Statistical Methods in Admixture Mapping: Mixed Model Based Testing and Genome-wide Significance Thresholds

Lisa A Brown

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Timothy Thornton, Chair
Sharon Browning, Chair
Bruce Weir

Program Authorized to Offer Degree:
Public Health: Biostatistics

University of Washington

Abstract

Statistical Methods in Admixture Mapping: Mixed Model Based Testing and Genome-wide Significance Thresholds

Lisa A Brown

Co-Chairs of the Supervisory Committee:
Sharon Browning
Biostatistics

Timothy Thornton
Biostatistics

Genetic admixture occurs when two or more previously isolated populations combine to form an admixed population. The study of admixed populations can provide valuable insights into the complex relationship between environmental exposures, genetic background and complex traits. Gene mapping by linkage admixture disequilibrium, or admixture mapping, is a powerful approach for the identification of genetic loci influencing complex traits in ancestrally diverse populations. Admixture mapping leverages genomic heterogeneity among sampled individuals for improved gene discovery, where genetic loci with unusual deviations in local ancestry and that are significantly associated with a trait are identified. Admixture mapping can serve both as a primary method for discovery of novel genetic variants and as a complement to association mapping. In this dissertation, we thoroughly investigate the performance of existing statistical methods used for admixture mapping and we develop new methods that improve upon existing approaches. We also characterize the correlation structure of genetic loci in admixed populations and develop new genome-wide significance

thresholds for admixture mapping under a range of models that should be useful for the future studies. Using real genotyping data in a large sample of African Americans, we find evidence of assortative mating, and in simulation studies with simulated phenotypes, we demonstrate that ancestry-related assortative can induce genome-wide inflation of admixture mapping test statistics and false positive associations. We also show how to appropriately adjust for this inflation and protect against spurious admixture associations. Finally, new linear and logistic mixed model methodology is developed for admixture mapping of quantitative and binary traits, respectively, in the presence of relatedness and population structure. We evaluate the performance of these methods through extensive simulation studies. The methods are applied to large-scale genetic studies of African American and Hispanic/Latino populations for genome-wide admixture mapping analyses where novel candidate loci for a variety of biomedical traits are identified.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Timothy Thornton and Dr. Sharon Browning for all their encouragement and direction during the dissertation process. Dr. Rakovski, thank you for mentoring me early in my years at Chapman; I wouldn't be here without your help and advice. I would also like to acknowledge the other professors working in statistical genetics at University of Washington who have contributed their ideas and expertise to my projects.

DEDICATION

In dedication to my parents, Karin and Philip Brown,
for all their love and support,
and for always believing in me.

Chapter 1

INTRODUCTION

The vast majority of genetic studies have been carried out in subjects from relatively homogenous European populations [1]. More recently, in an effort to capture additional genetic diversity that may be absent or present only at low frequencies in Europeans, a growing number of studies have been conducted in diverse populations [2-8]. Including individuals from diverse populations in genetic studies for diseases can help to ensure that clinical applications will benefit the greatest number of people, regardless of racial background. Genetic studies in non-Europeans often involve admixed populations, such as African Americans and Latinos, with recent ancestry derived from two or more ancestral populations from different continents [9-17]. Population admixture results in the combining of genomes from previously isolated populations that may have differing allele frequencies across the genome as a result of selection, mutation and genetic drift [18-20]. Admixed populations exhibit genetic variability at both the global and local level, with total proportional ancestry as well as inherited locus-specific ancestry varying from individual to individual [21-25].

Genetic studies in recently admixed populations can provide valuable insight into novel risk factors contributing to disease as they yield information on varying degrees of exposure to both genetic and environmental factors [26-30]. For example, a genetic variant that is fixed at opposite alleles in two populations cannot be identified as a risk variant by examining subjects from either population in isolation because the association with the variant is completely confounded by ancestry. An admixed population has the genetic variability required to separate genetic and environmental ancestry effects. Gene mapping by linkage admixture disequilibrium, or admixture mapping, is a powerful approach for the identification of genetic loci influencing complex traits in ancestrally diverse populations. Admixture mapping leverages genomic heterogeneity among sampled individuals for improved gene discovery, where genetic loci with unusual deviations in local ancestry and that are significantly associated with a trait are identified [31].

This dissertation develops methodology for admixture mapping in the presence of complex population structure and arbitrary relatedness patterns. The specific aims for this dissertation research are as follows:

- Compare the limitations of existing theoretical and simulated approaches to genome-wide testing significance thresholds for admixture mapping, and provide insight into approaches that can be used in realistic data settings for admixture mapping.
- Investigate the impact of long-range admixture linkage disequilibrium (LD) on the behavior of test statistics in admixture mapping and provide a method for correcting any potential inflation due to structure and/or non-random mating.
- Develop and evaluate linear and logistic mixed model frameworks for admixture mapping that allow for cryptic relatedness and population structure in admixed populations with ancestry derived from two or more ancestral populations.

1.1 Genome-wide Significance Thresholds in Admixture Mapping

In addition to background LD inherited from a population, admixed populations exhibit another form of LD, known admixture-LD or ancestry-LD, caused by the mosaic nature of admixed genomes [32]. Admixture-LD is the correlation of local ancestry values within a genomic region inherited from a non-admixed ancestor that has not been broken up by recombination [33]. In each successive generation of random mating, recombination will break up blocks of local ancestry. The admixture-LD within present day admixed populations, such as Latinos and African Americans, will be high for loci with a close genetic distance and lower for loci further apart. The critical value for genome-wide significance level of admixture mapping is substantially lower than that for a genotype test because the admixture-LD greatly reduces the number of independent tests in a genome-wide admixture mapping scan. Genome-wide association studies (GWAS) use a stringent p-value threshold, often 5×10^{-8}, which is derived from the consideration of a Bonferroni correction for one million SNP tests [34]. We consider several options for the genome-wide significance threshold including those based on real and simulated data, and discuss the justifications and advantages of each method.

Siegmund and Yakir [35] developed a theoretical framework for modeling the distribution of local ancestry and their corresponding admixture mapping test statistics within admixed populations derived from two ancestral populations. Assuming random mating, they show that the joint distribution of a pair of admixture mapping test statistics follows an Ornstein-Uhlenbeck process that depends on the number of generations since the admixture event and the recombination fraction between the two loci. In particular, this distribution does not depend on the overall admixture fractions or proportions of ancestry in the admixed sample population and is therefore generally applicable to all admixed populations that share the generation same time. This implies that we could use one genome-wide significance threshold for all African American populations if we believed they shared a common admixture event. This threshold approach is appealing because it lends itself to comparability of results across studies. We explore applying this threshold to admixture mapping settings when two or three ancestral populations are present. Sha et al. [36] proposed a similar model to Siegmund and Yakir which also modeled the underlying distribution of local ancestry but for a set of nearly uncorrelated sparse markers. However, we do not expect the correlation of the local ancestry (admixture-LD) to follow the same distribution as the correlation of alleles (allelic-LD). Therefore, significance thresholds based on the number of independent loci will likely be conservative for admixture mapping.

Some local ancestry inference programs, such as RFMix [54], call local ancestry within windows. For example, RFMix calls local ancestry within 0.2cM windows. As a result, adjacent local ancestry calls will be identical within short segments. In such instances, one option is to use a Bonferroni corrected significance threshold on the number of unique ancestry blocks in a specific sample of local ancestry calls, as the number of unique ancestry blocks will be less than the total number of genetic markers where local ancestry was inferred from a SNP chip panel. This approach is simple in that it does not make any modeling assumptions on the distribution of the underlying local ancestry in the sample. An approach similar to this uses the average number of switches in local ancestry per individual across the genome and accounts for correlation of markers within a chromosome using an autocorrelation (AR-1) model [37]. Other studies have used permutation tests [33], which are straightforward in data sets with unrelated

subjects, but become more complicated when relatives are present if one wishes to preserve the correlation structure in the sample.

1.2 Correcting for Confounding in Admixture Mapping Studies

Existing gene mapping methodologies often make an assumption of independence of loci across chromosomes. While this assumption is appropriate for randomly mating homogenous populations [38], in reality human populations do not mate at random, and previous studies have provided strong evidence of ancestry-related assortative mating in both intracontinental populations, such as European populations, and admixed populations with recent ancestry derived from multiple continents [39-41]. Previous work has shown that an admixed population that has been randomly mating for 7-10 generations will exhibit approximate independence of local ancestry values across chromosomes, with a degree of admixture LD within chromosomes depending on the number of generations [38].

We show that assortative-mating for ancestry can cause admixture-LD in local ancestry values far apart on a large chromosome and across chromosomes in recently admixed populations, and we assess the implications of this type of long-range and across chromosome LD in admixture mapping studies. In an analysis of real genotyping data from 8,421 African American (AA) women in the Women's Health Initiative SNP Health Association Resource (WHI SHARe) study, we find that African Americans have across chromosome admixture LD that can confound admixture mapping studies, where there can be false-positive associations on chromosomes that have no genetic effects. We also demonstrate that an admixture mapping analysis that conditions on local ancestry at the most significant genomic regions can provide protection against false-positives and allows for the identification of secondary admixture mapping signals that are not due to long range correlation. These results have implications, not only for admixture mapping studies, but also for other genetic applications that utilize local ancestry in studies of diverse populations.

1.3 Linear Mixed Models for Admixture Mapping in Related Samples

Current implementations of admixture mapping approaches perform association testing for one ancestral population at time, where the count of local ancestry from a reference ancestral group is included as a predictor in a regression framework and compared to a non-reference ancestry group [33-38]. The existing regression methodology for admixture mapping is valid only for unrelated subjects. Many genetic studies, however, include individuals with some degree of relatedness [48].

Mixed linear model (MLM) methods have become a popular method of choice for the analysis of genome-wide association studies, as they have been demonstrated to protect against spurious associations in structured samples by directly accounting for sources of dependence including cryptic relatedness and population stratification [49-51]. Here, we build on existing MLM methodology for GWAS and propose a mixed linear model-based approach for admixture mapping, AdmMix-LM (Admixture Mapping with Mixed Linear Models), in related samples. AdmMix-LM is implemented using local ancestry estimates inferred from genome-wide data and an empirical relatedness matrix, where sample structure is accounted for using both fixed and random effects. An important feature of AdmMix-LM is that the method can jointly test multiple ancestries for association. The performance of this methodology is assessed through simulation studies. We evaluate the type I error rate and power of our method across a range of single nucleotide polymorphism (SNP) effect sizes, additive genetic variance parameters and allele frequency differences across ancestral populations at the causal locus. We demonstrate that our method adequately controls for both continental and sub-continental ancestry admixture with the use of SNP array data and has appropriate type I error rates when applied to samples with known and/or cryptic relatedness.

As a real data application, we focus on urine albumin to creatinine ratio (uACR) in HSHC/SOL. Increased urine albumin excretion (albuminuria), a marker of kidney damage, is highly prevalent in Hispanic/Latinos. While the impact of genetic background on albuminuria risk remains elusive, previous studies have found an association between albuminuria and Amerindian ancestry in Hispanic/Latino populations. Our method, AdmMix-LM, identified three novel genome-wide significant signals at chromosomes 2, 11, and 16. The admixture mapping signal identified on chromosome 2, spanning q11.2-

14.1, has not been previously reported for albuminuria and is driven by Amerindian-ancestry ($p<5.7x\ 10^{-5}$). Within this locus, two common variants located at the pro-apoptotic *BCL2L11* gene were associated with albuminuria: rs116907128 (minor allele frequency [MAF] = 0.14, $p= 1.5x\ 10^{-7}$) and rs586283 (MAF=0.35, $p= 4.2x10^{-7}$). In a secondary analysis, rs116907128 was demonstrated to account for a majority of the admixture mapping signal observed in our primary analysis of albuminuria. The rs116907128 variant is common among Pima Indians (MAF=0.45) but is monomorphic in the 1000 Genomes European and African populations. We conducted a replication association analysis of rs116907128 in a sample of American Indians predominantly of Pima Indian heritage where rs116907128 was found to significantly associate with urinary albumin creatinine ratio (p=0.03). Our findings provide evidence for the presence of Amerindian-specific variants influencing the variation of albuminuria in Hispanic/Latinos.

1.4 Logistic Mixed Models for Admixture Mapping

Binary outcomes are common in genetic studies through case-control study designs. Although widely used linear mixed model methods assume that a quantitative trait is normally distributed, they are often utilized for the analysis of binary traits where a dichotomous outcome variable as treated as if it were continuous. The assumption of constant residual variance with linear mixed models for association mapping, however, can be violated when analyzing a binary outcome, and can result in inflated type I error rates in the presence of population structure. Recently, Chen et al. [52] developed a computationally efficient logistic mixed model, GMMAT, which correctly accounts for variance components of a binary trait. The problem of heteroskedastic residuals is not unique to association mapping and needs to be addressed in the admixture mapping framework as well. We extend GMMAT to allow for logistic mixed model based admixture mapping of binary traits. Our method has better calibrated test statistics under the null compared to a linear mixed model admixture mapping method when applied to analyze real genetic data with a binary outcome.

Chapter 2

BACKGROUND

In this chapter, we define some basic genetic concepts including admixture, local ancestry inference, admixture mapping, and provide an overview of sources of structure that are often encountered in real data from genetic studies. We present a commonly used model for admixture mapping, and briefly discuss the properties of the model, as it is essential for much of the work developed in the following chapters.

2.1 Admixture

Population admixture results in the combining of genomes from previously isolated populations that may have differing allele frequencies across the genome as a result of selection, mutation and genetic drift [42-44]. Admixed populations exist across the world, and the two largest minority populations in the United States, African Americans and Hispanic Americans, both have admixed ancestry. Admixed populations exhibit genetic variability at both the global and local level, with total proportional ancestry as well as inherited locus-specific ancestry varying from individual to individual [45-47]. The majority of the work presented here was motivated by problems seen within real admixed study populations.

2.2 Local Ancestry Inference

Recent computational and genomic advances allow for inference of an individual's genetic ancestry by marker position, using genome-wide SNP data [52]. This technique, known as local ancestry inference (LAI), estimates for a given subject, the number of alleles (0, 1, or 2) descended from each ancestral population of interest at each marker position. LAI relies on ancestral population reference panels to find the 'best match' population of origin for segments of admixed haplotypes. We note that Hispanic/Latino populations are descended from European, African and Native American populations and African Americans are descended from European and African populations.

Initial LAI was done using Ancestry Informative Marker (AIM) panels. Today, the availability of GWAS data allows for LAI at all marker positions genome-wide. For

all analyses presented here that were performed on real genetic data sets, we implemented a conditional random field-based approach, RFMix, to infer local ancestry at a set of SNPs genome-wide in common to the admixed sample panel and reference panel data sets, using selected populations from 1000 Genomes [55], Human Genome Diversity Project (HGDP) [56], and Hapmap3 [2] as a reference panels for Native American, European and West African populations, where applicable. RFMix requires phased data with no missing genotype values. Beagle [57] was employed for phasing and imputation of sporadic missing genotypes in the admixed sample and reference panel data sets.

2.3 Admixture Mapping

GWAS have been widely employed for the identification of associations between single nucleotide polymorphisms (SNPs) and a wide variety of complex traits of interest. GWAS rely on background linkage disequilibrium (LD) between a genotyped SNP and a causal variant to identify associations with traits [58]. Admixture mapping uses admixture-LD, which often extends further than the background LD inherited from an ancestral population, to identify genetic variants that affect trait levels and have differing allele frequencies across the ancestral populations that contributed to an admixed population. In admixture mapping studies, counts of locus-specific ancestry at genomic loci are used as predictors and tested for association with a trait. An advantage of admixture mapping studies is that local ancestry can capture associations with both common and rare variation [53,59], whereas the SNP genotyping arrays that are widely used in GWAS do not capture rare variation well as they have largely been designed for assaying common variants.

We focus our attention on a linear model for a quantitative phenotype, although regression models allow for other types phenotypic distributions such as binary, count and ordinal data. Suppose we collect N individuals who are admixed from two ancestral populations and record their value for a quantitative trait, $\boldsymbol{Y} = (y_{1,} y_{2,}, \dots, y_{N,})$. We will treat one population as the reference population and one population as the alternative population. Assuming the subjects are unrelated and typed at m SNP markers, we can model the relationship between the trait $\boldsymbol{Y}$ and local ancestry at position $j \in \{1, \dots, m\}$ as

$$\boldsymbol{Y} = \mathbf{1}\alpha + \boldsymbol{X_j^{REF}}\beta_j^{REF} + \boldsymbol{W\gamma} + \boldsymbol{\delta}, \quad (1)$$

where $\mathbf{1}$ is a vector of 1's of length N, $\boldsymbol{X_j^{REF}}$ is a vector of length N denoting the number of ancestral alleles (0,1 or 2) at locus j descended from the reference ancestral population, with corresponding effect size β_j^{REF}. $\boldsymbol{W}$ is a matrix that represents covariate adjustment variables such as principal components (PCs) with corresponding fixed effect vector $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ is a length N vector of independent and identically distributed random effects for residual error that are normally distributed with mean 0 and variance σ^2.

2.4 Sources of Structure

Admixed populations exhibit population structure due to variable ancestry from multiple ancestral populations, and this genetic diversity must be taken into account in studies of complex traits in population with admixed ancestry to protect against confounding. In the regression model described in equation (1) above, fixed effect adjustments often include PCs or proportions of global ancestry. Including ancestry representative PCs as covariates removes confounding due to ancestry [59,61]. This is especially important for admixed populations who have a multitude of ethnic, geographic and cultural backgrounds that contribute to any given phenotype. Furthermore, known and/or cryptic relatedness are common features of genetic studies, including those of admixed individuals [12,39-42].

PCs and the empirical relatedness matrix $\boldsymbol{\Phi}$ used throughout this dissertation are calculated from genome-wide genotype data and are obtained in a recursive manner. Relatedness is first estimated using KING-robust [13], which is robust to discrete population structure but not to admixture or departures from HWE within sub-populations. Using PCAir [62], the sample is then partitioned into a mutually unrelated set and the remaining who are relatives of the unrelated set. PCAir then performs standard principal components analysis (PCA) on the set of unrelated individuals and projects PC coordinate values for the related individuals. PCRelate [62] calculates pairwise relatedness conditional on the obtained PCs from PCAir. Both PCAir and PCRelate are run again, starting with the newly obtained relatedness matrix. The final kinship coefficients and sample eigenvectors from PCRelate and PCAiR, respectively,

are used to adjust for relatedness and population stratification in the models considered here.

Chapter 3

SIGNIFICANCE THRESHOLDS FOR ADMIXTURE MAPPING

3.1 Introduction

Ideally, we would like to develop a general framework for genome-wide significance thresholds to be used for all data sets in admixture mapping studies that takes into account the number of generations since admixture and the number of ancestral populations. Tang et al. reported 7×10^{-6} as an appropriate p-value significance level for admixture mapping in WHI-SHARe African Americans [63,64]. We show that a single admixture mapping threshold may not be appropriate for all samples of admixed populations, as many populations have unique histories resulting in different patterns of population structure. We evaluate genome-wide significance thresholds for admixture mapping based on the Ornstein-Uhlenbeck process developed by Siegmund and Yakir, and provide new insights into advantages and limitations of different significance threshold approaches using both simulated and real data.

3.2 Methods

3.2.1 Theoretical Genetic Data: Siegmund-Yakir Framework

Siegmund and Yakir developed a rigorous statistical theory for the determination of a genome-wide p-value threshold for admixture mapping when two ancestral populations are present, assuming independence of local ancestry values across chromosomes and using a test statistic similar in construction to a Wald statistic [65]. They showed that correlation of admixture mapping test statistics within a chromosome follows an Ornstein-Uhlenbeck process that depends on the number of generations since the admixture event but does not depend on the admixture proportions of the admixed population. Under this model, $Corr(X_s, X_t) = e^{-|s-t|g}$, where X_i is the test statistic at genetic map position $i \in \{s, t\}$, and g is the number of generations since admixture. This correlation can be rewritten in terms of the recombination fraction distance, θ, between two loci as $(1-\theta)^g$.

To explore the effect of marker spacing and generation time on the theoretical significance threshold obtained under this framework, we simulated test statistics using the Ornstein-Uhlenbeck model across the genome at markers spaced every 0.01, 0.02,

0.05 or 0.007 cMs at generation times $g \in \{6,8,10,12,14\}$. For reference, we note that African Americans are estimated to be 6-10 generations since admixture and a GWAS panel would be close to a spacing of 0.007 since GWAS panels for admixture mapping must be an intersection between the sample panel data and reference panels used for local ancestry inference.

The simulated phenotype distribution is drawn independently from a N(0,1) for each subject. We run a genome-wide admixture scan using linear regression with fixed effect adjustments for global ancestry proportions, recording the minimum p-value. The lower 5th percentile of minimum p-values across 10,000 independent replicates is chosen as the genome-wide p-value threshold.

3.2.2 Simulated Genetic Data: Siegmund-Yakir Framework

We directly model admixture mapping test statistics based on simulating local ancestry over generations of random mating and run a linear regression-based admixture mapping scan on null phenotypes. We compare the behavior of admixture mapping test statistics when two or three ancestral populations are present, varying the admixture proportions of the simulated population, to the theoretical behavior of the statistics based on the Siegmund-Yakir model.

To create simulated local ancestry values for an admixed individual included in our analysis, we simulate their entire set of ancestors dating back to the generation of founder non-admixed individuals involved in the admixture event. Starting with the founder generation, we simulate crossovers, transmission and mating leading to offspring at each subsequent generation. For a randomly mating population, all current generation admixed subjects' founders' ancestry haplotypes are drawn from a Multinomial($\mathbf{p}$), where $\mathbf{p}$ is a vector of admixture fractions that sum to 1. In the case where there is only two ancestry populations, this reduces to drawing founder ancestry haplotypes from a Binomial(p), where p is the admixture fraction of the population. For each case, we vary the admixture fractions of the simulated population obtain genome-wide significance thresholds unique to each admixture fraction set level. As described above, we simulate phenotype distribution drawn from a N(0,1) for each subject and run a genome-wide admixture scan with adjustments for global ancestry proportions, recording the minimum

p-value. A threshold set at the lower 5th percentile of minimum p-values across 10,000 independent replicates will control the family-wise type I error rate at 5%.

In obtaining a significance threshold via simulating local ancestry, we also need to consider how sensitive the threshold is to different randomly generated data sets. For the three-population case, we explore how the genome-wide significance threshold changes across simulated data sets and across admixture fraction levels. For each significance threshold, we create bootstrap 95% confidence intervals calculated across 1,000 bootstrap samples of size 5,000.

3.2.4 Real Genetic Data

Simulations make simplifying assumptions that may not be realistic in practice. In this instance, the simulated local ancestry framework models the stochastic nature of admixture mapping test statistics within a chromosome based on the expected behavior of an Ornstein-Uhlenbeck process. The process is simulated independently for each chromosome and no correlation across chromosomes is considered. As will be demonstrated in Chapter 4, admixed populations exhibit population structure patterns that can lead to violations of independence assumptions across chromosome assumptions. To this end, we calculate genome-wide p-value thresholds for data sets known to contain population structure. We explore genome-wide p-value thresholds based on real genotype data and simulated traits in two and three ancestral population cases using an African American and Hispanic data sets, respectively.

