Statistical Methods for the Analysis of Autosomal and X Chromosome Genetic Data in Samples with Unknown Structure

Caitlin P. McHugh

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Timothy A. Thornton, Chair

Bruce S. Weir

Ellen M. Wijsman

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Methods for the Analysis of Autosomal and X Chromosome Genetic Data in Samples with Unknown Structure

Caitlin P. McHugh

Chair of the Supervisory Committee:
Associate Professor Timothy A. Thornton
Department of Biostatistics

Genome-wide association studies (GWAS) and sequencing association studies are routinely conducted for the mapping of genes to complex traits. Genetic variants on the X chromosome could potentially play an important role in some complex traits, however, statistical methods for association studies have primarily been developed for variants on the autosomal chromosomes with significantly less attention given to the X chromosome. Existing association methods for variants on the autosomal chromosomes will typically not be valid for the analysis of X-linked variants due to the X chromosome having a different correlation structure than the autosomes as well as copy number differences for males and females on the X. This dissertation develops and applies new statistical methodology for genetic analysis of variants on the X chromosome. In particular, we focus on methods that are computationally feasible for large-scale genomic data for detecting genetic associations with common and rare variants from GWAS and sequencing studies. Furthermore, the proposed methods allow for valid genetic analysis in the presence of complex sample structures, such as population structure and cryptic relatedness among sampled individuals. Another aspect of this dissertation is the development of statistical methods for inference of heterogeneity in ancestry across the genome (including the X chromosome) in recently admixed populations, such as African Americans and Hispanics, who have experienced admixing within the past few hundred years from two or more continental groups that were previously isolated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Dr. Timothy Thornton guided me the entire way through this dissertation and offered support and mentorship without which I wouldn't have achieved this great accomplishment. I would like to acknowledge the other mentors I've had during my years at UW, most especially Drs. Bruce Weir and Cathy Laurie. Finally, I will always appreciate the support and teachings of my classmates, both inside the classroom and out.

DEDICATION

to my parents, who taught me math was cool

and to my husband, who has always loved me for my nerdiness

Chapter 1

INTRODUCTION

Recent advancements in high-throughput genotyping and whole-genome sequencing technologies have transformed human genetic research, enabling large amounts of genetic data to be produced both quickly and inexpensively. To date, genome-wide association studies (GWAS) have identified thousands of genetic polymorphisms associated with a variety of human diseases and traits. Very few of these associations are with variants on the X chromosome, and recent work has highlighted the lack of associations identified on the X in the GWAS era [59]. Indeed, in 2014 the GWAS catalog included only 135 X-linked associations out of over 19,000 published GWAS results [17]. Although the size and number of genes on the X chromosome are similar to chromosome 7, the X chromosome is often excluded from association analyses.

The uncovered variants from GWAS have largely been of small effect and explain only a small fraction of trait heritability. Rare and low frequency variants, those with minor allele frequencies less than 5%, likely play a significant role in many complex traits and may explain some of the missing heritability not explained by the common variants identified through GWAS. Detecting rare variant associations from GWAS data is difficult due to rare variants having low linkage disequilibrium (LD) with common variants on the single-nucleotide polymorphism (SNP) genotyping arrays used in GWAS. Whole-genome and whole-exome sequencing studies are now routinely conducted for the identification of rare and low frequency variants that are involved with complex traits. A commonly used statistical method for testing rare and low frequency variants is the burden test. However, burden tests offer little power when variants are associated with the trait in opposite directions. Kernel-based association methods gain power in settings where the burden test lacks power. The SKAT method [60], developed for unrelated samples, and a recent extension

of SKAT to pedigree-based samples, fam-SKAT [6] have been widely used. These methods are based upon a linear mixed effects model framework that jointly tests effects of multiple variants in a region on the autosomes. Both SKAT and fam-SKAT are computationally efficient and have been demonstrated to provide higher power in a variety of settings over previously proposed rare variant association approaches. An 'omnibus' framework has been proposed that assesses the linear combination of a SKAT and burden test, and chooses the most significant combination of the two. Such methods are SKAT-O [26] and MONSTER [19], the latter of which allows for a known pedigree structure. However, none of the published methods mentioned above are directly applicable for association testing with rare or low frequency variants on the X chromosome, nor are these methods suitable for samples with partially or completely unknown pedigree structure.

Statistical methods for detecting genetic associations with either rare or common variants have predominantly been developed for the analysis of markers on the autosomes, with significantly less attention given to the analysis of markers on the X chromosome. Autosomal association methods will typically not be valid for X chromosome analysis. Some of the methodological challenges for the X include (1) accounting for X chromosome copy number differences in females and males and (2) appropriately adjusting for genetic correlations among sampled individuals on the X, including pedigree and population structure (ancestry differences among sample individuals), which can be quite different from the autosomes. The relatively few methodology papers in the scientific literature addressing the problem of X chromosome association testing is disappointing given the importance of this topic. Novel statistical methods development is needed for X chromosome analysis of existing and future GWAS and sequencing studies.

Human genetic studies using GWAS data have primarily examined populations of European ancestry. More recent studies involve admixed populations, defined here as populations derived within the last few hundred years with ancestry from two or more continental groups that were previously isolated. Examples of admixed populations include the two largest minority populations in the United States: African Americans and Hispanics. Recent studies [63, 3, 35] have suggested that systematic differences in ancestry among admixed populations may arise as a result of selection and sex-specific patterns of non-random mat-

ing at the time of (or since) admixture. Identifying heterogeneity in population structure patterns across the genome, including the X chromosome, is important to applications in population genetics and genetic association studies of admixed populations.

The specific aims for this dissertation are as follows:

(1) To develop a statistical method to detect and quantify heterogeneity in population structure across chromosomes;

(2) To develop a mixed linear model (MLM) method to perform X chromosome association mapping in samples with related individuals and/or population structure; and

(3) To propose an extension of existing rare variant association methods to allow for X chromosome effects testing and adjustment.

1.1 Chromosomal Ancestry Differences in Admixed Populations

The genetic structure of both ancestrally homogeneous and admixed populations has largely been characterized using dense markers on the autosomal chromosomes. Heterogeneity in ancestry across the autosomes can be used to detect regions under selection, as was done in a sample of Puerto Rican individuals [54]. While it may be reasonable to assume that population structure patterns on the autosomes and the X chromosome are the same for structurally homogeneous populations, this may not be true for admixed populations. Sex-specific patterns of non-random mating as a consequence of historical events such as the transatlantic slave trade and the colonization of the Americas yielded differences between autosomal and X chromosome genetic patterns. A recent population structure study [3] found evidence of increased African ancestry on the X chromosome relative to the autosomes in a sample of African Americans. A more recent study [4] employed a simple pooled t-test to compare ancestry on the autosomes with X chromosome ancestry in a large sample from the United States. Along with confirming previous findings in African Americans, the authors identified increased Native American ancestry on the X chromosome as compared to the autosomes in the set of Hispanic Americans included for analysis. Although studies have begun to summarize the ancestry of individuals from admixed populations across the

genome, no formal statistical procedure for identifying and assessing systematic ancestry differences across the genome has been developed.

In Chapter 2 of this dissertation, we propose the chromosomal ancestry differences (CAnD) method for the detection of heterogeneity in population structure patterns across chromosomes. CAnD uses local ancestry inferred from SNP genotype data to identify chromosomes harboring genomic regions with ancestry contributions that are significantly different than expected. The CAnD method takes into account correlated ancestries among chromosomes within individuals for both valid testing and improved power for detecting heterogeneity in population structure across the genome. In simulation studies with real genotype data from Phase III of the International HapMap project [15], we demonstrate the validity and power of CAnD. We apply CAnD to the HapMap Mexican American (MXL) and African American (ASW) population samples; in this analysis the software RFMix [37] is used to infer local ancestry at genomic regions assuming admixing from Europeans, West Africans, and Native Americans. We found no significant differences in proportion ancestry among the autosomal chromosomes in either population sample. The X chromosome in the MXL, however, showed elevated levels of Native American ancestry proportions and deficiencies in European ancestry proportions that are highly significant. Furthermore, CAnD detects highly significant heterogeneity in ancestry when using either average local ancestry estimates across a chromosome from RFMix or global chromosomal ancestry estimates from FRAPPE. These results are consistent with a sex-biased pattern of gene flow with an excess of European male and Native American female ancestry in the MXL. We also applied the CAnD method to over 12,000 Hispanic/Latino individuals living in the United States recruited as part of the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Local ancestry estimates from RFMix [37], averaged by chromosome, were stratified by self-identified ancestry group background then analyzed. We detect a highly significant difference in all ancestries and self-identified background groups with the CAnD method comparing the X chromosome to the autosomes. We also detect heterogeneity in ancestry proportions when comparing each autosome to a pool of the remaining autosomes. We believe that genomic regions identified as having unusual ancestry distributions are strong candidates for being under selection and may provide new insight into the population history

of Hispanics/Latino populations.

The accompanying `R` package, `CAnD`, is published in BioConductor. With this package, users are able to easily apply the CAnD method to their datasets.

1.2 Association Mapping in Samples With Complex Structure Using Mixed Linear Models

Statistical methods for GWAS have primarily been developed for variants on the autosomal chromosomes with significantly less attention given to the X chromosome. A mixed linear model (MLM) approach in GWAS was first proposed in 2006 [62]. The MLM approach was promising because it offers powerful and accurate association testing in the presence of population structure and/or relatedness, however, it remained impractical due to intensive computation times. In 2010, a computationally feasible approach, EMMAX [22], was implemented and enabled MLMs to be applied to GWAS datasets. Despite major advancements in developing computationally efficient LMMs for genetic association testing "genome-wide," existing MLM approaches, , for example [22] [65] [18] [29], are not directly applicable for association testing on X chromosome markers using GWAS data. Some of the methodological challenges for association testing on the X include (1) accounting for X chromosome copy number differences in females and males and (2) appropriately adjusting for genetic correlations among sample individuals on the X, including pedigree and population structure, and ancestry admixture, which can be quite different from the autosomes.

Chapter 3 of this dissertation explores the mixed linear model framework for testing association with common variants on the X chromosome, in samples with unknown structure. To do so, we first explored estimating relatedness on the X chromosome. Simulation studies showed that with a realistic number of SNPs, we can reliably estimate relatedness on the X chromosome, similar to relatedness estimation from autosomal genetic markers. We propose the MLM-X method, which includes random effects for polygenic effects on both the X chromosome and the autosomes while appropriately taking into account unique correlation on the X chromosome. Variance components of the random effects in MLM-X are calculated via average information restricted maximum likelihood (AI-REML) with two empirical genetic relationship matrices (GRMs), one for the autosomes and one for the X

chromosome. Similar to LMMOPS [8], the inclusion of additional random effects, such as shared environmental exposures, is a simple extension. In simulation studies, we demonstrate that MLM-X can provide an improvement over existing MLM approaches, in terms of type I error and power, for complex trait mapping with both autosomal and X-linked variants. We applied MLM-X to a cohort of over 12,000 Hispanic/Latino individuals recruited from four metropolitan areas in the U.S. to identify genetic variants associated with red blood cell count (RBC). MLM-X detects genome-wide significant associations with RBC on the X chromosome and has a lower genomic control inflation factor than existing MLM approaches that ignore polygenic effects on the X chromosome.

1.3 Rare Variant Association Testing Genome-Wide in Samples with Complex Structure

GWAS focus on assaying and testing relatively common variants. Whole-genome and whole-exome sequencing assay rare genetic variants; here we define rare genetic variants as variants with minor allele frequency (MAF) less than 1% and low frequency variants as those with MAF between 1% and 5%. Rare variants likely play an important role in complex traits, and there is evidence that deleterious variants are more likely to be rare than common in a population. Single variant association testing methods that are often employed for common variants have poor power for detecting rare variant associations. In addition, there are many more rare variants across the genome than common variants, resulting in a prohibitive multiple testing correction penalty with single variant association testing with rare variants. As a result, gene-based region testing is usually employed in rare variant association studies [27].

'Optimal' or 'omnibus' tests have been proposed for rare variant association testing that are a convex combination of the burden and kernel-based tests [26, 19]. These methods capitalize on the fact that the burden and kernel-based association tests are most powerful in different scenarios and allow for adaptive weighting of the burden and kernel-based statistics, identifying the convex combination that yields the minimum p-value. An overall optimal p-value is then determined from the minimum p-value found from the combination of the two test statistics. Depending on the underlying genetic basis of the trait, the optimal

test identifies the burden test, kernel-based test, or a linear combination of the two as the 'optimal' test for a particular gene region and trait of interest. Extensions of previous methods allow for known pedigree structure, such as in MONSTER [19].

In Chapter 4, we propose an optimal kernel-based association testing method in structured samples (KEATS-O) for gene-based rare variant association testing that is valid in the presence of complex sample structure. The KEATS-O method is properly calibrated for testing both autosomal and X chromosome gene regions, and the optimal framework allows for results that are robust to the underlying genetic architecture of the trait of interest. We use an empirical genetic relatedness matrix, allowing for accurate adjustment of cryptic, unknown and documented relatedness. Our approach is flexible and allows for inclusion of any number of random effects in the model, such as known environmental correlation. We show via simulation that KEATS-O is properly calibrated in terms of type I error and yields high power when testing various configurations of causal variants when considering either X chromosome or autosomal variants. We apply KEATS-O to a large-scale genetic study of Hispanic/Latino individuals from the U.S. and consider genetic association testing with red blood cell count (RBC) and 15 gene regions previously published to be associated with RBC, none of which were studies focusing on Hispanic/Latino populations. We generalized association with six candidate genes to the Hispanic/Latino populations. Finally, we apply KEATS-O to gene regions genome-wide and identify six genes associated with RBC in the HCHS/SOL sample of Hispanic/Latino individuals that reach genome-wide significance.

Chapter 2

DETECTING HETEROGENEITY IN POPULATION STRUCTURE ACROSS THE GENOME IN ADMIXED POPULATIONS

2.1 Introduction

Technological advancements in genotyping and sequencing technologies have allowed for unprecedented insight into the genetic structure of human populations. Population structure studies have largely focused on populations of European descent, and ancestry differences among European populations have been well studied and characterized [44, 42]. Recent studies have also investigated the genetic structure of more diverse populations, including recently admixed populations, such as African Americans [63, 3] and Hispanics [35], who have experienced admixing within the past few hundred years from two or more ancestral populations from different continents.

Both continental and fine-scale genetic structure of human populations have largely been characterized by aggregating measures of ancestry across the autosomal chromosomes. While it may be reasonable to assume that population structure patterns across the genome are similar for populations with ancestry derived from a single continent, such as populations of European descent, this may not be a reasonable assumption for recently admixed populations who have ancestries from multiple continents. For example, a previous analysis of Puerto Rican samples identified multiple chromosomal regions with large, systematic ancestry differences, as compared to what would be expected based on genome-wide ancestry, and thus providing evidence of recent selection in this admixed population [54]. Sex-specific patterns of non-random mating at the time of or since admixture can also result in systematic differences in ancestry at genomic loci as well as across entire chromosomes, such as the X and Y chromosomes, in admixed populations. For example, a recent study compared the average ancestry on the autosomes to the X chromosome in a large sample of Hispanics and African Americans [4] and highly significant differences in ancestry were detected, with

increased Native American and African ancestry, respectively, on the X chromosome in the Hispanic and African American samples, and a deficit of European ancestry as compared to the autosomes.

Previous methods [54, 20, 1] have been proposed to identify signals of selection by detecting genomic regions in admixed populations that exhibit unusually large deviations in ancestry proportions compared to what is expected based on genome-wide ancestry. For assessing significance, however, these methods require strong assumptions about the evolution of the admixed population of interest, which will generally be partially or completely unknown, including (1) the relative contribution from each of the ancestral populations to the gene pool at the time of the admixture events, (2) the number of generations since the admixture events, (3) an assumed effective population size, and (4) random mating. Significance is then assessed either analytically or through simulation studies based on these evolutionary assumptions. Misspecification of these assumptions, however, can result in false positives due to an incorrect null distribution, and regions of the genome that appear to have large ancestry differences are actually not significantly different from what would be expected when sampling variation, genetic drift after admixture, and potential bias in local ancestry estimation are appropriately taken into account [1].

Here, we consider the problem of detecting heterogeneity in ancestry across the genome in admixed populations. We propose the Chromosomal Ancestry Differences (CAnD) test for the identification of chromosomes that harbor genomic regions with significantly different proportional ancestry as compared to the rest of the genome. For each sampled individual, CAnD incorporates ancestry inferred at genomic regions using local ancestry methods, such as HAPMIX [48] or RFMix [37], and tests for systematic differences in genetic contributions to the chromosomes from the underlying ancestral populations. The CAnD method takes into account correlated ancestries among chromosomes within individuals for improved power, and the method can be used for the detection of ancestry differences among the autosomes, as well as between the autosomes and the X chromosome.

We perform simulation studies using real genotype data from Phase III of the HapMap Project [15] to evaluate the type I error rate and power of CAnD. We also apply CAnD to the HapMap Mexican Americans from Los Angeles, California (MXL) and African Americans

from Southwest U.S.A. (ASW) population samples for the detection of heterogeneity in population structure. In this analysis, RFMix is used to infer European, Native American, and African ancestry at genomic locations across the autosomes and the X chromosome using RFMix. Finally, we applied CAnD to a cohort of over 12,000 Hispanic/Latino individuals recruited from four metropolitan areas in the United States to identify heterogeneity in population structure across these admixed individuals. Again here we infer local European, Native American and African ancestry across the genome and test for heterogeneity genome-wide. In both the simulation studies and real data applications, we compare heterogeneity testing of ancestry between the autosomes and the X chromosome with CAnD to the t-test that does not account for the correlated ancestry among chromosomes of an admixed individual.

2.2 Methods

2.2.1 Chromosomal and Genome-wide Ancestry Measures

Let n be the number of unrelated individuals sampled from a population derived from K ancestral subpopulations. For individual i, $i \in \{1, \ldots, n\}$, we define the overall, or genome-wide, ancestry of i as measured across the autosomal and X chromosomes. (For males, ancestry on the Y chromosome could also be included when calculating genome-wide ancestry if this information is available). Quantitatively, we denote the genome-wide ancestry vector for individual i as $\mathbf{a}_i = (a_{i1}, \ldots, a_{iK})^T$, where a_{ik} is the proportion of ancestry from subpopulation k for individual i, $a_{ik} \geq 0$ for all k, and $\sum_{k=1}^{K} a_{ik} = 1$.

Consider the set $\mathcal{G}$ of autosomal and X chromosomes, i.e., $\mathcal{G} = \{1, \ldots, 22, X\}$. Denote the genetic ancestry for individual i on a particular chromosome $c \in \mathcal{G}$ as $\mathbf{a}_i^c = (a_{i1}^c, \ldots, a_{iK}^c)^T$. For each chromosome c, denote $\mathcal{G}_{-c} = \mathcal{G} \setminus \{c\}$ to be the set of all chromosomes excluding c, i.e., $\mathcal{G}_{-c} = \{1, 2, \ldots, c-1, c+1, \ldots, 22, X\}$, $\mathcal{G}_{-1} = \{2, \ldots, 22, X\}$ and $\mathcal{G}_{-X} = \{1, 2, \ldots, 22\}$. Define $a_{ik}^{-c} = \frac{1}{22} \sum_{M \in \mathcal{G}_{-c}} a_{ik}^M$ to be the mean of all chromosomal ancestries with chromosome c excluded for subpopulation k and individual i. Note for individual i, $a_{ik}^{-X} = \frac{1}{22} \sum_{M \in \mathcal{G}_{-X}} a_{ik}^M$ is the average autosomal ancestry for subpopulation k. We define $D_{ik}^c = a_{ik}^{-c} - a_{ik}^c$ to be the difference in ancestry between a given chromosome c

and the mean ancestry of all other other chromosomes in individual i for subpopulation k. We denote $\overline{D_k^c}$ to be the mean of the D_{ik}^c values across all individuals $i \in \{1, \ldots, n\}$.

2.2.2 The CAnD Test

Consider the previously defined set $\mathcal{G}$ consisting of the autosomal and X chromosomes. To test for heterogeneity in ancestry from subpopulation k among a subset $\mathcal{G}_s$ of $\mathcal{G}$, where $\mathcal{G}_s$ could also be $\mathcal{G}$ i.e., $\mathcal{G}_s \subseteq \mathcal{G}$, that contains m chromosomes, we first calculate a statistic T_k^c for each chromosome $c \in \mathcal{G}_s$ that is the mean of the standardized proportional ancestry differences for population k between c and the pooled average ancestry of all other chromosomes in $\mathcal{G}_s$ within each of the n sampled individuals, where

$$T_k^c = \frac{\overline{D_k^c}}{\sigma_{ck}}, \tag{2.1}$$

and σ_{ck} is the standard deviation of $\overline{D_k^c}$ (defined in the previous subsection). Under the null hypothesis of no ancestry differences among the m chromosomes, T_k^c approximately follows a normal distribution with mean 0 and variance 1 for each $c \in \mathcal{G}_s$, and the multivariate statistic

$$\mathbf{T}_k = \begin{pmatrix} T_k^1 \\ T_k^2 \\ \vdots \\ T_k^m \end{pmatrix} \sim \mathcal{N}(0, \Sigma), \tag{2.2}$$

where Σ is the $m \times m$ covariance matrix of T_k, allowing for correlation among the T_k^c statistics. To test for heterogeneity in ancestry from population k among the m chromosomes in $\mathcal{G}_s$, we propose the chromosomal ancestry differences (CAnD) test statistic

$$\mathcal{CA}_k = \widehat{\mathbf{T}}_k^T \widehat{\Sigma}^{-1} \widehat{\mathbf{T}}_k, \tag{2.3}$$

where $\widehat{\mathbf{T}}_k$ is $\mathbf{T}_k$ calculated with estimated $\widehat{\sigma}_{ck}$ for σ_{ck} for each chromosome c, and $\widehat{\Sigma}$ is an estimate of Σ. Under the null hypothesis, $\mathcal{CA}_k$ approximately follows a χ^2 distribution with $m-1$ degrees of freedom. Details about the estimators $\widehat{\Sigma}$ and $\widehat{\sigma}_{ck}$ for Σ and σ_{ck}, respectively, that we propose are given in Appendix A.

2.2.3 Detecting Evidence of Assortative Mating for Ancestry

Suppose we have a set of L known independent mate pairs in a sample where $p_i = (f_i, m_i)$ is mating pair i, $1 \leq i \leq L$, and f_i and m_i refer to the female and male member of the pair, respectively. Similar to the notation in the previous subsection, we define $\mathbf{F}_k^{-X} = (a_{f_1k}^{-X}, \ldots, a_{f_Lk}^{-X})^T$ and $\mathbf{M}_k^{-X} = (a_{m_1k}^{-X}, \ldots, a_{m_Lk}^{-X})^T$ to be the autosomal ancestry vectors in subpopulation k for the females and males respectively, in the set of L mating pairs. Likewise, the vectors of X chromosome ancestry in subpopulation k for the females and males in the L mating pairs are defined as $\mathbf{F}_k^X = (a_{f_1k}^X, \ldots, a_{f_Lk}^X)^T$ and $\mathbf{M}_k^X = (a_{m_1k}^X, \ldots, a_{m_Lk}^X)^T$. We denote

$$\rho_k^X = \frac{cov(\mathbf{F}_k^X, \mathbf{M}_k^X)}{SD(\mathbf{F}_k^X)SD(\mathbf{M}_k^X)} \tag{2.4}$$

to be the correlation of X chromosome ancestry of all of the mating pairs in subpopulation k, where *cov* is the covariance and SD indicates the standard deviation. Equation 2.4 calculates the correlation of ancestry between all observed mate pairs in the population. We define the correlation of autosomal ancestry between mating pairs in an analogous manner and signify this value by ρ_k^{-X}.

A permutation test was performed to create an empirical distribution of ρ_k^X under the null hypothesis of no X chromosome ancestry correlation between mating pairs. We randomly shuffled the members across pairs p_i, $1 \leq i \leq L$, respecting the characteristic that one male and one female constitute a pair, and recalculated the correlation among the permuted pairs to approximate the distribution of ρ_k^X under the null hypothesis. We repeat this process B times. The null hypothesis of no correlation between mating pairs corresponds to a null hypothesis of random mating, and to test for ancestry-related assortative mating, we examine whether ρ_k^X is larger than expected from our empirical distribution. To test for non-random mating, i.e. assortative or disassortative mating for ancestry, we examine whether $|\rho_k^X|$ is larger than expected.

2.2.4 Simulation Studies

In order to assess type I error and power of the CAnD method, we performed simulation studies using real data from the HapMap CEU (Utah residents with ancestry from northern

and western Europe from the Centre dÉtude du Polymorphisme Humain collection) and YRI (Yoruba in Ibadan, Nigeria) populations. Each simulated replicate consisted of simulated chromosomes for 50 admixed individuals that were derived from 118 CEU and YRI haplotypes on chromosomes 1 and 2, where the chromosomal haplotypes consisted of 5,000 evenly spaced markers [15] across the chromosome.

Each simulated admixed individual $i \in \{1, \ldots, 50\}$ has admixture vectors for chromosomes 1 and 2 of the form $\mathbf{a}_i^1 = (a_{i1}^1, a_{i2}^1)^T$ and $\mathbf{a}_i^2 = (a_{i1}^2, a_{i2}^2)^T$, respectively, where a_{i1}^1 and a_{i1}^2 are the population 1 ancestry proportions on chromosomes 1 and 2, respectively, and $a_{i1}^j + a_{i2}^j = 1$ for $j = 1, 2$. We denote CEU and YRI to be populations 1 and 2, respectively, in the simulation study, and proportional CEU ancestry on chromosome 1 for individual i is $a_{i1}^1 = \alpha_{i1} + \epsilon_{i1}^1$, where α_{i1}^1 is drawn from uniform distribution on [0.05,0.45] and ϵ_{i1}^1 is drawn from a $N(0, 8.2\text{e}{-04})$ distribution. The variance of ϵ_{i1}^1 corresponds to an estimate of the average variance across the autosomal chromosomes for European ancestry within admixed individuals from the HapMap MXL. For chromosome 2, $a_{i1}^2 = \alpha_{i1} + \epsilon_{i1}^2$, where ϵ_{i1}^2 is a random ancestry effect for chromosome 2 that follows a $N(\mu, 8.2\text{e}{-04})$ distribution, where $0 \leq |\mu| \leq 1$. Under the null hypothesis, $\mu = 0$, i.e., there is no difference in mean ancestry between chromosomes 1 and 2, and $|\mu| > 0$ under the alternative hypothesis. Each chromosome 1 for individual i is constructed from the CEU and YRI haplotypes, where the chromosome has proportions a_{i1}^1 and $1 - a_{i1}^1$, respectively, from a randomly drawn CEU haplotype and a randomly drawn YRI haplotype. The two copies of chromosome 2 for individual i are similarly obtained.

For each simulated individual, chromosomal-wide ancestry proportions were estimated from the genotype data using the FRAPPE software program [55], which uses a likelihood-based model to infer each individual's ancestry proportions. Included as reference samples in the FRAPPE runs were 58 CEU and 57 YRI HapMap samples, and the number of reference populations was set to two. The reference samples used for the FRAPPE analyses were different from those used to simulate the admixed individuals genotypes. With the resulting FRAPPE proportions, we implemented the CAnD method to identify heterogeneity in population structure across two chromosomes. A variety of μ values were considered for the assessment of type I error and power at different significance levels.

2.2.5 HapMap MXL and ASW

We considered detection of heterogeneity in ancestry across the genome in unrelated HapMap MXL and ASW samples. REAP [56] was used to infer both known and cryptic relatedness in the MXL and ASW, and a subset of 53 MXL individuals and a subset of 45 ASW individuals with inferred relationships less than third degree were identified and included for the ancestry heterogeneity analysis. Of the unrelated subset of 53 MXL individuals, there were eight singletons, 20 families with two individuals included and one family with three individuals. Among the 45 unrelated ASW individuals, there were 23 singletons and 11 families with two individuals that were included. There were 27 females and 26 males in the unrelated HapMap MXL subset, and 25 females and 20 males in the unrelated HapMap ASW subset. We also performed CAnD tests stratified by sex to determine if there was any bias in the results due to copy number differences in the X chromosome for males and females.

We used the RFMix software [37] to estimate local ancestry across the autosomes and the X for all HapMap MXL and ASW samples. RFMix allows for more than two ancestral subpopulations and in both the HapMap MXL and ASW analyses, and we assumed ancestral contributions from African, European and Native American populations. The HapMap CEU and YRI samples were included as the reference population panels in the local ancestry analysis for European and African ancestry, respectively, and the Human Genome Diversity Project (HGDP) [5] samples from the Americas were included as the reference population panel for Native American ancestry. All samples were phased and sporadic missing genotypes were imputed using the BEAGLE v.3 software [2]. Recombination maps for each chromosome were downloaded from the HapMap website [15] and were converted to Human Genome Build 36. There was no phasing conducted on the X for males since a male only has one X chromosome. Only SNPs that were genotyped in both the HapMap and HGDP datasets were considered in the local ancestry analysis. For local ancestry on the X chromosome, SNPs on the non-pseudoautosomal regions, where there is no homology between the X and Y chromosomes, were considered.

We compared CAnD when using global ancestry for each chromosome estimated using

the FRAPPE software [55] to CAnD when using local ancestry estimated across the chromosomes with RFMix. For each chromosome, a supervised global ancestry analysis was conducted separately for the HapMap MXL and ASW population samples with FRAPPE. The number of ancestral populations was set to three and the same reference population samples used in the RFMix local ancestry analysis were also used with FRAPPE. Since males only have one allele at each of the X chromosome SNPs, one of the alleles at an X-linked SNP was coded to be missing in the FRAPPE analysis. We realize coding one allele as missing is artificial since the missing allele in males actually does not exist. FRAPPE is not designed to handle X chromosome genetic data and by coding one allele missing, we are able to obtain ancestry estimates that correspond to the observed data where there is only one allele for males at each SNP on the X. We found that coding male genotypes as homozygous in the FRAPPE analysis yielded nearly identical results to coding one of the alleles to be missing.

2.2.6 Characterization of Admixed Ancestry in the HCHS/SOL Dataset

The HCHS/SOL study [53] provides a unique opportunity to evaluate heterogeneity in ancestry across the genome in a large-scale study of Hispanic/Latino individuals. PC-Relate [11] was used to infer both known and cryptic relatedness in all samples, and a subset of 10,642 unrelated samples with inferred relationships less than fourth degree were identified and included in the ancestry heterogeneity analysis. PC-Relate is able to accurately infer relatedness in a set of samples in the presence of population structure. We used the RFMix [37] software to estimate local ancestry across the autosomes and the X chromosome for all 10,642 samples, assuming ancestral contributions from African, European and Native American populations. The reference population panels and subsequent steps are as described in the previous subsection for the HapMap analysis. We performed the CAnD method stratified by six self-identified background groups: Cuban, Dominican, Puerto Rican, Mexican, Central American, South American or Other. We grouped individuals with unknown background group with Other. We applied CAnD to the HCHS/SOL individuals using chromosomal-wide averaged local ancestry estimated with RFMix.

Table 2.1: CAnD Empirical Type I Error (95% CI) at significance levels $\alpha = 0.01$, 0.005, and 0.001 based on 5,000 simulated replicates. This simulation setting was conducted under the null hypothesis where the randomly drawn ancestry proportions of an admixed individual are the same for both chromosomes 1 and 2.

α	CAnD Empirical Type I Error
0.01	0.0118 (0.009, 0.015)
0.005	0.0053 (0.003, 0.007)
0.001	0.0004 (0, 0.0013)

2.3 Results

2.3.1 Assessment of Type I Error

In the simulation studies for detecting ancestry heterogeneity, FRAPPE was first used to estimate proportional ancestry on chromosomes 1 and 2 for each simulated admixed individual. To ensure that the FRAPPE estimates were accurate when using unphased genotypes from 5,000 SNPs on a chromosome, we first compared the FRAPPE ancestry estimates to the simulated ancestry. The differences between the FRAPPE estimates and the simulated ancestry proportion values have mean of -5.147e-06 (SD=0.018), indicating FRAPPE can accurately estimate chromosomal ancestry proportions when using a set of 5,000 markers (Figure 2.1).

To assess the type I error rate of CAnD, we simulated admixed chromosomes for 50 sampled individuals under the null hypothesis of no ancestry differences among the chromosomes, on average. The empirical type I error rates for the CAnD test at the $\alpha = 0.01$, 0.005, and 0.001 significance levels calculated using 5,000 simulated replicates are given in Table 2.1. The CAnD test is properly calibrated for all significance levels considered. Empirical type I error rates are not significantly different from the nominal levels, as can be seen from the 95% confidence intervals given in the table.

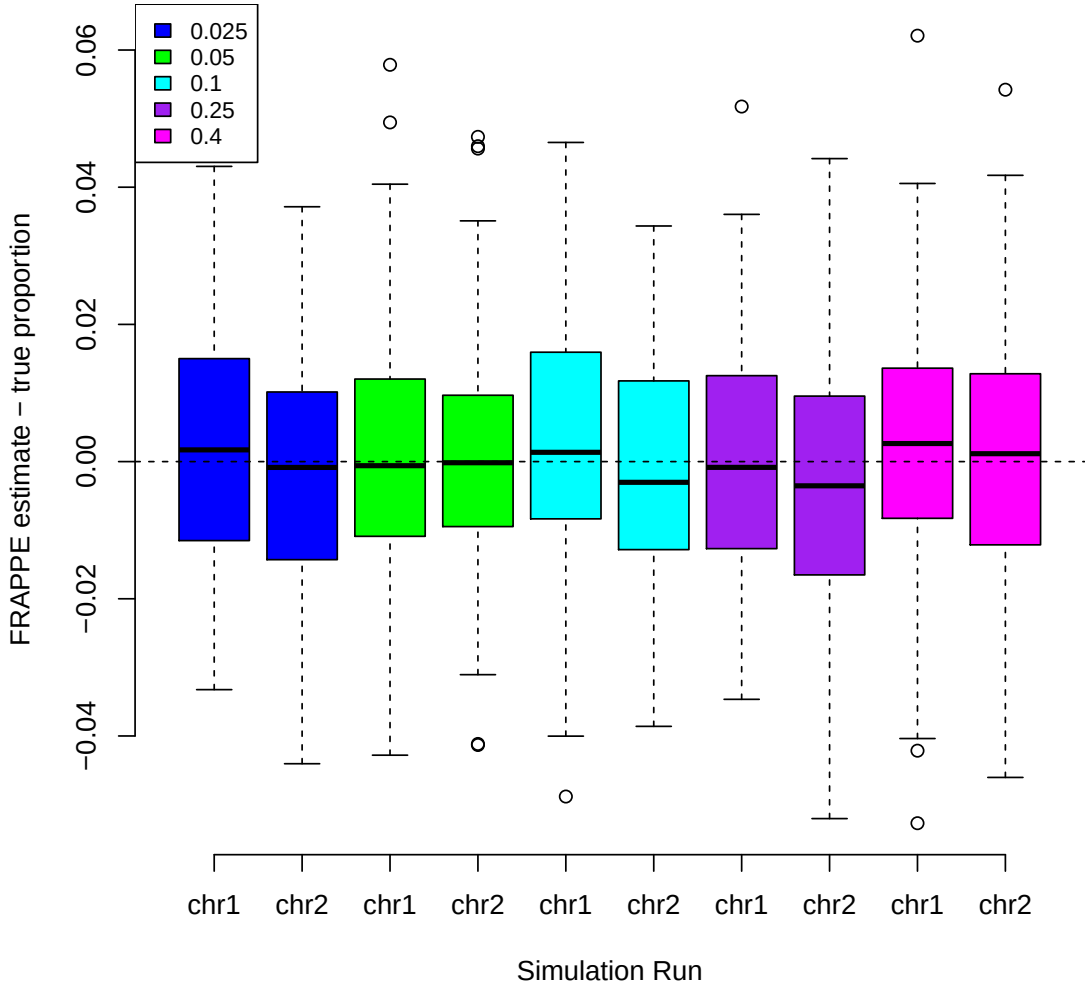


Figure 2.1: Boxplots of the difference between the simulated metrics and the FRAPPE estimates, for increasing values of expected ancestry differences, shown in differing colors, among 50 sample batches.

2.3.2 Power Evaluation and Comparison

We evaluated the power of the CAnD method for an admixed sample of 50 individuals. The values of μ, the mean difference in ancestry between chromosomes 1 and 2, ranged from 0.005 and 0.25. We also compared the power of CAnD to a pooled t-test that ignores the correlation of ancestry across chromosomes within an individual. Although ancestry across chromosomes is not independent within an individual, we present this method for comparison to CAnD, as it has been used in previous studies for the testing of ancestry differences between the autosomal chromosomes and the X chromosome in admixed populations [4].

Empirical power results at the $\alpha = 0.01$ significance level using the CAnD and pooled t-test are given in Figure 2.2. CAnD has higher power than the pooled t-test for all values of μ considered, and significantly higher values for low to moderate values of μ. For example, there is essentially no power to detect a mean difference in ancestry of 5% between the two chromosomes with the pooled t-test, while CAnD has power that is close to 1. The loss in power with the pooled t-test is due to the test not accounting for the correlation in ancestry between chromosomes within an individual. We recommend the CAnD test over the pooled t-test for improved power to detect ancestry differences among chromosomes.

2.3.3 HapMap ASW Ancestry

The predominant genome-wide ancestry in all 87 HapMap ASW subjects is African. Table 2.2 shows the mean and SD of the local ancestry estimates by chromosome in each of the ancestral populations and accompanying Figure 2.3A shows violin plots of the local ancestry results by chromosome. RFMix estimated 11 individuals to have no European ancestry on the X chromosome, and the maximum European ancestry on the X chromosome is 0.67. On the other hand, nine individuals are estimated to have an X chromosome entirely of African ancestry, where the proportion ranges from 0.33 to 1. We see these patterns in the barplots shown in Figure 2.4A which displays the proportional ancestry for each sample. The samples are ordered separately in increasing order of European ancestry. For comparison, Figure **??** shows the same estimates but only ordered by European ancestry estimated on the autosomes. Thus, we can compare the ancestry within an individual as estimated

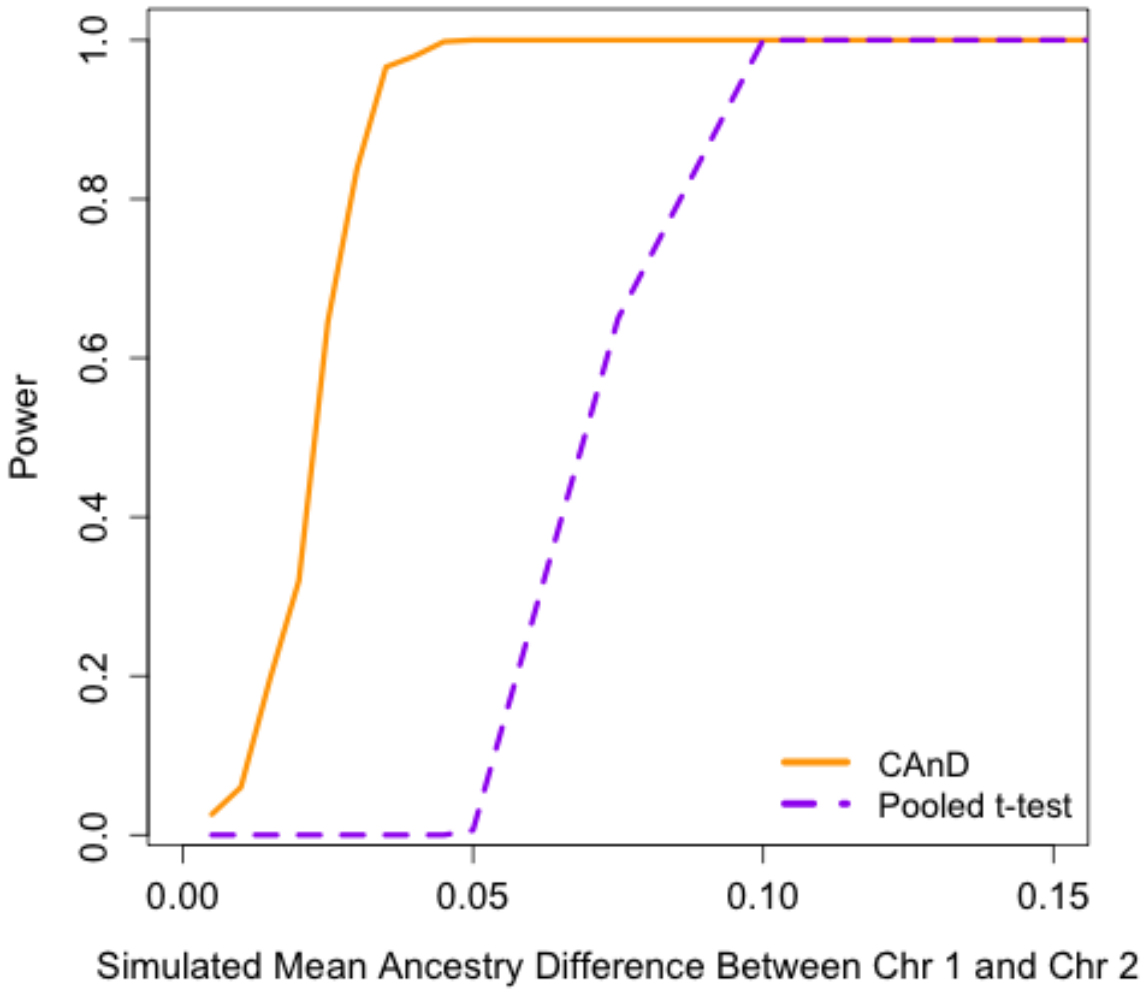


Figure 2.2: The proportion of tests rejected at a significance level of 0.01 when using the CAnD method as compared to the pooled t-test under increasing differences in ancestry proportion between chromosomes. For each simulated ancestry proportion difference, the proportion of tests rejected was calculated from 500 independent simulations of 50 samples each.

on the X chromosome as compared to the autosomes. Across both the autosomes and the X chromosome, the proportion of Native American ancestry is quite small over all samples. Fifty-seven individuals are estimated to have no Native American ancestry on the X chromosome. While the proportion of Native American ancestry is larger on the autosomes than the X chromosome, on average, it remains small in magnitude and we conclude that Native American ancestry is negligible in this sample of individuals. Furthermore, we detect more African and less European ancestry on the X chromosome than the autosomes, overall.

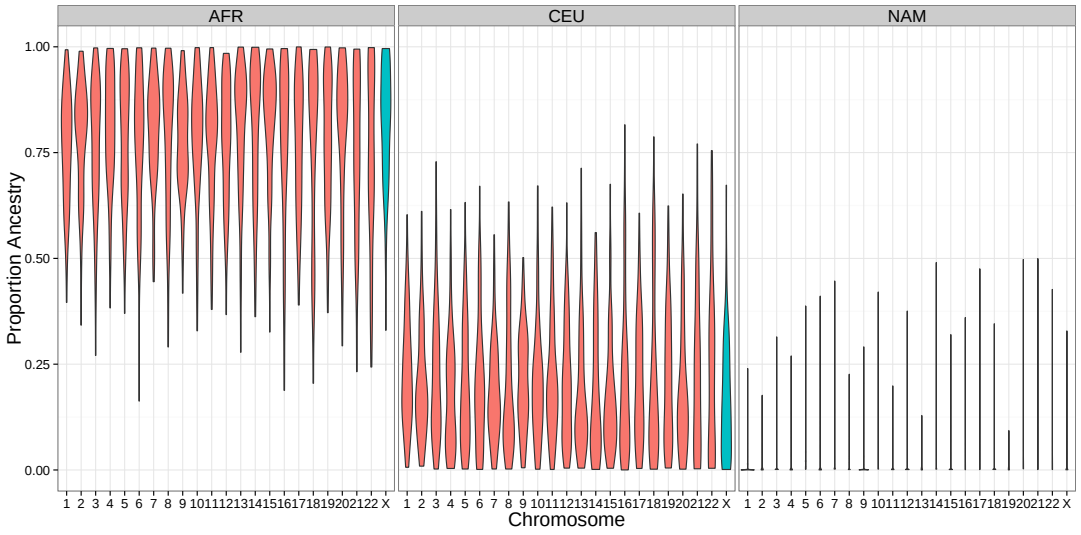
We calculated the correlation of ancestry proportions across the autosomes and X chromosome for each ancestral subpopulation. Correlation between the autosomal and X chromosome Native American ancestry is highest at 0.78. The European and African correlations between autosomal and X chromosome proportions are 0.20 and 0.17, respectively.

2.3.4 HapMap MXL Ancestry

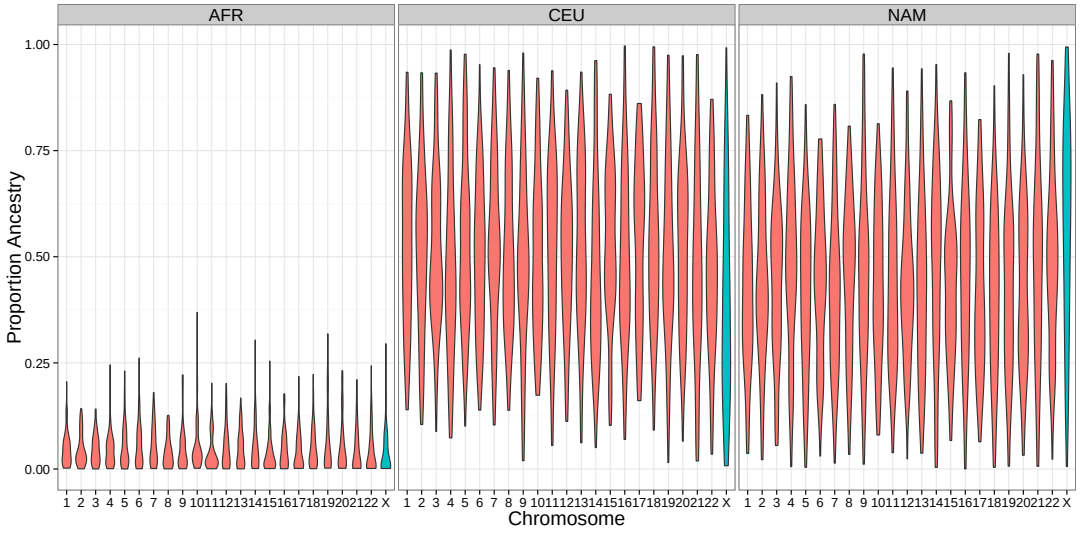
From our local ancestry analysis of the 86 HapMap MXL individuals, we found the predominant ancestries to be European and Native American, as expected based on previously reported results [4, 56], with African ancestry being quite modest with little variation. Table 2.2 shows the mean and SD of the average local ancestry estimates by chromosome and averaged across the autosomes within the MXL samples. Interestingly, Native American ancestry is highest on the X chromosomes, with a mean of 57.4% (SD=24.8%), while for the autosomes, European ancestry is highest with a mean of 50.8% (SD=14.9%). African ancestry on the autosomes and the X chromosome, however, are quite similar, with mean values of 4% and 5%, respectively. Figure 2.3B shows violin plots by chromosome of the RFMix local ancestry estimates in the MXL samples. The plots illustrate the marked increase in proportional European ancestry across the autosomes, and, correspondingly, a decrease in proportion of Native American ancestry on the autosomes as compared to the X chromosome. Estimates of ancestry on chromosome 21 and 22 are less variable than estimates across other chromosomes. Figure 2.4B shows barplots of the ancestral proportions within each individual. The proportion of both European and Native American ancestries on the X chromosome ranges from 0 to 1. The range and variation of the European and

Table 2.2: Mean (SD) of local ancestry estimates by chromosome, stratified by the ASW and MXL HapMap population samples.

