sum_{s \in \mathcal{S}_{ij}} [g_{is}^D - \widehat{\mu}_{is}(1-\widehat{\mu}_{is})(1+\widehat{f}_i^{PC})][g_{js}^D - \widehat{\mu}_{js}(1-\widehat{\mu}_{js})(1+\widehat{f}_j^{PC})]}{\sum_{s \in \mathcal{S}_{ij}} \widehat{\mu}_{is}(1-\widehat{\mu}_{is})\widehat{\mu}_{js}(1-\widehat{\mu}_{js})}. \tag{A.51}$$

As stated above, computing the limiting behavior of this estimator for relatives in the presence of ancestry admixture becomes intractable, so we start by focusing on the setting with outbred individuals from a population with discrete substructure. In this setting, $f_i = f_j = 0$, and assuming that i and j are relatives, they must come from the same subpopulation where $\mu_{is} = \mu_{js} \equiv \mu_{*s}$. We make the same set of relaxed

convergence assumptions as for $\widehat{\phi}_{ij}^{PC}$, and we assume that the true individual-specific allele frequencies are known so that $\widehat{\mu}_{is} = \widehat{\mu}_{js} = \mu_{*s}$. Using the expectation given in Equation (A.48) and the property that $\mathbb{E}[g_{is}^D g_{js}^D] = \mathbb{E}[\mathbb{E}[g_{is}^D g_{js}^D|\mathbf{p}_s]]$, as $|\mathcal{S}_{ij}| \to \infty$,

$$\begin{aligned}
\widehat{\delta}_{ij}^{PC} &\to \frac{\mathbb{E}[g_{is}^D g_{js}^D] - \mathbb{E}[g_{is}^D \mu_{*s}(1-\mu_{*s})] - \mathbb{E}[g_{js}^D \mu_{*s}(1-\mu_{*s})] + \mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2\right]}{\mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2\right]} \\
&= \frac{\mathbb{E}[g_{is}^D g_{js}^D] - \mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2\right]}{\mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2\right]} \\
&= \frac{\mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2(k_{ij}^{(2)}+1)\right] - \mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2\right]}{\mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2\right]} \\
&= k_{ij}^{(2)}, \qquad \text{(A.52)}
\end{aligned}$$

demonstrating that the PC-Relate estimator $\widehat{\delta}_{ij}^{PC}$ can also provide consistent estimation of $k_{ij}^{(2)}$ in populations with discrete substructure. The simplification in the second line comes from Equation (A.36) and recognition of the fact that μ_{*s} is a fixed quantity conditional on $\mathbf{p}_s$; therefore, $\mathbb{E}[g_{is}^D \mu_{*s}(1-\mu_{*s})] = \mathbb{E}[\mu_{*s}(1-\mu_{*s})\mathbb{E}[g_{is}^D|\mathbf{p}_s]] = \mathbb{E}\left[[\mu_{*s}(1-\mu_{*s})]^2\right]$. While we can not show consistency of this estimator for relatives in the presence of ancestry admixture, similar to the PC-Relate kinship coefficient estimator, simulations show that the bias is generally small.

On the other hand, we can consider a pair of unrelated individuals under general population structure and with possible inbreeding. When i and j are unrelated, $g_{is}^D \perp g_{js}^D$ and $\mu_{is} \perp \mu_{js}$. Additionally, we know from Equation (A.41) that the plug-in estimator $\widehat{f}_i^{PC}$ converges to the quantity $\frac{\mathbb{E}[g_{is}^D]}{\mathbb{E}[\mu_{is}(1-\mu_{is})]} - 1$. With these facts in mind,