We obtain a genome-wide significance threshold in the manner described in the previous section for simulated genetic data scenarios. We implemented admixture mapping with linear regression on a set of unrelated African Americans, adjusting for the first two PCs. The model used for analyzing the two Hispanic data sets needed to take into account the relatedness present in both data sets. The details and motivation for the linear and logistic mixed model used to analyze these data sets are described in Chapters 5 and 6. For the one data set, we simulated phenotypes from a $N(\mathbf{0}, \boldsymbol{\Phi}\sigma_a^2 + \boldsymbol{I}\sigma_e^2)$, where σ_a^2 and σ_e^2 represent additive genetic and environment variances, respectively, setting the parameters to their estimated quantity based on the phenotype data. For the second, we compared the Bonferroni correction on the number of unique local ancestry blocks

obtained from RFMix to one using direction simulation of phenotypes as described above. We also examined the extent to which the total heritability of the trait affected the p-value threshold in the second data set. To do this, we simulated phenotypes from a $N(\mathbf{0}, \boldsymbol{\Phi}\sigma_a^2 + \boldsymbol{I}\sigma_e^2)$, where σ_a^2 and σ_e^2 represent additive genetic and environment variances, respectively. We set $\sigma_e^2 = 1$, and varied σ_a^2 such that the total heritability of the simulated trait, $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, varied between 0 and 0.5.

We additionally record the genome-wide genomic control inflation factor, λ_{GC}, for the second data set's p-value threshold analyses. The genome-wide genomic control inflation factor, the ratio of the median observed test statistic to the median expected test statistic under the null hypothesis of no association, is often used in genetic studies to assess model fit. An inflation factor greater than 1 indicates inflation of the test statistics, which can be caused by model misspecification, confounding, or a skewed or heavy-tailed trait distribution. Inflation factors are commonly reported for GWAS studies but they are not often reported in admixture mapping studies [10-15] and the behavior of local ancestry test statistics with regard to genome-wide genomic control inflation factors has not been well studied.

3.2.5 WHI-SHARe African American Cohort

The WHI is a long-term health study of postmenopausal women in the U.S. A total of 161,808 postmenopausal women aged 50-79 years old were recruited, including 12,151 self-identified AAs. The WHI SHARe minority cohort includes 8,515 AA women who provided consent for DNA analysis. Genotype data are from the Affymetrix Genome-Wide Human SNP Array 6.0 that contains 906,000 single nucleotide polymorphisms (SNPs) and more than 946,000 probes for the detection of copy number variants. The genotype data were processed for quality control, including call rate, concordance rates for blinded and unblinded duplicates, and sex discrepancy, leaving 871,309 unflagged SNPs with a genotyping rate of 99.8%. After exclusions, there remain 8,421 AA women [66,67]. We identify a set of 551,025 SNPs common to the WHI-SHARe and HapMap data sets, with an overall genotyping rate of 99.9%.

3.2.6 Hispanic Community Healthy Study/Study of Latinos (HCHS/SOL)

The HCHS/SOL consists of 16,415 participants age 18-74 years who self-identified as Hispanic/Latino. Subjects were recruited from a random sample of households in defined communities in the Bronx, Chicago, Miami and San Diego from 2008 to 2011. Sampling was conducted so that there would be adequate representation from each of the backgrounds of interest: Cuban, Puerto Rican, Mexican, and Central/South American descent. Subjects were sampled within households and within census blocks. Genotyping was performed using the Illumina SOL HCHS Custom 15041502 B3 array for the 12,803 subjects who consented to have their DNA extracted to genetics studies. The array contains of a total of 2,536,661 SNPs, of which 2,427,090 are from a standard Illumina Omni2.5M array (HumanOmni2.5-8v1-1) and the remaining 109,571 are custom SNPs. The missing call rate for all subjects was less than 2% and the median call rate was 99% [68,69].

3.2.7 Alzheimer's Disease Sequencing Project (ADSP) Caribbean Hispanic Families

The ADSP is a large scale sequencing project to identify genetic variants contributing to Alzheimer's disease (AD) in multi-ethnic populations. The Family-Based study included 545 Caribbean Hispanic subjects from 68 families with available GWAS data. GWAS chip genotyping was performed on HumanOmniExpress12.v1.1.8, Human650Y.v2 and HumanOmni1-Quad.v1.0.H at Baylor, Broad, and Washington University. We pruned markers to a genotyping rate of 90%.

3.2.8 On the Use of True vs. Estimated Local Ancestry in Threshold Calculation

In practice, local ancestry is unknown and must be estimated and therefore multiple testing corrections should be based on estimated local ancestry, and not the true ancestry. It is possible that a small amount of error is accrued through estimation of local ancestry, for example, when a small bit of ancestry is missed-called by the algorithm as a neighboring ancestry instead. As a result, we expect the number of true unique ancestry blocks to be larger than the number estimated. To assess how large this difference is, we performed local ancestry inference on a set of simulated data and compared the number of ancestry switches in the true versus inferred local ancestry data sets.

3.3 Results

3.3.1 Theoretical Genetic Data: Siegmund-Yakir Framework

Table 1 displays the genome-wide p-value thresholds obtained via simulation of admixture mapping test statistics following an Ornstein-Uhlenbeck process. The significance thresholds become more stringent as generation time and marker density increase. Assuming 8 generations since admixture for African Americans and a marker spacing similar to a SNP panel that remained after local ancestry inference via reference panels, (which would be equivalent to ~533,430 markers genome-wide), we obtain significance thresholds between 4.06×10^{-6} and 5.75×10^{-6}.

Table 1: Significance Thresholds by Marker Density and Number of Generation

cM Distance Between Markers	**Generation Time**					
	6	**8**	**10**	**12**	**14**	**16**
0.007	6.03×10^{-6}	5.71×10^{-6}	3.24×10^{-6}	2.92×10^{-6}	2.44×10^{-6}	2.01×10^{-6}
0.01	6.44×10^{-6}	4.06×10^{-6}	3.38×10^{-6}	3.17×10^{-6}	2.63×10^{-6}	2.08×10^{-6}
0.02	7.24×10^{-6}	5.37×10^{-6}	3.91×10^{-6}	3.18×10^{-6}	2.65×10^{-6}	2.11×10^{-6}
0.05	7.39×10^{-6}	5.75×10^{-6}	4.73×10^{-6}	3.90×10^{-6}	3.24×10^{-6}	2.51×10^{-6}

3.3.2 Simulated Genetic Data: Siegmund-Yakir Framework

Figure 1 shows the correlation in test statistics for loci separated by a recombination fraction θ ranging from 0 to 0.5 for the two and three ancestral population case. The two-population case follows the Siegmund-Yakir theoretical line of $(1-\theta)^g$ but the three-population case does not. Results are shown for p =1/2 for the two population and p =(1/3,1/3,1/3) for the three population cases, respectively. Varying the admixture fraction level for either case did not change the level of correlation observed.

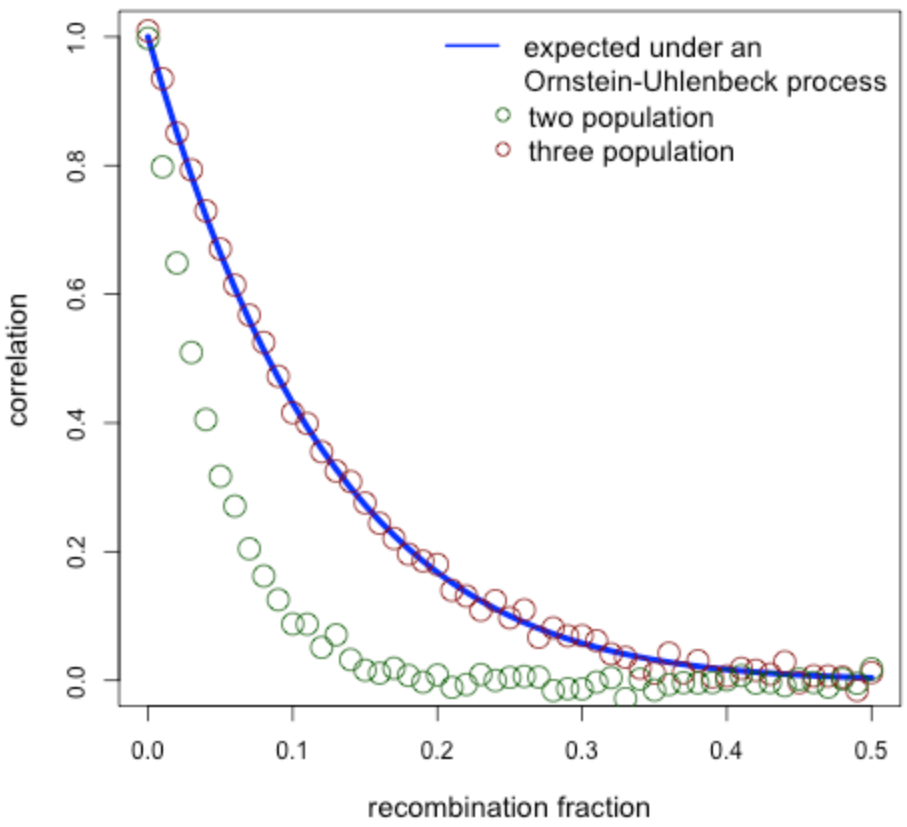


Figure 1: Correlation of Test Statistics

The average correlation across admixture mapping test statistics calculates within windows of 0.01 recombination fraction distance is shown with circles for simulated admixed populations with two ancestral populations (red) and three ancestral populations (green). The theoretical expected correlation assuming the test statistics follow an Ornstein-Uhlenbeck process, $(1-\theta)^g$, is shown in blue. The simulated population with two ancestral populations follows the expected line but the simulated population with three ancestral populations does not.

3.3.3 Simulated Genetic Data p-value Thresholds

Figure 2 shows the p-value thresholds for various population admixture fractions. Values range from 1.55×10^{-5} to 1.99×10^{-5}**,** with no clear trend towards more or less

significant thresholds at extreme of admixture fraction values. **Figure 3** displays the scatter plot for p-value thresholds comparing two data sets with identical admixture fractions. If results were concordant across data sets, we would expect to see points along the 1-to-1 line. There is a large amount of variability in p-value thresholds across data sets, with a correlation of 0.0045. Furthermore, the variability we see across admixture fractions is similar in magnitude to variability seen across data sets of the same admixture fraction, and thus there does not appear to be a difference of p-value thresholds for different admixture fractions. **Table 2** gives the confidence intervals for each data set and set of admixture fractions.

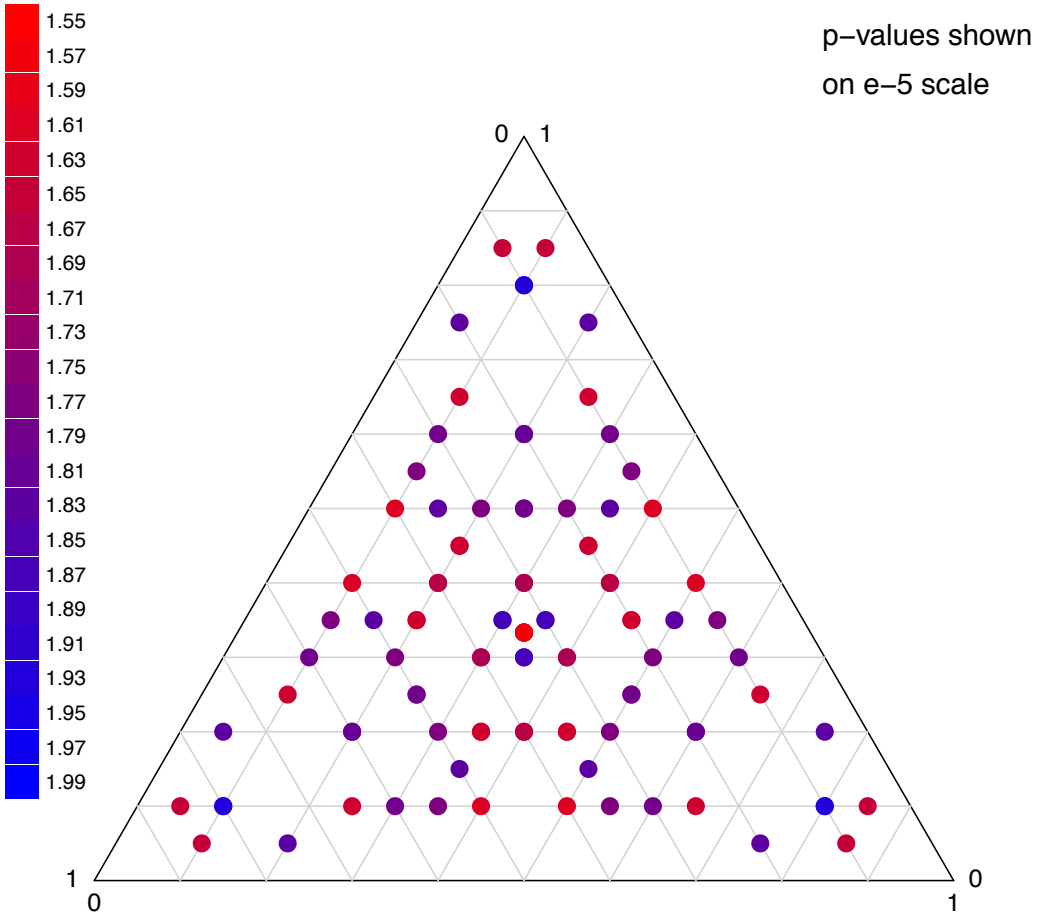


Figure 2: Genome-wide Significance Threshold p-values for Simulated Populations with Three Ancestral Populations

Admixture proportions within the convex hull of possible proportion sets are shown based on simulated local ancestry data sets with three ancestral populations, colored by gradients from red to blue denoting more or less stringent thresholds according to the bar of p-values $\times 10^{-5}$ to the left of the triangle. Each admixture proportion set (p) is represented three times because results do not depend on the order of the proportions.

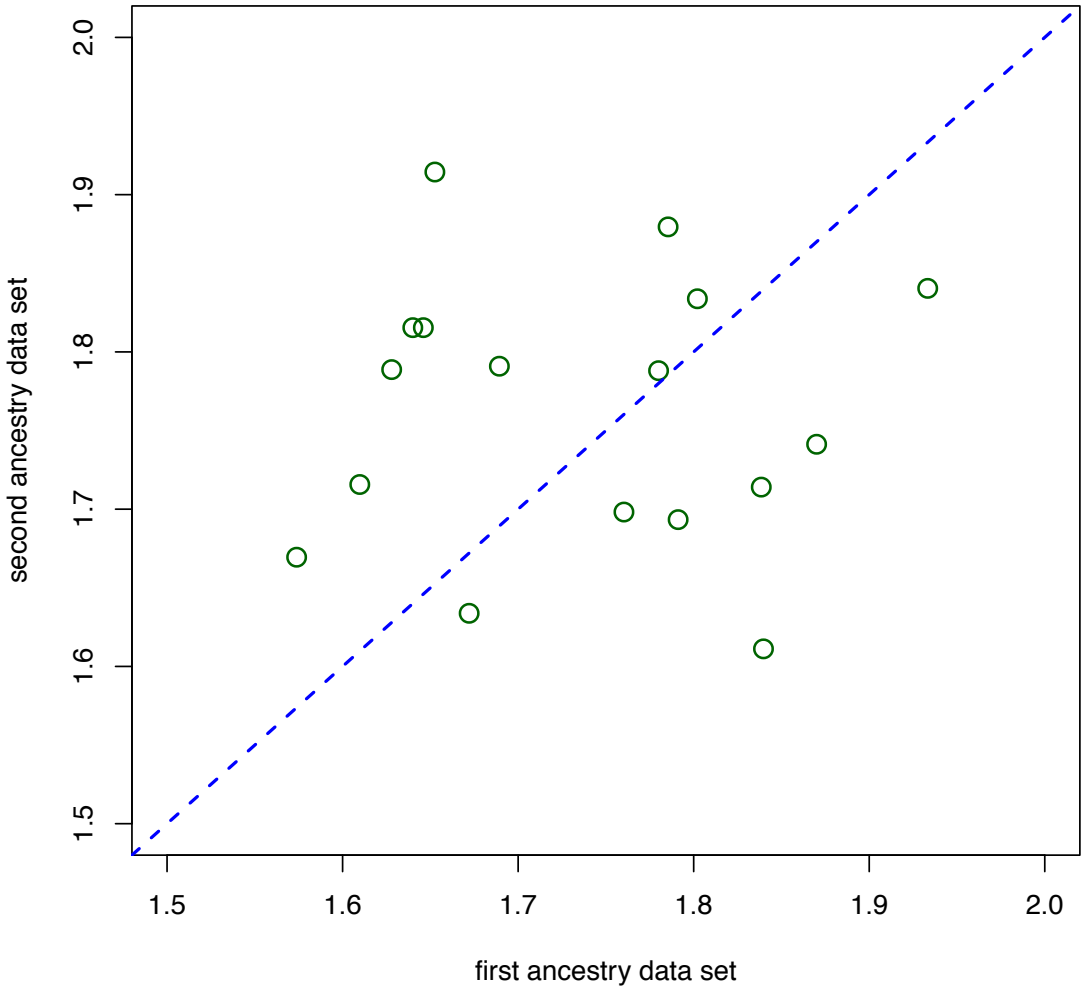


Figure 3: Genome-wide Significant p-value Thresholds

Each green circle represents genome-wide significance p-value thresholds for the first (x-axis) and second (y-axis) simulated data sets at the same admixture proportion combination. The 1-to-1 line dashed blue line is given for reference (i.e. assuming generated data sets gave concordant results at each chosen admixture proportion combination).

Table 2: Three Population p-value Thresholds for Genome-wide Significance with 95% Bootstrap Confidence Intervals for Two Sets of Simulated Data at Each Combination of Admixture Proportions.

Admixture Proportions			Data Set 1		Data Set 2	
			p-value	Confidence Interval	p-value	Confidence Interval
0.33	0.33	0.33	1.57	(1.43, 1.66)	1.67	(1.53, 1.80)
0.35	0.35	0.3	1.87	(1.76, 2.01)	1.74	(1.56, 1.90)
0.4	0.3	0.3	1.69	(1.60, 1.80)	1.59	(1.47, 1.75)
0.4	0.4	0.2	1.67	(1.56, 1.81)	1.35	(1.25, 1.48)
0.45	0.35	0.2	1.65	(1.51, 1.78)	1.82	(1.69, 1.94)
0.5	0.25	0.25	1.79	(1.64, 1.90)	1.88	(1.72, 2.02)
0.5	0.3	0.2	1.76	(1.61, 1.90)	1.60	(1.47, 1.75)
0.5	0.35	0.15	1.84	(1.71, 1.92)	1.61	(1.46, 1.74)
0.5	0.4	0.1	1.61	(1.49, 1.76)	1.70	(1.56, 1.83)
0.55	0.35	0.2	1.78	(1.64, 1.89)	1.79	(1.63, 1.91)
0.6	0.2	0.2	1.80	(1.66, 1.94)	1.83	(1.67, 1.99)
0.6	0.3	0.1	1.79	(1.68, 1.94)	1.69	(1.56, 1.81)
0.65	0.25	0.1	1.63	(1.53, 1.81)	1.79	(1.66, 1.90)
0.75	0.2	0.05	1.84	(1.69, 2.00)	1.71	(1.60, 1.82)
0.8	0.1	0.1	1.93	(1.80, 2.09)	1.84	(1.71, 1.99)
0.85	0.1	0.05	1.65	(1.54, 1.78)	1.91	(1.80, 2.08)

3.3.4 Real Genetic Data p-value Thresholds

The genome-wide p-value threshold based on simulating a null phenotypes in the WHI-SHARe AAs is 2.24×10^{-5}. This threshold is much less stringent than those obtained via Ornstein-Uhlenbeck simulation. The genome-wide p-value threshold based on simulating a null phenotype in the ADSP Caribbean Hispanics is 4.51×10^{-5}. In HCHS/SOL we had 236,456 genetic markers with local ancestry calls and tested 14,815 unique ancestry blocks. Applying a Bonferroni correction leads to a genome-wide p-value threshold of 3.6×10^{-6}. The p-value thresholds, corresponding bootstrap 95% confidence intervals and average inflation factors for various trait heritability levels are show in **Table 3**. Varying the trait heritability in simulations did not substantially change significance threshold, and values remained between 4.6 and 6.1 $\times10^{-5}$, with confidence intervals overlapping one another for all but the highest heritability considered. Interestingly, the average λ_{GC} decreased from 1.05 to 0.95 as heritability increased from 0 to 0.5.

Table 3: Significance Threshold Estimates and 95% Confidence Intervals and Mean Inflation Factor by Total Trait Heritability

Total Trait Heritability	**Significance Threshold**		**Mean Inflation Factor**
	Estimate	**95% CI**	
0.0	5.4×10^{-5}	(5.1, 5.8)	1.05
0.1	6.0×10^{-5}	(5.3, 6.0)	1.05
0.2	6.1×10^{-5}	(5.6, 6.8)	1.02
0.3	5.7×10^{-5}	(5.3,6.3)	1.01
0.4	5.5×10^{-5}	(5.1, 5.9)	0.98
0.5	4.6×10^{-5}	(4.3, 5.2)	0.95

3.3.5 On the Use of True vs. Estimated Local Ancestry in Threshold Calculation

Figure 4 shows a heat map of the count of local ancestry switches in inferred versus true ancestry on the same data. If local ancestry called with 100% accuracy, we would see all counts on the diagonal. There is a slight trend towards underestimation of switches, indicating that local ancestry inference is missing some small bits of ancestry.

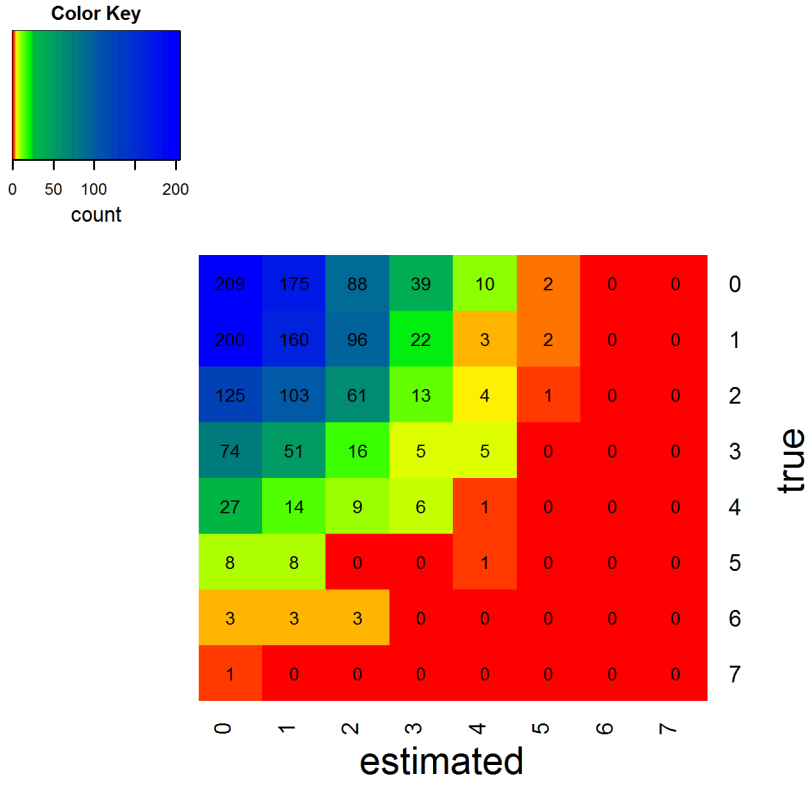


Figure 4: Heat Map of the Count of True and Estimated Number of Local Ancestry Switches.