	ASW			MXL		
Chr	African	European	Native American	African	European	Native American
X	0.82 (0.139)	0.163 (0.136)	0.017 (0.047)	0.0396 (0.0521)	0.387 (0.245)	0.574 (0.248)
Autosomal	0.783 (0.0861)	0.202 (0.0808)	0.0150 (0.0382)	0.0489 (0.0182)	0.508 (0.149)	0.444 (0.148)
1	0.762 (0.13)	0.228 (0.131)	0.00962 (0.0354)	0.047 (0.0389)	0.525 (0.192)	0.428 (0.191)
2	0.789 (0.132)	0.201 (0.128)	0.0106 (0.0241)	0.0457 (0.0379)	0.514 (0.195)	0.44 (0.188)
3	0.769 (0.155)	0.221 (0.154)	0.0102 (0.0369)	0.0462 (0.0345)	0.514 (0.18)	0.439 (0.183)
4	0.807 (0.136)	0.177 (0.134)	0.0164 (0.0398)	0.0461 (0.0408)	0.47 (0.212)	0.484 (0.206)
5	0.786 (0.149)	0.199 (0.146)	0.0148 (0.0536)	0.0539 (0.05)	0.528 (0.2)	0.418 (0.188)
6	0.774 (0.167)	0.201 (0.15)	0.0257 (0.0696)	0.0555 (0.0502)	0.5 (0.179)	0.445 (0.177)
7	0.804 (0.125)	0.184 (0.117)	0.012 (0.0539)	0.056 (0.047)	0.524 (0.193)	0.42 (0.188)
8	0.785 (0.163)	0.201 (0.16)	0.0141 (0.0419)	0.0397 (0.0349)	0.504 (0.187)	0.456 (0.179)
9	0.772 (0.12)	0.21 (0.116)	0.0175 (0.0506)	0.0499 (0.0521)	0.489 (0.21)	0.462 (0.213)
10	0.785 (0.145)	0.205 (0.14)	0.00997 (0.047)	0.059 (0.066)	0.502 (0.189)	0.439 (0.183)
11	0.778 (0.141)	0.212 (0.139)	0.00953 (0.0255)	0.0402 (0.0422)	0.525 (0.202)	0.435 (0.201)
12	0.779 (0.142)	0.202 (0.137)	0.0186 (0.0625)	0.0501 (0.0488)	0.511 (0.18)	0.439 (0.177)
13	0.804 (0.152)	0.18 (0.149)	0.0165 (0.032)	0.0488 (0.0424)	0.523 (0.199)	0.428 (0.201)
14	0.802 (0.162)	0.183 (0.155)	0.015 (0.0577)	0.0559 (0.0643)	0.47 (0.217)	0.474 (0.219)
15	0.817 (0.141)	0.172 (0.138)	0.0109 (0.0399)	0.0382 (0.0452)	0.528 (0.182)	0.434 (0.179)
16	0.778 (0.192)	0.201 (0.183)	0.0207 (0.0631)	0.0456 (0.0449)	0.498 (0.205)	0.457 (0.209)
17	0.772 (0.156)	0.207 (0.145)	0.0208 (0.079)	0.041 (0.043)	0.533 (0.195)	0.426 (0.192)
18	0.772 (0.21)	0.21 (0.196)	0.0184 (0.0605)	0.0501 (0.047)	0.537 (0.207)	0.413 (0.198)
19	0.78 (0.154)	0.213 (0.155)	0.00745 (0.0189)	0.0654 (0.0809)	0.506 (0.208)	0.429 (0.202)
20	0.801 (0.167)	0.187 (0.152)	0.0125 (0.0611)	0.0541 (0.0616)	0.52 (0.194)	0.426 (0.195)
21	0.76 (0.191)	0.22 (0.186)	0.0202 (0.0748)	0.0451 (0.0529)	0.475 (0.235)	0.48 (0.232)
22	0.747 (0.209)	0.234 (0.206)	0.0188 (0.0671)	0.0419 (0.0506)	0.47 (0.196)	0.488 (0.204)



(a) ASW Samples



(b) MXL Samples

Figure 2.3: Chromosomal averaged local ancestry estimates for HapMap individuals using the RFMix software. Ancestry was estimated for each marker then averaged across chromosomes. (A): Estimates for 87 HapMap ASW individuals. (B): Estimates for 86 HapMap MXL individuals. The reference samples for the European and African ancestries were HapMap CEU and YRI individuals, while the HGDP samples from the Americas were references for the Native American ancestry.

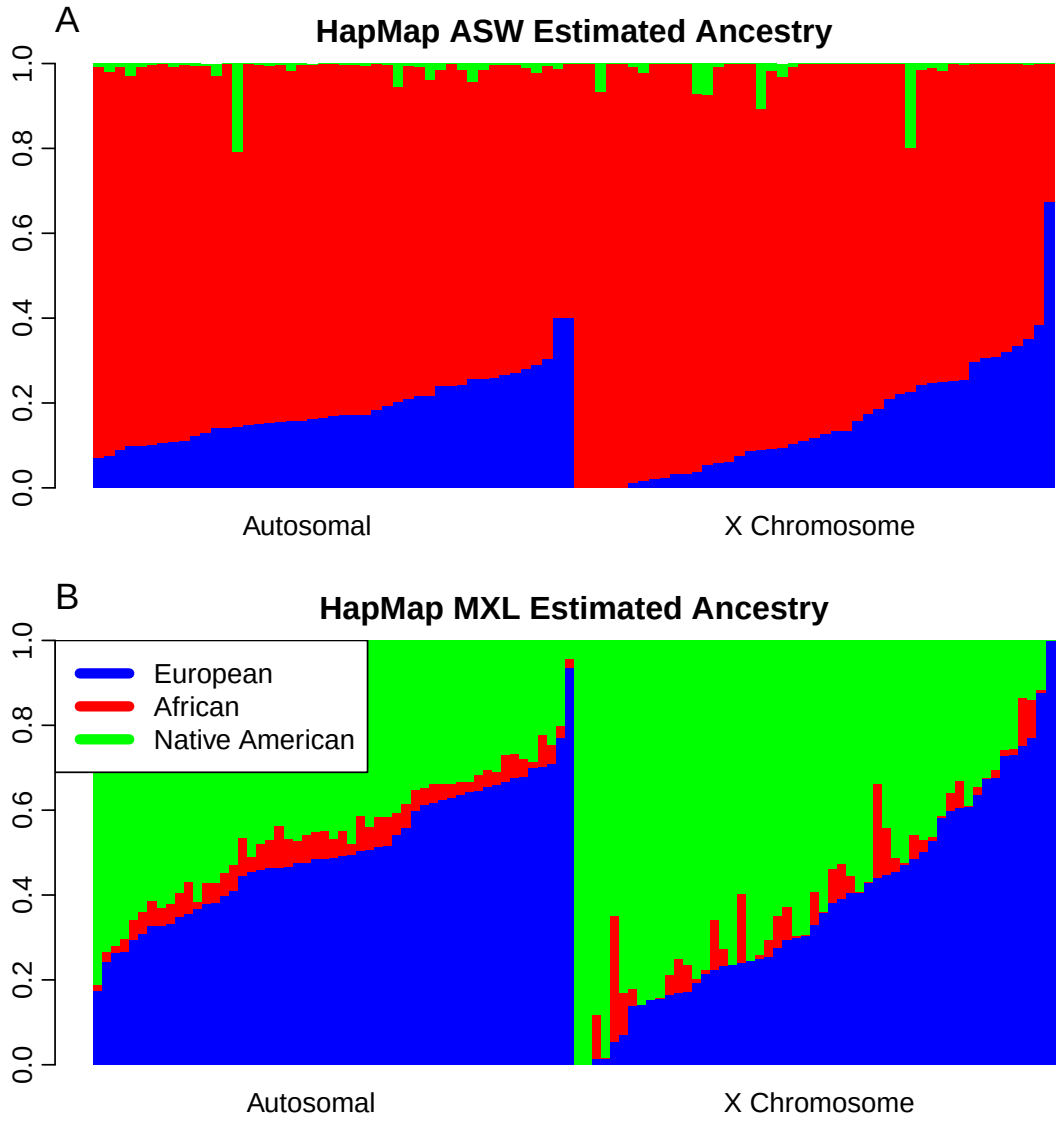


Figure 2.4: Local ancestry estimates for HapMap individuals using the RFMix software. Each individual is represented by a vertical bar, where the European, African and Native American ancestries are colored with blue, red, and green, respectively. The two panels represent the autosomal and X chromosome average. (A): Estimates for 87 HapMap ASW individuals. (B): Estimates for 86 HapMap MXL individuals. The reference samples for the European and African ancestries were HapMap CEU and YRI individuals, while the HGDP samples from the Americas were references for the Native American ancestry.

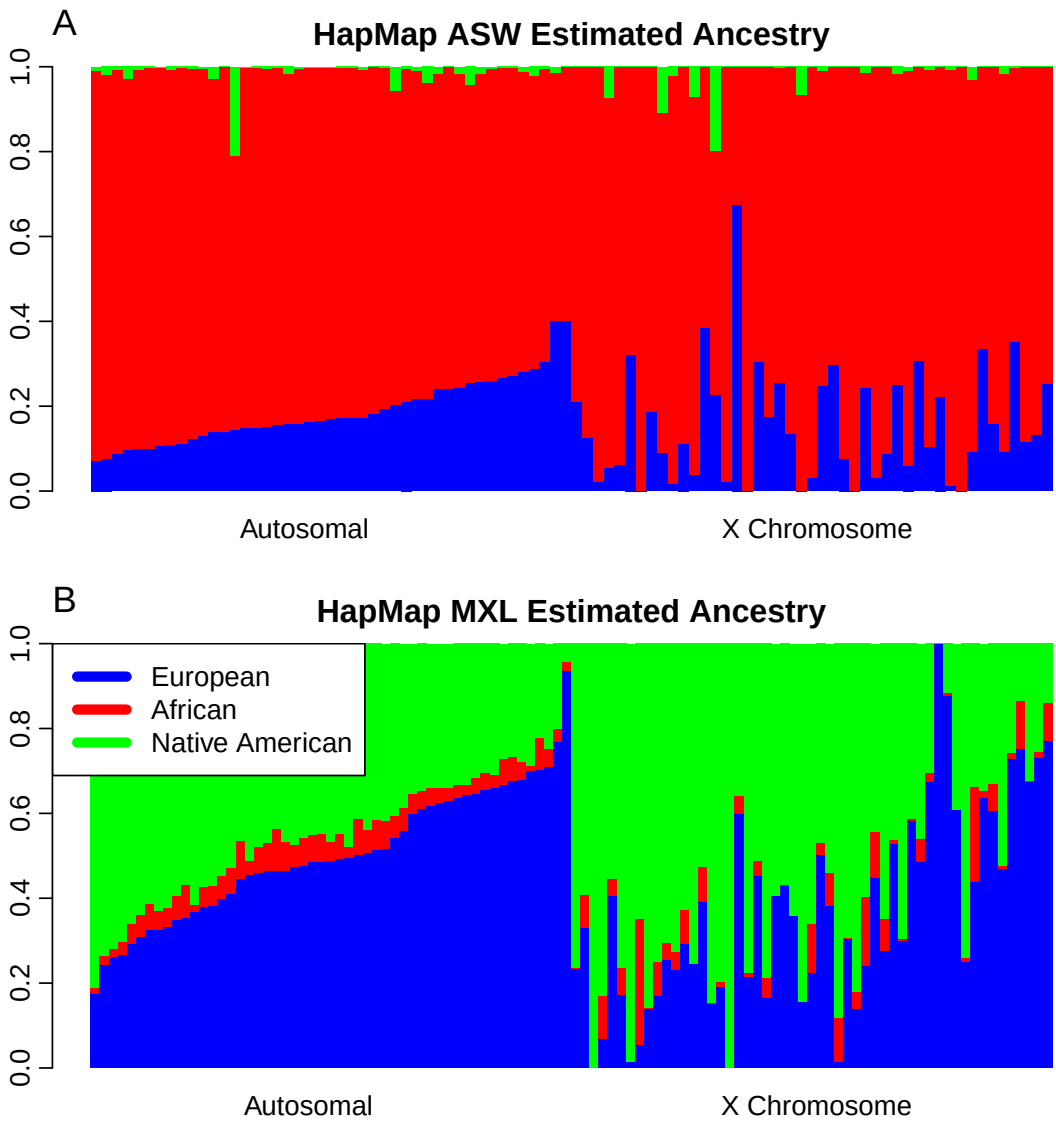


Figure 2.5: Local ancestry estimates for HapMap individuals using the RFMix software. Each individual is represented by a vertical bar, where the European, African and Native American ancestries are colored with blue, red, and green, respectively. The two panels represent the autosomal and X chromosome average, where the bars are ordered the same to compare each individual's autosomal and X chromosome ancestries. (A): Estimates for 87 HapMap ASW individuals. (B): Estimates for 86 HapMap MXL individuals. The reference samples for the European and African ancestries were HapMap CEU and YRI individuals, while the HGDP samples from the Americas were references for the Native American ancestry.

Native American ancestries on the X chromosome are larger than those estimated across the autosomes. Furthermore, Native American and European ancestries on the X chromosome are almost perfectly negatively correlated (corr=-0.98).

We also calculated correlation in autosomal and X chromosome ancestries. The correlation between the autosomal and X chromosome European ancestry is 0.71 and is the highest, and the Native American correlation is 0.67. We expected that the X chromosome and autosomal ancestry correlations would be similar between the European and Native American ancestries. This is because the correlation between the European and Native American ancestries is -0.98. With a correlation of 0.03, there is essentially no African ancestry correlation between the autosomes and the X chromosome, which likely is attributed to the small contribution of African ancestry to the HapMap MXL.

There is one male MXL individual who has an X chromosome that inferred to be completely Native American derived. The phased RFMix results of this individual's mother indicates that one of her X chromosomes is entirely Native American while her other X chromosome is 69% Native American and 31% European, with five ancestry switches on the chromosome. This male individual has genome-wide averages of 39% European ancestry, 4% African and 57% Native American.

2.3.5 Ancestry Heterogeneity Testing in HapMap MXL and ASW

Figure 2.6 shows histograms of the mean difference between the autosomal and X chromosome ancestry proportions for the sets of 45 unrelated ASW (Figure 2.6A) and 53 unrelated MXL (Figure 2.6B) individuals, with a smoothed density line overlaid. The mean difference in European ancestry between the autosomes and the X chromosome is 0.12, and the mean difference for Native American ancestry is -0.13. Based on our simulation studies, we expect to have high power to detect such large differences in ancestry between the autosomes and the X chromosome for a sample of this size. For the ASW samples, however, the mean difference between the X chromosome and the autosomes for the two predominant ancestries, African and European, is 0.04, which is much smaller than the predominant ancestry differences observed in the MXL. We expect the power to detect a mean difference in ancestry

between the X and the autosomes in the ASW to be much lower, as compared to the MXL, due to both smaller mean ancestry differences and smaller sample size.

We applied the CAnD test to a set of 53 unrelated MXL samples. The genome-wide combined CAnD p-values are 0.592, 4.01e-05 and 9.57e-06 for the African, European and Native American ancestries, respectively. To understand which chromosomes are driving the significance found in the European and Native American ancestries, Figure 2.7 shows, by chromosome, the unadjusted (Figure 2.7A) and Bonferroni-adjusted (Figure 2.7B) p-values from the CAnD test in the HapMap MXL for the three ancestries. Chromosome 7 and the X chromosome have a larger proportion of Native American ancestry as compared to the mean Native American ancestry of all other chromosomes pooled together, before adjustment for multiple testing. The same result holds for the X chromosome when considering European ancestry. Chromosome 8 has a larger proportion of African ancestry than a pool of all other chromosomes. After the Bonferroni multiple testing correction, the X chromosome remains significant in the European and Native American ancestries. No other chromosomes obtain statistical significance after correction for multiple testing. Ancestry as estimated from the X chromosome is statistically significantly different from the ancestry estimates across any and all of the autosomes.

CAnD applied to the set of 45 unrelated ASW samples yielded no significant results with genome-wide combined p-values of 0.122, 0.0858, 0.243 for the African, European and Native American ancestries, respectively (Figure 2.8). As previously mentioned, the autosomes and the X chromosome are predominantly African derived in the ASW, and a larger sample size is needed to achieve enough power to detect the smaller ancestry differences among chromosomes in the ASW. Indeed, in much larger population-based samples of African Americans [4, 3], increased African ancestry and decreased European ancestry has been reported for the X chromosome as compared to the autosomes.

To assess whether inclusion of the X chromosome biased the CAnD results for the autosomal chromosomes within the HapMap MXL individuals, we performed the analysis using only the autosomes. When excluding the X chromosome, African ancestry on chromosome 8 remains significant and Native American ancestry on chromosomes 4 and 22 are significant at a 0.05 threshold (Figure 2.9), similar to the results when the X chromosome is included

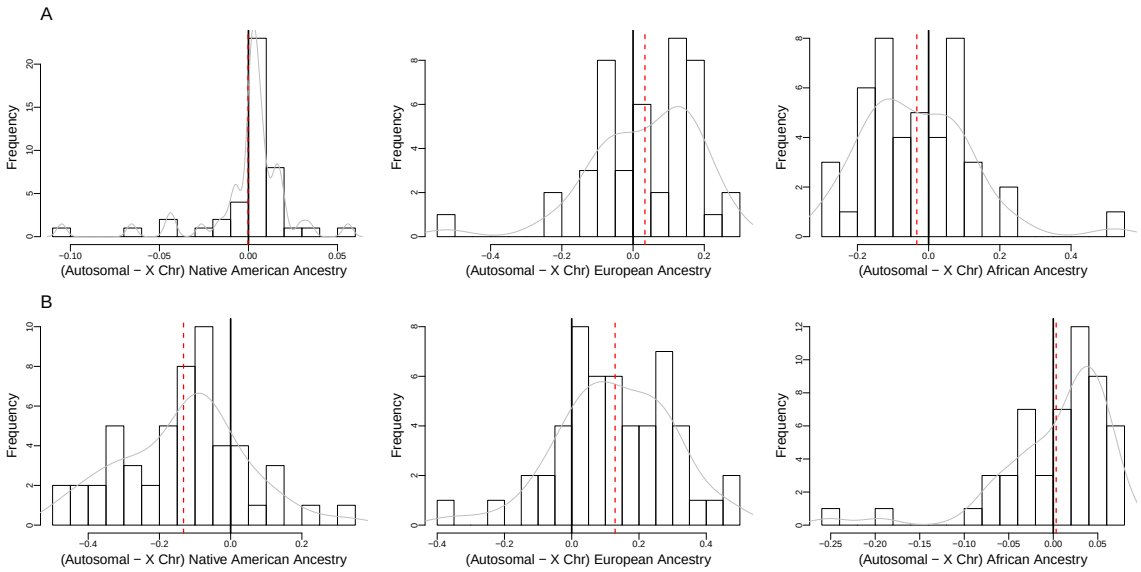


Figure 2.6: Histograms of the difference in autosomal and X chromosome ancestry proportions among the (A): 45 unrelated HapMap ASW and (B): 53 unrelated HapMap MXL samples. The dashed line indicates the mean difference, whereas the solid line indicates zero. A smoothed density line is overlaid on each histogram.

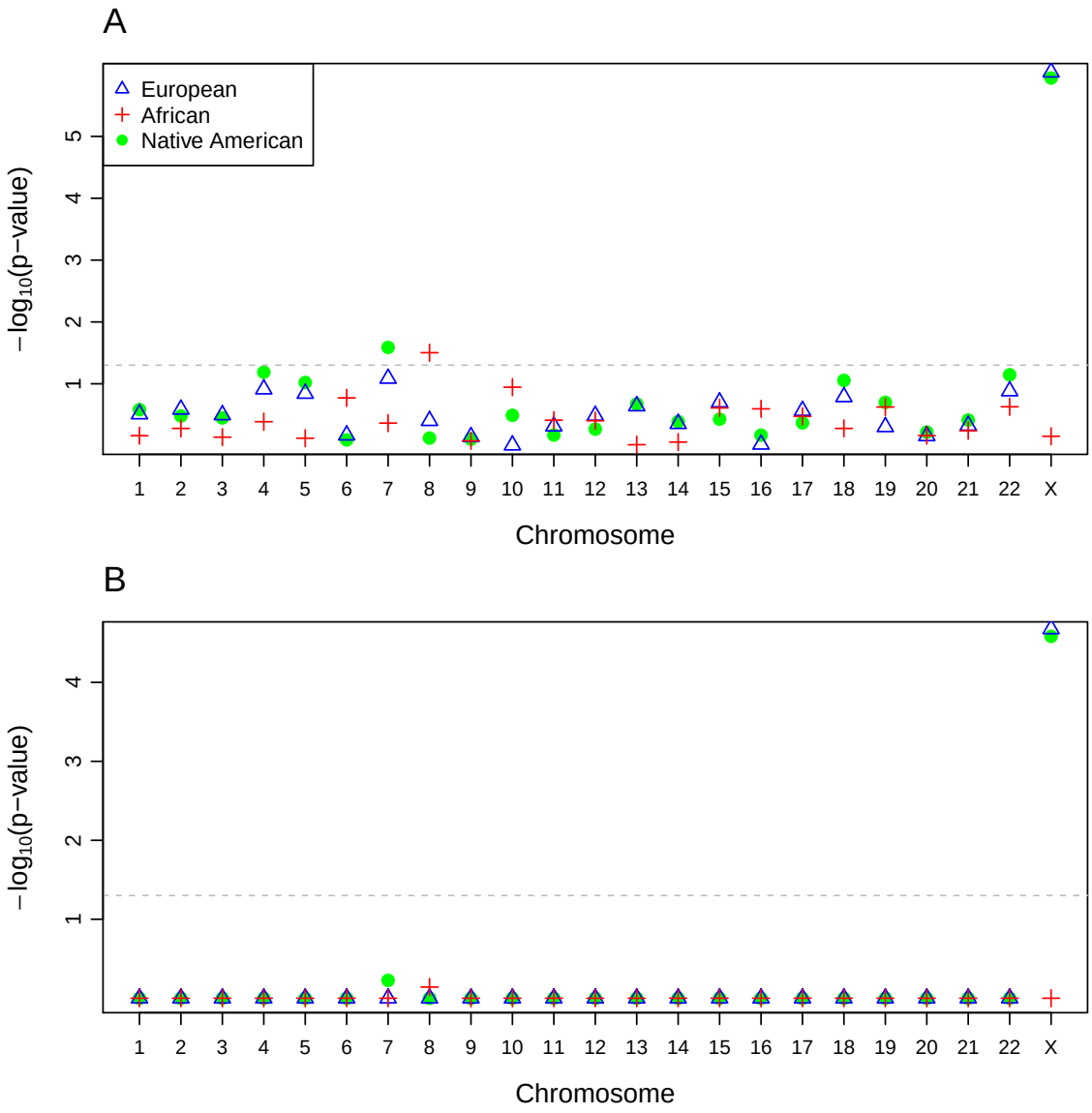


Figure 2.7: (A): Unadjusted and (B): adjusted p-values by chromosome obtained from the parametric CAnD test comparing the estimated ancestry for each chromosome with the mean ancestry of all remaining chromosomes, including the X chromosome, for the African, European and Native American ancestries in the HapMap MXL samples. The adjusted p-values were calculated using the Bonferroni multiple testing correction. The dotted line indicates the 0.05 significance level.

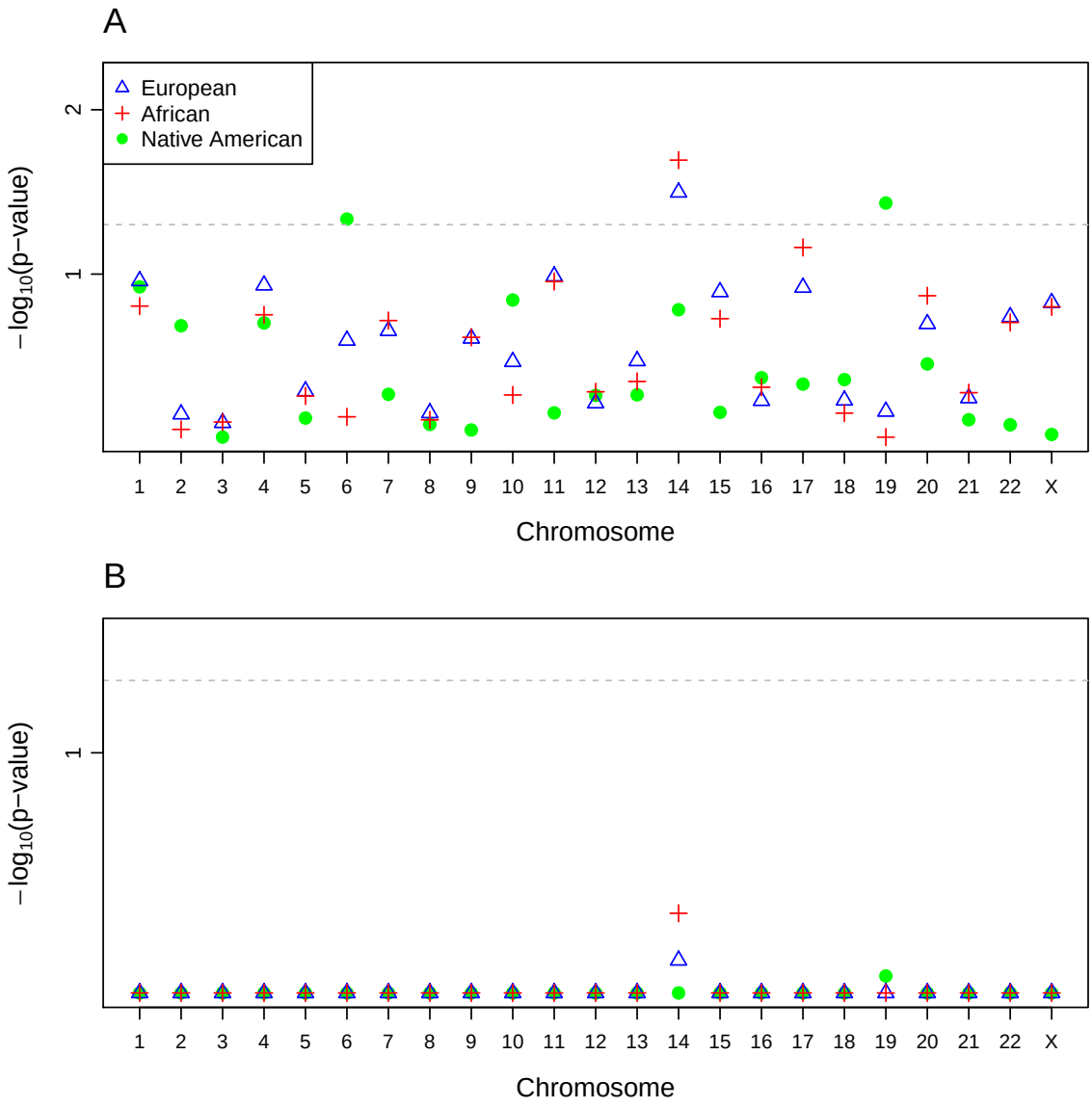


Figure 2.8: P-values from the parametric CAnD test on 45 unrelated HapMap ASW samples, comparing the estimated ancestry for each chromosome with the mean ancestry of all remaining chromosomes, including the X chromosome. (A): Uncorrected for multiple testing. (B): Bonferroni-adjusted. The dotted line indicates the 0.05 significance level.

in the CAnD analysis. After correction for multiple testing using the Bonferroni procedure, however, no estimates remain significant, indicating that the significance in heterogeneity detected with CAnD is being driven by ancestry differences between the X chromosome as compared to the autosomes.

Previous studies have identified a significant difference between autosomal and X chromosome ancestry proportions in individuals from admixed populations [4], where this difference has been assessed using a pooled t-test that ignores the correlation of ancestries among the autosomes and the X chromosome within an individual. We compare the performance of the pooled t-test to the CAnD test for detecting differences in ancestry between the X chromosome and the autosomes in the HapMap MXL samples. The pooled t-test finds significant differences in European ancestry and Native American ancestry between the autosomes and the X chromosome, with a p-value of 0.001 for both analyses. In comparison, the CAnD p-values comparing mean European ancestry and Native American ancestry on the X chromosome are 9.17e-07 and 1.13e-06, respectively, which is more than three orders of magnitude smaller than the pooled t-test. No significant differences in African ancestry were found using either method.

2.3.6 Comparison of CAnD Results Using Local Versus Global Ancestry Estimates

We performed a CAnD analysis in the HapMap MXL and ASW using ancestry estimates for each chromosome with FRAPPE that uses unphased genotype data and assumes independent markers. We compare these results to the CAnD results reported in the previous subsection that used local ancestry estimates from RFMix, which takes into account LD among SNPs and requires phased genotype data. With the FRAPPE estimates for the ASW, no chromosomal ancestry differences were detected with CAnD, similar to the CAnD analysis results with local ancestry estimates from RFMix. Interestingly, we found that the CAnD results are slightly more significant for the MXL when using ancestry estimates from FRAPPE as compared to the estimates from RFMix, particularly for detecting differences in European ancestry across the genome (Figure 2.10). Inference on population structure heterogeneity in the HapMap ASW and MXL, however, is qualitatively the same

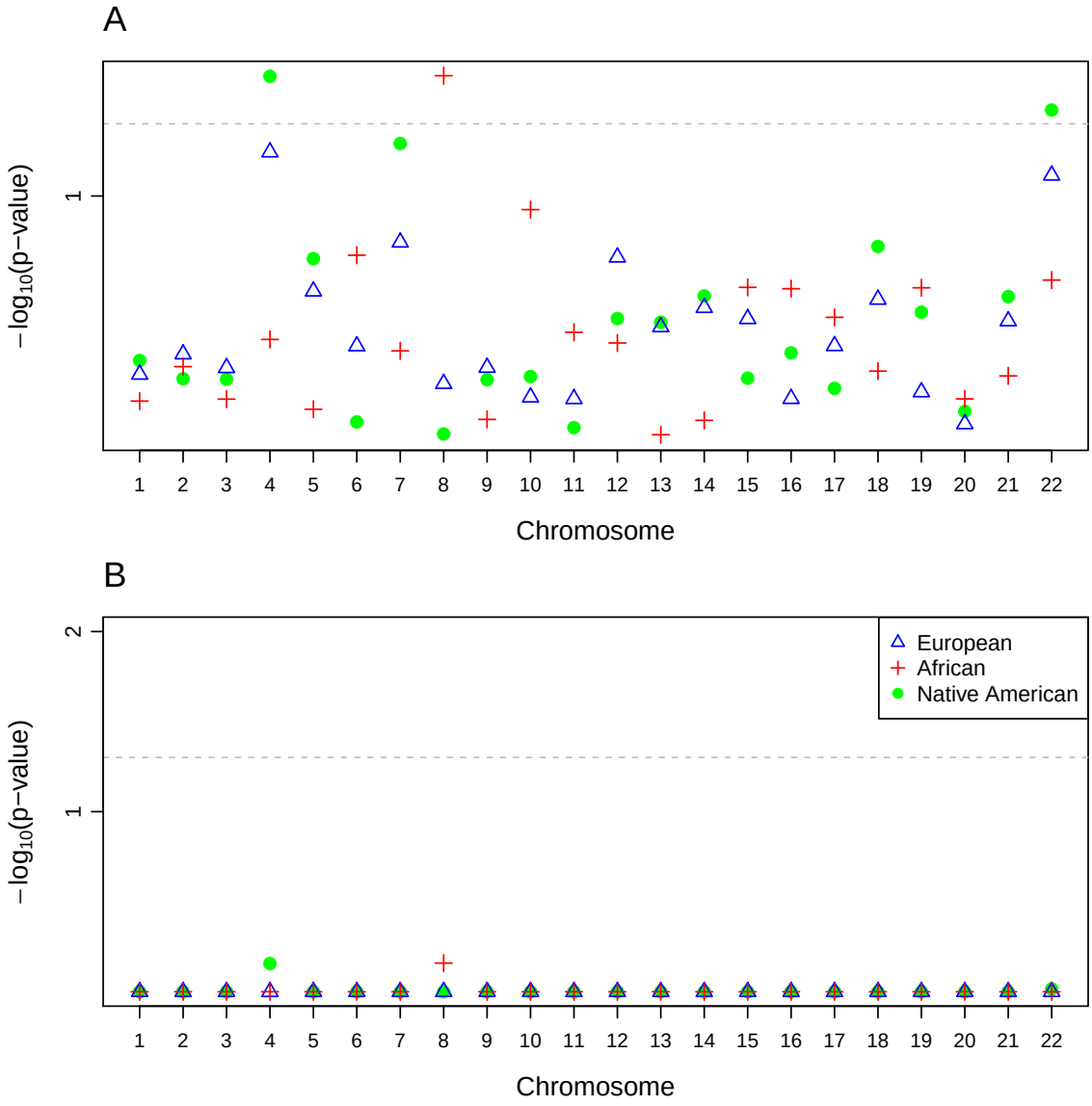


Figure 2.9: P-values from the parametric CAnD test comparing the ancestry at each autosome with the mean ancestry of all other autosomes, for the African, European and Native American ancestries in the HapMap MXL population sample. Adjustment for multiple testing was made using the Bonferroni multiple testing correction. (A): Uncorrected for multiple testing. (B): Bonferroni-adjusted. The dotted line indicates the 0.05 significance level.

with CAnD when using either local ancestry versus global ancestry estimates from RFMix or FRAPPE, respectively.

We also compared autosomal-wide and X chromosome ancestry estimates from RFMix and FRAPPE using genotype data for the HapMap MXL and ASW population samples. Table 2.3 shows the correlation of the ancestry estimates from the methods for each ancestral subpopulation. For the two predominant ancestries in the MXL (European and Native American) and ASW (African and European), the correlation between the ancestry estimates for the autosomes from RFMix and FRAPPE are all greater than 0.99, and is 0.95 or greater for the X chromosome. As previously mentioned, there is very little Native American ancestry and African ancestry in the ASW and MXL, respectively. Nevertheless, with a correlation of 0.99, Native American ancestry estimates on the autosomes are nearly perfectly correlated between RFMix and FRAPPE, and the correlation between the estimates is 0.90 for Native American ancestry on the X chromosome in the ASW. For proportional African ancestry in the MXL, the correlation between the two estimates is 0.893 for the autosomes and 0.93 for the X chromosome. So, for the predominant ancestries in the MXL and ASW, there appears to be little difference in estimating autosomal ancestries with FRAPPE or by averaging local ancestry estimates from RFMix. There is high concordance between the methods for the predominant ancestry in ASW and MXL for the X chromosome as well. In general, there is less concordance between the methods when estimating proportional ancestries from populations with relatively small contributions to the admixed population, and local ancestry estimates, such as RFMiX, are likely more accurate in inferring low levels of ancestral contribution, than global ancestry methods, such as FRAPPE.

2.3.7 Assortative Mating for Ancestry in the HapMap MXL

The CAnD test identified significant heterogeneity in ancestry among the HapMap MXL chromosomes. Systematic differences in ancestry at genomic loci on chromosomes can be due to sex-specific patterns of non-random mating at the time of or since admixture. We investigated assortative mating between pairs of individuals in the HapMap MXL for which there is a documented offspring; there are 24 such pairs. However, we excluded three mate

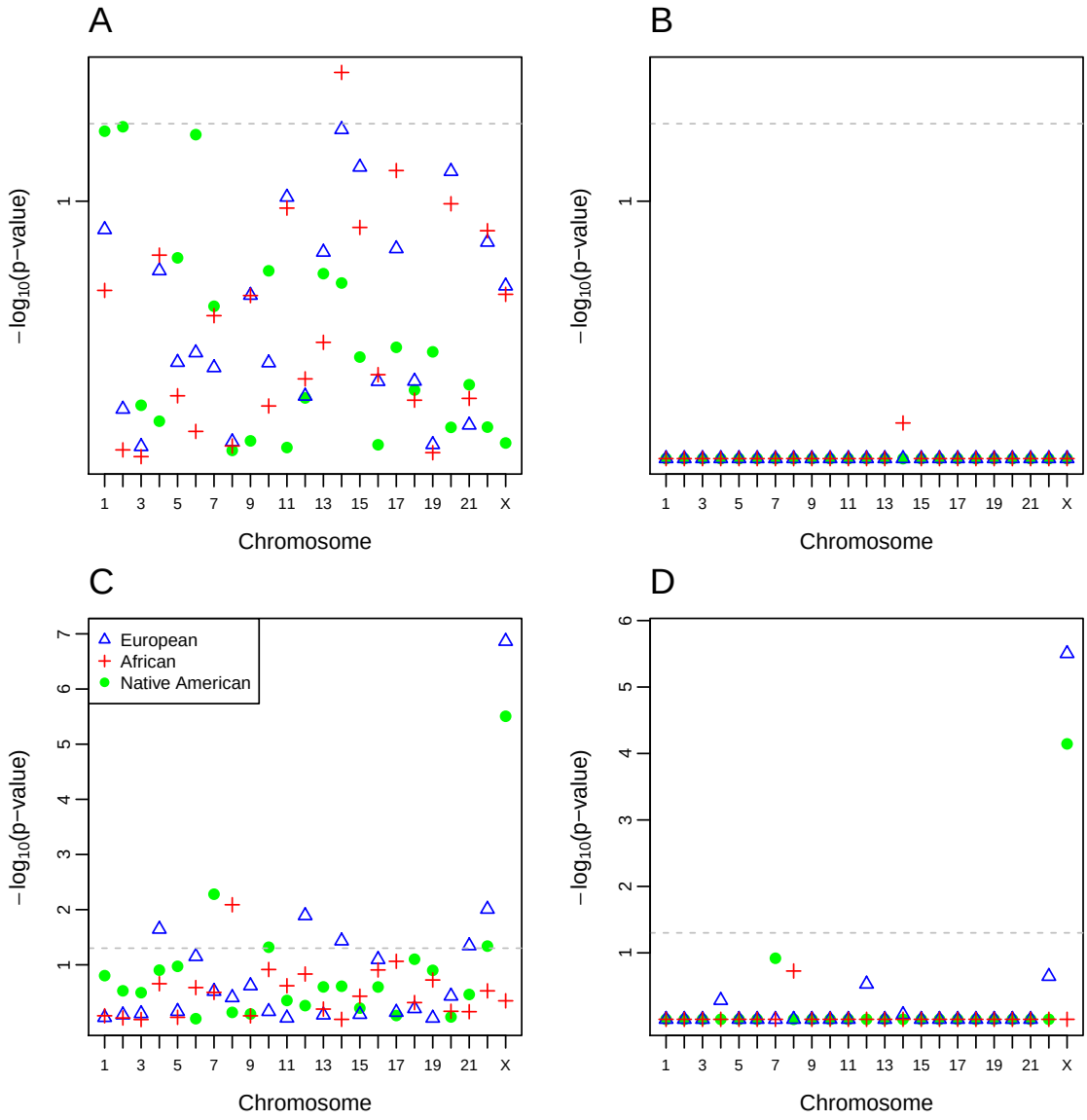


Figure 2.10: Unadjusted p-values from the parametric CAnD test comparing the ancestry at each autosome with the mean ancestry of all other autosomes, for the African, European and Native American ancestries in the (A): HapMap ASW and (C): HapMap MXL population samples. Adjustment for multiple testing was made using the Bonferroni multiple testing correction and the corresponding results for (B): HapMap ASW and (D): HapMap MXL are included. The dotted line indicates the 0.05 significance level.

Table 2.3: Correlation between ancestry estimates from RFMix and FRAPPE, stratified by autosomal and X chromosome estimates, in each of the population samples.

	Autosomal		X Chromosome	
	ASW	MXL	ASW	MXL
African	0.9990	0.8932	0.9697	0.9256
European	0.9979	0.9935	0.9548	0.9878
Native American	0.9963	0.9940	0.9001	0.9898

pairs due to cryptic relatedness (described earlier) with other mate pairs, resulting in a subset of 21 independent mate pairs included in our assortative mating analysis.

We used an empirical distribution to assess if the observed correlations of ancestry between mate pairs are significantly different from what would be expected under the null hypothesis of random mating. In particular, we randomly permuted the MXL mate pairs 5,000 times, and for each of the 5,000 permutations, we calculated correlations of the mate pairs for each of the three ancestries (European, Native American, and African). The correlations of each ancestry on the autosomes and the X chromosome between mate pairs from the 5,000 permutations were then used to construct empirical distributions under the null hypothesis of random mating in the MXL. The distributions of ancestry correlations among mate pairs are centered around zero when there is random mating, with a standard deviation around 0.2 for each of the three ancestries (Figure 2.11).

We first tested the null hypothesis of random mating versus an alternative hypothesis of assortative mating for ancestry using the observed correlations among mate pairs and the empirical null distributions. Table 2.4 shows the p-values for the autosomal and X chromosome correlations of African, European and Native American ancestry proportions calculated from the 21 MXL mate pairs. There is significant evidence of assortative mating for European and Native American ancestries on the autosomes in the HapMap MXL, with corresponding p-values of 0.015 and 0.017, respectively. There is also significant evidence for assortative mating based on European and Native American ancestry on X chromosome,

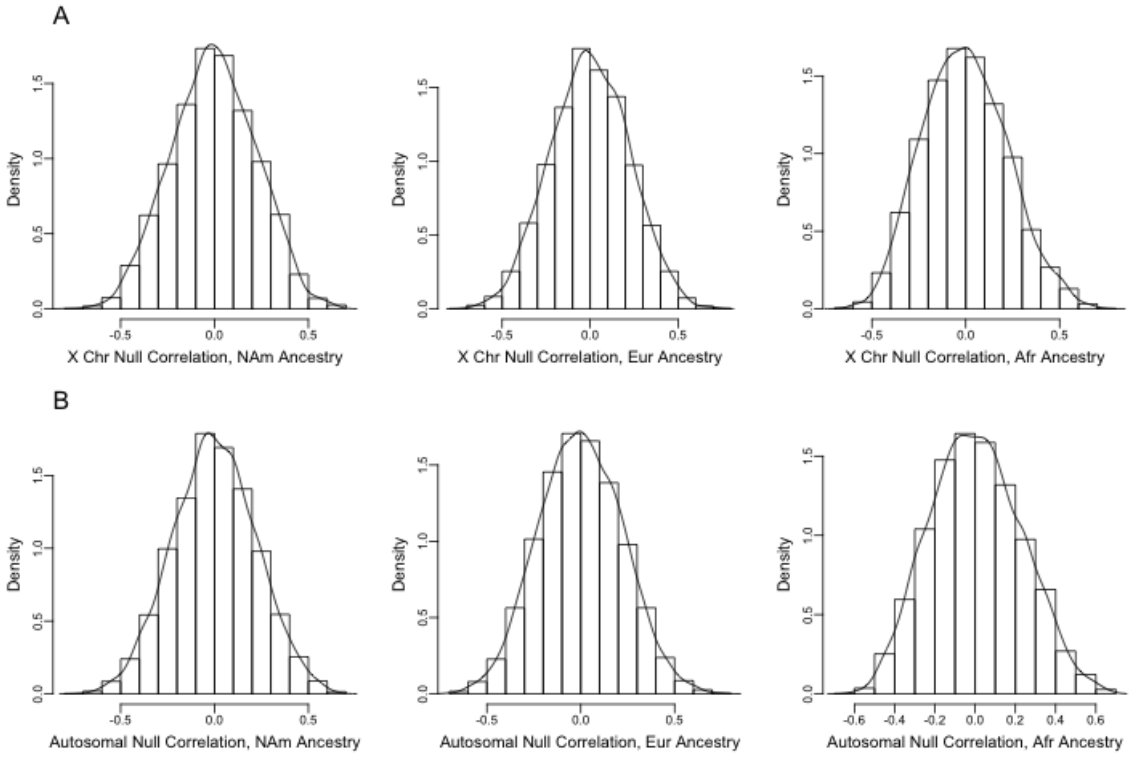


Figure 2.11: (A): X chromosome and (B): autosomal empirical null distributions of correlation of Native American, European and African ancestries between 5,000 mate pair permutations.

with p-values of 0.011 and 0.007, respectively. The p-values remain significant, even after Bonferroni correction for testing three ancestries. There is not significant evidence of assortative mating for African ancestry for either the autosomes or the X chromosomes (p=0.26 and 0.14, respectively). For testing the null hypothesis of random mating versus an alternative hypothesis of non-random, e.g., assortative or disassortative mating, a two-sided test can be conducted. The p-values for this test are given in Table 2.4 and are roughly twice the assortative mating p-values. We also performed permutation tests to assess evidence of assortative and non-random mating for 11 HapMap ASW mate pairs with a documented offspring. No significant evidence of assortative mating in the ASW was detected, and ASW p-values for the three ancestries tested are given in Table 2.4.

2.3.8 Ancestry Equilibrium on the X Chromosome Under Random Mating After Initial Admixture Event

We also investigated the number of generations required for males and females to reach ancestry equilibrium on the X chromosome in a randomly mating population. We considered the setting where there is admixing between two ancestral populations and where mate pairs at the initial admixture event consist of males with ancestry entirely from one of the populations and females having ancestry derived from the other population. We then performed a simple computation to estimate proportional ancestry for each generation assuming random mating, and after an initial admixing event between founder females and males with the most extreme setting of discordant ancestry between the two sexes at the time of admixture. Figure 2.12 shows the proportion ancestry by generation in the admixed population for males and females. A recent finding published a similar result, although the initial ancestry proportions considered did not include the extreme proportions as we did here [14]. We find an equilibrium of 1/2 is reached for autosomal ancestry in males and females in the first generation. Proportional ancestry on the X chromosome for males and females tends to an equilibrium of 2/3 and 1/3 of the founder female and male ancestries, respectively, that is achieved around eight generations after the initial admixing event. This result is not surprising since females contribute 2/3 of the X chromosomes in a population.

Table 2.4: P-values detecting assortative or disassortative mating for ancestry among 11 HapMap ASW and 21 HapMap MXL mate pairs, calculated on the autosomes and the X chromosome separately. The p-values are calculated from the empirical distribution created from sampling 5,000 mate pairs at random. Results presented under 'assortative mating' tested the hypothesis of no assortative mating, while 'non-random mating' tested the hypothesis of neither assortative nor disassortative mating.

		HapMap ASW			HapMap MXL		
		African	European	Native American	African	European	Native American
Autosomal	assortative mating	0.365	0.388	0.234	0.139	0.015	0.017
	non-random mating	0.871	0.888	0.532	0.268	0.028	0.032
X Chromosome	assortative mating	0.842	0.788	0.564	0.256	0.011	0.007
	non-random mating	1.000	1.000	1.000	0.530	0.024	0.013

A recent study developed a model that showed the 2/3 and 1/3 ancestry proportions on the X chromosome in admixed populations derived from two ancestries with a single admixture event may be accurate, but is not correct if the admixing is ongoing [14]. Nevertheless, whether a single admixture event or ongoing admixture is assumed, the X chromosome and the autosomal chromosomes will not have the same equilibrium ancestry proportions in an admixed population when males and females have different ancestries at the time of the admixture event(s).