we see that in this setting

$$\begin{aligned}
\widehat{\delta}_{ij}^{PC} &\rightarrow \frac{\mathbb{E}\left[[g_{is}^{D} - \mu_{is}(1-\mu_{is})(1+\widehat{f}_{i}^{PC})][g_{js}^{D} - \mu_{js}(1-\mu_{js})(1+\widehat{f}_{j}^{PC})]\right]}{\mathbb{E}[\mu_{is}(1-\mu_{is})\mu_{js}(1-\mu_{js})]} \\
&= \frac{\mathbb{E}\left[g_{is}^{D} - \mu_{is}(1-\mu_{is})\frac{\mathbb{E}[g_{is}^{D}]}{\mathbb{E}[\mu_{is}(1-\mu_{is})]}\right]\mathbb{E}\left[g_{js}^{D} - \mu_{js}(1-\mu_{js})\frac{\mathbb{E}[g_{js}^{D}]}{\mathbb{E}[\mu_{js}(1-\mu_{js})]}\right]}{\mathbb{E}[\mu_{is}(1-\mu_{is})\mu_{js}(1-\mu_{js})]} \\
&= \frac{(\mathbb{E}[g_{is}^{D}] - \mathbb{E}[g_{is}^{D}])(\mathbb{E}[g_{js}^{D}] - \mathbb{E}[g_{js}^{D}])}{\mathbb{E}[\mu_{is}(1-\mu_{is})\mu_{js}(1-\mu_{js})]} \\
&= 0. \qquad \text{(A.53)}
\end{aligned}$$

This demonstrates that $\widehat{\delta}_{ij}^{PC}$ will provide accurate estimates of 0 for unrelated pairs of individuals in any population structure setting as long as $\widehat{f}_{i}^{PC}$ and $\widehat{f}_{j}^{PC}$ are used to account for possible departures from HWE due to inbreeding or recent admixture events when centering the dominance genotype values.

A.5.3 Inbred Homogeneous Population

Finally, we return to a homogeneous population, but now with the possibility that individual i and/or j may be inbred. As discussed previously, in the presence of inbreeding there are no longer only three IBD states, but there are now nine condensed IBD states described by Jacquard. Assume that the true population allele frequencies are known, so that $\widehat{\mu}_{is} = \widehat{\mu}_{js} = p_s$, and that the true inbreeding coefficients are known, so that $\widehat{f}_{i}^{PC} = f_i$ and $\widehat{f}_{j}^{PC} = f_j$. Under the same convergence assumptions as presented

for the outbred setting, as $|\mathcal{S}_{ij}| \to \infty$, we see that

$$\begin{aligned}
\widehat{\delta}_{ij}^{PC} &\to \frac{\mathbb{E}[g_{is}^D g_{js}^D] - p_s(1-p_s)(1+f_j)\mathbb{E}[g_{is}^D] - p_s(1-p_s)(1+f_i)\mathbb{E}[g_{js}^D]}{[p_s(1-p_s)]^2} \\
&\quad + \frac{[p_s(1-p_s)]^2(1+f_i)(1+f_j)}{[p_s(1-p_s)]^2} \\
&= \frac{[p_s(1-p_s)]^2\left\{\frac{1-p_s(1-p_s)}{p_s(1-p_s)}\Delta_{ij}^{(1)} + 3\Delta_{ij}^{(2)} + \Delta_{ij}^{(3)} + \Delta_{ij}^{(4)} + \Delta_{ij}^{(5)} + \Delta_{ij}^{(6)} + \Delta_{ij}^{(7)} + 1\right\}}{[p_s(1-p_s)]^2} \\
&\quad + \frac{-[p_s(1-p_s)]^2(1+f_i)(1+f_j)}{[p_s(1-p_s)]^2} \\
&= c_p\Delta_{ij}^{(1)} + \Delta_{ij}^{(2)} + \Delta_{ij}^{(7)} - f_i f_j \qquad \text{(A.54)}
\end{aligned}$$