The number of true and estimated/inferred number of ancestry switches for a 10mB region is recorded for each simulated admixed individuals across a 10Mb region. We display the total count of each combination of true vs. estimated number of local ancestry switches, colored from highest count (blue) to lowest count (red).

Table 4 shows the p-value thresholds obtained across all methods discussed in this chapter. Methods such as the Siegmund-Yakir framework, simulated genetic data and Bonferroni correction, which assume independence across chromosomes and/or start from true local ancestry, are more conservative than methods that capture the structure and correlation present in real genetic data using inferred local ancestry.

Table 4: Comparison of p-value Thresholds Across Methods

	Two Population	**Three Population**
Siegmund-Yakir Framework	5.7×10^{-6}	
Simulated Local Ancestry Data		$1.6-2.0\times10^{-5}$
Real Genetic Data		
WHI-AA	2.2×10^{-5}	
ADSP		4.5×10^{-5}
HCHS/SOL Bonferroni		3.6×10^{-6}
HCHS/SOL simulated phenotype		$4.6\text{-}6.1\times10^{-5}$

3.4 Discussion

Through simulations and analysis of multiple real African American and Latino data sets, we examine a variety of methods for accurate control of type I error rates in multi-way admixed populations. In our simulations of local ancestry data via the Siegmund-Yakir framework for admixture mapping test statistics and direct simulation of local ancestry values, we assume independence across chromosomes and this may not be realistic in genetic data of admixed populations due to population structure and/or non-random mating. This difference in assumptions may be reflected in the magnitude of genome-wide significance thresholds obtained where methods that assume independence across chromosomes yield a more stringent p-value threshold compared methods used on real genetic data. In the data set where we investigated both types of methods, this difference was an order of magnitude apart, with real data on the order of 10^{-5} and simulated data on the order of 10^{-6}. The exception to this is the Bonferroni corrected threshold on real data for the number of unique ancestry blocks but this may not capture

long-range correlation in local ancestry beyond the identical ancestry surrounding an ancestry block. Simulations indicated that p-value threshold in the three population case did not depend on the admixture proportions of the population which is comparable to the two population case. Furthermore, the significance threshold did not depend strongly on the total heritability of a trait.

Genome-wide significance thresholds that reflect the actual amount of correlation in a given data set, such as those obtained via simulation of null phenotypes combined with real genetic data, are preferable to those obtained via simulating local ancestry because they make less assumptions about the underlying structure of the ancestry and are tailored to the specific population being studied. It is important to note that the two Hispanic data sets we used to come up with p-value thresholds did not give identical results. The HCHS/SOL subjects contained a mix of both mainland and island Hispanic populations compared to the ADSP Hispanic family data set, which was a set of only Caribbean Hispanics. The structure present in each of the samples was unique and the p-value threshold based on the specific data sets for each should not be applicable to the other. This should be true in general. That is, p-value thresholds calculated for specific data sets should not be used for other data sets, even if they are from the same population, such as Hispanic or African American. Genome-wide significance thresholds based on real data, while computationally expensive, more reflect the true structure of local ancestry in the sample compared to those calculated on a simulated local ancestry data set that is supposed to mimic a real population. The evidence given in this chapter suggests that there is not one single p-value threshold that should be used for all data sets from a the same general admixed population. We recommend using a p-value threshold for admixture mapping that is calculated from the study population one wishes to perform the analysis as this will take into account any unknown sources of structure within the sample.

Chapter 4

CONFOUNDING IN ADMIXTURE MAPPING STUDIES

4.1 Introduction

This work is motivated by global and local ancestry patterns observed in the WHI-SHARe AA data set. The WHI-SHARe AAs have similar patterns of genome-wide population structure and ancestry admixture to previously reported population genetic studies of AA [70-73]. Also, the WHI-SHARe AAs exhibit an increased amount of correlation in local ancestry both within and across chromosomes, as compared to a theoretically randomly mating admixed population. We demonstrate through simulation studies that this pattern of long-range and across chromosome admixture LD in the WHI-SHARe AAs is consistent with ancestry-related assortative mating.

We conduct simulation studies to assess the impact of long-range and across chromosome LD on widely used admixture mapping test statistics from a linear regression framework. In simulation studies with real genotype data from WHI-SHARe AAs and simulated phenotype data, we find that across chromosome admixture LD can confound admixture mapping studies. Skelly et al. [73] previously described how this same phenomenon can occur in association mapping when SNPs are not linked but are in LD. We find that inflation factors for these analyses increase linearly with simulated effect size at a true casual locus, and that false positives can be induced on chromosomes that have no genetic effects. We also demonstrate that an admixture mapping analysis that conditions on local ancestry at the most significant genomic regions in a regression model can control inflation, provide protection against false-positives on other chromosomes, and allow for the identification of secondary admixture mapping signals that are not due to long range correlation.

4.2 Methods

4.2.1 Characterization of Local and Global Ancestry Patterns in Real and Simulated Data

We created simulated local ancestry data sets for comparison with the WHI AA data by simulating crossovers and transmission over generations starting with non-admixed founders assuming random or non-random mating patterns. For the simulated randomly mating population, each founders' haplotype is African with probability p, and

European otherwise, where p is the observed average global African ancestry in WHI AA. To simulate an assortatively mating population with a global ancestry distribution similar to WHI AA, the ancestry of each of individual i's founding ancestors' haplotype is African with probability p_i, where p_i is drawn from the empirical distribution of global ancestry values in WHI. This means that some individuals have a higher proportion of African founder haplotypes than others, which gives rise to population structure. For each simulation scenario, we simulate 8 generations of mating along two chromosomes of length 250 cM, assuming crossovers occur at a rate of 0.01/cM, and we simulate a total of 8,421 subjects to match the number of subjects in the WHI AA cohort.

For all three data sets, the WHI-SHARe AAs and the simulated random and assortative mating subjects, we record (1) the distribution of global ancestry (2) the correlation in local ancestry values for loci separated by a cM distance between 0 and 250 (3) distribution of correlation in local ancestry values across chromosomes.

4.2.2 Assessing Admixture Mapping Test Statistics using Real Genotype Data and Simulated Phenotype Data

To assess the behavior of test statistics in admixture mapping, we used real genotypes from WHI-SHARe AAs and simulated phenotypes created by adding local ancestry effects to a randomly generated standard normal base phenotype for each subject. We selected 500 marker loci approximately equally spaced across the genome. For each selected locus, we drew phenotypes independently for each subject from a $N(0, X\beta)$, where X is the number of copies of European alleles the locus and the range of $\beta \in [0,1.5]$. We employed the software PLINK [74] to perform admixture mapping with linear regression on the set of unrelated subjects. Our primary analyses included fixed effect adjustment for the first four principal components. Our secondary analyses included local ancestry at the causal locus as an additional covariate. For both analyses we record the genome-wide genomic control inflation factor, λ_{GC}, at each simulated replicate.

4.3 Results

4.3.1 Characterization of Local and Global Ancestry Patterns in Real and Simulated Data Sets

Figure 5 illustrates the distribution of proportion of African ancestry observed within the WHI African Americans and simulated populations assuming assortative or random mating. The WHI AA subject distribution is right-skewed with a heavy left tail with some individuals having little to no African ancestry. The global ancestry distribution of the randomly mating simulated population is centered on the average admixture fraction of the WHI AAs, 0.77, with less overall variability compared to the WHI AAs. Our simulated population with assortative mating shows a trend similar to the WHI AAs with a wider dispersion of global ancestry proportions. **Figure 6A** and **Figure 6B** show the correlation in local ancestry values calculated across subjects for pairwise position separated by a given cM distance within chromosomes (A) and all pairwise positions across chromosomes (B) in the WHI AAs and simulated populations assuming assortative or random mating. The within-chromosome correlation of local ancestry in the simulated randomly mating population drops off to zero after 100cM, while the WHI AAs and simulated population with assortative mating show correlations of local ancestry that decay more slowly with distance and level off near 0.2 at approximately 100cM. As expected, the across chromosome correlation of local ancestry in the simulated randomly mating population is centered around zero, while the WHI AAs and simulated assortative mating population show an average correlation of 0.234 and 0.238, respectively, with little variability across positions.

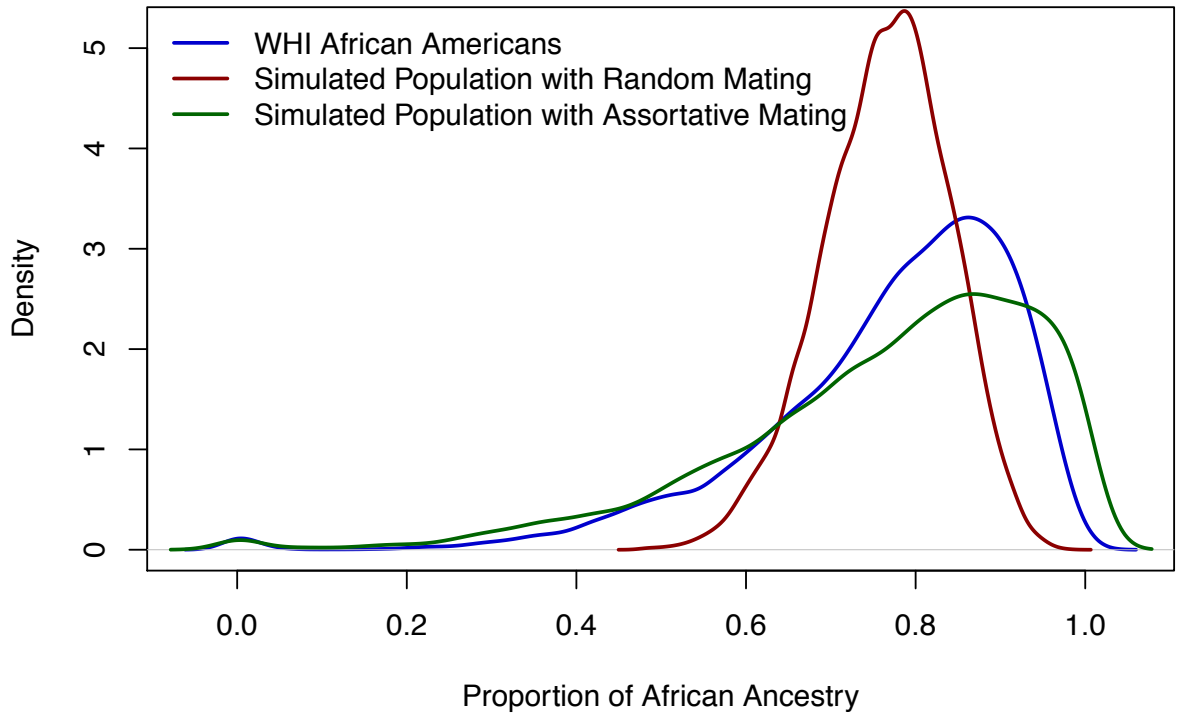


Figure 5: Distribution of Proportion of African Ancestry

Plot showing the smoothed density using an estimated kernel density of proportion of African ancestry observed within the WHI African Americans and simulated populations assuming assortative or random mating. The color represents population sample source.

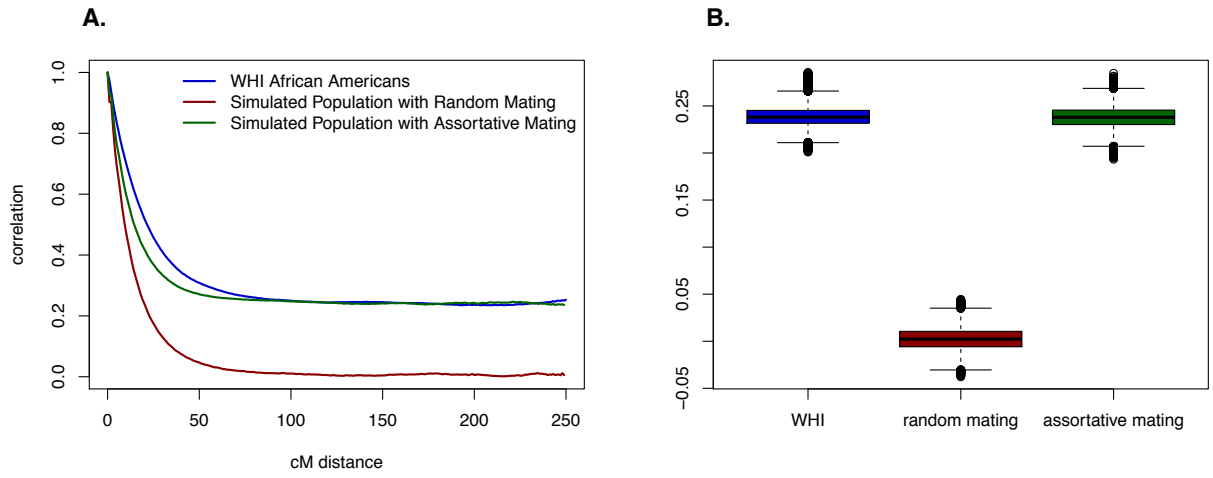


Figure 6: Local Ancestry Correlations Within and Across Chromosomes

Plots showing within (A) and across (B) chromosome correlation of local ancestry values within the WHI African Americans and simulated populations assuming assortative or random mating. (A) Correlation is represented on the y-axis and the cM distance between the locations along the x-axis. Correlation was calculated within 1cM windows and a fitted line was drawn connecting adjacent values. We consider a 250cM region in each simulated population sample and the first 250 cM of chromosome 1 in the WHI AA. (B) Boxplots are shown for pairwise correlations of local ancestry values for positions across chromosomes. The simulated populations' chromosomes were two independent chromosomes of length 250cM and the WHI AA chromosomes calculated for the first 250 cM of chromosomes 1 and 2. The color represents population sample source.

Figure 7 displays the average λ_{GC} for admixture mapping of inferred local ancestry in WHI AA across simulated replicates of the phenotype, calculated on test statistics for all chromosomes that did not have a simulated causal effect. Inflation factors for the primary analysis model increase linearly with effect size of the causal locus, with severe over-inflation at larger effect sizes. Secondary analyses with adjustment for local ancestry at the causal locus fully correct for the over-inflation, with an average inflation factor of 1.0 at all effect sizes. **Figure 8** shows the primary analysis model Manhattan plots (A-D) at effect sizes 0, 0.5, 1.0 and 1.5 for a randomly chosen replicate. In this replicate, the simulated effect is on chromosome 3 but as the simulated effect size

increases, signals on other chromosomes become more significant, creating false positive associations on chromosomes 1, 5 and 9. The significance threshold used for this assessment, 2.24×10^{-5}, matches the threshold obtained via simulation using the WHI-SHARe AA genotypes with simulated null phenotypes (results described in Chapter 2).

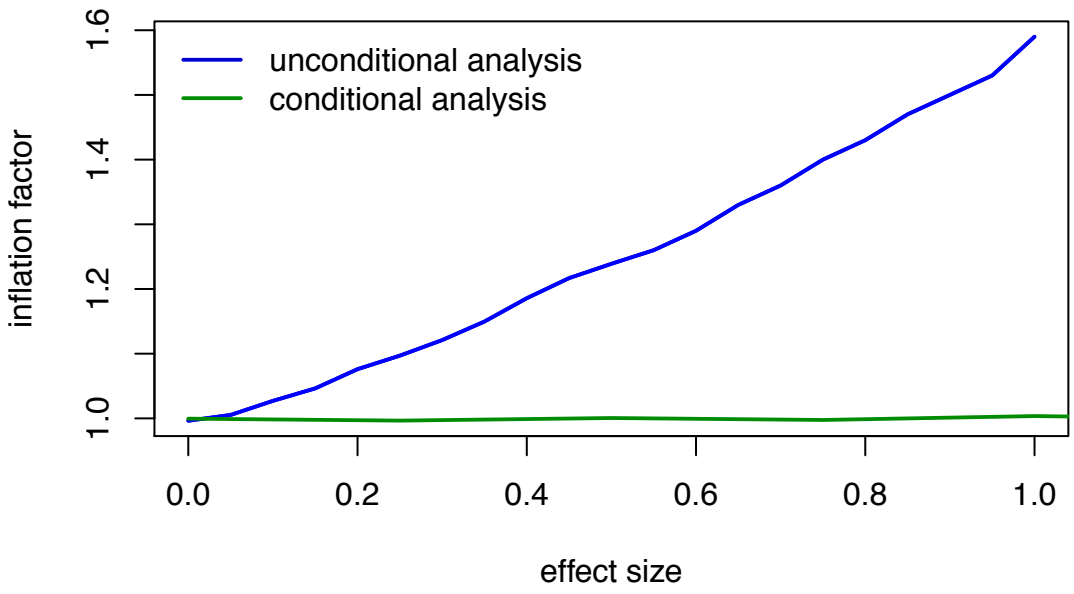


Figure 7: Inflation Factors by Effect Size

Each line represents the average inflation factor across simulated replicates observed at each simulated effect size, calculated for markers on a separate chromosome from the effect locus chromosome for each model. The primary analysis model tested dosage of local ancestry at each marker, adjusting for the first four principal components. The conditional analysis model included ancestry at the simulated causal locus as an additional covariate in the model. Color represents analysis model.

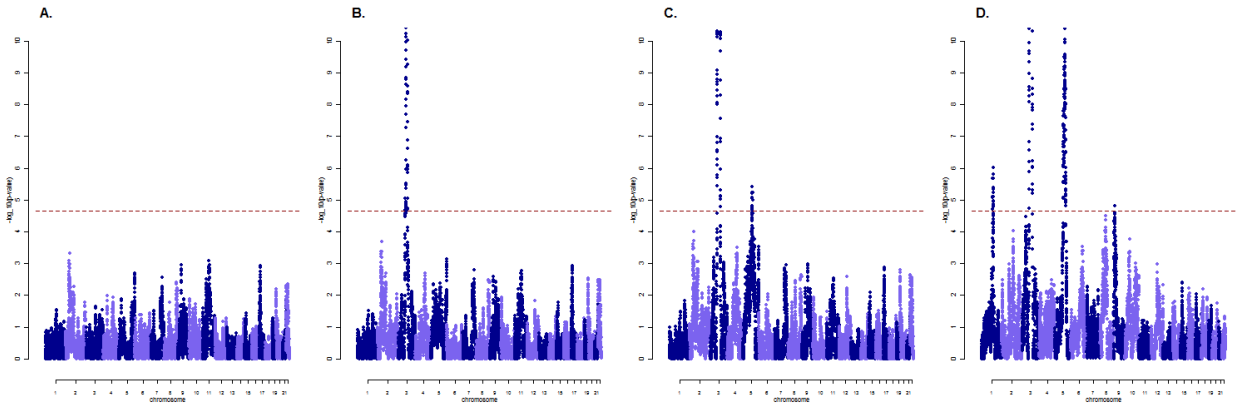


Figure 8: Manhattan plots for an Example Replicate

Plots showing the -$\log_{10}$ p-value for each position genome-wide for a randomly chosen replicate, ordered by position and colored by alternating chromosomes. Panels A-D show simulated effect sizes 0, 0.5, 1.0 and 1.5, respectively. The dotted horizontal line denotes the 2.24×10^{-5} p-value significance threshold.

4.4 Discussion

We conducted a thorough analysis of admixture mapping test statistics using 8,421 WHI-SHARe African Americans in order to further understand and characterize the behavior of local ancestry effects. We utilized HapMap samples as reference panels to conduct local ancestry analyses. For each study individual, local genetic ancestry at each marker was inferred and we used these local ancestry values to perform admixture mapping using linear regression, adjusting for principal components as fixed effects in primary analyses and including the local ancestry at the causal variant as an additional covariate in secondary analyses.

We observe over-inflation of test statistics when simulated local ancestry effects are sufficiently large. In particular, positions on chromosomes without a simulated effect show increased significance as the true local ancestry effect increases. These results illustrate that careful consideration should be given to both global and local ancestry patterns in a sample when performing admixture mapping. When ancestry-related assortative mating is present in admixed populations, the distribution of global ancestry can be more diverse and subjects are more likely to share ancestry across chromosomes

due to similar parentage as compared to admixed subjects where all individuals have the same proportional ancestry. Assortative mating for ancestry leads to a greater admixture-LD both within and across chromosomes, as we demonstrated in our simulation studies. When performing admixture mapping in admixed populations with population structure, we recommend performing secondary analyses that condition on local ancestry at the top hit to determine if the remaining signals are real or caused by confounding due to long range admixture-LD. In our simulations, conditioning on local ancestry at the causal variant eliminated over-inflation at all effect sizes.

Genetics studies of admixed populations present unique challenges and require prudence on behalf of the investigator. As demonstrated here, the complex patterns of local ancestry in admixed populations can affect the results of an admixture mapping analysis. We believe that the characterization of admixture mapping test statistics when there is long-range and across chromosome LD that is provided in this paper will be useful to future admixture mapping studies, as well as a variety of other genetic applications that rely on admixture-LD within admixed populations.

Chapter 5

ADMIXTURE MAPPING WITH LINEAR MIXED MODELS

5.1 Introduction

Current implementations of admixture mapping methods test one ancestry for association at time, where the count of local ancestry from a reference ancestral group is compared to the non-reference ancestral group [29-34] in a regression framework. When two ancestral populations are present, (e.g. African Americans), this is a direct comparison of both ancestral populations because the non-reference ancestral group consists of only one ancestral population. When three ancestral populations are present, the non-reference ancestral group consists of two ancestral populations combined. This model tests the effect of one ancestry compared to having either of other ancestries at that locus. Assessing the effect of each ancestry requires running multiple models and this is unsatisfactory.

Furthermore, regression assumes subjects are unrelated, however, many genetic studies now include individuals with some degree of relatedness [16]. Failure to properly correct for relatedness and population structure can result in inflation of the test statistics [17]. When applied to GWAS data, mixed model methods have been shown to protect against spurious associations in structured samples by directly accounting for sources of dependence including cryptic relatedness and population stratification [18,19].

We propose a linear mixed model-based approach for admixture mapping, AdmMix-LM, implemented using local ancestry estimates based on genome-wide data and an empirical relatedness matrix that can jointly test more than two ancestries, where population structure is accounted for with both fixed and random effects. We assess the power of our method across a range of SNP effect sizes, additive genetic variance parameters and allele frequency differences at the causal local across ancestral populations.

We apply our method to analyze traits in WHI-SHARe AA and HCHS/SOL subjects, comparing results to those acquired using regression. Analyses performed on WHI AA described in Chapter 2 include an unrelated set of WHI AAs. For analyses considered here, we include the full set of subjects. Similarly, when comparing results to regression in HCHS/SOL, we include the full set of subjects. In both data sets, regression

shows extreme over-inflation of the test statistics. For HCHS/SOL we compare results with MLM-based association mapping.

As a real data example, we look at uACR in HCHS/SOL. Increased urine albumin excretion, or albuminuria, is associated with a higher lifetime risk of end-stage renal disease (ESRD) and with increased cardiovascular disease risk [75,76]. Both albuminuria and ESRD differ by racial/ethnic groups in the U.S. with the lowest and highest risks noted in European and Amerindian populations, respectively. Using sex-specific cut-points, albuminuria prevalence in the U.S. is 10.3% in whites, 13.6% in African Americans, 9.9% in Mexicans Americans [77], over 20% in American Indians [78], and 12-14% in Hispanics/Latinos of the Hispanic Community Health Study / Study of Latinos (HCHS/SOL) [79]. Hispanic/Latinos also have an approximately two-fold higher risk of ESRD than whites [80]. However, Hispanics/Latinos are a heterogeneous group who show diversity in ancestry background including Amerindian, European and West African [81]. The percentage of African and Amerindian ancestry have previously been associated with albuminuria prevalence in Hispanic/Latino populations [82,83].