2.3.9 Application to HCHS/SOL Dataset

We applied CAnD to a cohort of over 12,000 self-identified Hispanic/Latino individuals recruited from four metropolitan areas in the U.S. In particular, we estimated ancestry proportions within a set of 10,642 unrelated samples. All individuals in HCHS/SOL provided a self-identified 'background group,' choosing one of Cuban, Dominican, Puerto Rican, Mexican, Central American, South American or Other. For purposes of this analysis, we grouped together those who have an unknown or missing background group with those who self-identified as 'Other.' We filtered SNPs with a sample minor allele frequency (MAF) < 0.05 from the HCHS/SOL genotypes and merged the HCHS/SOL genotyped markers with the Human Genome Diversity Panel (HGDP) [5] reference samples to obtain a set of 440,908 SNPs in common (431,143 autosomal and 9,694 X chromosome). We applied CAnD to local ancestry estimated using RFMix [37] averaged by chromosome.

Reference populations of African, European and Native American were used with RFMix to estimate local ancestry across autosomes and the X chromosome as described in Section 2.2.5. Figure 2.13 shows the average local ancestry across the autosomes and the X chromosome within each individual, stratified by self-identified background group. This Figure 2.13 displays the mean values outlined in Table 2.5, where the corresponding standard deviation estimates are presented. Even within the broad group of individuals who self-identify as Hispanic/Latino there is variation, which is most evident when we stratify the ancestry proportions by background group. The Caribbean (Cuban, Dominican and Puerto Rican) populations have a larger proportion of African ancestry than the Mainland (Mex-

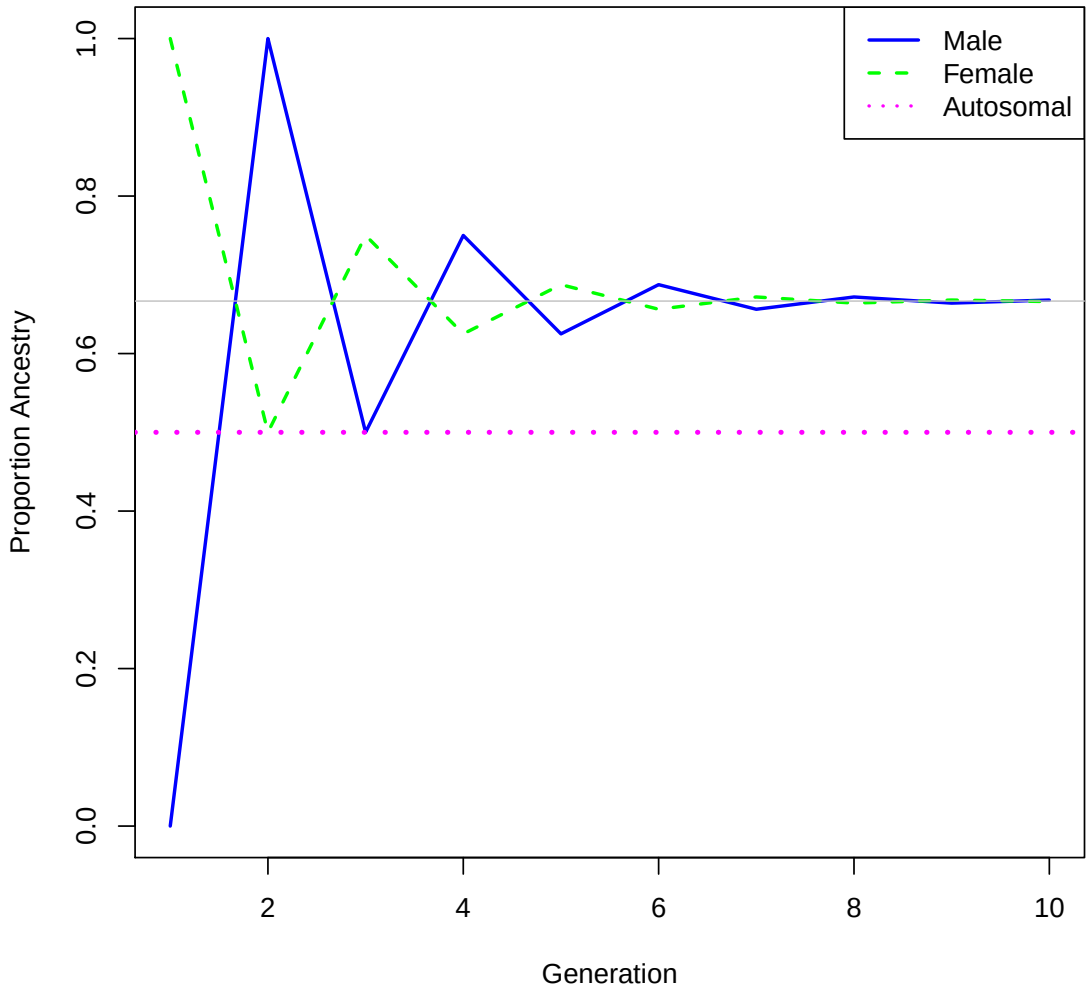


Figure 2.12: The proportion of ancestry for the autosomes and the X chromosome by sex, assuming females and males have opposite ancestries at the initial admixture event. After the initial admixture event, random mating is assumed. The gray line shows the equilibrium proportions on the X chromosome.

Table 2.5: Mean (SD) as calculated in 10,642 HCHS/SOL samples, stratified by self-identified background group and autosomal versus X chromosome.

Self-Identified	Autosomal			X Chromosomal		
Bkgd Group	African	European	N Am	African	European	N Am
Cuban	0.133 (0.172)	0.804 (0.183)	0.0625 (0.0624)	0.189 (0.235)	0.679 (0.26)	0.132 (0.139)
Dominican	0.435 (0.162)	0.495 (0.146)	0.0701 (0.0464)	0.508 (0.227)	0.371 (0.212)	0.121 (0.108)
PuertoRican	0.223 (0.135)	0.638 (0.122)	0.139 (0.0538)	0.252 (0.208)	0.518 (0.208)	0.23 (0.144)
Mexican	0.0455 (0.0305)	0.459 (0.17)	0.496 (0.175)	0.0631 (0.0783)	0.32 (0.218)	0.617 (0.227)
CentralAmerican	0.128 (0.137)	0.428 (0.159)	0.444 (0.165)	0.146 (0.175)	0.298 (0.211)	0.556 (0.234)
SouthAmerican	0.0744 (0.112)	0.476 (0.222)	0.449 (0.239)	0.0926 (0.148)	0.335 (0.263)	0.572 (0.295)
Other/Unknown	0.197 (0.209)	0.595 (0.22)	0.207 (0.19)	0.248 (0.267)	0.463 (0.278)	0.289 (0.247)

ican, Central American and South American) populations, across both the autosomes and the X chromosome. Correspondingly, the Caribbean populations have less Native American ancestry than the Mainland populations, overall. Comparing the autosomal to the X chromosome ancestry proportions within a background group, the Caribbean populations have larger proportions of African ancestry on the X chromosome while the Mainland populations have larger proportions of Native American ancestry.

We visualize the average local ancestry estimates as boxplots in Figure 2.14, directly comparing the autosomal and X chromosome ancestry proportions within a self-identified background group. The patterns recognized in the barplots (Figure 2.13) are more clearly displayed here. In particular, the median proportion Native American ancestry as estimated using X chromosome markers is higher than the autosomal median in all background groups. On the contrary, the median average local European ancestry is lower on the X chromosome than the autosomes in every background group. Furthermore, the proportion local ancestry ranges from zero to one in almost every ancestral subpopulation, background group and chromosome type.

Figure 2.15 shows the p-values from applying the CAnD and pooled t-test comparing the X chromosome to the autosomes for each ancestral subpopulation. A set of 10,642 unrelated HCHS/SOL samples was used. Each test was applied separately to the six self-identified

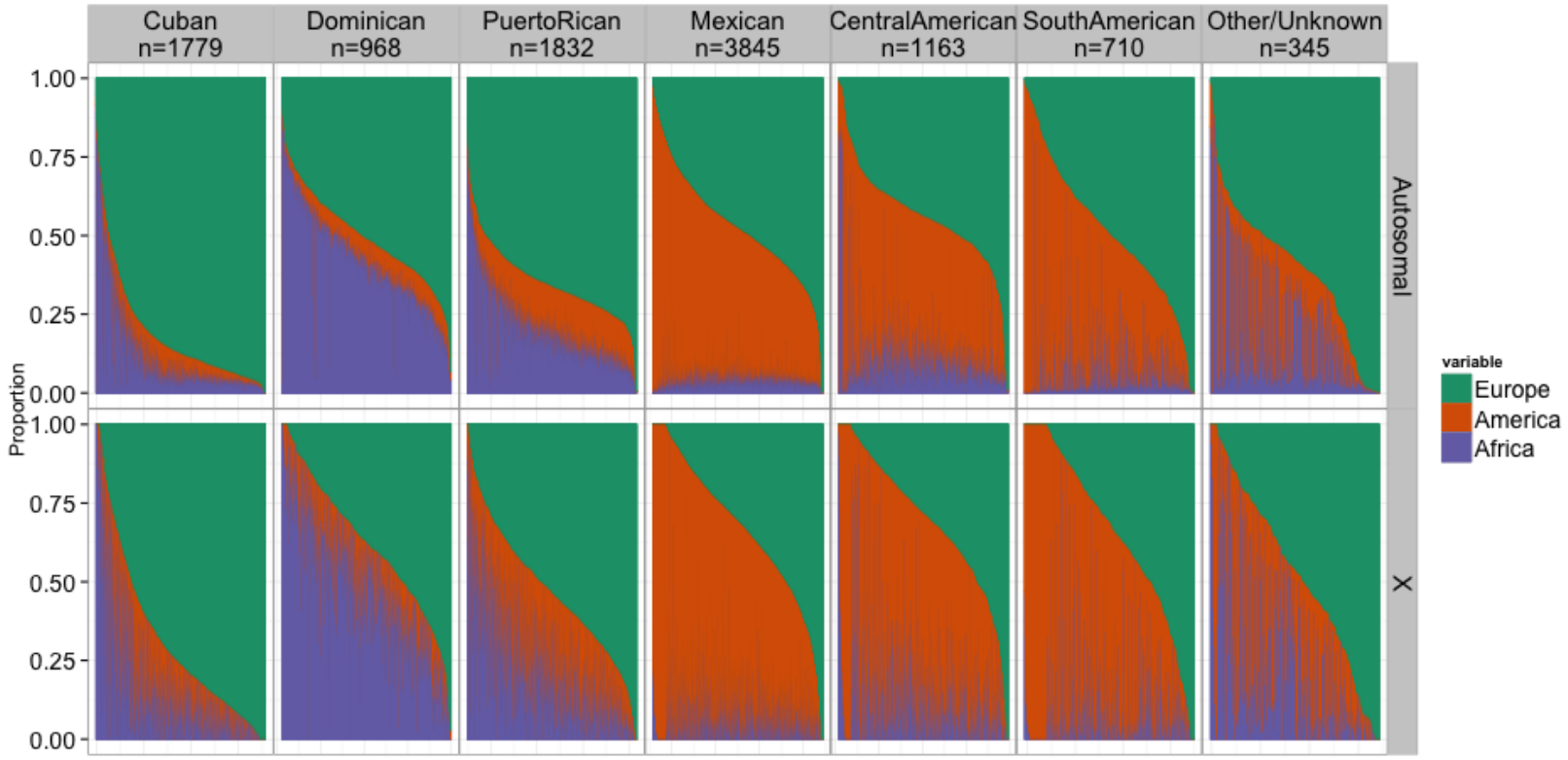


Figure 2.13: Barplot of average local ancestry within 10,642 HCHS/SOL samples, stratified by self-identified background group and autosomal versus X chromosome. Each bar is colored according to the proportion local ancestry of African, European and Native American estimated within each individual.

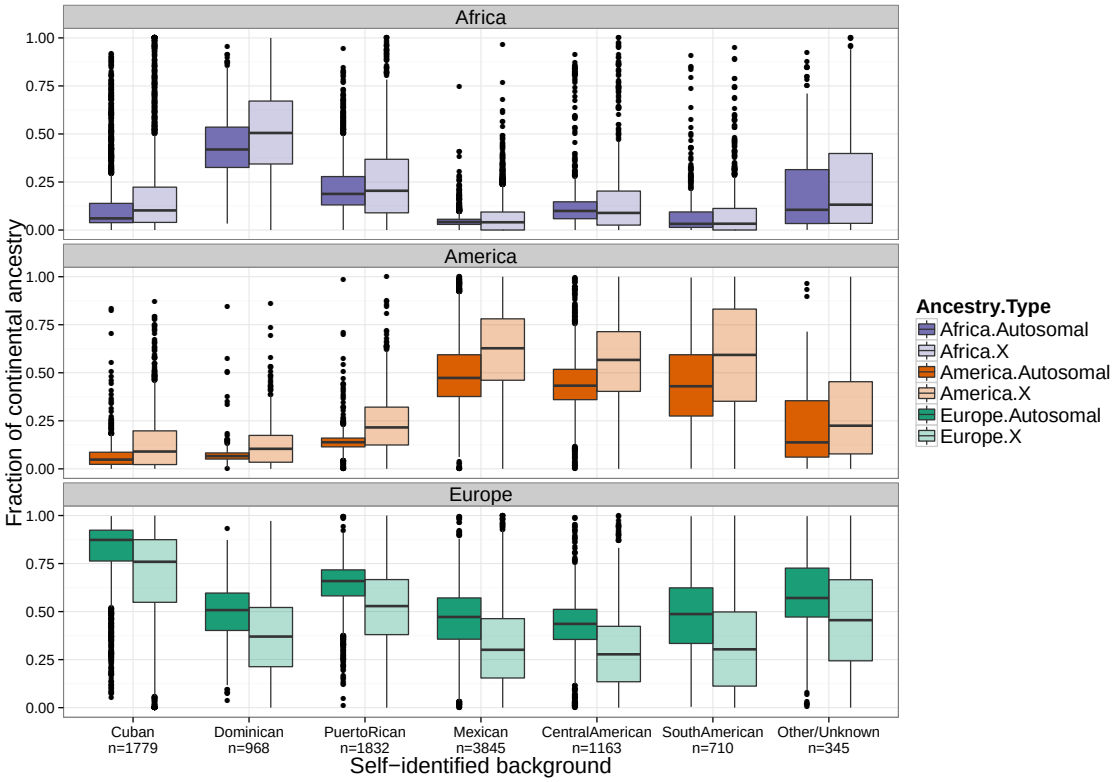


Figure 2.14: Boxplots of African, Native American and European ancestry in 10,642 HCHS/SOL samples stratified by self-identified background group and chromosome type (autosomal or X chromosome).

background groups, as well as the 'Other/Unknown' subgroup. The Mexican background group yields highly significant CAnD p-values for all three ancestral subpopulations. As the Mexican subgroup has over 3,500 samples, it is the largest group and thus achieves highest power. The Caribbean populations yield significant CAnD p-values in all ancestral subpopulations, and the Mainland groups yield significant CAnD p-values when considering the European and Native American ancestries. As seen in Figures 2.13 and 2.14, there is only a small proportion of African ancestry. As a result, we see only nominally significant differences between X chromosome and autosomal African ancestry proportions in the Central or South American background groups.

We applied CAnD comparing each autosome to a pool of the other autosomes within each self-identified background group (Figure 2.16). Unsurprisingly, the CAnD method yielded more significant results than the pooled t-test. Among the self-identified Mexican samples, we identify many autosomes, including chromosomes 2, 4, 5, 15, 16, 18, and 19, that have a statistically significant difference in either Native American or European ancestry as compared to a pool of the remaining autosomes. We also identify many significant autosomes in the Puerto Rican samples, with chromosomes 2, 3, 6, 7, 8, 9, 10, 16 and 21 exhibiting significant differences in ancestry as compared to a pool of the other autosomes. Other self-identified background groups have significant results, although only slightly so.

To further identify which region(s) of a chromosome are driving the significance of the CAnD test, we can perform CAnD on regions of a chromosome, comparing the estimated ancestral proportion in a chromosomal segment to a genome-wide average ancestry, excluding the chromosomal segment under analysis. Using a sliding window approach, we calculate CAnD p-values for a given chromosome allowing for finer scale identification of regions with ancestry differences. We performed the CAnD test on average local European ancestry estimates in the Puerto Ricans on chromosome 2, testing overlapping sliding windows of 100 segments and moving along 50 segments each iteration. The average within a given window was compared to the mean autosomal European ancestry excluding the chromosome segment under consideration. From this analysis, we find regions of chromosome 2 that have significant differences in European ancestry compared to the mean autosomal European ancestry in Puerto Ricans (Figure 2.17). One region has the highest p-value while regions

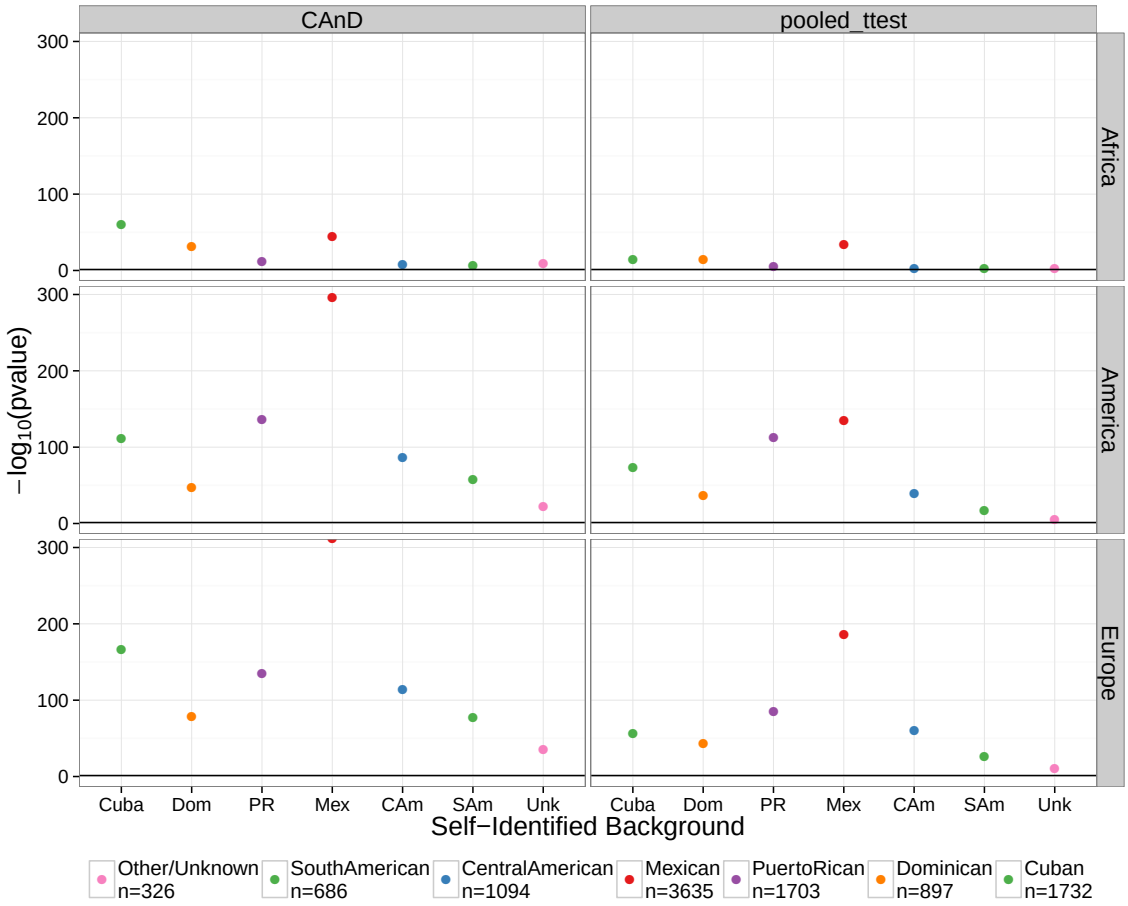
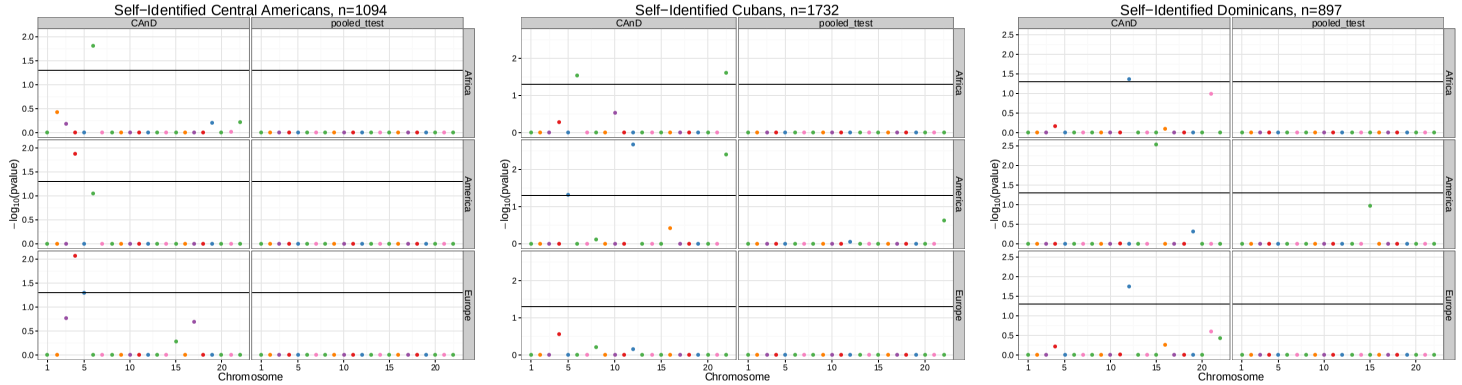
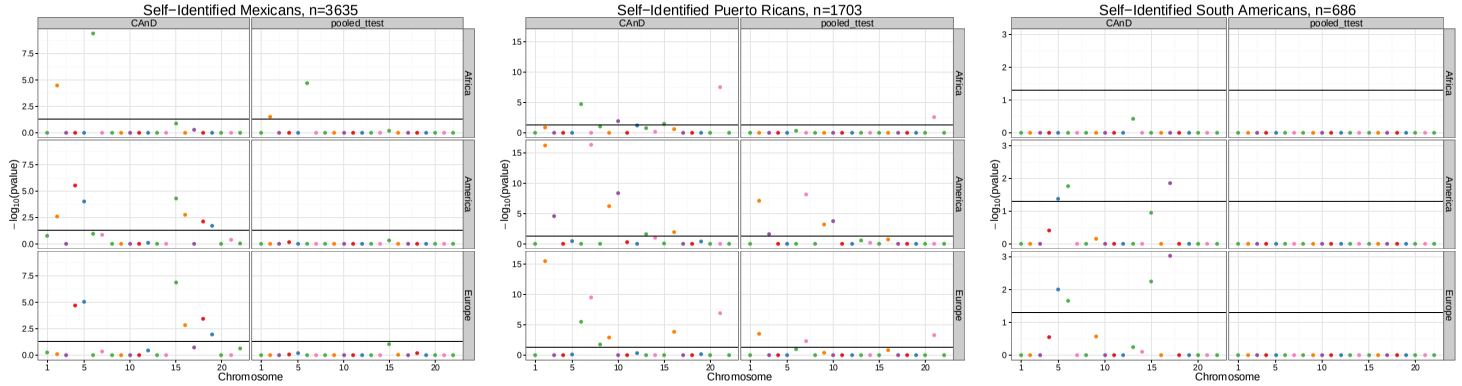


Figure 2.15: P-values from the CAnD and pooled t-test applied to 10,642 HCHS/SOL samples comparing average X chromosome ancestry to average autosomal ancestry. The samples were stratified by self-identified background group and ancestral subpopulation. The horizontal line indicates statistical significance after Bonferroni correction.



(a) Central Americans (b) Cubans (c) Dominicans



(d) Mexicans (e) Puerto Ricans (f) South Americans

Figure 2.16: P-values from the CAnD and pooled t-test applied to 10,642 HCHS/SOL samples comparing average ancestry on each autosome to the average autosomal ancestry excluding the autosome being tested. The samples were stratified by self-identified background group and ancestral subpopulation. The horizontal line indicates statistical significance after Bonferroni correction.

around it also show significance; this is expected as regions are correlated through linkage disequilibrium. This pattern offers support that indeed ancestry in this region is more different than ancestry across the remainder of the autosomes. The most significant region is close to the lactase gene, which includes variants that are unique to different ancestries [47].

2.4 Discussion

Systematic ancestry differences at genomic loci may arise in recently admixed populations as a result of selection and ancestry related assortative mating. Here, we developed the CAnD method for detecting heterogeneity in population structure across the genome in populations with admixed ancestry. CAnD uses ancestry inferred from SNP genotype data to identify chromosomes that have significantly different contributions from the underlying ancestral populations. The CAnD method takes into account correlated ancestries among chromosomes within individuals for both valid testing and improved power for detecting heterogeneity in population structure across the genome. Some additional features of the CAnD method are: (1) X chromosome data can easily be incorporated in the analysis; and (2) the method can be used for testing heterogeneity in ancestry among any subset of chromosomes in the genome.

We performed simulation studies with admixture and real genotype data from HapMap. We demonstrated that CAnD had appropriate type I error. We also showed in the simulation studies that the CAnD test has higher power to detect heterogeneity in ancestry between chromosomes than a pooled t-test that does not take account correlations in ancestry among chromosomes.

We applied the CAnD method to the HapMap MXL population sample where significant heterogeneity in European ancestry and Native American ancestry was detected across the genome (autosomes and the X chromosome), with p-values of 9e-07 and 1e-06, respectively. A subsequent analysis showed that the heterogeneity in ancestry across the MXL genomes detected by CAnD is largely due to elevated Native American ancestry and deficit of European ancestry on the X chromosomes. These results are consistent with previous reports for U.S. Hispanic/Latinos [4] and Latin Americans [3], where it has been suggested that the X versus autosomal differences are likely due to sex-specific patterns of gene flow in which

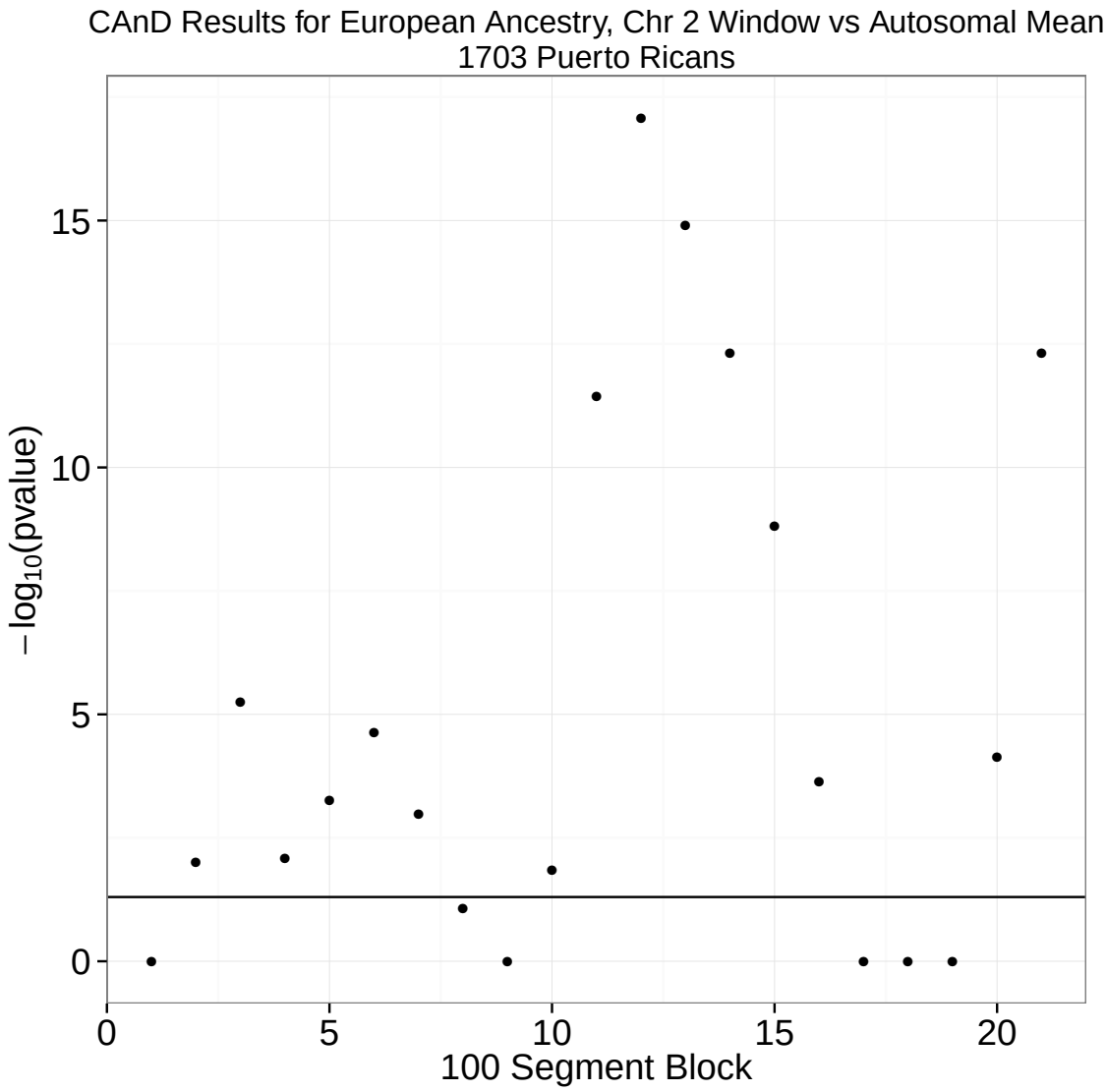


Figure 2.17: P-values from testing a 100-segment window on chromosome 2 to the average across the remainder of the autosomes, using estimated European ancestry within the self-identified Puerto Rican group. An overlapping sliding window was used, where the window slid 50 segments each iteration.

European male colonists contributed substantially more genetic material than European females at the time of admixture. There was no significant evidence of genetic heterogeneity in the HapMap ASW detected by CAnD, and no significant differences in ancestry between the autosomal chromosomes and the X chromosome were detected. The autosomal chromosomes and the X chromosome in the ASW are largely African derived, and a larger sample is required to have adequate power for the detection of chromosomal ancestry differences in this population.

To identify ancestry differences by chromosome, we applied CAnD to self-identified Hispanic/Latino individuals from the HCHS/SOL study. Highly significant heterogeneity in African, European and Native American ancestry was detected across the genome when testing each self-identified background group separately, comparing the X chromosome to the autosomes. Heterogeneity in population structure was detected across the autosomes as well, in all ancestral subpopulations and most self-identified background groups. In the self-identified Puerto Ricans, we followed up on the heterogeneity in European ancestry detected on chromosome 2. We were able to localize the structure in these samples, narrowing the regions of heterogeneity in order to find genes or regions of the genome that may be under selection. This procedure can be applied in order to fine-map regions of the genome that exhibit patterns of natural selection.

The CAnD method can incorporate estimates of local ancestry at specific locations across the genomes, using software such as RFMix, or proportional ancestry estimates for each chromosome with software such as FRAPPE or ADMIXTURE. We compared the CAnD results for the HapMap MXL when using local ancestry estimates from RFMix, which requires phased genotype data, to the results when using chromosomal ancestry estimates with FRAPPE where unphased genotype data was used. Heterogeneity in ancestry was detected with CAnD when using either local ancestry estimates from RFMix or chromosomal ancestry estimates from FRAPPE. Interestingly, p-values were slightly smaller when using estimates from FRAPPE that were based on unphased genotype data as compared to using local ancestry estimates from phased genotype data. This result may be an artifact from RFMix not taking into account uncertainty in the phasing when estimating local ancestry.

In the present paper, CAnD was used to identify entire chromosomes with ancestry

contributions that are significantly different than expected. If local ancestry estimates are available, CAnD can be used to follow up on the chromosomal findings by fine-mapping the specific regions that may be under selection. CAnD may be used with a sliding window or a set of genes within a chromosome to localize areas that exhibit heterogeneity in population structure. Future work will consider this approach.

We also investigated the number of generations required for ancestry on the X chromosome to reach equilibrium in males and females after a single admixing event with two populations. In the most extreme setting where all males are from one population and all females are from another population at the time of admixture, approximately eight generations are required under random mating between males and females to reach ancestry equilibrium on the X. Estimates of the number of generations since admixture in the Mexican population [21] range from 10 to 15. It is reasonable to assume that equilibrium on the X chromosome for males and females should have been reached in the Mexican population if mating in this population is at random. Previous studies [50, 52], however, have shown evidence of non-random mating in Mexican populations. In the HapMap MXL, between mate pairs that produced an offspring, we also detected significant evidence of assortative mating, where the correlation of European and Native American ancestries on both the autosomes and the X chromosome is significantly higher than what would be expected under the null hypothesis of random mating. Evaluating differences in ancestry on the X chromosome between males and females may potentially be a useful tool for the detection of non-random mating in recently admixed populations, since under the most extreme setting of discordant ancestry between males and females at the time of admixture, we find that there should be no difference in ancestry on the X chromosome between males and females after eight generations of random mating. This is future work to be considered.

The CAnD method is implemented in the R language and is available from Bioconductor (http://www.bioconductor.org) as part of the CAnD package.

Chapter 3

MIXED MODEL ASSOCIATION TESTING ON THE X CHROMOSOME IN GWAS WITH UNKNOWN SAMPLE STRUCTURE

3.1 Introduction

Genome-wide association studies (GWAS) are routinely conducted for the identification of genetic variants that influence complex traits and diseases. For valid association testing of a genetic marker with a trait of interest, methods for traditional case-control or population-based studies assume individuals are unrelated and of the same ancestral background. Due to the decreasing cost of genotyping, it is now common for genetic studies to have samples with thousands, or even tens of thousands of individuals. As a result, the assumption that all sampled individuals are independent and from a homogeneous population is often not valid. If not properly accounted for, sample structure can cause spurious association at a marker level [43, 16]. On a genome-wide scale, the genomic control factor can be inflated in the presence of sample structure [12].

Accurately detecting and subsequently accounting for sample correlation due to relatedness and population stratification has been a major focus of genetic research over the past few years. To estimate population structure, methods based upon principal components analysis (PCA) have been proposed, such as EIGENSTRAT [45], which assumes samples are independent. PC-based extensions have been developed that allow for related samples [66]. It is widely accepted that PC results mirror sample structure and adjustment for estimated PCs can aid in controlling for population stratification in association testing. Estimating sample correlation due to relatedness using observed genotype data is possible with methods such as REAP [56], which provides kinship estimates that are robust to population structure. More recent methods such as PC-AiR and PC-Relate [10, 11] deconvolute sample structure due to relatedness and population stratification without requiring a known pedigree structure, allowing for accurate adjustment for each of these components

in association testing.

Mixed linear models (MLMs) have emerged as a powerful and effective approach for association mapping in GWAS with population structure, family structure and/or cryptic relatedness [22, 29, 65]. MLMs use variance components and an empirical genetic relationship matrix (GRM) to simultaneously account for both population structure and relatedness among sample individuals. Based on the assumption that the effect of any given genetic variant on the trait is very small [36], the variance parameters are estimated once for each trait of interest and then the model is fit for each genetic marker. This simplifying assumption enables MLM methods to be applied genome-wide in a feasible amount of computation time.

Despite advancements in association testing methods, recent work highlighted the lack of associations identified on the X chromosome [59]. Although the size and number of genes on the X chromosome are similar to chromosome 7, the X chromosome is often excluded from association analyses, leading to a disproportionate number of published associations on the X [17]. Widely used MLM methods are not directly applicable to analyzing X chromosome markers. Some of the methodological challenges for association testing on the X include (1) accounting for X chromosome copy number differences in females and males and (2) appropriately adjusting for genetic correlations among sample individuals on the X, including pedigree and population structure, and ancestry admixture which can be quite different from the autosomes (see Chapter 2).

Here, we propose the MLM-X method for genome-wide association testing in samples with unknown structure. The MLM-X framework is valid and powerful for testing genetic variants across the autosomes and the X chromosome. MLM-X includes random effects for polygenic effects on both the X chromosome and the autosomes while appropriately taking into account unique correlation on the X chromosome. Variance components of the random effects in MLM-X are calculated via average information-restricted maximum likelihood (AI-REML) with two empirical GRMs, one for the autosomes and one for the X chromosome. MLM-X has a similar configuration as the LMMOPS [8] method in which the inclusion of additional random effects, such as shared environmental exposures, is a simple extension. In simulation studies, we demonstrate that MLM-X can provide an improve-

ment over existing MLM approaches, in terms of type I error and power, for complex trait mapping with both autosomal and X-linked variants. We applied MLM-X to a cohort of over 12,000 Hispanic/Latino individuals recruited from four metropolitan areas in the U.S. to identify genetic variants associated with red blood cell count (RBC). MLM-X detects genome-wide significant associations with RBC on the X chromosome and has a lower genomic control inflation factor than existing MLM approaches that ignore polygenic effects on the X chromosome.

3.2 Methods

The MLM-X method uses mixed linear models for association testing in the presence of sample structure genome-wide. We estimate PCs to correct for population stratification and pairwise kinship coefficients (KCs) to correct for familial structure. Both the PCs and KCs are stratified by the X chromosome and autosomes and we use AI-REML to estimate the variance components once per phenotype. Then, to perform the association test for each genetic marker of interest, we adjust for autosomal and X chromosome PCs using fixed effects and separate random effects for autosomal and X chromosome family structure. Our framework allows for additional fixed and random effects as desired. Here, we outline the methods used to estimate all proposed random and fixed effects in the MLM-X model on the X chromosome and autosomes, and how to apply them for a genome-wide association analysis.

3.2.1 Mixed Linear Model Framework for Association Testing

Consider a sample of individuals N and let $\mathbf{y}$ be a vector of values for some quantitative trait of interest. Assume each individual in the sample has been genotyped at a set $\mathcal{S}$ of SNP markers. Let bold uppercase letters denote matrices, while bold lowercase letters indicate vectors. When testing each marker individually under the mixed effects linear model as described in [8] and assuming an additive, polygenic background of the trait, we can fit the model testing association of $\mathbf{y}$ with a particular marker $k \in \mathcal{S}$,

$$\mathbf{y} = \beta_0 + \beta_k \mathbf{x}_k + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{g}_A + \boldsymbol{\epsilon} \tag{3.1}$$

where β_0 is the intercept, $\mathbf{x}_k$ is a vector of standardized genotype values at marker k, $\mathbf{Z}$ is a matrix of fixed effects covariates, $\boldsymbol{\alpha}$ is a vector of coefficients corresponding to the fixed effects, and $\boldsymbol{\epsilon}$ is a vector that captures the unaccounted-for environmental effects, where $\text{var}(\boldsymbol{\epsilon}) = \sigma^2_\epsilon \mathbb{I}$. When thinking about single SNP association testing in a MLM framework, we recall Fisher's polygenic model. In Equation 3.1, the background, additive polygenic effects of the trait across the autosomes is denoted by

$$\mathbf{g}_A = \sum_{m \neq k} \beta_m \mathbf{x}_m \tag{3.2}$$

We assume $\mathbf{g}_A \sim \mathcal{N}(0, \sigma^2_A \boldsymbol{\Phi}_A)$, where $\boldsymbol{\Phi}_A$ is the GRM estimated on the autosomes as described in the previous section. Further, we assume the effects of each locus act additively across the genome and thus σ^2_A corresponds to the total additive genetic variance of the trait.

We aim to test whether $\beta_k = 0$ for each SNP k in turn. The Wald test is performed to test the significance of β_k for each marker k. We refer to the model defined by Equation 3.1 as the 'simple MLM model.'

3.2.2 The MLM-X Method

We propose the MLM-X model, which includes specific adjustment for background polygenic effects across the X chromosome

$$\mathbf{y} = \beta_0 + \beta_k \mathbf{x}_k + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{g}_A + \mathbf{g}_X + \boldsymbol{\epsilon} \tag{3.3}$$

where $\mathbf{g}_X \sim \mathcal{N}(0, \sigma^2_X \boldsymbol{\Phi}_X)$. Note the polygenic effects across the autosomes, $\mathbf{g}_A$, are still included in the model. The total additive genetic variance of the trait across the X chromosome corresponds to σ^2_X. Now, we test a vector of standardized genotypes $\mathbf{x}_k$ which can be either autosomal or X chromosome genotypes. With X chromosome genotype codings of 0, 2 for males and 0, 1, 2 for females, the estimation of $\boldsymbol{\Phi}_X$ is as presented in Section 3.2.4. These genotypes correspond to the assumption that X chromosome inactivation occurs completely at random in females, and that a female with two copies of an allele is 'equivalent' to a male with one copy of that allele. Under different genotype codings, the pairwise entries of $\boldsymbol{\Phi}_X$ must be scaled accordingly. We define $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ in the following section. For each

marker in turn, we fit MLM-X with $\widehat{\mathbf{\Sigma}}$ and estimate $\hat{\beta}_k = (\mathbf{x}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}$. Finally, we perform the Wald test to assess significance of $\hat{\beta}_k$,

$$\frac{\hat{\beta}_k^{\,2}}{\text{var}(\hat{\beta}_k)} \sim \chi_1^2 \tag{3.4}$$

which follows a χ_1^2 distribution under the null hypothesis.

3.2.3 Variance of Y Under the Null Hypothesis

We test the null hypothesis that $\beta_k = 0$ and take into account the genetic correlation among all individuals using variance components of random polygenic effects. In the MLM-X model, we define the covariance matrix of the trait $\mathbf{Y}$ under the null hypothesis of no association between the trait and genetic marker to be

$$\mathbf{\Sigma} = \sigma_A^2\mathbf{\Phi}_A + \sigma_X^2\mathbf{\Phi}_X + \sigma_\epsilon^2\mathbb{I} \tag{3.5}$$

where we assume the unmodeled variance of the trait is independent across all individuals.

Since $\mathbf{\Sigma}$ is unknown, we must first estimate it. In a dense genome-wide setting, we approximate the additive polygenic effects with all markers once using AI-REML [8], separating the autosomal and X chromosome markers. We use all markers because we do not know *a priori* which markers are causal for a trait of interest. The standard genetic relatedness matrix is constructed and estimates of the variance components are obtained using

$$\hat{\mathbf{\Sigma}} = \widehat{\sigma_A^2}\widehat{\mathbf{\Phi}_\mathbf{A}} + \widehat{\sigma_X^2}\widehat{\mathbf{\Phi}_\mathbf{X}} + \widehat{\sigma_\epsilon^2}\mathbb{I} \tag{3.6}$$

The covariance structure remains fixed for all SNPs k across the genome and model fitting is performed using this estimated covariance structure.

3.2.4 Estimation of Relatedness in a Homogeneous Population Sample

Relatedness is routinely estimated among study samples using dense genome-wide data across the autosomes. Methods such as KING-robust [34], REAP [56] and PC-Relate [11] are computationally efficient, and accurately classify pairs of individuals for even distantly

related individuals using observed SNP genotypes. The autosomal kinship coefficient between individuals i and j, Φ_{ij}^{A}, is defined as the probability that an allele sampled at random from an autosomal locus in each individual is identical by descent (IBD). This $n \times n$ matrix of pairwise autosomal kinship coefficients $\mathbf{\Phi}_A$ is often referred to as the genetic relatedness matrix (GRM) [61].

Here, we describe the extension of the generally used kinship coefficient on the autosomes to X chromosome genetic markers. The X chromosome kinship coefficient between individuals i and j, Φ_{ij}^{X}, is defined to be the probability of sampling one allele IBD at random from individual i and individual j on the X chromosome. Since males only have one copy of the X chromosome, there is no randomness in sampling alleles from males at each location on the X chromosome and it is with certainty that we will sample the allele inherited from his mother. We calculate the $n \times n$ matrix of X chromosome kinship coefficients, $\mathbf{\Phi}_X$, using N observed genotypes across the X chromosome and allele frequencies calculated at each locus l, p_l. x_{il} denotes an observed genotype for a given locus l for individual i, coded as 0, 1, 2 for females and 0, 2 for males. The following equation estimates the X chromosome kinship coefficient between a pair of individuals i and j under this genotype coding

$$\Phi_{ij}^{X} = \frac{\sum_{i=1}^{N}(x_{il} - 2p_l)(x_{jl} - 2p_l)}{\sum_{i=1}^{N} 2p_l(1 - p_l)} \tag{3.7}$$

Note Equation 3.7 is not valid under other genotype coding schemes, but can be easily altered with the proper mean and standard deviation of genotype codings, stratified by sex. The autosomal kinship coefficient matrix Φ_A can also be estimated using Equation 3.7, where X_{il} denotes genotypes across the autosomes coded as 0, 1, 2 for all individuals. We know that in the presence of relatedness, the allele frequency estimates of p_l can be biased. Thus, a further correction for the biased nature of the allele frequencies could be explored.

Table 3.1 shows Φ_{ij}^{X} compared to Φ_{ij}^{A} for common relationships. Unlike the autosomal kinship coefficient, the X chromosome kinship coefficient varies by sex. If two individuals have a non-zero X chromosome kinship coefficient, the autosomal kinship coefficient will also be non-zero. When calculating Φ_{ij}^{X} for more distant relatives, we must consider both the number and and the sex of the ancestors connecting a pair of individuals.

		Autosomes	X Chromosome
	Self, Female	$\frac{1}{2}$	$\frac{1}{2}$
	Self, Male	$\frac{1}{2}$	1
	Mother-Daughter	$\frac{1}{4}$	$\frac{1}{4}$
	Mother-Son, Father-Daughter	$\frac{1}{4}$	$\frac{1}{2}$
	Father-Son	$\frac{1}{4}$	0
	Full sisters	$\frac{1}{4}$	$\frac{6}{16}$
	Full brothers	$\frac{1}{4}$	$\frac{1}{2}$
	Sister-Brother	$\frac{1}{4}$	$\frac{1}{4}$
Maternal	Aunt-Niece	$\frac{1}{8}$	$\frac{3}{16}$
	Aunt-Nephew	$\frac{1}{8}$	$\frac{6}{16}$
	Uncle-Niece	$\frac{1}{8}$	$\frac{1}{8}$
	Uncle-Nephew	$\frac{1}{8}$	$\frac{1}{4}$
	Grandmother-Granddaughter	$\frac{1}{8}$	$\frac{1}{8}$
	Grandmother-Grandson	$\frac{1}{8}$	$\frac{1}{4}$
	Grandfather-Granddaughter	$\frac{1}{8}$	$\frac{1}{4}$
	Grandfather-Grandson	$\frac{1}{8}$	$\frac{1}{2}$
Paternal	Aunt-Niece	$\frac{1}{8}$	$\frac{1}{8}$
	Aunt-Nephew	$\frac{1}{8}$	0
	Uncle-Niece	$\frac{1}{8}$	0
	Uncle-Nephew	$\frac{1}{8}$	0
	Grandmother-Granddaughter	$\frac{1}{8}$	$\frac{1}{4}$
	Grandmother-Grandson	$\frac{1}{8}$	0
	Grandfather-Granddaughter	$\frac{1}{8}$	0
	Grandfather-Grandson	$\frac{1}{8}$	0
Maternal-Maternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{3}{32}$
	First cousins, Girl-Boy	$\frac{1}{16}$	$\frac{3}{16}$
	First cousins, Boy-Boy	$\frac{1}{16}$	$\frac{6}{16}$
Paternal-Paternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{1}{32}$
	First cousins, Girl-Boy	$\frac{1}{16}$	0
	First cousins, Boy-Boy	$\frac{1}{16}$	0
Paternal-Maternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{1}{16}$
	First cousins, Girl-Boy	$\frac{1}{16}$	0
	First cousins, Boy-Boy	$\frac{1}{16}$	0

Table 3.1: The autosomal and X chromosome kinship coefficient for commonly observed relationships.