where we used the identities $f_i = \Delta_{ij}^{(1)} + \Delta_{ij}^{(2)} + \Delta_{ij}^{(3)} + \Delta_{ij}^{(4)}$ and $f_j = \Delta_{ij}^{(1)} + \Delta_{ij}^{(2)} + \Delta_{ij}^{(5)} + \Delta_{ij}^{(6)}$, and where $c_p = \frac{\sum_{s\in\mathcal{S}_{ij}} p_s(1-p_s)[1-3p_s(1-p_s)]}{\sum_{s\in\mathcal{S}_{ij}}[p_s(1-p_s)]^2}$ is a function of the allele frequency distribution across SNPs. Unlike all of the other estimators discussed, in this particular situation, the limiting value of this estimator is a function of the population allele frequencies, and $c_p = 1$ if $p_s = 0.5$ for every $s \in \mathcal{S}_{ij}$. This will not hold in most realistic settings; however, $\Delta_{ij}^{(1)} = 0$ for most pairs of individuals, and this should not be of major concern. Note that for a pair of independent, i.e. unrelated, individuals, $\Delta_{ij}^{(1)} = \Delta_{ij}^{(7)} = 0$, $\Delta_{ij}^{(2)} = f_i f_j$, and $\widehat{\delta}_{ij}^{PC} \to 0$. If either individual in a relative pair is outbred, then $\Delta_{ij}^{(1)} = \Delta_{ij}^{(2)} = f_i f_j = 0$, and $\widehat{\delta}_{ij}^{PC} \to \Delta_{ij}^{(7)}$, the analog to $k_{ij}^{(2)}$ in this setting.

Appendix B

SOFTWARE

All of the statistical methodology presented in this dissertation has been implemented efficiently in the R language and will be released as a freely downloadable R package. The implementation of these methods interfaces with the GWASTools package [13] and takes advantage of the GDS genotype format implemented in SNPRelate [71].

B.1 PC-AiR

A full implementation of the PC-AiR procedure for inferring ancestry representative principal components is available. Additionally, a function that only performs the partitioning of the entire sample into the ancestry representative mutually unrelated subset and the related set, i.e. performs Algorithm 3.1, is also available. Both of these functions allow for specification of different kinship coefficient and ancestry divergence measures, and they also allow for the specification of individuals, such as reference panel samples, that should be designated for inclusion in the unrelated set. An extension to the plot function has been created that allows for easy plotting and visualization of pairs of principal components. A useful tool for converting the default output from KING-robust into an R matrix is also included.

B.2 PC-Relate

A full implementation of the PC-Relate procedure for estimating genetic relatedness measures is available. By default, inbreeding coefficients as well as all pairwise measures including kinship coefficients, and probabilities of sharing 0, 1, and 2 copies of alleles IBD are estimated. An option to not calculate IBD sharing probabilities is

provided in order to speed up computation. A version of the function which allows for computation on a single chromosome enables easy parallelization. An additional function for combining the output across sets of chromosomes is provided.

B.3 MMAAPS

An analysis with MMAAPS requires use of a number of available functions that we have implemented. First, MMAAPS requires ancestry representative principal components as well as an ancestry adjusted kinship matrix. The ancestry representative PCs can be found using the PC-AiR function, and the ancestry adjusted kinship matrix can be found using a function that takes the output from PC-Relate as its input.

Second, the current implementation of MMAAPS performs an approximate test, estimating the variance components once under the null model with no genotype in the mean model. A function is provided to perform the variance component estimation using the average information REML procedure [12]. This is the same procedure that GCTA [65] uses, but extended to allow for environmental random effects. Another function is provided to compute confidence intervals for the variance components or the proportion of variability they explain from the output of the estimation procedure.

Finally, association testing is performed at each SNP using either a Wald or score test that is computationally similar to the linear mixed model methodology implemented in EMMAX [21]. For efficient computation, we extend the semi-parallelization technique implemented by Sikorska et al. [50] for linear regression to allow for correlated samples and perform generalized least squares to fit the linear mixed model. In brief, the phenotype covariance matrix estimated under the null hypothesis is treated as fixed and known, and the Cholesky decomposition of this matrix is used to "decorrelate" the outcome and covariate vectors. The regression of these decorrelated variables can then be fit efficiently using semi-parallelization of ordinary least squares.

BIBLIOGRAPHY

[1] Mark Abney. A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics*, 25(12):1561–1563, 2009.

[2] Mark Abney, Mary Sara McPeek, and Carole Ober. Estimation of variance components of quantitative traits in inbred populations. *The American Journal of Human Genetics*, 66(2):629–650, 2000.

[3] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

[4] 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.

[5] Gaurav Bhatia, Nick Patterson, Sriram Sankararaman, and Alkes L Price. Estimating and interpreting fst: The impact of rare variants. *Genome Research*, 23(9):1514–1521, 2013.