Despite the strong evidence for a role of ancestry in chronic kidney disease (CKD) susceptibility, few studies of kidney traits have leveraged the known genetic admixture in Hispanic/Latino populations to discover potential chromosomal regions that may harbor variants which confer risk for CKD traits such as albuminuria. Among the two genome-wide significant loci that have been identified and consistently replicated for albuminuria, *CUBN* (chromosome 10) genetic variants are associated with albuminuria in individuals of European ancestry and Hispanic/Latinos [84], and the *HBB* variant related to sickle cell trait (chromosome 11) is African-specific and associated with albuminuria in Hispanic/Latinos with African admixture [85]. An additional African-specific gene associated with albuminuria and CKD is *APOL1* [86]. Our recent work in the HCHS/SOL has confirmed a high proportion of Amerindian ancestry among Mainland Hispanics (Mexican, Central and South American), who also had low proportion of African ancestry [87]. However, Mainland Hispanics have similar mean albuminuria and frequency of increased albuminuria compared to individuals of Caribbean background (Cuban, Dominican, Puerto-Rican), in spite of the absence of African-specific risk

variants (*APOL1* or *HBB*). This evidence suggests the presence of Amerindian ancestry variants in Hispanic/Latinos influencing albuminuria.

There is great potential for genetic studies of CKD in Hispanics/Latinos to provide new insight into population-specific variants that confer risk for CKD traits but have not been uncovered through genome-wide association studies (GWAS). Admixture mapping leverages the known genomic heterogeneity of admixed individuals for improved genetic discovery, by identifying loci that contain genetic variants with highly-differentiated allele frequencies among ancestral populations that are also significantly associated with a trait. It can use local ancestry at genomic regions to capture both common and rare variants. Prior research leveraging admixture has successfully identified the *APOL1* alleles as a strong risk factor for hypertensive-attributable CKD, focal segmental glomerulosclerosis and HIV nephropathy in African Americans [86,88].

We use AdmMix-LM to identify loci that may harbor variants which increase albuminuria risk in a large population of Hispanic/Latinos. We identified a new locus at chromosome 2 which harbor Amerindian-specific variants associated with albuminuria and these findings were replicated in a cohort of Pima Indians.

5.2 Methods

5.2.1 Linear Mixed Model for Admixture Mapping with K Ancestral Populations

Suppose we collect N subjects that are admixed from K ancestral populations indexed by $k = 1, \dots, K$. We propose a linear mixed model (LMM) extension to (1) for quantitative traits:

$$\boldsymbol{Y} = \mathbf{1}\alpha + \boldsymbol{X_j}\boldsymbol{\beta_j} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \qquad (2)$$

where $\boldsymbol{X_j}$ is an $N \times (K-1)$ matrix of ancestry allelic dosages for locus j with corresponding effect size vector $\boldsymbol{\beta_j}$, which is a vector of length $K-1$. The matrix $\boldsymbol{W}$ represents covariate adjustment variables such as PCs with corresponding fixed effect vector $\boldsymbol{\gamma}$. We assume $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Phi}\sigma_a^2 + \boldsymbol{I}\sigma_e^2)$, where $\boldsymbol{\Phi}$ is a relatedness matrix and $\boldsymbol{I}$ is an

identity matrix. The parameters $\sigma_a^2$and σ_e^2 represent additive genetic and environmental variances, respectively. This model can extended to include additional random effects for more complex sampling designs. Generalized least squares can be used to fit this linear mixed model to test the null hypothesis H_0: $\boldsymbol{\beta_j} = \mathbf{0}$. The variance components, σ_a^2 and σ_e^2, are estimated once under the null using restricted maximum likelihood (REML).

The interpretation of admixture mapping coefficients when three ancestral populations are present is not immediately obvious but can be found by letting $(\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3})$ be three vectors of local ancestry calls at a locus of interest for subjects admixed from three ancestral populations. We will let the third ancestral population serve as the reference population. The model is then given by:

$$\boldsymbol{Y} = \mathbf{1}\alpha + \boldsymbol{a_1}\beta_1 + \boldsymbol{a_2}\beta_2 + \boldsymbol{W\gamma} + \boldsymbol{\epsilon}. \quad (3)$$

β_1 is the average increase in Y for each additional copy of an population 1 allele, holding all other covariates constant. Holding all other covariates constant implies $\boldsymbol{a_2}$ is held constant. Therefore, β_1 is effect of substituting one population 1 allele for one population 3 allele and β_2 is effect of substituting one population 2 allele for one population 3 allele. Finally, (β_2- β_1)= effect of substituting one population 2 allele for one population 1 allele.

5.2.2 Admixture Mapping in WHI-SHARe AA

The use of mixed models in association mapping is widely accepted for protection against false positives due to sources of structure. The problem has been assessed in the context of association mapping but not admixture mapping. To determine whether the problem exists in admixture mapping and to what degree, we perform the 'naïve' analysis including all WHI-SHARe AA subjects in a linear regression admixture mapping via PLINK for log white blood cell count (WBC), adjusting for global ancestry proportions.

5.2.3 Admixture Mapping in HCHS/SOL

We use genotype and local ancestry data from HCHS/SOL to analyze log urinary albumin-to-creatinine ratio (uACR) with AdmMix-LM. Our primary analysis included a joint test for all three local ancestries. The significance threshold used is described in the section below. As secondary analyses, for chromosomes containing SNPs that reached

genome-wide significance in our primary analysis, we conducted both single ancestry admixture mapping tests to identify the ancestry driving the signal and association mapping with linear mixed models as a comparison. For all analyses considered, the covariate adjustment included sex, age, the first five PCs, recruitment center, smoking, log of sampling weight and genetic analysis group (a six level categorical variable derived using self-identified background). We excluded any individuals with missing outcome or covariate data, Asian outliers identified with PCA, clinical conditions, blood/lymph malignant tumor, bone cancer, pregnancy, chronic kidney disease, chemotherapy, and blood cell count outliers (blasts or immature cell >5%). To obtain a more normalized trait distribution, uACR was log-transformed. Due to the sampling of participants in HCHS/SOL, we chose to add two additional random effect parameters for the association and admixture mapping mixed models: block group and shared household.

5.2.4 Assessment of Power

We assessed the power of our method using simulated trait data that reflected the complex correlation structure present in HCHS/SOL, varying the difference in allele frequencies across ancestral populations at the causal locus. To do this, we first calculated allele frequencies within ancestral populations using ADMIXTURE. Based on these frequencies, we separated variants into three groups, variants that were lowly differentiated, moderately differentiated and highly differentiated across populations based on the maximum allele frequency difference between any two of the three populations falling within differences of <0.1, 0.1 to 0.5, and >0.5. Within each of these groupings, we randomly selected 500 SNPs at which to simulate a causal SNP effect. For each SNP, we drew phenotypes, $\mathbf{Y}$, from a $N(\delta \boldsymbol{g}, \boldsymbol{\Phi}\sigma_a^2 + \boldsymbol{I}\sigma_e^2)$, where, $\boldsymbol{g}$ is the dosage vector of the causal SNP, $\delta \in [0.001,1.5]$ is the SNP effect size. We set $\sigma_e^2 = 1$, and varied σ_a^2 such that the heritability of the simulated trait, $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, varied between 0 and 0.5.

For the comparison with regression we drew phenotypes from N($\delta \boldsymbol{g}$, $\boldsymbol{I}$) and compared power at various effect sizes. To correct for any over-inflation in the regression

model, we compared power for each using an empirically derived honest false positive rate (with N($\mathbf{0}$, $\boldsymbol{I}$)) of 5% for each.

Our simulations use real genotypes, with simulated phenotypes that are different for each SNP being tested with the underlying difference in allele frequencies across ancestral populations varying across SNPs. This framework leads to more realistic admixture association signals to what would be expected in real study settings as compared to using simulated genotypes. Covariate and ancestry adjustment included principle components and genetic analysis group. We recorded power as the proportion of p-values exceeding our chosen significance threshold for locations falling within a 5 kB window surrounding the causal SNP. We record a genome-wide genomic control inflation factor, λ_{GC}, for each simulated replicate.

We assess power in terms of the amount of phenotypic variance explained by a simulated causal locus instead of effect size. One key quantity of interest in genetic studies is the amount of heritability explained by individually significant single-marker associations [59]. In simulation, one can calculate this quantity knowing the true effect size of the simulated causal variant using the formula from quantitative genetics $h^2_{causal} = 2p(1-p)\tau^2$ where p is the frequency of the SNP and $\tau^2 = \frac{\delta^2}{\sigma_a^2 + \sigma_e^2}$.

To obtain insight for when admixture mapping and association mapping results might differ at rare variants not tagged well by SNP data but that appear on an ancestral background, we performed a simple simulation study. We simulated genotypes and their local ancestry at two loci with correlation R ∈ [0,0.3] for 5,000 admixed subjects from two populations assuming a uniform distribution of global ancestry values across subjects. One locus was chosen as the causal locus. Phenotypes were simulated from N(0, gB) where g is the genotype at the causal locus and B taking values of 0.1 or 0.2. Frequency differences at the causal locus were constructed to vary between 0 and 0.9. We estimated the power (the proportion of p-values<0.05) using the genotype at the causal locus, the proxy locus (which was correlated with the causal locus) and the local ancestry at the causal SNP as the predictor of interest in a regression of the phenotype, adjusting for global ancestry proportions.

5.3 Results

5.3.1 Admixture Mapping in WHI AA

Figure 9 shows the admixture mapping results of WBC using regression with signs of serious inflation (λ_{GC}=1.6). **Figure 10** shows the analysis using AdmMix-LM still identifying the known signal on chromosome 1 in the Duffy gene region but with no signs of inflation (λ_{GC}=0.86). Results with adjustment for PCs were identical and are not shown.

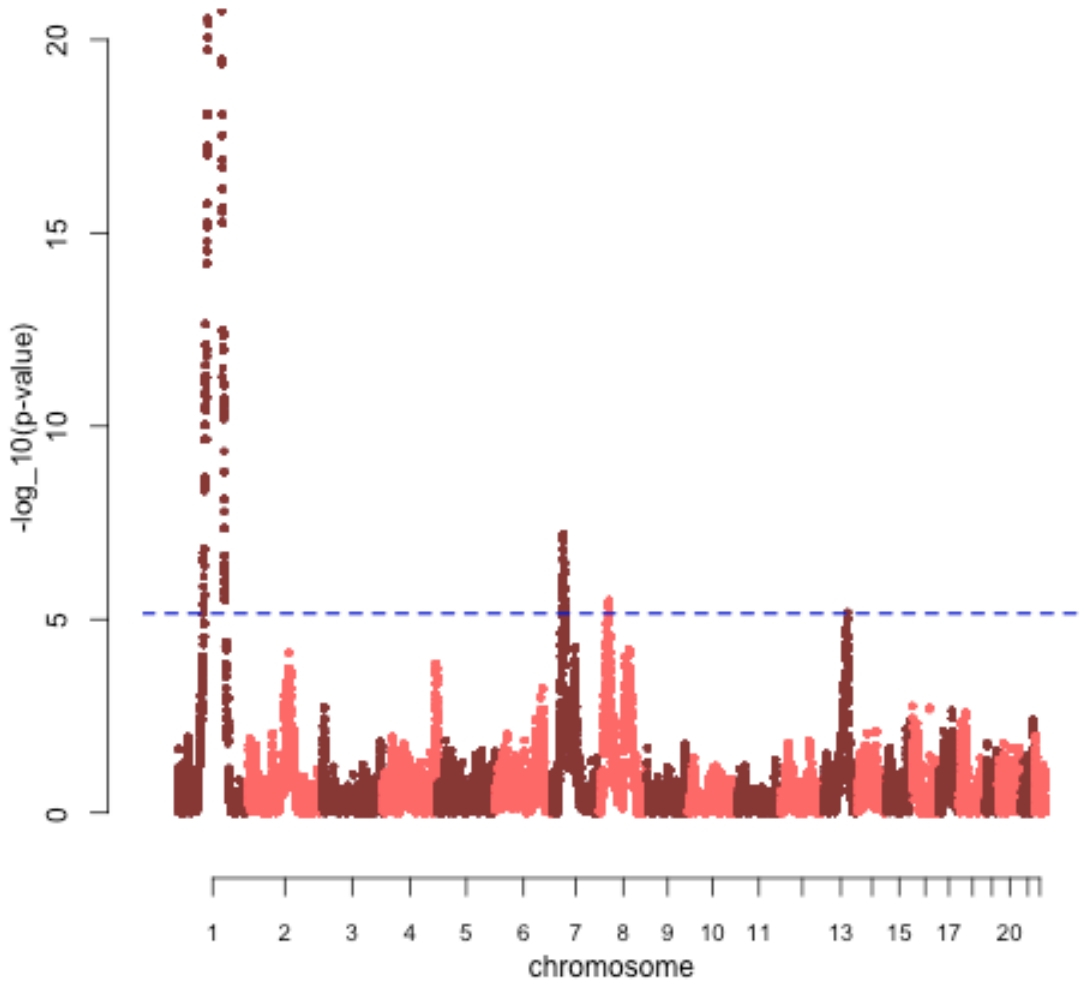


Figure 9: Manhattan Plot for log White Blood Cell Count using Linear Regression in Full WHI-SHARe African American Cohort

Manhattan plot of $-log_{10}(p\text{-values})$ at all unique local ancestry inferred positions from SNPs in common to WHI-SHARe and reference panels for WBC in the full African

American cohort of the WHI-SHARe study. The dotted grey line indicates genome-wide significance (p-value $< 7.6\times10^{-6}$). Genome-wide significant SNPs appear in regions on chromsomes 1, 7, 8 and 13.

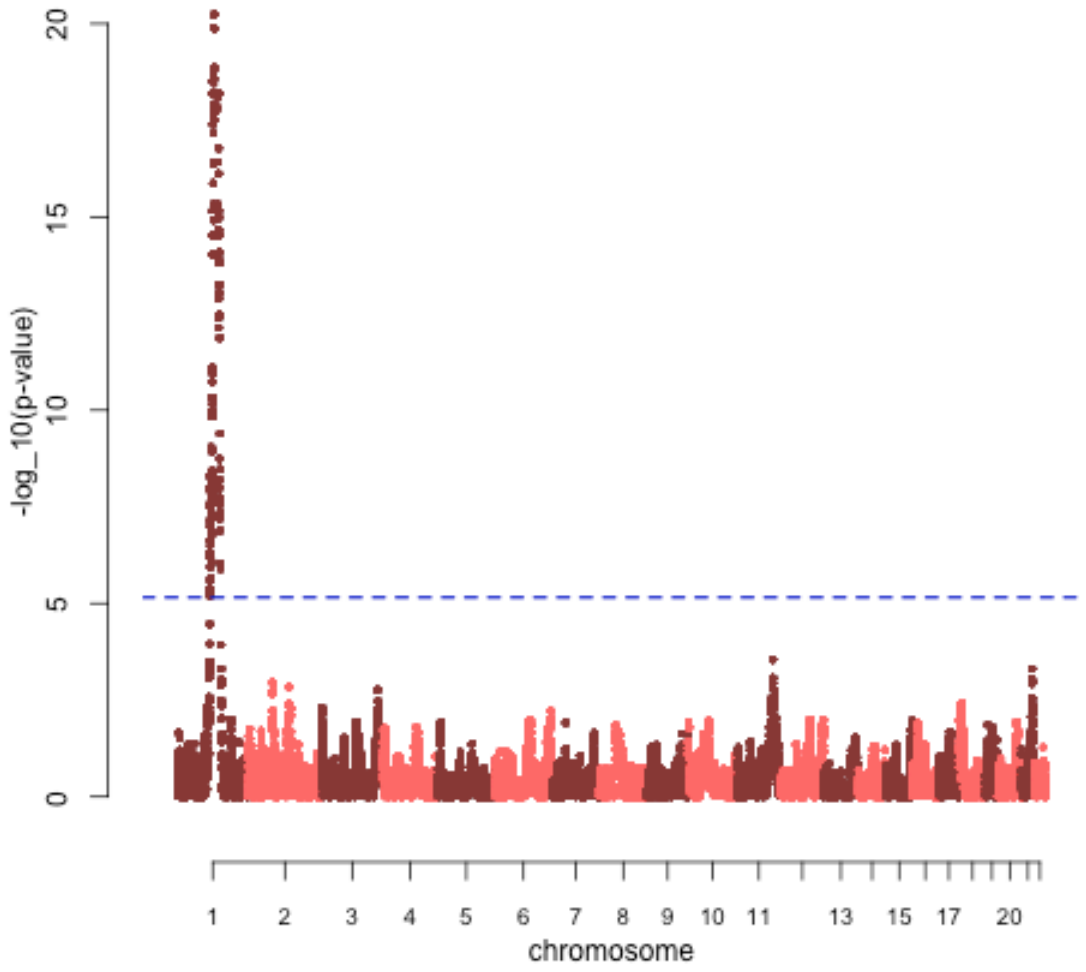


Figure 10: Manhattan Plot for log White Blood Cell Count using Linear Regression on PC-AiR PCRelate Unrelated Subjects

Manhattan plot of $-log_{10}(p\text{-values})$ at all unique local ancestry inferred positions from SNPs in common to WHI-SHARe and reference panels for log WBC in the unrelated African American cohort of the WHI-SHARe study. The dotted grey line indicates

genome-wide significance (p-value $< 7.6\times10^{-6}$). The only SNPs reaching genome-wide significance are in a region of chromosome 1.

5.3.2 Admixture Mapping in HCHS/SOL

The median urine albumin to creatinine ratio (UACR) was 6.54 mg/dl, interquartile 4.49 and 12.24. Three percent of individuals had an estimated glomerular filtration rate (eGFR) < 60 ml/min/1.73 m^2 and 14.1% had albuminuria (UACR ≥ 17 mg/g in men and ≥25 mg/g in women). We used AdmMix-LM on log-transformed albuminuria, conducting a joint test for three ancestries (Amerindian, West African, European) at genomic loci, adjusting for age, sex, study center, and principal components of ancestry. This admixture mapping analysis identified three genome wide significant regions (**Figure 11**), with no signs of inflation (λ_{GC} = 1.11). The most significant peak spans 2q11.2-q14.1 with a minimum p-value achieved (p=5.2 x 10^{-7}) at 2q13. The second spans 11q13.2-q13.4 with a peak at 11q13.3 (p=3.4 x 10^{-5}). The third peak is located in the 16p13.3 gene region (p= 3.6 x 10^{-5}). Secondary admixture mapping analyses that tested the local ancestry of each ancestral group separately revealed that the 2q13 locus was significantly associated with local Amerindian ancestry, whereas the 11q13 locus was associated with both local Amerindian and local European ancestry, and the 16p13 locus was associated with local African ancestry (**Figure 12).** Association analyses performed using imputed genotypes within a 10Mb window around 2q11.2-q14.1 yielded two SNPs significantly associated with UACR: rs116907128 (minor allele frequency [MAF] = 0.14, p= 1.5x10^{-7}) and rs586283 (MAF=0.35, p= 4.2x10^{-7}, linkage disequilibrium=0.11), located 5' upstream of *BCL2L11* (**Figure 13**).

To determine if these two identified SNPs on chromosome 2 account for the admixture mapping findings, we repeated the admixture mapping analyses using these SNPs as covariates. Adjusting for rs116907128 markedly reduced the admixture mapping signal (p=0.0018) at the chromosome 2 locus, whereas conditioning on rs586283 slightly attenuated the association findings (p=1.4x10^{-4}). Analyses including both SNPs showed similar results to conditional analysis on rs116907128. Neither of these SNPs were associated with eGFR in Hispanic/Latinos (rs116907128, p=0.97 and rs586283, p=0.15) or diabetes (p=0.81 and 0.46, respectively). SNP rs116907128 is monomorphic in 1000

Genome Project European and African samples, but it is a common variant among full heritage Pima Indians (allele frequency=0.55). In a replication analysis conducted in American Indians predominately of Pima Indian heritage, rs116907128 nominally associated with UACR as a continuous trait (N= 2,364; p=0.03) and type 2 diabetes (N= 7,635; p = 0.049; OR= 1.11 [95% CI 1.00-1.23]), but not with diabetic nephropathy in a case (UACR>300 mg/g or end stage renal disease) versus control (UACR<300 mg/g) analysis (N= 2,465; p= 0.23; OR= 1.10 [95% CI 0.94-1.30]).

We used Haploreg to query the evidence for regulatory function of SNPs on chromosome 2 in the Epigenome Roadmap Project and ENCODE data [89]. rs116907128 overlaps enriched regulatory annotations for dnase I hypersensitivity sites (DHS) in multiple tissues including fetal kidney and immune cells. This variant also overlaps transcription factor binding sites, including histone H3K4me1, H3K4me3, H3K27ac and H3K9ac in several cell lines.