3.2.5 Estimation of Relatedness in a Structured Population Sample

When estimating relatedness among a set of study samples that exhibit population stratification, kinship coefficients can be inflated for pairs of samples that have similar ancestry profiles [56, 8]. The PC-Relate [11] method calculates a GRM that is robust to population structure by using estimated individual-specific allele frequencies rather than the mean sample allele frequency. We can extend the PC-Relate method to X chromosome SNPs and infer relatedness accurately even in samples with complex population structure by using the genotype coding for the X as described above. Thus, PC-Relate can be use to estimate relatedness on both the autosomes and the X chromosome in the presence of population structure.

3.2.6 Genome-wide Estimation of Population Structure

The PC-AiR method provides estimates of population structure in the presence of relatedness [10]. We extended PC-AiR to X chromosome SNPs to obtain estimates of population structure on the X chromosome in samples that include relatives. Population structure inference is first performed on a set of unrelated individuals who were chosen based upon their ancestry profiles. By choosing the unrelated set in this way, the range of ancestries in the entire study sample is represented in the unrelated set of individuals. Then, the PCA values are predicted for the relatives based upon their genetic similarities to individuals in the unrelated set.

When extending the PC-AiR method to X chromosome genotypes, we code male genotypes as 0, 2 and female genotypes as 0, 1, 2. This genotype coding yields a common covariance between pairs of individuals, regardless of their sex (Appendix A). Thus, we can simply apply PC-AiR to our genotypes with this coding on the X chromosome as well as the autosomes.

3.2.7 Simulation Studies

We performed simulation studies using the `R` statistical package [49]. Samples were related through a pedigree with three generations, and independent SNPs from both the X chro-

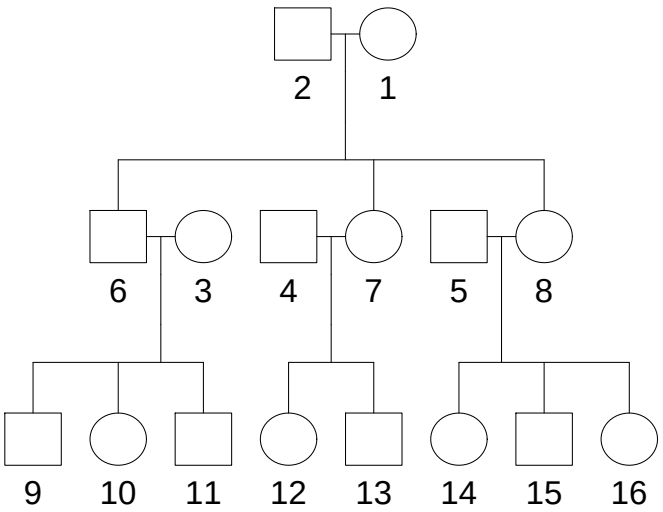


Figure 3.1: The 16-person pedigree used for the simulations.

mosome and an autosomal chromosome were generated for all founders at allele frequency 0.2. Genotypes for the remaining members of the pedigree were created by dropping alleles down the pedigree while properly taking into account the patterns of inheritance for the X chromosome. Male genotypes were coded 0, 2 and female genotypes were coded 0, 1, 2.

To assess the accuracy of estimating relatedness on the X chromosome, samples were simulated according to the 16-person pedigree shown in Figure 3.1. X chromosome genotypes were dropped down according to relatedness patterns as described above, and resulting genotypes were used to estimate pairwise relatedness on the X chromosome between all individuals in the pedigree. We considered increasing numbers of SNPs for the estimation.

The proposed MLM-X model was applied to simulated genotypes and phenotypes to calculate power and type I error. To do so, we considered two pedigree structures, one that has a substantially larger $\mathbf{\Phi}_X$ than $\mathbf{\Phi}_A$, and one that has similar values for $\mathbf{\Phi}_X$ and $\mathbf{\Phi}_A$ (Figure 3.2). We refer to these pedigrees as 'male-centric' and 'balanced,' respectively.

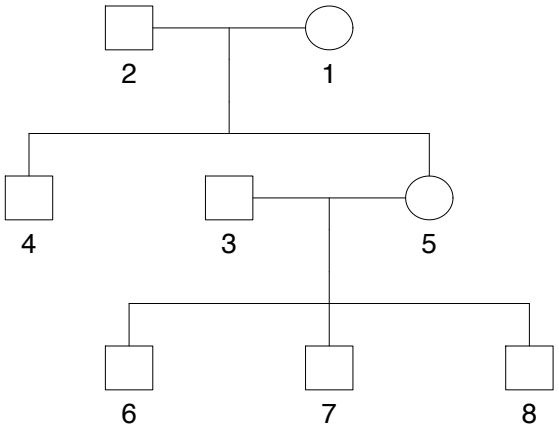
The 'male-centric' pedigree has six males and two females and the 'balanced' pedigree has four males and four females. These pedigree structures were considered in particular because the 'male-centric' pedigree has mean($\mathbf{\Phi}_X/\mathbf{\Phi}_A$) $\approx$ 1.8 and the 'balanced' pedigree has mean($\mathbf{\Phi}_X/\mathbf{\Phi}_A$) $\approx$ 1.1. Thus, we can assess the performance of MLM-X as well as the simple MLM under these two relatedness settings.

Variance components were estimated under different settings. Two values for σ^2_X and σ^2_A were considered, 0.3 and 0.5. A value of 3 was considered for each variance, in turn, to examine the model performance under extreme settings. The value for σ^2_ϵ was set to 1 in all simulation settings.

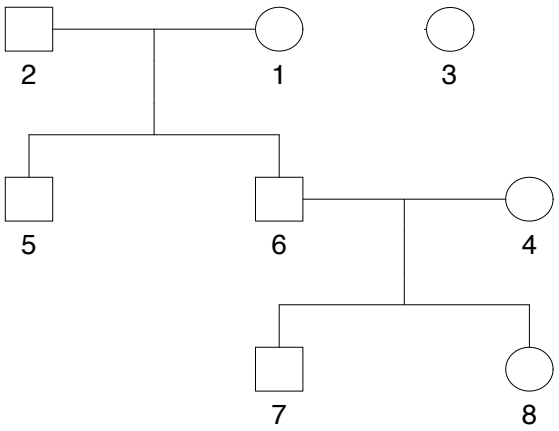
The simulated quantitative phenotypes had polygenic effects on both the X chromosome and the autosomes and an effect size of 0.05. Since the background autosomal and X chromosome polygenic effects are modeled as random effects, the effects were sampled from multivariate random normal distributions with mean zero and covariance matrices corresponding to the X chromosome and autosomal GRMs from the known pedigree structure. The error term was sampled from a normal distribution with mean zero and variance one, where all errors were independent between samples. With the simulated phenotypes, we performed association tests with both the causal SNP and null SNPs by fitting (1) the MLM-X model, (2) the simple MLM model, (3) the model including adjustment only for X chromosome random effects, and (4) a simple linear model that ignores all correlation among our samples.

3.2.8 Application to Subjects from the HCHS/SOL Study

The HCHS/SOL study incorporates a unique survey design [25] and includes baseline examinations on over 12,000 self-identified Hispanic/Latino individuals recruited from the Bronx, NY, Chicago, IL, Miami, FL and San Diego, CA [53]. The red blood cell count trait (RBC) considered for this study was measured at baseline. SNP genotype data was obtained from DNA extracted from blood and genotyping was done on an Illumina custom array, SOL HCHS Custom 15041502 B3, which is a superset of the Illumina Omni 2.5M array plus ancestry-informative custom SNPs. After quality [24] and informativeness filters, 2,232,944



(a) 'Male-centric' pedigree configuration



(b) 'Balanced' pedigree configuration

Figure 3.2: Pedigree structures considered for the simulation studies.

SNPs were used for analysis. The genotype data with quality control details are posted in the database of genotypes and phenotypes (dbGaP), accession numbers phs000880.v1.p1 and phs000810.v1.p1.

For association testing with the RBC trait, we excluded samples with known blood/lymph malignant tumor, bone cancer, pregnancy, or chronic kidney disease. We also excluded those who were currently undergoing chemotherapy, those with percentage blasts >5%, or percentage immature granulocytes >5%. We included fixed effects covariates of age, sex, recruitment center and autosomal and/or X chromosome ancestry eigenvectors. We included random effects of autosomal and/or X chromosome kinship, as well as household and block group membership. The latter two random effects were specifically included in the model because data were collected on each participant's city block and household membership. Thus, we have environmental correlation measures for all individuals and so these effects were modeled with random effects where the covariance matrices had entries of 1 where pairs of individuals share a household or live on the same block.

3.3 Results

3.3.1 Estimation of Relatedness on the X Chromosome in Simulated Samples

To investigate estimation of the pairwise X chromosome KC, varying numbers of X chromosome SNPs were simulated for one iteration of the 16-sample pedigree shown in Figure 3.1. The founder allele frequency of the SNPs was set at 0.4. Figure 3.3(a) shows the difference between the estimated and theoretical X chromosome KC for all sample pairs for increasing numbers of SNPs. The subplots show the results broken up by the composition of sex in the pair. Some relationships are underestimated but all are generally centered around the truth. The estimates of KC for all pairs and for female-male pairs have larger variances than the estimates of KC for female-female pairs or male-male pairs. With a larger number of SNPs, we are able to more accurately estimate the true X chromosome KC. We note from Figure 3.3 that the estimated KC is at most 0.06 away from the true value. The HCHS/SOL genotype data yielded 3,500 SNPs on the X chromosome after linkage disequilibrium pruning. In terms of the number of SNPs for which we have genotype information,

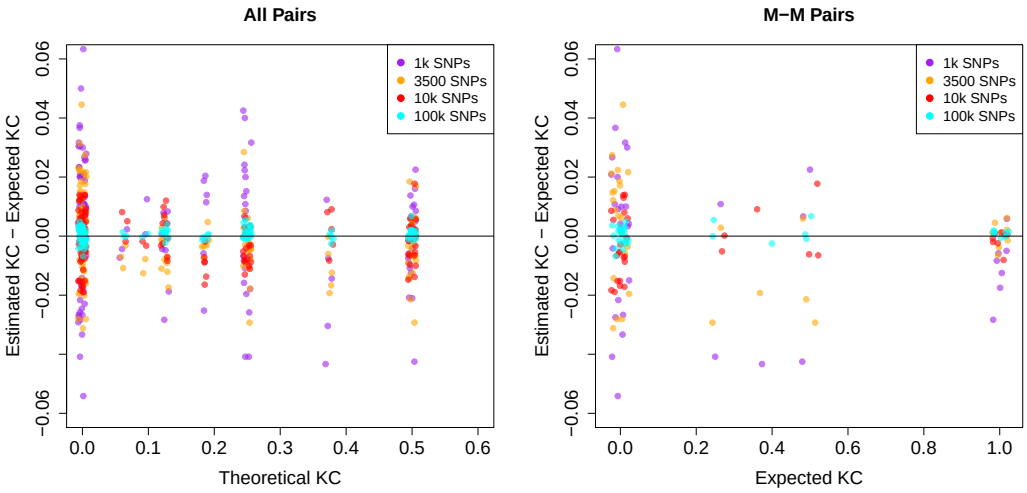
and after pruning based on LD, the results shown in orange in Figure 3.3 are most realistic. We conclude that we are sufficiently able to estimate the X chromosome KC from 3,500 independent, genotyped X chromosome SNPs.

3.3.2 Estimation of Variance Components in Simulated Samples

Phenotypes were simulated with polygenic effects across both the X chromosome and autosomes for a set of 8,000 samples. AI-REML [8] was used to estimate variance components for the MLM-X model, the simple MLM model, and the model including only X chromosome random effects. Samples were related through one of two eight-person, three-generation pedigrees, one that includes six males and two females (male-centric) and one that has an equal number of males and females (balanced). The pedigree structures are shown in Figure 3.2.

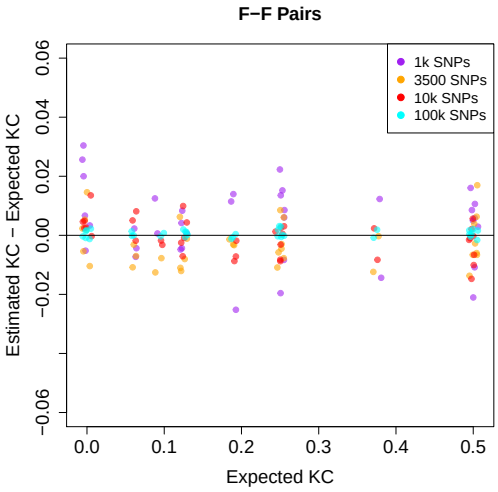
Figure 3.4 shows boxplots of the estimates of the three variance components from 10,000 independent simulations when fitting each of the three models under both pedigree configurations. Figure 3.4A and B show variance component estimates under the MLM-X model that adjusts for X chromosome and autosomal random effects under the 'male-centric' and 'balanced' pedigrees, respectively. For each simulated value of σ^2_X and σ^2_A, and under both pedigree configurations, the estimates are centered around the true value. When ignoring the X chromosome random effects and only fitting random effects on the autosomes, both pedigree configurations overestimate the total variance due to genetic effects (Figure 3.4C and D). The overestimation is approximately 30% when considering the male-centric pedigree, whereas the balanced pedigree only overestimates by about 5%. On the other hand, when ignoring the autosomal random effects and only fitting random effects on the X chromosome, both pedigree configurations underestimate the non-residual variance by approximately 40% (Figure 3.4E and F). The estimate of σ^2_X is less than the sum of the true values of σ^2_X and σ^2_A.

When misspecifying the correlation structure by including only one of the random effects in the model fit, the covariance that is being estimated attempts to account for the correlation structure as properly modeled as well as the correlation of the unmodeled ran-

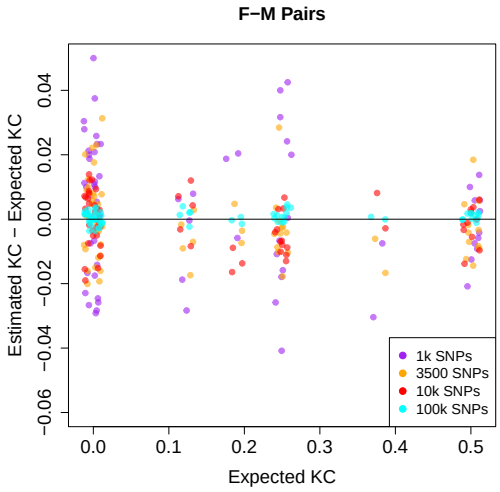


(a) All Samples

(b) Male-Male Pairs



(c) Female-Female Pairs



(d) Female-Male Pairs

Figure 3.3: The expected X chromosome KC versus the difference between the expected and estimated X chromosome KC, for all sample pairs and all sample pairs stratified by sex composition for the 16-person pedigree shown in Figure 3.1. The colors indicate the number of simulated X chromosome SNPs used to estimate the KC.

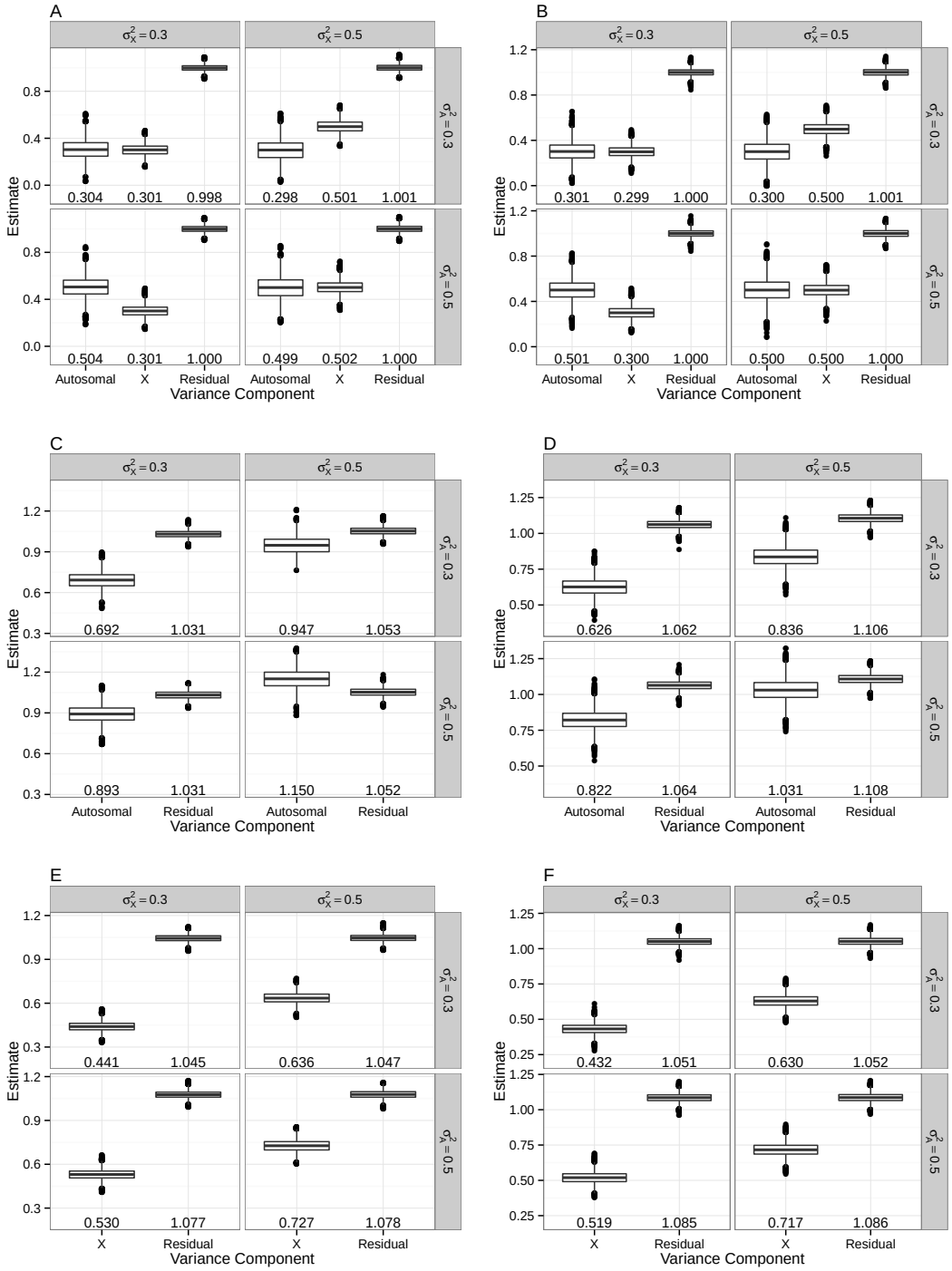


Figure 3.4: Estimates of variance components for 10,000 simulation iterations for two values each of σ^2_X and σ^2_A, and for $\sigma^2_\epsilon = 1$. Figures A, C, E correspond to the 'male-centric' pedigree whereas B, D, F are results from the 'balanced' pedigree. The first row are estimates under the MLM-X model, the second row shows estimates under the simple MLM model and the third row are results estimated under the model including only $\mathbf{g}_X$ random effects. The mean estimate is printed below each boxplot.

dom effect. Thus, depending on which pedigree setting and which random component(s) is included, we over- or underestimate our variance. For example, when only including $\mathbf{\Phi}_A$ in the male-centric pedigree setting, we overestimate the total variance since $\mathbf{\Phi}_A$ tries to scale up to match the covariance explained by $\mathbf{\Phi}_X$ (Figure 3.4C). On the other hand, when only including $\mathbf{\Phi}_A$ but in the balanced pedigree setting, our estimate of the variance is close to the truth because the difference in correlation values between $\mathbf{\Phi}_A$ and $\mathbf{\Phi}_X$ is not large under the balanced pedigree (Figure 3.4D). In both pedigree settings and when including X only, we underestimate the variance because $\mathbf{\Phi}_X$ is larger than $\mathbf{\Phi}_A$ so it is scaled down to try and incorporate the covariance structure explained by $\mathbf{\Phi}_A$ (Figures 3.4E and F).

3.3.3 Evaluation of Power and Type I Error

To assess the power and type I error rate of the MLM-X method, we performed association tests with simulated data. Independent autosomal and X chromosome genotypes for 8,000 individuals related through one of two known pedigree structures were simulated. For each iteration, a causal SNP was selected and a quantitative phenotype was generated with both autosomal and X chromosome random effects. Association tests were carried out with the causal SNP, as well as with non-causal SNPs. From the results, we can tabulate the proportion of false and true positives discovered with MLM-X, for varying effect sizes and significance levels. We compare the performance of MLM-X to the simple MLM model, the model that includes only polygenic effects on the X chromosome, and the simple linear model that ignores all correlation among the samples.

We tested association of one million null autosomal and one million null X chromosome SNPs with a quantitative phenotype from which we calculated the proportion of false positives for various significance levels (Table 3.2). We simulated a large number of markers to evaluate the type I error of the methods in a large-scale genotyping setting. Generally, GWAS are performed on the X chromosome only adjusting for autosomal genetic effects (simple MLM model). From our simulation studies, we observe nearly twice the amount of false positives expected in almost all simulation cases. In fact, the 95% confidence interval does not include the nominal level in these instances. In the opposite situation when we

Table 3.2: Type I error rate for various MLM approaches of one million simulations using the 'balanced' pedigree configuration. * implies type I error rates for which the 95% confidence interval does not include the nominal significance level.

			$\alpha = 1e-04$			$\alpha = 1e-05$		
	σ^2_X	σ^2_A	MLM-X	Simple MLM	X chr only	MLM-X	Simple MLM	X chr only
Autosomal	0.3	0.3	9.93e-05	1.03e-04	1.26e-04*	9.33e-06	8.00e-06	1.33e-05
	0.3	0.5	1.09e-04	1.13e-04	1.59e-04*	1.07e-05	1.00e-05	1.60e-05*
	0.5	0.3	1.08e-04	1.03e-04	1.26e-04*	1.33e-05	1.40e-05	1.60e-05*
	0.5	0.5	1.05e-04	1.03e-04	1.50e-04*	1.20e-05	1.20e-05	1.47e-05
X chr	0.3	0.3	1.14e-04	1.88e-04*	1.14e-04	1.19e-05	2.59e-05*	1.10e-05
	0.3	0.5	9.65e-05	1.66e-04*	8.66e-05	6.97e-06	1.29e-05	9.95e-06
	0.5	0.3	9.80e-05	2.19e-04*	1.05e-04	7.07e-06	2.12e-05*	5.05e-06
	0.5	0.5	1.07e-04	2.27e-04*	1.11e-04	1.19e-05	2.59e-05*	1.19e-05

test an autosomal SNP and only adjust for polygenic effects on the X, we observe increased false positives as well. However, the MLM-X model is properly calibrated in all simulation settings considered.

Power for the MLM-X model, the simple MLM model and the X chromosome only model is shown in Figure 3.5. We compared the false positive rate to the true positive rate for a sequence of significance levels ranging from 1×10^{-5} to 0.5. When testing either autosomal or X chromosome SNPs, MLM-X achieves similar power to the other methods considered under non-extreme variance settings. MLM-X has the highest power when testing both autosomal and X chromosome SNPs under extreme variance settings as compared to either method that only includes adjustment for one random effect. In particular, most analyses on autosomal genetic data are done only adjusting for the autosomal random effect. Additional power can be gained when we include adjustment for X chromosome random effects, even when testing an autosomal marker.

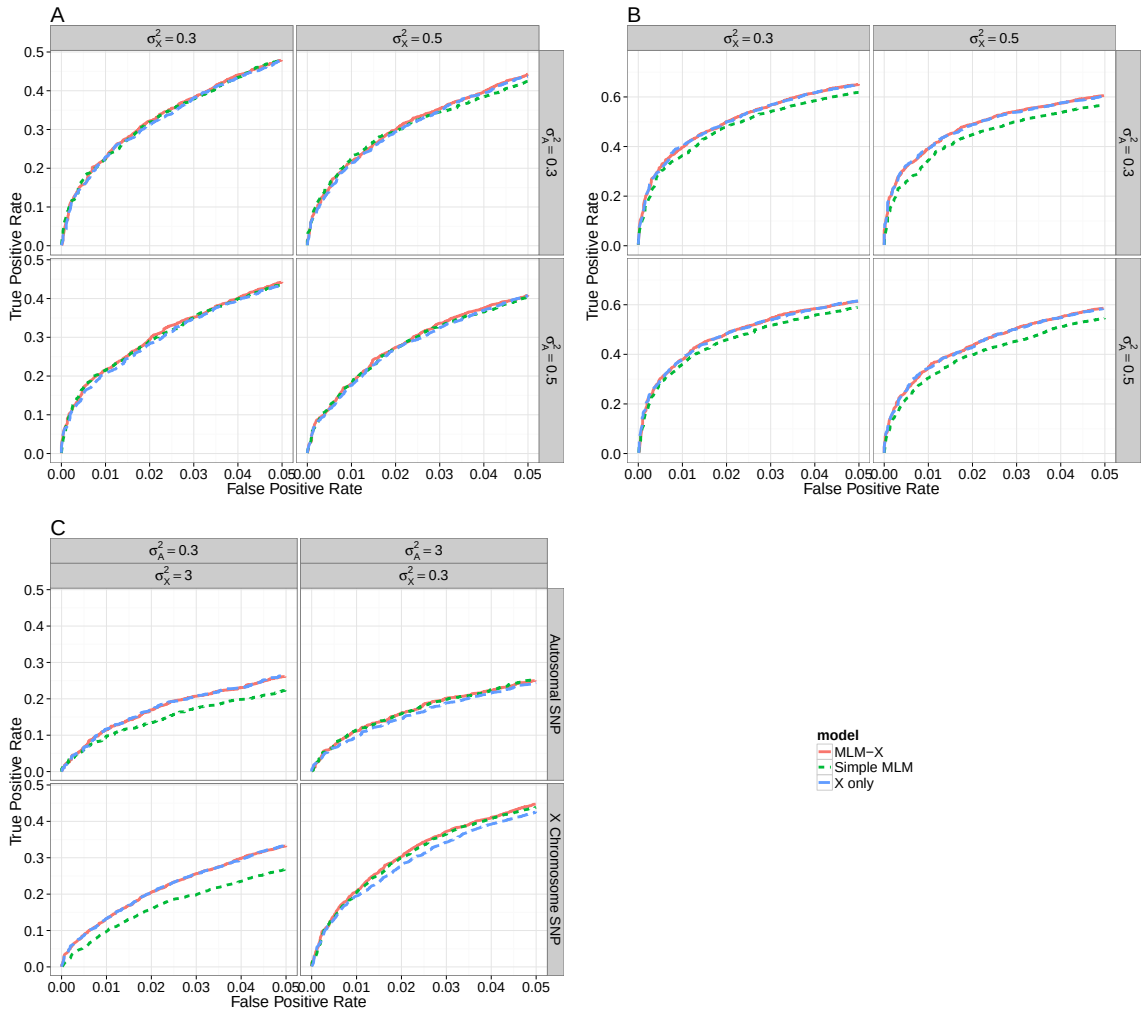


Figure 3.5: The proportion of true positives versus the proportion of false positives for one million simulation iterations. Results from the MLM-X model as compared to the simple MLM model and the model that only includes $\mathbf{g}_X$ random effects are shown. (A): testing an autosomal SNP for association and (B): testing an X chromosome SNP for association. (C) shows the results for testing under extreme variance settings.

3.3.4 Population Structure Estimation in HCHS/SOL Samples

It has been shown that population structure may confound association between a genetic marker and phenotype of interest [43, 16]. To control for population stratification in genetic studies, PCA is generally performed across a set of autosomal markers and a number of PCs are included in the association test as fixed covariates. In the MLM-X model, we included both autosomal and X chromosome-specific PCs as fixed effects to eliminate any possible confounding that may exist due to ancestry.

We estimated PCs using PC-AiR [10], which provides principal component estimates among related individuals. The autosomal PCA estimation among the HCHS/SOL samples was described in detail previously [9], so here we discuss estimation of X chromosome PCs. The X chromosome PCs were estimated using 3,600 observed X chromosome genotype markers, pruned for linkage disequilibrium (LD) in a sliding window fashion such that any pair of SNPs in the set has $r^2 < 0.32$. PC-AiR requires a set of unrelated individuals upon which the estimation of the related individuals' principal components is based. In the HCHS/SOL samples, we used a set of 10,287 unrelated individuals, where unrelated is defined up to fourth degree. We then projected the PC estimates onto the remaining 2,497 samples to obtain PC estimates for a total set of 12,784 individuals.

Figure 3.6 shows the first two principal components for the total set of 12,784 samples including relatives. Although a bit noisier than the autosomal PCA results [9], we detect large-scale continental structure using a relatively small set of SNPs solely from the X chromosome. PCs higher than the second do not display any structure among the HCHS/SOL samples (Figure 3.7). It is perhaps surprising that a sample of only X chromosome SNPs can identify large-scale continental ancestry among individuals from an admixed population. We advocate that when using MLM-X, PCs from both autosomal and X chromosome SNPs be included regardless of where the genetic marker being tested for association lies in the genome.

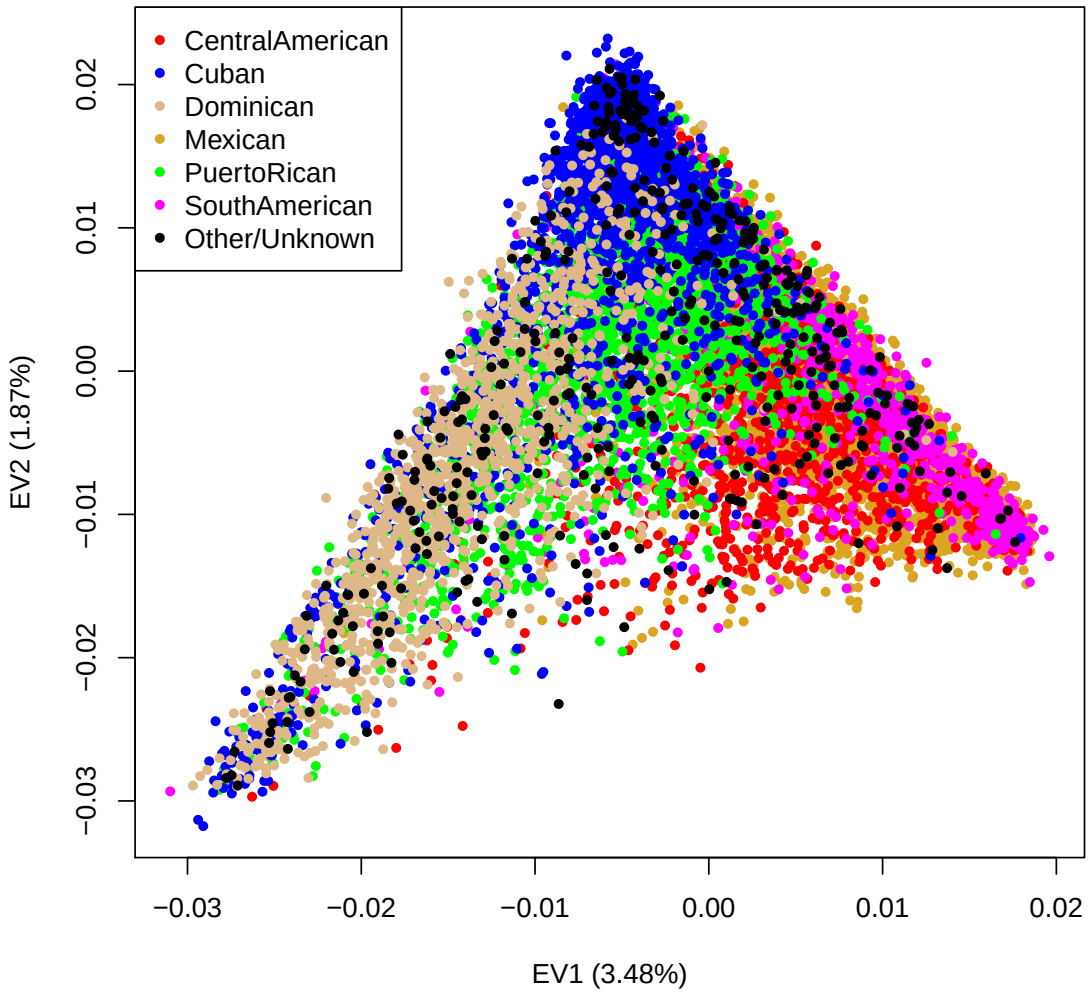


Figure 3.6: Principal components analysis results from PC-AiR estimated from a set of 10,287 unrelated individuals and a total set of 12,784 samples. A set of 3,600 LD-pruned X chromosome SNPs was used. Each sample is color-coded by the self-identified background group. The proportion variance explained by each of the PCs is shown in the axis labels.

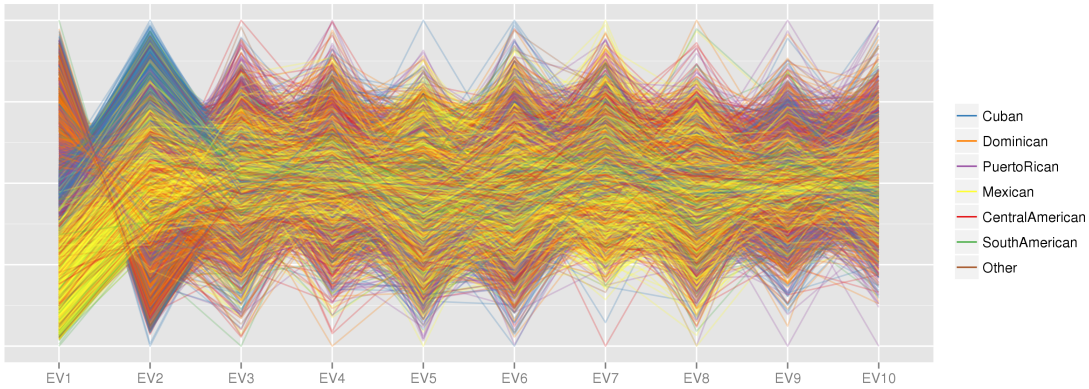


Figure 3.7: Parallel coordinates of 12,784 X chromosome eigenvectors 1-10, colored by the self-identified background group.

3.3.5 Relatedness Estimation in HCHS/SOL Samples

The pairwise X chromosome genetic relatedness matrix was estimated using PC-Relate [11], which provides accurate kinship estimates in the presence of population structure and is valid for X chromosome relatedness estimates with genotype coding of 0, 2 for males and 0, 1, 2 for females. We estimated $\mathbf{\Phi}_X$ for a set of 12,784 individuals in HCHS/SOL using 3,600 observed X chromosome genotype markers, pruned for LD as described in the previous section. PC-Relate requires principal components estimates which it uses to adjust for ancestry effects within individuals, aiding in the deconvolution of correlation due to ancestry rather than relatedness. Thus, we include the first two X chromosome PCs as estimated from PC-AiR (Figure 3.6). The corresponding pairwise autosomal kinship coefficients were estimated using the same set of 12,784 individuals and 153,470 SNPs pruned for LD across the autosomes. The autosomal PC-Relate estimates were adjusted for PCs 1-5 estimated from PC-AiR on the autosomes as described in detail in [9].

All pairs of study individuals were classified into relationships based upon the estimated autosomal kinship values. We compared $\widehat{\mathbf{\Phi}_X}$ to $\widehat{\mathbf{\Phi}_A}$ among all study samples related at fourth degree or higher as detected on the autosomes. Figure 3.8 shows boxplots of $\widehat{\mathbf{\Phi}_X}$ for each relationship type up to fourth degree, stratified by the sex composition of the

pairs. The $\widehat{\mathbf{\Phi}_X}$ value for parent-offspring relationships are the most separated among the sex composition of the pairs, and as the relationships grow more distant, the median $\widehat{\mathbf{\Phi}_X}$ becomes more similar between the pairs, regardless of their sex composition; this is what we expect (Table 3.1). It is perhaps unreasonable to group all pairs with a second degree (or more distant) relationship together, as the expected kinship depends on the type of relationship as well as whether the lineage is maternal or paternal. The variance for female-female pairs is smaller in all relationship types than for pairs that include at least one male. We expect this behavior, as the variance for male X chromosome genotypes is twice that of females. Male full sibling pairs have kinship coefficient estimates ranging from less than 0.25 to nearly 1. Full brothers have an average X chromosome kinship value of 0.5, however the true value can range from 0 to 1, depending on the rates of recombination and which X chromosome each brother inherits from his mother. We use $\widehat{\mathbf{\Phi}_X}$ as the empirical covariance matrix when adjusting for polygenic effects across the X chromosome.

3.3.6 Application to RBC Trait in HCHS/SOL Samples

We applied MLM-X to the red blood cell count (RBC) trait as measured at baseline in 12,502 individuals enrolled in the HCHS/SOL study. This sample size is smaller than the 12,784 samples for whom we obtained PC estimates since we include samples who have non-missing phenotype data for all measurements used in the association test. We compare the variance components estimates, calculated p-values and genomic control inflation factor [12] as computed from the MLM-X model to the simple MLM model which only includes adjustment for autosomal effects. A QC filtered [24] set of 43,868 X chromosome and 2,084,623 autosomal SNPs were tested for association. As described in Section 3.3.4, population structure was estimated across the autosomes using PC-AiR and separately using an extension of PC-AiR on the X chromosome, which yields principal components robust to the presence of relatedness among samples [10]. In both MLM-X and the simple MLM model, autosomal principal components 1-5 were included as fixed effects to account for population structure among the samples. In the MLM-X model, additional fixed effects of X chromosome principal components 1-2 were included. Pairwise genetic relatedness was estimated

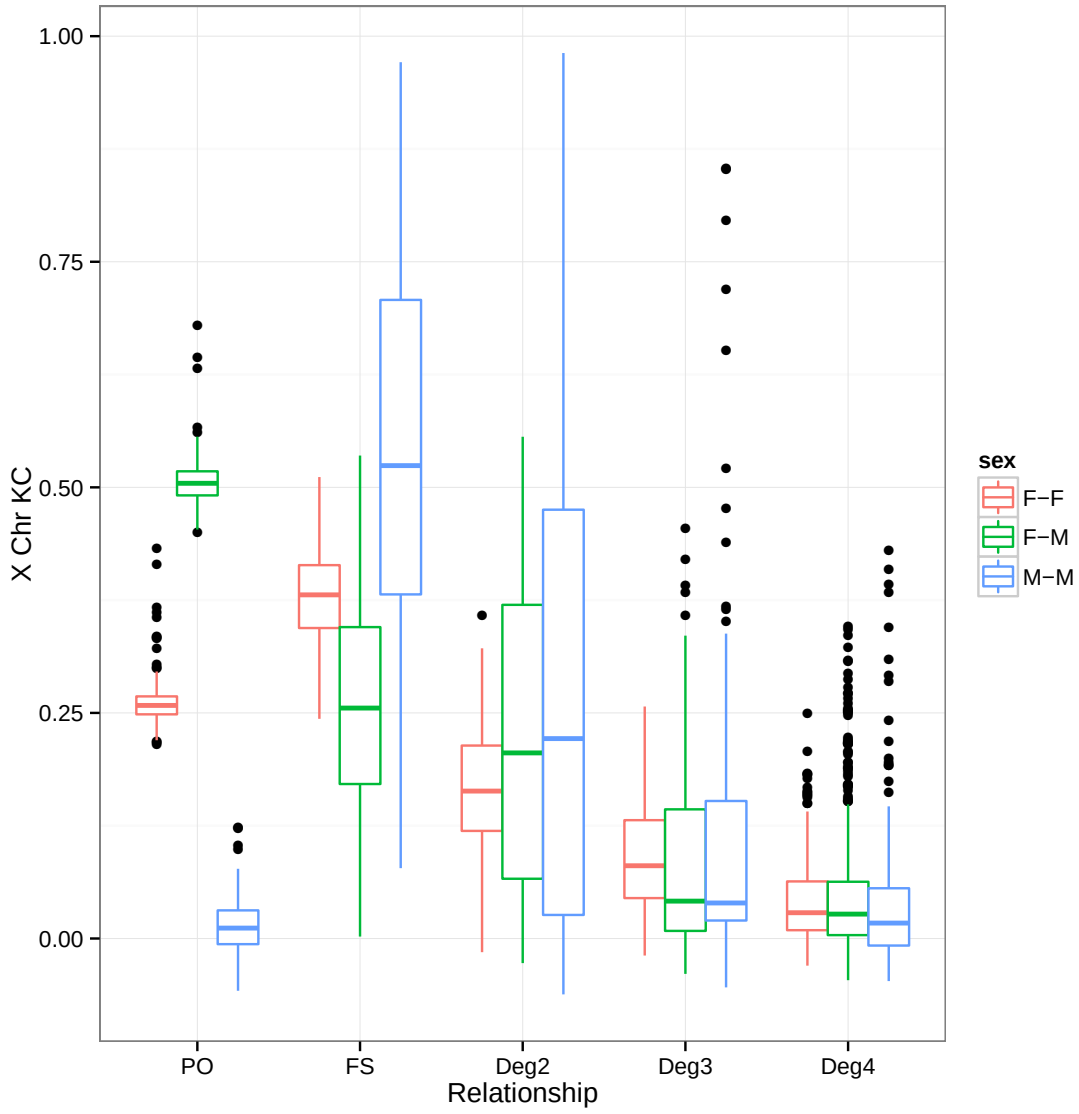


Figure 3.8: Boxplots of $\widehat{\Phi_X}$ estimated using PC-Relate with 3,600 LD-pruned X chromosome SNPs. Pairs of relatives up to fourth degree classified using the autosomes are shown stratified by the sex composition of each pair.

Table 3.3: Estimate (95% CI) of the proportion variance explained by each random effect, applied to red blood cell count measured in 12,502 HCHS/SOL study samples.

	MLM-X	Simple MLM
Block group	0.00396 (0, 0.0095)	0.00363 (0, 0.0090)
Household	0.0495 (0.0209, 0.0781)	0.0495 (0.0211, 0.0778)
Autosomal Kinship	0.285 (0.232, 0.338)	0.285 (0.232, 0.337)
X Chromosome Kinship	0.0294 (0.0137, 0.0450)	-
Residual	0.636 (0.578, 0.687)	0.662 (0.610, 0.715)

for all 12,502 study samples using PC-Relate [11] which estimates relatedness accurately in the presence of population structure. This procedure was described in Section 3.3.5. We obtain a GRM for the autosomes and the X chromosome separately.

The MLM-X model includes random effects for both autosomal and X chromosome polygenic effects, whereas the simple MLM model only includes autosomal polygenic effects. Due to the sampling scheme of the HCHS/SOL study as previously described [25], random effects for block group and household membership are both included in all models [9]. The HCHS/SOL study is unique in that we are able to model the portion of environmental correlation as captured through household and block group, along with the genetic correlation as detected through the observed genetic data.

The proportional variance (95% CI) estimated for each of the random effects is shown in Table 3.3 with accompanying Figure 3.9. The proportional variance of the RBC trait among HCHS/SOL samples that is explained by the block group, household and autosomal kinship remains essentially constant between the simple MLM and MLM-X model. The MLM-X model estimates approximately 3% of the variance of the RBC trait to be explained by polygenic effects on the X chromosome. Furthermore, the confidence interval for the X chromosome does not include zero. Interestingly, this variance is absorbed by the residual variance term in the simple MLM model, implying the variance detected by the X chromosome is unique and different from that detected using the autosomes.

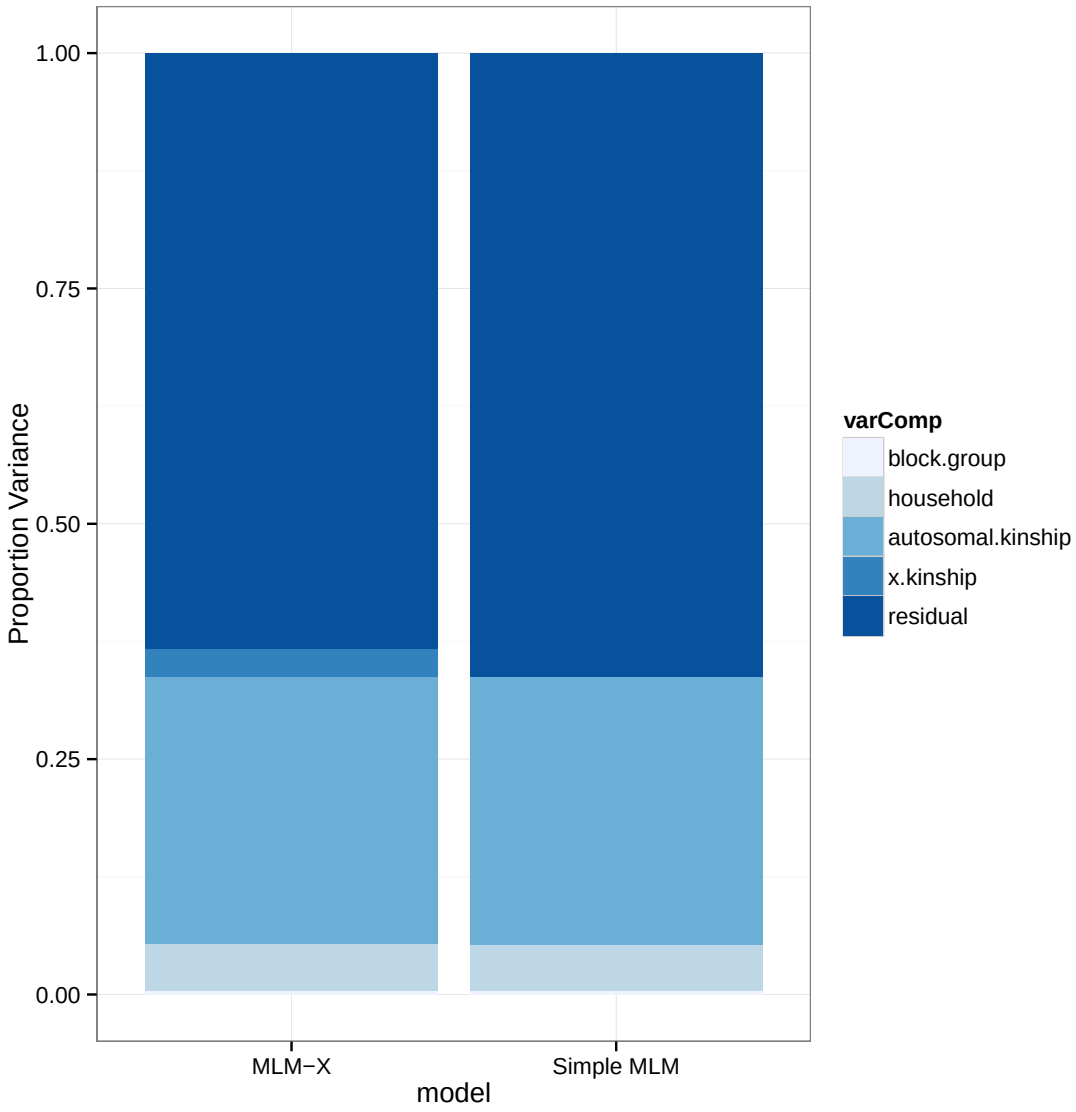


Figure 3.9: Bar charts of the estimated proportion variance for RBC in the HCHS/SOL samples from the MLM-X and simple MLM models.