[6] Chia-Yen Chen, Samuela Pollack, David J Hunter, Joel N Hirschhorn, Peter Kraft, and Alkes L Price. Improved ancestry inference using weights from external reference panels. *Bioinformatics*, 29(11):1399–1406, 2013.

[7] Riyan Cheng and Maintainer Riyan Cheng. Package 'qtlrel'. 2013.

[8] C Clark Cockerham and BS Weir. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics*, pages 157–164, 1984.

[9] B Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[10] Ronald Aylmer Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[11] LA García-Cortés, A Legarra, and MA Toro. The coefficient of dominance is not (always) estimable with biallelic markers. *Journal of Animal Breeding and Genetics*, 2014.

[12] Arthur R Gilmour, Robin Thompson, and Brian R Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450, 1995.

[13] Stephanie M Gogarten, Tushar Bhangale, Matthew P Conomos, Cecelia A Laurie, Caitlin P McHugh, Ian Painter, Xiuwen Zheng, David R Crosslin, David Levine, Thomas Lumley, et al. Gwastools: an r/bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, 28(24):3329–3331, 2012.

[14] Simon Gravel, Fouad Zakharia, Andres Moreno-Estrada, Jake K Byrnes, Marina Muzzio, Juan L Rodriguez-Flores, Eimear E Kenny, Christopher R Gignoux, Brian K Maples, Wilfried Guiblet, et al. Reconstructing native american migrations from whole-genome and whole-exome data. *PLoS Genetics*, 9(12):e1004023, 2013.

[15] Ben John Hayes, Peter M Visscher, and Michael E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, 91(01):47–60, 2009.

[16] Simon C Heath, Ivo G Gut, Paul Brennan, James D McKay, Vladimir Bencko, Eleonora Fabianova, Lenka Foretova, Michael Georges, Vladimir Janout, Michael Kabesch, et al. Investigation of the fine structure of european populations with applications to disease association studies. *European Journal of Human Genetics*, 16(12):1413–1429, 2008.

[17] WG Hill and BS Weir. Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genetics Research*, 93(01):47–64, 2011.

[18] International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[19] Albert Jacquard et al. Genetic structures of populations. *Structures Genetiques des Populations.*, 1970.

[20] Luc Janss, Gustavo de los Campos, Nuala Sheehan, and Daniel Sorensen. Inferences from genomic models in stratified populations. *Genetics*, 192(2):693–704, 2012.

[21] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

[22] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[23] BW Kennedy, LR Schaeffer, and DA Sorensen. Genetic properties of animal models. *Journal of Dairy Science*, 71:17–26, 1988.

[24] Cathy C Laurie, Kimberly F Doheny, Daniel B Mirel, Elizabeth W Pugh, Laura J Bierut, Tushar Bhangale, Frederick Boehm, Neil E Caporaso, Marilyn C Cornelis, Howard J Edenberg, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34(6):591–602, 2010.

[25] Seunggeun Lee, Fei Zou, and Fred A Wright. Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics*, 38(6):3605, 2010.

[26] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, 2008.

[27] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[28] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–526, 2012.

[29] Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA, 1998.

[30] Jianzhong Ma and Christopher I Amos. Principal components analysis of population admixture. *PloS One*, 7(7):e40115, 2012.

[31] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[32] Ani Manichaikul, Walter Palmas, Carlos J Rodriguez, Carmen A Peralta, Jasmin Divers, Xiuqing Guo, Wei-Min Chen, Quenna Wong, Kayleen Williams, Kathleen F Kerr, et al. Population structure of hispanics in the united states: the multi-ethnic study of atherosclerosis. *PLoS Genetics*, 8(4):e1002640, 2012.

[33] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[34] Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

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

[36] Brook G Milligan. Maximum-likelihood estimation of relatedness. *Genetics*, 163(3):1153–1167, 2003.

[37] Ida Moltke and Anders Albrechtsen. Relateadmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics*, 30(7):1027–1028, 2014.

[38] Jean Morrison. Characterization and correction of error in genome-wide ibd estimation for samples with population structure. *Genetic Epidemiology*, 2013.