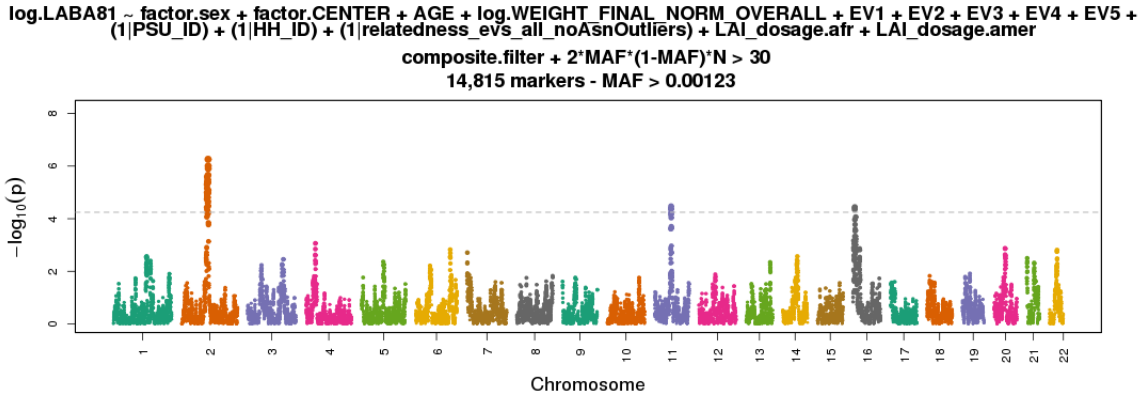
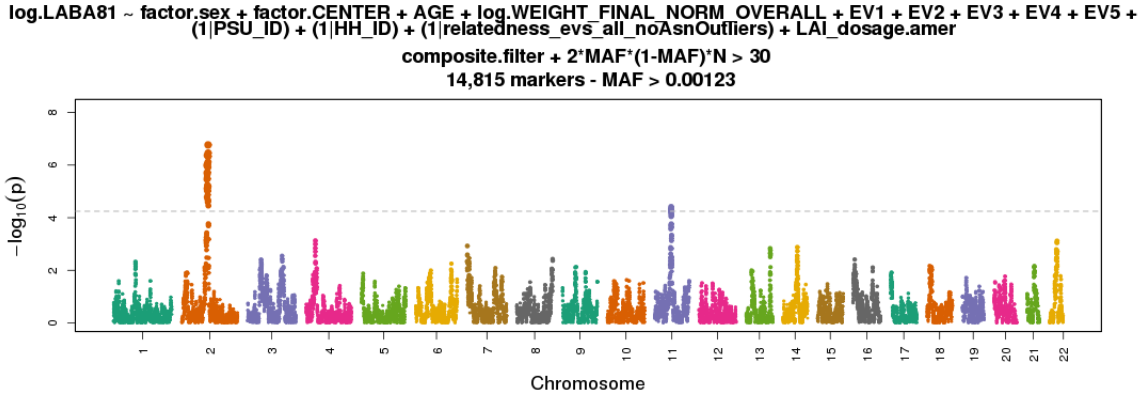


Figure 11: Manhattan Plot of Joint Test Admixture Mapping for log Urinary Albumin-to-Creatinine Ratio

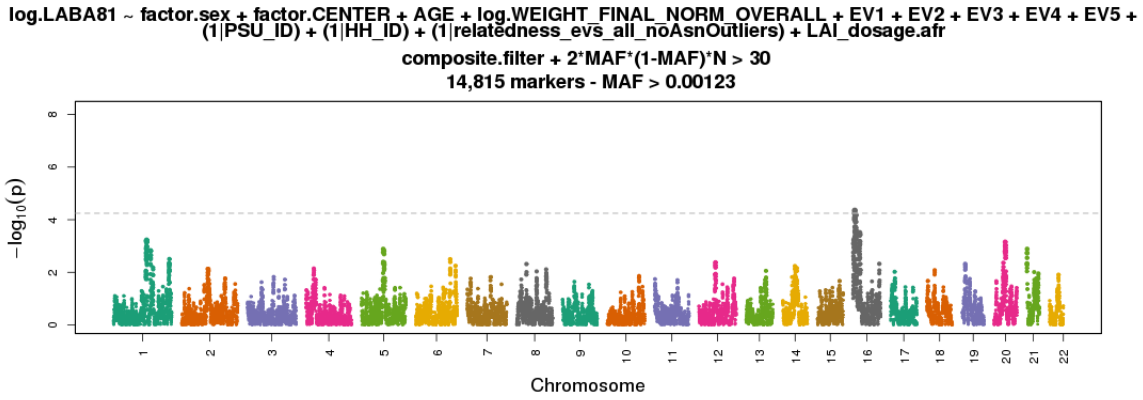
Manhattan plot of $-log_{10}(p$-values) at all unique local ancestry inferred positions from SNPs in common to HCHS/SOL and reference panels for log urinary albumin creatinine ratio in HCHS/SOL. The dotted grey line indicates genome-wide significance (p-value

$< 5\times10^{-5}$). Genome-wide significant SNPs appear in regions of chromosomes 2, 11 and 16.

Native American vs. Other



African vs. Other



European vs. Other

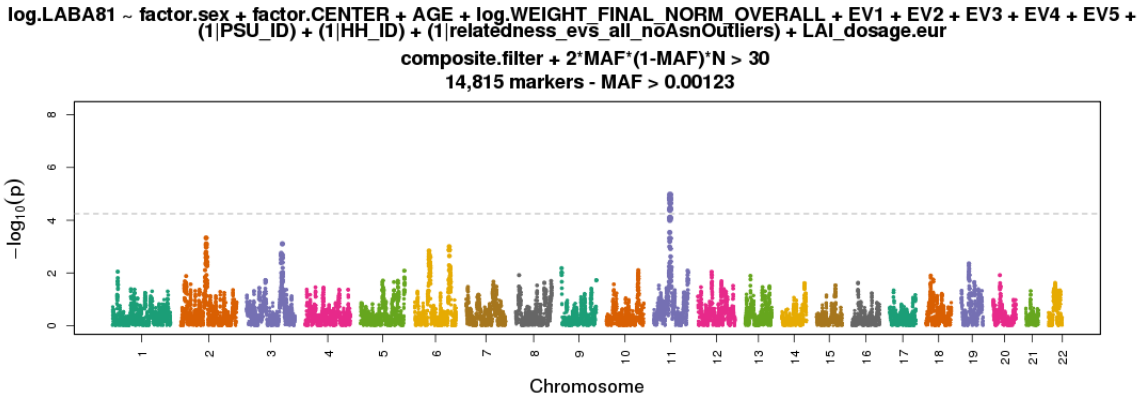
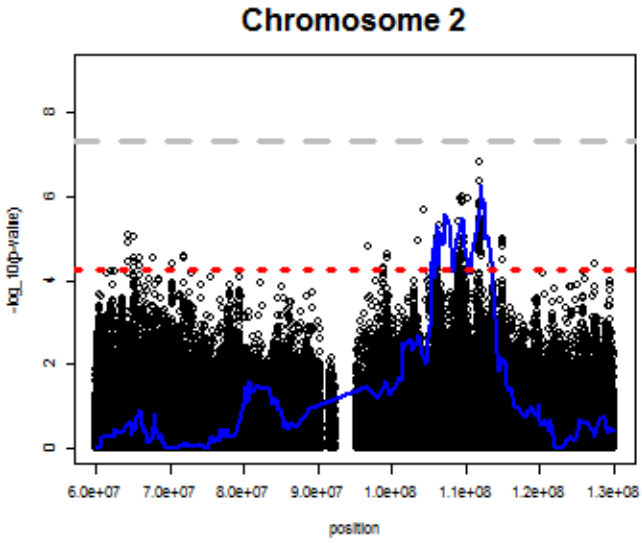
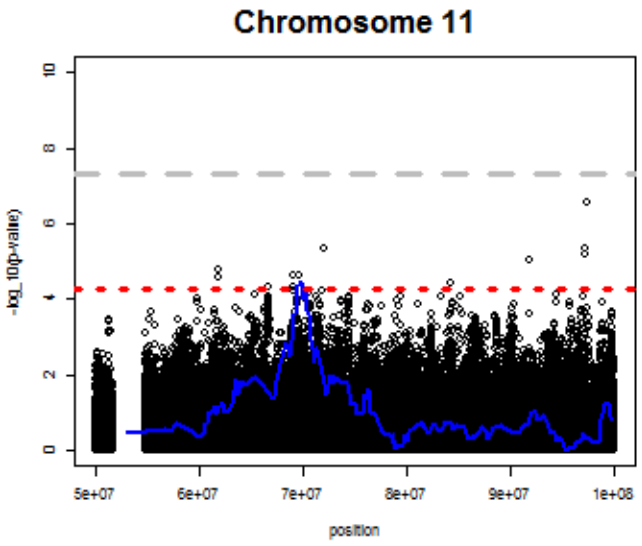


Figure 12: Manhattan Plot of Single Ancestry Tests for log Urinary Albumin-to-Creatinine Ratio

Panels showing the Manhattan plots of $-log_{10}(p$-values) at all unique local ancestry inferred positions from SNPs in common to HCHS/SOL and reference panels for log urinary albumin-to-creatinine ratio in HCHS/SOL. The dotted grey line indicates genome-wide significance (p-value $< 5\times10^{-5}$). Local ancestry associations at the most significant SNPs lying in a region of chromosome 2 are Native American driven. Tests for Native American and European ancestry yield SNPs that share a genome-wide significant region on chromosome 11. Tests for African ancestry yield genome-wide significant SNPs in a region on chromosome 16.


Chromosome 2
-log_10(p-value)
0
2
4
6
8
6.0e+07
7.0e+07
8.0e+07
9.0e+07
1.0e+08
1.1e+08
1.2e+08
1.3e+08
position


Chromosome 11
-log_10(p-value)
0
2
4
6
8
10
5e+07
6e+07
7e+07
8e+07
9e+07
1e+08
position

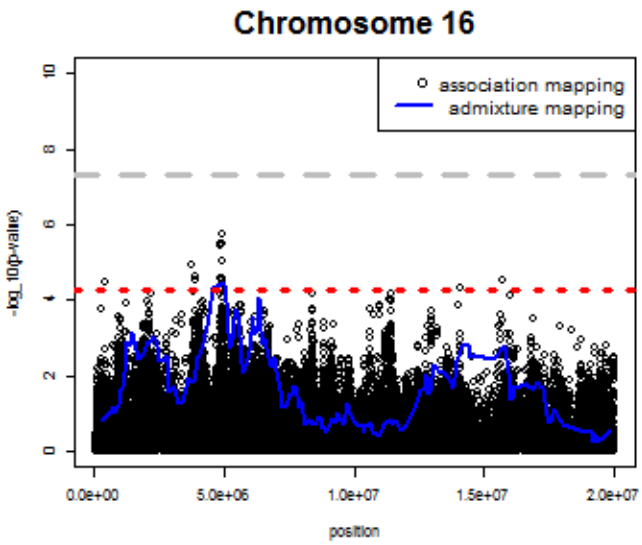

Chromosome 16
association mapping
admixture mapping
-log_10(p-value)
0
2
4
6
8
10
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
position

Figure 13: $-log_{10}(p$-values) for Association and Admixture Mapping of log Urinary Albumin-to-Creatinine Ratio on Chromosomes 2, 11 and 16

Plot showing the $-log_{10}(p$-values) for association and admixture mapping of log urinary albumin-to-creatinine ratio for positions on chromosomes 2, 11 and 16. Association results are filtered to exclude variants with MAF < 0.00123 and MAC < 30. Grey dashed and red dotted lines indicate genome-wide significance for association and admixture mapping, respectively. Admixture mapping results are shown in blue and the association mapping results are shown in black. SNPs in region 2q12.1-q14.1 show SNPs with p-values of similar magnitude across association mapping and admixture mapping.

5.3.2 Assessment of Power

Figure 14 illustrates the power and λ_{GC} for lowly, moderately and highly differentiated SNPs at various effect sizes using 5.4×10^{-5} as a p-value threshold for determining genome-wide significance. As expected, the power to detect a genome-wide statistically significant SNP effect increases as the difference in allele frequencies increases across ancestral populations.

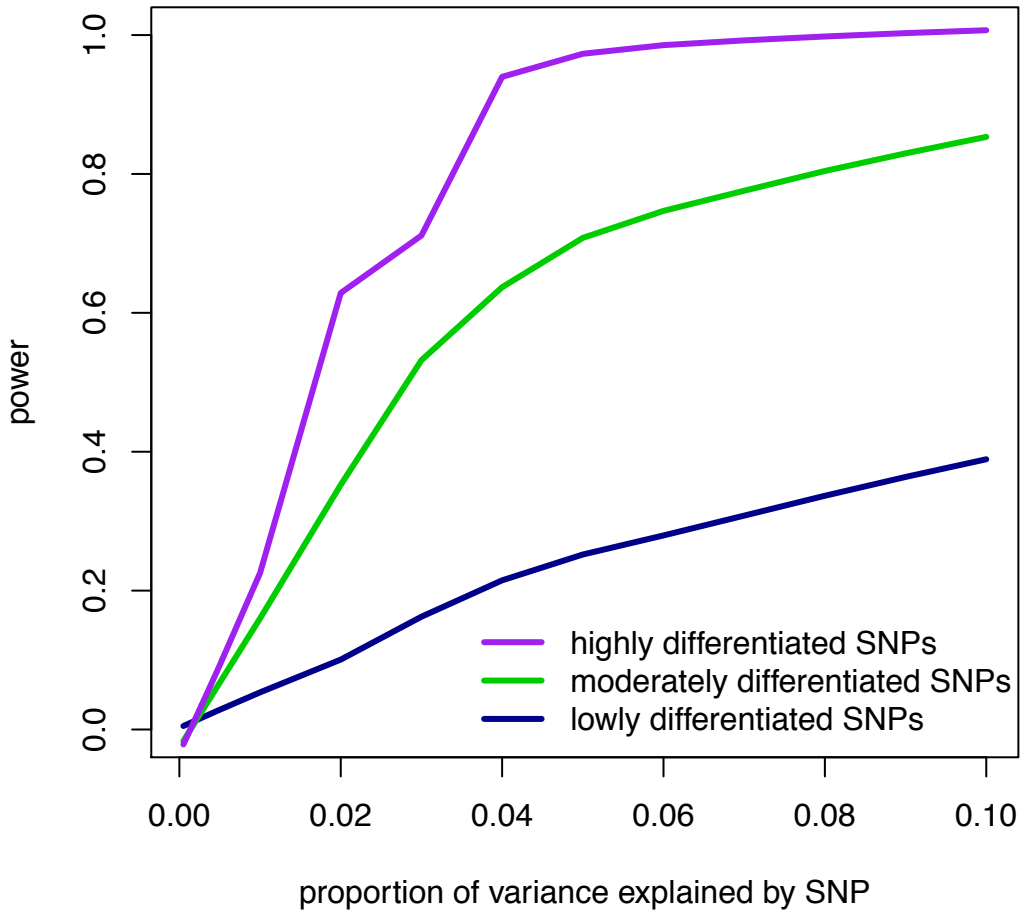


Figure 14: Power Curves for AdmMix-LM with $h^2 = 0.2$

The proportion of true positive associations identified (power) at a nominal significance level of $\alpha = 5\times10^{-5}$ by AdmMix-LM in all three classes of SNPs, is shown for each choice of h^2_{causal} for the causal SNP of interest.

Figure 15 illustrates the power and λ_{GC} for lowly, moderately and highly differentiated SNPs at various effect sizes comparing regression and AdmMix-LM at controlled 5% false positive rates. AdmMix-LM shows increased power over regression.

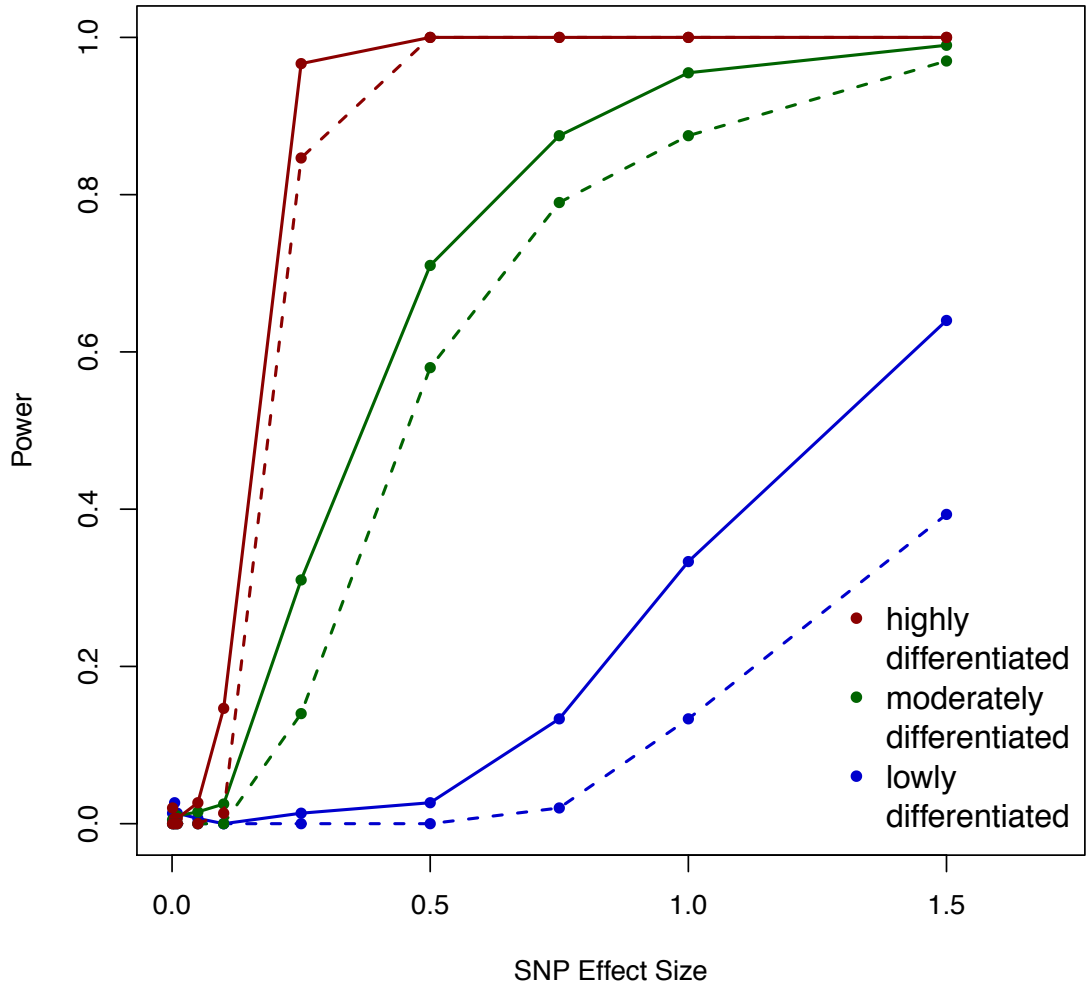


Figure 15: Power Curves for AdmMix-LM and Linear Regression with $h^2 = 0$

The proportion of true positive associations identified (power) at an honest false positive rate of 5% by AdmMix-LM (solid lines) and linear regression (dashed lines) is shown for each choice of effect size for the causal SNP of interest wide in all three classes of SNPs. The honest false positive rate for the linear regression model does not match the nominal level of the test, α, due to mis-calibration of the test at null SNPs.

Our secondary analysis revealed we had lower power to detect traits with high heritability (**Figure 16**), likely due to the increased correlation in trait values, which decreased the effective sample size. As observed in the null simulations, λ_{GC} decreased as heritability increased. Furthermore, we observe higher λ_{GC} with more differentiated SNPs.

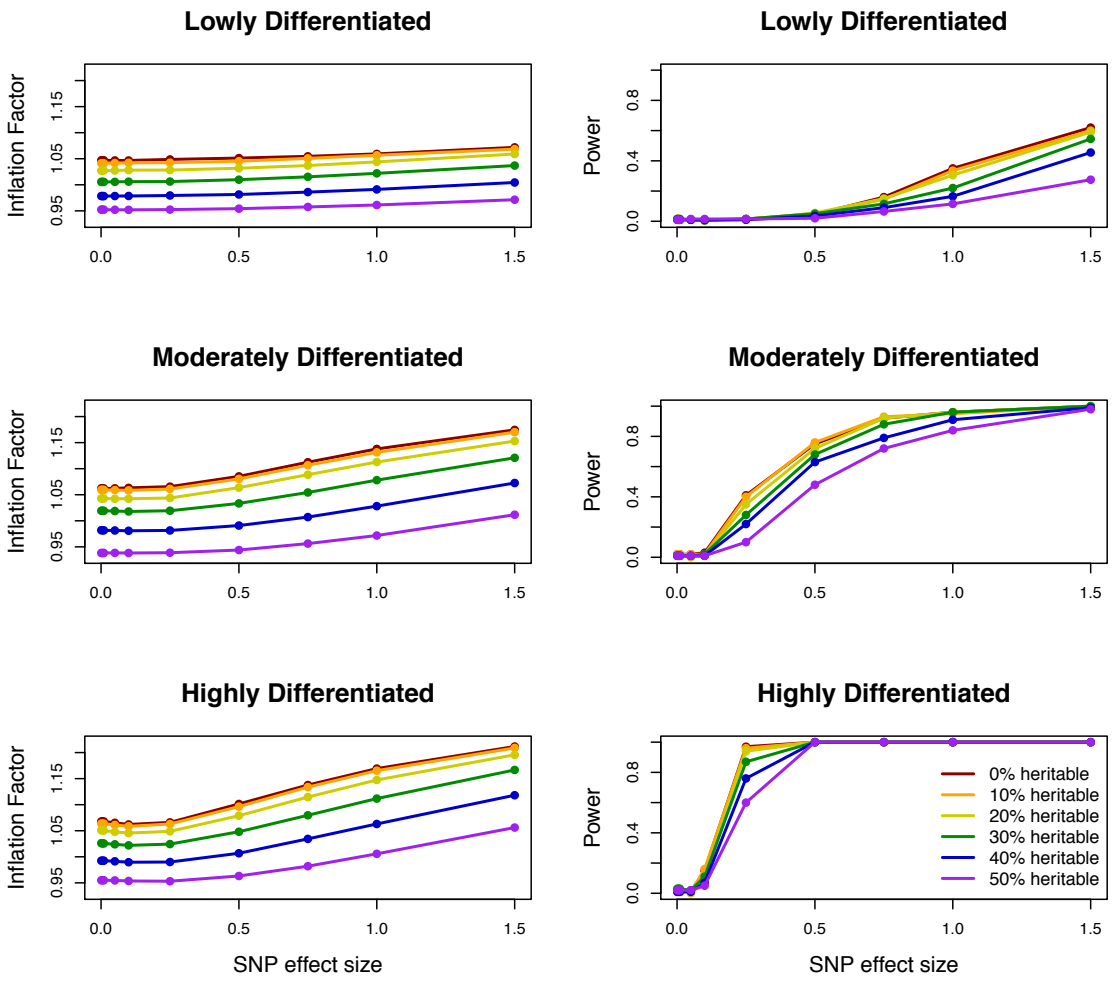


Figure 16: Power and Inflation Factor Curves for AdmMix-LM

Left panels show the genome-wide genomic control inflation factor, λ_{GC}, by AdmMix-LM in all three classes of SNPs and all six h^2 values considered at each choice of effect size for the causal SNP of interest. The right panels show the proportion of true positive

associations identified (power) at a nominal significance level of $\alpha = 5\times10^{-5}$ by AdmMix-LM in all three classes of SNPs and all six h^2 values considered for each choice of effect size for the causal SNP of interest.

Figure 17 shows the power to detect the causal locus using local ancestry versus a SNP that serves as a proxy for the true SNP. When r=0.2, local ancestry has greater power over the proxy SNP when the frequency difference between the ancestral populations is greater than 0.22. Local ancestry has higher power than using the marker (proxy) SNP when the allele frequency difference is greater than 0.46 when r=0.3. When r<0.2, local ancestry had higher power over the proxy SNP at all frequency differences (results not pictured).