We applied MLM-X and the simple MLM to test association with RBC genome-wide. Manhattan plots for p-values calculated from 2,128,491 SNPs using the MLM-X and the simple MLM model are shown in Figure 3.10, with accompanying QQ plot shown in Figure 3.11. The genomic control inflation factor [12], λ_{GC} is given in the QQ plot legend. The λ_{GC} value is similar between the two models, but slightly smaller for the MLM-X model (1.048 compared to 1.051). Although this difference is small, it implies that adjusting for X chromosome-specific effects, both fixed and random, may result in more properly controlled association tests. From both the Manhattan and QQ plots, we see that the models yield p-values that are much the same. Figure 3.12 compares the p-values from the simple MLM method to the MLM-X results across the set of 2,084,623 QC-filtered autosomal SNPs. The differences in the $-log_{10}$ p-values from the two methods have a mean of -7.5e-04 with standard deviation 0.37. The results are highly correlated ($corr = 0.997$) and there is not a loss of power due to inclusion of X chromosome effects when testing autosomal markers for association.

It is perhaps most interesting to compare the results from association testing with X chromosome SNPs. Manhattan plots of p-values for the X chromosome SNPs tested in the MLM-X and simple MLM models are shown in Figure 3.13. In both models, a peak in the Xq28 region, near the G6PD gene, yields highly significant p-values. Indeed, this gene has been previously published to be associated with RBC in a cohort of African Americans [7]. Figure 3.14 shows a QQ plot for the 43,868 X chromosome SNPs tested for association, colored by MLM-X and simple MLM results. The SNP with the most significant p-value was the same in both models, rs1050828, although the MLM-X p-value (1.82e-18) was less significant than that found from the simple MLM model (1.40e-19). However, as demonstrated from simulation studies, we believe the MLM-X model to be more properly calibrated.

The genomic control inflation factor [12], λ_{GC}, for the two models as calculated across the X chromosome SNPs is 1.116 and 1.043 for the simple MLM and MLM-X models, respectively (Figure 3.14); this difference is much larger than what we observed when considering all SNPs genome-wide. To address whether the inflation in λ_{GC} was due to the significant p-values in the G6PD gene, we re-calculated λ_{GC} for each model excluding the last 10 Mbs of the X chromosome. The resulting λ_{GC} values were 1.104 and 1.036 for the autosomal and

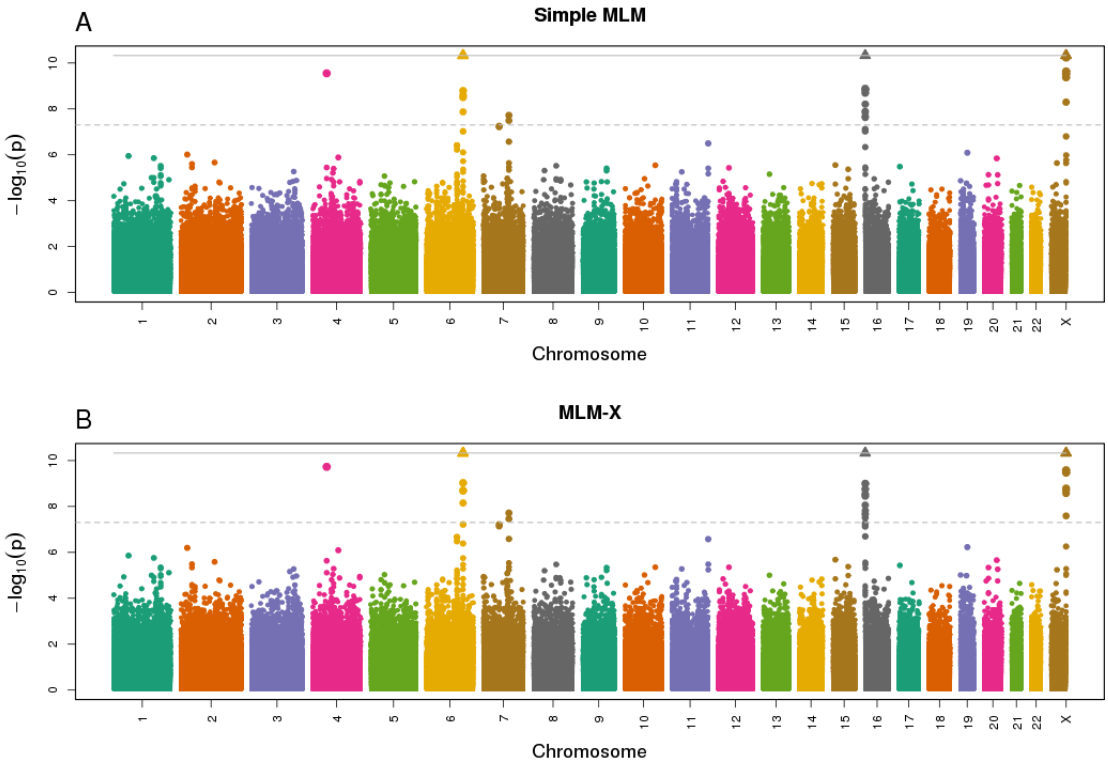


Figure 3.10: Manhattan plot of 2,128,491 SNPs tested for association with the RBC trait in HCHS/SOL samples using (A): the simple MLM model and (B): the MLM-X model.

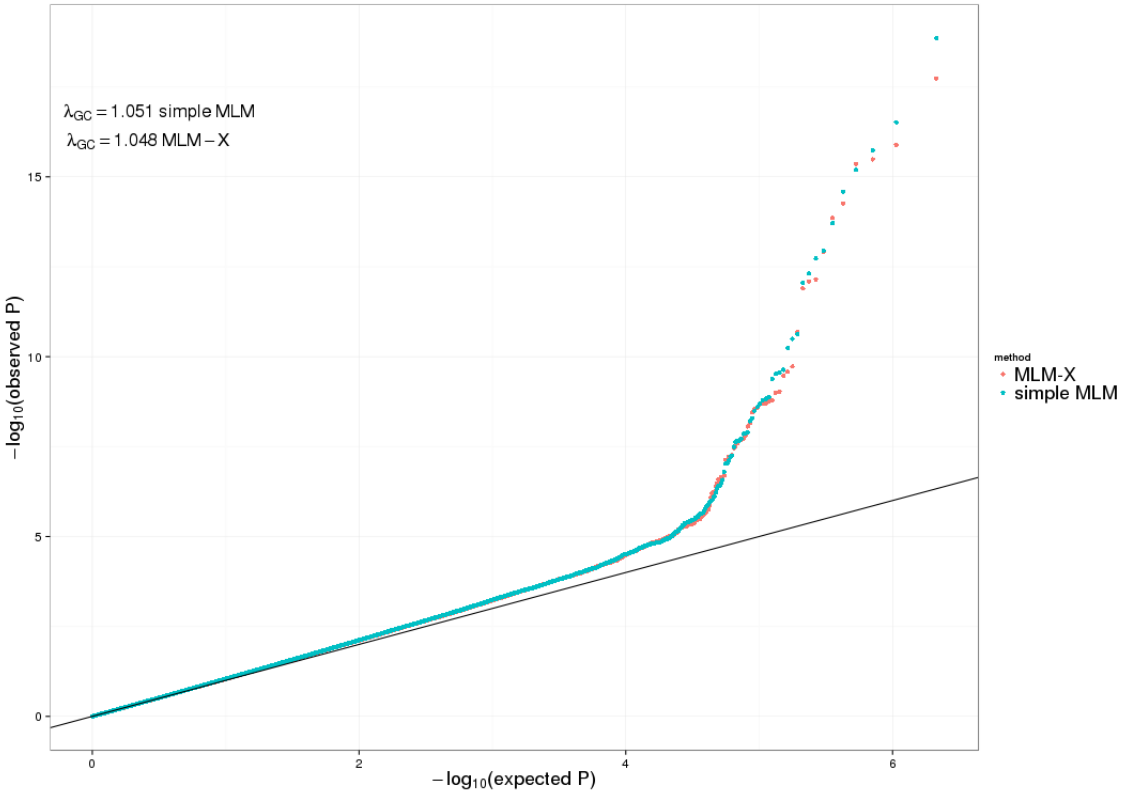


Figure 3.11: QQ plot of p-values from 2,128,491 SNPs, colored by MLM-X or simple MLM models. The solid black line indicates the x=y line.

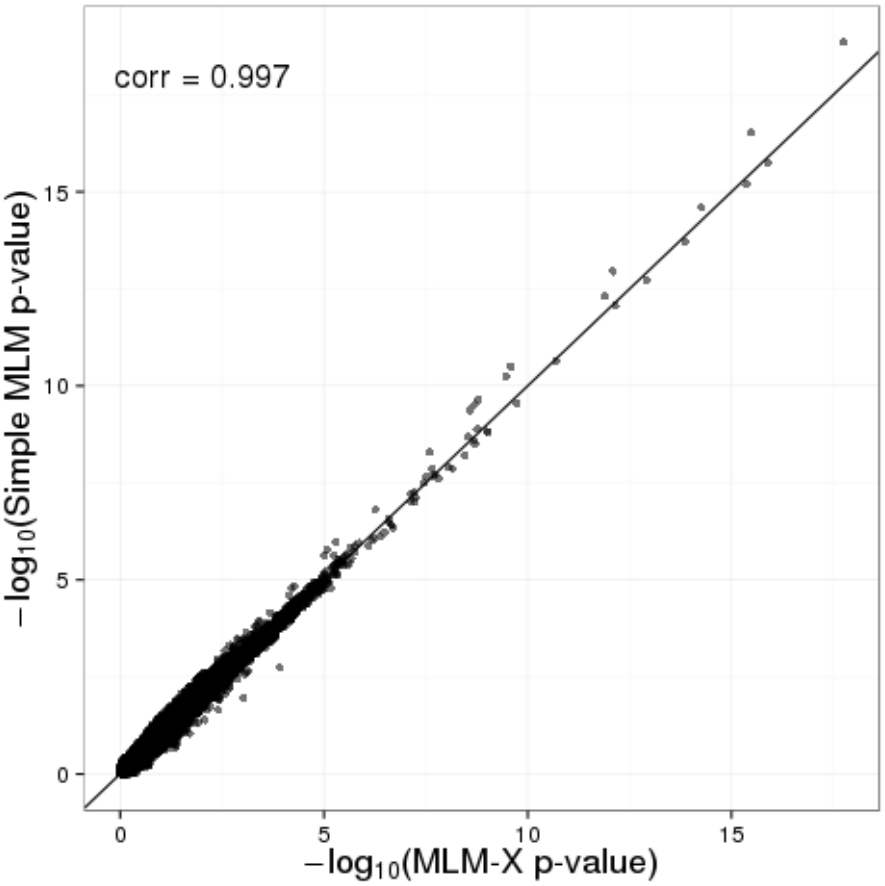


Figure 3.12: P-values from a set of 2,128,491 SNPs comparing the simple MLM model to the MLM-X model. The p-values have a correlation of 0.997.

MLM-X models, respectively. Thus, even after excluding the SNPs with the most significant p-values, we still observe a relatively high λ_{GC} from the simple MLM model. For the RBC trait in the HCHS/SOL samples, these results indicate that the MLM-X model yields a more properly calibrated test in terms of the genomic control inflation factor calculated from a set of QC-filtered X chromosome SNPs.

3.4 Discussion

The widespread use of MLMs for association testing in genetic studies has illuminated the strengths and limitations of this framework. The ability to efficiently and correctly account for correlation among study samples as measured through genetic and environmental factors in a computationally realistic timeframe is important for the application of MLMs to a set of markers across the genome. We have presented an approach that extends previous MLM models to simultaneously model correlation on the X chromosome and the autosomes to provide accurate and powerful association testing with not only X chromosome genetic markers, but across the genome. The MLM-X method can handle complex structure among study samples, such as relatedness or admixture. This structure is estimated empirically and is not required to be known or documented. Furthermore, additional random effects, such as shared environmental effects, can be easily added to the model.

In simulation studies that considered various pedigree structures, we demonstrated that MLM-X yields test results that are properly calibrated in terms of type I error. Furthermore, power gains can be achieved under certain settings. With an application to the RBC trait in the HCHS/SOL cohort, we were able to detect and generalize an association between a region on the X chromosome and the RBC trait previously published in a different admixed population. The MLM-X method yielded a more properly calibrated test in terms of λ_{GC} compared to the simple MLM method. With use of the MLM-X model for testing X chromosome SNPs, we can begin to alleviate the lack of published GWAS associations on the X chromosome.

Efforts should prioritize association testing with traits for which there are a large portion of unexplained heritability. Our results considering just one trait indicate that perhaps some missing heritability lies in genetic material on the X chromosome. As we perform more

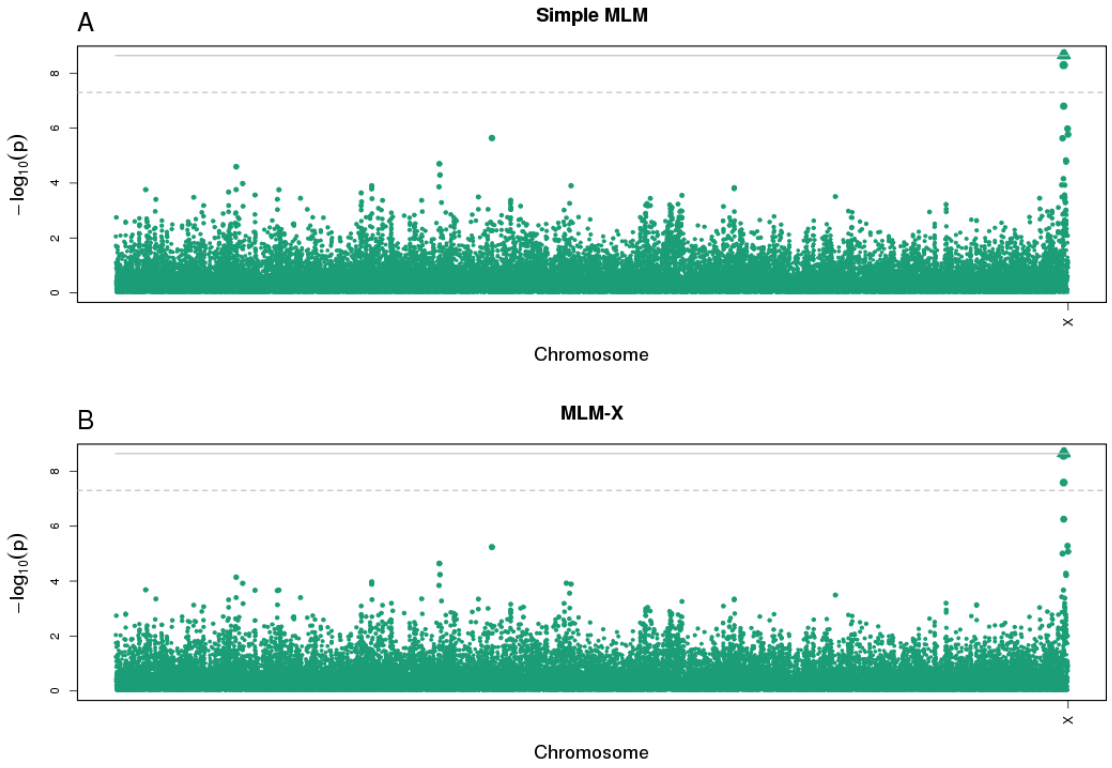


Figure 3.13: Manhattan plot of 43,868 X chromosome SNPs tested for association with the RBC trait in HCHS/SOL samples using (A): the simple MLM model and (B): the MLM-X model. The peak in the Xq28 region includes the G6PD gene.

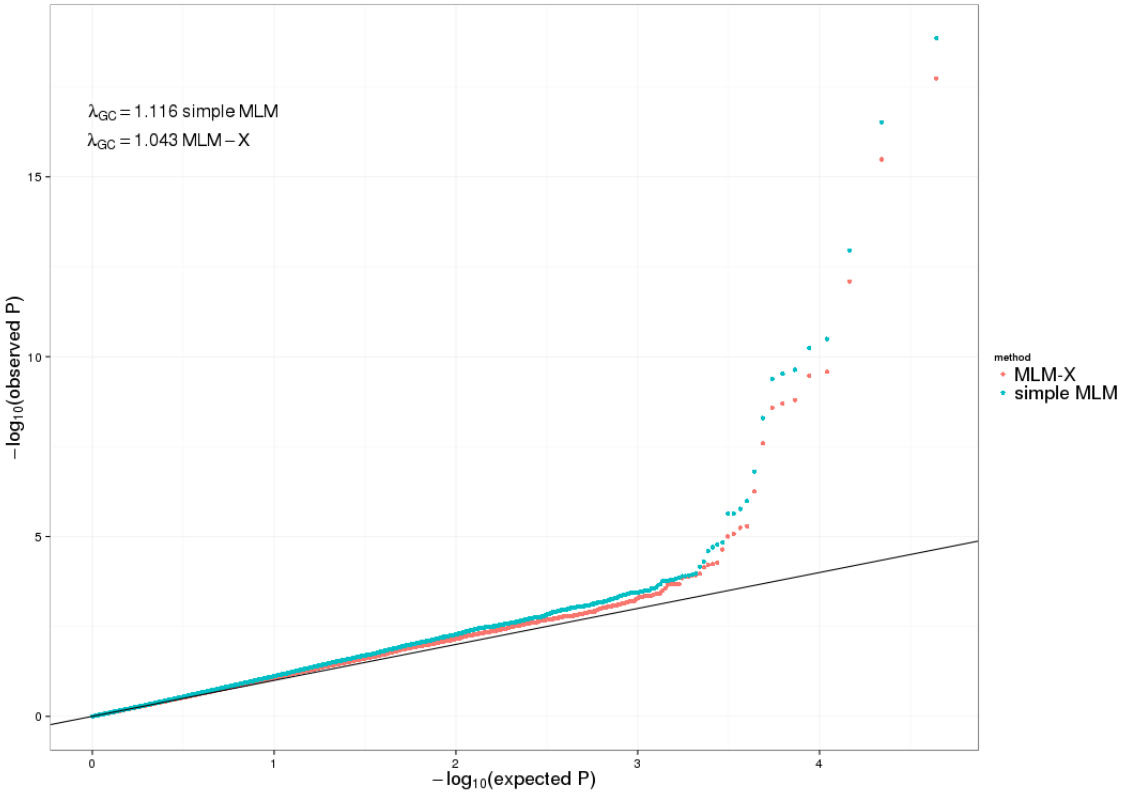


Figure 3.14: QQ plot of p-values from 43,868 X chromosome SNPs, colored by MLM-X or simple MLM models. The solid black line indicates the x=y line.

association tests considering the X chromosome properly, we can estimate the genetic heritability more accurately with the MLM-X method. Furthermore, with properly developed X chromosome association tests, we can identify gene variants associated with disease. The first step in discovering X chromosome genetic variants associated with complex disease is to ensure that we have a powerful test that is accurate.

By considering X chromosome SNPs, one begins to think about sex-specific effects in complex diseases. The MLM-X model assumes the random effects are the same for both males and females. Additional questions that address whether an interaction term of the polygenic effects across the X chromosome with sex should be included, or how we could examine sex-specific effects in the X chromosome by considering twin pairs, can be examined using MLM-X or extensions of the proposed model.

With the genotype coding of X chromosome SNPs at 0, 1, 2 in females and 0, 2 in males, we assume random X chromosome inactivation as well as additivity. To relax this assumption, X chromosome SNP coding could vary depending on the process of X chromosome inactivation or dominance effects. MLM-X can be extended to take into account other genotype codings which reflect varying biological processes on the X chromosome. However, we recognize the limitation that X chromosome inactivation can vary by tissue type, so it is important to have measurements from the correct tissue type. Future work can consider such an extension.

The basic idea of including a specific random effect to model polygenic effects on a particular chromosome could be extended to the autosomes. Estimates of variance would then be stratified by each chromosome, rather than a grouping across the autosomes, and could offer insight into regions of the genome that contribute a relatively large proportion of variance to a trait of interest. Additional work could investigate whether principal components should also be estimated by chromosome, or if accounting for population structure using autosomal PCs corrects for potential confounding properly.

Chapter 4

OMNIBUS GENE-BASED ASSOCIATION TESTING ACROSS THE GENOME IN SAMPLES WITH COMPLEX STRUCTURE

4.1 Introduction

As the cost of whole-genome sequencing continues to decrease, researchers are able to generate rich datasets that include both rare and common variants. Large-scale genomic studies often include individuals who are related through documented pedigrees and/or are cryptically related. Furthermore, genetic studies often include individuals from diverse populations with complex ancestry, such as admixed populations. In recent years, statistical methods have been proposed for association testing with rare, low frequency (and common) variants within a gene region of interest on the autosomes. Some methods allow for family structure, but only if this structure is known. A further limitation of existing methods is that they are not directly applicable to association testing of regions on the X chromosome. The ability to handle complex sample structure with empirical information is important for analysis of genetic variants across the genome.

Commonly used analysis methods for genome-wide association studies (GWAS) focus on testing relatively common variants, one at a time. Whole-genome and whole-exome sequencing assay rare and low frequency genetic variants, where here we define rare genetic variants as variants with minor allele frequency (MAF) less than 1% and low frequency variants as variants with MAF between 1% and 5%. Rare and low frequency variants have shown promise in helping to describe complex traits. The single variant association testing methods that are widely used with common variants have poor power for detecting rare variant associations. In addition, there are many more rare variants across the genome than common variants, resulting in a prohibitive multiple testing correction penalty for single variant association testing genome-wide. As a result, gene- or region-based region testing is usually employed in association studies that simultaneously test multiple variants with

MAF less than 5% [27].

Perhaps the most widely used method for testing rare and low frequency variants are burden tests, which aggregate variants in a region of interest into a summary 'dose' variable [33, 40, 39, 27, 46]. Burden tests are expected to have high power under an assumption that all variants are causal with equal effect sizes and direction of effect. If this is not the case, burden tests lose a substantial amount of power to detect an association [58, 41]. A further loss of power can occur when non-causal variants are included in the test [27]. Many extensions and variations of burden tests exist. A weighted burden test relaxes the assumption that the effect sizes of each variant are the same, where the pre-specified weights usually follow the belief that rarer variants yield larger effect sizes. For example, Madsen and Browning [33] proposed a weighted-sum test in which all variants are collapsed to one sum, with weights dependent on the MAF of the variant. However, it is most often unknown what the underlying genetic architecture of the trait of interest is. As a result, the burden test may not be the most powerful approach to testing a gene region of interest.

Kernel-based association testing methods are a powerful alternative to burden tests under some scenarios. Kernel-based testing methods use a variance component score test, which is more powerful than a burden test when the effect sizes of the variants are not in the same direction and when there are some variants in the test that are not causal [60]. The C-alpha test [41] is robust to variants with effects in the opposite direction, however permutations are required to assess the significance of an observed statistic. Furthermore, the C-alpha test was developed for binary traits and does not allow for adjustment of covariates. The SKAT [60] method, a generalization of the C-alpha test, computes a test statistic that is a weighted sum of single variant score statistics. SKAT computes a p-value through approximation and thus does not require expensive permutations to assess significance. An extension of SKAT to allow for known family structure, famSKAT, has been proposed [6].

Variance component tests are more powerful than burden tests in settings where the set of variants includes both negatively and positively associated variants and when some variants are non-causal [60]. On the other hand, in settings where all effects are of the same direction, similar magnitude, and there is a minimal number of non-causal variants, burden

tests achieve higher power than variance component tests [26, 58]. A method that exploits the strengths of these two approaches could provide the highest power, as the underlying genetic architecture is usually unknown in complex diseases.

'Optimal' or 'omnibus' tests have been proposed that are a convex combination of the burden and kernel-based tests [26, 19]. These methods capitalize on the fact that the burden and kernel-based association tests are most powerful in different scenarios and allow for adaptive weighting of the burden and kernel-based statistics, identifying the convex combination that yields the minimum p-value. An overall optimal p-value is then determined from the minimum p-value found from the combination of the two test statistics. Depending on the underlying genetic basis of the trait, the optimal test identifies the burden test, kernel-based test, or a linear combination of the two as the 'optimal' test for a particular gene region and trait of interest. Extensions of the SKAT method to SKAT-O [26] and the fam-SKAT method to MONSTER [19] take kernel-based association testing approaches and adapt them for the omnibus setting.

We propose an optimal **ke**rnel-based **a**ssociation **t**esting method in **s**tructured samples (KEATS-O) that is a gene-based test valid in the presence of complex sample structure. The KEATS-O method is properly calibrated for testing both autosomal and X chromosome gene regions and the optimal framework allows for results that are robust to the underlying genetic architecture of the trait of interest. We use two empirical genetic relatedness matrices, one for the autosomes and one for the X chromosome, providing accurate adjustment of both documented and cryptic relatedness. Our approach is flexible and allows for inclusion of any number of random effects in the model, such as known environmental correlation.

We show via simulation that KEATS-O is properly calibrated in terms of type I error and yields high power when testing various configurations of causal variants on either the X chromosome or the autosomes. We apply KEATS-O to a large-scale genetic study of Hispanic/Latino individuals from the U.S. and consider association of red blood cell count (RBC). Initially, we tested SNP genotype data mapped to 15 gene regions previously published to be associated with RBC, none of which were specifically studies focusing on individuals with Hispanic/Latino background. We were able to generalize association with six candidate genes. Finally, we apply KEATS-O to SNP genotypes mapped to gene

regions across the genome and identify six genes associated with RBC in the set of Hispanic/Latino individuals that reach genome-wide significance. One of the six genes found from the genome-wide analysis is not identified with single variant association testing (Chapter 3).

4.2 Methods

We first outline the burden test and variance component score test, then we present the omnibus test, which is a linear combination of the burden and variance component score tests. A bold uppercase letter symbolizes a matrix while a bold lowercase letter indicates a vector. In what follows, we indicate the matrix of genotype values as $\mathbf{G}$ for n samples and q variants of interest. We denote the matrix of fixed effect covariates for n samples, including the intercept, as $\mathbf{X}$. The pairwise kinship coefficients across the X chromosome and the autosomes are estimated as previously described (Section 3.2.4) and symbolized as $\mathbf{\Phi_X}$ and $\mathbf{\Phi_A}$, respectively.

4.2.1 The KEATS-bt Method

The KEATS burden test (KEATS-bt) method fits a mixed linear effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \gamma \sum_{j=1}^{q} \sqrt{w_j}\mathbf{G}_j + \mathbf{g}_A + \mathbf{g}_X + \boldsymbol{\epsilon} \tag{4.1}$$

where $\mathbf{G}$ is a matrix of observed genotypes for n samples and q variants of interest, $\mathbf{X}$ holds fixed effect covariates, $\mathbf{w} = \{w_1, \ldots, w_q\}$ is a prespecified vector of weights, where we assume γ follows an arbitrary distribution with mean zero and correlation matrix $\tau\mathbf{W}$, $\mathbf{g}_A \sim \mathcal{N}(0, \sigma_A^2\mathbf{\Phi_A})$, $\mathbf{g}_X \sim \mathcal{N}(0, \sigma_X^2\mathbf{\Phi_X})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2\mathbb{I})$, $\mathbf{\Phi_X}$ holds the pairwise kinship coefficients across the X chromosome and $\mathbf{\Phi_A}$ holds the pairwise kinship coefficients across the autosomes. Here, we adapt the weighted-sum test proposed by Madsen and Browning [33] by including additional random effects for polygenic effects across the autosomes and X chromosome. We use the score test statistic used to assess the null hypothesis $H_0 : \gamma = 0$

$$Q_{\text{KEATS-bt}} = \left((\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T \widehat{\mathbf{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{1/2}\mathbf{1}\right)^2 \tag{4.2}$$

where $\mathbf{1}$ is a vector of ones of length q. This test assumes one common γ parameter for all variants in the region and that all variants are causal in the same direction. Under the null hypothesis, $Q_{\text{KEATS-bt}} \sim \chi^2_1$.

4.2.2 The KEATS Method

The KEATS method is a kernel-based association method allowing for complex sample structure. We use the variance component score test that fits the model for q variants of interest and n samples

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{g}_A + \mathbf{g}_X + \boldsymbol{\epsilon} \tag{4.3}$$

where we assume γ has mean zero and covariance $\tau\mathbf{W}$, $\mathbf{g}_A \sim \mathcal{N}(0, \sigma^2_A\boldsymbol{\Phi}_{\mathbf{A}})$, $\mathbf{g}_X \sim \mathcal{N}(0, \sigma^2_X\boldsymbol{\Phi}_{\mathbf{X}})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_\epsilon\mathbb{I})$. The matrix of observed genotype values for q genetic variants in a region of interest is denoted $\mathbf{G}$, $\mathbf{W}$ is a prespecified, diagonal $q \times q$ matrix of variant weights, and the matrices of pairwise kinship coefficients across the autosomes and X chromosome are $\boldsymbol{\Phi}_{\mathbf{A}}$ and $\boldsymbol{\Phi}_{\mathbf{X}}$, respectively. The null hypothesis is $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_q = 0$. We can use a test with q degrees of freedom to test that the vector of length q is zero. Equivalently, we can rewrite the null hypothesis $H_0 : \tau = 0$ with the alternative hypothesis $H_1 : \tau > 0$. We recognize that since τ is a variance, the proposed test places the parameter of interest τ on the boundary. Further, note we do not impose distributional assumptions on the distribution of $\boldsymbol{\gamma}$, only that it has mean zero and variance $\tau\mathbf{W}$. The weight matrix $\mathbf{W}$ reflects the relative contribution of each weight to the score statistic, i.e. the larger the weight, the larger the variant will contribute to Q_{KEATS}. To test the null hypothesis, we use the variance component score statistic

$$Q_{\text{KEATS}} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{M}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{4.4}$$

where $\mathbf{M} = \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}^{-1}$. The $\mathbf{M}$ matrix is the 'kernel' matrix and measures the genetic similarity between all pairs of subjects in the region through the q markers being tested.

The variance of a quantitative trait $\mathbf{y}$ is

$$\begin{aligned} \text{var}(\mathbf{y}) &= \boldsymbol{\Sigma} && (4.5) \\ &= \tau\mathbf{G}\mathbf{W}\mathbf{G}^{\mathbf{T}} + \sigma^2_A\boldsymbol{\Phi}_{\mathbf{A}} + \sigma^2_X\boldsymbol{\Phi}_{\mathbf{X}} + \sigma^2_\epsilon\mathbb{I} && (4.6) \end{aligned}$$

Under the null hypothesis, $Q_{\text{KEATS}} \sim \sum_{i=1}^{p} \lambda_i \chi^2_{1,i}$ where λ_i are the p non-zero eigenvalues of $(\widehat{\boldsymbol{\Sigma}} - \widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}})^{1/2}\mathbf{M}(\widehat{\boldsymbol{\Sigma}} - \widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}})^{1/2}$ and $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$. The full derivation of the Q_{KEATS} score statistic and its distribution is outlined in Appendix B.

4.2.3 The KEATS-O Test

For a fixed value of ρ, we define the KEATS optimal (KEATS-O) statistic to be a convex combination of Q_{KEATS} and $Q_{\text{KEATS-bt}}$

$$T_\rho = (1-\rho)Q_{\text{KEATS}} + \rho Q_{\text{KEATS-bt}} \tag{4.7}$$

$$= (\mathbf{y} - \mathbf{X}\widehat{\beta})^T\mathbf{K}_\rho(\mathbf{y} - \mathbf{X}\widehat{\beta}) \tag{4.8}$$

where $\mathbf{K}_\rho = \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{1/2}\mathbf{R}_\rho\mathbf{W}^{1/2}\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}^{-1}$ and $\mathbf{R}_\rho = (1-\rho)\mathbb{I} + \rho\mathbf{1}\mathbf{1}^T$. We do not assume that our effects $\boldsymbol{\gamma}$ follow any particular distribution but do require that $\boldsymbol{\gamma}$ has mean zero and covariance $\tau\mathbf{R}_\rho$. With this covariance matrix, we assume the variants of interest follow an exchangeable correlation structure where all variants have correlation ρ. We notice that T_ρ is simply a kernel-based association test as we saw in Section 4.2.2 with a different kernel matrix that is a function of ρ. Under the null hypothesis $H_0 : \tau = 0$, $T_\rho \sim \sum_{i=1}^{p} \lambda_i \chi^2_{1,i}$ where λ_i are the p non-zero eigenvalues of $(\widehat{\boldsymbol{\Sigma}} - \widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}})^{1/2}\mathbf{K}_\rho(\widehat{\boldsymbol{\Sigma}} - \widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}})^{1/2}$. It is clear to see that when $\rho = 0$, T_ρ collapses to the Q_{KEATS} statistic (Equation 4.4), as $R_\rho = \mathbb{I}$ in this case. When $\rho = 1$, T_ρ collapses to the $Q_{\text{KEATS-bt}}$ statistic (Equation 4.2), as $R_\rho = \mathbf{1}\mathbf{1}^T$.

To perform the KEATS-O test, we calculate T_ρ for $\rho \in [0, 1]$ and corresponding p-values p_ρ. Then, we define the KEATS-O test statistic to be

$$Q = \min_{0 \le \rho \le 1} p_\rho \tag{4.9}$$

To find the p-value for Q, we integrate

$$1 - \int F(\delta(x)|\lambda) f(x|\chi^2_1) dx \tag{4.10}$$

where $F(\cdot|\lambda)$ is the distribution function of $\sum_{i=1}^{p} \lambda_i \chi^2_{1,i}$ as described above and $f(\cdot|\chi^2_1)$ is a χ^2_1 density. The full derivation of Equation 4.10 can be found in Appendix C.

In practice, we test prespecified ρ values ranging from 0 to 1. To find the p-value associated with the KEATS-O test statistic Q, we first calculate the KEATS-bt and KEATS

test statistics, and for each ρ value under consideration, record T_ρ. We then find the p-values associated with each T_ρ using Kuonen's saddlepoint method [23]. This method was chosen over Davies' method, as Davies' method has been shown to have less accuracy when estimating a very small p-value [6]. From these values, we can identify the KEATS-O test statistic Q. Based on Q, we calculate the eigenvalues of $(\widehat{\boldsymbol{\Sigma}} - \widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}})^{1/2}\mathbf{K}_\rho(\widehat{\boldsymbol{\Sigma}} - \widehat{\mathbf{P}}\widehat{\boldsymbol{\Sigma}})^{1/2}$ and the non-centrality parameter δ. Then, we have all components needed to calculate the one-dimensional integral shown in Equation 4.10.

4.2.4 Simulation Studies

To assess the power and type I error of the proposed KEATS-O method, we performed simulation studies under various parameter settings. We compared four approaches: KEATS-O, KEATS-O including only the X chromosome random effect (KEATS-O X only), MONSTER, and SKAT-O. Note MONSTER is identical to the KEATS-O method including only the autosomal random effect. The burden tests were fit using linear mixed effects models where the variants are weighted according to the prespecified weight matrix $\mathbf{W}$ that follow Wu's weighting scheme [60] where $\sqrt{w_j} \sim \text{Beta}(MAF_j; 1, 25)$. With this weighting scheme, variants with a lower MAF will get higher weight and vice versa. Other weighting schemes can be easily implemented, although Wu weights are commonly used and are feasible biologically. Previous studies have found no attributable difference to performance based upon weighting schemes [19, 26].

We simulated gene regions of either 20 or 50 variants, and considered variants across both the autosomes and the X chromosome. Minor allele frequency (MAF) for each variant was sampled from a Uniform(0.005, 0.05) distribution. We simulated sets of samples under three relatedness settings: (1) entirely unrelated, (2) related through an 8-person pedigree with equal number of males and females, where pairwise KC values on the X chromosome and autosomes are equal, on average, and (3) related through an 8-person pedigree with more males than females, where the pairwise KC values on the X chromosome are approximately twice the KC values on the autosomes, on average. Pedigree structures (2) and (3) are shown in Figure 3.2.

The linkage disequilibrium (LD) structure of the variants followed an AR-1 structure where neighboring variants have LD 0.5. To simulate markers that follow this specified LD pattern, we use the algorithm as outlined in HapSim [38], which simulates a multivariate Bernoulli variable such that each variable has a marginal distribution according to the specified minor allele frequency and the correlation between a pair of variables is exactly the specified LD between a pair of variants.

Finally, we simulated phenotypes with differing heritability. We simulated phenotypes with all four configurations of $\sigma^2_A, \sigma^2_X \in \{0, 0.5, 0.6, 1\}$ and $\sigma^2_\epsilon = 1$. We combined the values such that phenotypes were simulated with heritabilities of either $0, 1/2, 0.6$ or $2/3$. We can assess the performance of each method when ignoring a component with a relatively large proportion of genetic variance on the trait. Furthermore, we can assess the impact of each method when testing an autosomal or X chromosome region for association, whether ignoring or accounting for the main components of variance.

To estimate the power of our proposed method, we simulated effect sizes and direction of effects under various percentages. We varied the proportion of variants positively/negatively/not causal to follow these percentages: 10/0/90, 20/0/80, 50/0/50, 5/35/60, 15/25/60, 20/20/60. The causal variants were selected at random from each region with effect sizes calculated as $0.2|\log_{10}(\text{MAF})|$, where we multiply this by -1 for those variants chosen to be negatively correlated. To find the optimal ρ parameter, we did a grid search of 11 values equally spaced between 0 and 1: 0, 0.1, 0.2, ..., 0.9, 1. Previously published studies have shown that using a finer grid does not improve method performance, in terms of power or type I error [19].

4.2.5 Application to Subjects from the HCHS/SOL Study

The HCHS/SOL study incorporates a unique survey design [25] and includes baseline examinations on over 12,000 self-identified Hispanic/Latino individuals recruited from the Bronx, NY, Chicago, IL, Miami, FL and San Diego, CA [53]. The red blood cell count (x10e12) trait (RBC) considered for this study was measured at baseline. SNP genotype data was obtained from DNA extracted from blood and genotyping was done on an Illumina custom array, SOL HCHS Custom 15041502 B3, which is a superset of the Illumina Omni 2.5M

array plus ancestry-informative custom SNPs. After quality [24] and informativeness filters, 2,232,944 SNPs are used for analysis. The genotype data with quality control details are posted on dbGaP (accession numbers phs000880.v1.p1 and phs000810.v1.p1). Population structure and relatedness were estimated in the HCHS/SOL samples as described in Sections 3.3.4 and 3.3.5, respectively.

4.3 Results

4.3.1 Assessment of Type I Error

To assess the type I error rate of KEATS-O, we simulated phenotypes under the null hypothesis of no association. Sets of 2,000 samples were related under pedigree structure (2), which corresponds to an 8-person pedigree with equal number of males and females. Phenotypes had polygenic effects across the autosomes and X chromosome of $\sigma^2_A \in \{0, 1\}$ and $\sigma^2_X \in \{0, 1\}$. We consider association testing with both autosomal and X chromosome variants. In each case, either 20 or 50 variants were simulated with MAF$\sim$Uniform(0.005,0.05) and AR-1 LD structure with adjacent variants having LD 0.5. For each scenario, 100,000 replicates were performed.

Figure 4.1 shows the type I error rate when testing gene regions with 20 variants. When testing autosomal gene regions, KEATS-O is slightly conservative. Unsurprisingly, SKAT-O yields an excess of false positives when σ^2_A, σ^2_X are non-zero. However, when there is no heritability of the trait, but with relatedness among sampled individuals, ignoring the relatedness and using SKAT-O does not yield excess false positives. When testing X chromosome gene regions, KEATS-O and KEATS-O X only are properly calibrated. However, MONSTER has a high type I error rate when σ^2_X is non-zero. Again, we see that SKAT-O is properly calibrated when σ^2_A, σ^2_X are zero, but identifies many false positives when these variances are non-zero.

The distributions of the p-values from the four methods, stratified by autosomal and X chromosome gene regions, for the scenario where $\sigma^2_A = 1$ and $\sigma^2_X = 1$ are shown in Figure 4.2. KEATS-O has p-value distributions that are inflated towards one, providing an explanation of why the test appears to be slightly conservative. The p-value distribution of SKAT-O

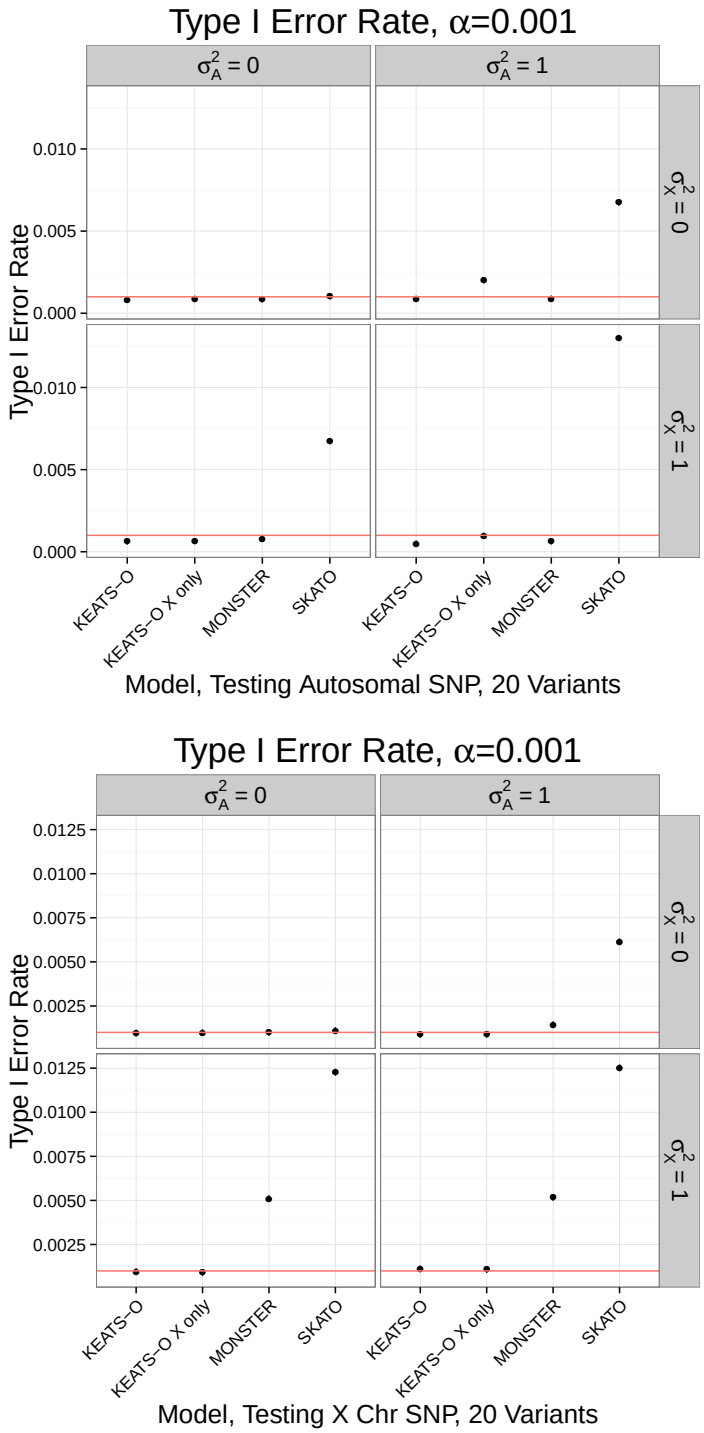


Figure 4.1: Type I error rate calculated when testing genes with 20 variants using KEATS-O, KEATS-O using only X chromosome random effect, MONSTER and SKAT-O. Both autosomal and X chromosome regions were considered. Samples are related through pedigree configuration (2), where 8-person pedigrees with equal number of females and males are simulated. The solid red line indicates the nominal type I error rate.

is highly skewed towards lower p-values, lending an explanation to why SKAT-O yields an inflated type I error rate. MONSTER is also slightly inflated towards one when testing autosomal SNPs. Strikingly, when testing an X chromosome SNP, the p-value distribution of MONSTER looks like that of SKAT-O. This is most likely because the covariance structure on the X chromosome is 'wrong,' so there is still unadjusted covariance in the data.

We consider the type I error rate when the samples are related under the same pedigree (2) configuration but when testing a larger set of 50 variants in each gene region. Figure 4.3 shows the type I error calculated when the gene regions include more variants. The tests are still properly calibrated, but KEATS-O is no longer conservative when testing autosomal SNPs. This implies that when testing a larger set of variants, although they are non-causal, we are able to more accurately account for the correlation among our samples and KEATS-O performs better in terms of type I error. We observe the same patterns with SKAT-O as we did with 20 variant genes; there are inflated false positives when testing both autosomal and X chromosome variants with $\sigma_A^2 = \sigma_X^2 = 0$. Furthermore, MONSTER remains inflated when testing X chromosome variants when $\sigma_X^2 = 1$.

Figure 4.4 shows the histogram of the optimal ρ parameter chosen for the 100,000 null simulations under pedigree configuration (2) and when testing 50 variants in each gene region. When testing either autosomal or X chromosome genes, nearly half the simulations choose $\rho = 0$, which corresponds to performing the kernel-based association test. The remaining simulation iterations yielded a ρ value distributed approximately uniformly between the 0.1 and 1 limits. We recall that under the null hypothesis, ρ is not identifiable. As a result, the choices of ρ make very little difference and tend to drift to the edge of the parameter space. These results are similar to those seen in simulation studies published in MONSTER [19].

