Statistical Inference in Admixed Populations

Kelsey Grinde

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Sharon Browning, Chair

Timothy Thornton

Kelley Harris

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Inference in Admixed Populations

Kelsey Grinde

Chair of the Supervisory Committee:
Sharon Browning
Department of Biostatistics

Understanding the genetic causes of human diseases and traits has long been of interest in the scientific community. However, the large majority of research in this area has been conducted in European populations. This dissertation focuses on developing statistical methods for genetic studies in admixed populations, such as African Americans and Hispanics/Latinos, that have been historically underrepresented in genetics research. The diverse, mixed ancestry of admixed populations presents unique opportunities for statistical inference, many of which are explored in this work. Here, we focus in particular on two important tasks: inferring genetic ancestry from genotype and sequence data, and identifying genetic variants associated with complex traits and diseases. We propose and evaluate methods for inferring local ancestry on chromosome X, correcting for multiple testing in genome-wide admixture mapping studies, and controlling for confounding by global ancestry in admixture mapping and genome-wide association studies in admixed populations. We motivate our proposed methods with theoretical results, simulation studies, and applications to genotype and whole genome sequence data from large studies of African American and Hispanic/Latino individuals. Our work provides solutions to a number of the statistical challenges posed by genetic studies in admixed populations, and we hope that our results will help guide future studies in these populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I feel extremely lucky to have spent the last five years working with and learning from some of the best and brightest in the field of biostatistics. First and foremost, I am forever thankful for my advisor, Sharon Browning: the first faculty member that I met at UW and a person who, since that day, has been an incredible mentor, teacher, and source of inspiration. I am also grateful to the rest of my committee—Tim Thornton, Kelley Harris, Ellen Wijsman, Amy Willis, and Erick Matsen—and the members of the Browning Lab, past and present, for all the time, feedback, and mentorship that they have contributed over the years. Many other faculty and staff have provided support and guidance throughout my time at UW, modeling how to be a good statistician, researcher, teacher, communicator, collaborator, and citizen; the list is too long to name every name. Finally, a huge "thank you!" to "the greatest cohort ever" and the other friends I've made along the way: the last five years have been a whole lot more enjoyable because of you, and I think perhaps I've learned as much from you as anyone else. Although I won't miss T-wing, I definitely will miss all of you!

Throughout my time at UW, I have been fortunate to receive funding from a number of sources, including the Department of Biostatistics, Browning Statistical Genetics Lab, UW Genetic Analysis Center, ARCS Foundation Seattle Chapter, UW NIGMS sponsored Statistical Genetics Training Grant (Grant No. NIGMS T32GM081062), and National Science Foundation Graduate Research Fellowship Program (Grant No. DGE-1256082). I am incredibly grateful for their support over these past five years. I am also grateful to the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), Women's Health Initiative (WHI), and NHLBI Trans-Omics for Precision Medicine Project (TOPMed) for providing

access to the data analyzed in this dissertation. In particular, I would like to thank the investigators, staff and participants of these studies for their contributions and for making my own work possible. Please note that any opinions, findings, ideas, interpretations, conclusions, or recommendations contained in this dissertation are mine and do not necessarily reflect the views of these studies, investigators, or funding sources.

Finally, I am thankful for all the people who helped me get here in the first place, and who made the last five years more manageable. My undergraduate professors and research advisors—especially Paul, Katie, and Nathan—inspired me to pursue a career in biostatistics and equipped me with the skills and knowledge that made the transition to graduate school far easier than it would have been otherwise. My family and friends have wholeheartedly supported my love of math (even if they haven't always understood it) and inspired me to find a way to use math for good; they have modeled kindness, hard work, and intellectual curiosity, and I am incredibly thankful for their (blind) faith and encouragement as I decided to pursue a PhD in a a field that none of us had ever heard of. Phone calls and visits from my parents, my sister, and other family and friends have been instrumental in keeping me sane, providing a sounding board as I navigated the ups and downs of this program, and reminding me of the light at the end of the grad school tunnel. Last, but definitely not least, I am eternally grateful for the support of my best friend and partner, Zack Meyer. Thank you for moving across the country to a city you had never visited; for five years of way more than your share of the cooking, grocery shopping, and cleaning; for encouraging me to slow down and take breaks when that was what I needed; and for generally making my life more balanced, calm, and fun. I am so lucky to have you in my life.

DEDICATION

to my family, for a lifetime of support and inspiration

Chapter 1
INTRODUCTION

1.1 Statistical Inference in Admixed Populations

Understanding the genetic causes of human diseases and traits has long been of interest in the scientific community. However, most of the research in this area has been conducted in populations of European descent. In 2009, an analysis by Need and Goldstein [1] found that 96% of participants in genome-wide association studies (GWAS)—one of the most widely employed approaches for looking for genetic variants associated with diseases/traits of interest—were of European descent. In recent years, there has been a growing international effort to increase the diversity of genetic studies, with modest improvements. In 2016, Popejoy and Fullerton [2] found that the proportion of non-European GWAS participants had risen to 19%. The majority of the improvements from 2009 to 2016 were seen with respect to representation of individuals of Asian ancestry, increasing from 3% of GWAS participants in 2009 to 14% in 2016. However, fewer than 4% of GWAS participants were from African, Hispanic/Latino, native/indigenous, or mixed ancestral backgrounds as of 2016, and when this analysis was repeated in 2019 [3], that proportion remained unchanged. Genetic studies in these historically underrepresented populations are imperative, not only to ensure that genomic medicine benefits more than just "a privileged few" [4], but to ensure a broader understanding of which genetic variants play a role in human disease and traits, including those variants that are more frequent or even exclusively present in non-European populations [5].

Populations with mixed ancestry, known as *admixed populations*, are historically underrepresented in genetic studies, yet their mixed and diverse ancestry presents unique opportunities for detecting genetic variants associated with complex traits and diseases. Admixed

populations are formed when two or more previously separated ancestral populations come together and form a new population with mixed genetic ancestry. Examples of admixed populations in the United States include African Americans and Hispanics/Latinos. The mixed ancestry of admixed populations leads to increased variability of genetic material and provides an opportunity to find genetic variants that would not be identifiable otherwise.

Due to the processes involved in the inheritance of genetic material, the genomes of individuals in admixed populations are a mosaic of segments with different ancestral origins (Figure 1.1). We refer to the ancestral origin of each of these segments as *local ancestry*, while *global ancestry* quantifies the overall proportion of genetic material inherited from each ancestral population. In an admixed population with K ancestral populations, we characterize local ancestry for individual i at locus j via the vector $\mathbf{a}_{ij} = \begin{pmatrix} a_{ij1} & \cdots & a_{ijK} \end{pmatrix}^\top$, where $\sum_{k=1}^{K} a_{ijk} = 2$ and a_{ijk} denotes the number of alleles inherited by individual i from ancestral population k at locus j. We characterize global ancestry via the vector of *admixture proportions* $\boldsymbol{\pi}_i = \begin{pmatrix} \pi_{i1} & \cdots & \pi_{iK} \end{pmatrix}^\top$, where $\sum_{i=1}^{K} \pi_{ik} = 1$ and π_{ik} is the genome-wide proportion of genetic material inherited by individual i from population k (i.e., $\pi_{ik} = \frac{1}{2m} \sum_{j=1}^{m} a_{ijk}$, where m is the total number of loci across the genome). Note that these definitions of local and global ancestry are restricted to consideration of the autosomes (chromosomes 1–22 in humans) and do not apply to the X or Y chromosomes.

Local and global ancestry are not features that can be directly observed; instead, they are inferred from genotype or sequence data. Various methods have been developed to infer global [6, 7, 8] and local [9, 10, 11] ancestry from genetic data, and companies such as Ancestry.com and 23andMe now offer direct-to-consumer ancestry inference, converting a saliva sample into an ancestry report in a matter of weeks. Many of these methods are *supervised*, requiring reference panels (i.e., genotype or sequence data for samples from relevant ancestral populations) in order to perform ancestry inference. Methods for local and global ancestry inference are described in further detail in Chapters 2 and 4, respectively.

Besides holding intrinsic scientific interest, ancestry inference in admixed populations is also useful for a number of applications, including mapping the genetic causes of diseases,

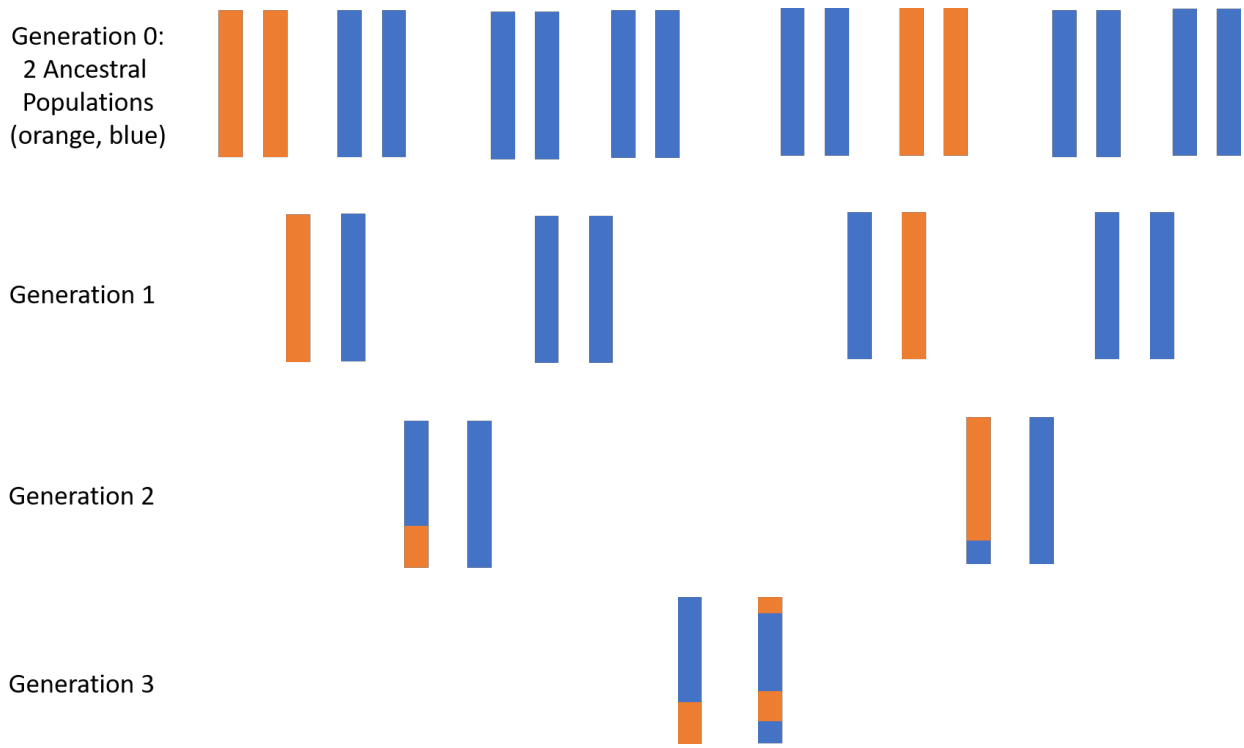

Figure 1.1: Inheritance of genetic material in an admixed population with two ancestral populations.
Each individual possesses a pair of chromosomes, one inherited from their mother and one from their father. In this figure, we track the ancestral origin of genetic material across a single chromosome by "painting" the chromosomes orange and blue according to their ancestral origin (orange = Ancestral Population 1, blue = Ancestral Population 2). In Generation 0, two previously separated *ancestral populations* come together and start mixing. In successive generations, genetic material is passed from parents to offspring, with *crossover* events resulting in inherited chromosomes with a mixture of genetic material from the maternal and paternal chromosomes, and, as a result, a mixture of ancestral origins. Over time these blocks of ancestry continue to be broken up by crossover events. The goal of *local ancestry inference* is to recover the chromosome painting, or infer the ancestral origin of each segment of the chromosome. *Global ancestry inference*, on the other hand, seeks to quantify the overall proportion of ancestry from each source. *Admixture mapping* looks for genetic variants associated with a disease or trait of interest by scanning the genome for associations with local ancestry, while *GWAS* looks for associations with the underlying genotypes (not depicted here).

adjusting for population structure in association testing, and making demographic inferences. Admixture mapping in populations with mixed ancestry is a powerful approach for identifying genetic variants associated with complex traits and diseases [12, 13, 14], and although the concepts behind admixture mapping have been around for many years, it has not been implemented or studied to the same extent as other approaches such as genome-wide association studies (GWAS) [15, 16, 17]. Like GWAS, the goal of admixture mapping is to determine whether any genetic variant is associated with a trait of interest. However, admixture mapping differs from GWAS in the approach it takes to find those causal variants, scanning the genome for associations between the trait and local ancestry, rather than between the trait and the genotype at each locus. We often observe differences among ancestral groups in disease prevalence and trait values—including asthma [18], prostate cancer [19], blood pressure [20], and kidney disease [21, 22]—which could result from a combination of genetic and environmental causes. By looking for associations between a trait and local ancestry, admixture mapping studies seek to identify the genetic variants that differ in frequency across ancestral groups and drive these observed phenotypic differences.

Early methods for admixture mapping were based on a small number of ancestry-informative markers [23, 14], but thanks to technological advances and growing international effort to increase the diversity of genetic studies, genome-wide admixture mapping studies are becoming more popular. Now, a common approach to admixture mapping utilizes marginal regression, regressing the trait of interest on inferred local ancestry at each measured locus across the genome [24, 25]. To perform admixture mapping, we regress the trait, $\mathbf{y}$, on each component ($k = 1, \ldots, K$) of the local ancestry vector at each locus ($j = 1, ..., m$), using the marginal regression model

$$E\left[y_i \mid a_{ijk}, \mathbf{w}_i\right] = \alpha + \beta_{jk} a_{ijk} + \boldsymbol{\gamma} \mathbf{w}_i, \tag{1.1}$$

where $\mathbf{w}_i$ is a vector of additional covariates. We test for association between the trait and local ancestry by testing the null hypotheses $H_0 : \beta_{jk} = 0 \ \forall j, k$. This is very similar to the

marginal regression framework used in GWAS:

$$E\left[y_i \mid a_{ijk}, \mathbf{w}_i\right] = \alpha + \beta_j g_{ij} + \boldsymbol{\gamma}\mathbf{w}_i, \tag{1.2}$$

where the trait is regressed instead on g_{ij}, the genotype for individual i at locus j, which is typically quantified as the number of copies of some pre-specified allele (e.g., the minor allele) carried by individual i at locus j. Admixture mapping and GWAS are complementary techniques, and each can be more powerful in different situations [26, 27, 28]. In many studies, both admixture mapping and GWAS are performed (e.g., [29, 30]); others have suggested methods for combining the two [31, 32].

A typical genetic association study, whether GWAS or admixture mapping, involves conducting hundred of thousands or millions of hypothesis tests. Multiple testing correction procedures are needed to control the overall type I error rate of these association studies. The p-value threshold 5×10^{-8} has become quite widely accepted as a control for multiple testing in GWAS [33, 34]—although recent work has suggested that more stringent thresholds should be used for association studies in whole genome sequence data [35]—but there is no such "established" significance threshold for admixture mapping studies. The correlation structure of test statistics is considerably different in admixture mapping than in GWAS. In particular, admixture mapping test statistics tend to be more highly correlated, implying that a less stringent significance threshold will be required to control for multiple testing in admixture mapping studies. However, modeling the complex correlation structure of admixture mapping test statistics and incorporating this knowledge into multiple testing correction procedures is a challenging task.

An additional challenge posed by genetic association studies in admixed populations relates to controlling for population structure. It has been widely documented that *population structure*, or heterogeneity in ancestral composition across a population of interest, can induce spurious associations in GWAS [36, 37]. In response, a number of approaches have been developed to control for population structure in genetic association studies. Among the most widely-implemented approaches are those that infer global ancestry and then adjust

for it as a potential confounder in association studies. There are two methods for inferring global ancestry that are often used. One option is to estimate admixture proportions $\hat{\boldsymbol{\pi}}_i$ using global ancestry inference programs (e.g., [6, 7]) or, if local ancestry was inferred, using the genome-wide average local ancestry (e.g., [38]). The other, perhaps more widely-implemented, approach involves running principal component analysis (PCA) on sample genotypes and including the first few principal components (PCs) in the regression model [37]. It has been shown that the top PCs typically capture global ancestry [39, 40], but choosing the number of PCs to include in a regression model and ensuring that the PCs actually capture global ancestry can be challenging. The consequences of adjusting for extraneous PCs or PCs that capture effects other than ancestry are not fully understood.

1.2 Dissertation Aims

In this dissertation, we aim to address the following open research questions related to statistical inference in admixed populations:

1. What are the best practices for inferring local ancestry on chromosome X, particularly when using supervised local ancestry inference programs such as `RFMix` [10] that require diploid individuals?

2. How can we estimate study-specific significance thresholds for admixture mapping studies that appropriately account for the correlation structure of test statistics in admixed populations with potentially complex population structure?

3. When do we need to control for population structure, or heterogeneity in global ancestry across a population, in genetic association studies (admixture mapping and GWAS) in admixed populations?

4. How should we control for population structure in genetic association studies in admixed populations? Is it better to adjust for estimated admixture proportions from model-based approaches or principal components in our regression models?

The first question is addressed in Chapter 2, the second is addressed in Chapter 3, and Chapter 4 addresses the remaining two questions. Finally, in Chapter 5 we apply the lessons learned from earlier chapters to a large-scale analysis of whole genome sequence data for African American and Hispanic/Latino subjects from the Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project. The following sections provide brief outlines of each of these chapters.

1.2.1 Chapter 2: Local ancestry inference on chromosome X

Over the past several years, a number of methods for inferring local ancestry from genotype data have been proposed [9, 10, 11]. `RFMix` [10] is particularly widely used for local ancestry inference (LAI) in admixed populations, such as Hispanics/Latinos, that have three or more ancestral populations. However, like many genetics software tools, `RFMix` seems to have been designed only for inference on the autosomes (chromosomes 1–22). Techniques for applying the program to chromosome X are not discussed in the original paper or in the user manual.

The X chromosome differs from the autosomes in that males have only a single copy of X, while females have two. This poses challenges for programs such as `RFMix` which require that all individuals be diploid. Naively coding male haplotypes as homozygous diploid on X (i.e., duplicating their X chromosome so they have two identical copies of the chromosome), as is often suggested for other analyses on chromosome X [41, 42, 43, 44], could pose problems for LAI, especially for supervised programs such as `RFMix` that rely on reference panels for model training. In particular, this homozygous diploid coding will make observed male haplotypes in the reference panel look twice as frequent as they really are, which could impact model training and, as a result, LAI accuracy.

Further complicating the application of `RFMix` to chromosome X is the fact that it requires that haplotypes be phased. Since males have only a single copy of X, they are already perfectly phased. However, females are diploid on X and their haplotypes will need to be phased using statistical methods (e.g., [41]). Existing techniques for statistically inferring phase are largely accurate, but not perfect, and any phasing errors could pose problems for

LAI. To account for possible phase uncertainty, `RFMix` can perform re-phasing of admixed haplotypes along with inferring local ancestry. On chromosome X, this re-phasing will likely be helpful for female haplotypes, but unnecessary for males.

In this chapter, we propose and evaluate four techniques for applying `RFMix` to chromosome X. These techniques consider different choices of coding schemes for admixed and reference panel male haplotypes, as well as whether to allow `RFMix` to perform re-phasing of admixed haplotypes. We compare the performance of these approaches in a large cohort of Hispanic/Latino individuals from the Hispanic Community Health Study/Study of Latinos [45, 46]. Based on our findings, we provide suggestions regarding which coding schemes and phasing options should be used when inferring local ancestry on chromosome X using `RFMix`. This work will prove useful for a wide range of downstream applications, including studying sex-biased patterns of admixture [47, 48] and conducting genetic association studies on chromosome X [44, 49].

This work is published, along with other HCHS/SOL local ancestry results, in *G3: Genes, Genomes, Genetics* [50]. If you wish to cite the work contained in this chapter, please cite our published paper rather than this dissertation.

1.2.2 Chapter 3: Genome-wide significance thresholds for admixture mapping

Admixture mapping studies have become more widely implemented in recent years, due in part to technological advances and growing international efforts to increase the diversity of genetic studies. However, many open questions remain about appropriate implementation of admixture mapping studies, including how best to control for multiple testing, particularly in the presence of population structure.

A handful of existing multiple testing correction techniques have been applied to the context of admixture mapping studies. Perhaps the best known approach is a Bonferroni correction on the total number of tests conducted ($m \times K$). Although easy to implement (and widely used), this approach does not take into account the complex correlation structure of admixture mapping studies and consequently yields conservative significance thresholds.

Related approaches involve a Bonferroni correction on the estimated effective number of *independent* tests [51, 52, 32]; however, it has been shown that these approaches do not always guarantee family-wise error rate control [53, 54, 55]. Permutation and simulation-based multiple testing correction procedures [31, 28, 30, 26], which are often considered to be the gold standard for controlling for multiple testing in genetic association studies, can yield more appropriate significance thresholds but are often very computationally intensive. In practice, many admixture mapping studies will avoid the issue of estimating a study-specific significance threshold altogether and simply use a threshold proposed in another study, even if that study considered a different admixed population [29, 56, 57].

In this chapter, we develop an analytic approach for controlling for multiple testing in admixed populations with any number of ancestral populations or distribution of admixture proportions. First, we derive the correlation of local ancestry and admixture mapping test statistics in admixed populations with arbitrary population structure. Based on this theoretical framework, we develop two approaches for estimating genome-wide significance thresholds for admixture mapping studies. In the case of admixed populations with two ancestral populations, we show that an analytic approximation to the family-wise error rate proposed by Siegmund and Yakir [58] can be used to estimate admixture mapping significance thresholds, even in the presence of heterogeneous admixture proportions, as long as the admixture mapping regression model adjusts for admixture proportions as a potential confounder (i.e., by including estimates $\hat{\boldsymbol{\pi}}_i$ in the vector of covariates $\mathbf{w}_i$ in Model 1.1). For admixed populations with more than two ancestral populations, we propose an approach based on simulating admixture mapping test statistics from their derived asymptotic joint distribution using a fast, recursive, low memory algorithm. We implement both approaches in an R package named `STEAM`: Significance Threshold Estimation for Admixture Mapping.

We validate our theoretical work and multiple testing correction procedures via analysis of simulated traits with real genotype data for 8,064 unrelated African American and 3,425 Hispanic/Latina women from the Women's Health Initiative SNP Health Association Resource (WHI SHARe). Compared to other commonly used multiple testing correction

procedures, our method is fast, easy to implement, and controls the family-wise error rate even in structured populations. Furthermore, our work demonstrates that the appropriate admixture mapping significance threshold depends on the number of ancestral populations, generations since admixture, and population structure of the sample; as a result, significance thresholds are not, in general, transferable across studies.

This work is published in *The American Journal of Human Genetics* [59] and our R package `STEAM` is publicly available on my GitHub page: https://github.com/kegrinde/STEAM. If you wish to cite the work contained in this chapter, please cite our published paper rather than this dissertation.

1.2.3 Chapter 4: Controlling for ancestral heterogeneity in genetic association studies in admixed populations

It is well-established that differences in global ancestry within a sample, often referred to as *population structure*, can lead to spurious associations in genome-wide association studies [36, 37]. It is generally understood that these spurious associations arise due to the fact that global ancestry can confound the association between genotypes and a phenotype of interest, particularly when single nucleotide variants (SNVs) have different allele frequencies across ancestral populations and global ancestry has a direct effect on the trait through environmental differences between ancestral groups. The issue of population structure has not been studied as extensively in the context of admixture mapping, and it is unclear whether the same approaches that we use to detect and control for population structure in GWAS can also be applied to admixture mapping. Some studies have suggested the importance of adjusting for global ancestry in admixture mapping studies [13, 24]; however, these suggestions are not universally implemented.

A number of approaches for detecting and controlling for population structure in genetic association studies have been proposed. Early approaches included genomic control [36], using family-based study designs [60], and restricting—or attempting to restrict—analyses to ancestrally homogeneous populations [61]. Although these approaches are still implemented

today, newer approaches are now generally preferred: using mixed models to control for both distant and close relationships [62, 63, 64] or adjusting for inferred global ancestry as a fixed effect in marginal regression models [65, 37, 24].

Methods for inferring global ancestry fall primarily into two classes: model-based approaches and unsupervised dimension reduction techniques such as principal component analysis (PCA). Model-based approaches such as `ADMIXTURE` [7] model the probability of observed genotypes given admixture proportions and allele frequencies in each ancestral population. Depending on the program, ancestry-specific allele frequencies can be estimated along with admixture proportions, or pre-estimated/trained based on reference panel data. The primary disadvantages of model-based approaches are that they typically perform ancestry inference at a continental level (e.g., African versus European, rather than South European versus North European; see `fineSTRUCTURE` [8] for an exception), the number of ancestral populations of interest (K) must be pre-specified, and reference panels are often required. Inferring global ancestry via principal component analysis, on the other hand, is an unsupervised approach that does not require reference panel data or pre-specification of the number of ancestral populations of interest, and it is capable of capturing sub-continental structure (e.g., [66]). It has been widely shown that, after running PCA on sample genotypes, the top principal components (PCs) typically reflect global ancestry [40, 39]. Although PCA does offer the advantages mentioned above, choosing the number of PCs necessary to capture global ancestry can be very challenging, and the PCs themselves are less interpretable than estimated admixture proportions. Furthermore, PCs can sometimes reflect features other than global ancestry, including relatedness across samples [67, 40], data quality issues [37, 68], and/or small regions of the genome with unusual patterns of linkage disequilibrium (LD) [69, 70].

In this chapter, we investigate the impact of heterogeneity of global ancestry on genome-wide association and admixture mapping studies in admixed populations. First, we present analytic results that provide insight into when genetic association studies must adjust for global ancestry. Second, we demonstrate that adjusting for global ancestry using PCs can

actually *induce* spurious associations in both GWAS and admixture mapping studies. We show that these spurious associations arise, due to the phenomenon known as *collider bias* [71], when PCs capture multiple small regions of the genome rather than global ancestry. We show that careful pre-processing of genotypes prior to running PCA (e.g., LD pruning) can help reduce the occurrence of these spurious associations, and that, at least in the applications we consider, problems can be avoided altogether by adjusting instead for estimated admixture proportions from model-based global ancestry inference approaches. Based on our results, we provide suggestions regarding best practices for appropriately controlling for population structure in genetic association studies in admixed populations. We anticipate that this work will have great impact, particularly given the wide-spread use of principal component analysis in our field.

We plan to submit the work in this chapter as two separate manuscripts, one devoted to genome-wide association studies and the other focused on admixture mapping.

1.2.4 Chapter 5: Local ancestry inference and genetic association testing in the TOPMed Whole Genome Sequencing Project

Reduced kidney function has serious, sometimes deadly, implications for human health. Current reports estimate that as many as 37 million people across the United States struggle with chronic kidney disease [72], with a higher prevalence of the disease among African Americans and Hispanics/Latinos [21, 22]. This difference in prevalence across ancestral groups makes chronic kidney disease an ideal trait for admixture mapping studies. Previous studies [73, 22, 25] have been successful in identifying genetic variants associated with kidney function that differ in frequency across these ancestral groups, and could help explain some of the observed differences in disease prevalence.

In this chapter, we build on this prior work and conduct genetic association studies of kidney traits in admixed populations using data from the Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project. Until recently, the majority of genetic association studies were conducted using genotype data from single nucleotide polymorphism

(SNP) arrays/chips rather than whole genome sequences, due largely to the initially precipitous costs of whole genome sequencing technology. However, in recent years the cost of sequencing studies has dropped drastically, and whole genome sequence data is starting to become more readily accessible. In studying whole genome sequences, we hope to capture important rare or population-specific genetic variation that may have been missed in prior studies based on genotype data.

Using whole genome sequence data for 20,048 African American and Hispanic/Latino subjects from the TOPMed sequencing project, we apply the lessons learned from Chapters 2–4 to search for and characterize rare and common genetic variation that may play a role in human kidney function. First, we infer autosomal local ancestry for each admixed individual using `RFMix` [10]. Next, we use this inferred local ancestry to estimate admixture proportions. Then, we conduct genome-wide admixture mapping and association studies of kidney traits (estimated glomerular filtration rate (eGFR), serum creatinine, and a binary indicator of chronic kidney disease), adjusting for admixture proportions as covariates in our marginal regression models as suggested by Chapter 4. For our admixture mapping studies, we use `STEAM` (Chapter 3) to estimate genome-wide significance thresholds. Finally, we use our local ancestry calls to estimate African, European, and Native American allele frequencies for candidate variants of interest from a concurrent study [Bridget Lin and Nora Franceschini, personal communication]; these ancestry-specific allele frequency estimates provide valuable insight into GWAS findings and guide replication studies.

Our work in this chapter highlights the potential for success, as well as the existing challenges, for local ancestry inference and genetic association studies in admixed populations using whole genome sequence data. Furthermore, the local ancestry calls produced for this analysis provide a valuable resource for the large community of investigators interested in using the TOPMed whole genome sequence data to study other phenotypes.

This work will soon be submitted for publication along with results from an association study of all admixed, European, and Asian American TOPMed samples with available kidney phenotype data.

Chapter 2

LOCAL ANCESTRY INFERENCE ON CHROMOSOME X

2.1 Introduction

Due to the processes involved in the inheritance of genetic material, the genomes of admixed individuals are a mosaic of segments with different ancestral origins (Figure 1.1). This mosaic nature of admixed genomes proves useful for a number of tasks. In particular, local ancestry—the ancestral origin of genetic material at each location along an admixed genome—is used in applications including admixture mapping [12, 13, 14], interpreting the results of genetic association studies [74, 27], adjusting for population structure in genome-wide association studies [75, 76], estimating recombination rates along the genome [77, 78], and making demographic inferences [79, 80, 81]. In order to perform these tasks, we must first infer local ancestry from genotype or sequence data.

Over the past few years, a number of local ancestry inference methods have been proposed. These methods can primarily be divided into two classes. First, there are methods that use classification tools such as random forests and support vector machines to identify the population to which a haplotype is most similar within a given window (Figure 2.1) [10, 82, 76]. Second, there are methods that explicitly model biological processes such as recombination and mutation using Hidden Markov Models or extensions of the Li and Stephens [83] population genetic model [84, 9, 8, 85, 11]. The vast majority of these methods are supervised, requiring reference haplotypes from ancestral populations of interest (e.g, Africans, Europeans, and Native Americans for a Hispanic/Latino population), and only a few are applicable to admixed populations such as Hispanics/Latinos with more than two ancestral populations.

RFMix [10] is a widely-used approach for local ancestry inference in admixed populations

Figure 2.1: Local ancestry inference in an admixed population with two ancestral populations.
The genomes of admixed individuals are split into windows. Within each window, we infer the ancestral origin of an admixed haplotype by comparing the pattern of genetic mutations in that window to the mutations carried by reference haplotypes from the ancestral populations of interest (or close proxies). In this example, the highlighted admixed haplotype is most similar to the haplotypes in Population 2, so local ancestry is assigned to that population.

with two or more ancestral populations. The method models the conditional probability of ancestry given the observed haplotype sequences using a Conditional Random Field (CRF). This CRF is parameterized by random forests trained on reference panels from each of the ancestral populations. When datasets are not too large, RFMix can use the ancestry information contained in the admixed samples themselves to augment small reference panels via an expectation-maximization (EM) approach; this is particularly useful when reference haplotypes that are well-representative of the ancestral population are unavailable or otherwise difficult to find. In large datasets, however, this EM option may not be computationally feasible [50]. An additional advantageous feature of RFMix is its ability to jointly model phase and ancestry, resulting in both improved phasing (when admixed haplotypes are statistically phased) and local ancestry inference.

Despite its many advantages, RFMix does not provide a specific X chromosome setting. This poses challenges for inferring local ancestry on X, which may be of interest for applications such as studying sex-biased admixture [47, 48, 86] or association testing on chromosome X [44, 49]. To use RFMix, the input admixed and reference haplotypes must first be phased and missing genotypes imputed. In the nonpseudoautosomal regions of X, males are haploid and thus already perfectly phased; females, on the other hand, are diploid and their phase will need to be inferred using statistical methods [87, 41, 88]. Missing data must be imputed for both males and females, and many imputation programs (e.g., [41, 42]) require that males be coded as homozygous diploid on X, such that they have two identical copies of their chromosome. This coding is common in many standard file formats [43], as well as other applications such as association testing on X [44]. However, we hypothesize that this coding could be problematic for inferring local ancestry on chromosome X. In particular, coding reference panel males as homozygous diploid on X will make the observed male reference haplotypes look twice as frequent as they really are, which will impact the training of the random forests on which the RFMix CRF is based.

In this chapter, we investigate this hypothesis and suggest alternative approaches for coding male haplotypes when performing local ancestry inference on chromosome X using

RFMix. We use these proposed strategies to infer local ancestry on X for 12,775 Hispanics/Latinos from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), and compare the performance of each approach. To conclude, we offer suggestions for researchers who hope to conduct local ancestry inference on X chromosome data using RFMix or other local ancestry inference programs that do not explicitly handle X.

2.2 Methods

2.2.1 Local Ancestry Inference on X with RFMix

To perform local ancestry inference on chromosome X using RFMix, we must make the following decisions: (1) how to phase female haplotypes and impute missing values for both males and females, (2) how to code male haplotypes, and (3) which version of RFMix to run.

Phasing and Imputation

Before running RFMix, haplotypes must be phased and sporadic missing genotypes imputed. As mentioned above, males are haploid on X (in the nonpseudoautosomal regions) and thus already perfect phased. However, females are diploid and their haplotypes will need to be statistically phased using one of many existing programs. Both males and females may have missing genotypes, so imputation of these sporadic missing genotypes will need to be performed on all samples. Programs such as Beagle [41] can be easily used to perform both steps. To run Beagle on chromosome X, males should be coded as homozygous diploids.

Male Coding Options

There are three options which we will consider for coding of males on chromosome X: haploid, homozygous diploid, and paired diploid (Figure 2.2). The *haploid* coding represents each male as a single haplotype, which is the truth for chromosome X, but, depending on the task and software, may not be allowed. The *homozygous diploid* coding duplicates each male haplotype so that each male becomes diploid with two identical copies of their X chromosome.

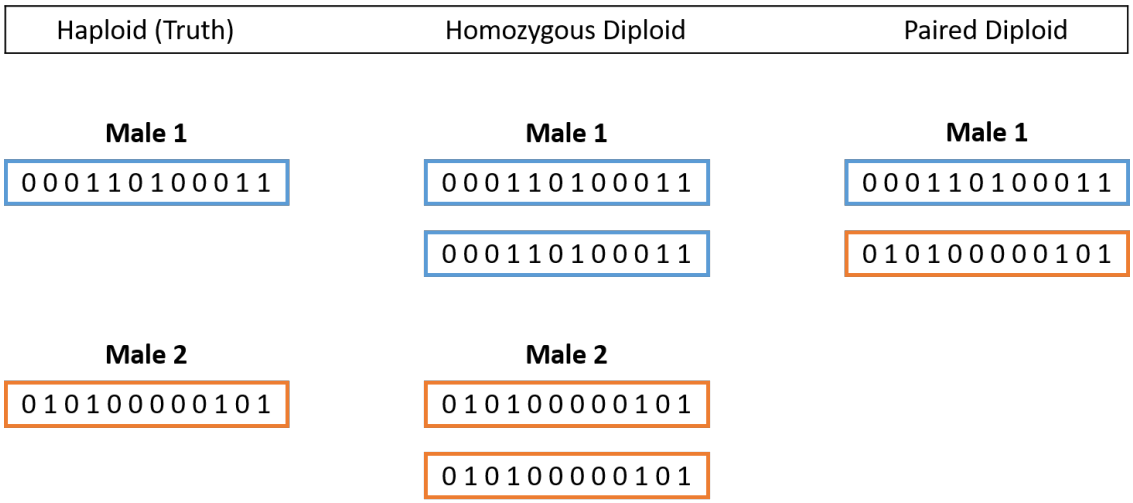

Figure 2.2: Haplotype coding options for males on chromosome X.
Each haplotype is represented by a sequence of zeros and ones, often with 1 representing the minor allele and 0 the major allele at each position. The haploid coding option represents the truth for males in the nonpseudoautosomal regions of chromosome X, but may not be allowed by all software.

This option is common in many file formats and suggested by software such as Beagle [41] for phasing and imputation. The final coding option, *paired diploid*, creates artificial, perfectly-phased diploid individuals by pairing male haplotypes. When diploid individuals are required by software but haplotypes are treated independently, this option is equivalent to the haploid coding (modulo a single haplotype that will need to be discarded if there is an odd number of males).

RFMix Versions

RFMix is implemented as two separate programs: *RFMix_PopPhased* and *RFMix_TrioPhased*. The *PopPhased* version jointly models phase and ancestry, accounting for uncertainty in the initial phasing of haplotypes. This option is recommended when admixed haplotypes are statistically phased using a program such as Beagle [41], but requires that all samples be diploid. The *TrioPhased* version of RFMix assumes that phase is known (or very accurately inferred, e.g., via trio phasing) and does not attempt to re-phase admixed haplotypes. Provided that

the EM option is not used, local ancestry is inferred independently for each haplotype. Given that re-phasing is not attempted with the *TrioPhased* option, it is considerably faster than *PopPhased*.

Running RFMix on Chromosome X

We consider the following combinations of male haplotype coding and RFMix versions for local ancestry inference on chromosome X. *A priori*, we anticipate that Options 1 and 2 will have superior performance; this will be discussed further in the *Results* and *Discussion*. Note that this is not an exhaustive list of all possible approaches: other options are either essentially equivalent to those considered here or are anticipated to have inferior performance (e.g., running *PopPhased* with paired diploid coding of admixed males, as any re-phasing would introduce errors) so are not considered. Table 2.1 summarizes the four options considered.

Option 1: Use the paired diploid coding for males in the reference panels (pairing males within ancestral populations and discarding the final male if the ancestral group has an odd number of males) and the haploid coding for admixed males. Run RFMix on the male and female admixed samples in parallel, using the *TrioPhased* option on the perfectly phased haploid males, the *PopPhased* option on the statistically phased diploid females, and the same reference panel for both analyses.

Option 2: Use the paired diploid coding for reference panel males and homozygous diploid coding for admixed males. Run *PopPhased* on the male and female admixed samples in a single analysis.

Option 3: Use the homozygous diploid coding for both reference panel and admixed males. Run *PopPhased* on the male and female admixed samples in a single analysis.

Option 4: Use the paired diploid coding for reference panel males and haploid coding for admixed males. Run *TrioPhased* on the male and female admixed samples in a single analysis.

Table 2.1: Summary of four approaches considered for local ancestry inference on chromosome X using RFMix.

Option	Refpanel Male Coding	Admixed Male Coding	RFMix Version
1	Paired diploid	Haploid	TrioPhased (M) + PopPhased (F)
2	Paired diploid	Homozygous diploid	PopPhased (all)
3	Homozygous diploid	Homozygous diploid	PopPhased (all)
4	Paired diploid	Haploid	TrioPhased (all)

2.2.2 Application to the Hispanic Community Health Study/Study of Latinos

The HCHS/SOL Data

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a community-based cohort study of 16,415 self-identified Hispanic/Latino participants aged 18–74 years old [45, 46]. Participants were recruited in four US field centers (Chicago, IL; Miami, FL; Bronx, NY; San Diego, CA) using a two-stage sampling scheme, with community block units sampled first, followed by households within the block units. Initial visits occurred between 2008 and 2011, and yearly telephone follow-up assessment lasted at least three years. Self-identified background groups vary across participants, with the largest groups being Mexican ($n = 6,471$), Puerto Rican ($n = 2,728$), Cuban ($n = 2,348$), Central American ($n = 1,730$), Dominican ($n = 1,460$), and South American ($n = 1,068$). The HCHS/SOL study was approved by institutional review boards at participating institutions, and written informed consent was obtained from all participants. More information about the study is available at https://www2.cscc.unc.edu/hchs/.

In total, 12,803 participants consented to genetic studies. Individuals were genotyped on an Illumina Omni 2.5M array with additional custom content, and the genotype and phenotype data are posted on dbGaP (accession numbers phs000880.v1.p1 and phs000810.v1.p1). After quality control and removal of 19 outlier individuals with significant Asian ancestry

[89] and 14 individuals with large amounts of missing data on chromosome X, our analyses included 12,775 individuals, 41% of whom are male (Table 2.2).