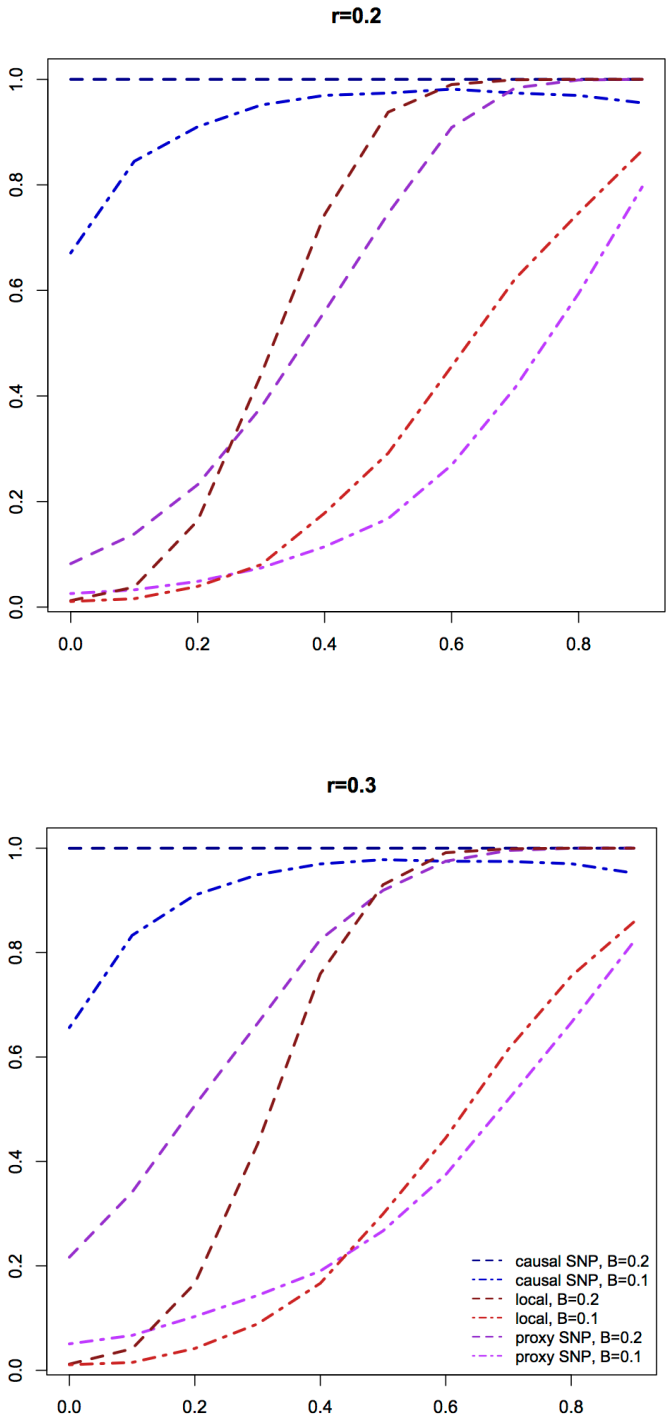


Figure 17: Power Curves for SNP and Local Ancestry Association Analyses
Panels showing the power to detect the causal SNP at p=0.05 when r=0.2 (upper panel) and r=0.3 (lower panel) between the causal and proxy SNP locus for each choice of allele frequency differences across ancestral populations. The parameter B represents SNP effect size. Color denotes predictor of interest used. Line type denotes effect size.

5.4 Discussion

We have described a powerful approach for gene mapping in admixed populations and have demonstrated (1) its ability to detect SNP associations not identified by mixed model association methods, (2) its ability to adequately control for sub-continental admixture and relatedness with the use of SNP array data, and (3) has adequate type I error rates when applied to data with relatedness. In this model, sample structure is addressed by both fixed and random effects. Population structure is accounted for by including ancestry representative PCs as fixed effects and relatedness is accounted for by modeling the trait covariance structure with an ancestry-adjusted relationship matrix.

Admixture mapping can yield complementary as well as different results from association mapping. This stems from the correlation structure of ancestry where ancestry is inherited in blocks whose length depends on the number of generations since admixture. Using local ancestry as a predictor of interest protects against unreliable behavior of SNPs with low minor allele frequency. As long as the average ancestry at each marker position is not rare for any of the ancestral groups, which will be the case for most Hispanic and African American populations, the test statistics do not need to be filtered. As a result, admixture mapping has the ability to capture both rare and common variation occurring on a genetic background. As shown here, admixture mapping can have greater power over association mapping for causal SNPs not well captured by genotype arrays when allele frequency differences across ancestral populations are substantial. Another advantage of admixture mapping is the reduced multiple-hypothesis testing burden compared to SNP association mapping, admixture mapping.

We used AdmMix-LM to identify novel loci which may harbor variants which confer risk for albuminuria. The main findings of our study are the identification of three novel genomic regions at chromosomes 2, 11 and 16 for UACR in this large and diverse population of Hispanic/Latinos. This approach identified an Amerindian-specific locus at chromosome 2, driven by a variant at 5' of *BCL2L11*, common in Amerindians but rare in other populations. To date, the only other validated genome-wide significant loci for UACR are *CUBN* and *HBB* related to sickle cell trait [84,85]. Two additional UACR loci

located at 2q21 (*HS6ST1*) and 11q14 (near *RAB38/CTSC*) were recently described in a GWAS of diabetic individuals (n=5,509 to 5,825) of European ancestry [90]. These regions do not overlap our identified loci. We previously have shown significant genome wide associations of *CUBN* and *HBB* with UACR in the HCHS/SOL study (Kramer, JASN, in press). The most significant *CUBN* variant has similar and very low allele frequency in Mainland and Caribbean Hispanic subgroups and therefore will not be expected to be identified in admixture mapping.

Our main admixture mapping finding is *BCL2L11,* which encodes a pro-apoptotic protein that belongs to the BCL-2 protein family. Several lines of evidence suggest a role of this gene and its protein in kidney development and kidney disease states. MicroRNA (miRNA)-mediated regulation of *Bcl1211* expression in mice plays an important role in nephron progenitor survival during kidney development [91]. MiRNA are involved in post-transcriptional repression of target mRNAs. Loss of miRNAs in nephron progenitors increased apoptosis and elevated expression of protein Bim (also known as BCL2L11) leading to a decrease in nephron number [91]. Bim null mice manifest systemic autoimmune disease and immune complex glomerulonephritis [92]. In an experimental study of diabetic nephropathy, increased advanced glycation end products (AGE) promoted *Bcl2l11* expression, leading to podocyte apoptosis [93].

The variant we identified, located at the 5' of *BCL2L11*, overlaps enriched regulatory annotations for DHS and histone marks in fetal and adult kidney, and in several other cell lines including immune cells. This evidence suggests a regulatory function on these cells and tissues. Among American Indians, this variant associated with urinary albumin creatinine ratio (N= 2,364; p=0.03) and type 2 diabetes (N= 7,635; p = 0.049) but not with eGFR or diabetic nephropathy. In HCHS/SOL Hispanics/Latinos, there was no association with either type 2 diabetes or eGFR. Since local ancestries of an individual are fixed over regions, the detected signal region is quite large. Therefore, to localize a region into a candidate SNP, we use association testing. In the region on chromosome 2, there was an apparent match between the admixture mapping signal region and the association-testing signal. However, in the regions detected on chromosome 11 and 16, there were no highly significant SNPs. This could be due to low power in association testing (e.g. due to relatively low combined MAF of the tag SNPs),

or due to lack of good proxies of the causal SNP. Further studies will require sequencing of the chromosomes 11 and 16 loci to identify variants (common and rare) accounting for the admixture signal in Hispanics/Latinos in addition to fine-mapping regions associated with Amerindian local ancestry in studies of American Indians.

Our study is limited by the available genomic markers (imputed and genotyped) for fine mapping of regions. Although we found strong associations of rs116907128 at the locus which accounted for the local ancestry association in the admixture mapping, it is possible that this SNP is correlated with one or more variants in the region, and it is not the "causal" variant. In addition, further studies examining longitudinal changes in kidney function and UACR are needed to better characterize the relevance of the *HS6ST1* gene to albuminuria and CKD, which may provide important clues on the clinical impact of this variant. The HCHS/SOL is currently examining the participants in second visit, which data could be used for these follow-up studies. Our study provides important information on the presence of Amerindian-specific genetic variants associated with albuminuria in Hispanics and American Indians.

The significance threshold varied slightly across heritability levels but remained on the order of 10^{-5}. Consistent with known results on admixture mapping, the power to detect loci increased as the maximum difference in allele frequencies across ancestral populations increased. While inflation factors are commonly reported for GWAS studies, they are not often reported in admixture mapping studies [10-15] and the behavior of local ancestry test statistics with regards to genome-wide genomic control inflation factors has not been well studied. In Chapter 3 we observed a positive linear relationship between the genome-wide genomic control inflation factor and local ancestry effect size for admixture mapping with linear regression. We observe a similar trend here, made more extreme as the allele frequency differences across ancestral populations increases. The protection against this, conditioning on a top result to elimination over-inflation and assess secondary signals, can be used in this context as well. We expect a largely polygenic trait to show a higher λ_{GC} so the decrease in λ_{GC} with increasing trait heritability was quite surprising.

Our linear mixed model method assumes that variances across subgroups are equal. We note that in various scenarios, this may not be true and may cause inflation

when prevalence differs across groups [46]. Although we have focused here on Hispanic Americans, in principle, our approach is also suitable for any admixed population.

Chapter 6

ADMIXTURE MAPPING WITH LOGISTIC MIXED MODELS

6.1 Introduction

Binary outcomes are common in genetic studies through case-control study designs. Linear mixed models assume a trait is quantitative but can be implemented to analyze binary traits by simply treating the binary outcome variable as if it were continuous. Within the linear mixed model, the assumption of constant residual variance can be violated when analyzing a binary outcome with covariates due to population structure. This can lead to inflated type I error rates. Recently, Chen et al. developed a logistic mixed model that correctly accounts for variance components of a binary trait. The same problems observed in the association mapping case for the analysis of binary traits using linear mixed models will also apply to admixture mapping. In Chapter 5, we developed the AdmMix-LM method for admixture mapping with linear mixed models that utilizes a joint test for all ancestries under consideration. Here, we implement a logistic mixed model for admixture mapping. We apply the logistic admixture mapping method to the ADSP Hispanic families where a genome-wide significant signal for Alzheimer's disease within the MYO16 gene is identified that is not detected by association mapping.

6.2 Methods

6.2.1 Logistic Mixed Model

Consider a sample of N individuals who are admixed from K ancestral populations. Let $\boldsymbol{Y} = (y_{1,}y_{2,}, \dots, y_{N,})$ be a binary trait vector, where $y_{i,}$ takes values of 1 or 0, if individual i is affected or unaffected, respectively. Assume that the subjects have been genotyped at m single nucleotide polymorphism (SNP) markers. We propose to model the relationship between the trait $\boldsymbol{Y}$ and the ancestry at position $j \in \{1, \dots, m\}$ using a logistic mixed model given by:

$$logit(\boldsymbol{\pi}) = \mathbf{1}\alpha + \boldsymbol{X_j\beta_j} + \boldsymbol{W\gamma} + \boldsymbol{\epsilon}, \quad (3)$$

where $\boldsymbol{\pi} = \Pr(\boldsymbol{Y} = \mathbf{1} \mid \boldsymbol{X_j}, \boldsymbol{W}, \boldsymbol{\epsilon})$ is the vector of probabilities of the binary outcome, $\boldsymbol{X_j}$ is an $N \times (K-1)$ matrix of ancestry allelic dosages for locus j with corresponding effect

size vector $\boldsymbol{\beta_j}$, which is a vector of length $K-1$. The matrix $\boldsymbol{W}$ represents covariate adjustment variables such as PCs with corresponding fixed effect vector $\boldsymbol{\gamma}$. We assume $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Phi}\sigma_a^2 + \boldsymbol{I}\sigma_e^2)$, where $\boldsymbol{\Phi}$ is a relatedness matrix and I is an identity matrix. The parameters $\sigma_a^2$and σ_e^2 represent additive genetic and environmental variances, respectively.

To test the association between the binary outcome $\boldsymbol{Y}$ and ancestry at position j, we first fit the null model:

$$logit(\boldsymbol{\pi}_0) = \mathbf{1}\alpha + \boldsymbol{W\gamma} + \boldsymbol{\epsilon}, \quad (4)$$

where $\boldsymbol{\pi}_0 = \Pr(\boldsymbol{Y} = \mathbf{1} \mid \boldsymbol{W}, \boldsymbol{\epsilon})$. Variance components are calculated using AI-REML. Once variance components are obtained, we use a Penalized Quasi-Likelihood method to iteratively update the fixed effects until convergence it reached, using a Cholesky decomposition to calculate matrix inversion at each iteration. Finally, we implement a score test for $\boldsymbol{\beta_j} = \mathbf{0}$.

6.2.2 Admixture Mapping in ADSP Hispanics

165 CEU and 203 YRI from Hapmap3 were used as reference for European and African populations, respectively. 63 samples from the Americas in the HGDP data were included as surrogates for Native American ancestry. Hapmap and HGDP data sets were merged, keeping SNPs in common to Hapmap and HGDP, leaving 603,611 SNPs with an overall genotyping rate of 0.998. We merged these reference panels with 545 Dominican Samples in the ADSP Family Study using PLINK, keeping SNPs in common to all data sets.

After applying a genotype missingness filter of 7%, this left a combined set of 273,523 SNPs with an overall genotyping rate of 0.996. We performed joint phasing of reference panels and Dominican Samples in the ADSP Family Study using Beagle version 3.3.2. We performed local ancestry estimation in Dominican Samples in the ADSP Family Study using RFMix version 1.5.4 assuming subjects are admixed from European, African and Native American populations, using Hapmap CEU, Hapmap YRI and HGDP Native American as reference. We calculated proportions of global ancestry from European, African and Native American populations for Dominican Samples in the

ADSP Family Study by averaging across local ancestry values in all 273,523 SNPs for which we had local ancestry estimates.

We implement with logistic mixed model admixture mapping with a joint test for all three ancestries described above with fixed effect adjustment for the first four principal components and a random effects for additive genetic and environmental variances. As a secondary analysis, we run single ancestry logistic mixed model admixture mapping to determine the ancestral group driving any observed signals. As a sensitivity analysis, we re-ran the logistic mixed model leaving out one family at a time. For comparison, we also run the analysis of AD in ADSP Caribbean Hispanic families using linear mixed model.

6.3 Results

Figure 18 displays the first two principal components for the ADSP Hispanic families and reference samples used for phasing and local ancestry estimation. ADSP Hispanic subjects cluster towards the European and African reference samples, reflecting their largely African America ancestry. **Figure 19** shows the average ancestry for each ADSP Hispanic subject, ordered by increasing European ancestry proportions. The proportion of European ancestry per subject ranges from 0.22 to 0.89 with an average of 0.65 and a SD of 0.13. African proportions ranged from 0.06 to 0.76 with an average of 0.27 and an SD of 0.14. Native American varied between 0.01 and 0.27 with an average of 0.09 and an SD of 0.03. Several outlying subjects showed particularly high Native American ancestry with one family in particular having more Native American ancestry than other families.

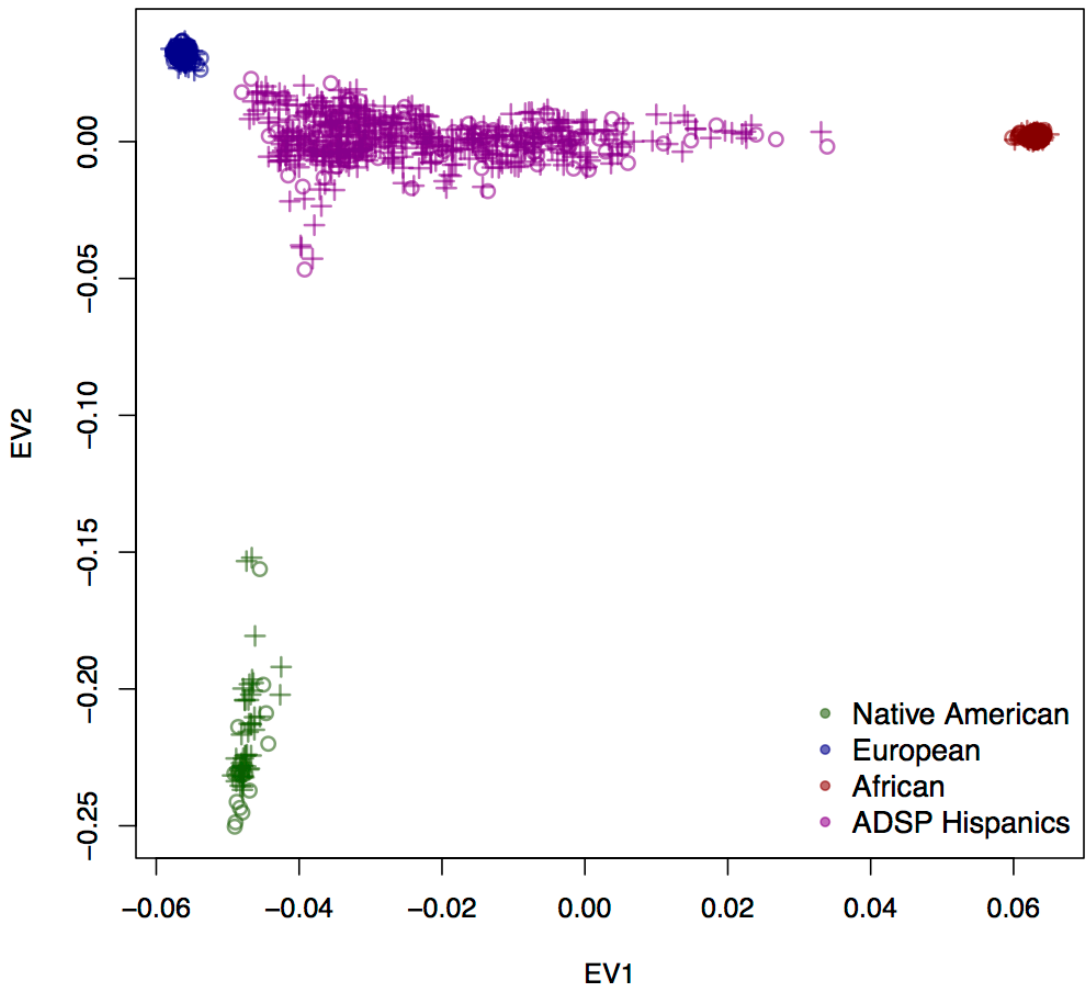


Figure 18: Principal Components

Plot of the first 2 EVs from PC-AiR. Each point represents one individual who is in the unrelated (o) or related (x) set. The color represents population source.

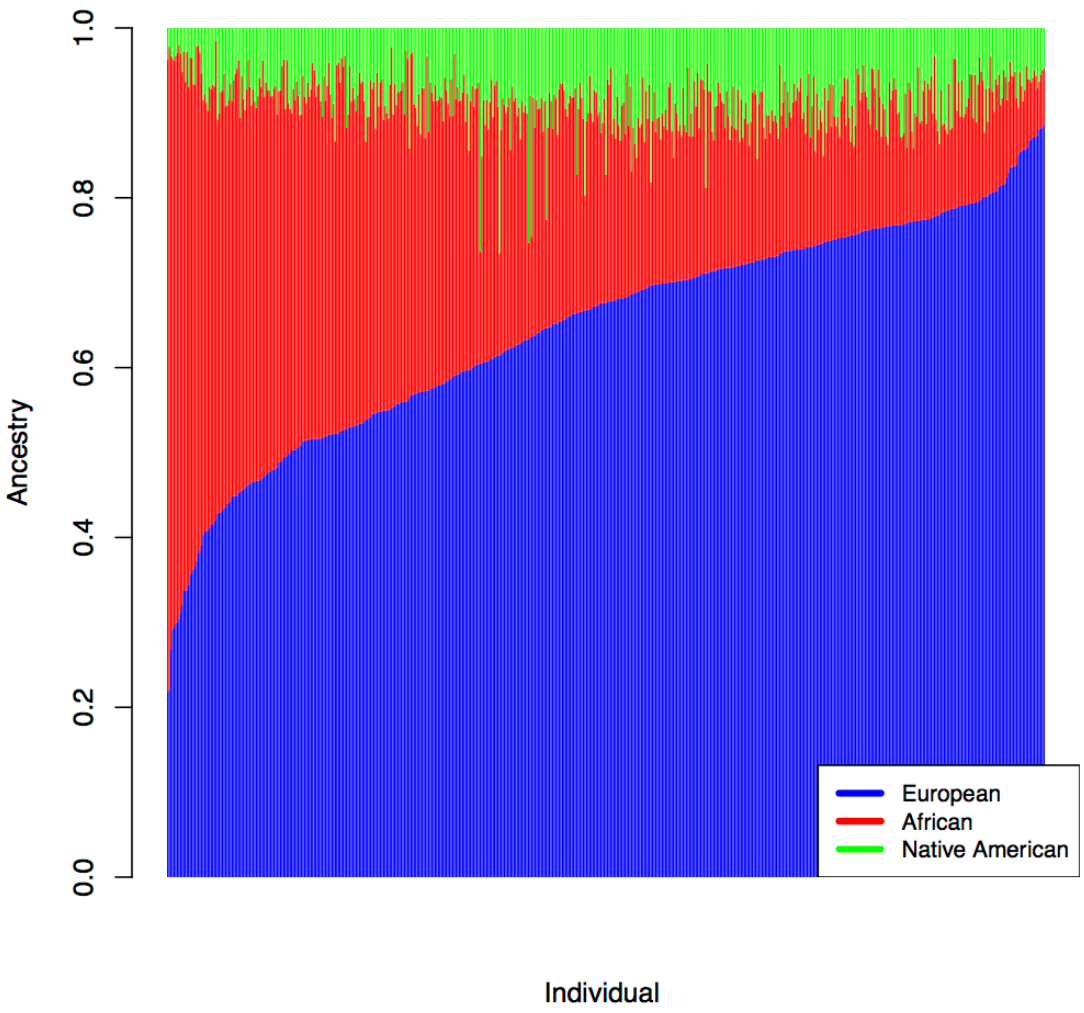


Figure 19: Global Ancestry Proportion Estimates
Bar plot illustrating the genome-wide global ancestry proportion estimates for each individual averaging estimated local ancestry from RFMix genome-wide. Each vertical bar represents one individual, and the proportion of that bar shown in each color represents the estimate of the genome-wide proportion of ancestry from the corresponding ancestral population. Individuals are arranged in order of increasing estimated European ancestry within group.

The p-value threshold obtained based on simulating a null phenotype was 4.5×10^{-5} (discussed in Chapter 3). **Figure 20** displays the Manhattan plot for admixture mapping with a joint test for all three ancestries. We find a genome-wide significant hit on chromosome 13 near the MYO16 gene region ($p=3.01\times10^{-5}$) which has been previously implicated in studies of AD [94] and dementia [95], and there exists evidence that it is expressed in the brain [96], making it a good candidate gene for AD. Our secondary analyses of admixture mapping testing single ancestries at a time revealed that the MYO16 region signal on chromosome 13 is driven by Native American ancestry (**Figure 21**). No single family appears to be driving the signal (**Figure 22**), as leaving each family out has a maximum change in p-value from 4.52 to 3.8 and 5.2 on the low and high end (on the –log10 scale). We had originally hypothesized that the heavily Native American family may be driving the signal but this does not appear to be the case. The family with the greatest influence on the signal, with a drop in p-value to 3.8, is not the family with the most Native American ancestry.