4.3.2 Evaluation of Power

To evaluate the power of KEATS-O, we simulated variants with non-zero effect sizes. We considered simulations in which the proportion of positive/negative/non causal variants were: 10/0/90, 20/0/80, 50/0/50, 5/35/60, 15/25/60, 20/20/60. Following the assumption

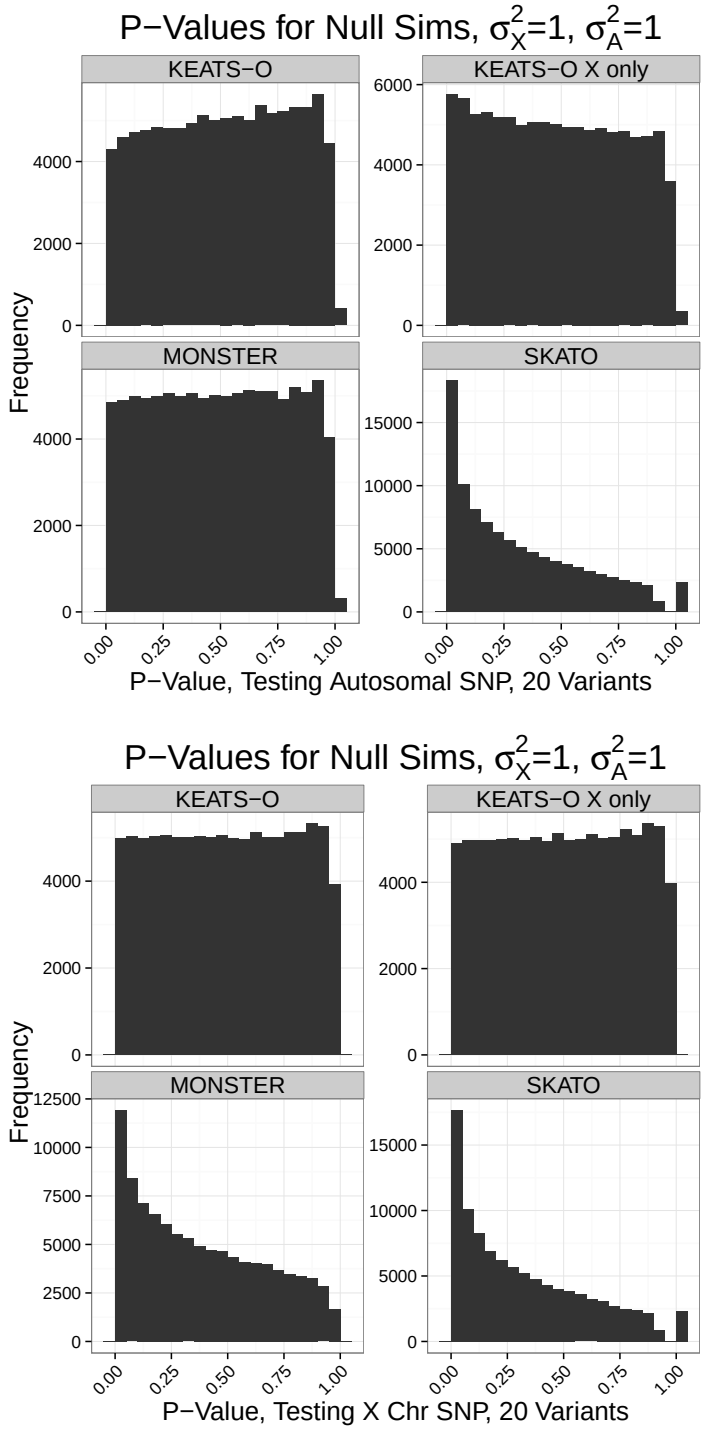


Figure 4.2: P-values from 100,000 null simulations calculated when testing 20 variants using KEATS-O, KEATS-O using only X chromosome random effect, MONSTER and SKAT-O. Both autosomal and X chromosome variant results are shown. Samples are related through pedigree configuration (2), where 8-person pedigrees with equal number of females and males are simulated.

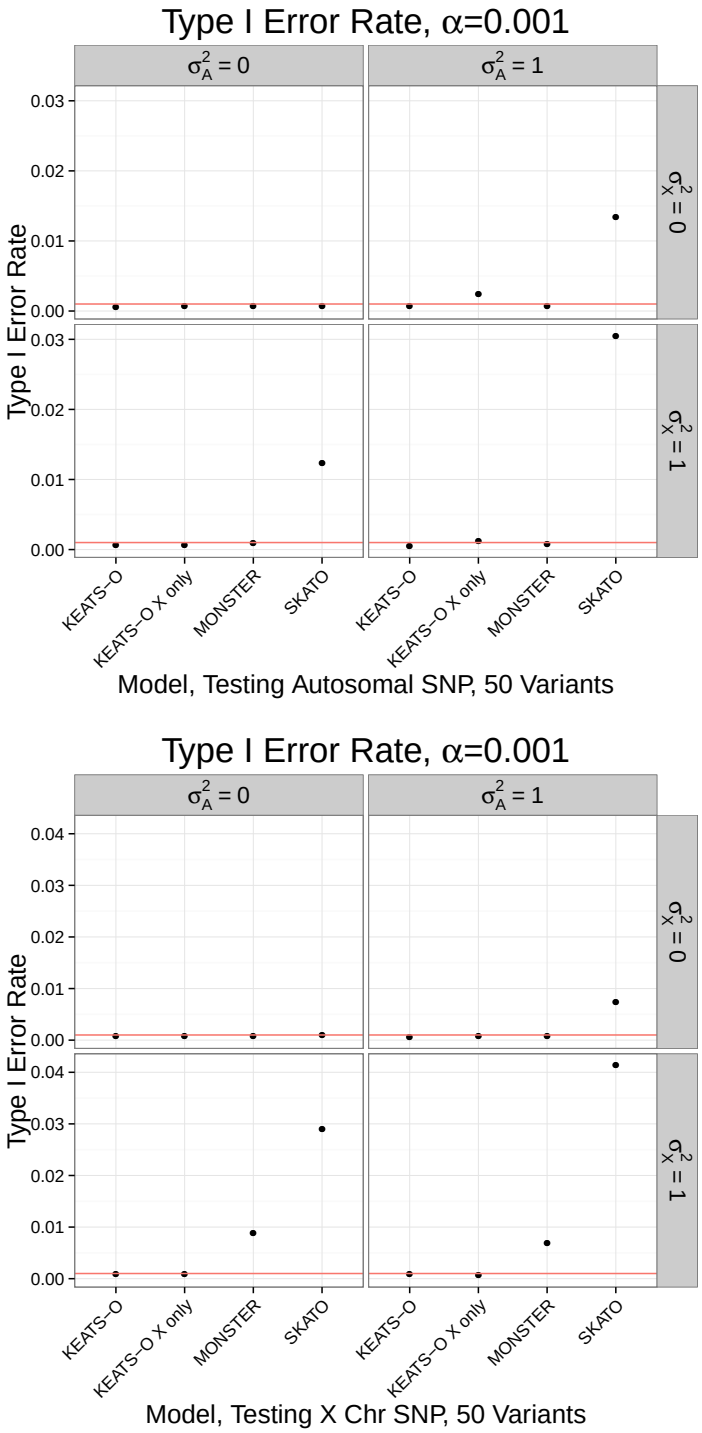


Figure 4.3: Type I error rate calculated when testing genes with 50 variants using KEATS-O, KEATS-O using only X chromosome random effect, MONSTER and SKAT-O. Both autosomal and X chromosome variants were considered. Samples are related through pedigree configuration (2), where 8-person pedigrees with equal number of females and males are simulated. The solid red line indicates the nominal type I error rate.

that variants with lower minor allele frequency are more likely to have larger effect sizes, we set the effect size to be $0.2|\log_{10}(\text{MAF})|$ for causal variants, where we multiply this by -1 for those variants chosen to be negatively correlated.

Figure 4.5 shows the false positive rate compared to the true positive rate for α significance levels ranging from 1e-10 to 0.25 as calculated from 10,000 simulation iterations in a set of 600 samples simulated under pedigree configuration (2). The genes were simulated to have 20 variants each which follow an AR-1 LD structure with neighboring variants having LD 0.5. The phenotypes have heritability 0.6, where $\sigma_X^2 = 0.9$, $\sigma_A^2 = 0.6$ and $\sigma_\epsilon^2 = 1$.

In scenarios when the effects are all in the same direction, and as the proportion of non-causal variants decreases, the SKAT-O method exhibits the lowest power. When the causal variants have some positive and some negative effects, the methods considered do not achieve as high a power than when all effects are in the same direction. Again, we see that KEATS-O, KEATS-O X only, and MONSTER achieve higher power when testing autosomal variants as compared to SKAT-O. The MONSTER method exhibits lower power when testing X chromosome variant regions, and on the contrary, the KEATS-O X only method exhibits low power when testing autosomal variant regions; this behavior is not surprising. In all scenarios considered, KEATS-O achieves the highest power.

In the optimal setting, a choice of ρ corresponds to which convex combination of the burden and variance component score test yields the most significant p-value. Figure 4.6 shows a histogram of the optimal ρ chosen by KEATS-O when testing both autosomal and X chromosome variants with varying proportions of variants positive/negative/not causal. We recall that when $\rho = 0$, the KEATS-O test corresponds to the variance component score test and when $\rho = 1$, the KEATS-O test corresponds to the burden test. In settings where all variants are causal in the same direction, the optimal ρ is zero much of the time, but there are simulation iterations in which $\rho > 0$. This is perhaps more apparent when we compare these values with the settings in which the causal variants have differing directions. In these simulation settings, the optimal ρ is almost always zero.

We assessed the power for 10,000 simulations in a set of 600 samples who are related through pedigree configuration (3), where the pedigree was chosen in such a way that the pairwise X chromosome kinship coefficients are, on average, twice the autosomal kinship

coefficients. Figure 4.7 shows the false positive rate compared to the true positive rate for α significance levels ranging from 1e-10 to 0.25. Again here, the genes were simulated to have 20 variants each which follow an AR-1 LD structure with neighboring variants having LD 0.5. The phenotypes have heritability 0.6, where $\sigma^2_X = 0.9$, $\sigma^2_A = 0.6$ and $\sigma^2_\epsilon = 1$.

Again in this pedigree setting, the SKAT-O method achieves the lowest power in all settings considered. The relative difference in power when testing autosomal gene regions is smaller for the KEATS-O, KEATS-O X only and MONSTER methods as compared to the X chromosome gene regions. This indicates that applying a method to X chromosome data that cannot specifically handle those data yields more loss of power than when testing autosomal methods. We notice that the KEATS-O X only method does not have the power loss when testing autosomal regions that we see when applying MONSTER to X chromosome gene regions. KEATS-O achieves the highest power in all simulation settings considered. All patterns observed in Figure 4.7 mirror those seen in Figure 4.5.

Finally, we consider the proportion of false to true positives when testing 20 variant gene regions in 600 unrelated samples (Figure 4.8). Perhaps unsurprisingly, all methods considered achieve essentially the same power under all configurations of positive/negative/not causal variants. The KEATS-O method achieves 95% power at a false positive rate of 0.05 in every scenario considered. The SKAT-O method, as well as the MONSTER and KEATS-O X only methods, are able to accurately identify which variants are causal and the p-values are properly calibrated when testing either autosomal or X chromosome variants. These results indicate that when samples are unrelated, the methods under consideration are equally effective in gene-based association testing with regard to power.

It is of interest to compare the results of KEATS-O to those found when using KEATS and KEATS-bt. Figure 4.9 shows the false positive to true positive rate comparing the omnibus test to the burden and kernel-based tests under pedigree simulation (3) using gene regions of 20 variants, stratified by autosomal and X chromosome gene regions. In the settings where we know that the burden test does not perform well, i.e., when the causal variants have different direction of effects, we see the KEATS-bt achieves a significant loss of power as compared to the other two methods. It is not surprising that we observe similar power between the KEATS-O and KEATS methods, as we saw in most settings the ρ value

is zero, corresponding to performing the KEATS method.

4.3.3 Application to HCHS/SOL Study, RBC Candidate Genes

We illustrate the KEATS-O method in an application to samples collected as part of the HCHS/SOL study. We investigated association of 15 previously reported genes with red blood cell count (RBC) as measured at baseline in 12,502 HCHS/SOL samples. Samples were excluded from the association test if they were missing RBC measurements, had bone cancer or a blood/lymph malignant tumor, were pregnant, had chronic kidney disease or were undergoing chemotherapy. Fixed effect covariates of sex, age, recruitment center, principal components eigenvectors 1-5 estimated from autosomal genetic variants and principal components eigenvectors 1-2 estimated from X chromosome genetic variants were included in the association model. Furthermore, we included random effects for autosomal and X chromosome kinship, as well as matrices with entries of 1 for pairs of individuals who are known to live in the same city block or household, reflecting known environmental correlation among the HCHS/SOL samples.

Association was considered with 15 gene regions: *LINC01221, TRIM58, PRKCE*, *PDGRFA*, *TERT, CCND3, MYB, ACTL6B, TFR2, ABO, CCND2, KITLG, ITFG3, NPRL3* and *G6PD*. These 15 gene regions were chosen based upon the SNPs published to be associated with RBC [17]. The SNPs were mapped to genes based upon annotated position, and variants genotyped in the HCHS/SOL study that lie within 100kb of the genes were grouped together and tested for association. We only considered polymorphic variants that were directly genotyped and that passed the composite quality filter [24] with MAF<0.05 as observed in the set of study samples. Summary values in Table 4.1 display the number of SNPs included in each test, along with the chromosome and region of each gene, listed in order across the genome.

Table 4.1 reports the p-value along with the optimal ρ as found using the KEATS-O method using Wu weights. There are three candidate genes that yield a significant KEATS-O p-value after adjusting for the 15 tests performed. These three candidate genes with significant p-values, *ITFG3, NPRL3* and *G6PD*, have been previously reported and

replicated [7, 28, 32], although none with a set of Hispanic/Latino individuals. We generalize these three published gene hits to the admixed population of Hispanic Americans. For only two of the 15 candidate genes does KEATS-O choose a ρ larger than zero. In these cases, the KEATS-O test is putting weight on both the variance component score and burden tests, rather than all weight on the variance component score test.

We compare the KEATS-O results to the minimum p-value of SNPs tested in the region as calculated using MLM-X (Chapter 3) in Table 4.1. In all cases, the unadjusted single test minimum p-value is more significant than the KEATS-O p-value. This is as expected since the gene region must 'include' the KEATS-O p-value, in a sense. We are testing all variants within a region and grouping them together, rather than hand-picking the most significant variant (which we do not know *a priori*). However, when adjusting the single test minimum p-value for multiple testing, we see that indeed KEATS-O is more significant than the single variant test in all but one case. One gene region, *ITFG3*, is still more significant after correction for multiple testing compared to KEATS-O. However, this gene region is highly significant in all cases.

4.3.4 Application to HCHS/SOL Study, Genome-Wide

The KEATS-O method was applied to 16,500 genes across the genome using SNP genotyping data collected as part of the HCHS/SOL study. Association was tested with the RBC phenotype as described previously (Section 4.3.3). Gene regions were defined with 897,163 SNPs that passed the composite quality filter [24] and with MAF< 0.05. The UCSC transcript database [51] was used to annotate SNPs based upon their position with respect to genomic location and gene; from this annotation SNPs were grouped into genes. Finally, gene regions with less than four variants were excluded from analysis. This resulted in a total set of 431,812 SNPs grouped into 16,500 genes.

Histograms of the number of variants per gene and the number of genes per chromosome are shown in Figure 4.10. The mean number of variants per gene is 27, with the largest gene, CSMD1 on chromosome 8, having over 1,500 variants. The number of genes per chromosome generally decreases as the chromosomes become smaller in length. However,

Gene	Chr	Region	# Vars	KEATS-O P-Value	Optimal ρ	Single Test Min. P-Value	Single Test Corrected P-Value
LINC01221	1	1q32.1	18	0.0224	0.3	1.30e-03	0.023
TRIM58	1	1q44	14	0.667	0.0	0.0464	0.650
PRKCE	2	2p21	236	0.0255	0.0	3.44e-03	0.812
PDGFRA	4	4q12	203	0.866	0.0	0.0155	1
TERT	5	5p15.33	15	0.340	0.0	0.0422	0.633
CCND3	6	6p21.1	39	0.808	0.0	0.113	1
MYB	6	6q23.3	18	0.413	0.0	0.0517	0.936
ACTL6B	7	7q22.1	3	0.104	0.0	0.0431	0.130
TFR2	7	7q22.1	8	0.346	0.4	0.0574	0.459
ABO	9	9q34.2	9	0.520	0.0	0.0171	0.154
CCND2	12	12p13.32	17	0.0136	0.0	2.25e-03	0.038
KITLG	12	12q21.32	35	0.354	0.0	0.0242	0.847
ITFG3	16	16p13.3	13	2.18e-12	0.0	5.43e-15	7.06e-14
NPRL3	16	16p13.3	21	2.63e-04	0.0	2.87e-05	6.03e-04
G6PD	X	Xq28	5	1.96e-18	0.0	1.82e-18	7.28e-18

Table 4.1: P-values for association of RBC with 15 candidate gene regions in the HCHS/SOL study using Wu weights. The corrected p-value was calculated using Bonferroni adjusted for the number of variants tested per gene region.

due to the ascertainment of SNPs on genome-wide SNP arrays, the distribution of SNPs may not be uniform across chromosomes. Figure 4.11 shows a histogram of the number of times a variant is mapped to a gene, where the y-axis is truncated at 100. The count of each bin is shown in text above each bar. Note the bars for variants mapped to 1, 2 and 3 genes are truncated from this plot. Since genes may be overlapping in the genome, some variants may be mapped to more than one gene. In this analysis, we find that 95.4% of variants are mapped to only one gene. The remaining 4.6%, 19,983 SNPs, of variants were mapped to between two and 22 genes.

Figure 4.12 shows a Manhattan plot of the 16,500 gene regions tested for association with RBC in the HCHS/SOL samples. Here, the 'genome-wide' significance level is simply calculated as the nominal 0.05 significance level Bonferroni-adjusted for the number of genes tested for association (0.05/16500=3e-06). Six genes reach the Bonferroni-corrected significance level (Table 4.2). Two of these genes, *ITFG3* and *G6PD*, have been previously published to be associated with RBC and were tested as part of the 15 candidate gene regions in the previous section (Section 4.3.3). The *F8* gene is close to the *G6PD* gene and was mentioned in a previous publication as a possible gene driving a hit in the Xq28 region [7]. Similarly, the chromosome 16 genes *LUC7L, RAB11FIP3* are close in location to the published gene *ITFG3*. An additional gene on chromosome 1, *GNB1*, is just slightly below the significance threshold with a p-value of 4.69e-06. This gene has no published associations with red blood cell traits known to the author at the time of writing.

One significant finding is on chromosome 11 in the *HBB* gene region. This gene has been previously published to be associated with mean corpuscular hemoglobin concentration in a cohort of African Americans [13] but was not significantly associated with RBC in that study. The *HBB*, 'hemoglobin, beta,' gene codes for making the beta-globin protein, which is a component of hemoglobin, a protein located inside red blood cells. Without further functional analyses, it appears this gene could indeed be involved with RBC, as it codes for a protein directly involved with red blood cells. Variants in the *HBB* gene have been published to be associated with sickle cell disease, production of hemoglobin C or production of hemoglobin E. Interestingly, hemoglobin C has been most commonly found among West African populations, while hemoglobin E is mostly seen among individuals of Southeast

Gene	Chr	Region	# Vars	KEATS-O P-Value	Optimal ρ	Single Test Min. P-Value	Single Test Corrected P-Value
HBB	11	11p15.5	4	5.52e-12	0.2	1.76e-11	7.03e-11
LUC7L	16	16p13.3	39	1.05e-16	0.0	5.43e-15	2.12e-13
ITFG3	16	16p13.3	13	2.18e-12	0.0	5.43e-15	7.06e-14
RAB11FIP3	16	16p13.3	43	1.58e-08	0.0	1.01e-09	4.33e-08
G6PD	X	Xq28	5	1.96e-18	0.0	1.82e-18	9.12e-18
F8	X	Xq28	15	1.32e-07	0.0	3.40e-10	5.09e-09

Table 4.2: Results from six genes that passed the genome-wide threshold when performing KEATS-O on 16,500 genes across the autosomes and the X chromosome testing for association of the RBC trait in 12,502 HCHS/SOL samples.

Asian descent.

Table 4.2 lists the KEATS-O p-value along with the optimal ρ value chosen for the six genes that reach significance at a Bonferroni-corrected level. We compare the KEATS-O p-value to the minimum p-value from each gene as calculated from the single SNP association tests as outlined in detail previously (Section 3.3.6). From the single test minimum p-value we calculate a Bonferroni-corrected p-value that adjusts for the number of variants in each gene. In two cases, the single test Bonferroni-corrected p-value is more significant than the KEATS-O p-value. The single test corrected p-value could be more significant, especially in the case when the optimal ρ is zero, as the single variant test does not search through several ρ values as well as correct for the number of variants tested per gene. When the optimal ρ value is non-zero, we expect the KEATS-O p-value to be more significant than the single test corrected p-value. In the one gene we see in Table 4.2 with a non-zero ρ value, this expectation is satisfied. Overall, the patterns of significance are mirrored in each method, with the KEATS-O method yielding a more significant p-value in most cases.

We compare the results using KEATS-O to performing either KEATS or KEATS-bt outside the omnibus framework. Table 4.3 compares the p-value found using KEATS-

Gene	Chr	Region	# Vars	KEATS-O P-Value	Optimal ρ	KEATS-bt P-Value	KEATS P-Value
HBB	11	11p15.5	4	5.52e-12	0.2	1.81e-10	2.88e-12
LUC7L	16	16p13.3	39	1.05e-16	0.0	1.48e-06	9.51e-18
ITFG3	16	16p13.3	13	2.18e-12	0.0	5.94e-07	1.86e-14
RAB11FIP3	16	16p13.3	43	1.58e-08	0.0	1.83e-04	1.30e-08
G6PD	X	Xq28	5	1.96e-18	0.0	1.58e-06	3.26e-17
F8	X	Xq28	15	1.32e-07	0.0	3.70e-04	2.24e-07

Table 4.3: Results from six genes that passed the genome-wide threshold when performing KEATS-O on 16,500 genes across the autosomes and the X chromosome testing for association of the RBC trait in 12,502 HCHS/SOL samples.

O to that found using KEATS and KEATS-bt for the six genes that reach genome-wide significance. When the optimal ρ parameter is zero, the KEATS p-value is more significant than the KEATS-bt p-value.

The optimal ρ parameter values chosen for each of the 16,500 gene tests are shown in a histogram in Figure 4.13. Similar to patterns observed in simulation studies, many of the tests chose $\rho = 0$, which corresponds to performing the variance component score test. However, there are approximately 6,500 genes that had a non-zero ρ. Further, over 1,000 genes yielded a test where $\rho = 1$, corresponding to the burden test. These results show that if we were to use the burden test, we would be performing the optimal test in less than 10% of our gene regions. By using the omnibus framework, we are able to capture the test that achieves the highest power for every gene tested for association.

We applied KEATS-O to 19,629 genes across the genome with 1,153,339 SNPs mapped to gene regions, where SNPs were not filtered based on MAF. The analysis included rare, low frequency and common variants, and SNPs were filtered based on quality. As before, we remove genes with fewer than four variants. The filtering and grouping resulted in 1,048,345 SNPs grouped into 19,629 genes. Histograms of the number of variants per gene and the

number of genes per chromosome are shown in Figure 4.14. The mean number of variants per gene is 59, with the largest gene, CSMD1 on chromosome 8, having 4,951 variants. Overall, the gene regions have more variants than the previous genome-wide KEATS-O test due to inclusion of variants with MAF>5%. Figure 4.15 shows a histogram, truncated at 300, of the number of times a variant is mapped to a gene. Of the variants included, 95.4% of them are only mapped to a single gene. The remaining 4.6% (48,109 SNPs) are mapped to between two and 22 genes.

Figure 4.16 shows a Manhattan plot of the 19,629 gene regions tested for association with RBC in the HCHS/SOL samples. Here, the 'genome-wide' significance level is calculated as the nominal significance level Bonferroni-adjusted for the number of genes tested for association (0.05/19629=2.5e-06) and seven genes reach the Bonferroni-corrected significance level. Table 4.4 displays the genes that reach genome-wide significance in the present association test, as well as those genes that reached significance in the KEATS-O analysis only considering rare and low frequency variants. The *HBB* gene was not significant when including common variants in the association test. Interestingly, the *MPP1* and *SMIM9* genes were not included in the rare and low frequency KEATS-O analysis; the variants in these genes all have MAF>5%. As the Wu weights depend on allele frequency, common variants also contribute to the signal detected. In some cases, by including the common variants in the test, a gene region has enough variants to be included in the analysis and provides sufficient signal to pass genome-wide significance.

4.3.5 Comparison of Gene-Based Tests to Single SNP Tests in HCHS/SOL

In practice, single variant tests are usually performed on SNP genotyping data and gene-based tests are performed on whole-genome or whole-exome sequence data. This difference in methods is due to the lack of power of single variant tests in rare or low frequency variants, such as those observed in whole-genome or whole-exome sequencing data. In this dissertation, we applied both single variant and gene-based tests to SNP genotyping data in the HCHS/SOL samples. The single SNP association analyses were outlined in detail previously (Chapter 3). In particular, we recall the genome-wide single SNP association

			Rare + LF KEATS-O			All Variants KEATS-O		
Gene	Chr	Region	# Vars	P-Value	Optimal ρ	# Vars	P-Value	Optimal ρ
HBB	11	11p15.5	4	5.52e-12*	0.2	9	3.79e-04	0.0
LUC7L	16	16p13.3	39	1.05e-16*	0.0	109	1.38e-16*	0.0
ITFG3	16	16p13.3	13	2.18e-12*	0.0	26	1.64e-13*	0.0
RAB11FIP3	16	16p13.3	43	1.58e-08*	0.0	77	9.84e-09*	0.0
G6PD	X	Xq28	5	1.96e-18*	0.0	12	3.26e-16*	0.0
MPP1	X	Xq28	-	-	-	10	1.72e-08*	0.0
SMIM9	X	Xq28	-	-	-	7	2.32e-10*	0.0
F8	X	Xq28	15	1.32e-07*	0.0	43	8.23e-08*	0.0

Table 4.4: Results from nine genes that passed the genome-wide threshold when performing either the rare + low frequency and/or all variants KEATS-O analysis testing for association of the RBC trait in 12,502 HCHS/SOL samples. A * indicates the gene was genome-wide significant in a particular analysis. A - indicates the gene was not included in the analysis.

test results in which four regions contained variants that reached genome-wide significance (Figure 3.10) on chromosomes 6, 7, 16 and the X chromosome. The chromosome 6 variants are located in the *AHI1* gene on the 6q23.3 region of the chromosome. The variants on chromosome 7 are in the *GATS* gene on chromosome 7q22.1. There are many variants on chromosome 16 in the 16p13.3 region that reach genome-wide significance. The most significant hit is a variant in the *LUC7L* gene, and the remaining variants are in genes around the *LUC7L* gene and are most likely in LD with the *LUC7L* region. Finally, the variants on the X chromosome that reach genome-wide significance are in the Xq28 region and the leading SNP is in the *G6PD* gene.

We contrast the single variant results with the gene-based test performed genome-wide on the HCHS/SOL SNP genotyping data (Figure 4.12). Figure 4.17 shows Venn diagrams of the variants on chromosome 16 and the X chromosome that were significant in the single SNP test along with the variants that were mapped to the genes that achieved significance

with the KEATS-O test. On chromosome 16, only 10 of the significant single SNP variants were mapped to one of the three significant genes. In this case, it does not appear that the single SNP variants are predicting the gene based results. By combining the rare and low frequency variants, we are able to identify significant signals on chromosome 16. Furthermore, all variants in the *ITFG3* gene were mapped to the *LUC7L* gene. Genes can overlap, thus, of the variants we had for association testing in the HCHS/SOL samples, the variant sites in the *ITFG3* gene were completely contained within the variant sites in the *LUC7L* gene. The X chromosome only has three variants that overlap between the single SNP and gene-based association results. There are no variants that are mapped to both the *F8* and *G6PD* gene.

Aside from looking at the overlap of variants in each gene, we calculate the pairwise LD between the three chromosome 16 genes and the two X chromosome genes that achieved significance in the rare and low frequency KEATS-O analysis. Figure 4.18 shows the three pairwise combinations of the significant genes on chromosome 16. As noted from Figure 4.17, the 13 variants mapped to the *ITFG3* gene are also mapped to the *LUC7L* gene. Thus, Figure 4.18a shows LD values of 1 for those 13 variants. All LD falls between 0.5 and 0.7 for the variants in the chromosome 16 genes. The pairwise LD for the two significant genes on the X chromosome are shown in Figure 4.19. The variants in the *G6PD* and *F8* genes have an LD of 0.5, approximately, for all variants.

Genes on chromosome 16 and the X chromosome remain significant between the single SNP and gene-based tests. However, the *HBB* gene on chromosome 11 did not yield variants significant in the single SNP association test. Appendix D includes tables of the significant KEATS-O genes and the variants that mapped to each of the genes, along with their single SNP test p-value and minor allele frequency. Interestingly, it appears that the gene-based test identified a signal from the *HBB* gene that was not discovered in the single SNP test. In the previous section, we found that the *HBB* gene was not significant when applying KEATS-O to genes defined using all variants. In this gene, there is not one single, common variant that tags the causal SNP well enough, but when all the rare and low frequency variants in the gene are combined, they jointly tag the region and produce a signal that is strong enough to pass genome-wide significance. Together, the variants with MAF less than

5% in the *HBB* region capture an association signal that is uncaptured by a single variant.

4.4 Discussion

Gene-based association testing in samples with complex, often unknown structure requires careful attention. Here, we developed the KEATS-O method for properly calibrated and powerful gene-based association testing valid for autosomal and X chromosome genetic variants. KEATS-O uses an optimal framework to choose the most powerful weighted linear combination of the burden and variance component score tests, adjusting for autosomal and X chromosome genetic random effects, yielding correct type I error and improved power. KEATS-O is easily generalizable to any number of random effects to control for other factors, such as environmental correlation.

We performed simulation studies under various settings in which we assessed the power and type I error of KEATS-O compared to existing methods. We demonstrated that KEATS-O has appropriate type I error. In settings where samples are related, KEATS-O achieves higher power when testing X chromosome gene regions than other methods that either ignore relatedness or are developed for testing autosomal gene regions.

In our application to SNP genotyping data collected as part of the HCHS/SOL study of Hispanic/Latino individuals, we first examined 15 genes previously identified to be associated with RBC, although not in Hispanic/Latinos. The KEATS-O p-value was more significant than the single test p-value corrected for the number of variants in the gene, in all but three genes tested. When applying KEATS-O genome-wide, we identified six genes that achieve significance at a Bonferroni-corrected threshold. At time of writing, it is not known that any of these gene regions have been published to be associated with RBC in Hispanic/Latino individuals. Furthermore, the *HBB* gene region does not appear to have been published to be associated with RBC in any population, although it has been linked with other red blood traits, such as hemoglobin levels.

It is important to note that in this dissertation we are using genotype data, rather than sequence data, for these tests. Thus, our set of variants to be tested are selected to be common, so *a priori* we may not expect them to be causal [19]. SNP ascertainment is non-uniform and was originally based upon regions of the genome that exhibited variation

on a relatively large scale within populations of European ancestry. Analysis with sequence data will allow us to use variants with lower MAF, as well as restrict to variants that are nonsynonymous or have been found to be functional. This may yield higher power.

To define gene regions, we mapped genotyped SNPs to annotated genes using position on the genome. There is no consensus agreement in the scientific community on the definition of a 'gene,' and the list of known genes in the human genome can even change depending on which database used. It is unclear if the variants should be grouped based upon gene expression, i.e. if a particular group of exons are usually expressed at the same time, or if exons should be split out into different genes. Furthermore, it is unclear how to handle overlapping genes and if flanking regions should be included in the gene-based association testing methods. Should variants be grouped within genes, then separated based upon annotated function? The present study examines only one approach to mapping assayed SNPs to genes. Further work should identify how sensitive the methods are to the definition of a gene region and perhaps use more sophisticated methods of defining genes. In addition, our set of variants includes SNPs with different functions. Grouping variants based upon their function could allow for more clear identification of genes that are associated with complex disease through their function. Using additional known information such as excluding SNPs from a region that are known to be non-causal could increase power to detect an association.

We observed in nearly half of our simulation iterations that the optimal ρ chosen was zero, corresponding to the variance-component based score test implemented in the KEATS method (Figure 4.6). Additional simulations could explore the mechanism driving these results. Is it the proportion of null variants? The direction of effects of the causal variants? The total number of variants in the gene region? Perhaps it is the LD structure among the variants being tested?

Further exploration could be done with regard to weighting variants in a gene region. In all analyses presented here, Wu weights were used, which put higher weight on rarer variants. A weighting scheme that considers the functional annotation as well as the observed frequency of the variants could be used instead to improve power.

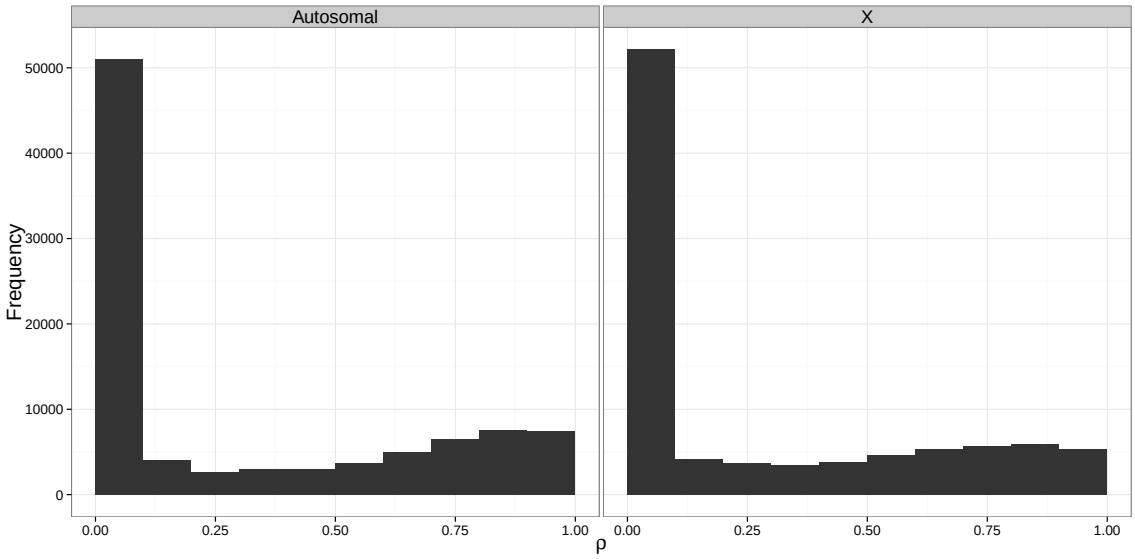


Figure 4.4: Histogram of optimal chosen ρ values for 100,000 null simulations. The samples are related under the pedigree configuration (2) and gene regions include 50 variants. Results testing autosomal and X chromosome gene regions are displayed.

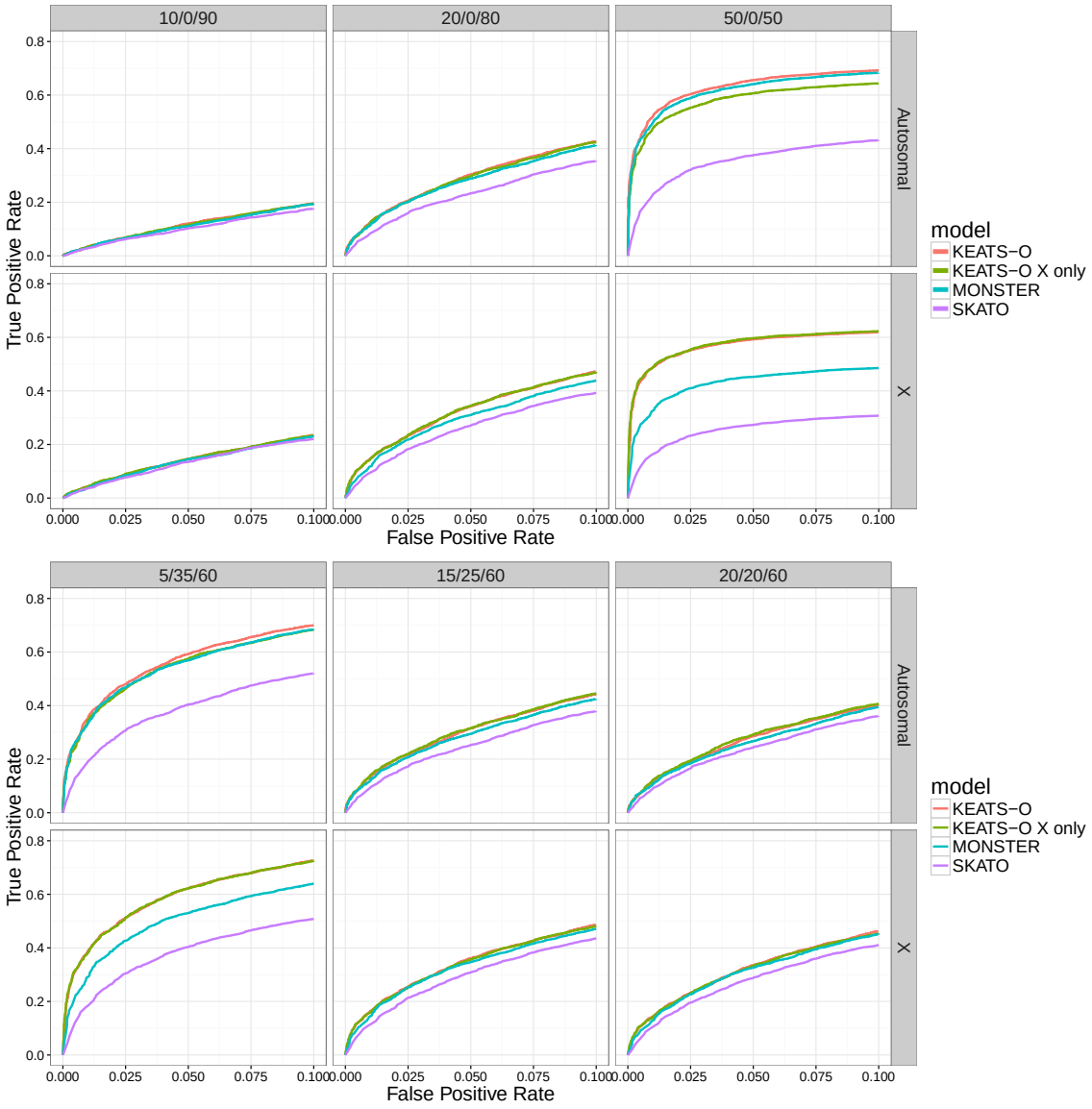


Figure 4.5: Power calculated when testing genes with 20 variants using KEATS-O, KEATS-O using only the X chromosome random effect, MONSTER and SKAT-O. Both autosomal and X chromosome variant results are shown here. The column titles indicate what percentage of variants are +/-/0 associated with the phenotype. Samples are related through pedigree configuration (2), where 8-person pedigrees with equal number of females and males are simulated.

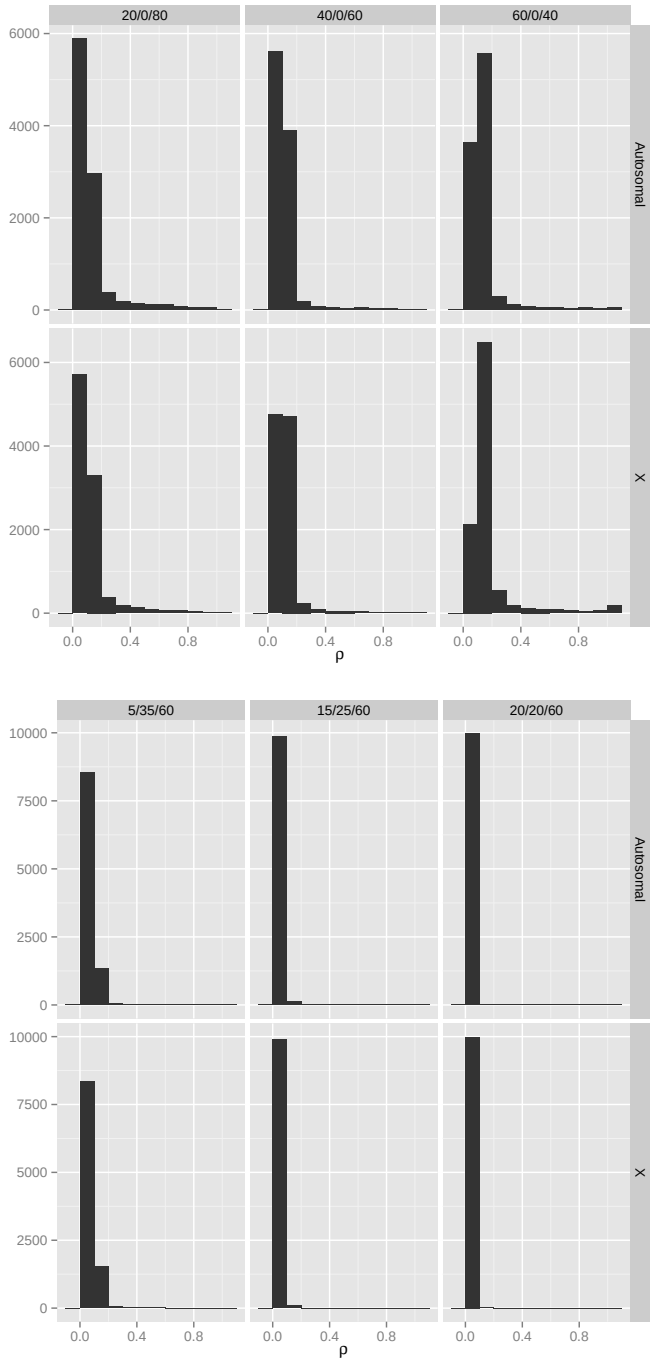


Figure 4.6: The optimal ρ parameter chosen in 10,000 simulation iterations when testing genes with 20 variants using KEATS-O. Both autosomal and X chromosome variant results are shown here. The column titles indicate what percentage of variants are +/-/0 associated with the phenotype. Samples are related through pedigree configuration (2), where 8-person pedigrees with equal number of females and males are simulated.

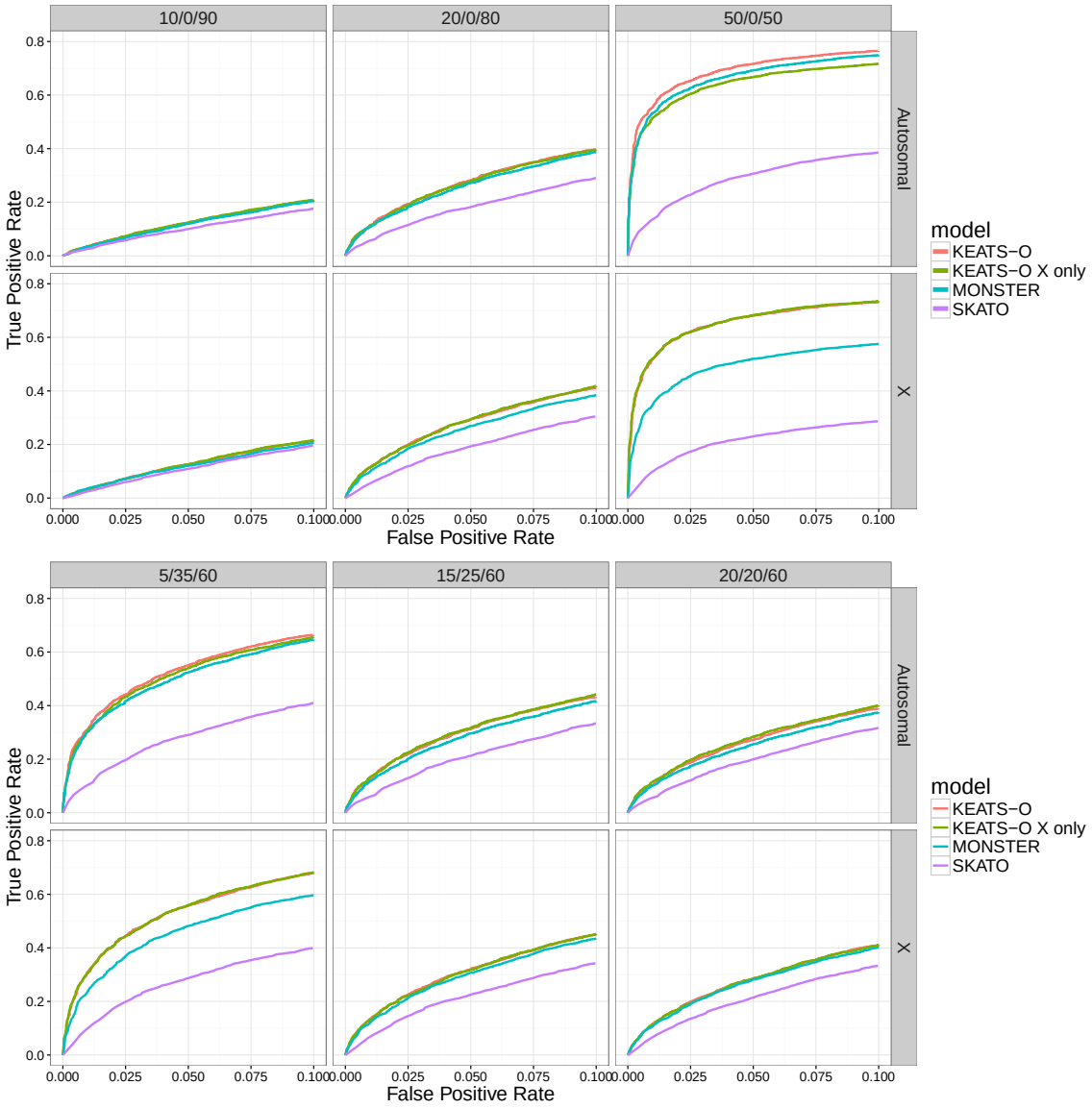


Figure 4.7: Power calculated when testing genes with 20 variants using KEATS-O, KEATS-O using only X chromosome random effect, MONSTER and SKAT-O. Both autosomal and X chromosome variant results are shown here. The column titles indicate what percentage of variants are +/-/0 associated with the phenotype. Samples are related through pedigree configuration (3), where 8-person pedigrees with 6 males and 2 females per 8-person family are simulated.