Reference Panel

To perform local ancestry inference using RFMix, a reference panel representing the ancestral populations of interest is required. We created a reference panel by combining individuals of Amerindian, European, and West African descent from the publicly available Human Genome Diversity Project (HGDP) [90] and 1000 Genomes Project (1000G) [91]. We ran an unsupervised ADMIXTURE analysis [7] on these individuals and kept those with at least 90% estimated ancestry from one of the inferred ancestral populations. After additionally excluding three individuals with large amounts of missing data on X, we were left with 63 Amerindians (from Peru, Mexico, Ecuador, Brazil, and Colombia), 524 Europeans (from Spain, Italy, the British Isles, France, and Utah), and 195 West Africans (from Nigeria, Senegal, and Barbados). The breakdown of males and females in this reference panel is presented in Table 2.2. The HGDP individuals were genotyped on the Illumina HumanHap650Y array and the 1000G individuals were genotyped on the Illumina Omni 2.5M array. After intersection with the HCHS/SOL variants and exclusion of single nucleotide polymorphisms (SNPs) with minor allele frequency smaller than 0.5%, a total of 9,081 SNPs remained on chromosome X for this analysis.

We also constructed a reference panel using data published in [92], combined with samples from 1000G. This reference panel has a larger number of individuals (154 Amerindians, 516 Europeans, and 198 West Africans) but a smaller number of SNPs that overlap with the HCHS/SOL data (236,736 versus 419,645 genome-wide). In a separate analysis, we found that using this reference panel for local ancestry inference on the autosomes led to slightly inferior performance compared to using the reference panel described above [50], so for the remainder of this chapter we will only present results using the reference panel constructed from HGDP and 1000G individuals.

Table 2.2: Sample sizes for inferring local ancestry on chromosome X in HCHS/SOL. Numbers of admixed (HCHS/SOL) and reference panel individuals included in our analysis.

		Reference Panel		
	Admixed	African	Amerindian	European
Males	5231	104	24	266
Females	7544	91	39	258
Total	12775	195	63	524

Local Ancestry Inference and Pre-Processing

Before running RFMix, we used Beagle 4.0 [41] to phase the HCHS/SOL and reference panel data and impute sporadic missing genotypes on chromosome X. The HCHS/SOL data include individuals with known relatedness, but we did not use the pedigree information in our phasing. All males (reference panel and admixed) were coded as homozygous diploid for these pre-processing steps. The phased and imputed haplotype data were then converted from VCF format into the specific RFMix input format described in the RFMix manual (available online: https://sites.google.com/site/rfmixlocalancestryinference/). For local ancestry inference, we used RFMix version 1.5.4. We performed analyses using the four options for male coding and RFMix version summarized in Table 2.1.

Assessing Performance of Local Ancestry Inference Approaches

We could not directly compare the accuracy of local ancestry inference approaches on the HCHS/SOL data because we do not know the true local ancestry for HCHS/SOL individuals. However, we do have known pedigree information for many of these individuals, and thanks to the household-based sampling design of the HCHS/SOL study there are large numbers of parent-offspring pairs ($n = 174$) and trios ($n = 203$). We used this pedigree information to compare the Mendelian inconsistency rates of each method, as has been done previously [11, 93]. Females inherit one X chromosome from their mother and one from their father,

while males inherit their single X chromosome from their mother; the inferred local ancestry should reflect these inheritance patterns. An inconsistency arises, for example, if a female is called as having two alleles of European origin at a position but her mother is called as having one African allele and one Amerindian allele. An additional example of inconsistent local ancestry calls is a position where a male is called as having one allele of European origin while his mother is called as having two Amerindian alleles. At every SNP along the genome, we classified a parent-offspring pair or mother-father-child trio as *Mendelian inconsistent* if their ancestry calls at the position were inconsistent with the laws of Mendelian inheritance, as in the examples mentioned above. We compared the rates of Mendelian inconsistencies across chromosome X for each of the four analysis options considered.

We also compared the posterior probabilities generated by RFMix for each approach. RFMix uses a forward-backward algorithm that produces posterior probabilities of each ancestry given the observed haplotype, and the ancestry with the largest posterior probability is the one that is ultimately called at that position [10]. In our analysis of the HCHS/SOL data, three posterior probabilities (for Amerindian, European, and West African ancestry) were generated for each haplotype at each position. We recorded the largest of these three probabilities, corresponding to the posterior probability for the called ancestry, and compared their averages across chromosome X. Ultimately, comparing these posterior probabilities allows us to compare how confident the RFMix algorithm was about the calls that it made.

Finally, we also compared the consistency of the local ancestry calls generated by the four approaches across chromosome X. In particular, for each pair of approaches we compared the proportion of calls, across all positions and haplotypes, that were identical.

2.3 Results

2.3.1 Mendelian Inconsistencies in HCHS/SOL

The Mendelian inconsistency rates for chromosome X are summarized in Figure 2.3 In general, Mendelian inconsistency rates were low for all four approaches. Inconsistency rates were

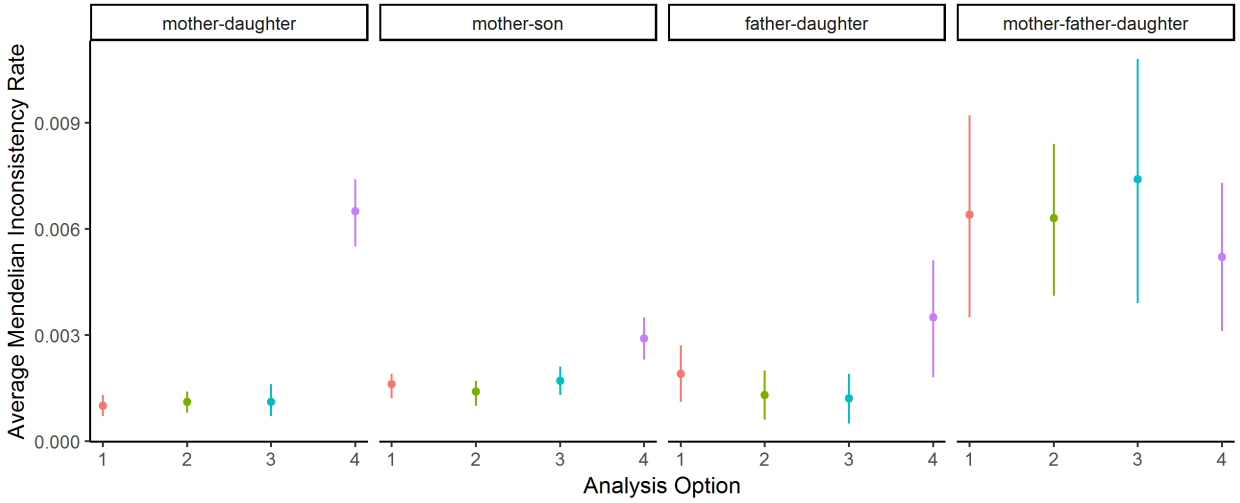

Figure 2.3: Mendelian inconsistency rates on chromosome X in HCHS/SOL.
For each chromosome X analysis option we present the average Mendelian inconsistency rate (with approximate 95% confidence interval) across HCHS/SOL parent-offspring pairs and trios.

lower in parent-offspring pairs than in trios due to the fact that errors in ancestry calls are more difficult to detect when data are missing for one parent. Note that mother-father-son trios are comparable to mother-son pairs since fathers do not contribute an X chromosome to their sons, so we combined these two groups into a single *mother-son* category. The approach with the highest Mendelian inconsistency rates was Option 4, which did not allow RFMix to correct potential phasing errors in the statistically phased admixed females. Among the remaining approaches, all of which allowed for re-phasing of females, the approaches that used paired diploid coding for reference panel males (Options 1 and 2) performed very similarly to one another and tended to slightly, but not significantly, outperform the approach that used homozygous diploid coding for reference panel males (Option 3). We note that the chromosome X mother-daughter Mendelian inconsistency rates for Options 1–3 are comparable to the rates observed for parent-offspring pairs on the autosomes [50].

Table 2.3: RFMix posterior probabilities on chromosome X in HCHS/SOL females. Average maximum posterior probability for each analysis option considered.

	Description			Posterior Probabilities			
Option	Ref. Males	Adm. Males	RFMix	Overall	African	Amerindian	European
1	Paired diploid	Haploid	Pop + Trio	0.9929	0.9937	0.9927	0.9924
2	Paired diploid	Hom. diploid	Pop	0.9930	0.9937	0.9927	0.9925
3	Hom. diploid	Hom. diploid	Pop	0.9931	0.9939	0.9927	0.9928
4	Paired diploid	Haploid	Trio	0.9872	0.9885	0.9866	0.9866

2.3.2 RFMix Posterior Probabilities in HCHS/SOL

On chromosome X, admixed males are not directly comparable across the four analysis options considered due to the different diploid and haploid coding schemes, so we only present posterior probability results for females (Table 2.3). We present overall results, as well as results broken down by ancestral population. Across all approaches and ancestral populations, the average maximum posterior probabilities were close to 1. Posterior probabilities were slightly higher on average among positions assigned to African origin, suggesting that RFMix was slightly more confident in calling African ancestry than European or Amerindian ancestry. On average, the lowest posterior probabilities were observed with the approach that did not allow for re-phasing of females (Option 4). Options 1–3 treated admixed females identically and had more similar posterior probabilities, with slightly larger posterior probabilities observed, on average, for Option 3 (which used homozygous diploid coding for reference panel males). This pattern could be explained by the fact that the duplication of male haplotypes in the reference panel under Option 3 leads the RFMix program to think that it has a larger number of reference haplotypes and then, as a result, it is (inappropriately) more confident in its calls. We note that posterior probabilities may be miscalibrated; this is a pitfall of using posterior probabilities to assess accuracy.

Table 2.4: Comparison of local ancestry calls generated by four chromosome X analysis options in HCHS/SOL.
Percent of identical local ancestry calls across all haplotypes and loci considered.

	Option 1	Option 2	Option 3	Option 4
Option 1	1.0000	0.9980	0.9960	0.9864
Option 2		1.0000	0.9961	0.9865
Option 3			1.0000	0.9851
Option 4				1.0000

2.3.3 Consistency of Calls in HCHS/SOL

Across chromosome X, the four approaches yielded largely similar local ancestry calls. For any two pairs of approaches, the local ancestry calls were identical for at least 98.5% of loci and haplotypes (Table 2.4). Options 1 and 2, which treated admixed females identically and used the same paired diploid coding for reference panel males, yielded the most consistent local ancestry calls with one another (99.8% identical) and provided similarly consistent calls with Option 3 (99.6% identical). Option 4—the only approach that did not allow for re-phasing of admixed females—was least similar to the other three approaches.

2.4 Discussion

Like many genetics software tools, RFMix was designed only for autosomal data. We investigated four approaches to analyzing the X chromosome with RFMix. We assessed the results using Mendelian inconsistency rates, and also compared RFMix posterior probabilities and the consistency of local ancestry calls generated by the four approaches. The two best approaches involved pairing haploid males into pseudodiploid individuals in the reference panel in order to avoid the double-counting that would occur if these individuals were coded as homozygous diploid. However, we found that coding reference males as homozygous diploids resulted in only a small loss of accuracy. Larger losses in accuracy were observed when we

did not allow RFMix to correct for potential phasing errors in the statistically phased female haplotypes.

Our comparisons are limited by the fact that we do not know the true local ancestry for the HCHS/SOL participants. However, we were able to compare accuracy using Mendelian inconsistency rates, which provides valuable insight into the accuracy of competing local ancestry inference approaches provided that the pedigree information used was correct. Future work might explore the performance of these approaches using simulated data where the underlying true local ancestry is known. In particular, it would be interesting to investigate the impact of reference panel male coding on the training of the random forests which parameterize the RFMix conditional random field. Furthermore, given the size of our dataset, we were unable to use the EM option for RFMix; whether our findings extend to the setting where RFMix is used with the EM option remains to be shown.

Our results clearly demonstrate the advantages posed by the joint modeling of phase and ancestry, and confirm that the *PopPhased* option should be used for haplotypes that are statistically phased (e.g., females on chromosome X). As hypothesized, our results also suggest that coding males as homozygous diploids in the reference panel could lead to losses in accuracy, although we did not observe as dramatic of a loss as initially anticipated. In terms of the treatment of admixed male haplotypes, our results are less conclusive: we did not see any practically significant differences in accuracy when haploid males were analyzed separately from females using the *TrioPhased* version of RFMix versus when homozygous diploid males were analyzed with females using *PopPhased*. However, there are some practical differences between these approaches in terms of the speed of computation (using *TrioPhased* on males and analyzing males and females in parallel was faster than running *PopPhased* on all samples) and ease of use (running a single *PopPhased* analysis with the standard homozygous diploid coding of admixed males was slightly easier to implement). The decision about which of these approaches to use may be left to personal preference.

The issue of extending existing software for the analysis of chromosome X is not unique to local ancestry inference with RFMix. As we have seen in this chapter, the standard ho-

mozygous diploid coding of male haplotypes may not always be appropriate. Instead, careful consideration should be given to both the application and the underlying algorithm of the software being used before proceeding with analysis of chromosome X. In the case of supervised local ancestry inference, particular care should be taken to avoid over-representation of male haplotypes in the reference panel. Two approaches that address this concern have been proposed here.

Chapter 3

GENOME-WIDE SIGNIFICANCE THRESHOLDS FOR ADMIXTURE MAPPING STUDIES

3.1 Introduction

Due to the processes involved in the inheritance of genetic material, the genomes of admixed individuals are a mosaic of segments with different ancestral origins (Figure 1.1). This mosaic pattern of locus-specific ancestry, or *local ancestry*, varies considerably across individuals within an admixed population and proves useful for identifying causal genetic variants via admixture mapping. Admixture mapping studies scan the genomes of admixed individuals for associations between local ancestry and a trait of interest [12, 13, 14]. Disease prevalence and trait values often differ across ancestral groups (e.g., asthma [18], prostate cancer [19], blood pressure [20]), due to a combination of genetic and environmental causes. By looking for associations between a trait and local ancestry, admixture mapping seeks to identify the genetic variants that differ in frequency across these ancestral groups and drive the observed phenotypic differences. In recent years, admixture mapping has become more widely used and has proven to be a powerful approach for localizing causal genetic variants [19, 25, 27, 29, 56, 94, 95, 96, 97].

A single genome-wide admixture mapping study will typically involve hundreds of thousands or millions of hypothesis tests, and multiple testing correction procedures are needed to control the overall type I error rate. Perhaps the best-known multiple testing correction procedure is the Bonferroni correction. In the context of admixture mapping, we can perform a Bonferroni correction based on the total number of loci tested, or on the (often considerably smaller) number of unique blocks of local ancestry [26, 98]. Although easy to implement, this approach is widely criticized for yielding conservative significance thresh-

olds in the presence of correlated tests. To address the issue of correlated hypothesis tests, various authors have proposed using a Bonferroni correction instead on the estimated effective number of independent tests [52, 99, 51, 32]; however, it has been shown that these approaches do not always guarantee family-wise error rate control in genome-wide associaiton studies [54, 55, 53]. Permutation- and simulation-based multiple testing correction procedures [100, 101, 31, 28, 30, 33, 26] are often considered to be the gold standard for genetic association studies, but can be very computationally intensive. Alternatives to these procedures, based on the multivariate normal distribution, have been suggested to speed up computation time [53, 102, 103].

A promising alternative to the above-mentioned approaches involves an analytic multiple testing correction [31, 104, 58]. In particular, Siegmund and Yakir [58] derived the correlation of admixture mapping test statistics and used that theoretical result to provide an analytic approximation to the appropriate significance threshold for admixture mapping studies in admixed populations with two ancestral populations and equal admixture proportions across individuals. However, many admixed populations have more than two ancestral populations (e.g., Hispanics/Latinos) and/or unequal admixture proportions across individuals in the population [47, 105, 89, 106, 107], the latter being a consequence of population structure [48].

In this chapter, we develop a theoretical framework that applies to admixed populations with any number of ancestral populations or distribution of admixture proportions, and then use that theoretical framework to develop multiple testing correction procedures for admixture mapping studies in admixed populations with population structure. We apply our proposed procedures to genotype data for individuals of African American and Hispanic/Latina ancestry from the Women's Health Initiative SNP Health Association Resource (WHI SHARe). Finally, we perform a simulation study using these WHI SHARe genotype data and simulated traits to validate our theoretical work and evaluate the performance of our approach relative to other commonly used multiple testing correction procedures.

3.2 Methods

3.2.1 Admixture Mapping Model

Following previous studies [14, 26, 24], we use marginal regression to perform admixture mapping in samples with unrelated individuals, regressing the trait of interest on inferred local ancestry at each observed locus across the genome. At each locus, we quantify local ancestry as the number of alleles (0, 1, or 2) inherited from each ancestral population at that locus. In an admixed population with K ancestral populations, we characterize the local ancestry for admixed individual i at locus j via the vector $\mathbf{a}_{ij} = \begin{pmatrix} a_{ij1} & \cdots & a_{ijK} \end{pmatrix}^\top$, where $\sum_{k=1}^{K} a_{ijk} = 2$ and the kth component of this vector, a_{ijk}, denotes the number of alleles inherited by individual i from ancestral population k at locus j. Similarly, we represent the admixture proportions (global ancestry) for each individual via the vector $\boldsymbol{\pi}_i = \begin{pmatrix} \pi_{i1} & \cdots & \pi_{iK} \end{pmatrix}^\top$, where $\sum_{k=1}^{K} \pi_{ik} = 1$ and the components of this vector represent the overall (genome-wide) proportion of genetic material inherited by individual i from each ancestral population. To perform admixture mapping, we regress the trait of interest, $\mathbf{y}$, on each component of the local ancestry vector ($k = 1, \ldots, K$) at each locus ($j = 1, \ldots, m$) using the marginal regression model

$$E\left[y_i \mid a_{ijk}, \boldsymbol{\pi}_i\right] = \alpha + \beta_{jk} a_{ijk} + \boldsymbol{\gamma}\boldsymbol{\pi}_{i,-K}, \tag{3.1}$$

where $\boldsymbol{\pi}_{i,-K} = \begin{pmatrix} \pi_{i,1} & \cdots & \pi_{i,K-1} \end{pmatrix}^\top$ includes the first $K-1$ components of the vector of admixture proportions. We fit separate regression models for each ancestral group in order to investigate which ancestral population(s) drive the association between the trait and local ancestry at each locus, and we adjust for estimated admixture proportions in all models to account for potential population structure [13, 24]. We test for association between the trait and local ancestry using a Wald test, where the test statistic is the ratio of the estimated regression coefficient for the local ancestry term and its standard error $\left(Z_{jk} = \frac{\hat{\beta}_{jk}}{\widehat{\text{se}}(\hat{\beta}_{jk})}\right)$, with one test statistic per locus and ancestral component.

3.2.2 Theoretical Framework: Joint Distribution of Admixture Mapping Test Statistics

Our goal is to derive a significance threshold that controls the family-wise error rate, or the probability of making at least one type I error, for a genome-wide admixture mapping study. In other words, we wish to find the genome-wide test statistic threshold Z^* such that

$$\Pr\left(\max_{1\leq j\leq m, 1\leq k\leq K} |Z_{jk}| > Z^* \mid \beta_{jk} = 0 \ \forall j,k\right) = \alpha^*,$$

for some pre-specified level a^* (e.g., 0.05). To derive this threshold, we must understand the asymptotic joint distribution of our admixture mapping test statistics $Z_{11}, \ldots, Z_{mK}$.

The first step is to characterize the correlation of local ancestry vectors at pairs of loci across the genome. For an admixed population with any number of ancestral populations, generations since admixture, or distribution of admixture proportions across the population, we can show that the correlation of local ancestry vectors $(\mathbf{a}_j, \mathbf{a}_{j'})$ at two loci (j, j') depends on the recombination fraction between the loci (θ), the number of generations since admixture (g), and the population mean, variance, and covariance of the admixture proportions.

Lemma 1. *Consider an admixed population with K ancestral populations, g generations since admixture, and global ancestry proportions distributed according to $\boldsymbol{\pi} \sim F$, where F has finite first and second moments. Then, the correlation of local ancestry vectors at two loci j, j' separated by recombination fraction θ is given by:*

$$Corr(a_{jk}, a_{j'k'}) = \begin{cases} (1-\theta)^g + [1-(1-\theta)^g] \frac{2V_F(\pi_k)}{E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)} & \text{if } k = k' \\ \frac{2Cov_F(\pi_k, \pi_{k'}) - (1-\theta)^g [Cov_F(\pi_k, \pi_{k'}) + E_F(\pi_k) E_F(\pi_{k'})]}{\sqrt{\left[E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)\right]\left[E_F(\pi_{k'}) - E_F^2(\pi_{k'}) + V_F(\pi_{k'})\right]}} & \text{if } k \neq k'. \end{cases}$$

A proof of Lemma 1 is available in Appendix A.1.1. Note that if all individuals in the population have the same admixture proportions (i.e., $\boldsymbol{\pi}_i = \boldsymbol{\pi}\ \forall i$) then the local ancestry correlation simplifies to $(1-\theta)^g$ when $k = k'$, as had been shown previously in the context of admixed populations with two ancestral populations [58].

Using Lemma 1, combined with the central limit theorem, it is straightforward to derive an approximation to the asymptotic joint distribution of our collection of admixture mapping test statistics $\mathbf{Z} = \begin{pmatrix} Z_{11} & \cdots & Z_{mK} \end{pmatrix}^\top$. For an admixed population with any number of ancestral populations, generations since admixture, or distribution of admixture proportions across the population, we can show that the asymptotic joint distribution of the test statistics $\mathbf{Z}$ can be approximated by a mean zero Gaussian process with a convenient covariance/correlation structure.

Theorem 1. *Consider an admixed population with K ancestral populations, g generations since admixture, and global ancestry proportions distributed according to $\boldsymbol{\pi} \sim F$, where F has finite first and second moments. For loci $j \in \{1, ..., m\}$ and ancestry components $k \in \{1, ..., K\}$, define test statistics $Z_{jk} = \frac{\hat{\beta}_{jk}}{\widehat{se}(\hat{\beta}_{jk})}$ based on Model (3.1). Then, under the universal null hypothesis ($\beta_{jk} = 0 \ \forall j, k$), the collection of test statistics $\mathbf{Z} = (Z_{11} \ \cdots \ Z_{mK})^T$ has an asymptotic multivariate normal distribution with mean $\mathbf{0}$ and covariance (and correlation) given by*

$$Cov(Z_{jk}, Z_{j'k'}) = \begin{cases} (1-\theta)^g & \text{if } k = k' \\ -(1-\theta)^g \frac{E_F(\pi_k)E_F(\pi_{k'})}{\sqrt{E_F(\pi_k)[1-E_F(\pi_k)]E_F(\pi_{k'})[1-E_F(\pi_{k'})]}} & \text{if } k \neq k', \end{cases}$$

where θ is the recombination fraction between loci j, j'.

A proof of Theorem 1 is available in Appendix A.1.2. Note that the covariance of test statistics simplifies conveniently when the admixed population has only two ancestral populations ($K = 2$):

$$\text{Cov}(Z_{jk}, Z_{j'k'}) = \begin{cases} (1-\theta)^g \approx \exp(-0.01g\delta) & \text{if } k = k' \\ -(1-\theta)^g \approx -\exp(-0.01g\delta) & \text{if } k \neq k', \end{cases}$$

where δ is the genetic distance in centimorgans (cM) between those loci. It follows that the distribution of admixture mapping test statistics can then be approximated by an Ornstein-Uhlenbeck process, as has been shown previously by Siegmund and Yakir in the context of

admixed populations with two ancestral populations and constant admixture proportions [58]. Our result in Theorem 1 generalizes the results from Siegmund and Yakir to admixed populations with any number of ancestral populations or distribution of admixture proportions.

3.2.3 Multiple Testing Correction Procedures

We propose two multiple testing correction procedures that use the asymptotic joint distribution of admixture mapping test statistics provided by Theorem 1 to derive a genome-wide significance threshold that will control the family-wise error rate in admixture mapping studies. Both approaches are implemented in our R package `STEAM` (Significance Threshold Estimation for Admixture Mapping).

Simulation-Based Approach

To estimate the appropriate genome-wide test statistic threshold for an admixture mapping study, we simulate test statistics from their asymptotic joint distribution (Theorem 1) and choose the threshold that controls the empirical family-wise error rate at the desired level. This approach differs from traditional simulation-based multiple testing approaches in that we simulate test statistics directly, rather than simulating traits and re-calculating test statistics at each locus for each simulation replicate. By simulating test statistics directly, computation time for our multiple testing correction procedure is considerably reduced and, importantly, is independent of sample size. We have developed a fast algorithm for simulating admixture mapping test statistics from this distribution which takes advantage of its convenient covariance structure (see Appendix A.2). As input, we only require the genetic distances between consecutive loci, the estimated admixture proportions for individuals in the sample, and an estimate of the number of generations since admixture (g). We suggest that g be estimated using our non-linear least squares approach described in Section 3.2.4.

Analytic Approximation Approach

An alternative approach for deriving genome-wide significance thresholds in the special case of admixed populations with two ancestral populations ($K = 2$) was developed previously. Siegmund and Yakir [58] showed that, under some assumptions, the asymptotic joint distribution of admixture mapping test statistics can be approximated by an Ornstein-Uhlenbeck process, and then used that result to provide an analytic approximation to the family-wise error rate:

$$\Pr\left(\max_{1\leq j\leq m,k} |Z_{jk}| > z\right) \approx 1 - \exp\{-2C[1-\Phi(z)] - 2\beta L z\phi(z)\nu(z\sqrt{2\beta\Delta})\}, \tag{3.2}$$

where C is the number of chromosomes analyzed, having total genetic length L cM; Δ is the marker density; Φ and ϕ are the cumulative distribution and density functions, respectively, of the standard normal distribution; $\beta = 0.01g$; and the function ν is an infinite sum which can be approximated by $\nu(y) \approx \frac{(2/y)[\Phi(y/2)-0.5]}{(y/2)\Phi(y/2)+\phi(y/2)}$. Although this analytic approximation was initially proposed for admixture mapping studies in populations with equal admixture proportions across individuals [58], our work (i.e., Theorem 1) shows that it is also applicable to populations with heterogeneous admixture proportions, provided that the admixture proportions are included as covariates in the regression analysis. As a result, we can use this analytic approximation to find the admixture mapping test statistic threshold that will control the family-wise error rate at the desired level (α^*) in an admixed population with two ancestral populations and any distribution of admixture proportions: we simply find the value z that sets the right hand side of Equation 3.2 equal to α^*. This involves an optimization step that can be quickly solved using existing tools (e.g., `uniroot` in R [108]). Simulation is not required for this approach, so the significance threshold can be derived in a matter of seconds.

3.2.4 Estimating the Number of Generations since Admixture

Both the analytic approximation and simulation-based multiple testing correction approaches rely on the number of generations since admixture. We can estimate the number of gener-

ations since admixture (g) using the observed pattern of local ancestry correlation in our sample. From Lemma 1 we know that g determines the rate of decay of this correlation:

$$\text{Corr}(a_{jk}, a_{j'k'}) = \begin{cases} a^* + (1-a^*)(1-\theta)^g & \text{when } k = k', \\ 2b^* - (b^* + c^*)(1-\theta)^g & \text{when } k = k', \end{cases} \tag{3.3}$$

where $a^*, b^*,$ and c^* are scalars that depend on $E_F(\boldsymbol{\pi})$ and $\text{Cov}_F(\boldsymbol{\pi})$. We propose an approach, similar in spirit to that of Hellenthal et al. [81], that uses this result, combined with non-linear least squares (NLS) estimation, to estimate the number of generations since admixture that provides the best fit (i.e., minimizes the sum of squared residuals) to the observed local ancestry correlation curves.

We consider two variations of this NLS approach. The *unconstrained* approach ignores the known form of a^*, b^*, c^* given by Lemma 1 and instead allows these scalars to be estimated via NLS along with g. The *constrained* approach, on the other hand, utilizes the known relationship between a^*, b^*, c^* and the population mean and covariance of the admixture proportions $\boldsymbol{\pi}$. First, we estimate a^*, b^*, c^* by replacing $E_F(\pi), \text{Cov}_F(\pi)$ with their sample equivalents (e.g., $\hat{a}^* = \frac{2\hat{V}(\hat{\pi}_k)}{\hat{E}(\hat{\pi}_k) - \hat{E}^2(\hat{\pi}_k) + \hat{V}(\hat{\pi}_k)}$). Then, we constrain the values of $a^*, b^*,$ and c^* in Equation (A.2) to be equal to these estimates, and use NLS to find the value of g that provides the best fit to the equations

$$\widehat{\text{Corr}}(a_{jk}, a_{j'k'}) = \begin{cases} \hat{a}^* + (1-\hat{a}^*)(1-\theta)^g & \text{when } k = k', \\ 2\hat{b}^* - (\hat{b}^* + \hat{c}^*)(1-\theta)^g & \text{when } k = k', \end{cases}$$

where now g is the only unknown parameter. More details are available in Appendix A.3.

Both constrained and unconstrained non-linear least squares approaches are implemented in our R package `STEAM`, along with our multiple testing correction procedures. We compare the constrained and unconstrained approaches in Section 3.3.2

3.2.5 Analysis of WHI SHARe Data

We applied our multiple testing correction procedures to two cohorts of admixed individuals with African American and Hispanic/Latina ancestry from the Women's Health Initiative

SNP Health Association Resource (WHI SHARe), and also used these data to perform simulation studies comparing the performance of our proposed multiple testing correction procedures to competing approaches.

The Data

The WHI is a long-term health study of postmenopausal women in the United States. A total of 161,808 postmenopausal women aged 50–79 years old were recruited, including 12,151 self-identified African Americans (AA) and 5,469 self-identified Hispanic Americans (HA) who had consented to genetic research. Study design details and cohort characteristics are described elsewhere [109]. A subsample of these women were selected for genotyping, using 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. In these analyses, we focus only on the SNPs. 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% and 12,008 (8,421 AA and 3,587 HA) women used in the current analysis [29].

Local Ancestry Inference

To implement and evaluate our proposed multiple testing correction procedures in the WHI SHARe data, we first needed to infer local ancestry. This process involved multiple steps. First, we formed reference panels for local ancestry inference using individuals of European, African, and Native American descent from the International HapMap Project (HapMap) [110] and the Human Genome Diversity Project (HGDP) [111]. In particular, the reference panels for both the AA and HA cohorts included 165 individuals of European descent (HapMap *CEU*, Utah residents with Northern and Western European ancestry) and 203 individuals of African descent (HapMap *YRI*, Yoruba in Nigeria), and the HA reference panel additionally included 63 individuals of Native American descent from HGDP. We identified a set of 551,025 and 536,374 SNPs common to the reference panels and the WHI AA and

HA samples, respectively. Second, we used an iterative procedure suggested by Conomos et al. [112] to identify sets of 8,064 and 3,425 mutually unrelated AA and HA individuals, respectively. To classify unrelated individuals, we used a kinship threshold of 0.044; this threshold corresponds to excluding first, second, and third degree relatives [112]. Third, we performed phasing and imputation of sporadic missing genotypes using Beagle version 3 [41]. Genetic distances were estimated using the publicly available HapMap genetic map [113]. After these pre-processing steps, we performed local ancestry inference using `RFMix` [10] to estimate the number of alleles inherited from each ancestral population at each locus across the genome.

Application of Multiple Testing Correction Procedures

We implemented the analytic approximation approach in the AA cohort and our test statistic simulation-based approach (with 10,000 replications) in both the AA and HA cohorts. Both approaches require the number of generations since admixture, which we estimated from the observed pattern of local ancestry correlation in these samples using our non-linear least squares approach described above. Our simulation-based approach additionally requires admixture proportions, which we estimated for each individual using the genome-wide average inferred local ancestry.

Simulation Study Using WHI SHARe Genotypes

To evaluate the performance of our proposed methods, we simulated 10,000 sets of traits for each individual according to the model $y_i \sim_{iid} N(0,1)$. We used `PLINK` v1.9 [114] to perform admixture mapping in each cohort, adjusting for estimated admixture proportions (Model 3.1). We calculated the observed correlation of these test statistics across simulation replicates to compare to our theoretical result (Theorem 1) and evaluated the empirical family-wise error rate of our multiple testing correction methods across the 10,000 simulation replicates.

We compared our approaches to three competing methods: the trait simulation approach with 10,000 replicates and a Bonferroni correction based on either the total number of hypothesis tests (the number of loci multiplied by the number of tests performed per locus: $n_{AA} = 551,025,\ n_{HA} = 1,609,122$) or the number of *unique* hypothesis tests performed (the number of local ancestry blocks multiplied by the number of tests performed at each block: $n_{AA} = 14,795,\ n_{HA} = 44,145$). Note that the Bonferroni correction based on local ancestry blocks may not always be relevant, depending on the choice of software used for local ancestry inference. Some programs, including `RFMix`, perform local ancestry inference within small windows (the default window length for `RFMix` is 0.2 cM) and produce identical local ancestry calls for every locus within the windows. As a result, all admixture mapping test statistics will be identical within each window and we can reduce the number of hypothesis tests performed by simply testing one locus per window.

For the two simulation-based multiple testing correction approaches (our test statistic simulation approach and the trait simulation approach), we also calculated a 95% bootstrap confidence interval for the genome-wide significance threshold. Specifically, we bootstrapped the vector of the largest (in absolute value) test statistics from each of the 10,000 simulation replicates and used the percentiles of the bootstrapped test statistics to generate a 95% confidence interval.

3.3 Results

3.3.1 Population Structure and Validation of Theoretical Results in WHI SHARe

The WHI SHARe African American (AA) and Hispanic American (HA) cohorts exhibit considerable heterogeneity in estimated admixture proportions (Figure 3.1), indicating that the theoretical work of previous authors [58] would not be applicable to these samples, even in the case of the AA cohort with just two ancestral populations. However, we do observe that the patterns of local ancestry and test statistic correlation in the WHI SHARe samples are consistent with our new theoretical results (Figure 3.2).

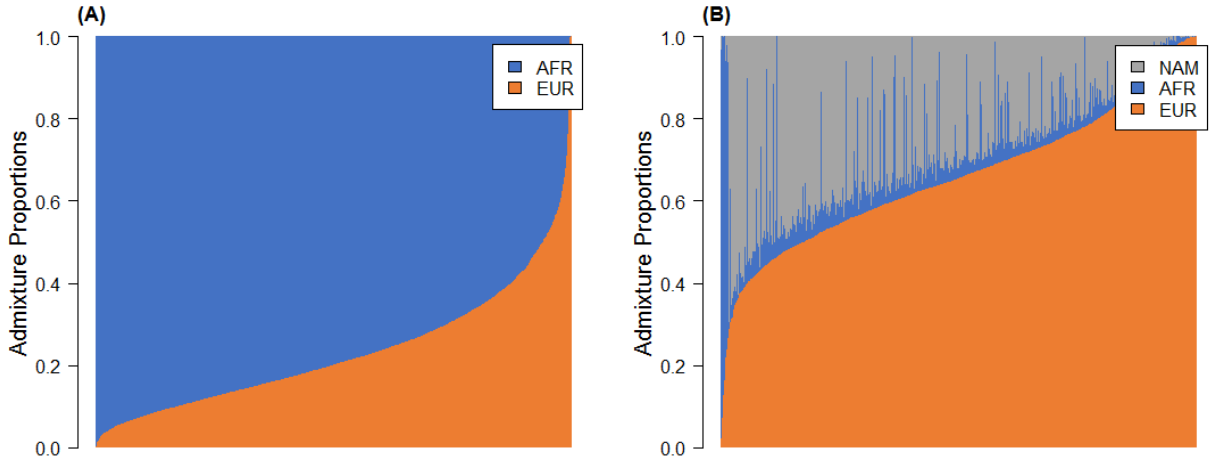

Figure 3.1: Estimated admixture proportions in WHI SHARe.
(A) Estimated proportions of genetic material inherited from African (AFR) and European (EUR) ancestral populations for the African American samples.
(B) Estimated proportions of Native American (NAM), African (AFR), and European (EUR) ancestry for the Hispanic American samples.

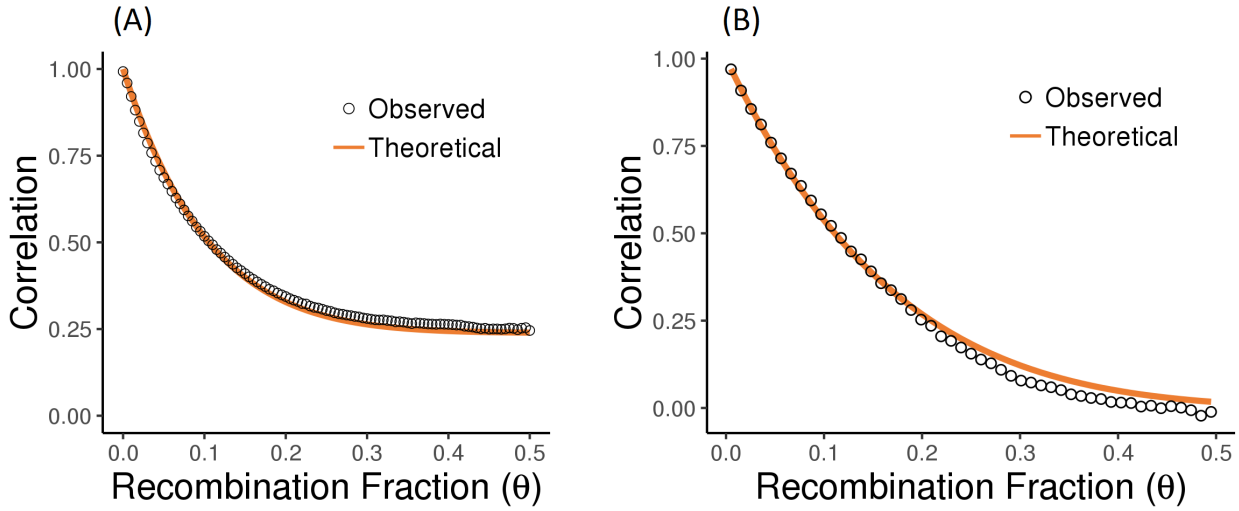

Figure 3.2: Correlation of local ancestry and test statistics in WHI SHARe.
(A) Comparison of the observed and expected (theoretical) correlation of the European component of local ancestry vectors in the Hispanic American cohort, averaging across pairs of markers falling into bins defined by their distance apart. The expected correlation comes from Lemma 1, with $g = 9.6$.
(B) Comparison of the observed and expected (theoretical) correlation of admixture mapping test statistics in the African American cohort, with expected correlation corresponding to Theorem 1, using $g = 5.9$.

3.3.2 Estimating the Generations since Admixture

Based on the observed patterns of local ancestry correlation in these data, our unconstrained non-linear least squares (NLS) regression approach yields estimates of 5.9 and 9.6 generations since admixture for the African American and Hispanic American cohorts, respectively. These estimates are consistent with previously published studies [48, 23, 115, 116, 9]. Using these estimated admixture times, we see a nice correspondence between the observed local ancestry and test statistic correlation and the expected correlation based on Lemma 1 and Theorem 1 (Figure 3.2). The constrained NLS approach yields similar estimates of the number of generations since admixture ($\hat{g}_{AA} = 6.2,\ \hat{g}_{HA} = 9.2$). We see slightly better correspondence between the observed and fitted local ancestry correlation curves using the unconstrained approach (see Appendix A.3), and this approach provides the additional advantage of not depending on the estimated admixture proportions, so we used the estimates $\hat{g}$ from the unconstrained approach for our multiple testing correction.

3.3.3 Comparison of Multiple Testing Correction Procedures in WHI SHARe

In the African American cohort, our multiple testing correction procedures yield genome-wide p-value thresholds of 2.1×10^{-5} and 2.0×10^{-5} for the test statistic simulation and analytic approximation approaches, respectively. Both thresholds are consistent with the threshold given by the trait simulation approach (see Table 3.1), and are orders of magnitude less stringent than the Bonferroni thresholds. The empirical family-wise error rate for each approach from a simulation study using simulated traits is reported in Table 3.2. By design, the trait simulation approach controls the empirical family-wise error rate exactly at the nominal level 0.05. Our proposed test statistic simulation and analytic approximation procedures also control family-wise error rate at the nominal level, while the Bonferroni corrections, as expected, are conservative.

The derived significance thresholds for the Hispanic American cohort are more stringent than those in the African American cohort. Our test statistic simulation procedure yields a

p-value threshold of 4.5×10^{-6}, which is again consistent with the trait simulation threshold (Table 3.1) and controls the empirical family-wise error rate at the nominal level (Table 3.2). As in the African American cohort, the Bonferroni correction based on the total number of loci tested yields a significance threshold that is orders of magnitude too conservative, and the Bonferroni correction based on unique ancestry blocks is somewhat less stringent, but still conservative.