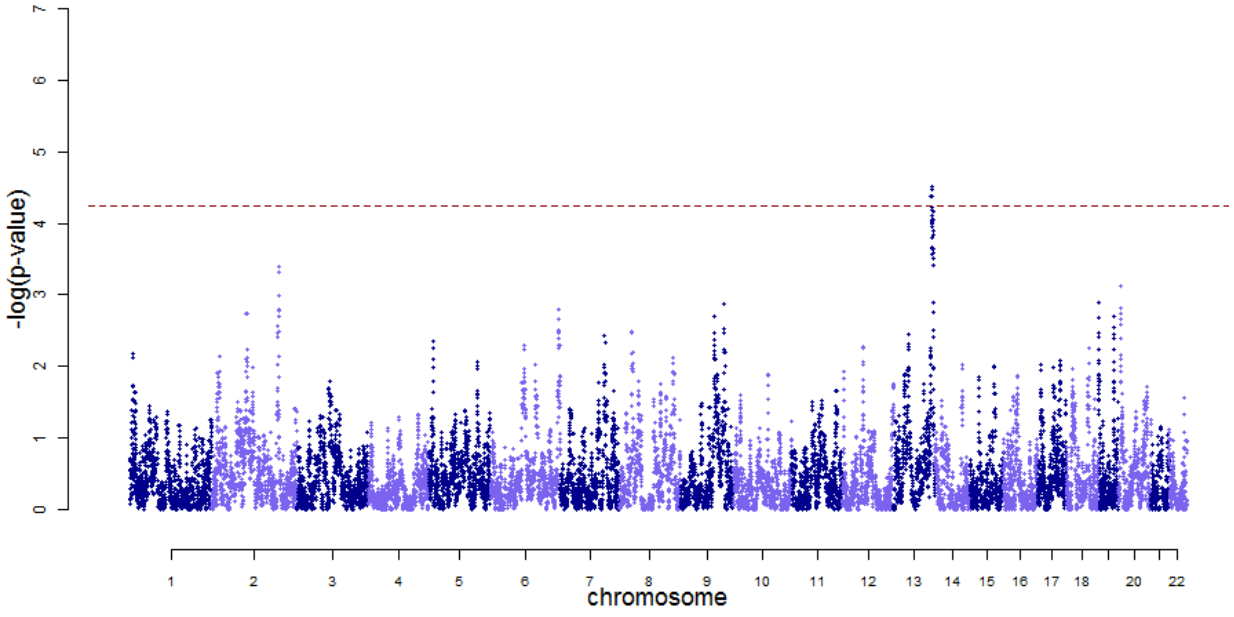


Figure 20: Manhattan Plot of Joint Test Logistic Mixed Model Admixture Mapping for Alzheimer's Disease

Manhattan plot of $-log_{10}(p\text{-values})$ at all local ancestry inferred positions from SNPs in common to ADSP and reference panels for Azheimer's Disease in the Caribbean Hispanic Family cohort of the ADSP. The dotted red line indicates genome-wide significance (p-value $< 4\times10^{-5}$). The only SNPs reaching genome-wide significance appear in a region of chromosome 13.

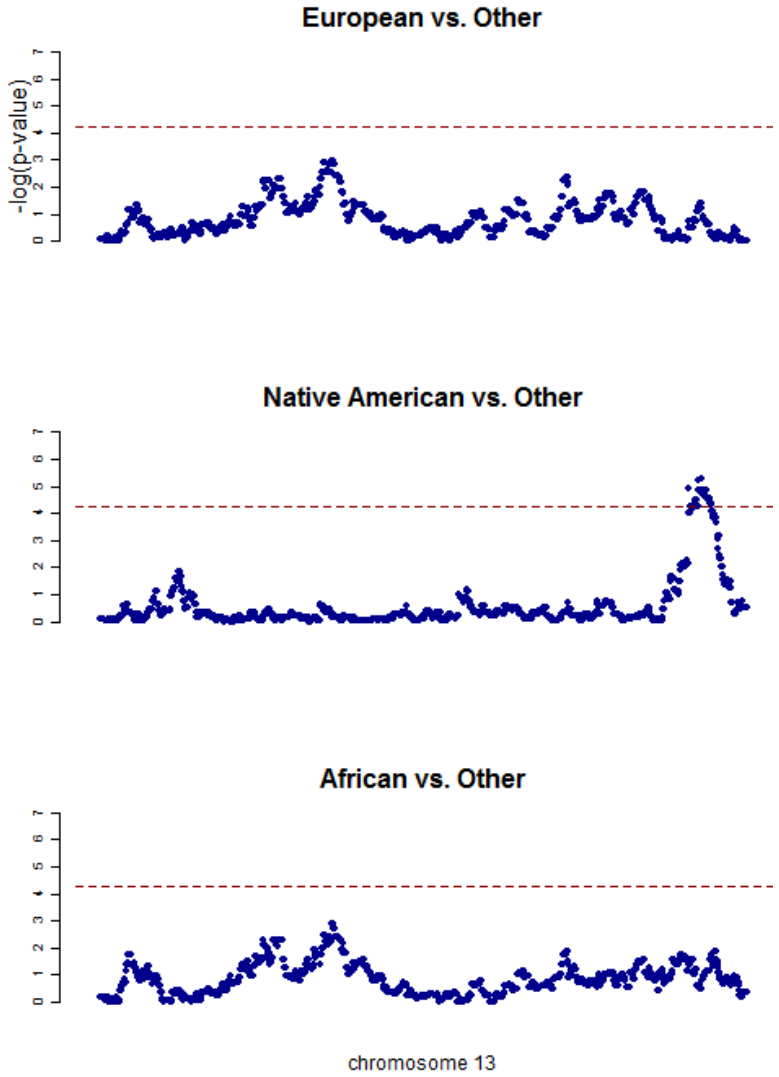


Figure 21: **$-log_{10}(p$-values) for Single Ancestry Logistic Mixed Model Admixture Mapping of Alzheimer's Disease on Chromosome 13**
Panels showing the plot of $-log_{10}(p$-values) at all local ancestry inferred positions from SNPs in common to ADSP and reference panels for Azheimer's Disease in the Caribbean Hispanic Family cohort of the ADSP. The dotted red line indicates genome-wide significance (p-value $< 5\times10^{-5}$). Local ancestry associations at the most significant SNPs lying in a region of chromosome 13 are Native American driven.

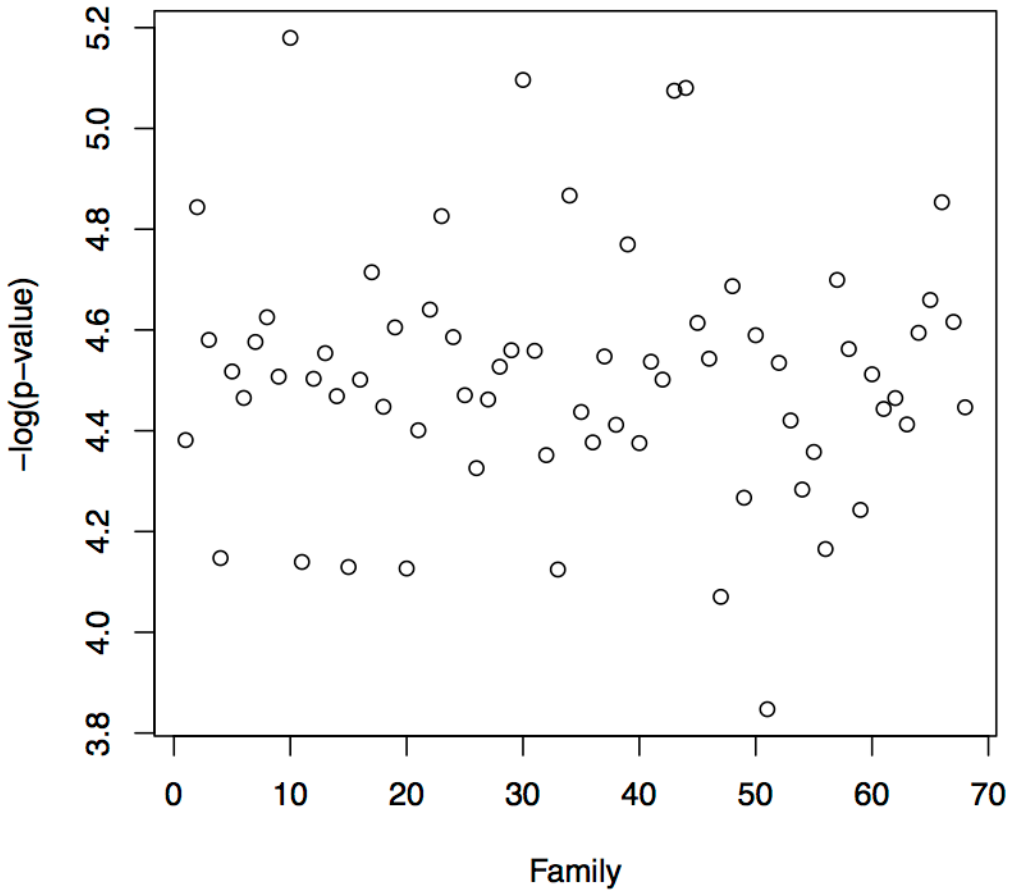


Figure 22: **Maximum $-log_{10}(p$-value) for Leave-One-Family-Out Logistic Mixed Model Admixture Mapping of Alzheimer's Disease on Chromosome 13**
Each circle represents one family. The y-axis denotes the maximum $-log_{10}(p$-value) from a logistic mixed model admixture mapping analysis excluding that family.

We compare the distribution of p-values using the logistic mixed model to the linear mixed model for admixture mapping, AdmMix-LM (**Figure 23**). For all p-values prior to the signal, the logistic mixed model stays closer to the expected 1-1 line than the linear mixed model. We find some slight deflation in the linear mixed model compared to the logistic mixed model, with the linear mixed model falling below the expected near p=0.001 with the logistic model gaining a slight power advantage over the linear mixed model near the top peak hit region. **Figure 24** shows the direct p-value comparison of the linear and logistic mixed model. While most values fall close to the 1-to-1 line, the models do not give identical results.

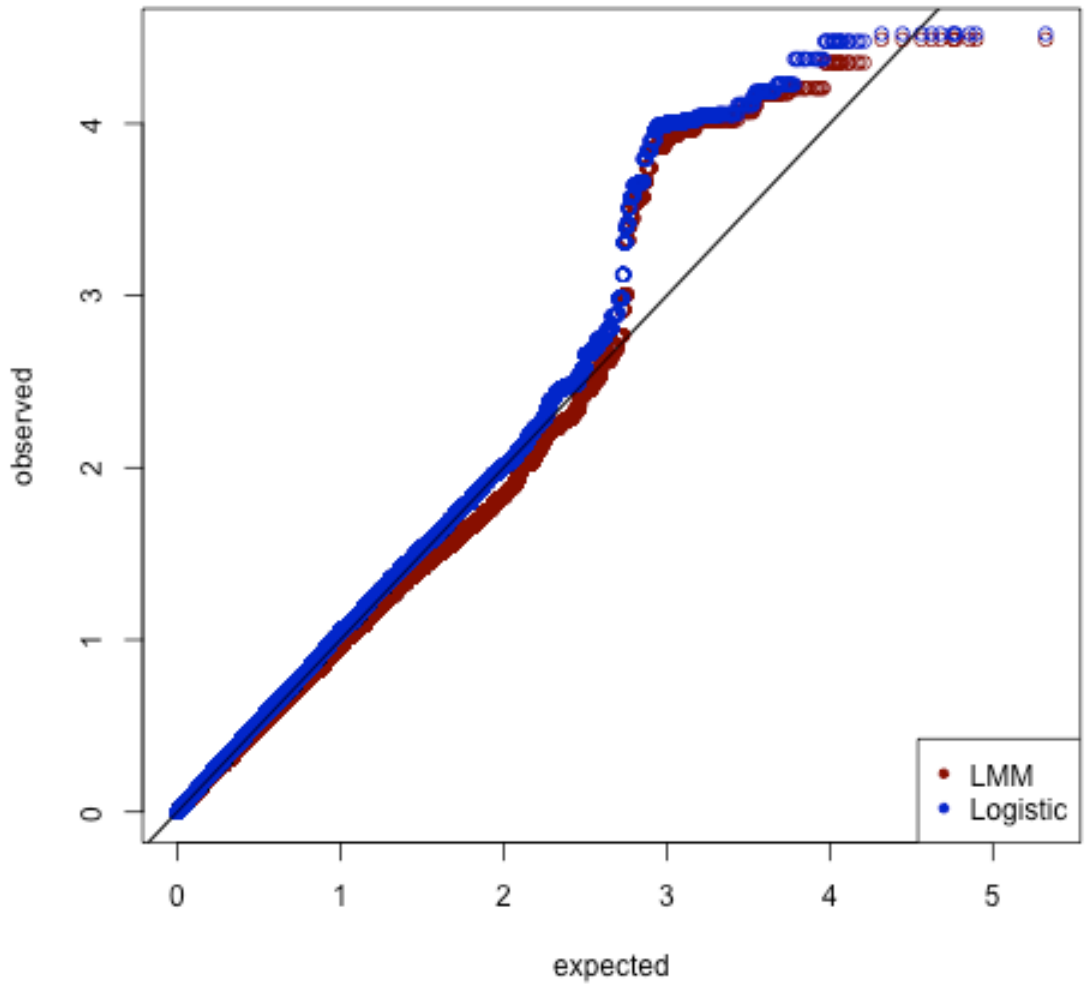


Figure 23: QQ-plot for Alzheimer's Disease
QQ-plot showing the distribution of $-log_{10}(p\text{-values})$ from linear and logistic mixed model admixture mapping for Alzheimer's Disease.

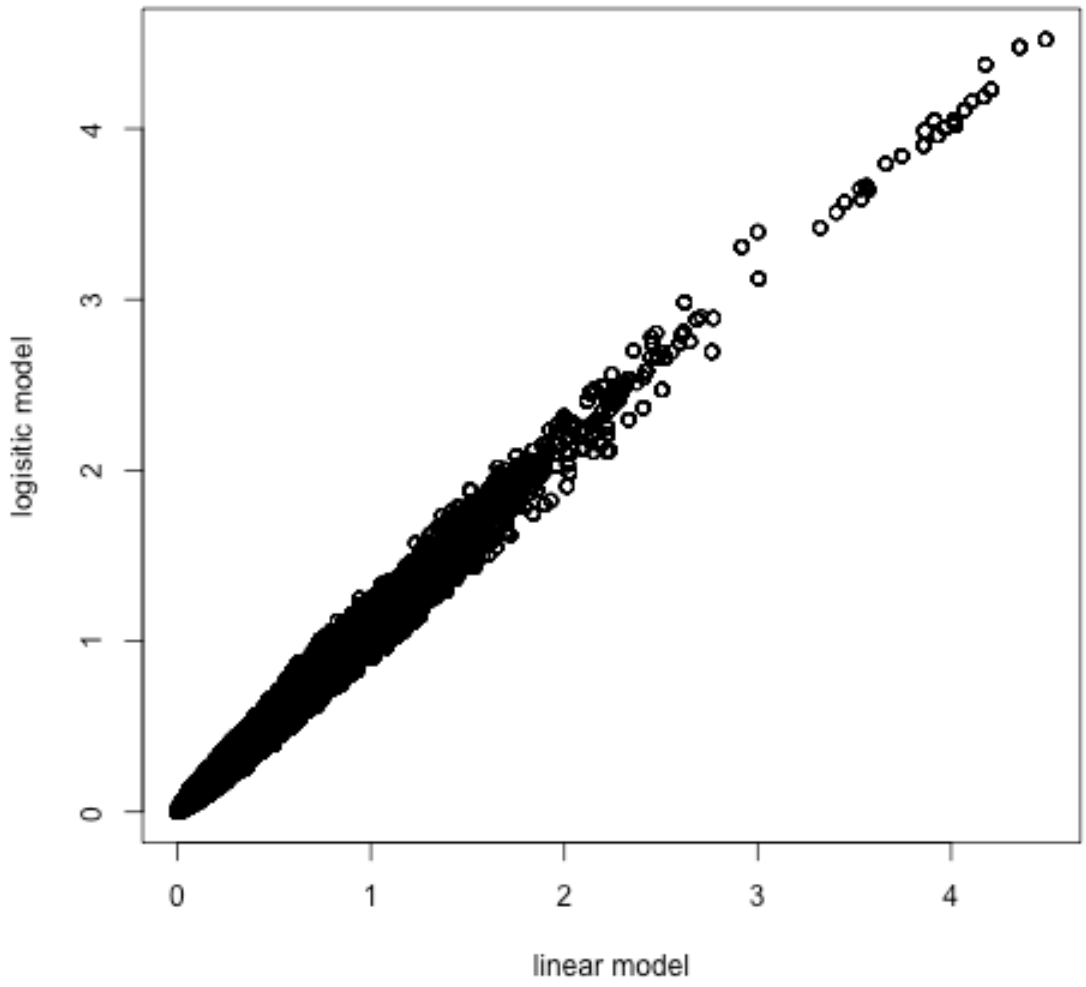


Figure 24: Logistic vs. LMM p-values
Each point represents a SNP with the $-log_{10}(p\text{-value})$ linear mixed model on the x-axis and the $-log_{10}(p\text{-value})$ logistic mixed model on the y-axis.

6.4 Discussion

We have proposed a logistic mixed model for admixture mapping that corrects for sources of structure and dependence in data including relatedness and population structure when analyzing binary traits. We demonstrate that our method had correct type 1 error rates and does not show the deflation that a linear mixed model does when

analyzing real data. In traits showing differential risk by ancestry subgroup, this method properly models the underlying variance of the trait and will not be inflated, where linear mixed models may run into issues of inflation. The logistic mixed model also gives a slightly more significant p-value at our region of interest compared to the linear mixed model. Given the wide use of the case-control study design in genetics, the methods presented here have large applicability to future admixture mapping studies.

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, we have shown how admixture mapping can be a powerful tool for discovering novel variants, extending previous work in a regression framework testing a single ancestral population at a time. We have developed a new method for implementing admixture mapping that substantially improves the robustness of admixture mapping methods to population structure and relatedness. We (1) demonstrate and compare methods for assessing genome-wide significance in admixture mapping (2) identify and show how to correct for long range admixture-LD in admixture mapping studies, and (3) implement linear and logistic mixed models to perform admixture mapping in the presence of relatedness and population structure. Our methods are designed for two or more ancestral populations and take into account the complex structure present in modern admixed populations.

Important directions for future work include extending the Siegmund-Yakir theory for two populations to directly model the correlation of ancestry and test statistics when three ancestral populations are present based on a multidimensional Ornstein-Uhlenbeck process. In addition, it would be valuable to further investigate the behavior of inflation factors in admixture mapping. Results presented here suggest that one should not over-interpret an inflation factor as they seem extremely sensitive to trait distributions and are not as stable compared to association mapping.

The problem of how to interpret admixture mapping results in conjunction with association mapping results presents an interesting problem. When genome-wide significant association and admixture mapping results agree, admixture mapping provides a nice complement to a main association mapping analysis. When admixture mapping identifies a genome-wide significant signal that is suggestive along the same order of magnitude in association mapping, one can conclude that it is possible the high multiple testing burden paid in association mapping dampened the signal and that the signal is real. Furthermore, when association mapping finds a result that admixture mapping does not, it is possible the signal is real given that admixture mapping needs a difference in allele frequencies to detect a results and this may not be present at the causal locus. Finally, in instances where admixture mapping finds a genome-wide significant SNP that

is not remotely suggestive in association mapping (e.g. p-values not less than 0.001), then follow-up and interpretation of admixture mapping results becomes tricky and less reliable. Careful consideration must be taken when performing and interpreting admixture mapping signals and results should never be considered alone without supporting information.

For the follow-up of association mapping in regions of interest identified by an admixture mapping scan, it is not immediately clear how far from a top peak one should search. In GWAS, regions directly surrounding top SNPs or at top SNPs only are investigated for functional relevance to the trait of interest. Peaks in admixture mapping are more continuous and cover a wider area. In admixture mapping, as with linkage mapping we may want to investigate an area surrounding a top SNP. The entire peak regions can be quite long, spanning several megabases. In linkage mapping, LOD-support intervals approximate 95% confidence intervals under modest assumptions are used to provide statistically precise and valid chromosomal regions for fine mapping following initial identification of broad regions that are identified via linkage. Given that we observe a locus, v_{max}, which maximizes the likelihood in a linkage scan, we define a LOD-support interval as all loci $V_{LOD,\varphi}$ such that

$log_{10}\left(LR\left(V_{LOD,\varphi}\right)\right) \leq \max\left(LR(v_{max})\right) + 1$ and

$log_{10}\left(LR\left(V_{LOD,\varphi}\right)\right) \geq \max\left(LR(v_{max})\right) - 1$

Preliminary simulation results shown below in **Figure 25** yield high coverage for LOD-support intervals based on one or two LOD difference from a top SNP when two or three ancestral populations are present in a linear regression. For these simulations, we simulate local ancestry for a 100cM region for 5,000 admixed individuals at admixture proportions of 50-50 and 50-40-10, respectively, creating phenotypes drawn from a N(XB,1) for B$\in$ [0.001,1], recording interval length and coverage of the true randomly chosen causal locus across 1,000 replicates. The results shown are for effect sizes that, on average, exhibited p-values that were large enough to be considered genome-wide significant in an admixture mapping scan.

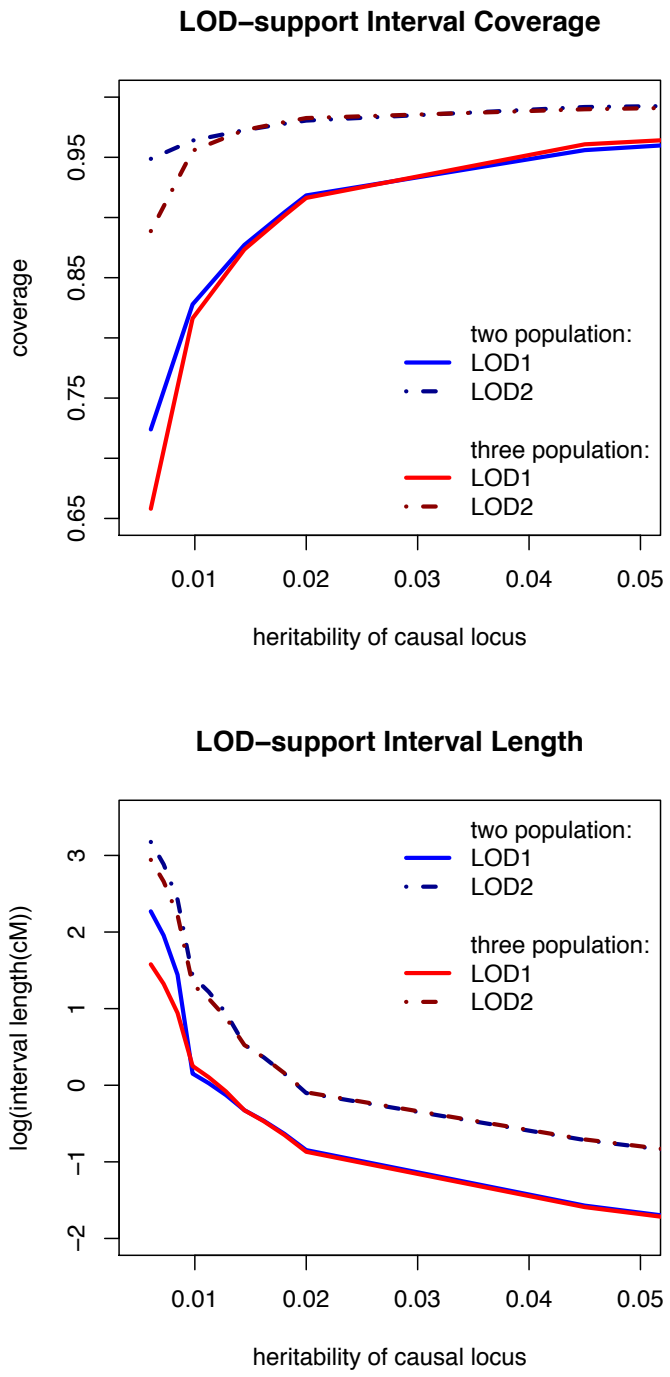


Figure 25: LOD-support Interval Coverage and Length
Panels showing the LOD-support interval coverage (upper panel) and interval length (lower panel) for LOD1 and LOD2 intervals across replicates of simulated phenotypes for simulated populations with two and three ancestral populations at each choice of h^2_{causal} for the causal SNP of interest.

This method for constructing LOD-support intervals is straightforward in the case of regression because the likelihood can be easily calculated at each SNP position. Extending this in the mixed model framework is an open and interesting problem that should be addressed for future admixture mapping studies.