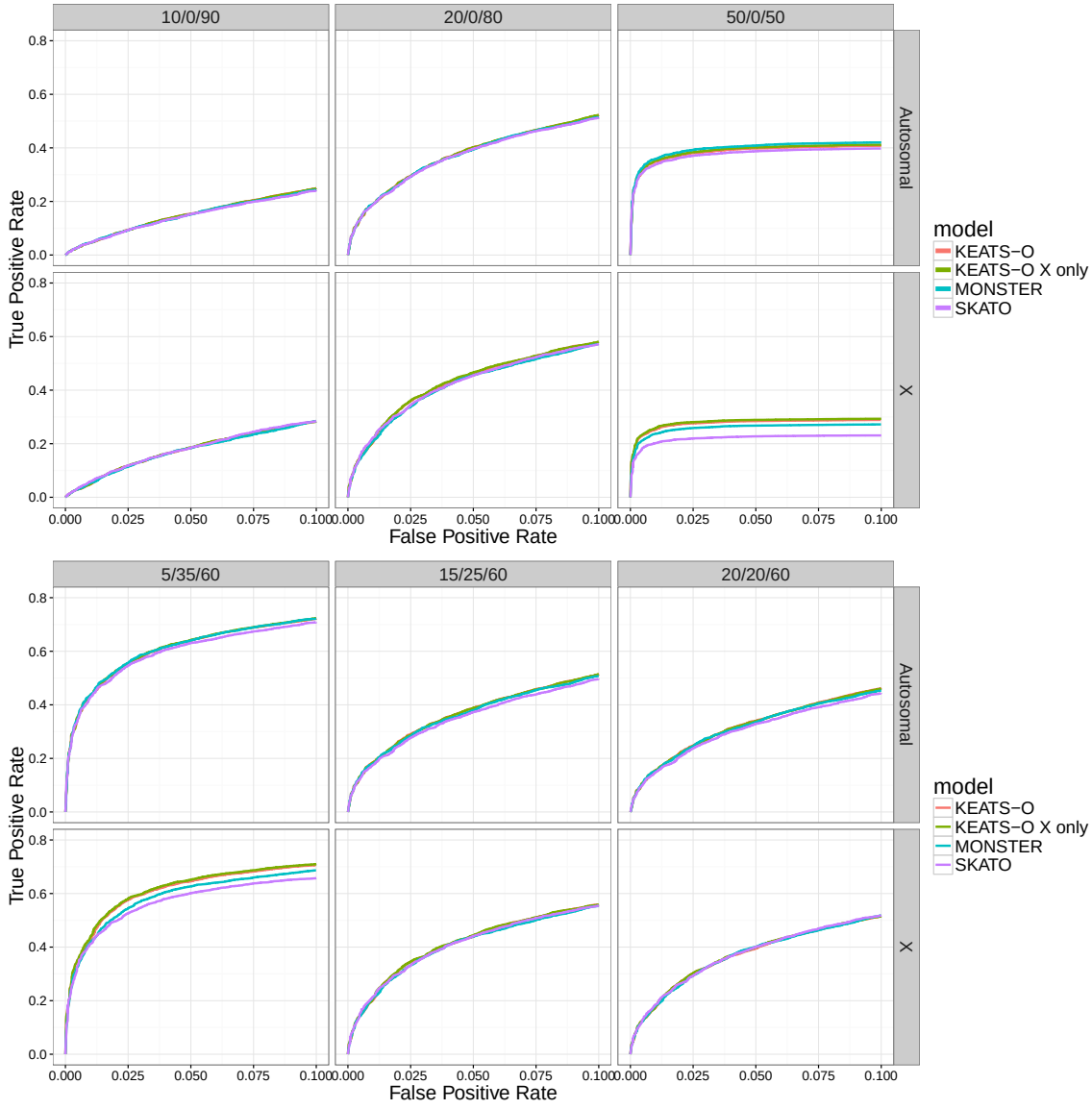


Figure 4.8: Power calculated when testing genes with 20 variants in a set of 600 unrelated samples using KEATS-O, KEATS-O using only X chromosome random effect, MONSTER and SKAT-O. Both autosomal and X chromosome variant results are shown here. The column titles indicate what percentage of variants are +/-/0 associated with the phenotype.

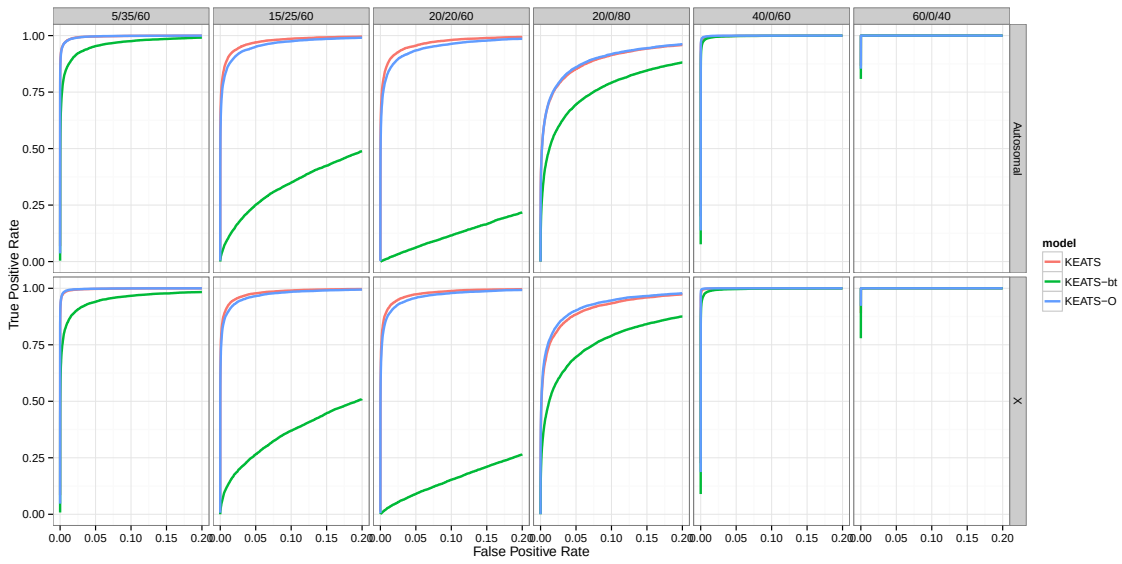


Figure 4.9: Power for the KEATS-O method compared to the KEATS-bt and KEATS methods. Gene regions of 20 variants were tested for autosomal and X chromosome genes. The percentage of positive/negative/non-causal variants in each gene are shown in the column titles.

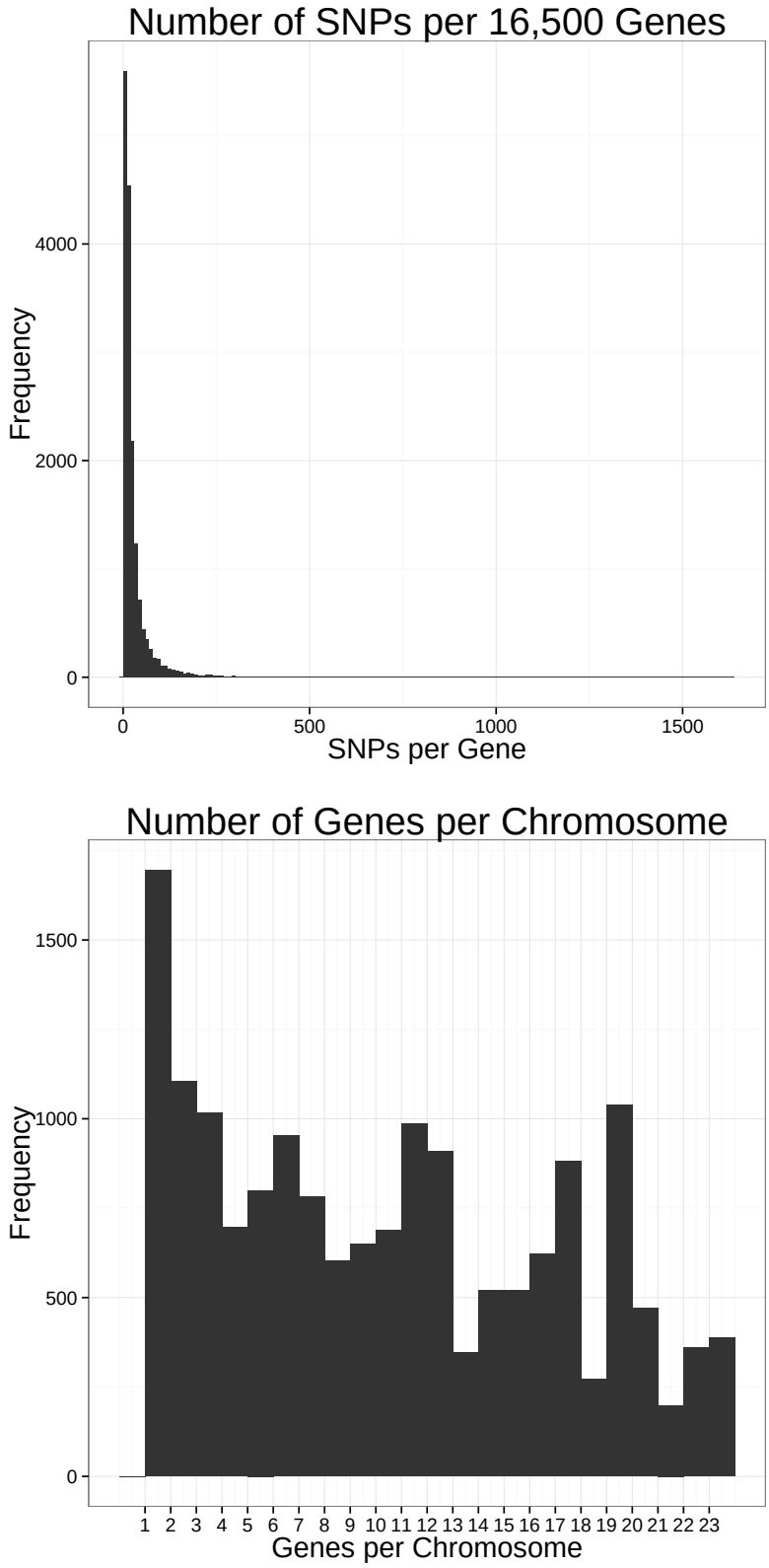


Figure 4.10: Histograms of variants per gene and genes per chromosome in the 16,500 genes mapped from SNP genotyping data in the HCHS/SOL samples.

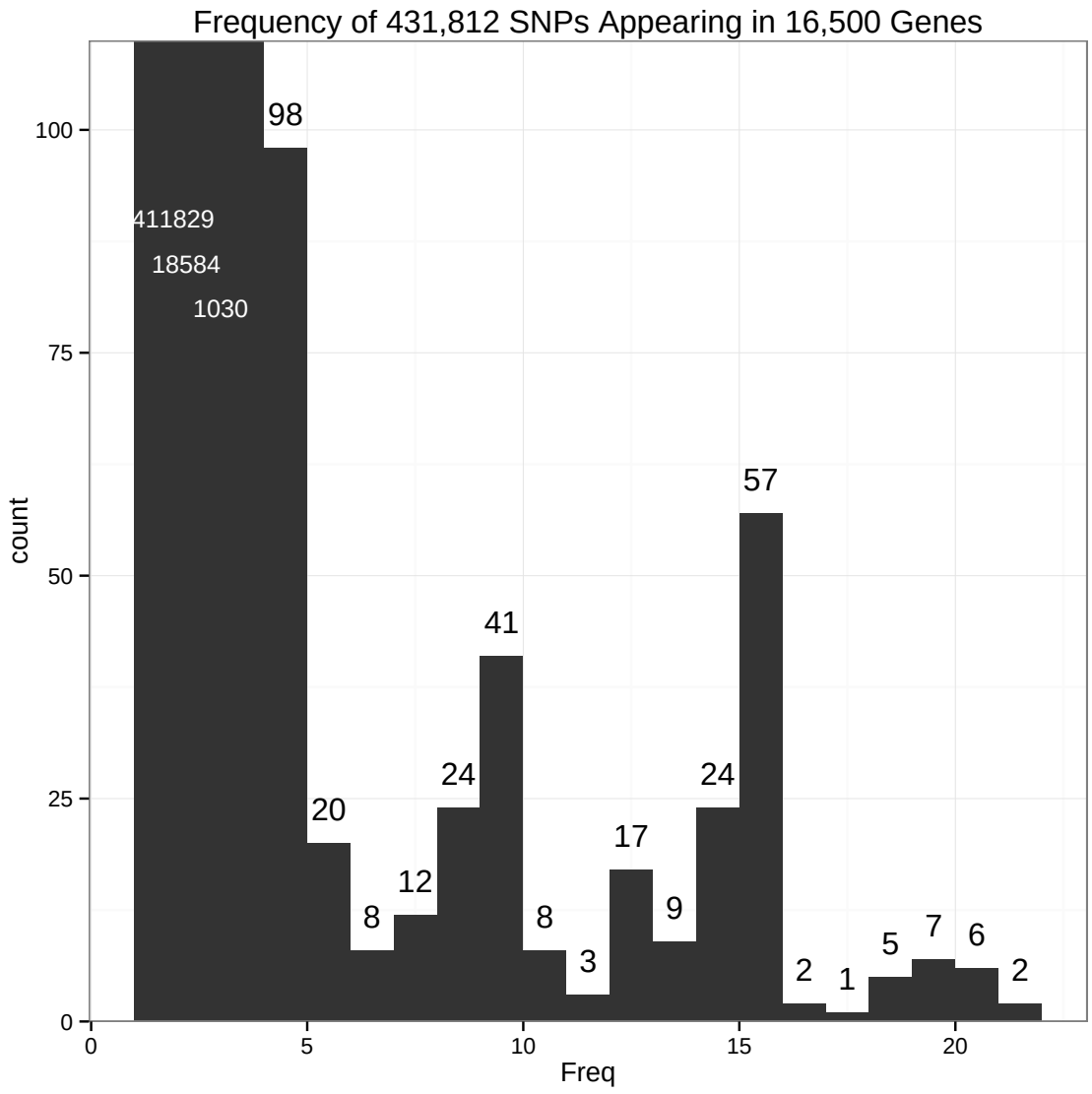


Figure 4.11: Histogram of the number of times a variant is mapped to a gene. Note this graph is truncated at 100 on the y-axis. The height of the bar is shown in the text; variants that mapped to 1, 2 or 3 genes are truncated from this plot but their count is shown in the text.

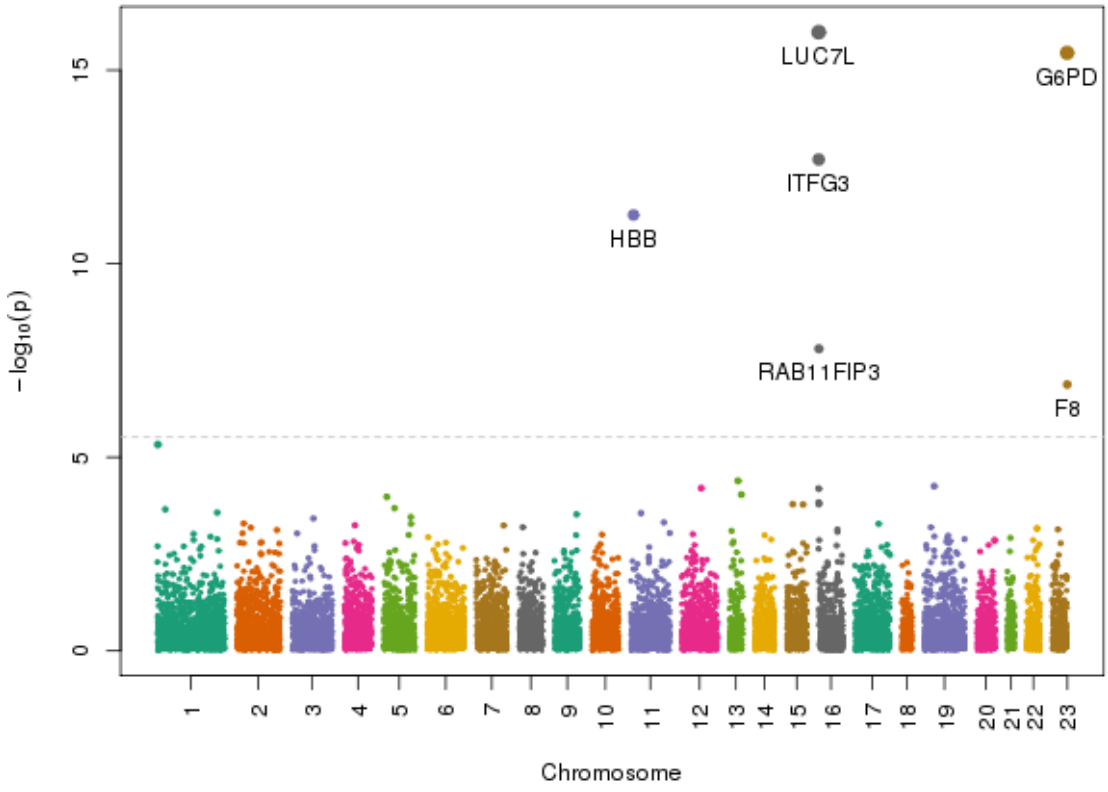


Figure 4.12: Manhattan plot of 16,500 genes genome-wide. P-values are results from KEATS-O testing the association of RBC in a set of 12,502 HCHS/SOL samples.

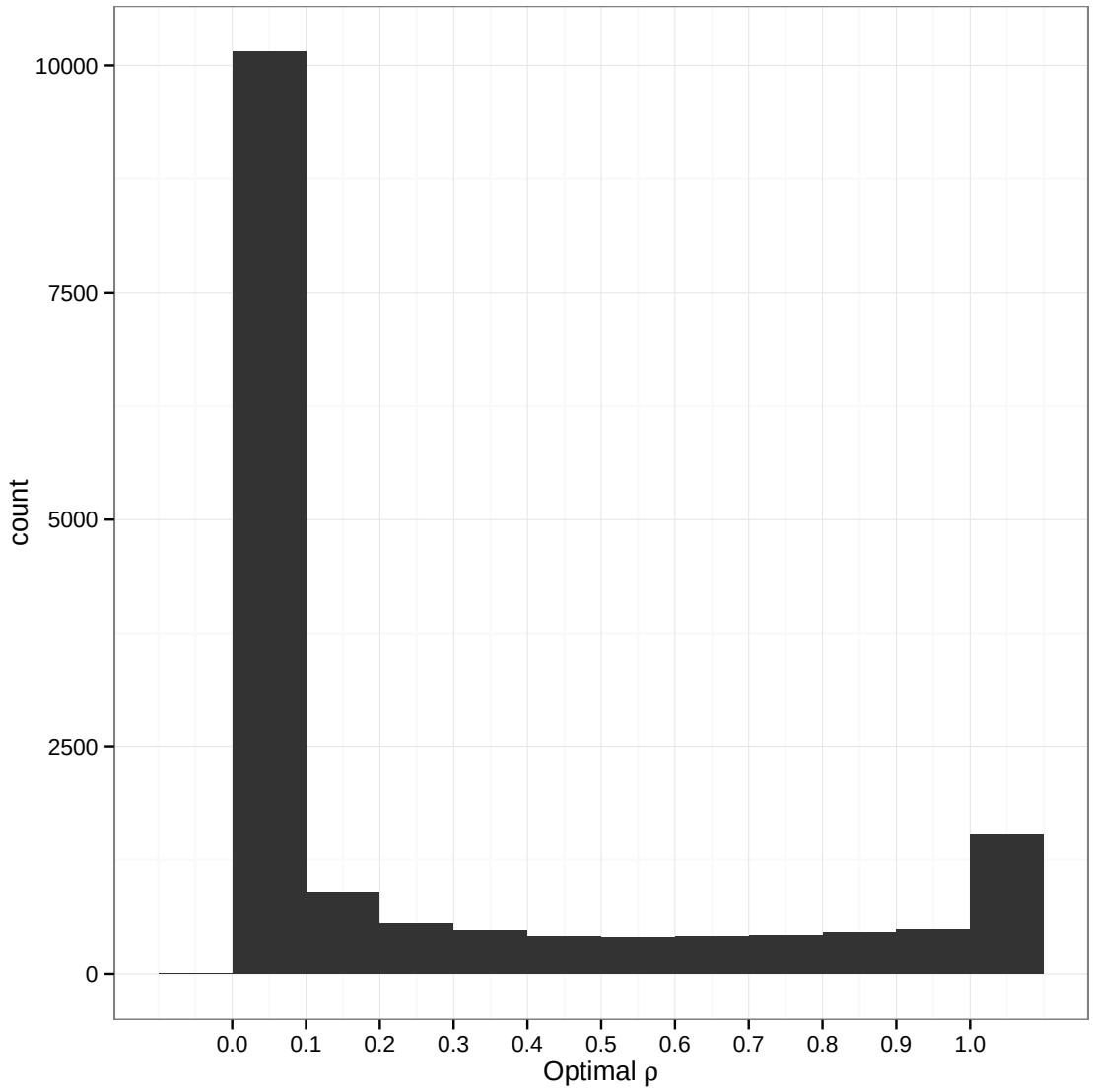


Figure 4.13: Histogram of the optimal ρ chosen from KEATS-O in 16,500 genes genome-wide testing the association of RBC in a set of 12,502 HCHS/SOL samples.

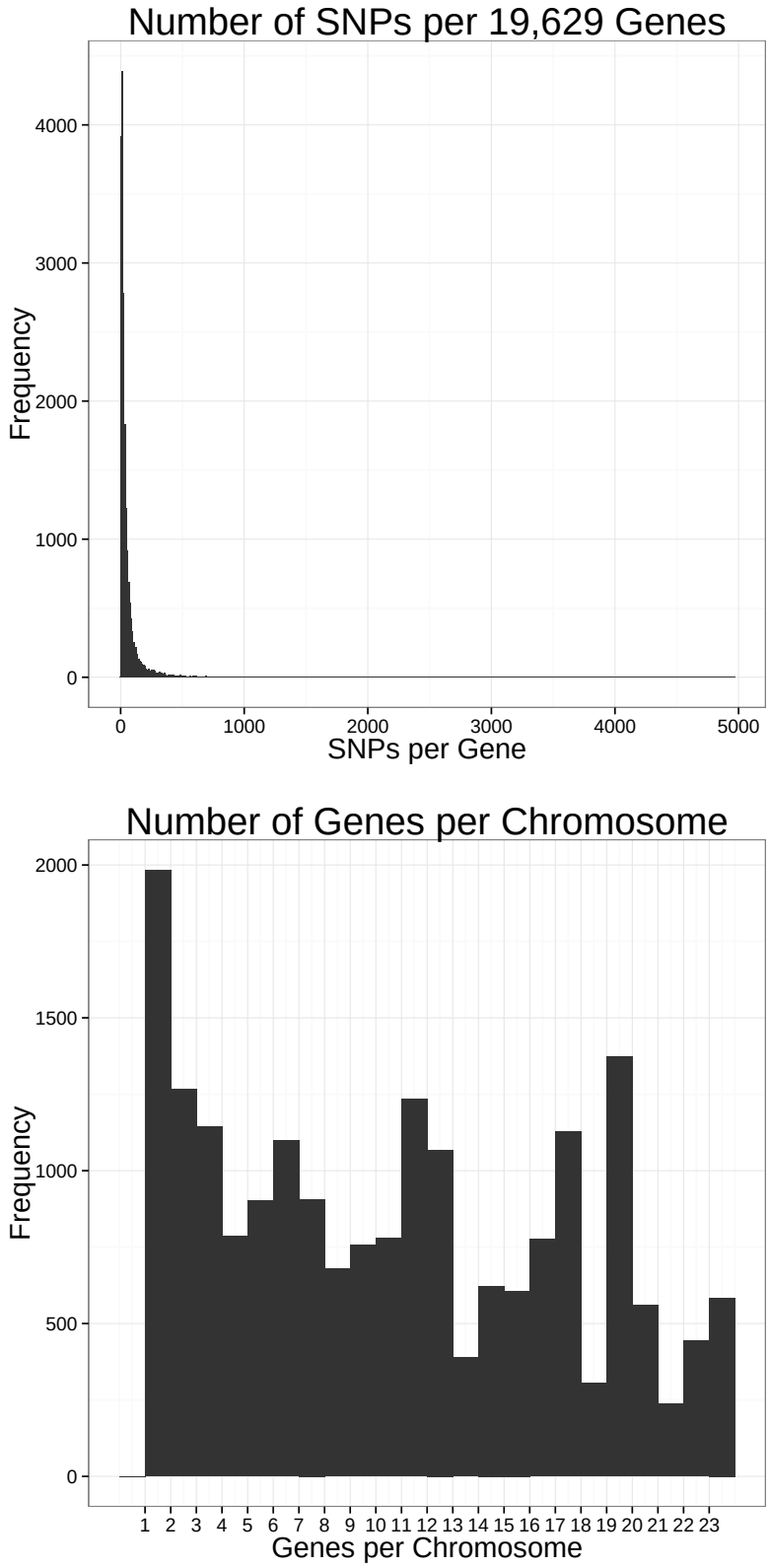


Figure 4.14: Histograms of variants per gene and genes per chromosome in the 19,629 genes mapped from SNP genotyping data in the HCHS/SOL samples.

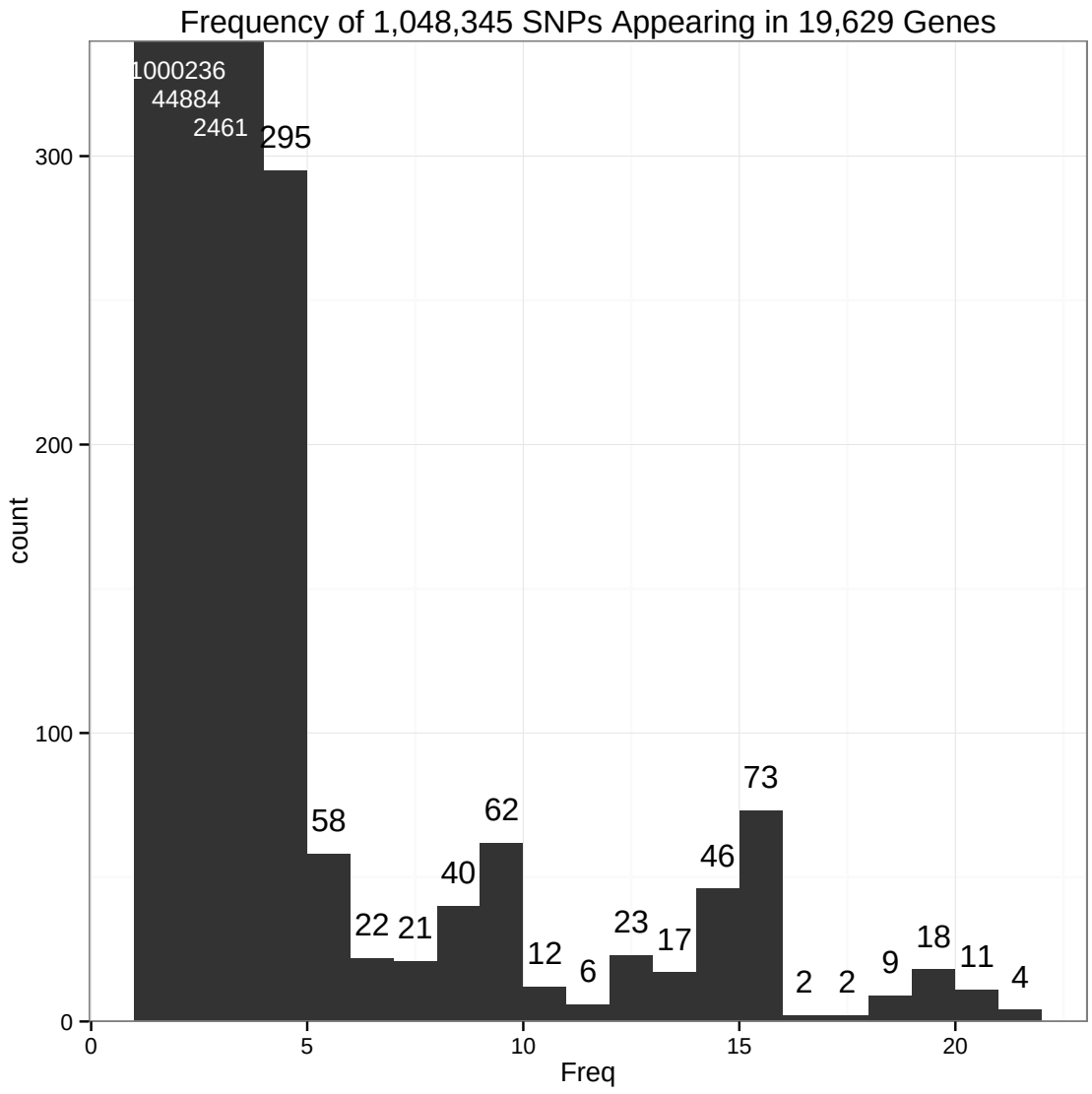


Figure 4.15: Histogram of the number of times a variant is mapped to a gene. Note this graph is truncated at 300 on the y-axis. The height of the bar is shown in the text; variants that mapped to 1, 2 or 3 genes are truncated from this plot but their count is shown in the text.

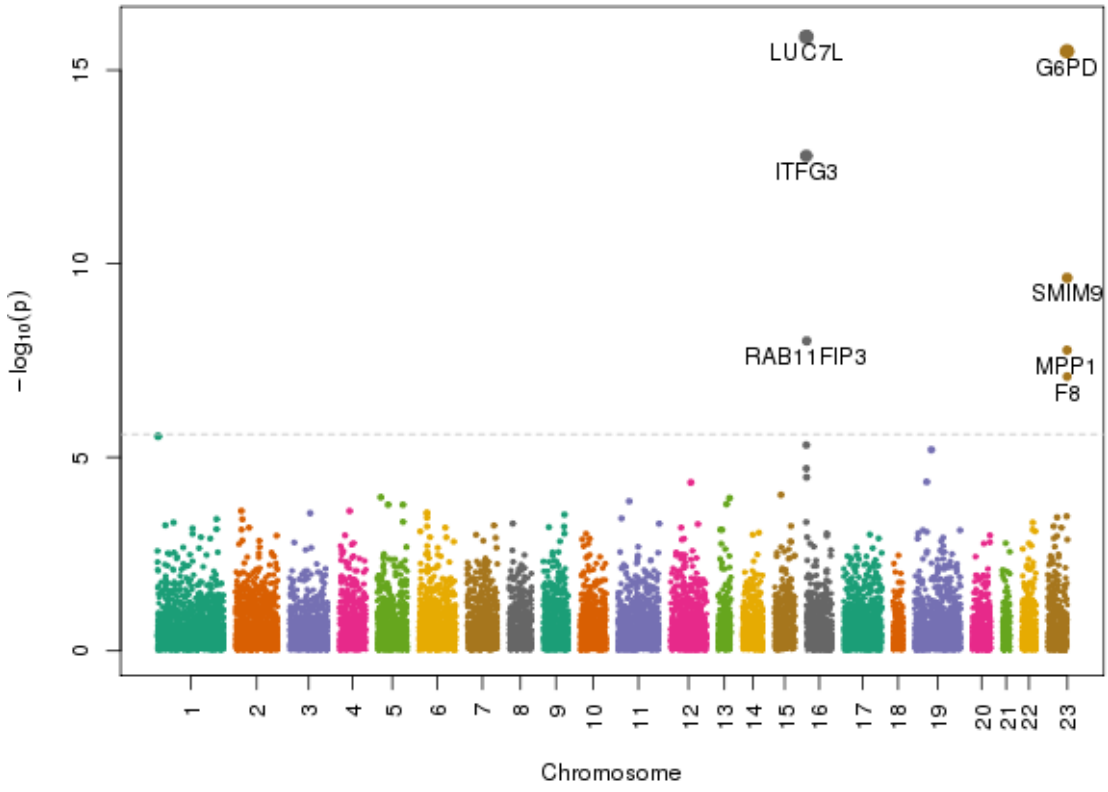
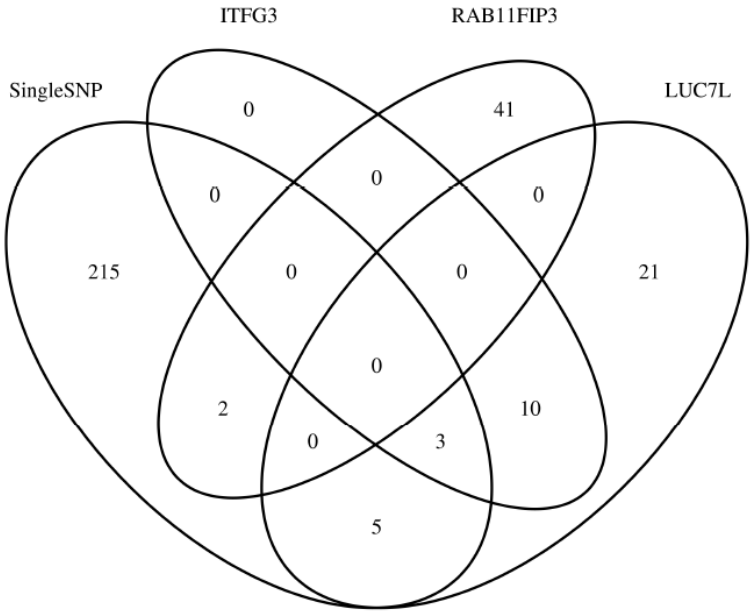
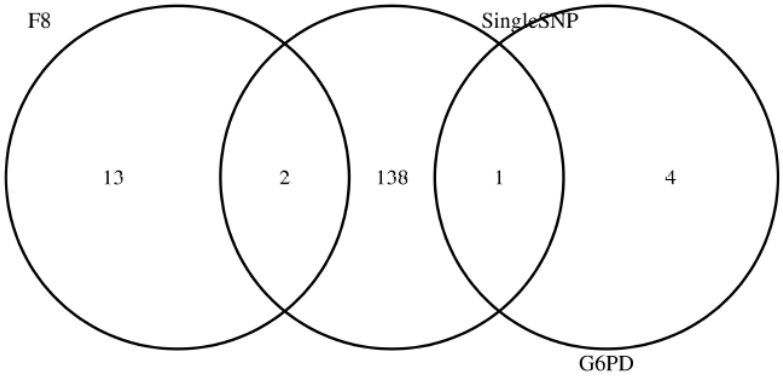


Figure 4.16: Manhattan plot of 19,629 genes genome-wide. P-values are results from KEATS-O testing the association of RBC in a set of 12,502 HCHS/SOL samples.

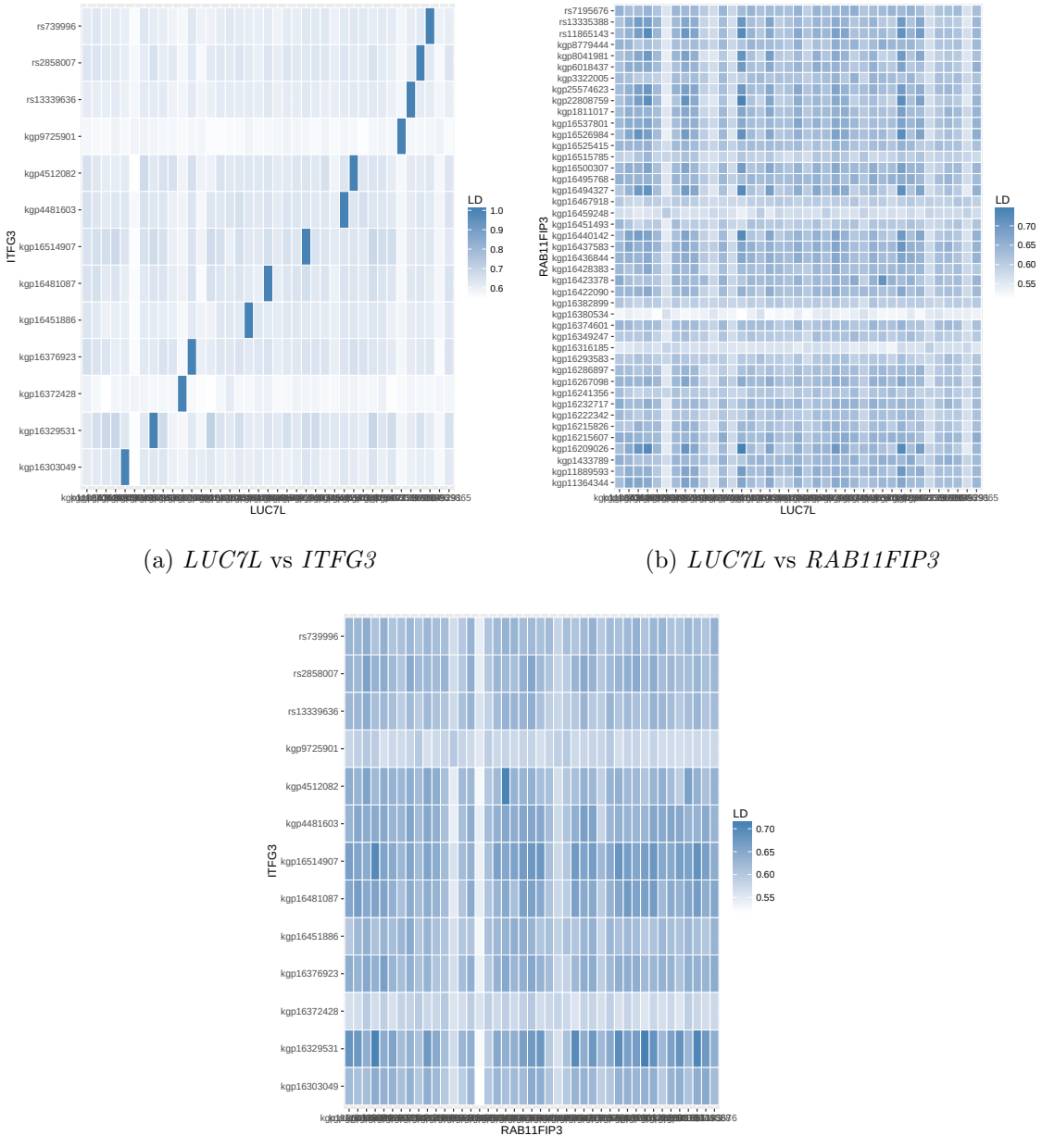


(a) Chromosome 16



(b) X Chromosome

Figure 4.17: Venn diagrams of the variants on chromosome 16 and the X chromosome that were significant in either the single SNP association test, or were mapped to genes that were significant in the KEATS-O analysis.



(a) *LUC7L* vs *ITFG3*

(b) *LUC7L* vs *RAB11FIP3*

(c) *RAB11FIP3* vs *ITFG3*

Figure 4.18: Pairwise LD between 39 variants in the *LUC7L* gene, 13 variants in the *ITFG3* gene and 43 variants in the *RAB11FIP3* gene.

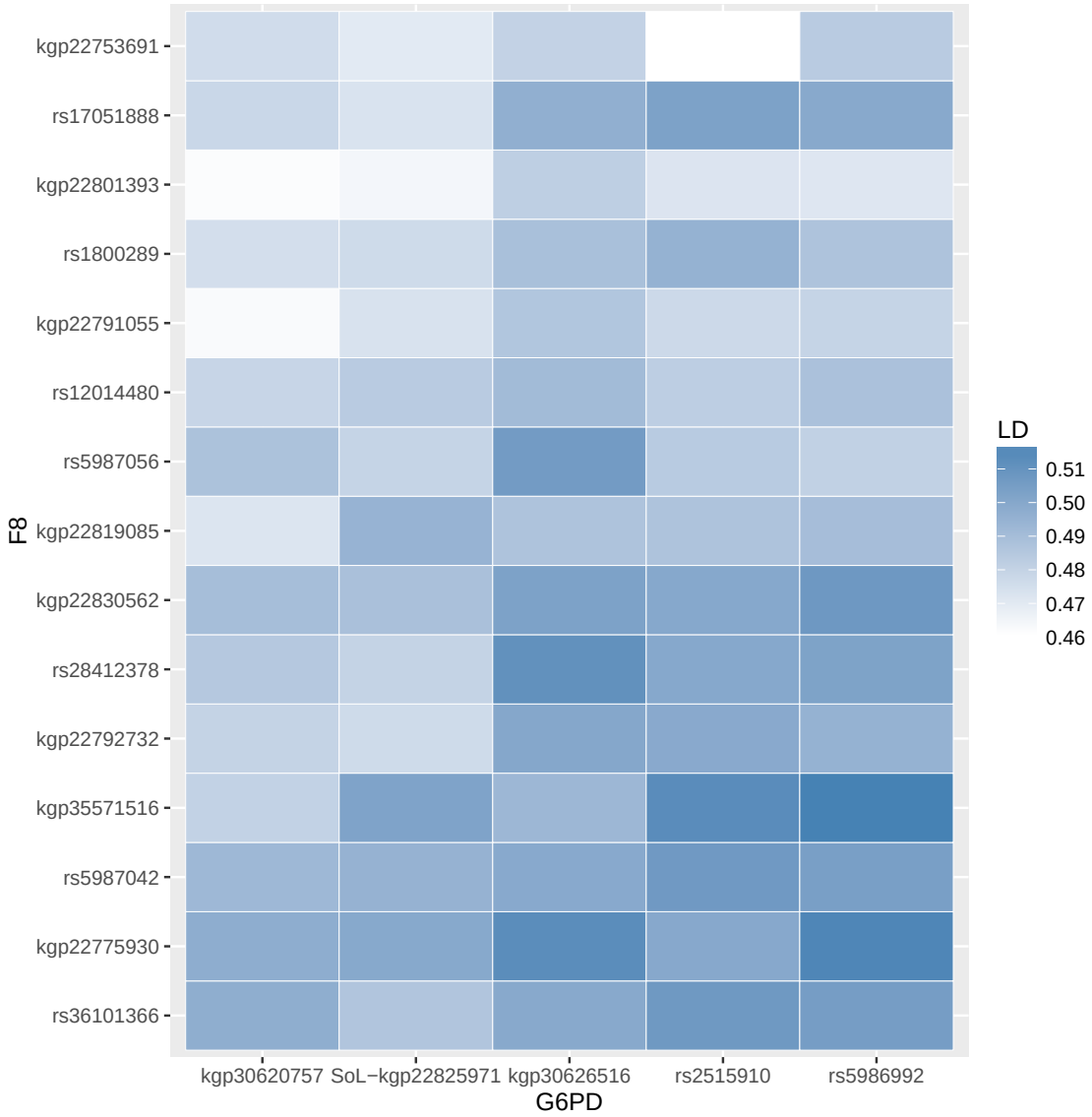


Figure 4.19: Pairwise LD between 15 variants in the *F8* gene and the 5 variants in the *G6PD* gene.

Chapter 5

CONCLUSIONS AND DISCUSSION

In this dissertation, we explored methods to detect and account for complex structure in genome-wide genetic analyses. We developed a method to detect heterogeneity in population structure, motivated by sex-specific non-random mating at the time of admixture in the colonization of the Americas, lending to differential population structure on the X chromosome as compared to the autosomes. We then focused on association testing methods for both rare and common genetic variants, paying special attention to the X chromosome. These methods can handle complex structure, such as ancestry or relatedness. Our proposed methods were compared with existing approaches, in both simulated and real data examples.

In Chapter 2, we developed a method that assesses heterogeneity in population structure across the genome in populations with admixed ancestry. Previous publications have detected differences in autosomal and X chromosome ancestry by using a t-test [4]. A t-test ignores correlation in ancestry within an individual between the autosomes and the X chromosome. Other models have been proposed [54, 20, 1] that require strong assumptions about the evolution of the population of interest in order to assess significance. Using inferred local ancestry, CAnD tests for systematic differences in genetic contributions to the chromosomes from underlying ancestral populations, without making any assumptions of the evolution of the population of interest. The CAnD method takes into account correlated ancestries among chromosomes within individuals for both valid testing and improved power for detecting heterogeneity in population structure across the genome. Furthermore, X chromosome data can easily be incorporated in the analysis and the method can be used for testing heterogeneity in ancestry among any subset of chromosomes in the genome.

One possible extension of CAnD is to localize the regions of the genome that exhibit heterogeneity to identify regions that may be under selection. CAnD is implemented in

chromosomal-wide fashion, but could be used with smaller regions of the genome with the current implementation. Using a sliding window or grouping by genes, we compare ancestry in small regions of a chromosome of interest to the average ancestry across the remainder of the genome. In this way, we can fine-map regions of the genome that may be under selection.

Another extension of CAnD is to consider equilibrium on the X chromosome in terms of ancestry between males and females to detect non-random mating. With mating completely at random and under the most extreme ancestry proportions at the initial admixture event, after approximately eight generations, we expect the female and male X chromosome ancestry proportions to be equal in an admixed population (Figure 2.12). Applying this technique to ancestry proportions in individuals from admixed populations will allow us to detect whether mating appears to be at random. Many analyses assume HWE, thus random mating, and as a result, being able to detect non-random mating will allow for accurate checking of model assumptions.

Chapter 3 examined single SNP association testing methods in GWAS data across the autosomes and the X chromosome. In particular, when samples have non-zero heritability due to X chromosome genetic effects, we find that existing methods have increased type I error. We developed the MLM-X method, which extends previous MLM models to simultaneously model correlation on the X chromosome and the autosomes. When accounting for X chromosome genetic effects in a mixed linear models framework, our test is properly calibrated in terms of type I error. The method is valid whether we are testing an X chromosome or autosomal SNP for association. Furthermore, we do not lose power compared to existing methods by correcting for X chromosome effects.

In Chapter 4, we examined methods for gene-based association analysis. With applications of current methods to simulated data, we demonstrated that when samples are related and have complex structure, increased type I error can occur. We developed the KEATS-O method, a gene-based test valid in the presence of complex sample structure. KEATS-O adjusts for autosomal and X chromosome relatedness, and we recommend adjustment for population structure as estimated on the autosomes and X chromosome separately, as well. Under the 'omnibus' framework, we fit a linear combination of the variance component score

test and the burden test in order to obtain the highest possible power without knowing the genetic architecture of the trait *a priori*.

We applied KEATS-O to both common and rare SNPs assayed in the HCHS/SOL samples, grouped into genes. Interestingly, the KEATS-O results identified gene regions that were not previously identified with single variant association tests using MLM-X. We demonstrated that SNP genotype data can be used for the more traditional single SNP analysis as well as gene-based association testing. Especially when SNP genotype data are imputed to a larger set of variants that may have smaller MAF, a gene-based method for detecting association can identify genes associated with a trait of interest without requiring whole-genome or whole-exome sequencing data.

One possible extension of the proposed association testing methods is adjustment for and testing of dominance effects. Complex trait mapping usually focuses on testing under the assumption of additivity. However, MLM-X could jointly test the additive and dominance effects. Separate kinship matrices that measure the additive and dominance genetic relatedness could be included in the model, and estimation of variance components could be done as described.

The genotype codings assumed for the methods we developed are 0, 1, 2 for females and 0, 2 for males. These codings assume X chromosome inactivation (XCI) completely at random. In fact, some regions of the X chromosome could escape inactivation. Furthermore, the skewness of the X chromosome could be associated with disease, and it also could change with age [57]. To relax the assumption of XCI completely at random, the heterozygous female genotype could be coded as a value between 0 and 2, depending on whether the inactivation occurs less often or more often than random; the genotype codings would then be <1 or >1, respectively. Then, we could perform the MLM-X or KEATS-O tests, more properly taking into account XCI (or lack thereof). Resulting association testing methods would better reflect the biological processes occurring on the X chromosome.

All analyses considered here excluded SNPs in the pseudo-autosomal regions (PARs) on the X chromosome when estimating population structure and relatedness. Future work could explore whether information can be gained by using genetic variants in the PARs. Additionally, analyses could extend to genetic data assayed on the Y chromosome. Although

smaller, more repetitive and only present in males, the Y chromosome can be included in male-only analyses and could exhibit patterns of population structure and association. However, since the Y is non-recombining and is essentially a single locus, the gains may be minimal.

There is a rich literature of analysis methods for genome-wide data that is ever-growing. In the present work, we aimed to contribute methods that provide useful implementations of statistical theory motivated by real data that can be feasibly applied to genetic studies. The results from these studies contribute to our understanding of complex diseases and the human genome that underlies them.

BIBLIOGRAPHY

[1] G Bhatia, A Tandon, N Patterson, MC Aldrich, CB Ambrosone, et al. Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *American Journal of Human Genetics*, 95(4):437–444, 2014.

[2] SR Browning and BL Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[3] K Bryc, A Auton, MR Nelson, JR Oksenberg, SL Hauser, S Williams, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Science*, 107:786–791, 2010.

[4] K Bryc, EY Durand, M Macpherson, D Reich, and JL Mountain. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *American Journal of Human Genetics*, 96:37–53, 2015.

[5] LL Cavalli-Sforza. The Human Genome Diversity Project: Past, present and future. *Nature Reviews Genetics*, 6:333–340, 2005.