The difference between the estimated genome-wide p-value thresholds for the African American and Hispanic American cohorts reflect the differences between the two cohorts in terms of the number of ancestral populations ($K = 2$ vs $K = 3$), number of generations since admixture ($\hat{g} = 5.9$ vs $\hat{g} = 9.6$), and distribution of admixture proportions. The difference in significance thresholds appears to be primarily driven by the different number of ancestral populations being considered in the two populations, since we consistently estimate a p-value threshold that is one order of magnitude more stringent in the Hispanic American cohort even if we use the same value of g for both groups. That said, the number of generations since admixture and variability of admixture proportions in the population do also impact the estimated significance threshold.

3.3.4 Computation Time

Computation time differs considerably across the five approaches. The Bonferroni correction can be used to compute the significance threshold nearly instantaneously. The analytic approximation approach is also very quick, taking under half a second on a 12-core 2.4 GHz computer with Intel Xeon E5-2630Lv2 processors and 128 GB of memory. The slowest is the trait simulation approach: for our WHI SHARe analyses, each replicate (which involved running a genome-wide admixture mapping study) took approximately five minutes on the same computer, for a total of more than 800 hours of computation time to run all 10,000 replicates. In comparison, our test statistic simulation approach took only a fraction of a second per replicate, amounting to less than ten minutes to run all 10,000 replicates in the African American and Hispanic American cohorts.

Table 3.1: Comparison of p-value thresholds from five multiple testing correction procedures in WHI SHARe African American (AA) and Hispanic American (HA) samples.
For simulation-based approaches, we also provide a 95% bootstrap confidence interval. Both simulation-based approaches used 10,000 replications. The nominal genome-wide type I error rate (α^*) is 0.05. From left to right, the five procedures being compared are a Bonferroni correction based on the total number of loci (*No. Loci*), Bonferroni based on the number of unique ancestry blocks (*No. Blocks*), the trait simulation approach (*Traits*), our test statistic simulation approach (*Test Stats*), and our analytic approximation approach (*Analytic Approx.*).

	Bonferroni		Simulation		
	No. Loci	No. Blocks	Traits	Test Stats	Analytic Approx.
AA	9.1×10^{-8}	3.4×10^{-6}	2.1×10^{-5}	2.1×10^{-5}	2.0×10^{-5}
			$(1.9, 2.3) \times 10^{-5}$	$(1.9, 2.2) \times 10^{-5}$	
HA	3.1×10^{-8}	1.1×10^{-6}	4.3×10^{-6}	4.5×10^{-6}	n/a
			$(4.0, 4.6) \times 10^{-6}$	$(3.9, 4.9) \times 10^{-6}$	

Table 3.2: Empirical family-wise error rate of five multiple testing correction procedures in simulation studies using WHI SHARe African American (AA) and Hispanic American (HA) genotype data.
Empirical family-wise error rate was calculated across 10,000 replications of a simulated null trait. The nominal genome-wide type I error rate (α^*) is 0.05. From left to right, the five procedures being compared are a Bonferroni correction based on the total number of loci (*No. Loci*), Bonferroni based on the number of unique ancestry blocks (*No. Blocks*), the trait simulation approach (*Traits*), our test statistic simulation approach (*Test Stats*), and our analytic approximation approach (*Analytic Approx.*).

	Bonferroni		Simulation		
	No. Loci	No. Blocks	Traits	Test Stats	Analytic Approx.
AA	5×10^{-4}	0.010	0.050	0.050	0.048
HA	8×10^{-4}	0.014	0.050	0.052	n/a

3.4 Discussion

We have developed a theoretical framework to characterize the correlation of local ancestry vectors and admixture mapping test statistics in admixed populations with any number of ancestral populations and distribution of admixture proportions. Our application to data from the Women's Health Initiative SNP Health Association Resource highlights the importance of this extension, as both the African American and Hispanic American samples display considerable heterogeneity in admixture proportions (Figure 3.1). Based on these new theoretical results, we show that an existing analytic approximation [58] can be used to derive significance thresholds for admixture mapping studies in admixed populations with two ancestral populations, even in the presence of population structure, as long as the admixture mapping model adjusts for admixture proportions. For admixed populations with any number of ancestral populations, we propose an approach that simulates test statistics directly from their asymptotic joint distribution, saving considerable computation time relative to the trait simulation approach, while still yielding an appropriate significance threshold that controls the family-wise error rate.

Our multiple testing correction procedures are based on theoretical work that explicitly models the correlation of admixture mapping test statistics, so are not conservative like the commonly used Bonferroni correction; this will translate to gains in power in genome-wide admixture mapping studies. Compared to the trait simulation approach, our correction procedures yield comparably appropriate significance thresholds but are far less computationally intensive, and we provide an R package for easy implementation. Furthermore, by simulating test statistics directly from their asymptotic distribution, the computation time of our simulation-based multiple testing approach does not increase with sample size, which will prove useful as future studies are able to recruit larger and larger numbers of individuals. We believe that our approaches provide an attractive alternative for researchers looking to control for multiple testing in genome-wide admixture mapping studies, particularly in admixed populations with population structure.

In this chapter, our theoretical work and data analyses have focused on genome-wide admixture mapping studies with quantitative traits and unrelated individuals. However, preliminary analyses indicate that our theoretical work extends easily to binary traits (see Appendix A.4). In the case of quantitative traits that are heavily skewed (or otherwise depart considerably from normality) larger sample sizes may be needed for asymptotic normality of the test statistics to be achieved; to address this problem, transformations such as rank-normalization [117, 118] could be considered. The presence of relatedness, accounted for by use of a mixed model [62], should not change the marginal distribution of admixture mapping test statistics, but would likely change their correlation structure. We expect that this will not have a large impact on the appropriate significance threshold, but further investigation is needed to confirm this hypothesis. The application of our approach to data with complex relatedness and population structure is considered in Chapter 5.

Our multiple testing correction procedures require estimates of the admixture proportions for each admixed individual, the number of generations since admixture, and the genetic distance between consecutive loci. To assess sensitivity to the choice of genetic map used to produce these pairwise genetic distances, we implemented `STEAM` in the WHI African American cohort using both the HapMap genetic map and an African American-specific genetic map [78]. Although these maps are quite different in some regions of the genome, we found that they still produce similar estimates of the number of generations since admixture (HapMap: 5.9, African American map: 5.7) and the genome-wide p-value threshold (HapMap: 2.1×10^{-5} (95% CI: $1.9\times10^{-5}, 2.2\times10^{-5}$); African American map: 2.0×10^{-5}, $(1.8\times10^{-5}, 2.3\times10^{-5})$). Our estimates of the number of generations since admixture (g) may be sensitive to assortative mating or departures from the assumption of a single instantaneous admixture event. Assortative mating can lead to increased variability in admixture proportions across a population [48, 119], which our approach accounts for by allowing these proportions to vary, and may additionally change the pattern of local ancestry correlation in the sample [119], which will impact our estimate of the number of generations since admixture. However, in application to real admixed populations (e.g., WHI SHARe) where departures from the assumption

of a single instantaneous admixture event and/or random mating (e.g., due to geographic constraints) are likely, we find that our approach still works well. In estimating the parameter g from observed data using our proposed method, we are able to appropriately capture the correlation structure of admixture mapping test statistics in the sample, which is what is important for estimating an appropriate genome-wide significance threshold. Extending our theoretical work to explicitly model assortative mating, multiple admixture times, or continuous admixture is an interesting area for future work.

The p-value threshold 5×10^{-8} has become quite widely adopted as a control for multiple testing in genome-wide association studies [33, 120, 121, 34], but there is no such "established" threshold for admixture mapping studies. Even in the specific context of the WHI SHARe genotype data, at least four different genome-wide p-value thresholds have been used in published admixture mapping analyses in the African American cohort (including 7×10^{-6} [29, 56], 1×10^{-5} [57], 1.5×10^{-5} [122], and 1.82×10^{-5} [100]), demonstrating the lack of consensus up to this point—even across analyses of the same dataset—on how best to derive significance thresholds for genome-wide admixture mapping studies. In practice, many admixture mapping studies cite the work of other studies (e.g., Tang et al. [31]) as the basis for their chosen significance threshold. However, our theoretical work and analysis of the African American and Hispanic American WHI SHARe cohorts demonstrate that admixture mapping significance thresholds are not necessarily transferable across studies. In particular, the appropriate significance threshold depends on the number of ancestral populations, generations since admixture (to which it is particularly sensitive), population structure (through the distribution of admixture proportions), and density of markers tested, all of which often differ from one study to another. We encourage investigators to take this important point into consideration when choosing a significance threshold for their own genome-wide admixture mapping study.

Chapter 4

CONTROLLING FOR ANCESTRAL HETEROGENEITY IN GENETIC ASSOCIATION STUDIES IN ADMIXED POPULATIONS

4.1 Introduction

Admixed populations have historically been vastly underrepresented in genetic studies [1, 4, 2, 123, 3, 124]. Although this underrepresentation has many causes, some authors have cited the statistical challenges posed by ancestrally heterogeneous populations as a possible contributing factor [1, 4, 2]. Heterogeneous ancestry is an example of *population structure*, which is known to pose challenges for genetic association studies. Population structure arises as a consequence of non-random mating—due to factors such as geographical separation or assortative mating based on physical traits (e.g., height), social factors (e.g., educational attainment), or ancestry [119, 125, 126]—and leads to increased variability of global ancestry within populations [119]. In this chapter, we consider the impact of variable global ancestry, or *ancestral heterogeneity*, on genetic association studies in admixed populations.

Considerable variability of inferred global ancestry—the genome-wide proportion of genetic material inherited from each ancestral population—has been observed in many studies of African American and Hispanic/Latino populations [106, 105, 47, 107, 89]. This ancestral heterogeneity must be addressed when performing genetic association studies in those populations. It has been widely documented that heterogeneous global ancestry can lead to spurious associations in genome-wide association studies (GWAS) [36, 37, 127, 128]. It is generally understood that these spurious associations arise due to the fact that global ancestry can confound the association between genotypes and a phenotype of interest, particularly when genetic variants are more frequent in some ancestral populations than in others and

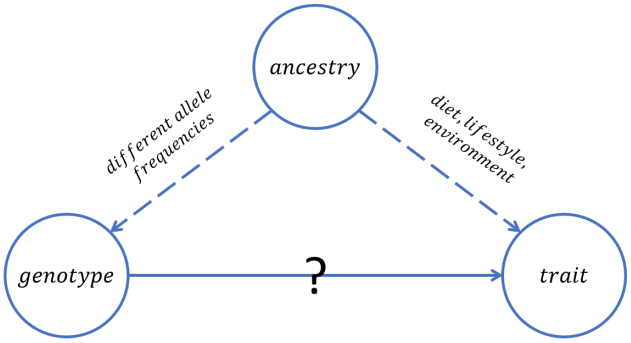

Figure 4.1: Global ancestry is a potential confounder in GWAS.

global ancestry has a direct effect on the trait through, for example, environmental differences across ancestral groups (Figure 4.1). This issue has not been studied as extensively in the context of admixture mapping, although some authors have suggested that ancestral heterogeneity can also induce spurious associations in admixture mapping studies [13, 24].

A number of methods for detecting and controlling for ancestral heterogeneity in genetic association studies have been proposed. Early approaches included restricting analyses to subsets of ancestrally homogeneous individuals [61], performing a genome-wide correction for test statistic inflation due to ancestral heterogeneity via *genomic control* [36], and using family-based study designs [60]. More recently, approaches based on mixed models have been proposed (e.g., [62, 63, 64]). These mixed model approaches use random effects to control for both close (e.g., due to family-based sampling) and distant (e.g., due to shared ancestry) relatedness across individuals. However, when studies do not include closely related individuals, a simpler approach is to include inferred global ancestry as a fixed effect in marginal regression models [65, 37, 24]. This fixed effects adjustment for global ancestry is currently used extensively throughout the literature, with global ancestry inferred using either model-based ancestry inference methods (e.g., `ADMIXTURE` [7]) or unsupervised dimension reduction techniques (e.g, principal component analysis (PCA) [37]).

Model-based approaches for global ancestry inference model the probability of observed genotypes given unobserved ancestry and allele frequencies in each ancestral population

[84, 6, 7, 8]. Most often, these approaches are used to estimate *admixture proportions* $\hat{\boldsymbol{\pi}}_i = \begin{pmatrix} \hat{\pi}_{i1} & \dots & \hat{\pi}_{iK} \end{pmatrix}^\top$ for individuals $i = 1, \dots, n$, where $\hat{\pi}_{ik}$ is the estimated proportion of genetic material inherited by individual i from ancestral population k. Once estimated, $\hat{\boldsymbol{\pi}}$ can then easily be included as a covariate in GWAS or admixture mapping models to adjust for ancestral heterogeneity. One of the challenges of using these model-based approaches to infer global ancestry is that the number of ancestral populations, K, usually needs to be pre-specified. In addition, some of these model-based approaches are *supervised*, requiring reference panel data from each ancestral population of interest to estimate allele frequencies. Furthermore, ancestry inference is typically conducted at a continental level (e.g., African versus European, rather than South European versus North European), so finer levels of population structure could be missed; recent efforts have consider global ancestry inference on a sub-continental scale [8, 82].

Principal component analysis (PCA), on the other hand, is a widely-implemented unsupervised approach for inferring global ancestry that does not require reference panel data or pre-specification of the number of ancestral populations of interest and is capable of capturing sub-continental structure (e.g., [66]). To infer global ancestry using PCA, we perform an eigenvalue decomposition of the genetic relationship matrix (GRM) $\hat{\boldsymbol{\Psi}} = \frac{1}{m}\mathbf{S}\mathbf{S}^\top$, where $\mathbf{S}$ is the $n \times m$ matrix of standardized genotypes for n individuals at m single nucleotide variants (SNVs). The top eigenvectors, or *principal components* (PCs), of $\hat{\boldsymbol{\Psi}}$ tend to reflect global ancestry [40, 39], so adjusting for PCs can be an effective approach for controlling for ancestral heterogeneity in genetic association studies [37]. In practice, however, determining the number of PCs needed to capture global ancestry can be difficult. Furthermore, it has been shown that PCs can sometimes capture features other than global ancestry, such as relatedness across samples [67, 40], data quality issues [37, 68], and/or small regions of the genome with unusual patterns of linkage disequilibrium (LD) [69, 70]. To address this last issue, some authors have suggested running PCA on a reduced subset of SNVs, after first removing regions of the genome that are known to have high or long-range LD [70] and/or performing LD pruning [129, 130]; however, these suggestions are not universally

implemented, and the downstream implications of adjusting for PCs that capture features other than global ancestry are not fully understood.

In this chapter, we investigate the impact of ancestral heterogeneity on genome-wide association and admixture mapping studies in admixed populations. Through both simulation studies and analytic results, we provide new insight into when genetic association studies must adjust for global ancestry. In addition, we compare two approaches for adjusting for global ancestry, using model-based estimates of admixture proportions or principal components, and show that using PCs can induce spurious associations in both GWAS and admixture mapping studies. To conclude, we provide suggestions regarding best practices for appropriately controlling for ancestral heterogeneity in genetic association studies in admixed populations.

4.2 Methods

4.2.1 Regression models for genetic association studies

To perform genetic association studies in samples of unrelated admixed individuals, we use marginal regression models. In genome-wide association studies (GWAS), we regress the trait of interest on the genotypes at each observed locus across the genome. For admixture mapping studies, we instead regress the trait on inferred local ancestry.

At a given locus j, we quantify genotype g_{ij} as the number of copies (0, 1, or 2) of some pre-specified allele (e.g., the minor allele) carried by individual i at that locus. Local ancestry is quantified by the vector $\mathbf{a}_{ij} = \begin{pmatrix} a_{ij1} & \dots & a_{ijK} \end{pmatrix}^\top$, where $\sum_{k=1}^{K} a_{ijk} = 2$ and the kth component of this vector, a_{ijk}, denotes the number of alleles (0, 1, or 2) inherited by individual i from ancestral population k at locus j. Considering a quantitative trait y_i, we can then represent GWAS and admixture mapping regression models via the general form

$$E[y_i \mid x_{ij}, \mathbf{w}_i] = \beta_0 + \beta_j x_{ij} + \boldsymbol{\beta}_w \mathbf{w}_i.$$

If we are conducting a GWAS, then $x_{ij} = g_{ij}$; for admixture mapping studies, $x_{ij} = a_{ijk}$, for some ancestral population of interest k. In both cases, $\mathbf{w}_i$ is a vector of additional covariates

(e.g., precision variables, potential confounders) that we want to include in the model. We fit these models at every locus $j = 1, \ldots, m$ across the genome and test for an association between the trait and genotype or local ancestry by testing the null hypothesis $H_0 : \beta_j = 0$.

4.2.2 Adjusting for ancestral heterogeneity in genetic association studies

We compare two methods for adjusting for ancestral heterogeneity in genetic association studies. The first approach uses model-based ancestry inference techniques to estimate admixture proportions $\hat{\boldsymbol{\pi}}_i$. This could involve estimating admixture proportions directly using a program such as `ADMIXTURE` [7], or by calculating the genome-wide average local ancestry (i.e., $\hat{\pi}_{ik} = \frac{1}{m}\sum_{j=1}^{m} a_{ijk}$), where local ancestry was first inferred using a program such as `RFMix` [10]. To adjust for ancestral heterogeneity, we include these estimated admixture proportions as covariates in our GWAS and admixture mapping regression models:

$$E[y_i \mid x_{ij}, \hat{\boldsymbol{\pi}}_i] = \beta_0 + \beta_j x_{ij} + \beta_{\pi,1}\hat{\pi}_{i,1} + \ldots \beta_{\pi,K-1}\hat{\pi}_{i,K-1}. \tag{4.1}$$

The second approach runs PCA on sample genotypes to generate the principal components $\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 & \ldots \mathbf{v}_n \end{pmatrix}$. We choose some number of principal components, P, needed to capture global ancestry ($1 \leq P < n$). A number of techniques have been proposed for selecting P, including formal significance tests based on Tracy-Widom theory [40, 37], examining the proportion of variance explained by each PC [131], comparing PCs to self-reported ancestry [89], and/or keeping PCs that are significantly associated with the trait [29, 132]. Once P is selected, we adjust for ancestral heterogeneity by including the top P principal components as covariates in our GWAS and admixture mapping regression models:

$$E[y_i \mid x_{ij}, v_{i1}, \ldots, v_{iP}] = \beta_0 + \beta_j x_{ij} + \beta_{v1} v_{i1} + \ldots \beta_{vP} v_{iP}. \tag{4.2}$$

We will also compare these approaches to models that do not make any adjustment for ancestral heterogeneity:

$$E[y_i \mid x_{ij}] = \beta_0 + \beta_j x_{ij}. \tag{4.3}$$

4.2.3 Genome-wide simulation study using WHI SHARe African American genotype data

We compared the performance of Models (4.1), (4.2), and (4.3) using genotype data and simulated traits for 8,064 unrelated African American individuals from the Women's Health Initiative SNP Health Association Resource (WHI SHARe).

The Women's Health Initiative genotype data

The WHI is a long-term health study of postmenopausal women in the United States. A total of 161,808 postmenopausal women aged 50–79 years old were recruited, including 12,151 self-identified African Americans who had consented to genetic research. Study design details and cohort characteristics are described elsewhere [109]. A subsample of these women were selected for genotyping, using 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. In these analyses, we focus only on the SNPs.

Quality control and local ancestry inference

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% and 8,421 African American women [29]. We used an iterative procedure suggested by Conomos et al. [112] to identify a subset of 8,064 mutually unrelated individuals, using a kinship threshold of 0.044 (i.e., excluding first, second, and third degree relatives [112]).

For local ancestry inference, we formed a reference panel using genotype data from the International HapMap Project (HapMap) [110]. In particular, the reference panel included 165 individuals of European descent (HapMap *CEU*, Utah residents with Northern and Western European ancestry) and 203 individuals of African descent (HapMap *YRI*, Yoruba in Nigeria). A total of 551,025 SNPs were present in both the reference panel and admixed genotype data. We performed phasing and imputation of sporadic missing genotypes using

Beagle version 3 [41] and estimated genetic positions for each SNP using the publicly available HapMap genetic map [113]. After these pre-processing steps, we performed local ancestry inference using `RFMix` [10] to estimate the number of alleles inherited from each ancestral population at each locus across the genome.

Global ancestry inference

To estimate admixture proportions $\hat{\boldsymbol{\pi}}_i$ for each WHI individual, we calculated the genome-wide average local ancestry using local ancestry calls from `RFMix`. We also ran supervised and unsupervised `ADMIXTURE` analyses with two ancestral populations ($K = 2$). We found that the estimated admixture proportions from `ADMIXTURE` were highly correlated with those based on average local ancestry (Pearson correlation = 0.9984 for both supervised and unsupervised analyses), so we decided to use only the local ancestry based admixture proportion estimates for the remainder of our analyses.

We ran PCA on the WHI SHARe genotype data using `SNPRelate` [133]. First, we applied PCA to the set of all 551,025 SNPs with available genotypes. We refer to these PCs as the *naively generated PCs*. We also applied PCA to subsets of SNPs based on the following filtering criteria: excluding SNPs falling into regions of the genome that have been cited in the literature as being potentially problematic for PCA (Table 4.1), LD pruning with an r^2 threshold of 0.1 and window size of 0.5 mega basepairs (Mb), or both Table 4.1 exclusions and LD pruning. After these filtering steps, a total of 536,668 (exclusions only), 49,723 (pruning only), and 48,794 (both exclusions and pruning) SNPs remained.

The regions in Table 4.1 have all been shown to have extended, high LD. Long-range LD can arise as a consequence of selection, as well as other factors such as inversions [70]. It has been previously demonstrated that PCA is sensitive to LD [136, 69, 70], so it is fairly common that researchers will exclude some subset of this list of high LD regions prior to running PCA [70, 129, 134, 89, 135]. However, these exclusions are not universally implemented.

Table 4.1: Regions of the genome with unusual patterns of linkage disequilibrium (LD). This list of regions was generated on the basis of an extensive literature review conducted in June 2017. Start and end physical (base pair) positions are provided with respect to genome build 36.

Chr	Start (bp)	End (bp)	References
1	48000000	52060567	[129, 70]
2	85941853	100500000	[129, 70]
2	134382738	138000000	[70, 134, 89]
2	182882739	190000000	[129, 70]
3	47500000	50000000	[129, 70]
3	83500000	87000000	[129, 70]
3	89000000	97500000	[70]
5	44000000	51500000	[129, 70, 135]
5	98000000	100500000	[70]
5	129000000	132000000	[129, 70]
5	135500000	138500000	[70]
6	24999925	33500000	[129, 70, 135, 134, 89]
6	57000000	64000000	[129, 70]
6	140000000	142500000	[129, 70]
7	55000000	66193285	[129, 70]
8	8000000	12000000	[129, 70, 135, 134, 89]
8	43000000	50000000	[129, 70]
8	112000000	115000000	[129, 70]
10	37000000	43000000	[129, 70]
11	45000000	57000000	[70, 135]
11	87500000	90500000	[129, 70]
12	33000000	40000000	[129, 70]
12	109500000	112021663	[70]
17	40000000	43000000	[89]
20	32000000	34500000	[129, 70]

Simulating quantitative traits

Traits for unrelated individuals $i = 1, \ldots, 8064$ were simulated such that they depended only on genotype g_{ij} at a single causal SNP with effect size β_j:

$$y_i = \beta_j g_{ij} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, 1).$$

We considered 7 choices of effect sizes ($\beta_j = 0, 0.25, 0.5, 1, 2, 4, 8$) and 473 choices for the position j of the causal SNP, varying the position of this causal SNP across all 22 chromosomes. To choose causal SNPs, we calculated the difference in observed allele frequencies for each SNP in HapMap CEU and YRI, as well as the SNP loadings—the contribution of a SNP to each PC [136]—for the first four naively generated PCs. Based on these calculations, we selected 473 SNPs with positions spread across the genome, high or low influence on the naively generated PCs, and large or small ancestral allele frequency differences. We first identified the 10 SNPs on each chromosome with the highest absolute SNP loadings for each of the top four PCs. In total, 373 unique SNPs were selected according to this procedure. For comparison, we also selected 100 SNPs across the autosomes with low SNP loadings ($|\text{loading}| < 0.0008$) for all of the first four PCs, with 85 of these SNPs selected such that they had different allele frequencies in the African and European ancestral populations ($|\hat{p}_{CEU} - \hat{p}_{YRI}| > 0.6$), and 15 SNPs with similar allele frequencies in the two ancestral populations ($|\hat{p}_{CEU} - \hat{p}_{YRI}| < 0.005$).

Evaluating ancestral heterogeneity adjustment approaches

For each simulated trait, we ran GWAS and admixture mapping studies using Models (4.1)–(4.3). For the principal component adjustment approach (Model 4.2), we considered four sets of PCs based on different filtering criteria (*none, exclusions only, pruning only, exclusions and pruning*) and two choices for the number of PCs included in the model: $P = 1$ or $P = 4$.

To evaluate these ancestral heterogeneity adjustment approaches, we compared the number of spurious associations that appeared when we used each model. We quantified spurious associations by counting the number of chromosomes, not including the chromosome on which

the causal SNP was located, with at least one p-value below the genome-wide significance threshold. For GWAS, the p-value threshold was set equal to the 5.0×10^{-8} threshold that is used extensively throughout the GWAS literature [33, 34]. For admixture mapping, we used a p-value threshold of 2.0×10^{-5}, which was estimated using our program `STEAM` [59] (see also: Chapter 3).

4.2.4 Software

All methods described above were implemented using freely available software: Beagle (phasing and imputation) [41], `RFMix` (local ancestry inference) [10], `ADMIXTURE` (global ancestry inference) [7], `SNPRelate` (LD pruning and PCA) [133], `PLINK` (GWAS and admixture mapping) [114], and `STEAM` (estimating admixture mapping significance thresholds) [59].

4.3 Results

4.3.1 Genetic association studies must adjust for global ancestry

Failing to adjust for ancestral heterogeneity can cause spurious associations in genetic association studies, even if global ancestry does not have a direct effect on the trait. In the WHI SHARe data, we simulated a trait depending only on genotype at the SNP *rs2036153* on chromosome 4 and conducted GWAS and admixture mapping studies using this simulated trait. Manhattan plots from these analyses are provided in Figures 4.2 and 4.3. Since only *rs2036153* is truly associated with the trait, we would like to see a single peak in the Manhattan plots on chromosome 4. However, if we use a GWAS or admixture mapping model that does not adjust for global ancestry, we observe statistically significant associations on *every* chromosome (Figures 4.2A, 4.3A). In stark contrast, models that adjust for estimated admixture proportions show a single genome-wide significant signal on chromosome 4, as desired (Figures 4.2D, 4.3D). This pattern of results is not unique to this simulation setting: when we repeat this analysis, moving the location of the causal SNP, we see that the unadjusted GWAS and admixture mapping models have severely inflated rates of spurious

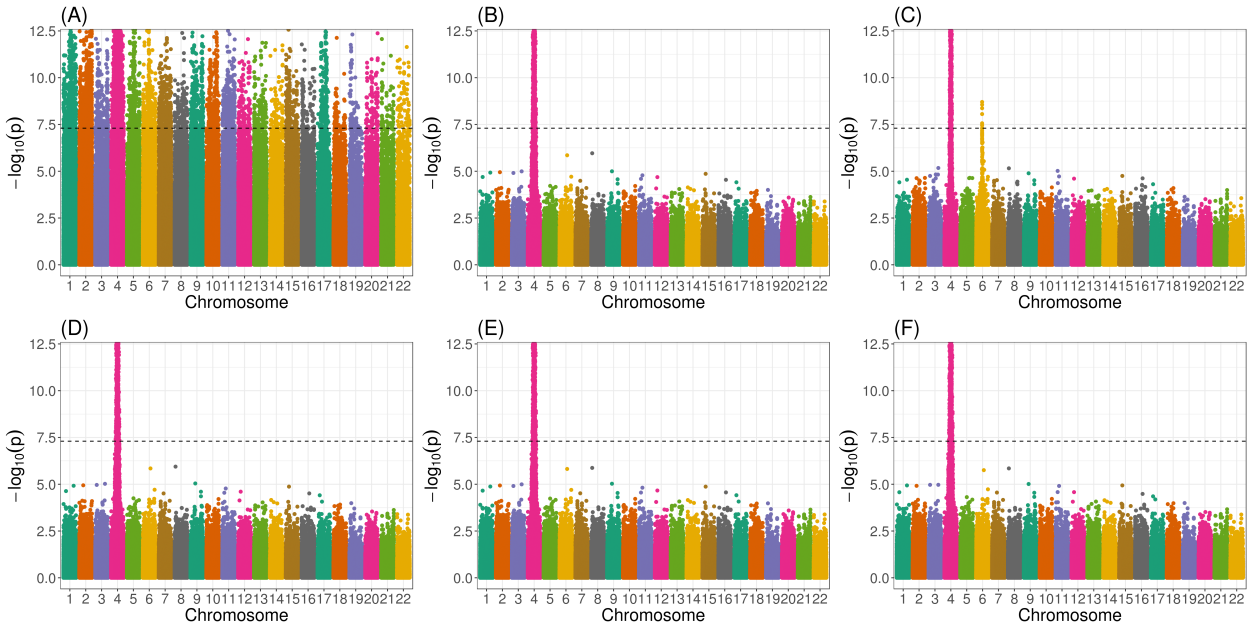

Figure 4.2: Manhattan plots from GWAS in WHI SHARe African Americans using six different approaches to adjust for ancestral heterogeneity.
GWAS in a sample of 8,064 unrelated African Americans, analyzing a simulated trait that depends only on genotype at a single SNP on chromosome 4: $y_i \sim N(g_{i,rs2036153}, 1)$. Panels present results based on different adjustment approaches: **(A)** = no adjustment, **(B)** = one PC, with PCs calculated using all SNPs, **(C)** = four PCs, with PCs calculated using all SNPs, **(D)** = estimated admixture proportions, **(E)** = one PC, with PCs calculated after LD pruning ($r^2 = 0.1$, window = 0.5 Mb) and Table 4.1 exclusions, and **(E)** = four PCs, with PCs calculated after LD pruning ($r^2 = 0.1$, window = 0.5 Mb) and Table 4.1 exclusions.

associations relative to models that adjust for global ancestry, except when the causal SNP has the same allele frequency in the two ancestral populations (Figure 4.4).

To better understand these patterns, we derived the expected effect size estimates from GWAS and admixture mapping models in admixed populations with two ancestral populations and potentially heterogeneous admixture proportions. We assume that the trait depends on genotypes only at a single causal SNP, as in our WHI simulations, and we compare the behavior of different models at this causal SNP, as well as a SNP on another chromosome that is not associated with the trait. At the causal locus (SNP 1) and the unlinked neutral locus (SNP 2), the expected effect size estimates from GWAS and admixture

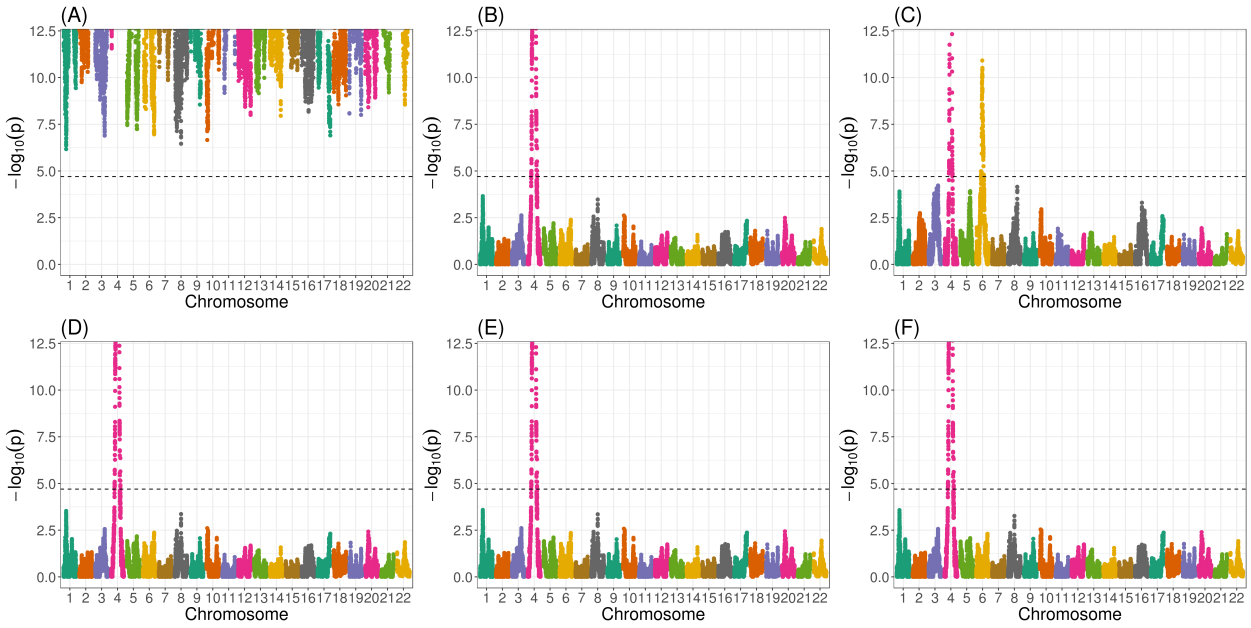

Figure 4.3: Manhattan plots from admixture mapping studies in WHI SHARe African Americans using 6 different approaches to adjust for ancestral heterogeneity.
Admixture mapping in a sample of 8,064 unrelated African Americans, analyzing a simulated trait that depends only on genotype at a single SNP on chromosome 4: $y_i \sim N(g_{i,rs2036153}, 1)$. Panels present results based on different adjustment approaches: **(A)** = no adjustment, **(B)** = one PC, with PCs calculated using all SNPs, **(C)** = four PCs, with PCs calculated using all SNPs, **(D)** = estimated admixture proportions, **(E)** = one PC, with PCs calculated after LD pruning ($r^2 = 0.1$, window = 0.5 Mb) and Table 4.1 exclusions, and **(E)** = four PCs, with PCs calculated after LD pruning ($r^2 = 0.1$, window = 0.5 Mb) and Table 4.1 exclusions.

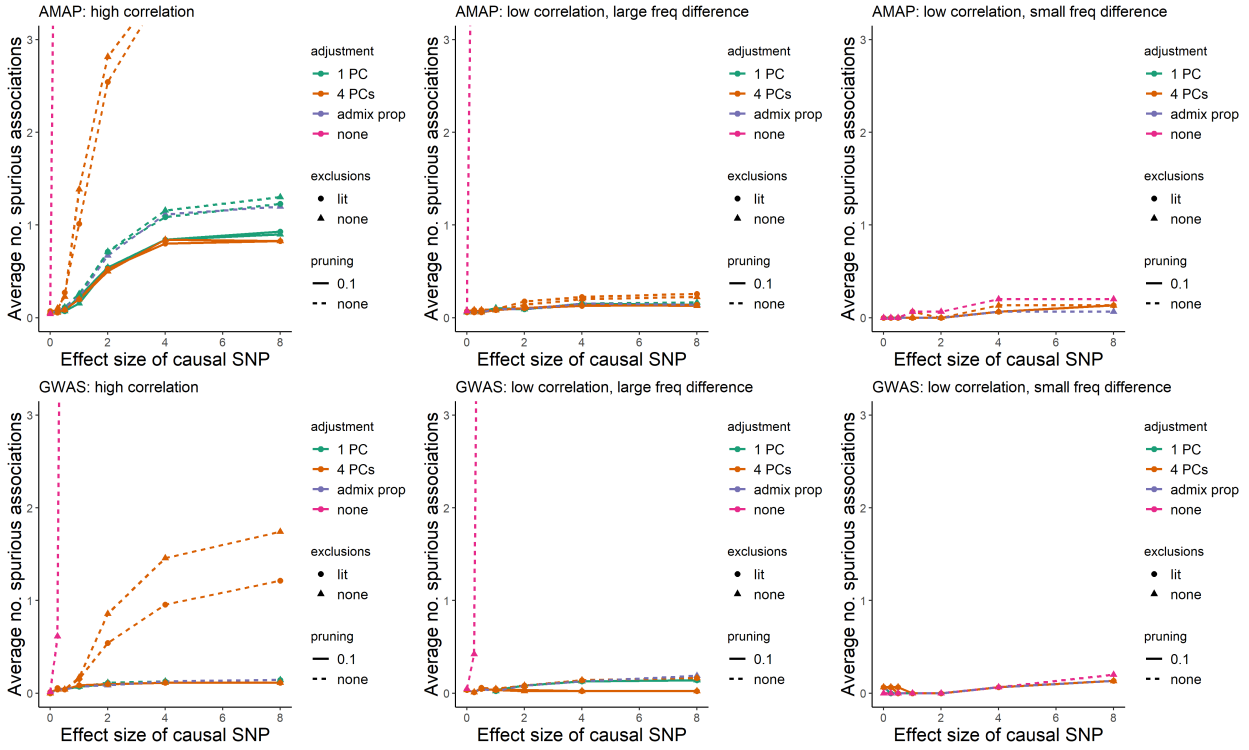

Figure 4.4: Comparison of the average number of spurious associations by ancestral heterogeneity adjustment technique and characteristics of the causal SNP.
For each simulation replicate, we count the number of spurious associations that arise in each genetic association study. Panels separate results by the type of genetic association study (top row = admixture mapping, bottom row = GWAS) and the characteristics of the causal SNP (left = causal SNP was highly correlated with the naively generated PCs, middle = causal SNP was uncorrelated with the naively generated PCs but had different allele frequencies in Africans and Europeans, right = causal SNP was uncorrelated with the PCs and had similar allele frequencies in Africans and Europeans). We compare the number of spurious associations from Model 4.1 (purple), Model 4.2 with $P = 1$ (green), Model 4.2 with $P = 4$ (orange), and Model 4.3 (pink). PCs were generated with (*lit*) or without (*none*) exclusion of regions from Table 4.1 and with ($r^2 =$ *0.1*, window = 0.5 Mb) or without (*none*) LD pruning.

mapping (AMAP) models that condition on the true admixture proportions (i.e., Model 4.1 with $\hat{\pi}_i = \pi_i$) are

$$E[\hat{\beta}_1] = \begin{cases} \beta_1 & \text{(GWAS with Model 4.1)} \\ \beta_1(p_{11} - p_{10}) & \text{(AMAP with Model 4.1)} \end{cases}$$

$$E[\hat{\beta}_2] = \begin{cases} 0 & \text{(GWAS with Model 4.1)} \\ 0 & \text{(AMAP with Model 4.1),} \end{cases}$$

where β_1 is the true effect size of the causal SNP (SNP 1), $\beta_2 = 0$ is the true effect size of the neutral SNP (SNP 2), and p_{11}, p_{10} are the allele frequencies of SNP 1 in the two ancestral populations. First, we note that these results support the observation by other authors that the power of admixture mapping studies depends on the difference in ancestral allele frequencies [137, 27, 138]. Second, it is clear from these analytic results that models that perfectly adjust for global ancestry will yield unbiased estimates of the effect size at the causal and unlinked neutral loci. As a result, models adjusting for global ancestry will control the rate of spurious associations. In comparison, GWAS and admixture mapping models that *do not* adjust for ancestral heterogeneity will yield the following estimated effect sizes at the causal and neutral SNPs:

$$E[\hat{\beta}_1] = \begin{cases} \beta_1 + \frac{(p_{11}-p_{10})V_\pi\beta_\pi}{p_{10}(1-p_{10})+(p_{11}-p_{10})(1-p_{11}-p_{10})E_\pi+(p_{11}-p_{10})^2(V_\pi+E_\pi-E_\pi^2)} & \text{(GWAS with Model 4.3)} \\ \beta_1(p_{11} - p_{10}) + \frac{V_\pi\beta_\pi}{V_\pi+E_\pi-E_\pi^2} & \text{(AMAP with Model 4.3)} \end{cases}$$

$$E[\hat{\beta}_2] = \begin{cases} 0 + \frac{(p_{21}-p_{20})V_\pi\{\beta_\pi+2\beta_1(p_{11}-p_{10})\}}{p_{20}(1-p_{20})+(p_{21}-p_{20})(1-p_{21}-p_{20})E_\pi+(p_{21}-p_{20})^2(V_\pi+E_\pi-E_\pi^2)} & \text{(GWAS with Model 4.3)} \\ 0 + \frac{V_\pi\{\beta_\pi+2\beta_1(p_{11}-p_{10})\}}{V_\pi+E_\pi-E_\pi^2} & \text{(AMAP with Model 4.3),} \end{cases}$$

where E_π and V_π are the population mean and variance of the admixture proportions, β_π is the direct effect of admixture proportions on the trait, and p_{21}, p_{20} are the ancestral allele frequencies at SNP 2. From these results, we see that the unadjusted model will in general yield biased estimates of the effect size at both the causal and neutral SNPs, unless there is no ancestral heterogeneity (i.e., $V_\pi = 0$). Furthermore, the unadjusted model will yield biased

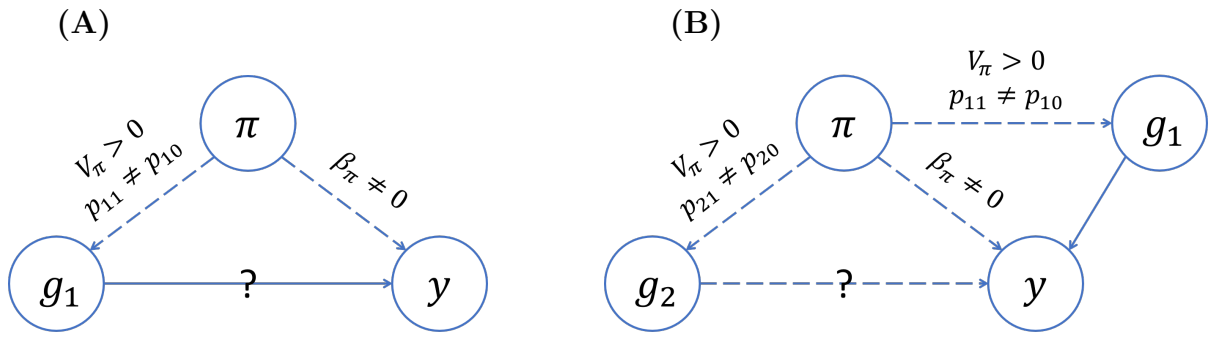

Figure 4.5: Conditions for confounding by global ancestry in GWAS.
(A) Global ancestry confounds the association at the causal locus (locus 1) if there is ancestral heterogeneity in the population ($V_\pi > 0$), the causal locus has different allele frequencies in the ancestral populations ($p_{11} \neq p_{10}$), and global ancestry has a direct effect on the trait ($\beta_\pi \neq 0$).
(B) Global ancestry can confound the association at an unlinked neutral locus (locus 2) even if global ancestry does not have a direct effect on the trait, provided that there is ancestral heterogeneity and the causal locus has different allele frequencies in the ancestral populations.

estimates of the effect size at the unlinked neutral locus (SNP 2) even if global ancestry does not have a direct effect on the trait (i.e., $\beta_\pi = 0$), unless the causal SNP (SNP 1) has the same frequency in the two ancestral populations (i.e., $p_{11} = p_{10}$). This bias away from zero of effect size estimates at the neutral locus will translate into spurious associations as sample sizes increase, as we saw in the WHI analyses (Figures 4.2A, 4.3A).