[39] Nick Patterson, Neil Hattangadi, Barton Lane, Kirk E Lohmueller, David A Hafler, Jorge R Oksenberg, Stephen L Hauser, Michael W Smith, Stephen J O'Brien, David Altshuler, et al. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

[40] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

[41] 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.

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

[43] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[44] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[45] David Reich, Michael A Nalls, WH Linda Kao, Ermeg L Akylbekova, Arti Tandon, Nick Patterson, James Mullikin, Wen-Chi Hsueh, Ching-Yu Cheng, Josef Coresh, et al. Reduced neutrophil count in people of african descent is due to a regulatory variant in the duffy antigen receptor for chemokines gene. *PLoS Genetics*, 5(1):e1000360, 2009.

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

[47] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008.

[48] Peter H Schönemann and Robert M Carroll. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35(2):245–255, 1970.

[49] Daniel Shriner. Overview of admixture mapping. *Current Protocols in Human Genetics*, pages 1–23, 2013.

[50] Karolina Sikorska, Emmanuel Lesaffre, Patrick FJ Groenen, and Paul HC Eilers. Gwas on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics*, 14(1):166, 2013.

[51] Michael W Smith, Nick Patterson, James A Lautenberger, Ann L Truelove, Gavin J McDonald, Alicja Waliszewska, Bailey D Kessing, Michael J Malasky, Charles Scafe, Ernest Le, et al. A high-density admixture map for disease gene

discovery in african americans. *The American Journal of Human Genetics*, 74(5):1001–1013, 2004.

[52] Jeffrey Staples, Deborah A Nickerson, and Jennifer E Below. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genetic Epidemiology*, 37(2):136–141, 2013.

[53] Gulnara R Svishcheva, Tatiana I Axenovich, Nadezhda M Belonogova, Cornelia M van Duijn, and Yurii S Aulchenko. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, 44(10):1166–1170, 2012.

[54] Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005.

[55] E. A. Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39(2):173–188, 1975.

[56] Timothy Thornton, Hua Tang, Thomas J Hoffmann, Heather M Ochs-Balcom, Bette J Caan, and Neil Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, 2012.

[57] Timothy A Thornton and Justo Lorenzo Bermejo. Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genetic Epidemiology*, 38(S1):S5–S12, 2014.

[58] Zulma G Vitezica, Luis Varona, and Andres Legarra. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, 195(4):1223–1230, 2013.

[59] Chaolong Wang, Zachary A Szpiech, James H Degnan, Mattias Jakobsson, Trevor J Pemberton, John A Hardy, Andrew B Singleton, and Noah A Rosenberg. Comparing spatial maps of human population-genetic variation using procrustes analysis. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[60] Chaolong Wang, Xiaowei Zhan, Jennifer Bragg-Gresham, Hyun Min Kang, Dwight Stambolian, Emily Y Chew, Kari E Branham, John Heckenlively, The FUSION Study, Robert Fulton, et al. Ancestry estimation and control of population stratification for sequence-based association studies. *Nature Genetics*, 46(4):409–415, 2014.

[61] Bruce S Weir and C Clark Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution*, pages 1358–1370, 1984.

[62] Sewall Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, 1949.

[63] Chengqing Wu, Andrew DeWan, Josephine Hoh, and Zuoheng Wang. A comparison of association methods correcting for population stratification in case–control studies. *Annals of Human Genetics*, 75(3):418–427, 2011.

[64] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

[65] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[66] Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza de Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature Genetics*, 43(6):519–525, 2011.

[67] Jian Yang, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'Connell, Massimo Mangino, et al. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19(7):807–812, 2011.

[68] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, 2014.

[69] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355–360, 2010.

[70] Q. Zheng and C. Bourgain. Kininbcoef: Calculation of kinship and inbreeding coefficients. *(http://galton.uchicago.edu/ mcpeek/software/index.html)*, 2009.

[71] Xiuwen Zheng, David Levine, Jess Shen, Stephanie M Gogarten, Cathy Laurie, and Bruce S Weir. A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, 28(24):3326–3328, 2012.

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

[73] Xiaofeng Zhu, Shengchao Li, Richard S Cooper, and Robert C Elston. A unified association analysis approach for family and unrelated samples correcting for stratification. *The American Journal of Human Genetics*, 82(2):352–365, 2008.