[6] H Chen, JB Meigs, and J Dupuis. Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology*, 37:196–204, 2013.

[7] Z Chen, H Tang, R Qayyum, UM Schick, MA Nalls, R Handsaker, et al. Genome-wide association analysis of red blood cell traits in African Americans: The COGENT Network. *Human Molecular Genetics*, 22:2529–2538, 2013.

[8] M Conomos. *Inferring, Estimating and Accounting for Population and Pedigree Structure in Genetic Analyses*. PhD thesis, University of Washington, 2013.

[9] MP Conomos, CA Laurie, AM Stilp, SM Gogarten, CP McHugh, SC Nelson, et al. Genetic diversity and association studies in U.S. Hispanic/Latino populations: Applications in the Hispanic Community Health Study/Study of Latinos. *American Journal of Human Genetics*, 98:165–184, 2016.

[10] MP Conomos, MB Miller, and TA Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39:276–293, 2015.

[11] MP Conomos, AP Reiner, BS Weir, and TA Thornton. Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics*, 98:127–148, 2016.

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

[13] K Ding, M de Andrade, TA Manolio, DC Crawford, LJ Rasmussen-Torvik, MD Ritchie, et al. Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: An electronic medical record-based genome-wide association study. *G3*, 3(7):1061–1068, 2013.

[14] A Goldberg and NA Rosenberg. Beyond 2/3 and 1/3: The complex signatures of sex-biased admixture on the X chromosome. *Genetics*, 201:263–279, 2015.

[15] International HapMap 3 consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.

[16] A Helgason, B Yngvadttir, B Hrafnkelsson, J Gulcher, and K Stefnsson. An Icelandic example of the impact of population structure on association studies. *Nature Genetics*, 37:90–95, 2005.

[17] LA Hindorff, J MacArthur, J Morales, HA Junkins, PN Hall, AK Klemm, and TA Manolio. A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies.

[18] J Jakobsdottir and MS McPeek. Mastor: Mixed-model association mapping of quantitative traits in samples with related individuals. *American Journal of Human Genetics*, 92:652–666, 2013.

[19] D Jiang and MS McPeek. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genetic Epidemiology*, 38:10–20, 2014.

[20] W Jin, S Xu, H Wang, Y Yu, Y Shen, et al. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Research*, 22:27–519, 2012.

[21] NA Johnson, MA Coram, MD Shriver, I Romieu, GS Barsh, et al. Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genetics*, 7(12):e1002410, 2011.

[22] HM Kang, JH Sul, SK Service, NA Zaitlen, S Kong, NB Freimer, C Sabatti, and E Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42:348–354, 2010.

[23] D Kuonen. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):929–935, 1999.

[24] CC Laurie, KF Doheny, DB Mirel, EW Pugh, LJ Bierut, T Bhangale, F Boehm, NE Caporaso, MC Cornelis, and HJ et al Edenberg. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34:591–602, 2010.

[25] LM Lavange, WD Kalsbeek, PD Sorlie, LM Aviles-Santa, RC Kaplan, J Barnhart, K Liu, A Giachello, DJ Lee, J Ryan, et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, 20:642–649, 2010.

[26] S Lee, MC Wu, and X Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012.

[27] B Li and S Leal. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, 83:311–321, 2008.

[28] J Li, JT Glessner, H Zhang, C Hou, Z Wei, JP Bradfield, FD Mentch, et al. GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. *Human Molecular Genetics*, 22(7):1457–1464, 2013.

[29] C Lippert, J Listgarten, Y Liu, CM Kadie, RL Davidson, and D Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8:833–835, 2011.

[30] D Liu, X Lin, and D Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63:1079–1088, 2007.

[31] H Liu, Y Tang, and HH Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis*, 53:853–856, 2009.

[32] KS Lo, JG Wilson, LA Lange, AR Folsom, G Galarneau, SK Ganesh, et al. Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. *Human Genetics*, 129(3):307–317, 2011.

[33] BE Madsen and SR Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2), 2009.

[34] A Manichaikul, JC Mychaleckyj, SS Rich, K Daly, M Sale, and WM Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26:28672873, 2010.

[35] A Manichaikul, W Palmas, CJ Rodriguez, CA Peralta, J Divers, X Guo, WM Chen, et al. Population structure of Hispanics in the United States: The Multi-Ethnic Study of Atherosclerosis. *PLoS Genetics*, 8:e1002640, 2012.

[36] TA Manolio et al. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

[37] BK Maples, S Gravel, EE Kenny, and CD Bustamante. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics*, 93:278–288, 2013.

[38] G Montana. HapSim: A simulation tool for generating haplotype data with prespecified allele frequencies and LD coefficients. *Bioinformatics*, 21(23):4309–4311, 2005.

[39] S Morgenthaler and WG Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56, 2007.

[40] AP Morris and E Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193, 2010.

[41] BM Neale, MA Rivas, BF Voight, D Altshuler, B Devlin, M Orho-Melander, et al. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322, 2011.

[42] M Nelis, T Esko, R Magi, F Zimprich, A Zimprich, D Toncheva, et al. Genetic structure of Europeans: A view from the North-East. *PLoS One*, 4(5):e5472, 2009.

[43] DL Newman, M Abney, MS McPeek, C Ober, and NJ Cox. The importance of genealogy in determining genetic associations with complex traits. *American Journal of Human Genetics*, 69:1146–1148, 2001.

[44] J Novembre, T Johnson, K Bryc, Z Kutalik, AR Boyko, A Auton, et al. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.

[45] AL Price et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

[46] AL Price, GV Kryukov, PIW de Bakker, SM Purcell, J Staples, L-J Wei, and SR Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics*, 86(6):832–838, 2010.

[47] AL Price, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, and D Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

[48] AL Price, A Tandon, N Patterson, KC Barnes, N Rafaels, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6), 2009.

[49] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[50] N Risch, S Choudhry, M Via, A Basu, R Sebro, C Eng, K Beckman, et al. Ancestry-related assortative mating in Latino populations. *Genome Biology*, 10:R132, 2009.

[51] KR Rosenbloom, J Armstrong, GP Barber, J Casper, H Clawson, M Diekhans, TR Dreszer, PA Fujita, L Guruvadoo, M Haeussler, et al. The UCSC genome browser database: 2015 update. *Nucleic Acids Research*, 43:D670–81, 2015.

[52] R Sebro, TJ Hoffman, C Lange, JJ Rogus, and NJ Risch. Testing for non-random mating: Evidence for ancestry-related assortative mating in the Framingham heart study. *Genetic Epidemiology*, 34:674–679, 2010.

[53] PD Sorlie, LM Aviles-Santa, S Wassertheil-Smoller, RC Kaplan, ML Daviglus, AL Giachello, N Schneiderman, L Raij, G Talavera, M Allison, et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, 20:629–641, 2010.

[54] H Tang, S Choudhry, R Mei, M Morgan, W Rodriguez-Cintron, EG Burchard, and NJ Risch. Recent genetic selection in the ancestral admixture of Puerto Ricans. *American Journal of Human Genetics*, 81:626–633, 2007.

[55] H Tang, J Peng, P Wang, and NJ Risch. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28:289–301, 2005.

[56] T Thornton, H Tang, TJ Hoffman, HM Ochs-Balcom, BJ Caan, and N Risch. Estimating kinship in admixed populations. *American Journal of Human Genetics*, 91(1):122–138, 2012.

[57] J Wang, R Yu, and S Shete. X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genetic Epidemiology*, 38(6):483–93, 2014.

[58] Y Wang, Y-H Chen, and Q Yang. Joint rare variant association test of the average and individual effects for sequencing studies. *PLoS ONE*, 7(3):e32485, 2012.

[59] AL Wise, L Gyi, and TA Manolio. eXclusion: Toward integrating the X chromosome in genome-wide association analyses. *American Journal of Human Genetics*, 92:643–647, 2013.

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

[61] J Yang, MN Weedon, S Purcell, G Lettre, K Estrada, and CJand others Willer. Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics*, 19:807812, 2011.

[62] J Yu, G Pressoir, WH Briggs, IV Bi, M Yamasaki, JF Doebley, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38:203–208, 2006.

[63] F Zakharia, A Basu, D Absher, TL Assimes, AS Go, MA Hlatky, et al. Characterizing the admixed African ancestry of African Americans. *Genome Biology*, 10:R141, 2009.

[64] D Zhang and X Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 1:57–74, 2003.

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

[66] X Zhu, S Li, RS Cooper, and RC Elston. A unified association analysis approach for family and unrelated samples correcting for stratification. *American Journal of Human Genetics*, 82:352365, 2008.

Appendix A

DERIVATION OF THE COVARIANCE OF GENOTYPES

We can calculate the variance of a SNP genotype under prespecified genotype codings. For an autosomal or X chromosome SNP, Table A.1 displays the genotype coding and allele frequency for a particular genotype, assuming the frequency of the 'A' allele is $(1-p)$ and the 'B' allele is p. Under the assumption of Hardy-Weinberg Equilibrium (HWE), the expectation and variance of an autosomal SNP genotype is

$$\begin{aligned}
\mathbb{E}(\text{SNP}) &= 0(1-p)^2 + 1(2p(1-p)) + 2p^2 && \text{(A.1)}\\
&= 2p && \text{(A.2)}\\
\text{var}(\text{SNP}) &= \mathbb{E}(\text{SNP}^2) - \mathbb{E}(\text{SNP})^2 && \text{(A.3)}\\
&= 2p(1-p) + 4p^2 - 4p^2 && \text{(A.4)}\\
&= 2p(1-p) && \text{(A.5)}
\end{aligned}$$

It is of interest to calculate the mean and variance of an X chromosome SNP genotype. Assume female genotype codings are the same as described for an autosomal SNP. For males, let genotype codings be 0 and 2. Table A.1 shows the frequencies for each observed genotype under these codings. The mean and variance of an X chromosome SNP can be

Genotype	Autosomal	X chromosome, Female	X chromosome, Male
0	$(1-p)^2$	$(1-p)^2$	$(1-p)$
1	$2p(1-p)$	$2p(1-p)$	
2	p^2	p^2	p

Table A.1: Genotype codings for autosomal and X chromosome SNPs where the 'A' allele has frequency $(1-p)$ and the 'B' allele has frequency p assuming HWE.

calculated conditionally based on whether the sample is male or female

$$\mathbb{E}(\text{SNP}_x^F) = 2p^2 + 2p(1-p) = 2p \tag{A.6}$$

$$\mathbb{E}(\text{SNP}_x^M) = 2p \tag{A.7}$$

$$var(\text{SNP}_x^F) = \mathbb{E}((\text{SNP}_x^F)^2) - \mathbb{E}^2(\text{SNP}_x^F) \tag{A.8}$$

$$= 4p^2 + 2p(1-p) - (2p)^2 \tag{A.9}$$

$$= 2p(1-p) \tag{A.10}$$

$$var(\text{SNP}_x^M) = \mathbb{E}((\text{SNP}_x^M)^2) - \mathbb{E}^2(\text{SNP}_x^M) \tag{A.11}$$

$$= 4p - (2p)^2 \tag{A.12}$$

$$= 4p(1-p) \tag{A.13}$$

To calculate the covariance of genotypes between a pair of individuals, we must consider their sex. In what follows, I am denoting the X chromosome kinship value between a pair of males as $\Phi^X_{M,M}$, a pair of females as $\Phi^X_{F,F}$ and one male and one female as $\Phi^X_{F,M}$. First,

we calculate the covariance for a SNP between a pair of males as

$$cov(\text{SNP}_x^M, \text{SNP}_x^M) = \mathbb{E}(\text{SNP}_x^M \text{SNP}_x^M) - \mathbb{E}^2(\text{SNP}_x^M) \tag{A.14}$$
$$= \mathbb{E}(\text{SNP}_x^M \text{SNP}_x^M | \text{IBD})\mathbb{P}(\text{IBD}) \tag{A.15}$$
$$+ \mathbb{E}(\text{SNP}_x^M \text{SNP}_x^M | \text{no IBD})\mathbb{P}(\text{no IBD}) - (2p)^2 \tag{A.16}$$
$$= \mathbb{E}(\text{SNP}_x^M \text{SNP}_x^M | \text{IBD})\Phi^X_{M,M} \tag{A.17}$$
$$+ \mathbb{E}(\text{SNP}_x^M \text{SNP}_x^M | \text{no IBD})(1 - \Phi^X_{M,M}) - 4p^2 \tag{A.18}$$
$$= 4p\Phi^X_{M,M} + 4p^2(1 - \Phi^X_{M,M}) - 4p^2 \tag{A.19}$$
$$= 4p(1-p)\Phi^X_{M,M} \tag{A.20}$$

Next we consider the covariance between a pair of female genotypes

$$cov(\text{SNP}_x^F, \text{SNP}_x^F) = \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^F) - \mathbb{E}^2(\text{SNP}_x^F) \tag{A.21}$$
$$= \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^F | \text{IBD})\mathbb{P}(\text{IBD}) \tag{A.22}$$
$$+ \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^F | \text{no IBD})\mathbb{P}(\text{no IBD}) - (2p)^2 \tag{A.23}$$
$$= (2(2p(1-p)) + 4p^2)\Phi^X_{F,F} + 4p^2(1 - \Phi^X_{F,F}) - 4p^2 \tag{A.24}$$
$$= 4p(1-p)\Phi^X_{F,F} \tag{A.25}$$

Finally, we calculate the covariance between a pair of genotypes where one is a female and one is a male

$$cov(\text{SNP}_x^F, \text{SNP}_x^M) = \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^M) - \mathbb{E}(\text{SNP}_x^F)\mathbb{E}(\text{SNP}_x^M) \tag{A.26}$$
$$= \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^M | \text{IBD})\mathbb{P}(\text{IBD}) \tag{A.27}$$
$$+ \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^M | \text{no IBD})\mathbb{P}(\text{no IBD}) - (2p)^2 \tag{A.28}$$
$$= [4p^2 + 2(2p(1-p))]\Phi^X_{F,M} + [4p^3 + 4p^2(1-p)](1 - \Phi^X_{F,M}) - 4p^2 \tag{A.29}$$
$$= 4p(1-p)\Phi^X_{F,M} \tag{A.30}$$

We can now see that using the self X chromosome kinship values from Table 3.1, the variance for a female and male SNP is indeed as calculated in Equations A.10 and A.13

after incorporating the self-kinship values. Finally, we find the variance for a given individual i on an X chromosome SNP, in the MLM-X model, to be

$$var(y_i) = \beta_k^2 4p(1-p)\Phi_i^X + \sigma_X^2 + \sigma_A^2 + \sigma_\epsilon^2 \tag{A.31}$$

Appendix B

DERIVATION OF THE KEATS SCORE STATISTIC

We assume the KEATS model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{g}_A + \mathbf{g}_X + \boldsymbol{\epsilon} \tag{B.1}$$

where $\boldsymbol{\gamma} \sim (0, \tau\mathbf{W}), \mathbf{g}_A \sim \mathcal{N}(0, \sigma_A^2\boldsymbol{\Phi}_\mathbf{A}), \mathbf{g}_X \sim \mathcal{N}(0, \sigma_X^2\boldsymbol{\Phi}_\mathbf{X}), \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2\mathbb{I})$. Denote the matrix of observed genotype values for genetic variants in a region of interest $\mathbf{G}$ and the matrices of pairwise kinship coefficients across the autosomes and X chromosome by $\boldsymbol{\Phi}_\mathbf{A}$ and $\boldsymbol{\Phi}_\mathbf{X}$, respectively. To test a gene region, we examine whether we have evidence to reject the hypothesis that none of the variants in the region are associated with the disease outcome of interest. Mathematically, we can write this hypothesis as $H_0 : \boldsymbol{\gamma} = 0$. Equivalently, we can test $H_0 : \tau = 0$ against $H_1 : \tau > 0$ using a variance component score test under the mixed linear models framework. The variance of our phenotype $\mathbf{y}$ is

$$\begin{aligned}
\text{var}(\mathbf{y}) &= \mathbf{G}\text{var}(\gamma)\mathbf{G}^T + \sigma_A^2\boldsymbol{\Phi}_\mathbf{A} + \sigma_X^2\boldsymbol{\Phi}_\mathbf{X} + \sigma_\epsilon^2\mathbb{I} && \text{(B.2)}\\
&= \tau\mathbf{G}\mathbf{W}\mathbf{G}^\mathbf{T} + \sigma_A^2\boldsymbol{\Phi}_\mathbf{A} + \sigma_X^2\boldsymbol{\Phi}_\mathbf{X} + \sigma_\epsilon^2\mathbb{I} && \text{(B.3)}\\
&\equiv \boldsymbol{\Sigma} && \text{(B.4)}
\end{aligned}$$

Recall some properties of matrices, in particular

$$\begin{aligned}
\frac{\partial\boldsymbol{\Sigma}}{\partial\tau} &= \mathbf{G}\mathbf{W}\mathbf{G}^T && \text{(B.5)}\\
\frac{\partial}{\partial x}\ln|\mathbf{A}| &= \text{tr}(\mathbf{A}^{-1}\frac{d\mathbf{A}}{dx}) && \text{(B.6)}\\
|\boldsymbol{\Sigma}|^T &= \text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^T)|\boldsymbol{\Sigma}| && \text{(B.7)}\\
(\log|\boldsymbol{\Sigma}|)^T &= \text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^T) && \text{(B.8)}
\end{aligned}$$

and

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} = \mathbb{I} \tag{B.9}$$

$$\text{take derivative wrt } \tau\text{: } \boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \tau} + \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \tau}\boldsymbol{\Sigma} = 0 \tag{B.10}$$

$$\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \tau}\boldsymbol{\Sigma} = -\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \tau} \tag{B.11}$$

$$\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \tau} = -\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \tau}\boldsymbol{\Sigma}^{-1} \tag{B.12}$$

Under the KEATS model, we can find the likelihood, log-likelihood and first derivative of the log-likelihood with respect to the parameter τ,

$$f(\mathbf{y}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta)) \tag{B.13}$$

$$\ell = \log(f(\mathbf{y})) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta) \tag{B.14}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \tau} &= -\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \tau}) - \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T(-\boldsymbol{\Sigma}^{-1})\frac{\partial \boldsymbol{\Sigma}}{\partial \tau}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta) \\ &= -\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T) + \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta)\end{aligned} \tag{B.15}$$

We will use REML to estimate the parameters in our model to adjust for possible small sample bias. To do so, we restrict our n observations to $n-p$ observations, where $p = \text{rank}(\mathbf{X})$. The REML log-likelihood equivalent to the maximum log-likelihood (Equation B.14) is

$$\begin{aligned}\ell_{\text{REML}} = &-\frac{1}{2}(n-p)\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| \\ &-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta) - \frac{1}{2}\log(|\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}|)\end{aligned} \tag{B.16}$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial \tau} =& -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \tau}) - \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T(-\boldsymbol{\Sigma}^{-1})\frac{\partial \boldsymbol{\Sigma}}{\partial \tau}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta) \\
& -\frac{1}{2}\mathrm{tr}((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\frac{\partial \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}}{\partial \tau}) \\
=& -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T) + \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta) \\
& -\frac{1}{2}\mathrm{tr}((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T(-\boldsymbol{\Sigma}^{-1})\frac{\partial \boldsymbol{\Sigma}}{\partial \tau}\boldsymbol{\Sigma}^{-1}\mathbf{X}) \\
=& -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T) + \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta) \\
& -\frac{1}{2}\mathrm{tr}((\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(-\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T\mathbf{X}^T\mathbf{X}) \\
=& -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T(\mathbb{I}-\boldsymbol{\Sigma}^{-1}\mathbf{X}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X})) \\
& +\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\beta) \\
=& -\frac{1}{2}\mathrm{tr}(\mathbf{M}\boldsymbol{\Sigma}-\mathbf{M}\mathbf{X}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X})) + \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{M}(\mathbf{y}-\mathbf{X}\beta)
\end{aligned}
\tag{B.17}
$$

where $\mathbf{M} = \boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}^T\boldsymbol{\Sigma}^{-1}$. We can ignore the first term, as it doesn't depend on $\mathbf{y}$. Twice the second term is the REML version of the score statistic with weighted linear kernel matrix $\mathbf{M}$

$$
Q_{\mathrm{KEATS}} = (\mathbf{y}-\mathbf{X}\beta)^T\mathbf{M}(\mathbf{y}-\mathbf{X}\beta) \tag{B.18}
$$

Q_{KEATS} is a quadratic function of $\mathbf{y}$ and follows a mixture of chi-squared random variables under H_0 [30].

To get the p-values for this statistic under the REML framework, we can think of first normalizing the residuals $\mathbf{y}-\mathbf{X}\beta$ so we define

$$
\tilde{\mathbf{y}} = \mathbf{y}-\mathbf{X}\beta \tag{B.19}
$$

$$
\sim \mathcal{N}(0, \boldsymbol{\Sigma}-\mathbf{P}\boldsymbol{\Sigma}) \tag{B.20}
$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}$ is the projection matrix. It follows

$$
(\boldsymbol{\Sigma}-\mathbf{P}\boldsymbol{\Sigma})^{1/2}\tilde{\mathbf{y}} \sim \mathcal{N}(0, \mathbb{I}) \tag{B.21}
$$

Then, replacing $\mathbf{y}$ by $\tilde{\mathbf{y}}$ in Equation B.17 we find

$$\begin{aligned}
\ell_{\text{REML}} =& -\frac{1}{2}\text{tr}(\mathbf{M}\boldsymbol{\Sigma} - \mathbf{M}\mathbf{X}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X})) + \frac{1}{2}\tilde{\mathbf{y}}^T(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\mathbf{M}(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\tilde{\mathbf{y}} \\
=& \frac{1}{2}\text{tr}(\mathbf{M}(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})) + \frac{1}{2}\tilde{\mathbf{y}}^T(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\mathbf{M}(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\tilde{\mathbf{y}} \\
=& \frac{1}{2}\text{tr}((\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\mathbf{M}(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}) + \frac{1}{2}\tilde{\mathbf{y}}^T(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\mathbf{M}(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\tilde{\mathbf{y}} \\
=& \frac{1}{2}\text{tr}(\widetilde{\mathbf{M}}) + \frac{1}{2}\tilde{\mathbf{y}}^T\widetilde{\mathbf{M}}\tilde{\mathbf{y}}
\end{aligned} \tag{B.22}$$

where $\widetilde{\mathbf{M}} = (\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}\mathbf{M}(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma})^{1/2}$. Now, if we find the p non-zero eigenvalues of $\widetilde{\mathbf{M}}$, $\lambda_1, \ldots, \lambda_p$ we know that $\ell_{\text{REML}} \stackrel{d}{=} \sum_{i=1}^{p} \lambda_i \chi^2_{1,i}$ [64] where the degrees of freedom depend on the index i. The p-value of this statistic can be found numerically using Satterthwaite's approximation method as implemented in `R` [49] and described by Kuonen [23]. The general method to find the p-value is to invert the characteristic, or moment generating, function and then to calculate the p-value analytically. Another analytic approach is to use moment-matching.

Appendix C

DERIVATION OF THE KEATS-O TEST STATISTIC P-VALUE

The proposed KEATS-O test statistic is

$$T_\rho = (\mathbf{y} - \mathbf{X}\hat{\beta})^T \widehat{\mathbf{\Sigma}}^{-1} \mathbf{G} \mathbf{W}^{1/2} \mathbf{R}_\rho \mathbf{W}^{1/2} \mathbf{G}^T \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{C.1}$$

Define the usual hat/projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \widehat{\mathbf{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\mathbf{\Sigma}}^{-1}$ such that

$$(\mathbf{y} - \mathbf{X}\hat{\beta}) = (\mathbb{I} - \mathbf{P})\mathbf{y} = (\mathbb{I} - \mathbf{P})(\mathbf{y} - \mathbf{X}\beta) \tag{C.2}$$

Now, we know that since $\widehat{\mathbf{\Sigma}}$ is PSD, a matrix $\widehat{\mathbf{\Sigma}}^{1/2}$ exists such that $\widehat{\mathbf{\Sigma}}^{-1/2} \widehat{\mathbf{\Sigma}} (\widehat{\mathbf{\Sigma}}^{-1/2})^T = \mathbb{I}$. Further, we know as $n \to \infty$,

$$(\mathbf{y} - \mathbf{X}\hat{\beta}) \sim \mathcal{N}(0, (\mathbb{I} - \mathbf{P})\widehat{\mathbf{\Sigma}}) \tag{C.3}$$

$$\widetilde{\mathbf{y}} = ((\mathbb{I} - \mathbf{P})\widehat{\mathbf{\Sigma}})^{1/2} (\mathbf{y} - \mathbf{X}\hat{\beta}) \sim \mathcal{N}(0, \mathbb{I}) \tag{C.4}$$

Thus, we find

$$T_\rho = \widetilde{\mathbf{y}}^T ((\mathbb{I} - \mathbf{P})\widehat{\mathbf{\Sigma}})^{1/2} \widehat{\mathbf{\Sigma}}^{-1} \mathbf{G} \mathbf{W}^{1/2} \mathbf{R}_\rho \mathbf{W}^{1/2} \mathbf{G}^T \widehat{\mathbf{\Sigma}}^{-1} ((\mathbb{I} - \mathbf{P})\widehat{\mathbf{\Sigma}})^{1/2} \widetilde{\mathbf{y}} \tag{C.5}$$

$$= \widetilde{\mathbf{y}}^T \mathbf{Z} \mathbf{R}_\rho \mathbf{Z}^T \widetilde{\mathbf{y}} \tag{C.6}$$

where $\mathbf{Z} = (\widehat{\mathbf{\Sigma}}^{1/2})(\mathbb{I} - \mathbf{P})\widehat{\mathbf{\Sigma}}^{-1} \mathbf{G} \mathbf{W}^{1/2}$. Note $\mathbf{W}$ is a symmetric matrix such that $\mathbf{W} = \mathbf{W}^T$. By writing T_ρ in this way, we see that T_ρ is asymptotically distributed as $\sum_{i=1}^m \lambda_i \chi^2_{1,i}$ where λ_is are the m non-zero eigenvalues of $\mathbf{Z}\mathbf{R}_\rho\mathbf{Z}^T$; we know how to approximate the p-values of T_ρ using the same methods we do for KEATS (see Appendix B and [64]).

Now, we want to rewrite T_ρ as a weighted linear combination of Q_{KEATS} and $Q_{\text{KEATS-bt}}$, $T_\rho = (1-\rho)Q_{\text{KEATS}} + \rho Q_{\text{KEATS-bt}}$. We note that $R_\rho = (1-\rho)\mathbb{I} + \rho \mathbf{1}\mathbf{1}^T$. We then find

$$T_\rho = \widetilde{\mathbf{y}}^T \mathbf{Z} \mathbf{R}_\rho \mathbf{Z}^T \widetilde{\mathbf{y}} \tag{C.7}$$

$$= \widetilde{\mathbf{y}}^T \mathbf{Z} \{(1-\rho)\mathbb{I} + \rho \mathbf{1}\mathbf{1}^T\} \mathbf{Z}^T \widetilde{\mathbf{y}} \tag{C.8}$$

$$= (1-\rho)\widetilde{\mathbf{y}}^T \mathbf{Z}\mathbf{Z}^T \widetilde{\mathbf{y}} + \rho \widetilde{\mathbf{y}} \mathbf{Z} \mathbf{1}\mathbf{1}^T \mathbf{Z}^T \widetilde{\mathbf{y}} \tag{C.9}$$

We exploit the properties of $\mathbf{1}$ and define $\bar{\mathbf{z}} = \frac{1}{a}\mathbf{Z1}$ as the row means of the matrix $\mathbf{Z}$. It follows $\mathbf{Z1} = a\bar{\mathbf{z}}$. Then, Equation C.9 becomes

$$T_\rho = (1-\rho)\widetilde{\mathbf{y}}^T\mathbf{Z}\mathbf{Z}^T\widetilde{\mathbf{y}} + \rho\widetilde{\mathbf{y}}^T(a\bar{\mathbf{z}})(a\bar{\mathbf{z}})^T\widetilde{\mathbf{y}} \tag{C.10}$$

$$= (1-\rho)\widetilde{\mathbf{y}}^T\mathbf{Z}\mathbf{Z}^T\widetilde{\mathbf{y}} + \rho a^2\widetilde{\mathbf{y}}^T\bar{\mathbf{z}}\bar{\mathbf{z}}^T\widetilde{\mathbf{y}} \tag{C.11}$$

We yet again rewrite T_ρ, this time aiming to find terms of asymptotically quadratic forms. In order to do that, we focus on the first term, and in particular $\mathbf{Z}\mathbf{Z}^T$. Define $\mathbf{A}$ to be the matrix that projects into the space spanned by the row means of $\mathbf{Z}$ such that $\mathbf{A} = \bar{\mathbf{z}}(\bar{\mathbf{z}}^T\bar{\mathbf{z}})^{-1}\bar{\mathbf{z}}^T$.

$$\mathbf{Z}\mathbf{Z}^T = \mathbf{Z}\mathbf{Z}^T - 2\mathbf{Z}\mathbf{Z}^T\mathbf{A} + \mathbf{M}\mathbf{Z}\mathbf{Z}^T\mathbf{A} + 2\mathbf{Z}\mathbf{Z}^T\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^T\mathbf{A} \tag{C.12}$$

$$= (\mathbf{Z}\mathbf{Z}^T - \mathbf{A}\mathbf{Z}\mathbf{Z}^T)(\mathbb{I} - \mathbf{A}) + 2\mathbf{Z}\mathbf{Z}^T\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^T\mathbf{A} \tag{C.13}$$

$$= (\mathbb{I} - \mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I} - \mathbf{A}) + 2(\mathbf{Z}\mathbf{Z}^T\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^T\mathbf{A}) + \mathbf{A}\mathbf{Z}\mathbf{Z}^T\mathbf{A} \tag{C.14}$$

$$= (\mathbb{I} - \mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I} - \mathbf{A}) + 2(\mathbb{I} - \mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A} + \mathbf{A}\mathbf{Z}\mathbf{Z}^T\mathbf{A} \tag{C.15}$$

Now, we note that $\mathbf{A}\mathbf{Z}\mathbf{Z}^T\mathbf{A} = \frac{\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{(\bar{\mathbf{z}}^T\bar{\mathbf{z}})^2}||\mathbf{1}^T\mathbf{Z}^T\mathbf{Z}||_2^2$ so Equation C.15 becomes

$$\mathbf{Z}\mathbf{Z}^T = (\mathbb{I} - \mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I} - \mathbf{A}) + 2(\mathbb{I} - \mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A} + \frac{\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{(\bar{\mathbf{z}}^T\bar{\mathbf{z}})^2}||\mathbf{1}^T\mathbf{Z}^T\mathbf{Z}||_2^2 \tag{C.16}$$

To get the final form of T_ρ, we can substitute Equation C.16 into the first term of Equa-

tion C.11 to get

$$T_\rho = (1-\rho)\widetilde{\mathbf{y}}^T\Big((\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A}) + 2(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A} + \frac{\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{(\bar{\mathbf{z}}^T\bar{\mathbf{z}})^2}||\mathbf{1}^T\mathbf{Z}^T\mathbf{Z}||_2^2\Big)\widetilde{\mathbf{y}} \tag{C.17}$$

$$+ \rho a^2\widetilde{\mathbf{y}}^T\bar{\mathbf{z}}\bar{\mathbf{z}}^T\widetilde{\mathbf{y}}$$

$$= (1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A})\widetilde{\mathbf{y}} + 2(1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A}\widetilde{\mathbf{y}} \tag{C.18}$$

$$+ (1-\rho)\widetilde{\mathbf{y}}^T\frac{\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{(\bar{\mathbf{z}}^T\bar{\mathbf{z}})^2}||\mathbf{1}^T\mathbf{Z}^T\mathbf{Z}||_2^2\widetilde{\mathbf{y}} + \rho a^2\widetilde{\mathbf{y}}^T\bar{\mathbf{z}}\bar{\mathbf{z}}^T\widetilde{\mathbf{y}}$$

$$= (1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A})\widetilde{\mathbf{y}} + 2(1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A}\widetilde{\mathbf{y}} \tag{C.19}$$

$$+ \widetilde{\mathbf{y}}^T\Big(\frac{(1-\rho)\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{(\bar{\mathbf{z}}^T\bar{\mathbf{z}})^2}||\mathbf{1}^T\mathbf{Z}^T\mathbf{Z}||_2^2 + \rho a^2\bar{\mathbf{z}}\bar{\mathbf{z}}^T\Big)\widetilde{\mathbf{y}}$$

$$= (1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A})\widetilde{\mathbf{y}} + 2(1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A}\widetilde{\mathbf{y}} \tag{C.20}$$

$$+ \widetilde{\mathbf{y}}^T\frac{\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{\bar{\mathbf{z}}^T\bar{\mathbf{z}}}\Big(\frac{1-\rho}{\bar{\mathbf{z}}^T\bar{\mathbf{z}}}||\mathbf{1}^T\mathbf{Z}^T\mathbf{Z}||_2^2 + \rho a^2\bar{\mathbf{z}}^T\bar{\mathbf{z}}\Big)\widetilde{\mathbf{y}}$$

$$= (1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A})\widetilde{\mathbf{y}} + 2(1-\rho)\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A}\widetilde{\mathbf{y}} + \tau(\rho)\widetilde{\mathbf{y}}^T\frac{\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{\bar{\mathbf{z}}^T\bar{\mathbf{z}}}\widetilde{\mathbf{y}} \tag{C.21}$$

$$= (1-\rho)\Big(\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A})\widetilde{\mathbf{y}} + 2\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A}\widetilde{\mathbf{y}}\Big) + \tau(\rho)\widetilde{\mathbf{y}}^T\frac{\bar{\mathbf{z}}\bar{\mathbf{z}}^T}{\bar{\mathbf{z}}^T\bar{\mathbf{z}}}\widetilde{\mathbf{y}} \tag{C.22}$$

where $\tau(\rho) = \frac{1-\rho}{\bar{\mathbf{z}}^T\bar{\mathbf{z}}}||\mathbf{1}^T\mathbf{Z}^T\mathbf{Z}||_2^2 + \rho a^2\bar{\mathbf{z}}^T\bar{\mathbf{z}}$. This is the final form of T_ρ we desire, in that the first and second terms are asymptotically independent quadratic (recall $\widetilde{\mathbf{y}} = ((\mathbb{I}-\mathbf{P})\widehat{\mathbf{\Sigma}})^{1/2}(\mathbf{y} - \mathbf{X}\beta) \to \mathcal{N}(0,\mathbb{I})$). The first term is asymptotically a mixture of χ^2 distributions, similar to what we saw in Equation C.6, and the asymptotic distribution is $\sum_{i=1}^{l}\lambda_i\chi^2_{1,i}$ where λ_i are the l non-zero eigenvectors of $(\mathbb{I}-\mathbf{M})\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{M})$ which are equivalent to the non-zero eigenvectors of $\mathbf{Z}^T(\mathbb{I}-\mathbf{M})\mathbf{Z}$. The second term is also asymptotically quadratic although doesn't follow a mixture of weighted χ^2 distributions, as we don't have a valid kernel matrix. We can, however, use what we know about quadratic forms to find the asymptotic variance and mean. From properties of matrices and MVN distributions, we know $\text{var}(\mathbf{X}^T\mathbf{B}\mathbf{X}) = 2\text{tr}(\mathbf{B}^2)$ when $\text{var}(\mathbf{X}) = \mathbb{I}$ and $\text{mean}(\mathbf{X}) = 0$. Thus, in our case, $\text{var}(\widetilde{\mathbf{y}}^T(\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A}\widetilde{\mathbf{y}}) = 2\text{tr}\Big(((\mathbb{I}-\mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A})^2\Big) = 2\text{tr}(\mathbf{Z}^T\mathbf{A}\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A})\mathbf{Z})$ since $\mathbf{A}$ and $(\mathbb{I}-\mathbf{A})$ are idempotent. Thus, the asymptotic variance of the second term is $4\text{tr}(\mathbf{Z}^T\mathbf{A}\mathbf{Z}\mathbf{Z}^T(\mathbb{I}-\mathbf{A})\mathbf{Z})$. Taking these facts into account, the asymptotic distribution of T_ρ is approximately

$$(1-\rho)(\sum_{i=1}^{l}\lambda_i\chi^2_{1,i} + \xi) + \tau(\rho)\chi^2_1 \tag{C.23}$$

where $\xi = 2\widetilde{\mathbf{y}}^T(\mathbb{I} - \mathbf{A})\mathbf{Z}\mathbf{Z}^T\mathbf{A}\widetilde{\mathbf{y}}$ has the asymptotic mean of zero and variance as derived above.

The next step is assess significance of an observed test statistic T_ρ, which we will do for each of the terms in the approximate asymptotic distribution shown in Equation C.23.

$$\mathbb{P}(\sum_{i=1}^{l} \lambda_i \chi^2_{1,i} + \xi < c) \approx \mathbb{P}\Big(\sum_{i=1}^{l} \lambda_i \chi^2_{1,i} < \frac{c - (\mu_A + \mu_\xi)}{\sqrt{\sigma_A^2 + \sigma_\xi^2}} \sigma_A + \mu_A + \mu_\xi \Big) \tag{C.24}$$

by Equation (4) in [31], where

$$\mu_A = \mathbb{E}(\sum_{i=1}^{l} \lambda_i \chi^2_{1,i}) = \sum_{i=1}^{l} \lambda_i \tag{C.25}$$

$$\mu_\xi = 0 \tag{C.26}$$

$$\sigma_A^2 = \text{var}(\sum_{i=1}^{l} \lambda_i \chi^2_{1,i}) = 2\sum_{i=1}^{l} \lambda_i^2 \tag{C.27}$$

$$\sigma_\xi^2 = 4\text{tr}(\mathbf{Z}^T\mathbf{M}\mathbf{Z}\mathbf{Z}^T(\mathbb{I} - \mathbf{M})\mathbf{Z}) \tag{C.28}$$

Now, we have all components needed to calculate the p-value for our overall optimal test statistic, $T = \min\{p_{\rho_1}, \ldots, p_{\rho_b}\}$ where p_ρ is the p-value corresponding to the test statistic T_ρ as shown in Equation C.1. Let $q_{min}(\rho)$ be the critical value corresponding to the $(1-T)^{th}$ percentile of the distribution of T_ρ. Then,

$$\mathbb{P}(T > c) = \mathbb{P}(\min\{p_{\rho_1}, \ldots, p_{\rho_b}\} > c) \tag{C.29}$$

$$= \mathbb{P}(p_{\rho_1} > c, \ldots, p_{\rho_b} > c) \tag{C.30}$$

$$= 1 - \mathbb{P}(p_{\rho_1} < c, \ldots, p_{\rho_b} < c) \tag{C.31}$$

$$= 1 - \mathbb{P}(T_{\rho_1} < q_{min}(\rho_1), \ldots, T_{\rho_b} < q_{min}(\rho_b)) \tag{C.32}$$

$$= 1 - \mathbb{E}\Big[\mathbb{P}\Big(\sum_{i=1}^{l} \lambda_i \chi^2_{1,i} + \xi < \min\{\frac{q_{min}(\rho_\nu) - \rho_\nu \eta_0}{1 - \rho_\nu}\} | \eta_0 \Big)\Big] \tag{C.33}$$

$$= 1 - \mathbb{E}\Big[\mathbb{P}\Big(\sum_{i=1}^{l} \lambda_i \chi^2_{1,i} + \xi < c(\eta_0) | \eta_0 \Big)\Big] \tag{C.34}$$

$$= 1 - \int F(\delta(x)|\lambda) f(x|\chi^2_1) dx \tag{C.35}$$

where $\delta(x)$ is the critical value shown in Equation C.24 with $c = c(\eta_0)$, namely $\delta(x) =$

$\left(\min\{\frac{q_{min}(\rho_\nu)-\rho_\nu x}{1-\rho_\nu}\} - \mu_A\right)\frac{\sigma_A}{\sqrt{\sigma_A^2+\sigma_\xi^2}} + \mu_A$, $\eta_0 \sim \chi_1^2$, $F(\cdot|\lambda)$ is the distribution function of $\sum_{i=1}^{l} \lambda_i \chi_{1,i}^2$, approximated as described previously, and $f(\cdot|\chi_1^2)$ is a χ_1^2 density.

Appendix D

TABLES OF VARIANTS IN SIGNIFICANT KEATS-O GENES

rsID	Single Test P-Value	MAF
rs33985472	3.01e-01	3.91e-05
rs35004220	1.76e-11	2.34e-04
SoL-rs34598529	2.71e-02	1.17e-04
SoL-rs35724775	2.86e-04	3.91e-05

Table D.1: Results for variants that map to the *HBB* gene on chromosome 11. The KEATS-O p-value for this gene was 5.52e-12.

rsID	Single Test P-Value	MAF
kgp1108431	9.28e-02	4.34e-02
kgp1170149	9.13e-02	3.91e-05
kgp16208856	3.20e-01	1.25e-02
kgp16269203	2.43e-02	1.66e-02
kgp16303049	5.43e-15	1.20e-02
kgp16304904	1.60e-04	9.32e-03
kgp16305219	5.22e-01	8.42e-03
kgp16329531	2.67e-01	4.88e-03
kgp16343602	8.83e-01	1.35e-02
kgp16367388	3.54e-09	5.00e-03
kgp16372428	5.34e-01	5.59e-03
kgp16376923	9.76e-01	3.95e-03
kgp16405110	6.41e-01	8.83e-03
kgp16414310	1.77e-09	2.38e-02
kgp1642633	2.03e-01	3.39e-02
kgp16434305	4.39e-01	4.77e-03
kgp16435416	3.09e-08	3.71e-03
kgp16451886	2.19e-08	3.68e-03
kgp16472449	7.56e-01	3.42e-02
kgp16481087	6.17e-01	7.15e-03
kgp16491497	1.41e-01	1.10e-03
kgp16494133	2.90e-09	3.59e-03
kgp16497982	1.54e-01	1.05e-02
kgp16514907	5.09e-01	1.02e-02
kgp2903335	8.74e-09	1.87e-03
kgp3574526	7.31e-01	1.86e-02
kgp3786628	8.42e-01	3.07e-02
kgp4481603	5.62e-01	4.01e-02
kgp4512082	3.03e-01	1.17e-02
kgp5852815	1.00e-01	1.78e-02
kgp7587593	2.54e-01	4.53e-02
kgp883492	2.19e-01	3.87e-02
kgp9062773	8.74e-01	6.09e-03
kgp9725901	7.86e-01	3.01e-03
rs13339636	1.27e-12	1.66e-02
rs2858007	4.88e-01	4.42e-02
rs739996	7.54e-01	3.99e-02
rs8045291	2.56e-01	3.99e-02
rs9939865	2.63e-01	3.86e-02

Table D.2: Results for variants that map to the *LUC7L* gene on chromosome 16. The KEATS-O p-value for this gene was 1.05e-16.

rsID	Single Test P-Value	MAF
kgp16303049	5.43e-15	1.20e-02
kgp16329531	2.67e-01	4.88e-03
kgp16372428	5.34e-01	5.59e-03
kgp16376923	9.76e-01	3.95e-03
kgp16451886	2.19e-08	3.68e-03
kgp16481087	6.17e-01	7.15e-03
kgp16514907	5.09e-01	1.02e-02
kgp4481603	5.62e-01	4.01e-02
kgp4512082	3.03e-01	1.17e-02
kgp9725901	7.86e-01	3.01e-03
rs13339636	1.27e-12	1.66e-02
rs2858007	4.88e-01	4.42e-02
rs739996	7.54e-01	3.99e-02

Table D.3: Results for variants that map to the *ITFG3* gene on chromosome 16. The KEATS-O p-value for this gene was 2.18e-12.

rsID	Single Test P-Value	MAF
kgp11364344	1.28e-01	3.70e-02
kgp11889593	8.77e-01	2.79e-02
kgp1433789	4.53e-01	2.07e-03
kgp16209026	1.77e-02	7.55e-03
kgp16215607	4.27e-02	7.51e-03
kgp16215826	6.76e-02	5.09e-03
kgp16222342	1.79e-04	1.47e-02
kgp16232717	2.07e-02	1.65e-02
kgp16241356	3.13e-01	8.55e-03
kgp16267098	5.92e-01	3.44e-03
kgp16286897	4.91e-03	6.41e-03
kgp16293583	1.45e-02	1.01e-02
kgp16316185	1.01e-09	1.41e-02
kgp16349247	1.11e-01	4.38e-03
kgp16374601	3.78e-01	7.12e-03
kgp16380534	1.87e-01	5.39e-03
kgp16382899	9.79e-01	5.43e-03
kgp16422090	3.13e-01	1.38e-02
kgp16423378	3.50e-01	3.17e-02
kgp16428383	7.16e-02	2.25e-02
kgp16436844	8.52e-01	2.19e-03
kgp16437583	2.08e-01	6.25e-04
kgp16440142	4.53e-03	1.67e-02
kgp16451493	1.37e-01	4.36e-03
kgp16459248	1.85e-01	4.18e-03
kgp16467918	2.59e-02	1.70e-02
kgp16494327	8.98e-01	1.65e-02
kgp16495768	3.20e-01	3.93e-05
kgp16500307	5.82e-01	8.79e-03
kgp16515785	9.77e-01	1.96e-02
kgp16525415	5.45e-01	3.64e-03
kgp16526984	1.45e-01	7.85e-05
kgp16537801	8.13e-01	2.66e-03
kgp1811017	1.10e-01	2.14e-02
kgp22808759	9.54e-01	2.93e-03
kgp25574623	9.43e-01	2.81e-02
kgp3322005	8.18e-02	3.30e-02
kgp6018437	2.52e-01	1.87e-02
kgp8041981	6.89e-02	3.97e-02
kgp8779444	8.38e-02	2.10e-02
rs11865143	9.48e-01	4.20e-02
rs13335388	4.88e-06	2.35e-02
rs7195676	1.52e-08	3.12e-02

Table D.4: Results for variants that map to the *RAB11FIP3* gene on chromosome 16. The KEATS-O p-value for this gene was 1.58e-08.

rsID	Single Test P-Value	MAF
kgp30620757	3.90e-01	1.03e-02
kgp30626516	1.82e-18	2.01e-02
rs2515910	3.13e-01	2.46e-04
rs5986992	5.79e-01	5.41e-03
SoL-kgp22825971	7.97e-01	1.97e-04

Table D.5: Results for variants that map to the *G6PD* gene on the X chromosome. The KEATS-O p-value for this gene was 1.96e-18.

rsID	Single Test P-Value	MAF
kgp22753691	3.54e-01	4.89e-02
kgp22775930	6.07e-01	3.93e-04
kgp22791055	1.13e-01	1.97e-04
kgp22792732	3.92e-01	9.83e-05
kgp22801393	9.66e-01	4.92e-05
kgp22819085	1.98e-01	1.47e-04
kgp22830562	1.02e-01	2.46e-04
kgp35571516	9.13e-01	3.11e-02
rs12014480	3.40e-10	4.17e-02
rs17051888	6.56e-01	3.15e-02
rs1800289	7.17e-02	7.08e-03
rs28412378	1.62e-09	4.10e-02
rs36101366	2.94e-01	2.02e-03
rs5987042	1.31e-01	7.72e-03
rs5987056	1.08e-01	6.39e-03

Table D.6: Results for varaints that map to the *F8* gene on the X chromosome. The KEATS-O p-value for this gene was 1.32e-07.