Proofs and simulations validating these analytic results are available in Appendix B.1. Our results provide insight into the scenarios under which global ancestry confounds the association between the trait and genotypes or local ancestry. These conditions can be summarized by the direct acyclic graphs (DAGs) in Figure 4.5 (GWAS) and Figure 4.6 (admixture mapping). Note that fewer conditions are required for confounding by global ancestry in admixture mapping studies compared to GWAS, and the magnitude of the bias of unadjusted admixture mapping models can be more substantial, as is reflected by the extreme inflation observed in the unadjusted WHI admixture mapping analysis (Figure 4.3A). However, ancestral heterogeneity still poses a problem for GWAS.

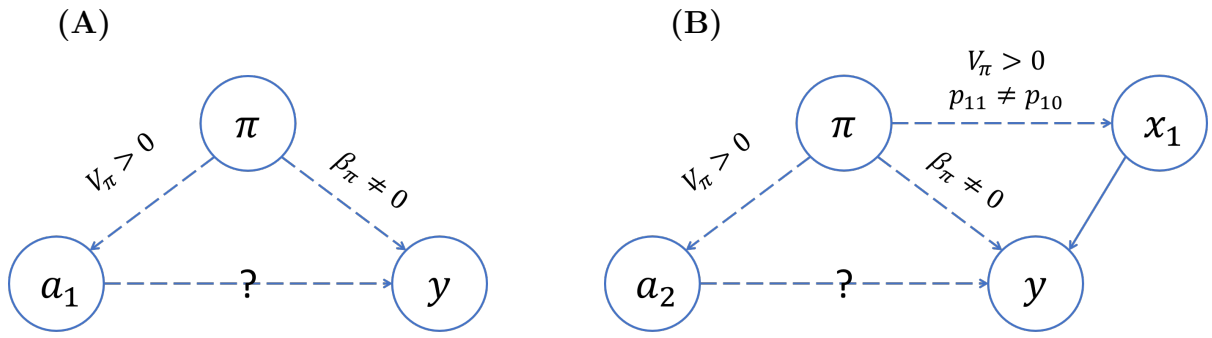

Figure 4.6: Conditions for confounding by global ancestry in admixture mapping.
(A) Global ancestry confounds the association at the causal locus (locus 1) if there is ancestral heterogeneity in the population ($V_\pi > 0$) and global ancestry has a direct effect on the trait ($\beta_\pi \neq 0$).
(B) Global ancestry can confound the association at an unlinked neutral locus (locus 2) even if global ancestry does not have a direct effect on the trait, provided that there is ancestral heterogeneity and the causal locus has different allele frequencies in the ancestral populations.

4.3.2 Comparison of ancestral heterogeneity adjustment techniques

In Section 4.3.1 we demonstrated the need to adjust for global ancestry in genetic association studies in ancestrally heterogeneous populations. In this section, we compare approaches that adjust for ancestral heterogeneity using either model-based estimates of admixture proportions (Model 4.1), one principal component (Model 4.2 with $P = 1$), or four principal components (Model 4.2 with $P = 4$). PCs were naively generated using all SNPs, or based on reduced subsets of SNPs after LD pruning and/or exclusion of potentially problematic regions (e.g., Table 4.1).

Adjusting for a single PC or admixture proportions performs similarly

In genetic association studies in WHI SHARe, models adjusting for model-based estimated admixture proportions performed nearly identically to models that adjusted for a single principal component. In Figures 4.2 and 4.3, the Manhattan plots from GWAS and admixture mapping models adjusting for admixture proportions (panel D) are indistinguishable from the Manhattan plots based on adjusting for a single PC (panels B and E). Furthermore,

the overall rates of spurious associations across all simulation studies are nearly identical for these models (Figure 4.4). This similar performance is observed regardless of whether PCs were generated using all SNPs or a reduced subset of SNPs. Comparing the estimated admixture proportions to the principal components for each individual, we see very high correlation between admixture proportions and the first PC (Figure 4.7), which explains the similar performance of genetic association studies adjusting for these covariates.

Adjusting for extra PCs can induce spurious associations

In Figure 4.7 we see that only the first PC is highly correlated with estimated admixture proportions in the WHI SHARe African American cohort. However, in practice it is very common that investigators will include additional PCs in their regression models just to be sure that they have fully captured global ancestry, recognizing that this may incur a small loss in power [37] but should not otherwise cause any problems. In fact, in prior analyses using these same WHI SHARe data, authors adjusted for two [139], four [29, 57], or even ten [52, 140] PCs in their genetic association studies.

Surprisingly, in our analysis of WHI SHARe data we see that adjusting for additional PCs can actually *induce* spurious associations in genetic association studies. For example, comparing GWAS and admixture mapping models that adjust for naively generated PCs (i.e., PCs that were calculated using all SNPs), we see a spurious association that appears on chromosome 6 when we include 4 PCs in the model (Figures 4.2C and 4.3C), but that spurious signal disappears if we only adjust for the first PC (Figures 4.2B and 4.3B). Of particular interest is the fact that both the true causal SNP (on chromosome 4) and the spurious signal (on chromosome 6) are located in regions of the genome that are highly correlated with the naively generated PCs (Figure 4.8, first column). This unusual behavior is not unique to this simulation setting. From Figure 4.8 we see that there are in fact many regions of the genome that are highly correlated with PCs 2–4, particularly when PCs are naively generated using all available SNPs. When the causal SNP is located in one of these regions, GWAS and admixture mapping models that adjust for naively generated PCs have

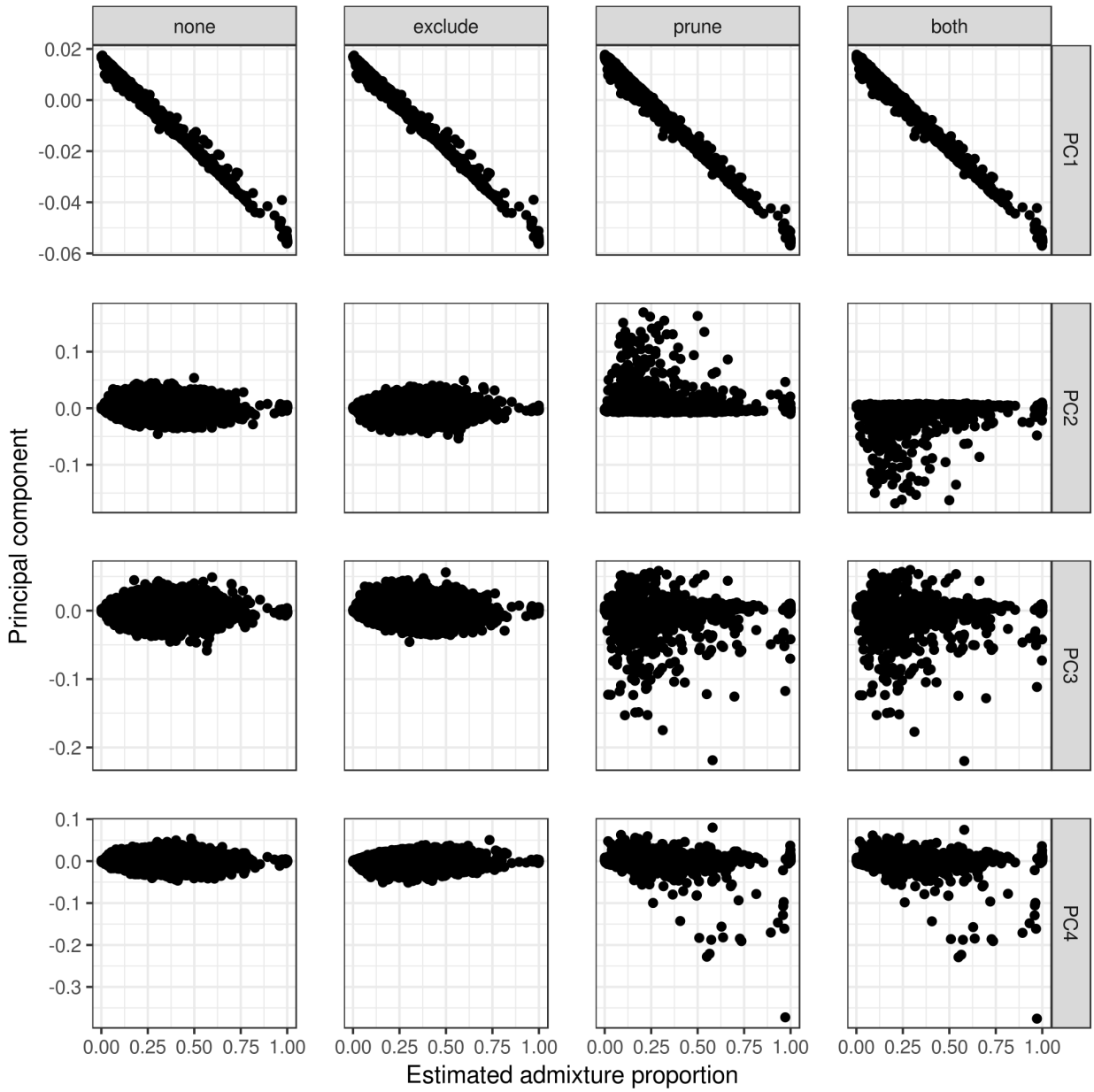

Figure 4.7: Comparison of estimated admixture proportions and PCs in WHI SHARe African Americans.

Each panel plots estimated admixture proportions (x-axis) versus principal components (y-axis). Panels are stratified according to which PC is being plotted on the y-axis (1, 2, 3, or 4) and what level of filtering was applied prior to running PCA: *none* (all SNPs), *exclude* (after excluding regions in Table 4.1), *prune* (after LD pruning with $r^2 = 0.1$ and window size = 0.5 Mb), or *both* (after exclusions and LD pruning).

considerably elevated rates of spurious associations when they include four PCs compared to when they include only the first PC (Figure 4.4, first column). However, if the causal SNP is located in a region of the genome that is not highly correlated with any of the PCs included in our model, then there is little difference between models that adjust for one or four PCs (Figure 4.4, second and third columns).

To understand this behavior, we derived the expected effect sizes from genetic association studies adjusting for principal components in an admixed population with two ancestral populations. We assume that only the first PC is needed to capture global ancestry in this population (i.e., $\mathbf{v}_1 \approx \boldsymbol{\pi}$), but we also include a second principal component in our model. Suppose that the second PC captures some feature other than global ancestry (i.e., $\mathbf{v}_2 = \mathbf{z}$ for some variable z) and the quantitative trait depends on genotype only at a single locus, as in our WHI simulations. At the causal locus (SNP 1) and an unlinked neutral locus (SNP 2), we can show that the expected effect size estimates from GWAS and admixture mapping models that include two principal components (i.e., Model 4.2 with $P = 2$) are

$$E[\hat{\beta}_1] = \begin{cases} \beta_1 & \text{(GWAS)} \\ \beta_1(p_{11} - p_{10}) + \frac{\beta_1 V_\pi E\{\mathrm{Cov}(a_1,z|\pi)\}[E\{\mathrm{Cov}(x_1,z|\pi)\} - (p_{11}-p_{10})E\{\mathrm{Cov}(a_1,z|\pi)\}]}{V_z(V_\pi V_{a_1} - C^2_{a_1,\pi}) - V_\pi C^2_{a_1,z} + C_{\pi,z}(2C_{a_1,\pi}C_{a_1,z} - V_{a_1}C_{\pi,z})} & \text{(AMAP)} \end{cases}$$

$$E[\hat{\beta}_2] = \begin{cases} 0 + \beta_1 \frac{-V_\pi E\{\mathrm{Cov}(x_1,z|\pi)\}E\{\mathrm{Cov}(x_2,z|\pi)\}}{V_z(V_\pi V_{x_2} - C^2_{x_2,\pi}) - V_\pi C^2_{x_2,z} + C_{\pi,z}(2C_{x_2,\pi}C_{x_2,z} - V_{x_2}C_{\pi,z})} & \text{(GWAS)} \\ 0 + \beta_1 \frac{-V_\pi E\{\mathrm{Cov}(x_1,z|\pi)\}E\{\mathrm{Cov}(a_2,z|\pi)\}}{V_z(V_\pi V_{a_2} - C^2_{a_2,\pi}) - V_\pi C^2_{a_2,z} + C_{\pi,z}(2C_{a_2,\pi}C_{a_2,z} - V_{a_2}C_{\pi,z})} & \text{(AMAP)} \end{cases}$$

where $V_a = \mathrm{Var}(a)$ and $C_{a,b} = \mathrm{Cov}(a, b)$. At the causal SNP, the estimated effect size will always be unbiased using GWAS models that adjust for two PCs, but this is not in general true for admixture mapping. Furthermore, the estimated effect sizes from both GWAS and admixture mapping will be biased away from the zero at the unlinked neutral locus, unless there is no ancestral heterogeneity ($V_\pi = 0$), the second PC is not correlated with genotype at the causal SNP, or the second PC is not correlated with genotype (GWAS) or local ancestry (admixture mapping) at the tested neutral SNP. In other words, these results indicate that if a model adjusts for a PC that captures genotype at the causal SNP as well as a second SNP that is not associated with the disease, then spurious associations will arise at that

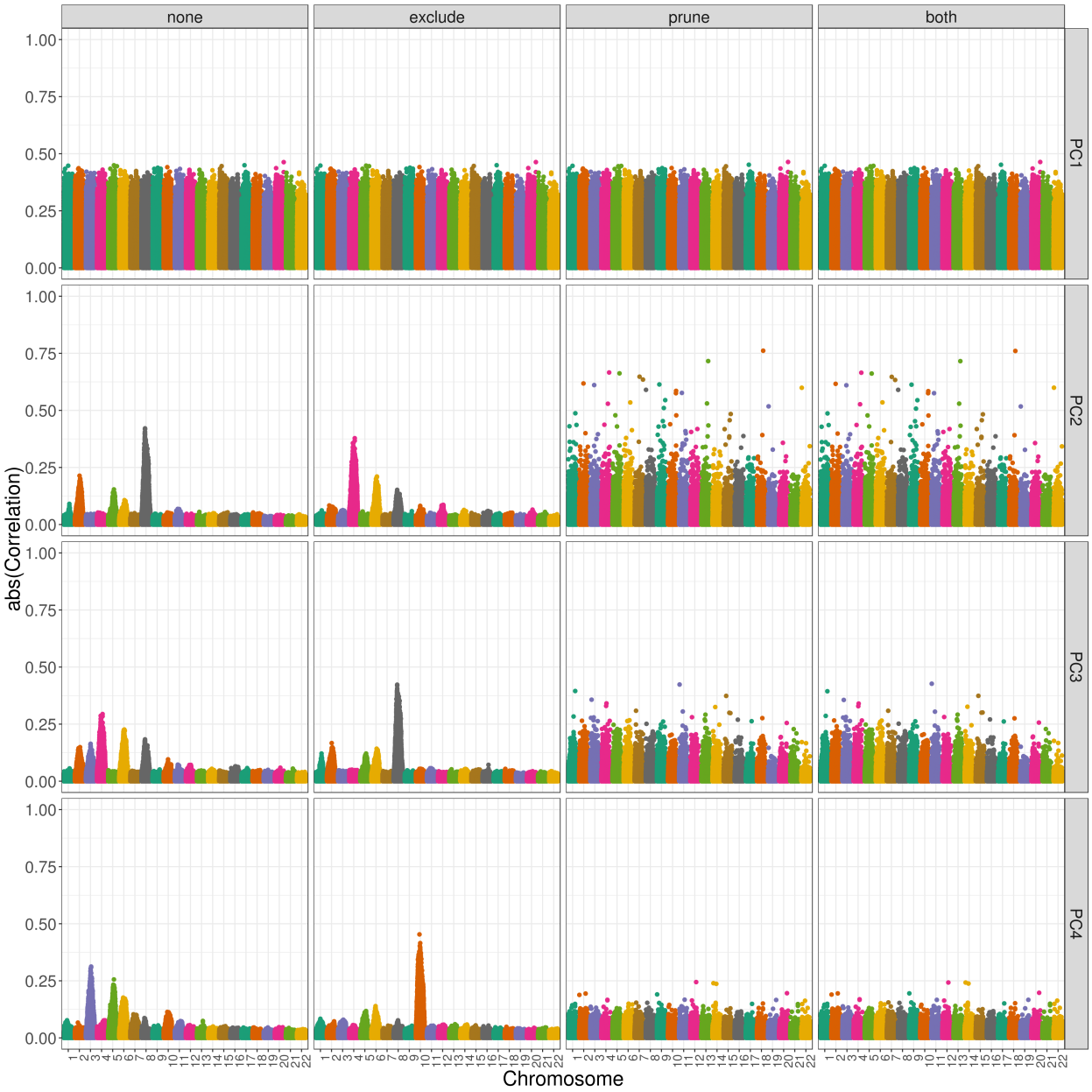

Figure 4.8: Correlation between PCs and genotypes in WHI SHARe African Americans. Each panel plots the absolute value of the correlation (y-axis) between principal components and genotypes at each position along the genome (x-axis). Panels are stratified according to which PC is being investigated (1, 2, 3, or 4) and what level of filtering was applied prior to running PCA: *none* (all SNPs), *exclude* (after excluding regions in Table 4.1), *prune* (after LD pruning with $r^2 = 0.1$ and window size = 0.5 Mb), or *both* (after exclusions and LD pruning).

neutral SNP in large enough samples. This is precisely what we observe in the WHI analyses (Figures 4.2, 4.3). Furthermore, our analytic results demonstrate that if the extra PC does not capture genotype at the causal SNP, then no spurious associations will arise in GWAS or admixture mapping studies (as we saw in Figure 4.4, second and third columns), although the estimated effect size at the causal SNP may still be biased in an admixture mapping study. Proofs and simulations validating these analytic results are available in Appendix B.1

We have seen that spurious associations can be induced by including extra PCs in a genetic association study, particularly when those extra PCs are driven by small regions of the genome rather than global ancestry or other genome-wide features. This suggests, then, that we could avoid some of these problems by pre-processing genotypes, prior to running PCA, such that PCs will no longer be driven by small regions of the genome. Many authors have previously suggested that PCA can be sensitive to LD among SNPs, and they recommend that SNPs in high LD with one another be removed prior to running PCA [141, 129, 68, 130, 131, 142]. This filtering could be accomplished via LD pruning and/or by excluding regions of the genome that have been previously shown to have high or long-range LD patterns (Table 4.1). In the WHI SHARe data, we see that simply excluding SNPs failling into the regions listed in Table 4.1 does not address the problems that we are seeing: PCs 2–4 are still strongly correlated with small regions of the genome (Figure 4.8, second column) and models that adjust for four of these PCs have similar rates of spurious associations to models adjusting for four naively generated PCs (Figure 4.4). However, if we run PCA after first performing LD pruning with an r^2 threshold of 0.1 (which is stricter than the default threshold for many software programs), then our top PCs no longer capture small regions of the genome (Figure 4.8, third and fourth columns). Furthermore, if we use these PCs generated after LD pruning in our genetic association studies, then the spurious association observed on chromosome 6 disappears (Figure 4.2F, 4.3F) and the overall rate of spurious associations across all simulation settings is dramatically reduced (Figure 4.4), regardless of whether one or four PCs are included in the model.

4.4 Discussion

In this chapter, we investigated the impact of ancestral heterogeneity on genetic association studies in admixed populations. Through simulation studies and analytic results, we demonstrated that we must adjust for ancestral heterogeneity in genetic association studies in order to avoid spurious associations. In addition, we showed that adjusting for ancestral heterogeneity using principal component analysis (PCA) can actually *induce* spurious associations in genetic association studies. These results are particularly concerning given the current wide-spread use of PCA to control for ancestral heterogeneity in genetic association studies.

Although it is generally understood that adjusting for ancestral heterogeneity is important for GWAS, the conditions under which this adjustment is required are not fully understood. It is often assumed that global ancestry is a confounding variable in GWAS only when it has a direct effect on the trait (Figure 4.1). However, our analytic and simulation results in Section 4.3.1 show that adjusting for global ancestry is necessary even when it does not have a direct effect on the trait (e.g., through environmental differences across ancestral groups), provided that there is a SNP that is truly associated with the trait that has different allele frequencies across the ancestral populations of interest. This fact has been recognized previously (e.g., [143]), but it seems that it has been forgotten by many, and to our knowledge no other group has provided analytic results such as ours that explicitly demonstrate the factors that impact the magnitude of the bias incurred by GWAS models that fail to adjust for global ancestry. There has also been less attention paid in the literature to the issues posed by ancestral heterogeneity to admixture mapping studies. Our results, however, have shown that admixture mapping is just as sensitive as GWAS (if not more) to ancestral heterogeneity. We hope that our results will serve as a reminder to researchers of the various ways in which global ancestry can confound genetic association studies and the care that needs to be taken to ensure that both GWAS and admixture mapping studies appropriately adjust for ancestral heterogeneity.

In Section 4.3.2, we show that when adjusting for ancestral heterogeneity using principal components, it is crucial to ensure that only PCs that capture global ancestry are used. If additional PCs that capture features other than global ancestry are included in the model, this can induce spurious associations, particularly when the PCs are correlated with genotype or local ancestry at a select number of SNPs. This phenomenon can be explained by the concept of *collider bias* (Figure 4.9). Suppose that a PC captures genotypes or local ancestry at two SNPs rather than global ancestry, and one of these SNPs (SNP 1) is associated with the trait. Then, this PC is by definition a *collider variable*, and adjusting for the PC will induce a spurious association between the trait and the neutral SNP (SNP 2) [71]. Prior work has shown that adjusting for heritable covariates (e.g., height, body mass index) can induce collider bias in GWAS [144, 145], and very recent work showed that principal components can induce collider bias in gene expression studies [146], but the issues posed by PCs to GWAS and admixture mapping studies have not been previously demonstrated.

In our analysis of genotype data from 8,064 unrelated WHI SHARe African Americans, we found that all but the first principal component were largely driven by small regions of the genome—and thus have the potential to be collider variables—unless careful pre-processing of genotype data was preformed prior to running PCA. Previous studies have found that PCs can capture small regions of the genome, and have suggested that these regions be excluded (see Table 4.1) and/or that LD pruning be performed prior to running PCA [141, 129, 68, 130, 131, 142]. However, the motivation for this LD-based filtering was framed in terms of the ability of the principal components to capture global ancestry and the computational complexity of running PCA, rather than the downstream implications on association testing. Our work has shown that the downstream implications are of great concern. Furthermore, we have found that excluding the regions listed in Table 4.1 does not actually resolve any issues in the WHI SHARe data, that identifying and removing potentially problematic regions based on our own data is tedious and does not provide any benefit beyond LD pruning (see Appendix B.2), and that a stricter threshold ($r^2 = 0.1$) is needed for LD pruning than is often suggested in the literature ($r^2 = 0.2$) (see Appendix

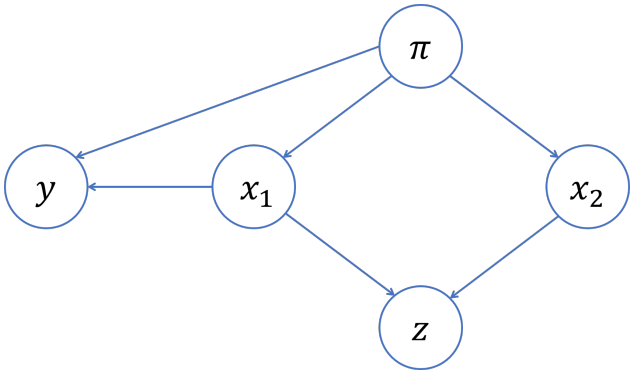

Figure 4.9: Collider bias in genetic association studies adjusting for PCs.
Consider an ancestrally heterogeneous population, where admixture proportions π are associated with genotype or local ancestry at two SNPs x_1, x_2. Suppose only the first SNP is causal, and that the trait y may also be directly affected by admixture proportions. Let z represent a principal component that captures genotypes/local ancestry at these SNPs rather than global ancestry. Then z is a *collider variable*, and conditioning on z will induce a spurious association between the neutral SNP (x_2) and the trait.

B.2). The vast majority of previous recommendations were based on studies in European populations, but LD patterns can be very different in admixed populations. In particular, LD typically extends much further (even across chromosomes) in admixed populations [147, 148], so it makes sense that stricter levels of LD-based filtering would be required in admixed populations.

The patterns that we have observed in our analysis of genotype data from WHI SHARe African Americans are not unique to these data. Results are not shown here, but we have observed similar patterns in analysis of sequence data for African American and Hispanic/Latino individuals from the Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project. Our work demonstrates the problems that can arise if principal component analysis is used to control for ancestral heterogeneity in admixed populations. In analysis of the WHI SHARe data, we found that strict LD pruning could resolve these issues, or that issues could be avoided altogether by simply adjusting for model-based estimates of admixture proportions rather than principal components. For populations such as African Americans where we have a good idea of the number of ancestral populations of

interest and relevant reference panel data is readily available, we suggest that genetic association studies adjust for estimated admixture proportions rather than principal components. If investigators feel strongly that PCA is needed to capture relevant levels of ancestral heterogeneity, then great care must be taken to address the issues discussed in this chapter. In particular, we suggest careful pre-processing of data prior to running PCA, combined with thorough diagnostics (e.g., by calculating and carefully examining the correlation between PCs and genotypes via plots like Figure 4.8), to ensure that models do not include principal components that could induce spurious associations.

Chapter 5

ANCESTRY INFERENCE AND GENETIC ASSOCIATION TESTING IN THE TRANS-OMICS FOR PRECISION MEDICINE WHOLE GENOME SEQUENCING PROJECT

5.1 Introduction

The kidney plays a vital role in the human body. Chronic kidney disease (CKD), characterized by low kidney function, is a precursor for end-stage renal disease and an important risk factor for other diseases, including cardiovascular disease, as well as early death [21]. Recent reports estimate that as many as 15% of adults across the United States—37 million people—have CKD [72], with higher prevalence of CKD among African Americans and Hispanics/Latinos compared to individuals of European ancestry [21, 22]. The difference in disease prevalence across these ancestral groups has been attributed to environmental differences, as well as genetic factors such as the increased frequency of *APOL1* risk variants among individuals with African ancestry [22].

The observed difference in chronic kidney disease prevalence across ancestral groups makes it an ideal candidate for admixture mapping studies. By looking for associations between the disease and local ancestry across the genome, admixture mapping studies search for causal genetic variants that differ in frequency across ancestral groups and drive, at least in part, the observed phenotypic differences across populations. Previous admixture mapping studies have been successful in identifying genetic variants that are more frequent in admixed populations and putatively involved in kidney function and disease [25, 73]. Furthermore, an ongoing genome-wide association study of kidney traits in a large multi-ethnic cohort from the Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project has also identified genome-wide significant variants that are more frequent among,

or even exclusive to, admixed subjects [Lin and Franceschini, personal communication].

In this chapter, we build on this prior work and conduct admixture mapping studies of kidney traits in whole genome sequence data, focusing our investigation on African American and Hispanic/Latino TOPMed subjects in hopes of gaining more insight into the genetic architecture of kidney disease in these admixed populations. Whole genome sequences and kidney phenotypes (quantitative levels of serum creatinine and estimated glomerular filtration rate (eGFR), as well as a binary indicator of chronic kidney disease) were available for 9,479 admixed individuals through the TOPMed whole genome sequencing project. We inferred autosomal local ancestry for these individuals using `RFMix`, and then used that inferred local ancestry to conduct genome-wide admixture mapping studies, control for ancestral heterogeneity in association mapping, and estimate ancestry-specific allele frequencies for candidate variants of interest. Our analyses highlight the potential for success, as well as the existing challenges, for local ancestry inference and genetic association studies in admixed populations using whole genome sequence data.

5.2 Methods

5.2.1 The Trans-Omics for Precision Medicine (TOPMed) Data

TOPMed Whole Genome Sequencing Project

The Trans-Omics for Precision Medicine (TOPMed) whole genome sequencing project, sponsored by the National Heart, Lung, and Blood Institute (NHLBI), is an ongoing project working toward the collection of whole-genome sequences, other omics data, and rich phenotypic information for over 100,000 individuals from diverse backgrounds. Samples and phenotypic data are contributed from a number of pre-existing NHLBI-funded studies, and whole genome sequencing (WGS) of all samples is in progress. In TOPMed *freeze 5b*, WGS data are available for approximately 55,000 samples with European, African, Hispanic/Latino, Asian, or other ancestry/ethnicity.

Sequencing and quality control

High coverage WGS was performed by several sequencing centers—Baylor College of Medicine Human Genome Sequencing Center, Broad Institute, Illumina, Macrogen Corp., McDonnell Genome Institute, New York Genome Center, and Northwest Genomics Center—using DNA from blood, PCR-free library construction, and Illumina HiSeq X technology. The average sequencing depth was 38X. Variant discovery and genotype calling for all samples was performed by the Informatics Research Center using the `GotCloud` pipeline [149]. For variant-level quality control (QC), a support vector machine filter was trained using known variants (positive controls) and variants with Mendelian-inconsistencies (negative controls). Sample-level QC was performed by the TOPMed data coordinating center and included additional checks for pedigree errors, sex discrepancies, and genotype concordance. Phasing was performed using Eagle 2.4 [150]. Additional details are reported in [151] and on the TOPMed website: `https://www.nhlbiwgs.org/data-sets`.

After filtering, 438 million single nucleotide variants (SNVs) and 33 million insertions/deletions remained. In *freeze 5b* of the TOPMed sequencing project, statistically phased whole genome sequences were available for 54,035 individuals from 31 contributing studies. A subset of 20,048 of these individuals self-identify as African American, African Caribbean, or Hispanic/Latino.

Kidney function phenotypes and covariates

Demographic data and kidney phenotypes were collected from all studies using standardized protocols. Demographic information included sex, age, and self-identified race/ethnicity. Serum creatinine levels were collected from each study, along with information on the serum creatinine assay and year of the assay to allow for calibration across studies. Based on serum creatinine and demographic information, we calculated estimated glomerular filteration rate (eGFR) using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation [152]. A binary indicator of chronic kidney disease (CKD) was generated by identifying sub-

Table 5.1: Number of African American and Hispanic/Latino subjects in TOPMed freeze 5b.
Sample sizes are stratified by race/ethnic group and the name of the contributing study.

	ARIC	GeneSTAR	GENOA	HyperGEN	JHS	MESA	WHI
African American	195	190	1010	1734	3005	960	1209
Hispanic/Latino	0	0	0	0	0	924	252

jects with eGFR < 60. After removing duplicate entries for subjects appearing in more than one study, phenotypes were available for a total of 8,303 African Americans and 1,176 Hispanics/Latinos from seven studies: Atherosclerosis Risk in Communities (ARIC), Genetic Study of Atherosclerosis Risk (GeneSTAR), Genetic Epidemiology Network of Ateriopathy (GENOA), Hypertension Genetic Epidemiology Network (HyperGEN), Jackson Heart Study (JHS), Multi-Ethnic Study of Atherosclerosis (MESA), and Women's Health Initiative (WHI). Table 5.1 summarizes the number of samples from each study. Additional details about processing of phenotype data are available in Appendix C.1.

5.2.2 Statistical Methods

Inferring ancestry and genetic relatedness

We inferred local ancestry for all 20,048 TOPMed subjects who self-identified as African American, African Caribbean, or Hispanic/Latino. The reference panel for local ancestry inference included 37 African, 35 European, and 20 Native American individuals with phased sequence data (for chromosomes 1–22) available from the Simons Genome Diversity Project (SGDP) [153]. For a full list of SGDP populations included in our reference panel, see Appendix C.2.1. We used `liftOver` [154] to update the SGDP data to build 38 and identified the set of bi-allelic variants present in both the reference panel and TOPMed. After removing very low frequency variants (minor allele count < 2 in SGDP or < 5 in TOPMed), a total of 9,137,968 autosomal variants remained for this analysis. We used the HapMap genetic map

[113], lifted over to build 38, to estimate genetic positions for each variant.

For each admixed TOPMed sample, we used `RFMix` version 1.5.4 [10] to infer the number of alleles inherited from each ancestral population (African, European, and Native American) at each locus. We ran `RFMix` using the suggested input parameters, with the exception of a smaller window size, 0.1 centimorgans (cM), to assist with computational costs of analyzing this large dataset, and a choice of 6, 8, or 10 generations since admixture for the African American, African Caribbean, and Hispanic/Latino samples, respectively, to reflect estimates from previous studies [155, 59]. Example `RFMix` commands are available in Appendix C.2.2. To estimate admixture proportions for each individual, we calculated the genome-wide average local ancestry.

We assessed the quality of our local ancestry calls by comparing our estimated admixture proportions to those generated on the same data by another group, and by looking across the genome for abnormal deviations in the proportion of local ancestry calls assigned to each ancestral population, maximum posterior probabilities generated by `RFMix`, or location of local ancestry segment breakpoints. We did not identify any notable quality issues. Details are available in Appendix C.2.3.

We used the iterative procedure suggested by Conomos et al. [112] to estimate kinship coefficients adjusted for population structure and admixture. In the final step of this procedure, we provided `PC-Relate` with estimated admixture proportions in place of principal components (PCs) since we planned to adjust for admixture proportions rather than PCs in our regression model. We used these estimated kinship coefficients to adjust for relatedness in our admixture mapping and genome-wide association mapping analyses.

Association mapping of kidney traits in large multi-ethnic cohort

In a concurrent study, our collaborators have conducted a large whole genome sequence-based analysis investigating the association between eGFR and genotypes among all TOPMed subjects with available kidney phenotype information. This multi-ethnic cohort includes 9,479 admixed individuals, as well as an additional 13,983 subjects of European and East

Asian ancestry. Detailed methods and results for this analysis will be reported elsewhere.

Estimating ancestry-specific allele frequencies

We used our local ancestry calls to estimate ancestry-specific allele frequencies for loci of interest. In particular, we estimated European, African, and Native American allele frequencies for four variants that were identified by our collaborators in their large multi-ethnic association study [Lin and Franceschini, personal communication].

To estimate ancestry-specific allele frequencies, we first filled in missing local ancestry calls at loci of interest. `RFMix` infers local ancestry only at bi-allelic variants that are present in both the admixed sequence data and the reference panel, so inferred local ancestry may not be available at all loci. However, given that local ancestry segments extend over multiple loci and we inferred local ancestry within 0.1 cM windows using `RFMix`, we were able to fill in the missing local ancestry calls at loci of interest with reasonable confidence by looking at the inferred local ancestry at neighboring loci. For a given locus, allele, and ancestral population of interest, we then calculated the frequency of the allele among TOPMed haplotypes with local ancestry assigned to that ancestral population at that locus. To account for uncertainty in the phase of genotypes relative to the local ancestry calls (particularly at loci where local ancestry was not inferred directly by `RFMix`), we used the expectation-maximization (EM) algorithm approach implemented in `ASAFE` [74]. We ran `ASAFE` using all 20,048 admixed individuals in TOPMed with local ancestry calls, including the 9,479 subjects in our admixture mapping and association mapping analyses, as well as additional African American, African Caribbean, and Hispanic/Latino subjects that were excluded from these association studies due to missing kidney phenotype data.

Genetic association studies of kidney traits in admixed subjects

We used `GENESIS` [156] to perform genome-wide admixture mapping studies using all 9,479 TOPMed admixed subjects with available kidney phenotype data. At each locus, we fit linear mixed models investigating each ancestral group (African, European, and Native

American) separately. In our primary analysis, the outcome variable was eGFR and we adjusted for sex, age, a combined indicator of study and race/ethnic group (e.g., JHS African Americans, MESA Hispanics/Latinos), and admixture proportions as fixed effects. To account for relatedness across samples, we included ancestry-adjusted kinship estimates as a random effect, allowing for heterogeneous variance within groups defined by study and race/ethnicity. We performed an inverse normal transformation of the trait [118] within these same study/race groups. To account for multiple testing, we used the genome-wide p-value threshold 5.4×10^{-6}. We estimated this significance threshold using the test statistic simulation approach proposed in Chapter 3 and implemented in our R package `STEAM` [59]. This multiple testing correction approach requires as input an estimate of the number of generations since admixture. We used the non-linear least squares regression approach implemented in `STEAM` to estimate this value. For more details, see Appendix C.3.

As secondary analyses, we also performed admixture mapping analyses separately in the African American and Hispanic/Latino subjects, testing only the African ancestral component in the African American cohort. Significance thresholds in the African American and Hispanic/Latino subsets were estimated using `STEAM` to be 1.6×10^{-5} and 3.5×10^{-6}, respectively. In addition to investigating eGFR, we also performed admixture mapping analyses using serum creatinine or chronic kidney disease (CKD) as the outcome. We used the same significance threshold for the serum creatinine and CKD analyses as we did for the eGFR analysis. Finally, to complement our admixture mapping analyses, we also implemented a genome-wide association study investigating eGFR in our reduced subset of admixed individuals. We performed association mapping using the same linear mixed model as in the admixture mapping analyses described above, with the only difference being that we replaced local ancestry with genotype as the predictor of interest.