Appendix

Mathematical Derivations

Consider an admixed population that was formed g generations ago w/ admixture fractions π and $(1-\pi)$. For loci r and s separated by a recombination fraction θ, let

$$X_j = \begin{cases} 0 & \text{gamete at locus } j \text{ from pop. } 2 \\ 1 & \text{gamete at locus } j \text{ from pop. } 1 \end{cases}, \text{ for } j \in \{r, s\}$$

Then

$$\begin{aligned} Corr(X_r, X_s) &= \frac{Cov(X_r, X_s)}{(Var(X_s)Var(X_r))^{1/2}} \\ &= \frac{Pr(X_r = X_s = 1) - Pr(X_r = 1)Pr(X_s = 1)}{(\pi(1-\pi)\pi(1-\pi))^{1/2}} \\ &= \frac{\pi(1-\theta)^g - \pi^2(1-\theta)^g}{\pi(1-\pi)} \\ &= (1-\theta)^g \end{aligned}$$

The correlation between local ancestry haplotypes can be derived in a more straight forward manner as:

$$\begin{aligned} Corr(X_r, X_s) &= Pr(\text{same ancestry at positions } r \ \& \ s) \\ &= Pr(\text{no recombination}) + Pr(\text{recombining back to original ancestry}) \\ &= (1-\theta)^g + [1-(1-\theta)^g]p_{\text{ancestry}} \end{aligned}$$

where p_{ancestry} is the admixture proportion of that ancestry in the admixed population. The second quantity in the formula will likely be negligible in practice.

For maternal and paternal haplotypes, m and p, respectively, let

$$a_j = X_{j,p} + X_{j,m}$$

Using the known admixture fractions, we can construct a Wald statistic to test for an association between a_j and Y:

$$\begin{aligned} Z_j &= \frac{\hat{\beta}}{SD(\hat{\beta})} \\ &= \frac{Cov(a_j, Y)/Var(a_j)}{(Var(a_j)^{-1}Var(Y))^{1/2}} \\ &= \frac{E[(a_j - E[a_j])(Y - E[Y])]}{(Var(a_j)Var(Y))^{1/2}} \\ &= \frac{\sum_{i=1}^{N}(a_{ij} - 2\pi)(Y_i - \bar{Y})}{(2\pi(1-\pi)\hat{\sigma}_Y^2)^{1/2}} \end{aligned}$$

The correlation between statistics Z_r and Z_s is given by

$$\begin{aligned} Corr(Z_r, Z_s) &= Corr(a_r, a_s) \\ &= \frac{Cov(a_r, a_s)}{(Var(Z_s)Var(Z_r))^{1/2}} \\ &= \frac{Cov(X_{r,p}, X_{s,p}) + Cov(X_{r,m}, X_{s,m})}{2\pi(1-\pi)} \\ &= (1-\theta)^g \end{aligned}$$

Therefore,

$$Corr(Z_r, Z_s) = Corr(X_r, X_s) = (1-\theta)^g \approx e^{-\theta g} \approx e^{-0.01\Delta g}$$

assuming $\theta \approx 0.01\Delta$ where Δ is the cM distance between locus r and s. Then $Corr(Z_r, Z_s)$ follows an Ornstein-Uhlenbeck process with parameter $0.01g$ and we can simulate from this directly to determine the behavior of admixture mapping test statistics within a chromosome.

For a multivariate extension to this, suppose we have a target population that was formed g generations ago by the blending of K ancestral populations at a proportions $\pi_1, \pi_2, ..., \pi_K$. Let $\tilde{\mathbf{a}}_\mathbf{k} = (a_{1k}, a_{2k}, ..., a_{Nk}) - 2\pi_k$ represent the vector of normalized ancestry value counts for the kth population for the N admixed subjects. The regression coefficient is given by

$$\hat{\boldsymbol{\beta}} = \left(\begin{pmatrix} \mathbf{1}^T \\ \tilde{\mathbf{a}}_\mathbf{1}^T \\ \tilde{\mathbf{a}}_\mathbf{2}^T \\ \vdots \\ \tilde{\mathbf{a}}_{\mathbf{K-1}}^T \end{pmatrix} (\mathbf{1}, \tilde{\mathbf{a}}_\mathbf{1}, \tilde{\mathbf{a}}_\mathbf{2}, ...\tilde{\mathbf{a}}_{\mathbf{K-1}}) \right)^{-1} \begin{pmatrix} \mathbf{1}^T \\ \tilde{\mathbf{a}}_\mathbf{1}^T \\ \tilde{\mathbf{a}}_\mathbf{2}^T \\ \vdots \\ \tilde{\mathbf{a}}_{\mathbf{K-1}}^T \end{pmatrix} \mathbf{Y}$$

Let $\hat{\boldsymbol{\beta}}^*$ be the $K-1$ entries of $\hat{\boldsymbol{\beta}}$ corresponding to The Wald statistic for testing for an ancestry effect is

$$\hat{\boldsymbol{\beta}^*}^T \hat{Var}(\hat{\boldsymbol{\beta}^*})\hat{\boldsymbol{\beta}^*}$$

where $\hat{\boldsymbol{\beta}}^*$ is the final $K-1$ entries of $\hat{\boldsymbol{\beta}}$ and

$$\hat{Var}(\hat{\boldsymbol{\beta}^*}) = \left(\begin{pmatrix} \mathbf{1}^T \\ \tilde{\mathbf{a}}_\mathbf{1}^T \\ \tilde{\mathbf{a}}_\mathbf{2}^T \\ \vdots \\ \tilde{\mathbf{a}}_{\mathbf{K-1}}^T \end{pmatrix} (\mathbf{1}, \tilde{\mathbf{a}}_\mathbf{1}, \tilde{\mathbf{a}}_\mathbf{2}, ...\tilde{\mathbf{a}}_{\mathbf{K-1}}) \right)^{-1}_{(2:K)\times(2:K)} \hat{\sigma}_Y^2$$

The correlation of the regression-based statistics above will not be identical to the correlation between the test statistics in the proposed mixed model framework used in practice. However, ttreatment of the correlation shown here can be used as a good surrogate for what we can expect in real applications of mixed models for admixture mapping. The theoretical calculation of the correlation of test statistics based on the mixed model framework is outside the scope of this project and will not be considered.

BIBLIOGRAPHY

1. Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I., Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Review Genetics 11*(5), 356-366.
2. International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature 467*, 52-58.
3. Hsieh, M.M., Everhart, J.E., Byrd-Holt, D.D., Tisdale, J.F., Rodgers, G.P. (2007). Prevalence of neutropenia in the U.S. population: age, sex, smoking status, and ethnic differences. *Annals of Internal Medicine 146*(7), 486-492.
4. Lim, E.M., Cembrowski, G., Cembrowski, M., Clarke, G. Race-specific WBC and neutrophil count reference intervals. *International Journal of Laboratory Hematology 32*(6 Pt 2), 590-597.
5. Haddy, T.B., Rana, S.R., and Castro, O. Benign ethnic neutropenia: what is a normal absolute neutrophil count? *J Lab Clin Med. 133* (1999),15–22.
6. Hsieh, M.M., Everhart, J.E., Byrd-Holt, D.D., Tisdale, J.F., Rodgers, G.P. Prevalence of neutropenia in the U.S. population: age, sex, smoking status, and ethnic differences. *Annals of Internal Medicine 146* (2007), 486-492.
7. Whitfield, J.B. and Martin, N.G. Genetic and environmental influences on the size and number of cells in the blood. *Genet Epidemiol. 2* (1985), 133–144.
8. Evans DM, Frazer IH, Martin NG. Genetic and environmental causes of variation in basal levels of blood cells. *Twin Res. 2* (1999), 250–257.
9. Pilia G, Chen WM, Scuteri A, Orru M, Albai G, et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet. 2* (2006), e132. doi: 10.1371/journal.pgen.0020132.
10. Crosslin, D.R., McDavid, A., Weston, N., Nelson, S.C., Zheng, X., et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE *Network. Hum Genet 131* (2012), 639-652.
11. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., et al. Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). *PLoS Genetics 7* (2011), e1002108.
12. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., Chen, W.M. Robust relationship inference in genome-wide association studies. *Bioinformatics 26*, 22 (2010), 2867-2873.
13. Schick, U., Jain, C., Hodonsky, C., Morrison, J. V., Davis, J. P. Brown, L., Sofer, T., Conomos, M. P., et al. (2016). Genome-wide Association Study of Platelet Count Identifies Ancestry- Specific Loci in Hispanic/Latino Americans. *American Journal of Human Genetics 98* (2), 229−242.
14. Christian Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W. Porcu, E. Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y. et al. New gene functions in megakaryoisis and platelet information. *Nature 480*, 7376 (2011), 201-208.
15. Nalls, M.A., Couper, D.J., Tanaka, T., Van Rooij, F.J., Chen, M.H., Smith, A.V., Toniolo, D., Zakai, N.A., Yang ,Q., Greinacher, A., et al. Multiple loci are associated with white blood cell phenotypes. *PLoS Genet 6* (2011), e1002113.

doi: 10.1371/journal.pgen.1002113.
16. Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., Lichtner, P., Klopp, N., Stephens, J., Watkins, N.A. et al. A genome wide association study identifies three loci. *AJHG 84*, 1 (2009), 66-71.
17. Kamatani, Y., Matsuda, K., Okada, T., Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nat Commun. 7*, 10531 (2016) doi: 10.1038/ncomms10531.
18. Li, J., D. Absher, H. Tang, A. Southwick, A. Casto et al., Worldwide human relationships inferred from genome- wide patterns of variation. *Science 319*, 5866 (2008), 1100-1104.
19. Wall, J., K. Lohmueller, and V. Plagnol, Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol. 26*, 8 (2002), 1823- 1827.
20. Gravel, S., B. Henn, R. Gutenkunst, A. Indap, G. Marth et al. Demographic history and rare allele sharing among human populations. *PNAS 108*, 29 (2011), 11983- 11988.
21. Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland et al. Ancient admixture in human history. *Genetics 192* (2012), 1065-1093.
22. Reich, D., K. Thangaraj, N. Patterson, A. Price, and L. Singh. Reconstructing Indian population history. *Nature 461*, 7263 (2009), 489-494.
23. Pugach, I., R. Matveyev, A. Wollstein, M. Kayser, and M. Stoneking Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol. 12*, 2 (2011), R19.
24. Green, R., J. Krause, A. Briggs, T. Maricic, U. Stenzel et al. A draft sequence of the Neandertal genome. *Science 328*, 5979 (2010), 710-722.
25. Gravel, S. Population genetics models of local ancestry. *Genetics 191* (2012), 607-619.
26. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature 491*(7422), 56-65.
27. Adeyemo, A., Gerry, N., Chen, G, Herbert, A., Doumatey, A., et al. (2009). A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genetics 5*(7), e1000564.
28. Hancock, D.B., Romieu, I., Shi, M., Sienra-Monge, J.J., Wu, H., et al. (2009).Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in mexican children. *PLoS Genetics 5*(8), e1000623. doi:10.1371/ journal.pgen.1000623.
29. Simons, Y.B., Turchin, M.C., Pritchard, J.K., Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature Genetics 46*, 220-224.
30. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., and Bustamante, C.D. (2011). 1000 Genomes Project. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences 108*(29), 11983-11988.

31. Falush, D., Stephens, M., and Pritchard, J.K. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics Society of America 164* (2003), 1567–158.
32. McKeigue, P. M. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *AJHG 63* (1998), 241-251.
33. Qin, Huaizhen, and Xiaofeng Zhu. "Power Comparison of Admixture Mapping and Direct Association Analysis in Genome-Wide Association Studies." *Genetic epidemiology 36.3* (2012): 235-243. Siegmund, D., Yakir, B. (2007). The statistics of gene mapping. New York: Springer.
34. Risch, Neil, and Kathleen Merikangas. "The future of genetic studies of complex human diseases." *Science* 273.5281 (1996): 1516-1517.
35. Siegmund, D., Yakir, B. (2007). The statistics of gene mapping. New York: Springer.
36. Sha Q, Zhang X, Zhu X, Zhang S. Analytical correction for multiple testing in admixture mapping. Hum Hered. 2006; 62:55–63.
37. Shriner D, Adeyemo A, Rotimi CN. Joint ancestry and association testing in admixed individuals. PLoS Comput Biol. 2011b; 7:e1002325.
38. Montana, G. Statistical methods in genetics. *Briefings in Bioinform 7*, 3 (2006), 297-308.
39. Zou, J. Y., Park, D. S., Burchard, E. G., Torgerson, D. G., Pino-Yanes, M., Song, Y. S., Sankararaman, S., Halperin, E., and Zaitlin, N. Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns. *PNAS 112*, 44 (2015), 13621-13626.
40. Sebro, R., Hoffman, T. J., Lange, C., Rogus, J. J., Risch, N. J. Testing for non-random mating: evidence for ancestry-related assortative mating in the Framingham heart study. *Genet. Epi. 34*, 7 (2010), 674-679.
41. Houseworth, C. and Chiswick, B. R. Ethnic intermarriage among immigrants: human capital and assortative mating. *Rev. Econ. Household 9*, (2011), 149-180.
42. Loh, P., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics 193* (2015), 1233-1254.
43. Bull, L.N., Hu, D., Shah, S., Temple, L., Silva, K. et al. Intrahepatic Cholestasis of Pregnancy (ICP) in U.S. Latinas and Chileans: Clinical features, Ancestry Analysis, and Admixture Mapping. *PLoS One 10*, 6 (2015), e0131211.
44. Tandon, A., Chen, C.J., Penman, A., Hancock, H., James, M. et al. African Ancestry Analysis and Admixture Genetic Mapping for Proliferative Diabetic Retinopathy in African Americans. *Invest Ophthalmol Vis Sci. 56*, 6 (2015), 3999-4005.
45. Gomez, F., Wang, L., Abel, H., Zhang, Q., Province, M.A., Borecki, I.B. Admixture mapping of coronary artery calcification in African Americans from the NHLBI family heart study. *BMC Genet. 16*, 42 (2015), doi: 10.1186/s12863-015-0196-x.
46. Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., and McKeigue, P.M. Design and analysis of admixture mapping studies. *AJHG 74* (2004), 965-978.

47. Gomez, F., Wang, L., Abel, H. Zhang, Q., Province, M.A., and Borecki, I.B. Admixture mapping of coronary artery calcification in African Americans from the NHLBI family heart study. *BMC Genet. 16*, 24 (2015) doi: 10.1186/s12863-015-0196-x.
48. Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michele Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics 26*, 22 (2010) 2867-2873.
49. Yang, J. Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Alkes, P. L. (2014). Advantages and pitfalls in the application of mixed model association methods. *Nature Genetics 46*(2), 100-106.
50. Kang, H.M., Sul, J.H., Service, S. K., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics 42*(4), 348-354.
51. Zhou, X. and Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics 44*, 7 (2012), 821-824.
52. Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., Redline, S., Papanicolaou, G.J., Thornton, T.A., Laurie, C.C., Rice, K., Lin, X. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet. 98*, 4 (2016), 653-66. doi: 10.1016/j.ajhg.2016.02.012.
53. Cavalli-Sforza, L. L., and W. F. Bodmer. The genetics of human populations. W. H. Freeman and Company, San Francisco, 2007.
54. Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *AJHG 93* (2013), 278-288.
55. McVean et al. An integrated map of genetic variation from 1,092 human genomes. *Nature 491* (2012), 56-65. doi:10.1038/nature11632.
56. L. L. Cavalli-Sforza. The Human Genome Diversity Project: past, present and future. Nat. Rev. Genet., 6(4):333-340, Apr 2005.
57. Browning, S. R. and Browning, B. L. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *AJHG 81* (2007), 1084-1097.
58. Visscher, Peter M., et al. "Five years of GWAS discovery." *The American Journal of Human Genetics* 90.1 (2012): 7-24.
59. Winkler, C.A., Nelson, G.W., and Smith, M.W. Admixture mapping comes of age. *Ann. Rev. of Genomics and Human Genet. 11* (2010), 65-89.
60. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics 38*, 8 (2006), 904-909.
61. Patterson, N., Price, A. L., Reich, D. Population structure and eigenanalysis. *PLoS Genetics 2*, 12 (2006), e190.
62. Conomos, M.P., Miller, M.B., Thornton, T.A. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology* (2015), doi:10.1002/gepi.21896
63. Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. Reconstructing genetic ancestry blocks in admixed individuals. *AJHG 79* (2006), 1-12.

64. Reiner, A.P., Beleza, S., Franceschini, N., Auer, P.L., Robinson, J.G., et al. Genome-wide Association and Population Genetic Analysis of C-Reactive Protein in African American and Hispanic American Women. *AJHG 91* (2012), 502-512.
65. Wald, A. (1949) Statistical decision functions. *Annals of Mathematical Statistics 20*(2), 165-205.
66. Hays, J., Hunt, J.R., Hubbell, F.A., Anderson, G.L., Limacher, M., Allen, C., and Rossouw, J.E. The Women's Health Initiative recruitment methods and results. *Annals of Epi. 13*, 9 Suppl. (2003), S18-S77.
67. Sorlie, P.D., Aviles-Santa, L.M., Wassertheil-Smoller, S., Kaplan, R.C., Daviglus, M.L., Giachello, A.L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology 20* (2010), 629-641.
68. Lavange, L.M., Kalsbeek, W.D., Sorlie, P.D., Aviles-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology 20* (2010), 642-649.
69. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., Mountain, J. L. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *AJHG. 96*, 1 (2015), 37–53. doi: 10.1016/j.ajhg.2014.11.010
70. McKeigue, P. M., Carpenter, J. R., Parra, E. J., Shriver, M. D. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann. Hum. Genet. 64* (2000), 171-186.
71. Reiner, A. P., Ziv, E. Link, D. L., Neivergelt, C. M., Shork, N. J., Cummings, S. J., Phong, A., Buchard E. G., Harris, T. B. Psaty, B. M., Kwok, P. Population Structure, Admixture, and Aging-Related Phenotypes in African American Adults: The Cardiovascular Health Study. *AJHG 76*, 3 (2005), 463-477.
72. Bryc, K. Auton, A., Nelson, M. R. Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J., Wambebe, C., Tishkoff, S., Bustamante, C. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *PNAS 107*, 2 (2009), 786–791.
73. Skelly, D.A., Magwene, P.M., and Stone, E.A. Sporadic, Global Linkage Disequilibrium Between Unlinked Segregating Sites. *Genetics 202*, 2 (2016), 427-437.
74. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. PLINK: A tool set for whole- genome association and population-based linkage analyses. *AJHG 81* (2007), 559-575.

75. Gerstein, HC, Mann, JF, Yi, Q, Zinman, B, Dinneen, SF, et al. Albuminuria and risk of cardiovascular events, death, and heart failure in diabetic and nondiabetic individuals. *Jama,* 286**:** 421-426, 2001.
76. Astor, BC, Matsushita, K, Gansevoort, RT, van der Velde, M, Woodward, M, et al. Lower estimated glomerular filtration rate and higher albuminuria are associated with mortality and end-stage renal disease. A collaborative meta-analysis of kidney disease population cohorts. *Kidney Int,* 79**:** 1331-1340, 2011.

77. Mattix, HJ, Hsu, CY, Shaykevich, S, Curhan, G: Use of the albumin/creatinine ratio to detect microalbuminuria: implications of sex and race. *J Am Soc Nephrol,* 13**:** 1034-1039, 2002.
78. Robbins, DC, Knowler, WC, Lee, ET, Yeh, J, Go, OT, et al. Regional differences in albuminuria among American Indians: an epidemic of renal disease. *Kidney Int,* 49**:** 557-563, 1996.
79. Ricardo, AC, Flessner, MF, Eckfeldt, JH, Eggers, PW, Franceschini, Net al. Prevalence and Correlates of CKD in Hispanics/Latinos in the United States. *Clin J Am Soc Nephrol,* 10**:** 1757-1766, 2015.
80. System, USRD: 2014 USRDS annual data report: Epidemiology of kidney disease in the United States. In: NATIONAL INSTITUTES OF HEALTH, N. I. O. D. A. D. A. K. D. (Ed.) Bethesda, MD, 2014.
81. Manichaikul, A, Palmas, W, Rodriguez, CJ, Peralta, CA, Divers, J, et al. Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. *PLoS Genet,* 8**:** e1002640, 2012.
82. Peralta, CA, Li, Y, Wassel, C, Choudhry, S, Palmas, W, et al. Differences in albuminuria between Hispanics and whites: an evaluation by genetic ancestry and country of origin: the multi-ethnic study of atherosclerosis. *Circ Cardiovasc Genet,* 3**:** 240-247, 2010.
83. Gonzalez Burchard, E, Borrell, LN, Choudhry, S, Naqvi, M, Tsai, HJ, et al. Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *Am J Public Health,* 95**:** 2161-2168, 2005.
84. Boger, CA, Chen, MH, Tin, A, Olden, M, Kottgen, A, et al. CUBN is a gene locus for albuminuria. *J Am Soc Nephrol,* 22**:** 555-570, 2011.
85. Naik, RP, Derebail, VK, Grams, ME, Franceschini, N, Auer, et al. Association of sickle cell trait with chronic kidney disease and albuminuria in African Americans. *Jama,* 312**:** 2115-2125, 2014.
86. Genovese, G, Friedman, DJ, Ross, MD, Lecordier, L, Uzureau, P, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science,* 329**:** 841-845, 2010.
87. Schick, UM, Jain, D, Hodonsky, CJ, Morrison, JV, Davis, JP, et al. Genome-wide Association Study of Platelet Count Identifies Ancestry-Specific Loci in Hispanic/Latino Americans. *Am J Hum Genet,* 98**:** 229-242, 2016.
88. Kopp, JB, Smith, MW, Nelson, GW, Johnson, RC, Freedman, BI, et al. MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat Genet,* 40**:** 1175-1184, 2008.
89. Ward, LD, Kellis, M: HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res,* 40**:** D930-934, 2012.
90. Teumer, A, Tin, A, Sorice, R, Gorski, M, Yeo, NC, et al. Genome-wide Association Studies Identify Genetic Loci Associated With Albuminuria in Diabetes. *Diabetes,* 65**:** 803-817, 2016.
91. Ho, J, Pandey, P, Schatton, T, Sims-Lucas, S, Khalid, M, et al. The pro-apoptotic protein Bim is a microRNA target in kidney progenitors. *J Am Soc Nephrol,* 22**:** 1053-1063, 2011.

92. Bouillet, P, Metcalf, D, Huang, DC, Tarlinton, DM, Kay, TW, Kontgen, F, Adams, JM, Strasser, A: Proapoptotic Bcl-2 relative Bim required for certain apoptotic responses, leukocyte homeostasis, and to preclude autoimmunity. *Science,* 286**:** 1735-1738, 1999.
93. Chuang, PY, Dai, Y, Liu, R, He, H, Kretzler, at al. Alteration of forkhead box O (foxo4) acetylation mediates apoptosis of podocytes in diabetes mellitus. *PLoS One,* 6**:** e23566, 2011.
94. Sherva, Richard, et al. "Genome-wide association study of the rate of cognitive decline in Alzheimer's disease." *Alzheimer's & Dementia* 10.1 (2014): 45-52.
95. Zhao, Yongzhong, et al. "Molecular and genetic inflammation networks in major human diseases." *Molecular BioSystems* 12.8 (2016): 2318-2341.
96. Rodriguez-Murillo, Laura, et al. "Fine mapping on chromosome 13q32–34 and brain expression analysis implicates MYO16 in schizophrenia." *Neuropsychopharmacology* 39.4 (2014): 934-943.