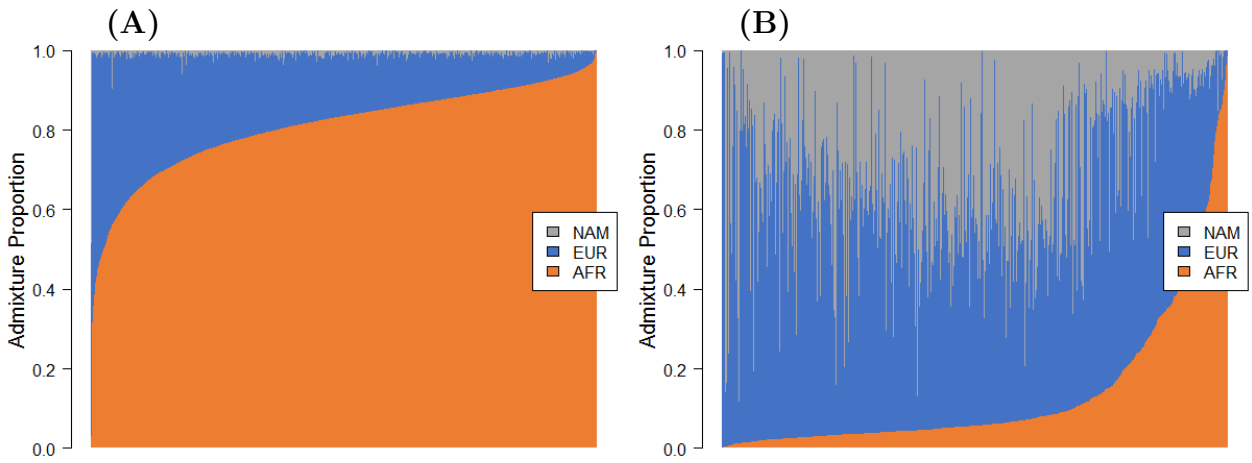

Figure 5.1: Barplots of estimated admixture proportions in admixed TOPMed samples. (A) Estimated admixture proportions for African Americans included in kidney association studies. (B) Estimated admixture proportions for Hispanics/Latinos included in kidney association studies.

5.3 Results

5.3.1 Characteristics of TOPMed admixed subjects

Local ancestry calls and complete phenotype data were available for 9,479 admixed individuals. The majority of these subjects self-identified as African American ($n = 8,303$). Subject characteristics are summarized in Table 5.2. Only 7% of subjects ($n = 656$) had low enough eGFR to be classified as having chronic kidney disease. On average, subjects with CKD tended to be older and had slightly more African ancestry and less European and Native American ancestry.

5.3.2 Population structure and relatedness in TOPMed admixed samples

We observed considerable heterogeneity in the estimated admixture proportions across samples (Figure 5.1). We also observed close relationships between many of our samples (Figure 5.2). These findings confirm the need to adjust for both population structure and relatedness in our admixture mapping and association mapping analyses.

Table 5.2: Characteristics of TOPMed admixed subjects.
Summaries are presented across all samples, as well as stratified by chronic kidney disease status (eGFR less than or greater than 60). For categorical variables, we present the proportion of subjects in each category. For continuous variables, we present the mean (standard deviation, minimum–maximum).

	All ($n = 9479$)	eGFR < 60 ($n = 656$)	eGFR ≥ 60 ($n = 8823$)
Admixture Prop.			
African	0.72 (0.26, 0.00–1.00)	0.76 (0.22, 0.00–0.98)	0.72 (0.26, 0.00–1.00)
European	0.23 (0.18, 0.00–0.99)	0.21 (0.17, 0.02–0.99)	0.23 (0.18, 0.00–0.99)
Native American	0.05 (0.12, 0.00–1.00)	0.03 (0.09, 0.00–0.71)	0.05 (0.13, 0.00–1.00)
Age (years)	56.5 (12.4, 18–94)	66.8 (9.6, 25.6–94)	55.7 (12.3, 18–91)
Sex			
Male	33.1%	30.8%	33.3%
Female	66.9%	69.2%	66.7%
Study & Race			
African American			
ARIC	2.1%	1.1%	2.1%
GeneSTAR	2.0%	0.2%	2.1%
GENOA	10.7%	17.2%	10.2%
HyperGEN	18.3 %	20.3%	18.1%
JHS	31.7%	29.7%	31.8%
MESA	10.1%	9.1%	10.2%
WHI	12.8%	13.9%	12.7%
Hispanic/Latino			
MESA	9.7%	6.1%	10.0%
WHI	2.7%	2.4%	2.7%

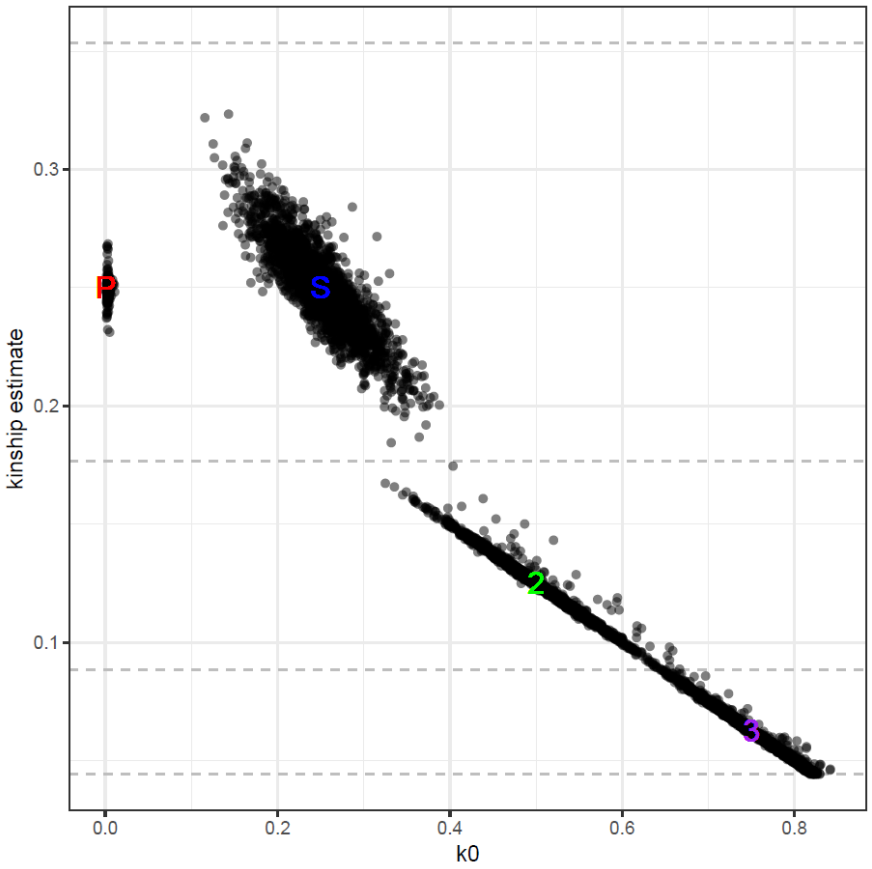

Figure 5.2: Inferred genetic relatedness in admixed TOPMed samples.
Scatterplot of estimated kinship coefficients versus estimated probabilities of sharing zero alleles identical by descent ($k0$) for pairs of TOPMed admixed samples included in kidney association studies. Expected coordinates for parent/offspring (P), full siblings (S), second degree relatives (2) and third degree relatives (3) are included for reference. Pairs of samples inferred to be more distantly related are excluded from this figure.

5.3.3 Ancestry-specific allele frequency estimates for variants of interest from a large multi-ethnic association study

In a concurrent association study, investigators identified a number of genetic variants that are significantly associated with eGFR in a large sample of African American, East Asian, European American, and Hispanic/Latino individuals from the TOPMed Whole Genome Sequencing Project [Lin and Franceschini, personal communication]. Here, we focus on four variants of interest from this multi-ethnic association study: *rs539182790* on chromosome 19, *rs190658489* on chromosome 6, *rs149589493* on chromosome 11, and *rs78902137* on chromosome 17. The first of these variants was significantly associated with eGFR in single variant tests; the remaining three were driving variants in top genes from gene-based testing using `SKAT` [157]. All four variants are of particular interest for our work given that they have been observed to be more frequent or exclusively present in TOPMed admixed populations, as compared to TOPMed Europeans and Asians [Lin and Franceschini, personal communication]. We used our local ancestry calls to estimate ancestry-specific allele frequencies for these variants of interest. These allele frequency estimates provide additional insight into the association findings and help motivate replication studies. Results are summarized in Table 5.3.

Of particular interest is rs539182790, an indel on chromosome 19 that was identified as genome-wide significant in single variant tests in the combined multi-ethnic TOPMed whole genome sequence-based analysis. This variant is low frequency (alternate allele frequency = 0.02) in the 1000 Genomes Admixed Americans *AMR* population—individuals with Mexican ancestry from Los Angeles USA, Puerto Ricans from Puerto Rico, Colombians from Medellin, Colombia and Peruvians from Lima, Peru—and non-existent (i.e., monomorphic) in the 1000 Genomes African, East Asian, European, and South Asian populations [158]. In TOPMed, the variant is mostly present in Hispanics/Latinos [Lin and Franceschini, personal communication]. Using our local ancestry calls, we estimated that this variant is low frequency in the Native American ancestral population (alternate allele frequency = 0.026) but

Table 5.3: Estimated ancestry-specific allele frequencies for variants identified in single variant or gene-based (SKAT [157]) tests in WGS analyses using all admixed, European American, and Asian American TOPMed samples.
List of variants of interest courtesy of Lin and Franceschini [personal communication]. Ancestry-specific allele frequencies were estimated using `ASAFE` on all 20,048 admixed TOPMed samples. Physical positions for each variant are based on build hg38.

			Allele Frequencies		
Chr:position	rsID	Test	African	European	Native American
6:44376382	rs190658489	SKAT	0.0005	3.2×10^{-8}	1.1×10^{-7}
11:102593550	rs149589493	SKAT	0.0067	9.2×10^{-9}	2.3×10^{-13}
17:82056475	rs78902137	SKAT	0.0265	9.0×10^{-9}	2.2×10^{-19}
19:3799817	rs539182790	Single	2.0×10^{-11}	0.0004	0.0257

essentially non-existent in African and European populations. These findings suggest that we might look for a Native American or a Hispanic/Latino population with larger amounts of Native American ancestry for replicating this association finding.

Looking at the estimated ancestry-specific allele frequencies for the variants driving the top gene-based test results (Table 5.3), we note that the three variants on chromosomes 6, 11, and 16 are low frequency (rs78902137) or rare (rs149589493, rs190658489) in the African ancestral population but non-existent in European and Native American populations. Again, these results are useful for planning replication studies, suggesting that we should look to replicate these signals in either African or African American populations.

5.3.4 Admixture mapping and WGS association mapping in TOPMed admixed subjects

For our primary analysis, we conducted a genome-wide admixture mapping study looking for an association between eGFR and African, European, or Native American local ancestry in a smaller subset of TOPMed admixed individuals. No regions of the genome reached genome-wide significance (Figure 5.3).

In secondary analyses, we repeated this admixture mapping analysis using just the African

Figure 5.3: Manhattan plots for eGFR admixture mapping analysis using 9,479 African American and Hispanic/Latino individuals.
Admixture mapping p-values at each variant for eGFR versus African local ancestry (top panel; $\lambda = 1.082$), European local ancestry (middle panel; $\lambda = 0.992$), and Native American local ancestry (bottom panel; $\lambda = 0.954$)

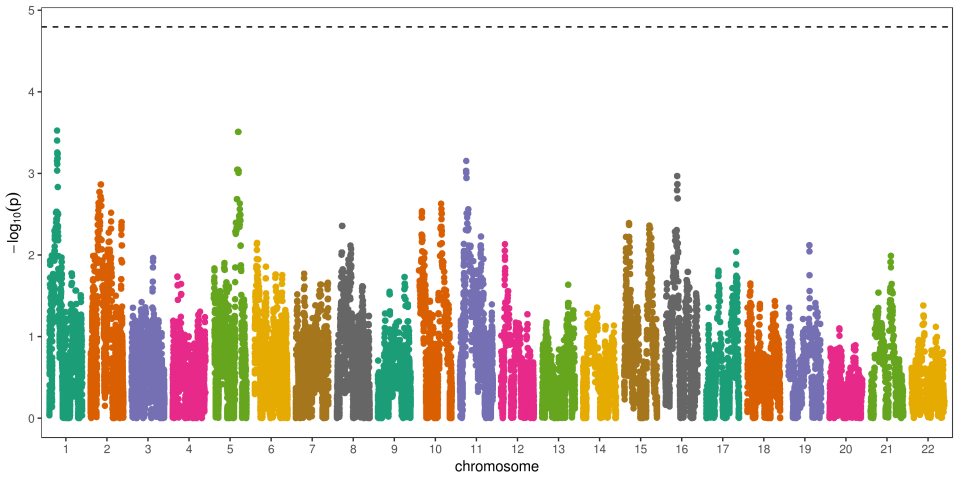

Figure 5.4: Manhattan plots for eGFR admixture mapping analysis using 8,303 African American individuals.
Admixture mapping p-values at each variant for eGFR versus African local ancestry ($\lambda = 1.022$).

American (Figure 5.4) or Hispanic/Latino (Figure 5.5) samples. Again, no genome-wide significant results were found, although there is a peak on chromosome 10 that is approaching genome-wide significance in the analysis of European local ancestry in the Hispanic/Latino subjects. Finally, we conducted an admixture mapping study using all samples and serum creatinine (Figure 5.6) and chronic kidney disease (Figure 5.7) as our trait. As with eGFR, there are no significant findings to report. Serum creatinine and CKD are strongly related to eGFR, and CKD has a low prevalance in our data, so it is not surprising that we also failed to identify genome-wide significant associations in these secondary admixture mapping analyses.

We also implemented a whole genome sequence-based association study of eGFR in the same subset of admixed individuals (Figure 5.8). One locus on chromosome 6 reaches genome-wide significance using a p-value threshold of 5×10^{-9}. This significance threshold has been suggested for whole genome sequence-based association studies in samples of admixed ancestry [35]. We note that this variant was not significantly associated with eGFR in the larger multi-ethnic association study conducted by our collaborators [Lin and Franceschini,

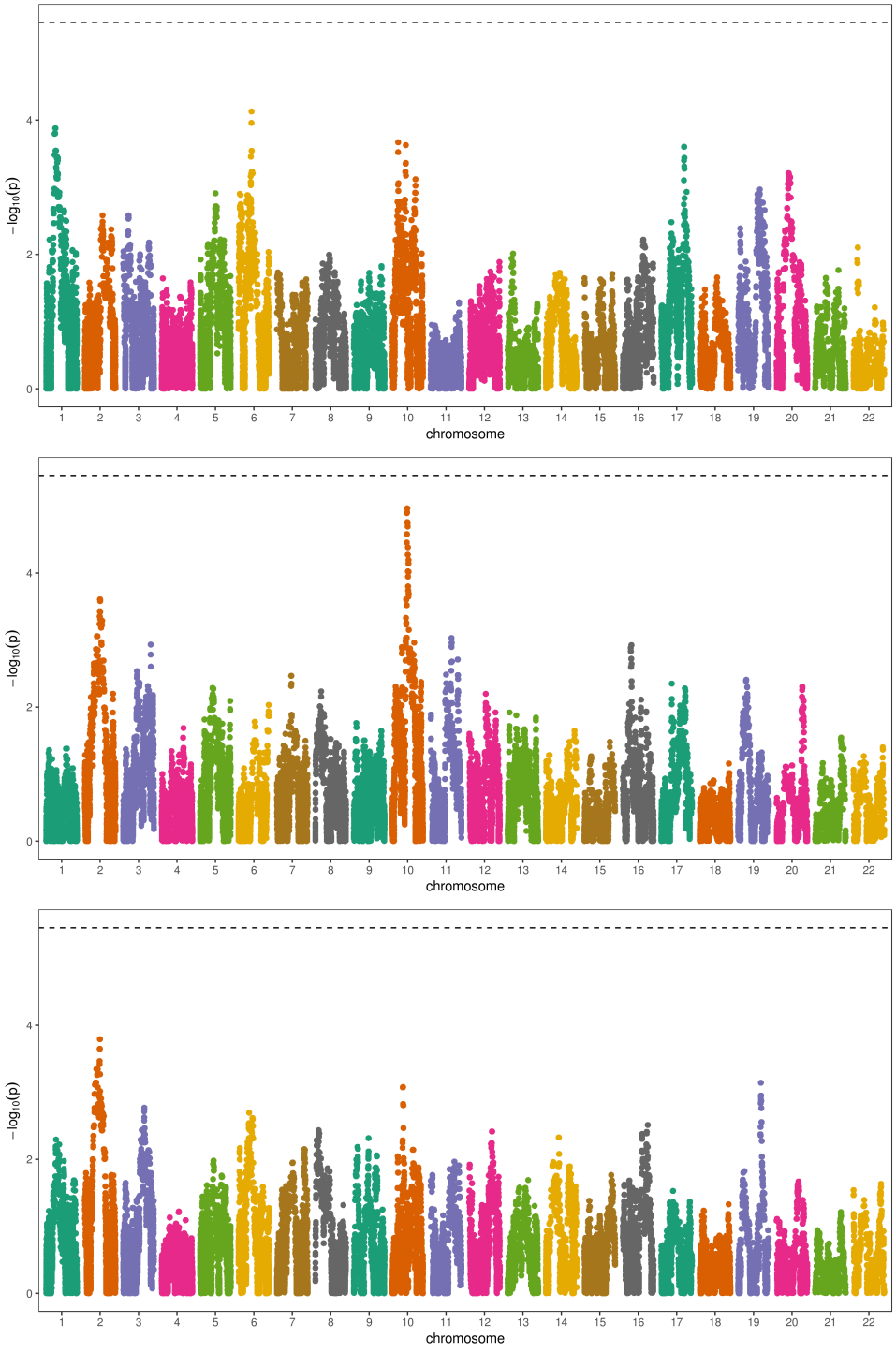

Figure 5.5: Manhattan plots for eGFR admixture mapping analysis using 1,176 Hispanic/Latino individuals.

Admixture mapping p-values at each variant for eGFR versus African local ancestry (top panel; $\lambda = 1.143$), European local ancestry (middle panel; $\lambda = 0.951$), and Native American local ancestry (bottom panel; $\lambda = 1.035$)

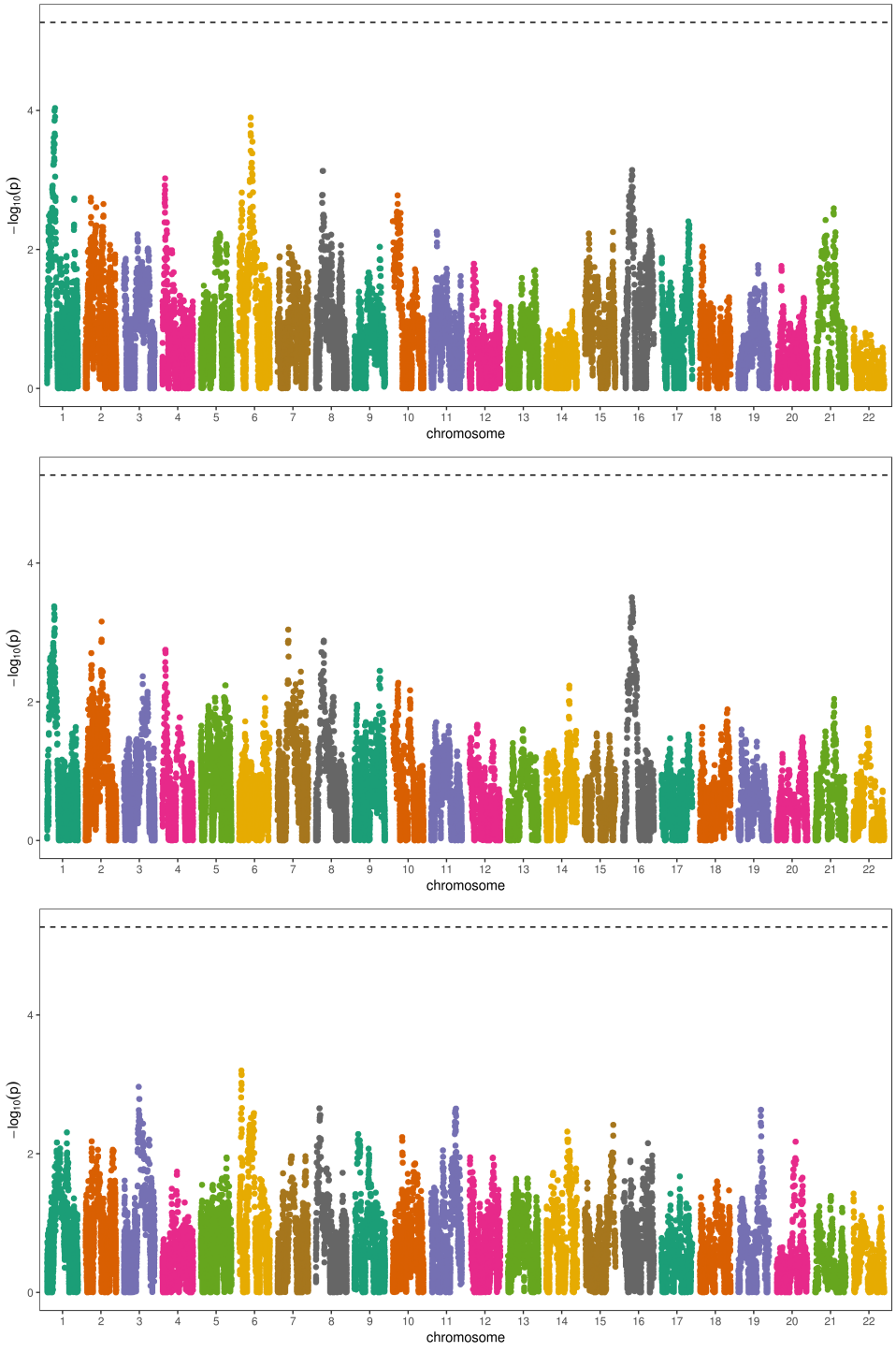

Figure 5.6: Manhattan plots for serum creatinine admixture mapping analysis using 9,479 African American and Hispanic/Latino individuals.

Admixture mapping p-values at each variant for serum creatinine versus African local ancestry (top panel; $\lambda = 1.075$), European local ancestry (middle panel; $\lambda = 1.009$), and Native American local ancestry (bottom panel; $\lambda = 0.999$)

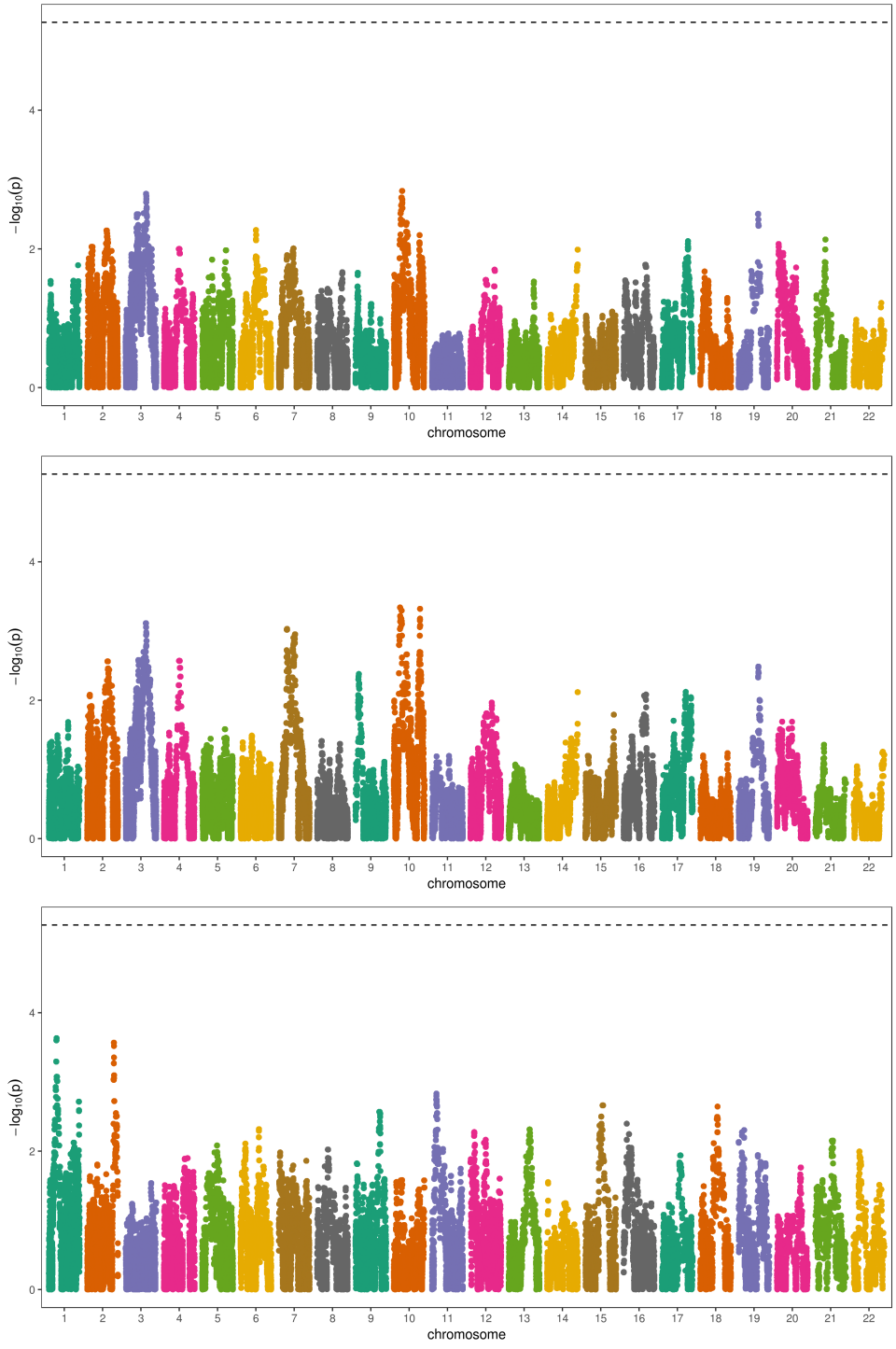

Figure 5.7: Manhattan plots for chronic kidney disease admixture mapping analysis using 9,479 African American and Hispanic/Latino individuals.

Admixture mapping p-values at each variant for chronic kidney disease versus African local ancestry (top panel; $\lambda = 0.906$), European local ancestry (middle panel; $\lambda = 0.876$), and Native American local ancestry (bottom panel; $\lambda = 0.941$)

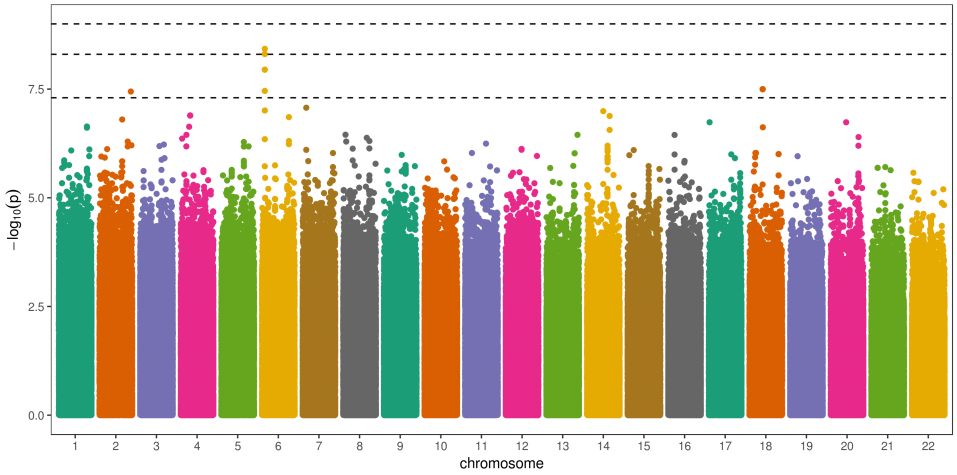

Figure 5.8: Manhattan plots for eGFR association mapping analysis using 9,479 African American and Hispanic/Latino individuals.
Association mapping p-values at each variant for eGFR versus genotype ($\lambda = 1.015$). Dashed lines represent p-value thresholds of $5 \times 10^{-8}, 5 \times 10^{-9}$, and 1×10^{-9}.

personal communication].

Quantile-quantile (QQ) plots for all analyses are available in Appendix C.4. The corresponding inflation factors λ—defined as the ratio of the median observed test statistic to the median expected test statistic under the (simplified) null hypothesis—are also provided in the figure legends for the Manhattan plots in this section.

5.4 Discussion

We used whole genome sequence data from the Trans-Omics for Precision Medicine project to investigate the genetic architecture of kidney disease in admixed populations. We inferred local ancestry for over twenty thousand admixed individuals and used that inferred local ancestry for ancestry-specific allele frequency estimation and genetic association studies of kidney traits. Although no variants reached genome-wide significance in our admixture mapping studies of eGFR, serum creatinine, or chronic kidney disease, we were still able to use our local ancestry calls to provide insight into variants of interest identified through

multi-ethnic whole genome sequence-based association mapping in a concurrent study [Lin and Franceschini, personal communication]. Furthermore, this work provides a valuable example of the application of local ancestry inference, admixture mapping, and the methods developed in Chapters 2–4 to whole genome sequence data.

In a concurrent study investigating the genetic architecture of kidney disease in a larger multi-ethnic cohort of TOPMed subjects, our collaborators identified a number of genetic variants significantly associated with eGFR [Lin and Franceschini, personal communication]. However, in our admixture mapping analyses in a reduced subset of admixed individuals, we did not identify any genetic variants significantly associated with eGFR, or related traits. In TOPMed, four of the top variants from the larger multi-ethnic association study are more common or exclusively present in admixed individuals compared to European or Asian Americans, making them seemingly good candidates for discovery via admixture mapping. Using our local ancestry calls, we estimated ancestry-specific allele frequencies for these variants to provide insight into the ancestral origin of these putatively causal genetic variants (Table 5.3). These estimated ancestry-specific allele frequencies could also provide insight into our lack of genome-wide significant findings in our admixture mapping studies: although we see a clear difference in allele frequencies across ancestral populations, the absolute magnitude of these differences is quite small. The power of admixture mapping studies depends on the difference in allele frequencies across ancestral populations [27] (see also: Chapter 4), so it may be that our sample size was too small to be able to detect a variant such as rs539182790 using admixture mapping. Alternatively, it is possible that the association mapping results from the larger multi-ethnic study are spurious. Further investigation is needed and in fact is already underway, including replication studies investigating rs539182790 in populations with larger amounts of Native American, as motivated by our ancestry-specific allele frequency estimates.

A number of challenges were presented by the use of these whole genome sequence (WGS) data. Currently, there are a limited number of publicly available whole genome sequences that could be used to construct a reference panel for local ancestry inference in WGS data.

In this analysis, we used whole genome sequences from the Simons Genome Diversity Project (SGDP) to form our reference panel. Using WGS data in the reference panel increased the number of variants at which local ancestry inference could be performed. However, a smaller number of samples are available in SGDP relative to other potential reference panel data sources, such as the Human Genome Diversity Project (HGDP) genotype data [111] used in Chapter 2. We decided to use the SGDP sequence data for this analysis because we found previously that local ancestry inference accuracy improves when more variants are present in the reference panel, even if fewer samples are included [50] (see also: Chapter 2). However, the optimal trade-off between the number of reference panel samples and variants remains an open question for local ancestry inference in WGS data. An additional challenge posed by these WGS data was computational in nature. In particular, we found that local ancestry inference was computationally challenging on a dataset of this size, likely due to the fact that existing methods for local ancestry inference were not designed for WGS data. To use `RFMix` on these data, we were required to split the admixed samples into very small subsets (250–500 individuals) and perform local ancestry inference separately within each subset. While this should not affect the accuracy of local ancestry inference using `RFMix`, it posed practical challenges.

This study highlights the need for future work. As suggested by the previous paragraph, there is a need for local ancestry inference methods that are computationally efficient in WGS data. Furthermore, the genome-wide significance thresholds used in our admixture mapping studies were estimated using `STEAM`, which was not designed for studies, such as this one, with related individuals. Preliminary analyses showed that using `STEAM` to estimate the number of generations since admixture yielded reasonable results in these data (see Appendix C.3), but further work is needed to verify that this approach appropriately controls for multiple testing in the presence of relatedness. Finally, this analysis has highlighted the need for a better understanding of the factors that influence the power of admixture mapping studies, as well as affirming the need to continue efforts in the recruitment of larger numbers of diverse samples in genetic association studies. It will be interesting to re-run the admixture

mapping analysis presented in this chapter when larger numbers of admixed samples are available for analysis in later freezes of the TOPMed WGS data.

Our work in this chapter provides important insight and ground-work for future studies. First, we have demonstrated the utility of local ancestry calls for tasks other than just admixture mapping. We used our local ancestry calls to estimate admixture proportions, which we included as fixed effects in our linear mixed model to adjust for population structure in association mapping. We also used local ancestry to estimate ancestry-specific allele frequencies for candidate variants of interest; these allele frequency estimates are useful for understanding disease etiology and for planning replication studies in independent datasets. Second, the local ancestry calls produced for our analysis provide a valuable resource for TOPMed investigators. We inferred local ancestry for all 20,048 admixed samples in TOPMed, including the 9,479 African American and Hispanic/Latino individuals included in our kidney association analyses, as well as additional African American, African Caribbean, and Hispanic/Latino samples who did not have available kidney phenotype data. The local ancestry calls for all 20,048 admixed individuals are available to other TOPMed investigators for studies of other phenotypes. Additional admixture mapping studies using our local ancestry calls are already underway.

Chapter 6

CONCLUSIONS AND FUTURE WORK

In this dissertation, we have proposed and evaluated statistical methods for ancestry inference and genetic association studies in admixed populations. First, in Chapter 2 we suggested approaches for extending existing local ancestry inference software to analysis of chromosome X. Then, in Chapter 3 we developed two related methods for estimating genome-wide significance thresholds for admixture mapping studies, motivated by a theoretical framework that is generally applicable to any admixed population, regardless of the number of ancestral populations or distribution of admixture proportions across the population. Chapter 4 investigated techniques for controlling for ancestral heterogeneity in genome-wide association and admixture mapping studies, providing new insight into the scenarios under which global ancestry confounds genetic association studies and showing that principal component analysis, a widely-used ancestral heterogeneity adjustment technique, can induce spurious associations in genetic association studies. Finally, in Chapter 5, we applied the lessons learned in earlier chapters to infer local and global ancestry and perform genetic association studies using whole genome sequence data from the Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project.

Our work has highlighted a number of existing challenges that arise in genetic studies in admixed populations. Local ancestry inference is limited by the availability of appropriate reference panel data, particularly in application to whole genome sequence data. For example, in our analysis of data from the TOPMed Whole Genome Sequencing Project in Chapter 5, the number of individuals included in our reference panel was small, and we were unable to perform local ancestry inference on chromosome X due to the fact that phased X chromosome sequence data was not available from the Simons Genome Diversity Project (SGDP)

[153] resource from which we formed our reference panel. The accuracy of local ancestry inference should improve as larger collections of whole genome sequence data from diverse populations become more readily available. New methods that are explicitly designed for local ancestry inference in whole genome sequence data will also be greatly beneficial, as the application of existing methods to sequence data is currently computationally cumbersome.

Another challenge highlighted by our work with TOPMed sequence data relates to the availability of data for admixed samples. The TOPMed project is a very large study that has prioritized the recruitment of samples from diverse backgrounds. Yet, it seems that even in this study our sample sizes may not be large enough to perform well-powered admixture mapping studies. In particular, our colleagues identified a variant on chromosome 19 that is putatively associated with kidney function [Bridget Lin and Nora Franceschini, personal communication], and we used our TOPMed local ancestry calls to show that this variant is more common in Native American populations than in Africans and Europeans; however, we were unable to detect this variant in our admixture mapping study. The absolute magnitude of the difference in allele frequencies at this variant is small, and our work in Chapter 4 suggests that the power of admixture mapping studies depends on the size of this difference. A more thorough investigation of the factors that impact the power of admixture mapping studies, and perhaps the development of a power calculator for these studies, would prove very useful for planning future admixture mapping studies and motivating the recruitment of larger numbers of admixed individuals for genetic association studies.

An important area of future work will involve extending our multiple testing work in Chapter 3. In that chapter, we developed methods for estimating significance thresholds for admixture mapping studies that investigate each locus $j \in \{1, \ldots, m\}$ and ancestral population $k \in \{1, \ldots, K\}$ individually, by fitting $m \times K$ total regression models

$$E[y_i \mid a_{ijk}, \boldsymbol{\pi}_i] = \beta_0 + \beta_{jk} a_{ijk} + \boldsymbol{\beta}_{\pi} \boldsymbol{\pi}_{i,K-1}.$$

Our multiple testing correction procedure accounts for the fact that we are testing K hypotheses at each locus, and explicitly models the correlation among tests at the same and

neighboring loci. However, in some cases it may be of interest to conduct a joint test for the association between the trait and local ancestry, by instead fitting a single model at each locus

$$E[y_i \mid \mathbf{a}_{ij}, \boldsymbol{\pi}_i] = \beta_0 + \beta_{j1} a_{ij1} + \cdots + \beta_{j,K-1} a_{ij,K-1} + \boldsymbol{\beta}_{\pi} \boldsymbol{\pi}_{i,K-1},$$

and testing the joint null hypothesis $H_0 : \boldsymbol{\beta}_j = \mathbf{0}$, where $\boldsymbol{\beta}_j$ is a vector of regression coefficients for the first $K-1$ local ancestry components (i.e., $\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{j1} & \dots \beta_{j,K-1} \end{pmatrix}^{\top}$). Our multiple testing correction method is not currently applicable to admixture mapping tests conducted in this way, but this extension is certainly of interest as many investigators utilize this joint modeling approach in the literature [26, 25, 27, 30]. Another extension that would be interesting to pursue would be developing an analytic approximation to the significance threshold for admixed populations with three or more ancestral populations. In Chapter 3, we proposed a simulation-based multiple testing correction approach that is applicable to admixed populations with any number of ancestral populations ($K \geq 2$) and an analytic approximation approach that is only applicable to populations with $K = 2$. Although we developed a very fast algorithm for implementing the simulation-based approach, the analytic approximation is still considerably faster, so it would be nice if we could derive an analytic approximation for admixed populations with $K \geq 3$ as well.

Finally, it will be important to extend our work in both Chapters 3 and 4 to consider the use of mixed models, which are widely used to perform genetic association studies while accounting for both close and distant relatedness across samples [64, 62, 63, 159]. Our application to TOPMed data in Chapter 5 seems to suggest that our proposed approaches for correcting for multiple testing (i.e., using our program `STEAM`) and adjusting for ancestral heterogeneity (i.e., using estimated admixture proportions rather than principal components) still work well in samples with related individuals. However, a more thorough investigation is warranted. Recent work related to correcting for multiple testing with linear mixed models [160], the sensitivity of mixed model based approaches to linkage disequilibrium (LD) [161, 162], and the potential of mixed model based approaches for controlling for confounding to induce spurious associations in gene expression studies [146] will provide useful starting

points for our own future investigations.

BIBLIOGRAPHY

[1] Anna C Need and David B Goldstein. Next generation disparities in human genomics: Concerns and remedies. *Trends in Genetics*, 25(11):489–494, 2009.

[2] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 2016.

[3] Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019.

[4] Carlos D Bustamante, Francisco M De La Vega, and Esteban G Burchard. Genomics for the world. *Nature*, 475(7355):163–165, 2011.

[5] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, David L Altshuler, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

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

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

[8] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):e1002453, 2012.

[9] Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6):e1000519, 2009.

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

[11] Yael Baran, Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10):1359–1367, 2012.

[12] David C Rife. Populations of hybrid origin as source material for the detection of linkage. *The American Journal of Human Genetics*, 6(1):26, 1954.

[13] Paul M McKeigue. Mapping genes that underlie ethnic differences in disease risk: Methods for detecting linkage in admixed populations, by conditioning on parental admixture. *The American Journal of Human Genetics*, 63(1):241–251, 1998.

[14] Daniel Shriner. Overview of admixture mapping. *Current Protocols in Human Genetics*, 76(1):1.23.1–1.23.8, 2013.

[15] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[16] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[17] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176, 2010.

[18] Erick Forno and Juan C Celedón. Asthma and ethnic minorities: Socioeconomic status and beyond. *Current Opinion in Allergy and Clinical Immunology*, 9(2):154–160, 2009.

[19] Matthew L Freedman, Christopher A Haiman, Nick Patterson, Gavin J McDonald, Arti Tandon, Alicja Waliszewska, Kathryn Penney, Robert G Steen, Kristin Ardlie, Esther M John, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences*, 103(38):14068–14073, 2006.

[20] Aram V Chobanian, George L Bakris, Henry R Black, William C Cushman, Lee A Green, Joseph L Izzo, Daniel W Jones, Barry J Materson, Suzanne Oparil, Jackson T Wright, et al. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension*, 42(6):1206–1252, 2003.

[21] Daniel Murphy, Charles E McCulloch, Feng Lin, Tanushree Banerjee, Jennifer L Bragg-Gresham, Mark S Eberhardt, Hal Morgenstern, Meda E Pavkov, Rajiv Saran, Neil R Powe, et al. Trends in prevalence of chronic kidney disease in the United States. *Annals of Internal Medicine*, 165(7):473–481, 2016.

[22] Afshin Parsa, WH Linda Kao, Dawei Xie, Brad C Astor, Man Li, Chi-yuan Hsu, Harold I Feldman, Rulan S Parekh, John W Kusek, Tom H Greene, et al. APOL1 risk variants, race, and progression of chronic kidney disease. *New England Journal of Medicine*, 369(23):2183–2196, 2013.

[23] Michael W Smith, Nick Patterson, James A Lautenberger, Ann L Truelove, Gavin J McDonald, Alicja Waliszewska, Bailey D Kessing, Michael J Malasky, Charles Scafe,

Ernest Le, et al. A high-density admixture map for disease gene discovery in African Americans. *The American Journal of Human Genetics*, 74(5):1001–1013, 2004.

[24] David T Redden, Jasmin Divers, Laura Kelly Vaughan, Hemant K Tiwari, T Mark Beasley, José R Fernández, Robert P Kimberly, Rui Feng, Miguel A Padilla, Nianjun Liu, et al. Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. *Plos Genetics*, 2(8):e137, 2006.

[25] Lisa A Brown, Tamar Sofer, Adrienne M Stilp, Leslie J Baier, Holly J Kramer, Ivica Masindova, Daniel Levy, Robert L Hanson, Ashley E Moncrieft, Susan Redline, et al. Admixture mapping identifies an Amerindian ancestry locus associated with albuminuria in Hispanics in the United States. *Journal of the American Society of Nephrology*, 28(7):2211–2220, 2017.

[26] Lisa Anne Brown. *Statistical Methods in Admixture Mapping: Mixed Model Based Testing and Genome-wide Significance Thresholds*. PhD thesis, University of Washington, 2016.

[27] Tamar Sofer, Leslie J Baier, Sharon R Browning, Timothy A Thornton, Gregory A Talavera, Sylvia Wassertheil-Smoller, Martha L Daviglus, Robert Hanson, Sayuko Kobes, Richard S Cooper, et al. Admixture mapping in the Hispanic Community Health Study/Study of Latinos reveals regions of genetic associations with blood pressure traits. *PloS One*, 12(11):e0188400, 2017.

[28] Huaizhen Qin and Xiaofeng Zhu. Power comparison of admixture mapping and direct association analysis in genome-wide association studies. *Genetic Epidemiology*, 36(3):235–243, 2012.

[29] Alex P Reiner, Sandra Beleza, Nora Franceschini, Paul L Auer, Jennifer G Robinson, Charles Kooperberg, Ulrike Peters, and Hua Tang. Genome-wide association and

population genetic analysis of C-reactive protein in African American and Hispanic American women. *The American Journal of Human Genetics*, 91(3):502–512, 2012.

[30] Ursula M Schick, Deepti Jain, Chani J Hodonsky, Jean V Morrison, James P Davis, Lisa Brown, Tamar Sofer, Matthew P Conomos, Claudia Schurmann, Caitlin P McHugh, et al. Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. *The American Journal of Human Genetics*, 98(2):229–242, 2016.

[31] Hua Tang, David O Siegmund, Nicholas A Johnson, Isabelle Romieu, and Stephanie J London. Joint testing of genotype and ancestry association in admixed families. *Genetic Epidemiology*, 34(8):783–791, 2010.

[32] Daniel Shriner, Adebowale Adeyemo, and Charles N Rotimi. Joint ancestry and association testing in admixed individuals. *PLoS Computational Biology*, 7(12):e1002325, 2011.

[33] Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, 32(4):381–385, 2008.

[34] Anne-Sophie Jannot, Georg Ehret, and Thomas Perneger. $P < 5 \times 10 - 8$ has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology*, 68(4):460–465, 2015.

[35] Sara L Pulit, Sera AJ de With, and Paul IW de Bakker. Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genetic Epidemiology*, 41(2):145–151, 2017.

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

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

[38] Daniel Yorgov, Karen L Edwards, and Stephanie A Santorico. Use of admixture and association for detection of quantitative trait loci in the Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Ethnic Samples (T2D-GENES) study. *BMC Proceedings*, 8(1):S6, 2014.

[39] Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10):e1000686, 2009.

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

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

[42] Christian Fuchsberger, Gonçalo R Abecasis, and David A Hinds. minimac2: Faster genotype imputation. *Bioinformatics*, 31(5):782–784, 2014.

[43] Jan Graffelman and BS Weir. Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome. *Heredity*, 116(6):558–568, 2016.

[44] Caitlin Patricia McHugh. *Statistical Methods for the Analysis of Autosomal and X Chromosome Genetic Data in Samples with Unknown Structure*. PhD thesis, University of Washington, 2016.

[45] Lisa M LaVange, William D Kalsbeek, Paul D Sorlie, Larissa M Avilés-Santa, Robert C Kaplan, Janice Barnhart, Kiang Liu, Aida Giachello, David J Lee, John Ryan, et al.

Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, 20(8):642–649, 2010.

[46] Paul D Sorlie, Larissa M Avilés-Santa, Sylvia Wassertheil-Smoller, Robert C Kaplan, Martha L Daviglus, Aida L Giachello, Neil Schneiderman, Leopoldo Raij, Gregory Talavera, Matthew Allison, et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, 20(8):629–641, 2010.

[47] Katarzyna Bryc, Adam Auton, Matthew R Nelson, Jorge R Oksenberg, Stephen L Hauser, Scott Williams, Alain Froment, Jean-Marie Bodo, Charles Wambebe, Sarah A Tishkoff, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, 107(2):786–791, 2010.

[48] Katarzyna Bryc, Eric Y Durand, J Michael Macpherson, David Reich, and Joanna L Mountain. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, 96(1):37–53, 2015.

[49] Ching-Yu Cheng, WH Linda Kao, Nick Patterson, Arti Tandon, Christopher A Haiman, Tamara B Harris, Chao Xing, Esther M John, Christine B Ambrosone, Frederick L Brancati, et al. Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. *PLoS Genetics*, 5(5):e1000490, 2009.

[50] Sharon R Browning, Kelsey Grinde, Anna Plantinga, Stephanie M Gogarten, Adrienne M Stilp, Robert C Kaplan, M Larissa Avilés-Santa, Brian L Browning, and Cathy C Laurie. Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *G3: Genes, Genomes, Genetics*, 6(6):1525–1534, 2016.

[51] J Li and L Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, 2005.

[52] Xiaofeng Zhu, JH Young, Ervin Fox, Brendan J Keating, Nora Franceschini, Sunjung Kang, Bamidele Tayo, Adebowale Adeyemo, Yun V Sun, Yali Li, et al. Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: Contributions from the CARe consortium. *Human Molecular Genetics*, 20(11):2285–2295, 2011.

[53] Karen N Conneely and Michael Boehnke. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168, 2007.

[54] Frank Dudbridge and Bobby PC Koeleman. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *The American Journal of Human Genetics*, 75(3):424–435, 2004.

[55] Daria Salyakina, Shaun R Seaman, Brian L Browning, Frank Dudbridge, and Bertram Muller-Myhsok. Evaluation of Nyholt's procedure for multiple testing correction. *Human Herediddty*, 60(1):19–25, 2005.

[56] Marc A Coram, Qing Duan, Thomas J Hoffmann, Timothy Thornton, Joshua W Knowles, Nicholas A Johnson, Heather M Ochs-Balcom, Timothy A Donlon, Lisa W Martin, Charles B Eaton, et al. Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *The American Journal of Human Genetics*, 92(6):904–916, 2013.

[57] Cara L Carty, Nicholas A Johnson, Carolyn M Hutter, Alexander P Reiner, Ulrike Peters, Hua Tang, and Charles Kooperberg. Genome-wide association study of body height in African Americans: The Women's Health Initiative SNP Health Association Resource (SHARe). *Human Molecular Genetics*, 21(3):711–720, 2012.

[58] David Siegmund and Benjamin Yakir. *The Statistics of Gene Mapping*. Springer Science & Business Media, NY, 2007.

[59] Kelsey E Grinde, Lisa A Brown, Alexander P Reiner, Timothy A Thornton, and Sharon R Browning. Genome-wide significance thresholds for admixture mapping studies. *The American Journal of Human Genetics*, 104(3):454–465, 2019.

[60] Richard S Spielman, Ralph E McGinnis, and Warren J Ewens. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *The American Journal of Human Genetics*, 52(3):506–516, 1993.

[61] Eric S Lander and Nicholas J Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–2048, 1994.

[62] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2006.

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

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

[65] Jonathan K Pritchard, Matthew Stephens, Noah A Rosenberg, and Peter Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000.

[66] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

[67] Matthew P Conomos, Michael B Miller, and Timothy A Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39(4):276–293, 2015.

[68] Michael E Weale. Quality control for genome-wide association studies. In *Genetic Variation*, pages 341–372. Springer, 2010.

[69] Chao Tian, Robert M Plenge, Michael Ransom, Annette Lee, Pablo Villoslada, Carlo Selmi, Lars Klareskog, Ann E Pulver, Lihong Qi, Peter K Gregersen, et al. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genetics*, 4(1):e4, 2008.

[70] Alkes L Price, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, Jerome I Rotter, Esther Torres, Kent D Taylor, et al. Long-range LD can confound genome scans in admixed populations. *The American Journal of Human Genetics*, 83(1):132–135, 2008.

[71] Felix Elwert and Christopher Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53, 2014.

[72] Centers for Disease Control and Prevention. Chronic kidney disease in the United States, 2019. US Department of Health and Human Services, Centers for Disease Control and Prevention, 2019.

[73] Liran I Shlush, Sivan Bercovici, Walter G Wasser, Guennady Yudkovsky, Alan Templeton, Dan Geiger, and Karl Skorecki. Admixture mapping of end stage kidney disease genetic susceptibility using estimated mutual information ancestry informative markers. *BMC Medical Genomics*, 3(1):47, 2010.

[74] Qian S Zhang, Brian L Browning, and Sharon R Browning. ASAFE: Ancestry-specific allele frequency estimation. *Bioinformatics*, 32(14):2227–2229, 2016.

[75] Xuexia Wang, Xiaofeng Zhu, Huaizhen Qin, Richard S Cooper, Warren J Ewens, Chun Li, and Mingyao Li. Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics*, 27(5):670–677, 2010.

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

[77] Daniel Wegmann, Darren E Kessner, Krishna R Veeramah, Rasika A Mathias, Dan L Nicolae, Lisa R Yanek, Yan V Sun, Dara G Torgerson, Nicholas Rafaels, Thomas Mosley, et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics*, 43(9):847, 2011.

[78] Anjali G Hinch, Arti Tandon, Nick Patterson, Yunli Song, Nadin Rohland, Cameron D Palmer, Gary K Chen, Kai Wang, Sarah G Buxbaum, Ermeg L Akylbekova, et al. The landscape of recombination in African Americans. *Nature*, 476(7359):170, 2011.

[79] John E Pool and Rasmus Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, 2009.

[80] Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012.

[81] Garrett Hellenthal, George BJ Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

[82] Eric Y Durand, Chuong B Do, Joanna L Mountain, and J Michael Macpherson. Ancestry composition: A novel, efficient pipeline for ancestry deconvolution. *biorxiv*, page 010512, 2014.

[83] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[84] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

[85] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1):1–12, 2006.

[86] Caitlin McHugh, Lisa Brown, and Timothy A Thornton. Detecting heterogeneity in population structure across the genome in admixed populations. *Genetics*, 204(1):43–56, 2016.

[87] Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

[88] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834, 2010.

[89] Matthew P Conomos, Cecelia A Laurie, Adrienne M Stilp, Stephanie M Gogarten, Caitlin P McHugh, Sarah C Nelson, Tamar Sofer, Lindsay Fernández-Rhodes, Anne E Justice, Mariaelisa Graff, et al. Genetic diversity and association studies in US Hispanic/Latino populations: Applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*, 98(1):165–184, 2016.

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

[91] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

[92] David Reich, Nick Patterson, Desmond Campbell, Arti Tandon, Stéphane Mazieres, Nicolas Ray, Maria V Parra, Winston Rojas, Constanza Duque, Natalia Mesa, et al. Reconstructing Native American population history. *Nature*, 488(7411):370, 2012.

[93] Bogdan Pasaniuc, Sriram Sankararaman, Dara G Torgerson, Christopher Gignoux, Noah Zaitlen, Celeste Eng, William Rodriguez-Cintron, Rocio Chapela, Jean G Ford, Pedro C Avila, et al. Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, 29(11):1407–1415, 2013.

[94] Xiaofeng Zhu, Amy Luke, Richard S Cooper, Tom Quertermous, Craig Hanis, Tom Mosley, C Charles Gu, Hua Tang, Dabeeru C Rao, Neil Risch, et al. Admixture mapping for hypertension loci with genome-scan markers. *Nature Genetics*, 37(2):177, 2005.

[95] David Reich, Nick Patterson, Vijaya Ramesh, Philip L De Jager, Gavin J McDonald, Arti Tandon, Edwin Choy, Donglei Hu, Bani Tamraz, Ludmila Pawlikowska, et al. Admixture mapping of an allele affecting interleukin 6 soluble receptor and interleukin 6 levels. *The American Journal of Human Genetics*, 80(4):716–726, 2007.

[96] Joshua M Galanter, Christopher R Gignoux, Dara G Torgerson, Lindsey A Roth, Celeste Eng, Sam S Oh, Elizabeth A Nguyen, Katherine A Drake, Scott Huntsman, Donglei Hu, et al. Genome-wide association study and admixture mapping identify

different asthma-associated loci in Latinos: The Genes-environments & Admixture in Latino Americans study. *Journal of Allergy and Clinical Immunology*, 134(2):295–305, 2014.

[97] Melissa L Spear, Donglei Hu, Maria Pino-Yanes, Scott Huntsman, Celeste Eng, Albert M Levin, Victor E Ortega, et al. A genome-wide association and admixture mapping study of bronchodilator drug response in African Americans with asthma. *The Pharmacogenomics Journal*, 2018.

[98] Calvin Chi, Xiaorong Shao, Brooke Rhead, Edlin Gonzales, Jessica B Smith, Anny H Xiang, Jennifer Graves, Amy Waldman, Timothy Lotze, Teri Schreiner, et al. Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. *PLoS Genetics*, 15(1):e1007808, 2019.

[99] Felicia Gomez, Lihua Wang, Haley Abel, Qunyuan Zhang, Michael A Province, and Ingrid B Borecki. Admixture mapping of coronary artery calcification in African Americans from the NHLBI family heart study. *BMC Genetics*, 16(1):42, 2015.

[100] Ayush Giri, Katherine E Hartmann, Melinda C Aldrich, Renee M Ward, Jennifer M Wu, Amy J Park, Mariaelisa Graff, Lihong Qi, Rami Nassir, Robert B Wallace, et al. Admixture mapping of pelvic organ prolapse in African Americans from the Women's Health Initiative Hormone Therapy Trial. *PLoS One*, 12(6):e0178839, 2017.

[101] Ayush Giri, Todd L Edwards, Katherine E Hartmann, Eric S Torstenson, Melissa Wellons, Pamela J Schreiner, and Digna R Velez Edwards. African genetic ancestry interacts with body mass index to modify risk for uterine fibroids. *PLoS Genetics*, 13(7):e1006871, 2017.

[102] Shaun R Seaman and Bertram Müller-Myhsok. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *The American Journal of Human Genetics*, 76(3):399–408, 2005.

[103] Buhm Han, Hyun Min Kang, and Eleazar Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genetics*, 5(4):e1000456, 2009.

[104] Qiuying Sha, Xihuan Zhang, Xiaofeng Zhu, and Shuanglin Zhang. Analytical correction for multiple testing in admixture mapping. *Human Heredity*, 62(2):55–63, 2006.

[105] Sarah A Tishkoff, Floyd A Reed, Françoise R Friedlaender, Christopher Ehret, Alessia Ranciaro, Alain Froment, Jibril B Hirbo, et al. The genetic structure and history of Africans and African Americans. *Science*, 324(5930):1035–1044, 2009.

[106] Esteban J Parra, Amy Marcini, Joshua Akey, Jeremy Martinson, Mark A Batzer, Richard Cooper, Terrence Forrester, David B Allison, Ranjan Deka, Robert E Ferrell, and Mark D Shriver. Estimating African American admixture proportions by use of population-specific alleles. *The American Journal of Human Genetics*, 63(6):1839–1851, 1998.

[107] Katarzyna Bryc, Christopher Velez, Tatiana Karafet, Andres Moreno-Estrada, Andy Reynolds, Adam Auton, Michael Hammer, Carlos D Bustamante, and Harry Ostrer. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):8954–8961, 2010.

[108] R Core Team. R: A language and environment for statistical computing., 2018.

[109] Jennifer Hays, Julie R Hunt, F Allan Hubbell, Garnet L Anderson, Marian Limacher, Catherine Allen, and Jacques E Rossouw. The Women's Health Initiative recruitment methods and results. *Annals of Epidemiology*, 13(9):S18–S77, 2003.

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

[111] Howard M Cann, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

[112] Matthew P Conomos, Alexander P Reiner, Bruce S Weir, and Timothy A Thornton. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, 98(1):127–148, 2016.

[113] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

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

[115] Clive J Hoggart, Mark D Shriver, Rick A Kittles, David G Clayton, and Paul M McKeigue. Design and analysis of admixture mapping studies. *The American Journal of Human Genetics*, 74(5):965–978, 2004.

[116] Alkes L Price, Nick Patterson, Fuli Yu, David R Cox, Alicja Waliszewska, Gavin J McDonald, Arti Tandon, et al. A genomewide admixture map for Latino populations. *The American Journal of Human Genetics*, 80(6):1024–1036, 2007.

[117] T. Mark Beasley, Stephen Erickson, and David B. Allison. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, 39(5):580, 2009.

[118] Tamar Sofer, Xiuwen Zheng, Stephanie M Gogarten, Cecelia A Laurie, Kelsey Grinde, John R Shaffer, Dmitry Shungin, Jeffrey R O'Connell, Ramon A Durazo-Arvizo, Laura

Raffield, Leslie Lange, Solomon Musani, Ramachandran S Vasan, L. Adrienne Cupples, Alexander P Reiner, Cathy C Laurie, and Kenneth M Rice. A fully-adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, 43(3):263–275, 2019.

[119] Noah Zaitlen, Scott Huntsman, Donglei Hu, Melissa Spear, Celeste Eng, Sam S Oh, Marquitta J White, et al. The effects of migration and assortative mating on admixture linkage disequilibrium. *Genetics*, 205(1):375–383, 2017.

[120] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[121] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[122] Heather M Ochs-Balcom, Leah Preus, Jean Wactawski-Wende, Jing Nie, Nicholas A Johnson, Fouad Zakharia, Hua Tang, Chris Carlson, Cara Carty, Zhao Chen, et al. Association of DXA-derived bone mineral density and fat mass with African ancestry. *The Journal of Clinical Endocrinology & Metabolism*, 98(4):E713–E717, 2013.

[123] Joannella Morales, Danielle Welter, Emily H Bowler, Maria Cerezo, Laura W Harris, Aoife C McMahon, Peggy Hall, Heather A Junkins, Annalisa Milano, Emma Hastings, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biology*, 19(1):21, 2018.

[124] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, 2019.

[125] Ronnie Sebro, Gina M Peloso, Josée Dupuis, and Neil J Risch. Structured mating: Patterns and implications. *PLoS Genetics*, 13(4):e1006655, 2017.

[126] Emily T Norris, Lavanya Rishishwar, Lu Wang, Andrew B Conley, Aroon T Chande, Adam M Dabrowski, Augusto Valderrama-Aguirre, and I King Jordan. Assortative mating on ancestry-variant traits in admixed Latin American populations. *Frontiers in Genetics*, 10:359, 2019.

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

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

[129] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573, 2010.

[130] Abdel Abdellaoui, Jouke-Jan Hottenga, Peter De Knijff, Michel G Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, Erik A Ehli, Yueshan Hu, Gareth E Davies, et al. Population structure, migration, and diversifying selection in the Netherlands. *European Journal of Human Genetics*, 21(11):1277–1285, 2013.

[131] Eric Reed, Sara Nunez, David Kulp, Jing Qian, Muredach P Reilly, and Andrea S Foulkes. A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, 34(28):3769–3792, 2015.

[132] Michelle Daya, Nicholas Rafaels, Tonya M Brunetti, Sameer Chavan, Albert M Levin, Aniket Shetty, Christopher R Gignoux, Meher Preethi Boorgula, Genevieve Wojcik, Monica Campbell, et al. Association study in African-admixed populations across the

Americas recapitulates asthma risk loci in non-African populations. *Nature Communications*, 10(1):880, 2019.

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

[134] Paola Raska, Edwin Iversen, Ann Chen, Zhihua Chen, Brooke L Fridley, Jennifer Permuth-Wey, Ya-Yu Tsai, Robert A Vierkant, Ellen L Goode, Harvey Risch, et al. European American stratification in ovarian cancer case control data: The utility of genome-wide data for inferring ancestry. *PloS One*, 7(5):e35235, 2012.

[135] Jacques Fellay, Kevin V Shianna, Dongliang Ge, Sara Colombo, Bruno Ledergerber, Mike Weale, Kunlin Zhang, Curtis Gumbs, Antonella Castagna, Andrea Cossarizza, et al. A whole-genome association study of major determinants for host control of HIV-1. *Science*, 317(5840):944–947, 2007.

[136] Fei Zou, Seunggeun Lee, Michael R Knowles, and Fred A Wright. Quantification of population structure using correlated SNPs by shrinkage principal components. *Human Heredity*, 70(1):9–22, 2010.

[137] Paul M McKeigue. Prospects for admixture mapping of complex traits. *The American Journal of Human Genetics*, 76(1):1–7, 2005.

[138] Michael W Smith and Stephen J O'Brien. Mapping by admixture linkage disequilibrium: Advances, limitations and guidelines. *Nature Reviews Genetics*, 6(8):623–633, 2005.

[139] Thomas J Hoffmann, Hua Tang, Timothy A Thornton, Bette Caan, Mary Haan, Amy E Millen, Fridtjof Thomas, and Neil Risch. Genome-wide association and admixture analysis of glaucoma in the Women's Health Initiative. *Human Molecular Genetics*, 23(24):6634–6643, 2014.

[140] Ellen W Demerath, Ching-Ti Liu, Nora Franceschini, Gary Chen, Julie R Palmer, Erin N Smith, Christina TL Chen, Christine B Ambrosone, Alice M Arnold, Elisa V Bandera, et al. Genome-wide association study of age at menarche in African-American women. *Human Molecular Genetics*, 22(16):3329–3346, 2013.

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

[142] Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael GB Blum. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16):2781–2787, 2018.

[143] Sholom Wacholder, Nathaniel Rothman, and Neil Caporaso. Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiology and Prevention Biomarkers*, 11(6):513–520, 2002.

[144] Hugues Aschard, Bjarni J Vilhjálmsson, Amit D Joshi, Alkes L Price, and Peter Kraft. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, 96(2):329–339, 2015.

[145] Felix R Day, Po-Ru Loh, Robert A Scott, Ken K Ong, and John RB Perry. A robust example of collider bias in a genetic association study. *The American Journal of Human Genetics*, 98(2):392–393, 2016.

[146] Andy Dahl, Vincent Guillemot, Joel Mefford, Hugues Aschard, and Noah Zaitlen. Adjusting for principal components of molecular phenotypes induces replicating false positives. *Genetics*, 211(4):1179–1189, 2019.

[147] Ranajit Chakraborty and Kenneth M Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences*, 85(23):9119–9123, 1988.

[148] Benjamin A Rybicki, Sudha K Iyengar, Trent Harris, Rachael Liptak, Robert C Elston, Mary J Maliarik, and Michael C Iannuzzi. Prospects of admixture linkage disequilibrium mapping in the African-American genome. *Cytometry: The Journal of the International Society for Analytical Cytology*, 47(1):63–65, 2002.

[149] Goo Jun, Mary Kate Wing, Gonçalo R Abecasis, and Hyun Min Kang. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, 25(6):918–925, 2015.

[150] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448, 2016.

[151] Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo, Stephanie M. Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*, 563866, 2019.

[152] Lesley A Inker, Christopher H Schmid, Hocine Tighiouart, John H Eckfeldt, Harold I Feldman, Tom Greene, John W Kusek, Jane Manzi, Frederick Van Lente, Yaping Lucy Zhang, et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. *New England Journal of Medicine*, 367(1):20–29, 2012.

[153] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. The

Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, 2016.

[154] Maximilian Haeussler, Ann S Zweig, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Angie S Hinrichs, Jairo Navarro Gonzalez, et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, 47(D1):D853–D858, 2018.

[155] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[156] Matthew P Conomos, Stephanie M Gogarten, Lisa Brown, Han Chen, Ken Rice, Tamar Sofer, Timothy Thornton, and Chaoyu Yu. GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R package, 2019.

[157] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[158] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[159] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284, 2015.

[160] Jong Wha J Joo, Farhad Hormozdiari, Buhm Han, and Eleazar Eskin. Multiple testing correction in linear mixed models. *Genome Biology*, 17(1):62, 2016.

[161] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[162] Alexander Gusev, Gaurav Bhatia, Noah Zaitlen, Bjarni J Vilhjalmsson, Dorothée Diogo, Eli A Stahl, Peter K Gregersen, Jane Worthington, Lars Klareskog, Soumya Raychaudhuri, et al. Quantifying missing heritability at known GWAS loci. *PLoS Genetics*, 9(12):e1003993, 2013.

[163] Simon N Wood. *Generalized additive models: An introduction with R*. Chapman and Hall/CRC, 2017.

[164] Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

[165] James Xue, Todd Lencz, Ariel Darvasi, Itsik Pe'er, and Shai Carmi. The time and place of European admixture in Ashkenazi Jewish history. *PLoS Genetics*, 13(4):e1006644, 2017.

[166] Hua Tang, Shweta Choudhry, Rui Mei, Martin Morgan, William Rodriguez-Cintron, Esteban González Burchard, and Neil J Risch. Recent genetic selection in the ancestral admixture of Puerto Ricans. *The American Journal of Human Genetics*, 81(3):626–633, 2007.

[167] Analabha Basu, Hua Tang, Xiaofeng Zhu, C Charles Gu, Craig Hanis, Eric Boerwinkle, and Neil Risch. Genome-wide distribution of ancestry in Mexican Americans. *Human Genetics*, 124(3):207–214, 2008.

[168] Georg B Ehret. Genome-wide association studies: Contribution of genomics to un-

derstanding blood pressure and essential hypertension. *Current Hypertension Reports*, 12(1):17–25, 2010.

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

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Proofs of Theoretical Results

A.1.1 Lemma 1: Local Ancestry Correlation

Lemma 1. *Consider an admixed population with K ancestral populations, g generations since admixture, and admixture proportions $\boldsymbol{\pi} = \begin{pmatrix}\pi_1 & \dots & \pi_K\end{pmatrix}^\top$ distributed according to $\boldsymbol{\pi} \sim F$, where F has finite first and second moments. Then, the correlation of local ancestry vectors at two loci j, j' separated by recombination fraction θ is given by:*

$$Corr(a_{jk}, a_{j'k'}) = \begin{cases} (1-\theta)^g + [1-(1-\theta)^g]\,\frac{2V_F(\pi_k)}{E_F(\pi_k)-E_F^2(\pi_k)+V_F(\pi_k)} & \text{if } k = k' \\ \frac{2Cov_F(\pi_k,\pi_{k'})-(1-\theta)^g[Cov_F(\pi_k,\pi_{k'})+E_F(\pi_k)E_F(\pi_{k'})]}{\sqrt{\left[E_F(\pi_k)-E_F^2(\pi_k)+V_F(\pi_k)\right]\left[E_F(\pi_{k'})-E_F^2(\pi_{k'})+V_F(\pi_{k'})\right]}} & \text{if } k \neq k', \end{cases}$$

where $E_F(\pi_k)$ and $V_F(\pi_k)$ are the mean and variance, with respect to the distribution F, of the admixture proportion from ancestral population k, and $Cov_F(\pi_k, \pi_{k'})$ is the covariance between the admixture proportions from ancestral populations k and k'.

Proof. Let individuals, indexed by $i = 1, ..., n$, come from an admixed population with K ancestral populations and g generations since admixture. Denote the admixture proportions for each individual by $\boldsymbol{\pi}_i = \begin{pmatrix}\pi_{i1} & \cdots & \pi_{iK}\end{pmatrix}^T$, $\sum_{k=1}^K \pi_{ik} = 1$, and let that vector be drawn from a distribution F with finite first and second moments. Let $\mathbf{a}_{ij} = \begin{pmatrix}a_{ij1} & \cdots & a_{ijK}\end{pmatrix}^T$ be the local ancestry vector for individual i at locus j, where $\sum_{k=1}^K a_{ijk} = 2$ and a_{ijk} denotes the number of alleles inherited by individual i from ancestral population k at locus j. Similarly, let $\mathbf{a}_{ij}^P = \begin{pmatrix}a_{ij1}^P & \cdots & a_{ijK}^P\end{pmatrix}^T$ be the parental local ancestry vector, where now $\sum_{k=1}^K a_{ijk}^P = 1$ and a_{ijk}^P denotes the number of alleles inherited by individual i from parent P ($P = M, F$ for mother and father, respectively) that are derived from ancestral population k at locus j.

Henceforth we will drop the subscript i for the sake of simplicity.

Consider two loci j, j' separated by recombination fraction θ or, equivalently, genetic distance δ cM. We wish to derive the correlation of local ancestry vectors $\mathbf{a}_j, \mathbf{a}_{j'}$ at these loci, but first we will consider the correlation of the parental local ancestry vectors $\mathbf{a}_j^P, \mathbf{a}_{j'}^P$.

Note that for the parental local ancestry vector $\mathbf{a}_j^P$, exactly one of the components takes the value 1 and the other $K-1$ components must take the value 0. Then, conditional on the vector of admixture proportions $\boldsymbol{\pi}$, the correlation of components k, k' of the parental local ancestry vectors at loci j, j' is:

$$\begin{aligned}
\mathrm{Corr}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi}) &= \frac{\mathrm{Cov}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi})}{\sqrt{\mathrm{V}(a_{jk}^P \mid \boldsymbol{\pi})\mathrm{V}(a_{j'k'}^P \mid \boldsymbol{\pi})}} \\
&= \frac{E(a_{jk}^P a_{j'k'}^P \mid \boldsymbol{\pi}) - E(a_{jk}^P \mid \boldsymbol{\pi})E(a_{j'k'}^P \mid \boldsymbol{\pi})}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}} \\
&= \frac{\Pr(a_{jk}^P = 1, a_{j'k'}^P = 1 \mid \boldsymbol{\pi}) - \pi_k\pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}}.
\end{aligned}$$

To reduce this further, we must consider two cases. First, suppose that $k = k'$. Then, $\Pr(a_{jk}^P = 1, a_{j'k'}^P = 1 \mid \boldsymbol{\pi}) = (1-\theta)^g\pi_k + [1-(1-\theta)^g]\pi_k^2$, where θ is the recombination fraction between loci j, j'. Second, suppose that $k \neq k'$. Now, $\Pr(a_{jk}^P = 1, a_{j'k'}^P = 1 \mid \boldsymbol{\pi}) = (1-\theta)^g \times 0 + [1-(1-\theta)^g]\pi_k\pi_{k'}$. After simplifying, it follows that

$$\mathrm{Corr}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi}) = \begin{cases} (1-\theta)^g & \text{if } k = k' \\ (1-\theta)^g \frac{-\pi_k\pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}} & \text{if } k \neq k'. \end{cases} \tag{A.1}$$

At each locus, we can separate the local ancestry vector $\mathbf{a}_j$ into the sum of the parental local ancestry vectors $\mathbf{a}_j^P$, such that $\mathbf{a}_j = \mathbf{a}_j^M + \mathbf{a}_j^F$. The parental local ancestry vectors are conditionally independent given the admixture proportions $\boldsymbol{\pi}$. Thus, $\mathrm{Cov}(a_{jk}, a_{j'k'} \mid \boldsymbol{\pi}) = \mathrm{Cov}(a_{jk}^M, a_{j'k'}^M \mid \boldsymbol{\pi}) + \mathrm{Cov}(a_{jk}^F, a_{j'k'}^F \mid \boldsymbol{\pi}) = 2\mathrm{Cov}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi})$, and the conditional correlation

of components of the local ancestry vectors $\mathbf{a}_j$ is the same as that of the parental vectors $\mathbf{a}_j^P$:

$$\begin{aligned}\text{Corr}(a_{jk}, a_{j'k'} \mid \boldsymbol{\pi}) &= \frac{2\text{Cov}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi})}{\sqrt{2V(a_{jk}^P \mid \boldsymbol{\pi})2V(a_{j'k'}^P \mid \boldsymbol{\pi})}} \\ &= \frac{\text{Cov}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi})}{\sqrt{V(a_{jk}^P \mid \boldsymbol{\pi})V(a_{j'k'}^P \mid \boldsymbol{\pi})}} \\ &= \text{Corr}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi}).\end{aligned}$$

We use the laws of total expectation, variance, and covariance to find the marginal correlation of local ancestry vectors:

$$\begin{aligned}\text{Corr}(a_{jk}, a_{j'k'}) &= \frac{\text{Cov}(a_{jk}, a_{j'k'})}{\sqrt{V(a_{jk})V(a_{j'k'})}} \\ &= \frac{E[\text{Cov}(a_{jk}, a_{j'k'} \mid \boldsymbol{\pi})] + \text{Cov}(E[a_{jk} \mid \boldsymbol{\pi}], E[a_{j'k'} \mid \boldsymbol{\pi}])}{\sqrt{\{E[V(a_{jk} \mid \boldsymbol{\pi})] + V(E[a_{jk} \mid \boldsymbol{\pi}])\}\{E[V(a_{j'k'} \mid \boldsymbol{\pi})] + V(E[a_{j'k'} \mid \boldsymbol{\pi}])\}}} \\ &= \frac{E[2\text{Cov}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi})] + \text{Cov}(2\pi_k, 2\pi_{k'})}{\sqrt{\{E[2\pi_k(1-\pi_k)] + V(2\pi_k)\}\{E[2\pi_{k'}(1-\pi_{k'})] + V(2\pi_{k'})\}}} \\ &= \frac{E[\text{Cov}(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi})] + 2\text{Cov}(\pi_k, \pi_{k'})}{\sqrt{\{E(\pi_k) - E(\pi_k^2) + 2V(\pi_k)\}\{E(\pi_{k'}) - E(\pi_{k'}^2) + 2V(\pi_{k'})\}}}\end{aligned}$$

$$= \begin{cases} \frac{E\left[(1-\theta)^g(\pi_k - \pi_k^2)\right] + 2V(\pi_k)}{E(\pi_k) - E^2(\pi_k) + V(\pi_k)} & \text{if } k = k' \\ \frac{E[-(1-\theta)^g \pi_k \pi_{k'}] + 2\text{Cov}(\pi_k, \pi_{k'})}{\sqrt{\{E(\pi_k) - E^2(\pi_k) + V(\pi_k)\}\{E(\pi_{k'}) - E^2(\pi_{k'}) + V(\pi_{k'})\}}} & \text{if } k \neq k' \end{cases}$$

$$= \begin{cases} \frac{(1-\theta)^g\left[E(\pi_k) - V(\pi_k) - E^2(\pi_k)\right] + 2V(\pi_k)}{E(\pi_k) - E^2(\pi_k) + V(\pi_k)} & \text{if } k = k' \\ \frac{-(1-\theta)^g[\text{Cov}(\pi_k, \pi_{k'}) + E(\pi_k)E(\pi_{k'})] + 2\text{Cov}(\pi_k, \pi_{k'})}{\sqrt{\{E(\pi_k) - E^2(\pi_k) + V(\pi_k)\}\{E(\pi_{k'}) - E^2(\pi_{k'}) + V(\pi_{k'})\}}} & \text{if } k \neq k' \end{cases}$$

$$= \begin{cases} (1-\theta)^g + [1-(1-\theta)^g] \frac{2V_F(\pi_k)}{E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)} & \text{if } k = k' \\ \frac{2Cov_F(\pi_k, \pi_{k'}) - (1-\theta)^g[\text{Cov}_F(\pi_k, \pi_{k'}) + E_F(\pi_k)E_F(\pi_{k'})]}{\sqrt{\left[E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)\right]\left[E_F(\pi_{k'}) - E_F^2(\pi_{k'}) + V_F(\pi_{k'})\right]}} & \text{if } k \neq k', \end{cases}$$

as desired. □

A.1.2 Theorem 1: Distribution of Test Statistics

Theorem 1. *Consider an admixed population with K ancestral populations, g generations since admixture, and admixture proportions distributed according to $\boldsymbol{\pi} \sim F$, where F has finite first and second moments. For loci $j \in \{1, ..., m\}$ and ancestry components $k \in \{1, ..., K\}$, define test statistics $Z_{jk} = \frac{\hat{\beta}_{jk}}{\widehat{se}(\hat{\beta}_{jk})}$ based on Model (3.1) in Chapter 3. Then, under the universal null hypothesis ($\beta_{jk} = 0 \ \forall j, k$), the collection of test statistics $\mathbf{Z} = (Z_{11} \ \cdots \ Z_{mK})^T$ has an asymptotic multivariate normal distribution with mean $\mathbf{0}$ and covariance (and correlation) given by*

$$Cov(Z_{jk}, Z_{j'k'}) = \begin{cases} (1-\theta)^g & \text{if } k = k' \\ -(1-\theta)^g \frac{E_F(\pi_k)E_F(\pi_{k'})}{\sqrt{E_F(\pi_k)[1-E_F(\pi_k)]E_F(\pi_{k'})[1-E_F(\pi_{k'})]}} & \text{if } k \neq k', \end{cases}$$

where θ is the recombination fraction between loci j, j'.

Proof. Let individuals, indexed by $i = 1, ..., n$, come from an admixed population with K ancestral populations and g generations since admixture. Denote the admixture proportions for each individual by $\boldsymbol{\pi}_i = \begin{pmatrix} \pi_{i1} & \cdots & \pi_{iK} \end{pmatrix}^T$, $\sum_{k=1}^K \pi_{ik} = 1$, and let that vector be drawn from a distribution F with finite first and second moments. Let $\mathbf{a}_{ij} = \begin{pmatrix} a_{ij1} & \cdots & a_{ijK} \end{pmatrix}^T$ be the local ancestry vector for individual i at locus j, where $\sum_{k=1}^K a_{ijk} = 2$ and a_{ijk} denotes the number of alleles inherited by individual i from ancestral population k at locus j. Define Wald test statistics $Z_{jk} = \hat{\beta}_{jk}/\widehat{se}(\hat{\beta}_{jk})$ based on the marginal linear regression model

$$E[y_i \mid a_{ijk}, \boldsymbol{\pi}_i] = \alpha + \beta_{jk} a_{ijk} + \boldsymbol{\gamma} \boldsymbol{\pi}_{i,-K},$$

for some quantitative trait of interest $\mathbf{y}$. Suppose that the universal null hypothesis holds, such that $\beta_{jk} = 0 \ \forall j \in \{1, ..., m\}, k \in \{1, ..., K\}$. We must show that the collection of test statistics $\mathbf{Z}$ is asymptotically multivariate normal with mean $\mathbf{0}$ and covariance as defined above.

It is straightforward to show (e.g., by using the asymptotic equivalence between Wald tests and score tests, combined with existing results about the asymptotic distribution of

score tests from such a model [53, 102]) that the test statistics $\mathbf{Z}$ are asymptotically multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$ with elements $\mathbf{\Sigma}_{jk,j'k'} = \text{Cov}(Z_{jk}, Z_{j'k'})$. Furthermore, we can show (e.g., as in Joo et al. [160]) that test statistics $\mathbf{Z}$ from the unadjusted admixture mapping model (the regression model defined above without $\boldsymbol{\pi}$) have covariance $\text{Cov}(Z_{jk}, Z_{j'k'}) = \text{Corr}(a_{jk}, a_{j'k'})$. From the adjusted model (the regression model defined above *with* $\boldsymbol{\pi}$), however, the covariance of test statistics is simply the correlation of local ancestry conditioned on the covariates $\boldsymbol{\pi}$ [53]: $\text{Cov}(Z_{jk}, Z_{j'k'}) = \text{Corr}(a_{jk}, a_{j'k'} \mid \boldsymbol{\pi})$. Combining these results with Lemma 1, we see that asymptotically the test statistics $\mathbf{Z}$ will have covariance defined by $\text{Cov}(Z_{jk}, Z_{j'k'}) = \text{Corr}(a_{jk}, a_{j'k'} \mid \boldsymbol{\pi} = E_F[\boldsymbol{\pi}])$, so

$$\text{Cov}(Z_{jk}, Z_{j'k'}) = \begin{cases} (1-\theta)^g & \text{if } k = k', \\ (1-\theta)^g \frac{-E[\pi_k]E[\pi_{k'}]}{\sqrt{E[\pi_k](1-E[\pi_k])E[\pi_{k'}](1-E[\pi_{k'}])}} & \text{if } k \neq k', \end{cases}$$

as desired. □

Corollary 1. *When $K = 2$,*

$$Cov(Z_{jk}, Z_{j'k'}) = \begin{cases} (1-\theta)^g \approx \exp(-0.01g\delta) & \text{if } k = k' \\ -(1-\theta)^g \approx -\exp(-0.01g\delta) & \text{if } k \neq k', \end{cases}$$

where δ is the genetic distance (cM) between loci j, j', and the distribution of admixture mapping test statistics can be approximated by an Ornstein-Uhlenbeck process.

Proof. Consider an admixed population with two ancestral populations (i.e., $K = 2$). Then from Theorem 1 we have

$$\text{Cov}(Z_{jk}, Z_{j'k'}) = \begin{cases} (1-\theta)^g & \text{if } k = k', \\ (1-\theta)^g \frac{-E[\pi_1]E[\pi_2]}{\sqrt{E[\pi_1](1-E[\pi_1])E[\pi_2](1-E[\pi_2])}} & \text{if } k \neq k'. \end{cases}$$

But $\frac{E[\pi_1]E[\pi_2]}{\sqrt{E[\pi_1](1-E[\pi_1])E[\pi_2](1-E[\pi_2])}} = \frac{E[\pi_1]E[\pi_2]}{\sqrt{E[\pi_1](E[\pi_2])E[\pi_2](E[\pi_1])}} = 1$ since $\sum_{k=1}^{K} E(\pi_k) = 1$. Furthermore, from Siegmund and Yakir [58] we know that $(1-\theta)^g \approx \exp(-0.01g\delta)$, where θ, δ are

the recombination fraction and genetic distance (in cM) between two loci, respectively. It follows that

$$\mathrm{Cov}(Z_{jk}, Z_{j'k'}) = \begin{cases} (1-\theta)^g \approx \exp(-0.01g\delta) & \text{if } k = k' \\ -(1-\theta)^g \approx -\exp(-0.01g\delta) & \text{if } k \neq k'. \end{cases}$$

Finally, since $\mathbf{Z}$ has this covariance structure and is also a Gaussian process with mean $\mathbf{0}$ (from Theorem 1), the distribution of admixture mapping test statistics $\mathbf{Z}$ can be approximated by an Ornstein-Uhlenbeck process [58], as desired. □

A.2 Test Statistic Simulation Algorithm

In Chapter 3, we propose a simulation-based multiple testing correction approach that simulates admixture mapping test statistics $\mathbf{Z} = \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{mK} \end{pmatrix}^\top$ from their asymptotic joint distribution provided by Theorem 1. There are many approaches that could be taken to simulate tests statistics from this distribution. We propose the following recursive algorithm, which takes advantage of the convenient form of the covariance matrix of this asymptotic distribution to simulate test statistics quickly and with low memory costs. It is easy to show that this recursive simulation algorithm generates a collection of test statistics $\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_m \end{pmatrix}^\top = \begin{pmatrix} Z_{11} \cdots Z_{1K} & \cdots & Z_{m1} \cdots Z_{mK} \end{pmatrix}^\top$ with the appropriate distribution given by Theorem 1.

A.2.1 The Algorithm

1. Set $\mathbf{Z}_1 = \begin{pmatrix} Z_{11} & \cdots & Z_{1K} \end{pmatrix}^\top = \mathbf{L}\mathbf{X}_1$, where $\mathbf{X}_1 \sim N_{K-1}(\mathbf{0}, \mathbf{I}_{K-1})$.

2. Set $\mathbf{Z}_2 = a_{12}\mathbf{Z}_1 + b_{12}\mathbf{L}\mathbf{X}_2$, where $\mathbf{X}_2 \sim N_{K-1}(\mathbf{0}, \mathbf{I}_{K-1})$.

3. Set $\mathbf{Z}_3 = a_{23}\mathbf{Z}_2 + b_{23}\mathbf{L}\mathbf{X}_3$, where $\mathbf{X}_3 \sim N_{K-1}(\mathbf{0}, \mathbf{I}_{K-1})$

...

m. Set $\mathbf{Z}_m = a_{m-1,m}\mathbf{Z}_{m-1} + b_{m-1,m}\mathbf{L}\mathbf{X}_m$, where $\mathbf{X}_m \sim N_{K-1}(\mathbf{0}, \mathbf{I}_{K-1})$

Here, K is the number of ancestral populations and m is the number of loci. The matrix $\mathbf{L}$ is the $K \times (K-1)$-dimensional square root of $\mathbf{\Sigma}$ (i.e., $\mathbf{L}\mathbf{L}^\top = \mathbf{\Sigma}$), where $\mathbf{\Sigma}$ is the covariance matrix for admixture mapping test statistics at a single locus:

$$\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{Z}_j) = \begin{pmatrix} 1 & \cdots & \frac{-E(\pi_1)E(\pi_K)}{\sqrt{E(\pi_1)[1-E(\pi_1)]E(\pi_K)[1-E(\pi_K)]}} \\ \vdots & \ddots & \vdots \\ \frac{-E(\pi_1)E(\pi_K)}{\sqrt{E(\pi_1)[1-E(\pi_1)]E(\pi_K)[1-E(\pi_K)]}} & \cdots & 1 \end{pmatrix}.$$

For admixed populations with 2 or 3 ancestral populations, we have derived the form of $\mathbf{L}$:

$$\mathbf{L}_{K=2} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\mathbf{L}_{K=3} = \begin{pmatrix} 1 & 0 \\ \frac{-E(\pi_1)E(\pi_2)}{\sqrt{E(\pi_1)[1-E(\pi_1)]E(\pi_2)[1-E(\pi_2)]}} & \sqrt{1-\left(\frac{-E(\pi_1)E(\pi_2)}{\sqrt{E(\pi_1)[1-E(\pi_1)]E(\pi_2)[1-E(\pi_2)]}}\right)^2} \\ \frac{-E(\pi_1)E(\pi_3)}{\sqrt{E(\pi_1)[1-E(\pi_1)]E(\pi_3)[1-E(\pi_3)]}} & -\sqrt{1-\left(\frac{-E(\pi_1)E(\pi_3)}{\sqrt{E(\pi_1)[1-E(\pi_1)]E(\pi_3)[1-E(\pi_3)]}}\right)^2} \end{pmatrix}.$$

When $K > 3$ we can find $\mathbf{L}$ by taking the Cholesky decomposition of Σ (e.g., using the `mroot` function in the `mgcv` package [163]). In practice, we replace the population mean admixture proportions $E(\pi_k)$ in $\mathbf{L}, \mathbf{\Sigma}$ with their sample equivalents ($\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ik}$). Finally, the scalars a_{ij}, b_{ij} depend on the number of generations since admixture, g, and the genetic distance (in cM), δ_{ij}, between consecutive loci i and j:

$$a_{ij} = \exp(-0.01g\delta_{ij})$$
$$b_{ij} = \sqrt{1-\exp(-2\times 0.01g\delta_{ij})}$$

A.2.2 Computation Time and Recommendations

Others [53, 160] have proposed multiple-testing correction procedures that similarly utilize knowledge of the asymptotic distribution of test statistics; however, our approach takes advantage of the specific, convenient form of this distribution for admixture mapping studies

to speed up computation time. Note that this algorithm scales linearly with the number of loci m, but run time does not depend on the number of samples n (except through calculation of the first and second moments of the sample admixture proportions). Run time does increase slightly, but not drastically, with increasing number of ancestral populations K. Running 10,000 replicates on the WHI SHARe data took approximately 8 and 9 min for the African American ($K = 2$) and Hispanic American ($K = 3$) samples, respectively.

Computation time can be drastically reduced by considering just a single locus per unique ancestry block. This is particularly relevant when local ancestry is inferred using software that calls local ancestry within windows (e.g., `RFMix` [10]). In the WHI SHARe data we also saw that this consideration of a single locus per ancestry block led to better estimates of the genome-wide significance threshold, so we recommend that this thinning step be utilized in future studies.

We have implemented this algorithm in our R package `STEAM` (Significance Threshold Estimation for Admixture Mapping), which is available on GitHub.

A.3 Estimating the Generations Since Admixture

In Chapter 3 we propose an approach that uses non-linear least squares (NLS) estimation, combined with our theoretical results from Lemma 1, to estimate the number of generations since admixture g from the observed patterns of local ancestry correlation in a sample of admixed individuals. Briefly, we wish to find the value of g that minimizes the equations

$$\begin{aligned} &\sum_{j,j'} \left(\text{Corr}(a_{jk}, a_{j'k'}) - [a^*_k + (1 - a^*_k)(1 - \theta_{jj'})^g]\right)^2 \text{ when } k = k' \\ &\sum_{j,j'} \left(\text{Corr}(a_{jk}, a_{j'k'}) - [2b^*_{kk'} - (b^*_{kk'} + c^*_{kk'})(1 - \theta_{jj'})^g]\right)^2 \text{ when } k \neq k', \end{aligned} \tag{A.2}$$

where $\text{Corr}(a_{jk}, a_{j'k'})$ is the correlation of local ancestry components k, k' at loci j, j' separated by recombination fraction $\theta_{jj'}$. In practice, we replace $\text{Corr}(a_{jk}, a_{j'k'})$ and $\theta_{jj'}$ with their estimates, and then use either the *constrained* or *unconstrained* approach to estimate the scalars $a^*_k, b^*_{kk'}, c^*_{kk'}$.

A.3.1 Constrained Non-Linear Least Squares

The *constrained* NLS approach takes advantage of the known form of the scalars $a_k^*, b_{kk'}^*, c_{kk'}^*$ given by Lemma 1:

$$a_k^* = \frac{2V(\pi_k)}{E(\pi_k) - E^2(\pi_k) + V(\pi_k)},$$

$$b_{kk'}^* = \frac{\mathrm{Cov}(\pi_k, \pi_{k'})}{\sqrt{[E(\pi_k) - E^2(\pi_k) + V(\pi_k)][E(\pi_{k'}) - E^2(\pi_{k'}) + V(\pi_{k'})]}},$$

and

$$c_{kk'}^* = \frac{E(\pi_k)E(\pi_{k'})}{\sqrt{[E(\pi_k) - E^2(\pi_k) + V(\pi_k)][E(\pi_{k'}) - E^2(\pi_{k'}) + V(\pi_{k'})]}}.$$

We set the parameters $a_k^*, b_{kk'}^*, c_{kk'}^*$ in Equation (A.2) equal to these forms, replacing population means and covariances with their sample equivalents. This leaves a single unknown parameter, g, to be estimated using non-linear least squares.

A.3.2 Unconstrained Non-Linear Least Squares

The *unconstrained* approach ignores our knowledge of the form of $a_k^*, b_{kk'}^*, c_{kk'}^*$. Instead, we treat these scalars as parameters to be estimated, along with g, using non-linear least squares.

A.3.3 Comparison of Constrained and Unconstrained Approaches in WHI SHARe Data

We calculated the correlation of inferred local ancestry vectors in WHI SHARe at a thinned set of pairs of loci. Then, we ran constrained and unconstrained non-linear least squares on the observed local ancestry correlation curves. Results are summarized in Table A.1 and Figures A.1, A.2. The two approaches yielded similar estimates of the number of generations since admixture and genome-wide p-value thresholds in the African American and Hispanic American cohorts (Table A.1). The unconstrained approach led to a slightly better correspondence between the fitted values from NLS and the observed local ancestry correlation curves (Figures A.1, A.2). This result fits with our intuition, as the unconstrained approach allows us to estimate the scalars a^*, b^*, c^* that provide the best fit to the data, along with estimating g. We used the estimates from the unconstrained approach for further analyses

Table A.1: Comparison of the constrained and unconstrained non-linear least squares estimation approaches in WHI SHARe African Americans and Hispanic Americans.
We present the estimated number of generations since admixture ($\hat{g}$), the mean squared error (MSE) comparing fitted values from NLS to the observed local ancestry correlation curves, and the genome-wide p-value threshold from our test statistic simulation approach using each estimate $\hat{g}$.

	African Americans			Hispanic Americans		
	$\hat{g}$	MSE	Threshold	$\hat{g}$	MSE	Threshold
Constrained	6.2	8.9×10^{-5}	1.97×10^{-5}	9.2	1.4×10^{-4}	4.51×10^{-6}
Unconstrained	5.9	3.0×10^{-5}	2.06×10^{-5}	9.6	5.5×10^{-5}	4.47×10^{-6}

of the WHI SHARe data, but both approaches are implemented in our R package `STEAM`, and the decision about which approach to use is left to the user's discretion.

A.4 Consideration of Binary Traits

A.4.1 Simulation Methods

To explore the validity of our theoretical work for binary traits, we simulated traits and local ancestry at pairs of loci for admixed individuals in a variety of populations. For each admixed individual $i = 1, \ldots, n$, we first drew admixture proportions from a pre-specified distribution F representing different population structure scenarios: no structure ($\boldsymbol{\pi}_i = \boldsymbol{\pi}\ \forall i$), subpopulations ($\boldsymbol{\pi}_i = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_S\}$ with probability $\{p_1, p_2, \ldots p_S\}$, where $\sum_{s=1}^{S} p_s = 1$), or Dirichlet admixture proportions ($\boldsymbol{\pi}_i \sim_{iid} \text{Dirichlet}(\boldsymbol{\alpha})$). We considered various choices of number of individuals (n), number of ancestral populations (K), and hyperparameters for the distribution of admixture proportions F. For each haplotype, we independently simulated crossover events between two loci separated by recombination fraction θ across g generations according to a Poisson process. We simulated ancestry at the first locus according to a Multinoulli (categorical) distribution with probabilities equal to the vector of admixture proportions $\boldsymbol{\pi}_i$. Using the simulated crossover history, we determined whether any recombination had occurred between the two loci; if so, we independently simulated

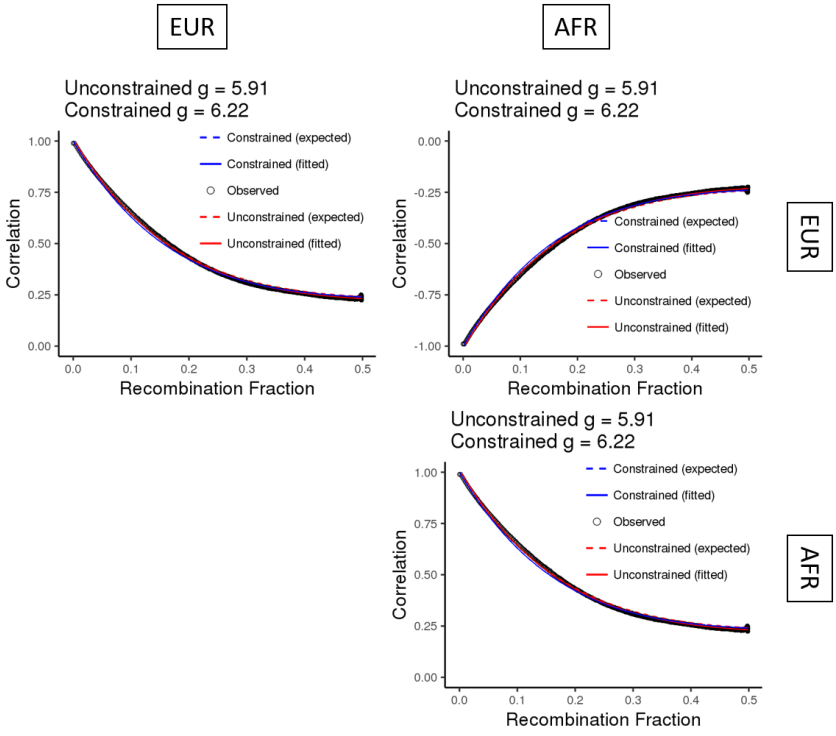

Figure A.1: Correspondence between observed local ancestry correlation in WHI SHARe African Americans and expected and fitted values based on non-linear least squares estimation.

Each panel presents the local ancestry correlation curves for a pair of ancestry components (European ancestry at both loci, European at one locus and African at the other, or African at both loci). The black dots represent the *observed* local ancestry correlation. Dashed lines represent the *expected* correlation based on Lemma 1, setting $g = \hat{g}$ from each non-linear least squares approach (blue = constrained, red = unconstrained). Solid lines represent the *fitted* values from each non-linear least squares approach (i.e., Lemma 1, replacing g with $\hat{g}$ and the terms depending on $E(\boldsymbol{\pi}), \text{Cov}(\boldsymbol{\pi})$ with $\hat{a}, \hat{b}, \hat{c}$ defined in Section A.3.1 (constrained) or estimated by NLS (unconstrained)). For the constrained approach, the *expected* and *fitted* values are identical.

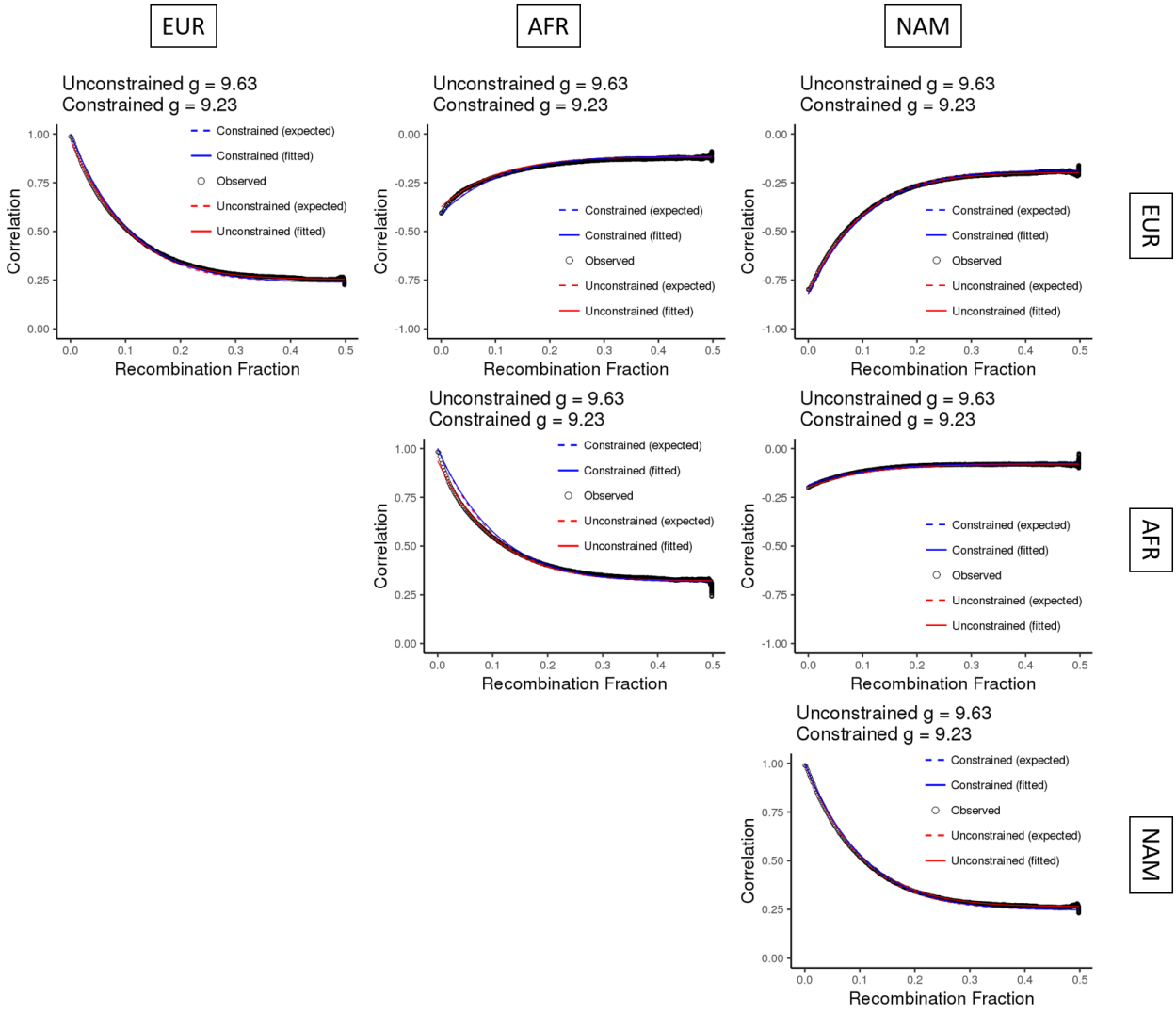

Figure A.2: Correspondence between observed local ancestry correlation in WHI SHARe Hispanics/Latinos and expected and fitted values based on non-linear least squares estimation.

Each panel presents the local ancestry correlation curves for a pair of ancestry components (European ancestry at both loci, European at one locus and African at the other, European at one locus and Native American at the other, etc.). The black dots represent the *observed* local ancestry correlation. Dashed lines represent the *expected* correlation based on Lemma 1, setting $g = \hat{g}$ from each non-linear least squares approach (blue = constrained, red = unconstrained). Solid lines represent the *fitted* values from each non-linear least squares approach.

ancestry at the second locus according to the same Multinoulli distribution; if not, we set ancestry at the second locus equal to ancestry at the first. We paired haplotypes to create diploid individuals and then simulated binary traits for each individual according to the model $y_i \sim_{iid} \text{Bernoulli}(0.2)$ and quantitative traits according to $y_i \sim_{iid} N(0, 1)$. Using these simulated traits, we calculated admixture mapping test statistics at each locus. We repeated this process 10,000 times and calculated the correlation of admixture mapping test statistics at the two loci across simulation replicates, then compared the observed patterns of correlation to the expected correlation given by our theoretical results (Theorem 1).

A.4.2 Simulation Results

Results for a simulated admixed population with three ancestral populations, 10,000 individuals, and admixture proportions drawn from a uniform distribution are presented in Figure A.3. We see nearly identical correspondence between the observed test statistic correlation and the expected correlation based on Theorem 1, regardless of whether the trait is binary (panel A) or quantitative (panel B).

A.5 Software Availability

An R package, `STEAM` (Significance Threshold Estimation for Admixture Mapping), that implements the methods described in Chapter 3 and this Appendix is available on GitHub: `https://github.com/kegrinde/STEAM`.

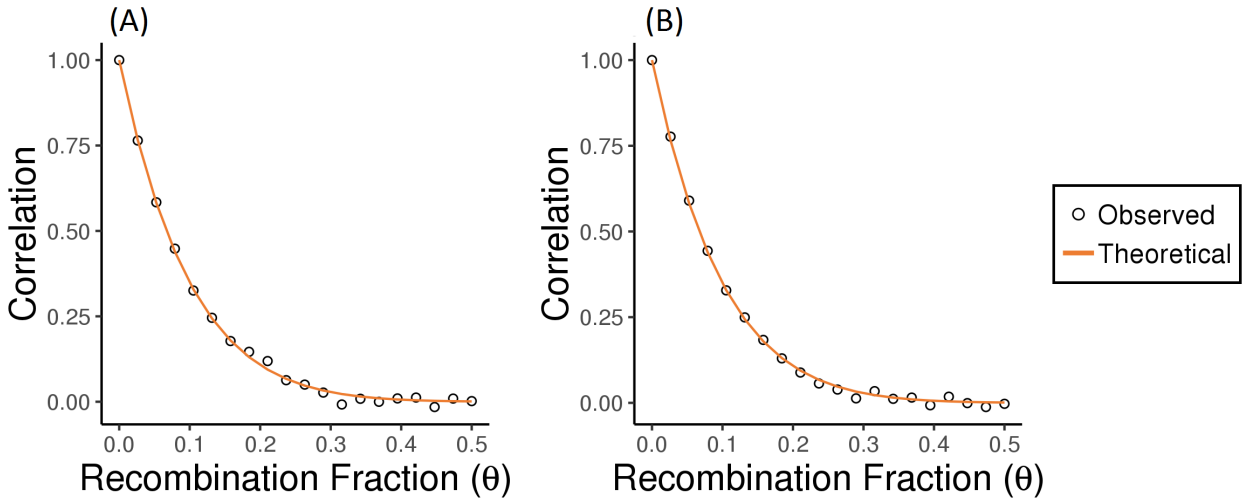

Figure A.3: Correlation of admixture mapping test statistics in simulated data with a quantitative or binary trait.
(A) Comparison of the observed and expected (theoretical) correlation of admixture mapping test statistics, testing the first ancestry component at two loci separated by a recombination fraction θ, in a simulated admixed population with $K = 3, n = 10,000, \boldsymbol{\pi}_i \sim_{iid} \text{Dirichlet}\left(\begin{pmatrix}1 & 1 & 1\end{pmatrix}\right)$, and a simulated binary trait.
(B) Comparison of the observed and expected (theoretical) correlation of admixture mapping test statistics, testing the first ancestry component at two loci separated by a recombination fraction θ, in a simulated admixed population with $K = 3, n = 10,000, \boldsymbol{\pi}_i \sim_{iid} \text{Dirichlet}\left(\begin{pmatrix}1 & 1 & 1\end{pmatrix}\right)$, and a simulated quantitative trait.

Appendix B

APPENDIX FOR CHAPTER 4

B.1 Validation of Theoretical Results

In Chapter 4, we provided the expected effect size estimates from genetic association studies at causal and neutral SNPs using different techniques for adjusting for ancestral heterogeneity. In this Appendix, we provide details and simulation studies validating these analytic results.

B.1.1 The data-generating mechanism

We consider an admixed population with two ancestral populations, n individuals, and admixture proportions $\boldsymbol{\pi}_i = \begin{pmatrix} \pi_i & 1-\pi_i \end{pmatrix}^\top$ that are allowed to vary across the population. We refer to the two ancestral populations as *Ancestral Population 1* and *Ancestral Population 2*, with π_i representing the genome-wide proportion of genetic material inherited by individual i from Ancestral Population 1 and $1-\pi_i$ representing the proportion of genetic material inherited from Ancestral Population 2. We denote local ancestry by $\mathbf{a}_{ij} = \begin{pmatrix} a_{ij} & 2-a_{ij} \end{pmatrix}^\top$, where a_{ij} and $2-a_{ij}$ are the number of alleles inherited by individual i from Ancestral Populations 1 and 2, respectively, at locus j. Genotypes, quantified as the number of copies of some pre-specified allele carried by individual i at locus j, are denoted represented by g_{ij}. We consider two *unlinked* loci $j = 1, 2$ (e.g., loci on separate chromosomes) and assume that data are generated according to the following hierarchical model:

$$\begin{aligned} \pi_i &\overset{\text{i.i.d.}}{\sim} F \text{ for some distribution } F \\ a_{ij} \mid \pi_i &\overset{\text{i.i.d}}{\sim} \text{Binomial}(2, \pi_i),\ j = 1, 2 \\ g_{ij} \mid a_{ij}, \mathbf{p}_j &\overset{\text{ind.}}{\sim} \text{Binomial}(a_{ij}, p_{j1}) + \text{Binomial}(2-a_{ij}, p_{j2}),\ j = 1, 2 \end{aligned}$$

where p_{j1}, p_{j2} are allele frequencies at locus j in Ancestral Populations 1 and 2, respectively. Note that since the two loci under consideration are unlinked, we assume that local ancestry and genotypes at these loci are conditionally independent.

We assume that our quantitative trait of interest $\mathbf{y}$ depends only on the genotype at locus 1, and we allow for the possibility that the admixture proportions $\boldsymbol{\pi}$ have a direct effect on the trait (e.g., through environmental differences across ancestral populations). More specifically, we assume that this trait is generated according to

$$y_i = \beta_0 + \beta_1 g_{i1} + \beta_\pi \pi_i + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d}}{\sim} (0, \sigma^2_\epsilon).$$

We refer to β_1 and β_2 as the true *effect sizes* of loci 1 and 2, respectively. Since the trait only depends on the genotype at locus 1, the true effect size of locus 2 is $\beta_2 = 0$. We are interested in using genome-wide association studies (GWAS) and admixture mapping studies to investigate the association between loci 1 and 2 and the trait of interest.

Assuming that data are generated according to the above-described mechanisms, and defining $E_\pi := E(\pi)$ and $V_\pi := \text{Var}(\pi)$, then the following statements are true:

- $E(a_j) = 2E_\pi,\ j = 1, 2$
- $V(a_j) = 2\{V_\pi + E_\pi(1 - E_\pi)\},\ j = 1, 2$
- $\text{Cov}(a_1, a_2) = 4V_\pi$
- $\text{Cov}(a_j, \pi) = 2V_\pi,\ j = 1, 2$
- $E(g_j) = 2\{p_{j2} + (p_{j1} - p_{j2})E_\pi\},\ j = 1, 2$
- $V(g_j) = 2[p_{j2}(1-p_{j2})+(p_{j1}-p_{j2})(1-p_{j1}-p_{j2})E_\pi+(p_{j1}-p_{j2})^2\{V_\pi+E_\pi(1-E_\pi)\}],\ j = 1, 2$
- $\text{Cov}(g_1, g_2) = 4(p_{11} - p_{12})(p_{21} - p_{22})V_\pi$
- $\text{Cov}(g_j, g_j) = 2(p_{j1} - p_{j2})\{V_\pi + E_\pi(1 - E\pi)\},\ j = 1, 2$

- $\text{Cov}(g_j, g_k) = 4(p_{j1} - p_{j2})V_\pi,\ j \neq k$
- $\text{Cov}(g_j, \pi) = 2(p_{j1} - p_{j2})V_\pi,\ j = 1, 2$

Furthermore, suppose we define a random variable $z_g = h(g_1, g_2) + e,\ e \sim (\mu_e, \sigma_e^2)$ for some function h. Then:

- $E(z_g) = \mu_e + E\{h(g_1, g_2)\}$
- $V(z_g) = \sigma_e^2 + V\{h(g_1, g_2)\}$
- $\text{Cov}(\pi, z_g) = \text{Cov}[\pi, E\{h(g_1, g_2) \mid \pi\}]$
- $\text{Cov}(a_j, z_g) = 2\text{Cov}(\pi, z_g) + E[\text{Cov}\{a_j, h(g_1, g_2) \mid \pi\}],\ j = 1, 2$
- $\text{Cov}(g_j, z_g) = 2(p_{j1} - p_{j2})\text{Cov}(\pi, z_x) + E[\text{Cov}\{g_j, h(g_1, g_2) \mid \pi\}],\ j = 1, 2$

These results are straightforward to derive, using our assumed hierarchical data-generating model and the laws of total expectation

$$E[x] = E\{E[x \mid y]\},$$

total variance

$$V[x] = V\{E[x \mid y]\} + E\{V[x \mid y]\},$$

and total covariance

$$\text{Cov}[x, y] = \text{Cov}\{E[x \mid z], E[y \mid z]\} + E\{\text{Cov}[x, y \mid z]\}.$$

B.1.2 General result

Suppose we fit an *unadjusted*, *admixture proportion adjusted*, or *principal component adjusted* genetic association model, as defined in Chapter 4. Then, the theory of linear models (combined with a lot of algebra) tells us that the effect size estimates from these models will take the following forms in expectation....

Unadjusted model

Model: $E[y_i \mid x_{ij}] = \beta_0 + \beta_j x_{ij}$.

Expected effect size estimate at SNP j:

$$E[\hat{\beta}_j] = \frac{\beta_1 \widehat{\mathrm{Cov}}(g_1, x_j) + \beta_\pi \widehat{\mathrm{Cov}}(\pi, x_j)}{\widehat{\mathrm{Var}}(x_j)},$$

where $\widehat{\mathrm{Var}}$ and $\widehat{\mathrm{Cov}}$ are the sample variance and covariance (for example, $\widehat{\mathrm{Var}}(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$).

Proof. Let $\boldsymbol{\pi}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{g}_1, \mathbf{g}_2$ be drawn from the hierarchical model specified above. Assume that the trait $\mathbf{y}$ is generated such that $\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{g}_1 + \beta_\pi\boldsymbol{\pi} + \boldsymbol{\epsilon}$, where ϵ_i are drawn *i.i.d.* from some distribution with mean 0 and variance σ_ϵ^2. Suppose that at locus j we fit the unadjusted genetic association regression model $E[\mathbf{y} \mid \mathbf{x}_j] = \beta_0\mathbf{1} + \beta_j\mathbf{x}_j$, for some predictor of interest $\mathbf{x}_j$ ($\mathbf{x}_j = \mathbf{g}_j$ for GWAS, $\mathbf{x}_j = \mathbf{a}_j$ for admixture mapping). Then, the estimated regression coefficients for this model will take the form

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_j \end{pmatrix} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \text{ for } \mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{x}_j \end{pmatrix},$$

with expected value

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}^*\boldsymbol{\beta}, \text{ for } \mathbf{X}^* = \begin{pmatrix} \mathbf{1} & \mathbf{g}_1 & \boldsymbol{\pi} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_\pi \end{pmatrix}.$$

But

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{1}^\top\mathbf{1} & \mathbf{1}^\top\mathbf{x}_j \\ \mathbf{x}_j^\top\mathbf{1} & \mathbf{x}_j^\top\mathbf{x}_j \end{pmatrix}^{-1} = \frac{1}{n\widehat{\mathrm{Var}}(x_j)} \begin{pmatrix} \widehat{\mathrm{Var}}(x_j) + \hat{E}(x_j)^2 & -\hat{E}(x_j) \\ -\hat{E}(x_j) & 1 \end{pmatrix}$$

and

$$\mathbf{X}^\top\mathbf{X}^* = \begin{pmatrix} \mathbf{1}^\top\mathbf{1} & \mathbf{1}^\top\mathbf{g}_1 & \mathbf{1}^\top\boldsymbol{\pi} \\ \mathbf{x}_j^\top\mathbf{1} & \mathbf{x}_j^\top\mathbf{g}_1 & \mathbf{x}_j^\top\boldsymbol{\pi} \end{pmatrix} = n \begin{pmatrix} 1 & \hat{E}(g_1) & \hat{E}(\pi) \\ \hat{E}(x_j) & \widehat{\mathrm{Cov}}(g_1, x_j) + \hat{E}(g_1)\hat{E}(x_j) & \widehat{\mathrm{Cov}}(\pi, x_j) + \hat{E}(\pi)\hat{E}(x_j) \end{pmatrix}.$$

It follows that

$$E[\hat{\boldsymbol{\beta}}] = \frac{1}{\widehat{\text{Var}}(x_j)} \begin{pmatrix} \widehat{\text{Var}}(x_j) & \widehat{\text{Var}}(x_j)\hat{E}(g_1) - \widehat{\text{Cov}}(x_j, g_1)\hat{E}(x_j) & \widehat{\text{Var}}(x_j)\hat{E}(\pi) - \hat{E}(x_j)\widehat{\text{Cov}}(x_j, \pi) \\ 0 & \widehat{\text{Cov}}(x_j, g_1) & \widehat{\text{Cov}}(x_j, \pi) \end{pmatrix} \boldsymbol{\beta},$$

and thus

$$E[\hat{\beta}_j] = \frac{\beta_1 \widehat{\text{Cov}}(x_j, g_1) + \beta_\pi \widehat{\text{Cov}}(\pi, x_j)}{\widehat{\text{Var}}(x_j)},$$

as desired. □

Admixture proportion adjusted model

Model: $E[y_i \mid x_{ij}, \boldsymbol{\pi}_i] = \beta_0 + \beta_j x_{ij} + \beta_\pi \pi_i$.

Expected effect size estimate at SNP j:

$$E[\hat{\beta}_j] = \beta_1 \frac{\widehat{\text{Var}}(\pi)\widehat{\text{Cov}}(g_1, x_j) - \widehat{\text{Cov}}(g_1, \pi)\widehat{\text{Cov}}(x_j, \pi)}{\widehat{\text{Var}}(\pi)\widehat{\text{Var}}(x_j) - \widehat{\text{Cov}}(x_j, \pi)^2}$$

Proof. (sketch) The proof of this result follows from similar arguments to that for the unadjusted model, simply replacing the design matrix $\mathbf{X}$ with $\begin{pmatrix} \mathbf{1} & \mathbf{x}_j & \boldsymbol{\pi} \end{pmatrix}$. With a bit of algebra, the rest follows. □

Principal component adjusted model

Model: $E[y_i \mid x_{ij}, \mathbf{v}_i] = \beta_0 + \beta_j x_{ij} + \beta_{v1} v_{i1} + \beta_{v2} v_{i2}$, where $v_{i1} = \pi_i \ \forall i$ and $v_{i2} = z_i$ for some random variable $\mathbf{z}$.

Expected effect size estimate at SNP j:

$$E[\hat{\beta}_j] = \beta_1 \frac{V_z(V_\pi C_{g_1,x_j} - C_{g_1,\pi} C_{x_j,\pi}) - V_\pi C_{g_1,z} C_{x_j,z} + C_{\pi,z}(C_{g_1,\pi} C_{x_j,z} + C_{g_1,z} C_{x_j,\pi} - C_{g_1,g} C_{\pi,z})}{V_z(V_\pi V_{x_j} - C_{x_j,\pi}^2) - V_\pi C_{x_j,z}^2 + C_{\pi,z}(2 C_{x_j,\pi} C_{x_j,z} - V_{x_j} C_{\pi,z})},$$

where $V_a = \widehat{\text{Var}}(a)$ and $C_{a,b} = \widehat{\text{Cov}}(a, b)$.

Proof. (sketch) Again, this proof follows from similar arguments to that for the unadjusted model, now replacing the design matrix $\mathbf{X}$ with $\begin{pmatrix} \mathbf{1} & \mathbf{x}_j & \boldsymbol{\pi} & \mathbf{z} \end{pmatrix}$. After making this substitution, the rest follows. □

B.1.3 GWAS and admixture mapping results

Combining the results from Sections B.1.1 and B.1.2, we can derive the estimated effect size estimates from GWAS and admixture mapping models in admixed populations with two ancestral populations. We consider studies with large sample sizes, such that we can replace the sample variance $\widehat{\text{Var}}$ and covariance $\widehat{\text{Cov}}$ with their population equivalent.

Unadjusted model

Below, we provide the expected effect size at the causal SNP (SNP 1) and a neutral unlinked SNP (SNP 2) from GWAS and admixture mapping models that do not adjust for ancestral heterogeneity. For notational simplicity, we drop the subscript i.

At the causal SNP (SNP 1),

$$\begin{aligned}
E[\hat{\beta}_1^{GWAS}] &= \frac{\beta_1 \text{Cov}(g_1, g_1) + \beta_\pi \text{Cov}(\pi, g_1)}{\text{Var}(g_1)} \\
&= \beta_1 + \frac{\beta_\pi V_\pi (p_{11} - p_{12})}{p_{12}(1-p_{12}) + (p_{11}-p_{12})(1-p_{11}-p_{12})E_\pi + (p_{11}-p_{12})^2(V_\pi + E_\pi - E_\pi^2)} \\
E[\hat{\beta}_1^{AMAP}] &= \frac{\beta_1 \text{Cov}(g_1, a_1) + \beta_\pi \text{Cov}(\pi, a_1)}{\text{Var}(a_1)} \\
&= \beta_1 (p_{11} - p_{12}) + \frac{\beta_\pi V_\pi}{V_\pi + E_\pi - E_\pi^2},
\end{aligned}$$

and at the unlinked neutral SNP (SNP 2),

$$\begin{aligned}
E[\hat{\beta}_2^{GWAS}] &= \frac{\beta_1 \text{Cov}(g_1, g_2) + \beta_\pi \text{Cov}(\pi, g_2)}{\text{Var}(g_2)} \\
&= \frac{(p_{21} - p_{22}) V_\pi [2\beta_1 (p_{11} - p_{12}) + \beta_\pi]}{p_{22}(1-p_{22}) + (p_{21}-p_{22})(1-p_{21}-p_{22})E_\pi + (p_{21}-p_{22})^2(V_\pi + E_\pi - E_\pi^2)} \\
E[\hat{\beta}_2^{AMAP}] &= \frac{\beta_1 \text{Cov}(g_1, a_2) + \beta_\pi \text{Cov}(\pi, a_2)}{\text{Var}(a_2)} \\
&= \frac{V_\pi [2\beta_1 (p_{11} - p_{12}) + \beta_\pi]}{V_\pi + E_\pi - E_\pi^2}.
\end{aligned}$$

Admixture proportion adjusted model

Next, we consider the GWAS and admixture mapping models that adjust for the true admixture proportions π_i. Again, we drop the subscript i for simplicity.

At the causal SNP (SNP 1),

$$\begin{aligned}
E[\hat{\beta}_1^{GWAS}] &= \beta_1 \frac{\text{Var}(\pi)\text{Cov}(g_1, g_1) - \text{Cov}(g_1, \pi)\text{Cov}(g_1, \pi)}{\text{Var}(\pi)\text{Var}(g_1) - \text{Cov}(g_1, \pi)^2} \\
&= \beta_1 \\
E[\hat{\beta}_1^{AMAP}] &= \beta_1 \frac{\text{Var}(\pi)\text{Cov}(g_1, a_1) - \text{Cov}(g_1, \pi)\text{Cov}(a_1, \pi)}{\text{Var}(\pi)\text{Var}(a_1) - \text{Cov}(a_1, \pi)^2} \\
&= \beta_1(p_{11} - p_{12}),
\end{aligned}$$

and at the unlinked neutral SNP (SNP 2),

$$\begin{aligned}
E[\hat{\beta}_2^{GWAS}] &= \beta_1 \frac{\text{Var}(\pi)\text{Cov}(g_1, g_2) - \text{Cov}(g_1, \pi)\text{Cov}(g_2, \pi)}{\text{Var}(\pi)\text{Var}(g_2) - \text{Cov}(g_2, \pi)^2} \\
&= 0 \\
E[\hat{\beta}_2^{AMAP}] &= \beta_1 \frac{\text{Var}(\pi)\text{Cov}(g_1, a_2) - \text{Cov}(g_1, \pi)\text{Cov}(a_2, \pi)}{\text{Var}(\pi)\text{Var}(a_2) - \text{Cov}(a_2, \pi)^2} \\
&= 0.
\end{aligned}$$

Principal component adjusted model

Finally, we consider GWAS and admixture mapping models that adjust for two principal components $(\mathbf{v}_{,1}, \mathbf{v}_{,2})$, where the first PC captures global ancestry ($v_{i,1} = \pi_i \ \forall i$) and the second PC captures some other feature quantified by the random variable $\mathbf{z}$ ($v_{i,2} = z_i \ \forall i$).

First, we provide results considering a general form of $\mathbf{z}$. At the causal SNP (SNP 1),

$$\begin{aligned}
E[\hat{\beta}_1^{GWAS}] &= \beta_1 \frac{V_z(V_\pi C_{g_1,g_1} - C_{g_1,\pi}C_{g_1,\pi}) - V_\pi C_{g_1,z}C_{g_1,z} + C_{\pi,z}(C_{g_1,\pi}C_{g_1,z} + C_{g_1,z}C_{g_1,\pi} - C_{g_1,g_1}C_{\pi,z})}{V_z(V_\pi V_{g_1} - C_{g_1,\pi}^2) - V_\pi C_{g_1,z}^2 + C_{\pi,z}(2C_{g_1,\pi}C_{g_1,z} - V_{g_1}C_{\pi,z})} \\
&= \beta_1 \\
E[\hat{\beta}_1^{AMAP}] &= \beta_1 \frac{V_z(V_\pi C_{g_1,a_1} - C_{g_1,\pi}C_{a_1,\pi}) - V_\pi C_{g_1,z}C_{a_1,z} + C_{\pi,z}(C_{g_1,\pi}C_{a_1,z} + C_{g_1,z}C_{a_1,\pi} - C_{g_1,a_1}C_{\pi,z})}{V_z(V_\pi V_{a_1} - C_{a_1,\pi}^2) - V_\pi C_{a_1,z}^2 + C_{\pi,z}(2C_{a_1,\pi}C_{a_1,z} - V_{a_1}C_{\pi,z})} \\
&= \beta_1(p_{11} - p_{12}) + \frac{V_\pi \beta_1 E[\text{Cov}(a_1, z \mid \pi)]\{E[\text{Cov}(g_1, z \mid \pi)] - (p_{11} - p_{12})E[\text{Cov}(a_1, z \mid \pi)]\}}{V_z(V_\pi V_{a_1} - C_{a_1,\pi}^2) - V_\pi C_{a_1,z}^2 + C_{\pi,z}(2C_{a_1,\pi}C_{a_1,z} - V_{a_1}C_{\pi,z})},
\end{aligned}$$

and at the unlinked neutral SNP (SNP 2),

$$\begin{aligned}
E[\hat{\beta}_2^{GWAS}] &= \beta_1 \frac{V_z(V_\pi C_{g_1,g_2} - C_{g_1,\pi}C_{g_2,\pi}) - V_\pi C_{g_1,z}C_{g_2,z} + C_{\pi,z}(C_{g_1,\pi}C_{g_2,z} + C_{g_1,z}C_{g_2,\pi} - C_{g_1,g_2}C_{\pi,z})}{V_z(V_\pi V_{g_2} - C_{g_2,\pi}^2) - V_\pi C_{g_2,z}^2 + C_{\pi,z}(2C_{g_2,\pi}C_{g_2,z} - V_{g_2}C_{\pi,z})} \\
&= \beta_1 \frac{-V_\pi E[\mathrm{Cov}(g_1, z \mid \pi)]E[\mathrm{Cov}(g_2, z \mid \pi)]}{V_z(V_\pi V_{g_2} - C_{g_2,\pi}^2) - V_\pi C_{g_2,z}^2 + C_{\pi,z}(2C_{g_2,\pi}C_{g_2,z} - V_{g_2}C_{\pi,z})} \\
E[\hat{\beta}_2^{AMAP}] &= \beta_1 \frac{V_z(V_\pi C_{g_1,a_2} - C_{g_1,\pi}C_{a_2,\pi}) - V_\pi C_{g_1,z}C_{a_2,z} + C_{\pi,z}(C_{g_1,\pi}C_{a_2,z} + C_{g_1,z}C_{a_2,\pi} - C_{g_1,a_2}C_{\pi,z})}{V_z(V_\pi V_{a_2} - C_{a_2,\pi}^2) - V_\pi C_{a_2,z}^2 + C_{\pi,z}(2C_{a_2,\pi}C_{a_2,z} - V_{a_2}C_{\pi,z})} \\
&= \beta_1 \frac{-V_\pi E[\mathrm{Cov}(g_1, z \mid \pi)]E[\mathrm{Cov}(a_2, z \mid \pi)]}{V_z(V_\pi V_{a_2} - C_{a_2,\pi}^2) - V_\pi C_{a_2,z}^2 + C_{\pi,z}(2C_{a_2,\pi}C_{a_2,z} - V_{a_2}C_{\pi,z})}.
\end{aligned}$$

Now, suppose $\mathbf{z} = \mathbf{z}_g = z_1\mathbf{g}_1 + z_2\mathbf{g}_2 + \mathbf{e},\ \mathbf{e} \sim (\mu_e, \sigma_e^2)$ for some scalars z_1, z_2. In other words, the 2nd PC captures genotypes at two SNPs, one of which is the causal SNP (SNP 1) and the other is an unlinked neutral SNP (SNP 2). Then, at the causal SNP (SNP 1),

$$\begin{aligned}
E[\hat{\beta}_1^{GWAS}] &= \beta_1 \\
E[\hat{\beta}_1^{AMAP}] &= \beta_1(p_{11} - p_{12}) \\
&\quad + \beta_1 \frac{-4z_1^2(p_{11} - p_{12})V_\pi(E_\pi - E_\pi^2 - V_\pi)[p_{12}(1 - p_{12}) + (p_{11} - p_{12})(1 - p_{11} - p_{12})E_\pi]}{V_z(V_\pi V_{a_1} - C_{a_1,\pi}^2) - V_\pi C_{a_1,z}^2 + C_{\pi,z}(2C_{a_1,\pi}C_{a_1,z} - V_{a_1}C_{\pi,z})},
\end{aligned}$$

and at the unlinked neutral SNP (SNP 2),

$$\begin{aligned}
E[\hat{\beta}_2^{GWAS}] &= \beta_1 \frac{-4z_1z_2V_\pi}{V_z(V_\pi V_{x_2} - C_{x_2,\pi}^2) - V_\pi C_{x_2,z}^2 + C_{\pi,z}(2C_{x_2,\pi}C_{x_2,z} - V_{x_2}C_{\pi,z})} \\
&\quad \times \prod_{j=1}^{2}[p_{j2}(1 - p_{j2}) + (p_{j1} - p_{j2})(1 - p_{j1} - p_{j2})E_\pi + (p_{j1} - p_{j2})^2(E_\pi - E_\pi^2 - V_\pi)] \\
E[\hat{\beta}_2^{AMAP}] &= \beta_1 \frac{-4z_1z_2(p_{21} - p_{22})V_\pi(E_\pi - E_\pi^2 - V_\pi)}{V_z(V_\pi V_{a_2} - C_{a_2,\pi}^2) - V_\pi C_{a_2,z}^2 + C_{\pi,z}(2C_{a_2,\pi}C_{a_2,z} - V_{a_2}C_{\pi,z})} \\
&\quad \times [p_{12}(1 - p_{12}) + (p_{11} - p_{12})(1 - p_{11} - p_{12})E_\pi + (p_{11} - p_{12})(E_\pi - E_\pi^2 - V_\pi)].
\end{aligned}$$

Figure B.1: Barplot of simulated admixture proportions.

B.1.4 Two locus simulations

To support these theoretical results, we performed a small simulation study. We generated data according to the data generating mechanism described in Section B.1.1, and then we compared the observed effect size estimates from GWAS and admixture mapping to the derived expected values provided in Section B.1.3.

We considered a variety of simulation settings, but present results from just a single setting here. Admixture proportions for $n = 5000$ individuals were generated from the distribution $F = \text{Beta}(7, 2)$ (see Figure B.1), allele frequencies at the causal SNP were 0.7 (ancestral population 1) and 0.2 (ancestral population 2), allele frequencies at the neutral SNP were 0.7 (ancestral population 1) and 0.2 (ancestral population 2), admixture proportions did not have a direct effect on the trait ($\beta_\pi = 0$), and the 2nd PC was generated according to $z_i = z_1 g_{i1} + z_2 g_{i2} + e_i$ for a range of values z_1, z_2 and added noise $e_i \overset{iid}{\sim} N(0, 0.25^2)$.

In Figures B.2, B.3, B.4, and B.5 we plot the effect size estimates from GWAS and admixture mapping models that we observe in our simulation study, and compare these observed effect size estimates to the expected effect sizes based on our analytic results (Section B.1.3), as well as the true effect sizes based on the data-generating mechanism (Section B.1.1). We see a perfect correspondence between the observed effect sizes and the expected effect

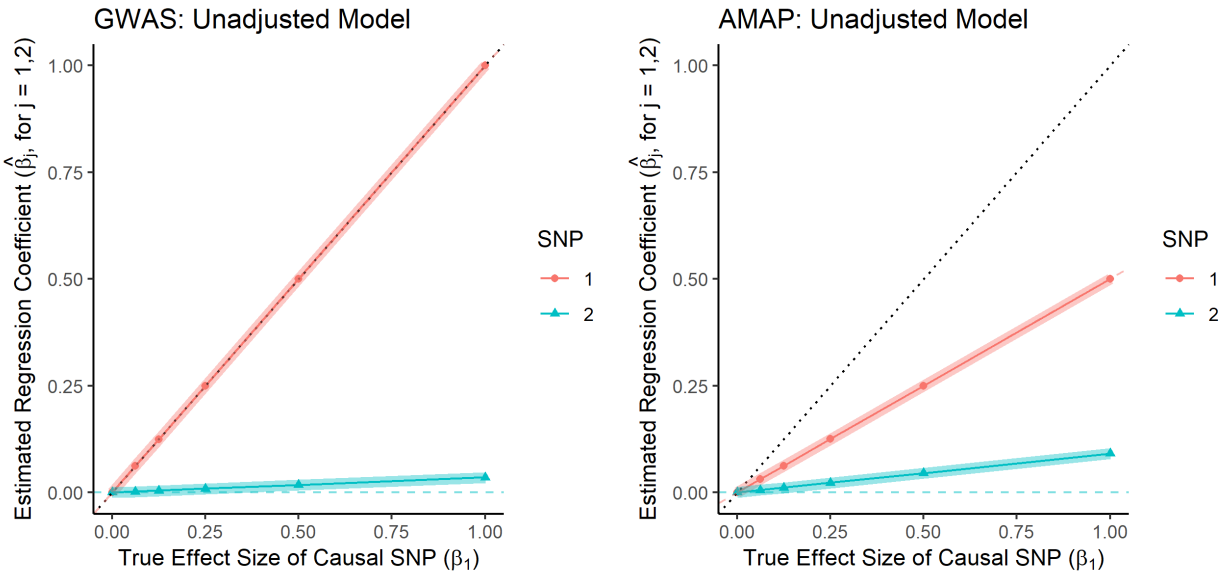

Figure B.2: Observed versus expected and true effect sizes from unadjusted GWAS and admixture mapping models.
Observed effect sizes = solid lines with points (red with dots = SNP 1, blue with triangles = SNP 2). Expected effect sizes = wider and faintly colored solid lines (red = SNP 1, blue = SNP 2). True effect sizes = dashed lines (red = SNP 1, blue = SNP 2). The $y = x$ line is also provided for reference (dotted black line).

sizes provided by our analytic results. Comparing these observed and expected effect sizes to the truth provides insight into the magnitude of bias that can be expected from each model.

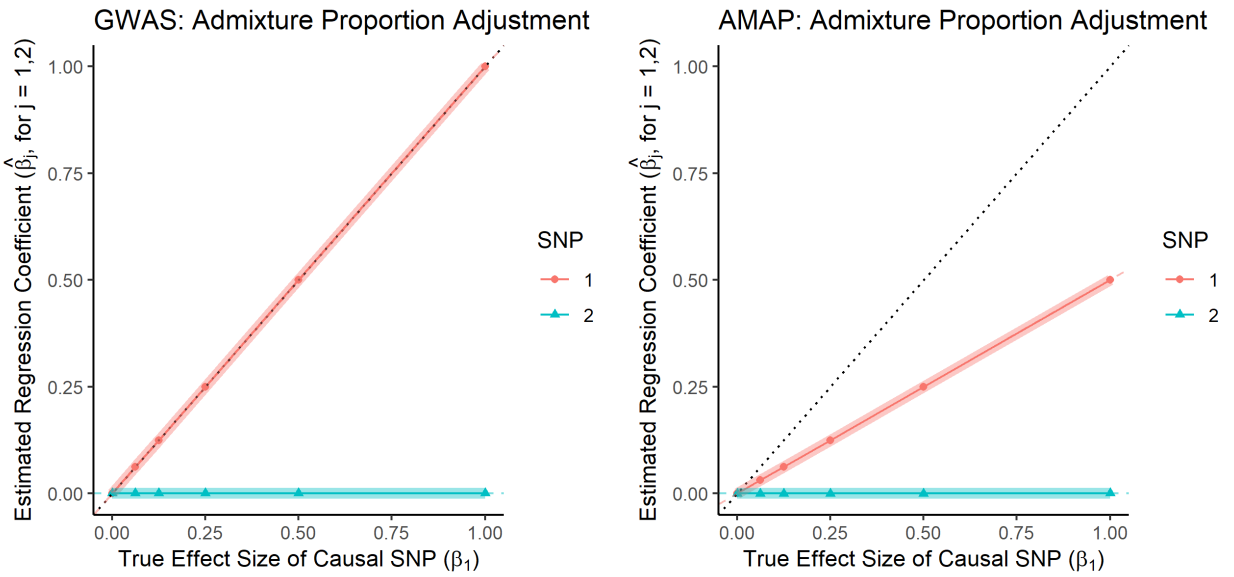

Figure B.3: Observed versus expected and true effect sizes from admixture proportion adjusted GWAS and admixture mapping models.
Observed effect sizes = solid lines with points (red with dots = SNP 1, blue with triangles = SNP 2). Expected effect sizes = wider and faintly colored solid lines (red = SNP 1, blue = SNP 2). True effect sizes = dashed lines (red = SNP 1, blue = SNP 2). The $y = x$ line is also provided for reference (dotted black line).

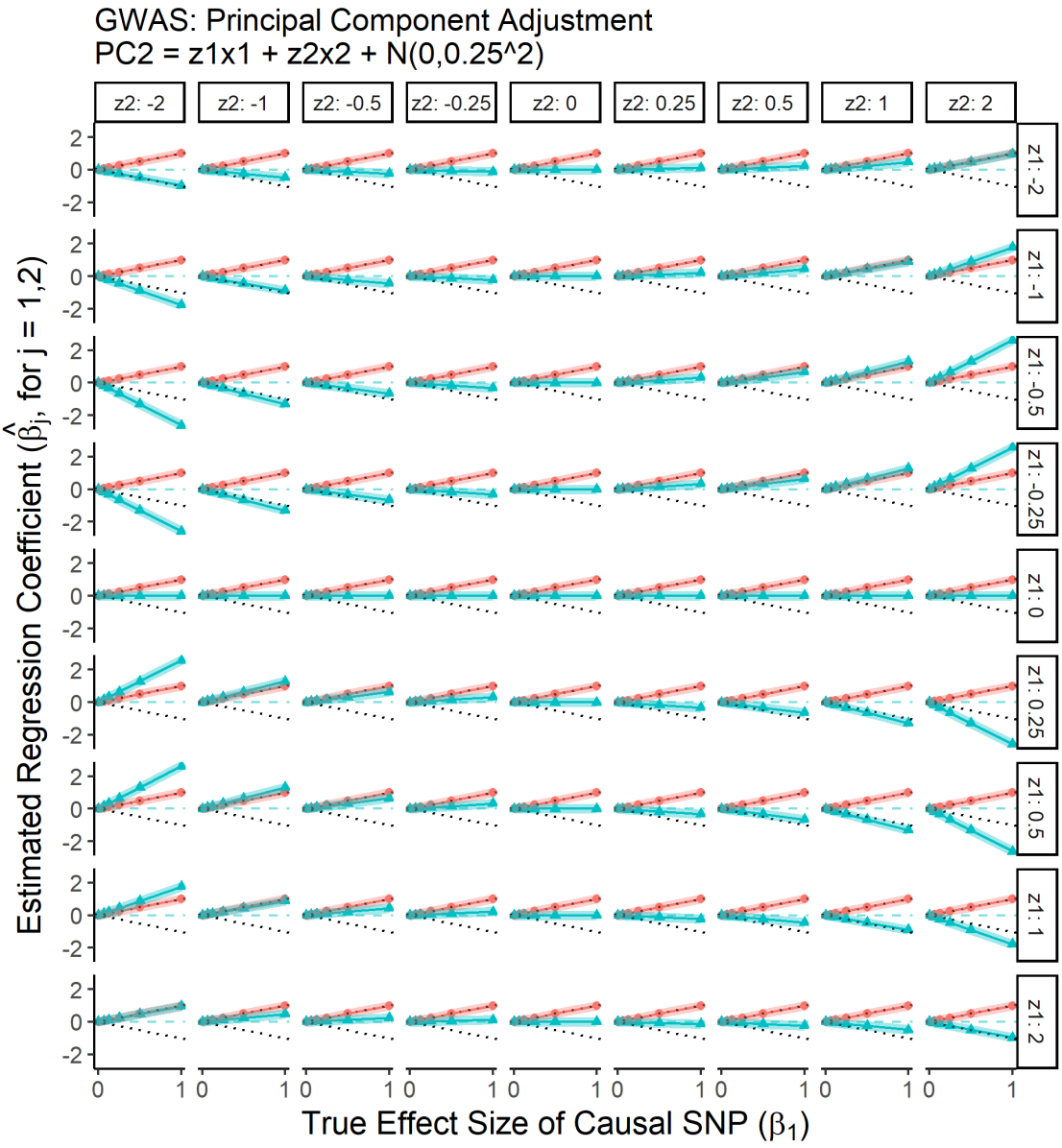

Figure B.4: Observed versus expected and true effect sizes from principal component adjusted GWAS models.

Each panel represents a different simulation setting where the second PC was generated according to the equation $z_1 g_{i1} + z_2 g_{i2} + N(0, 0.25^2)$, changing the scalars z_1, z_2 in each panel. Observed effect sizes = solid lines with points (red with dots = SNP 1, blue with triangles = SNP 2). Expected effect sizes = wider and faintly colored solid lines (red = SNP 1, blue = SNP 2). True effect sizes = dashed lines (red = SNP 1, blue = SNP 2). The $y = x$ line is also provided for reference (dotted black line).

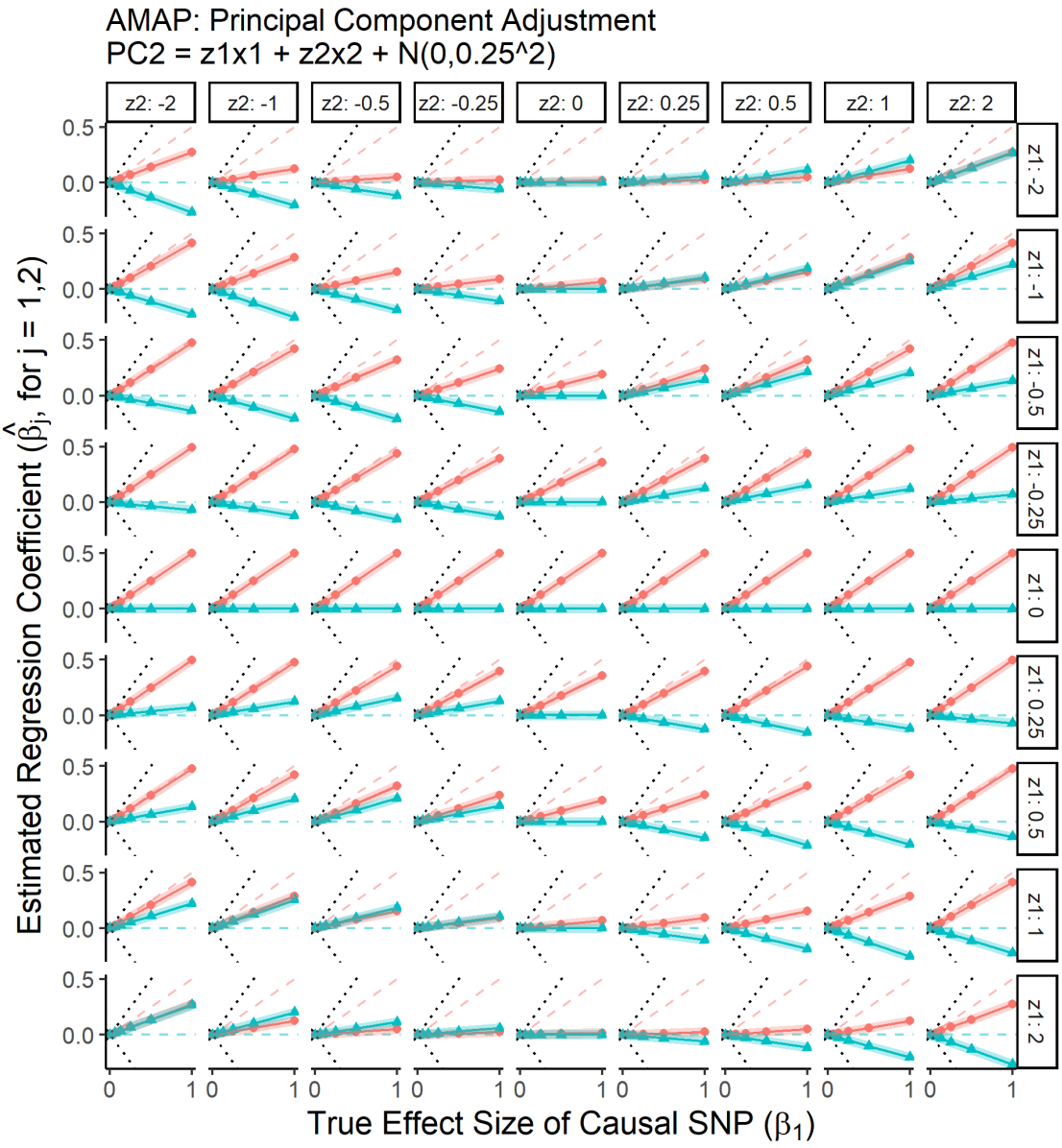

Figure B.5: Observed versus expected and true effect sizes from principal component adjusted admixture mapping models.
Each panel represents a different simulation setting where the second PC was generated according to the equation $z_1 g_{i1} + z_2 g_{i2} + N(0, 0.25^2)$, changing the scalars z_1, z_2 in each panel. Observed effect sizes = solid lines with points (red with dots = SNP 1, blue with triangles = SNP 2). Expected effect sizes = wider and faintly colored solid lines (red = SNP 1, blue = SNP 2). True effect sizes = dashed lines (red = SNP 1, blue = SNP 2). The $y = x$ line is also provided for reference (dotted black line).

B.2 Comparison of pre-PCA Filtering

In Chapter 4, we showed that adjusting for PCs that capture small regions of the genome rather than global ancestry can induce spurious associations in genetic association studies. We showed that this problematic behavior occurred in our analysis of genotype data from WHI SHARe African Americans when PCs were generated using all 551,025 available SNPs, or if we excluded regions identified in the literature as being potentially problematic for PCA. However, problems were ameliorated when we used PCs that were generated after strict LD pruning, using an r^2 threshold of 0.1 and window size of 0.5 Mb. In this Appendix, we investigate the behavior of PCs generated after different filtering techniques.

Many authors have suggested using an r^2 threshold of 0.2 for LD pruning prior to running PCA [142, 68, 164, 136, 135, 131, 66, 129]. Furthermore, this threshold is the default for LD pruning software such as `SNPRelate` [133]. However, in our analysis of WHI SHARe data, we found that using an r^2 threshold of 0.2 prior to running PCA still led to one of the top four PCs being highly correlated with small regions of the genome (Figure B.6), while if we used a stricter r^2 threshold of 0.1, the peaks have disappeared (at least for the top four PCs).

We also compared different choices for the window size to use in LD pruning. In the literature, various window sizes have been suggested, including 10 Mb [89], 2 Mb [68], or 0.5 Mb (the `SNPRelate` default), although others have suggested that window size may not have a big impact [136]. In our analysis of WHI SHARe data, we see little difference in the correlation between PCs and genotypes regardless of the choice of window (Figure B.7). Using a smaller window size is less computationally involved, so we used the window size 0.5 Mb for the remainder of our analyses.

Finally, we also considered filtering out regions that were highly correlated with PCs in our own data, as has been done previously [66, 142]. To implement this data-based filtering, we investigated the SNP *loadings* for each of the top four PCs (the SNP loadings are proportional to the correlation between the PCs and genotypes at each SNP). Starting with the second PC, we found the SNP on each chromosome with the largest loading: if this loading was

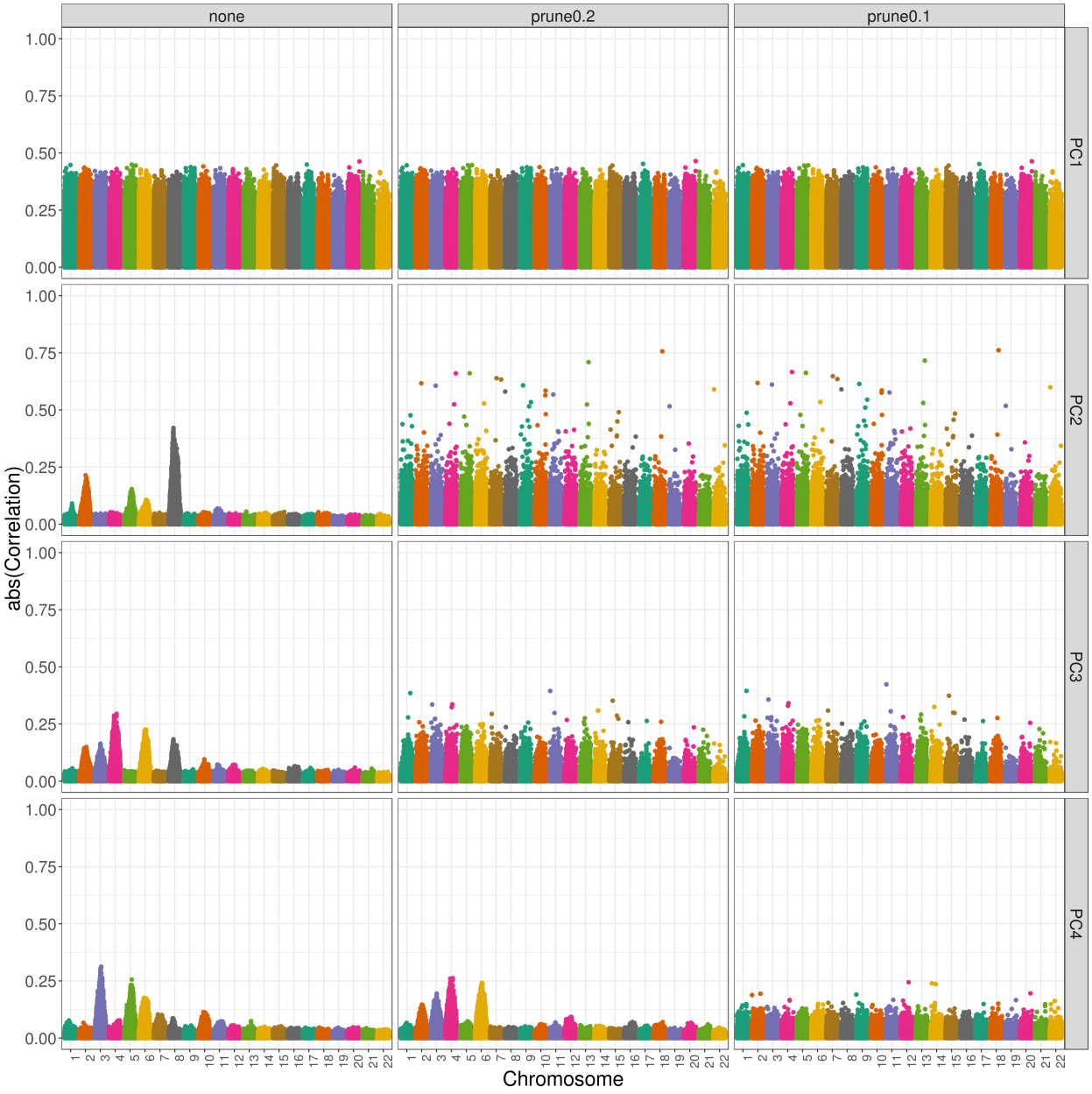

Figure B.6: Correlation between PCs and genotypes in WHI SHARe African Americans using different LD pruning thresholds.

Each panel plots the absolute value of the correlation (y-axis) between principal components and genotypes at each position along the genome (x-axis). Panels are stratified according to which PC is being investigated (1, 2, 3, or 4) and what r^2 threshold was used when running LD pruning prior to running PCA: *none* (no LD pruning), *prune0.2* (LD pruning with $r^2 = 0.2$ and window size = 0.5 Mb), or *prune0.1* (after LD pruning with $r^2 = 0.1$ and window size = 0.5 Mb).

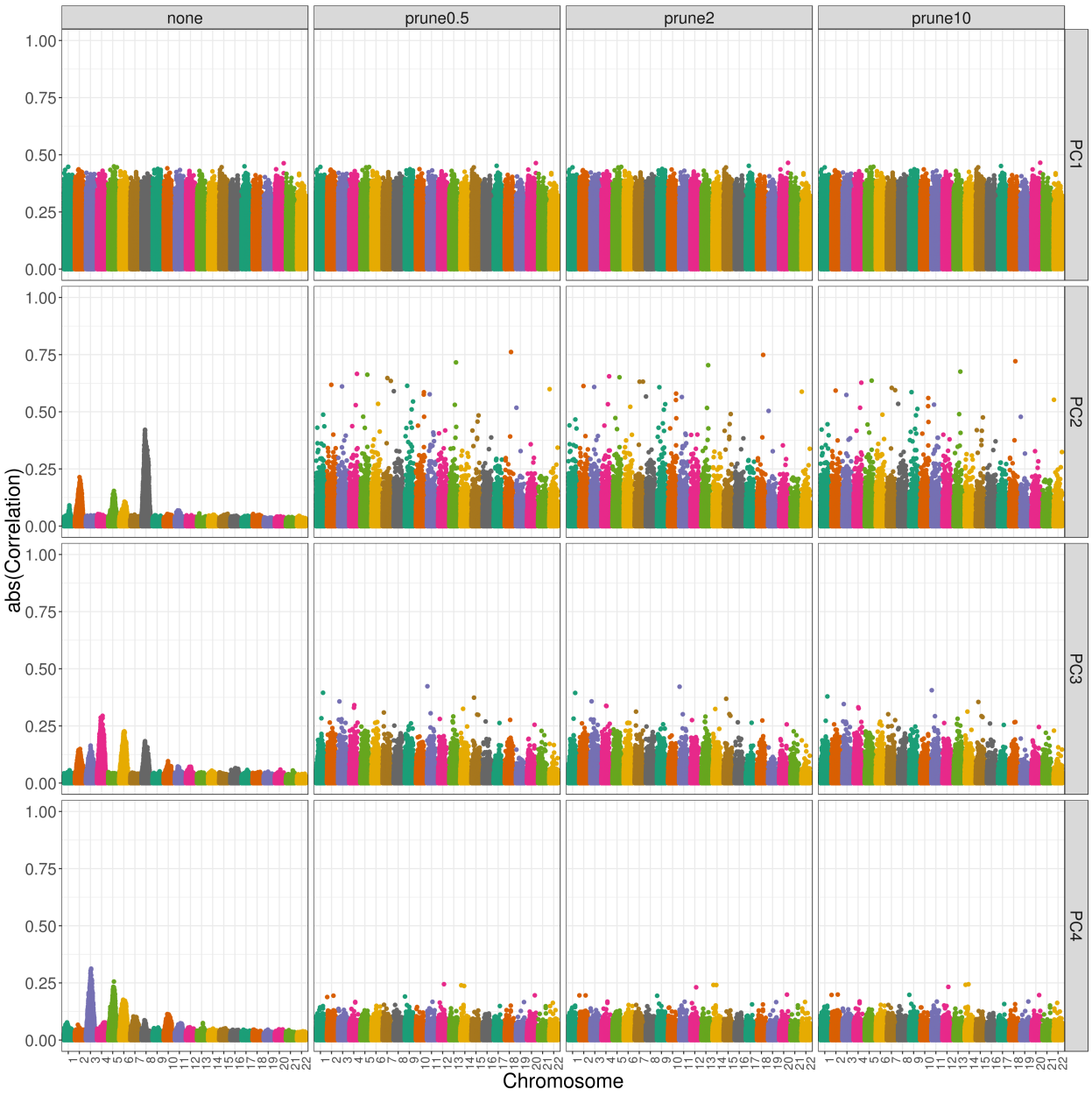

Figure B.7: Correlation between PCs and genotypes in WHI SHARe African Americans using different LD pruning window sizes.
Each panel plots the absolute value of the correlation (y-axis) between principal components and genotypes at each position along the genome (x-axis). Panels are stratified according to which PC is being investigated (1, 2, 3, or 4) and what window size was used when running LD pruning prior to running PCA: *none* (no LD pruning), *prune0.5* (LD pruning with $r^2 = 0.1$ and window size = 0.5 Mb), *prune2* (LD pruning with $r^2 = 0.1$ and window size = 2Mb), or *prune10* (after LD pruning with $r^2 = 0.1$ and window size = 10 Mb).

larger than 0.005, we excluded the SNP and all SNPs within M Mb; if the loading was small, we kept all SNPs on the chromosome. (We considered $M = 1, 5, 10,$ and 20 Mb.) We repeated this process for PCs 3 and 4, and then re-ran PCA using the SNPs that remained. Using these new PCs, we re-calculated SNP loadings and looked to see if there were still regions of the genome that were driving the PCs. If so, we repeated this entire process. This data-based filtering process is very tedious, and even after four rounds of exclusions with $M = 5$ Mb we found that the problematic behavior did not totally go away (Figure B.8). In WHI SHARe data, at least, strict LD pruning is the most effective in eliminating the correlation between PCs and genotypes in small regions of the genome.

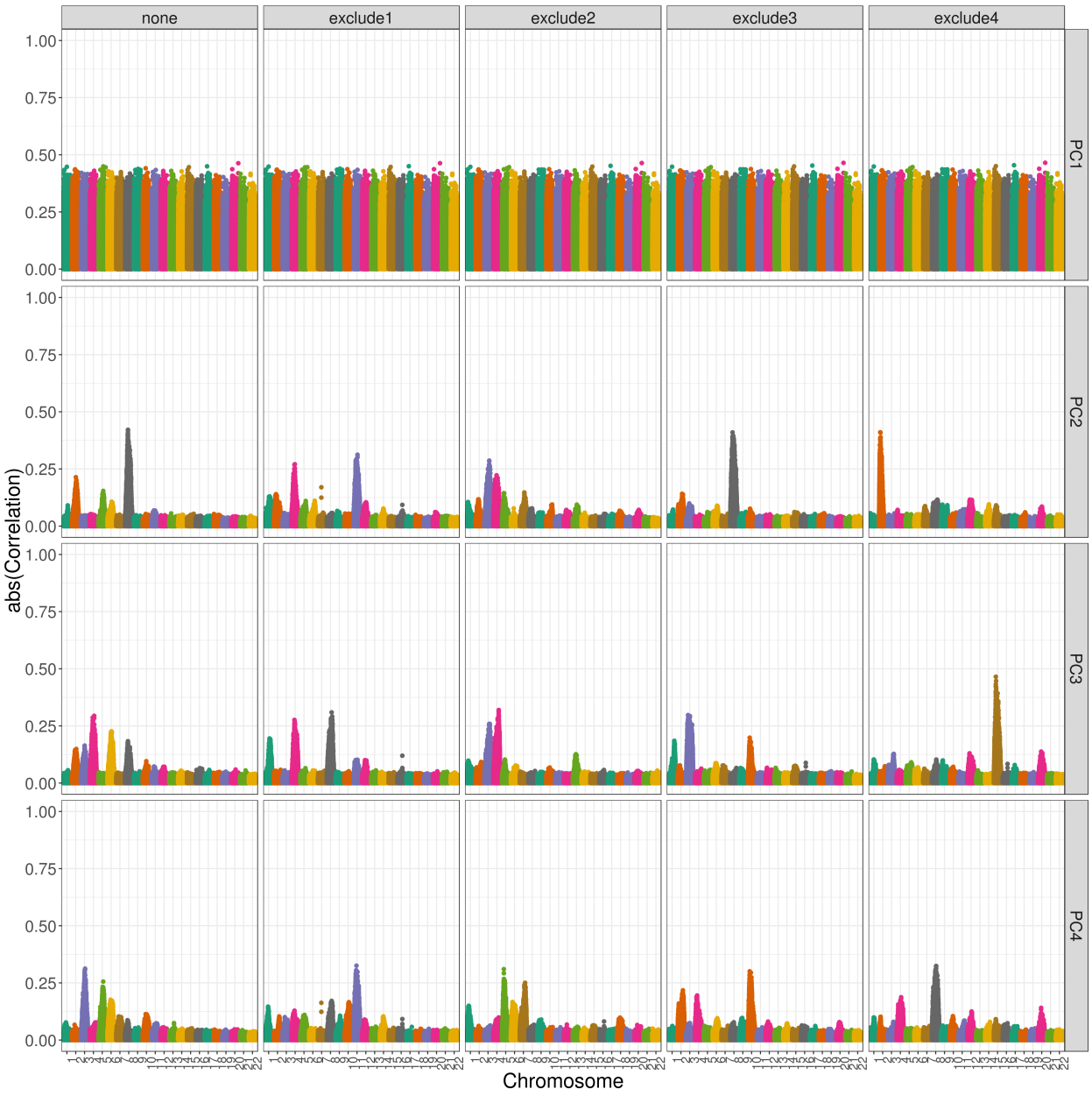

Figure B.8: Correlation between PCs and genotypes in WHI SHARe African Americans after multiple rounds of data-based exclusions.
Each panel plots the absolute value of the correlation (y-axis) between principal components and genotypes at each position along the genome (x-axis). Panels are stratified according to which PC is being investigated (1, 2, 3, or 4) and how many iterations of our data-based procedure for excluding potentially problematic regions were implemented prior to running PCA: *none* (no exclusions), *exclude1* (one round of exclusions, using a window size of 5 Mb), *exclude2* (two rounds of exclusions, using window size = 5Mb), *exclude3* (three rounds of exclusions, using window size = 5Mb), or *exclude4* (four rounds of exclusions, using window size = 5Mb).

Appendix C

APPENDIX FOR CHAPTER 5

C.1 Kidney Phenotype Processing

To account for differences in serum creatinine assays over time, we calibrated the reported serum creatinine levels from each study according to standard techniques. See Table C.1 for details.

Estimated glomerular filtration rate (eGFR) was calculated from serum creatinine (scr) and demographic variables (age, sex, race/ethnicity) using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation [152]:

$$\text{eGFR} = \begin{cases} 144 \times \left(\frac{\text{scr}}{0.7}\right)^{-0.329} \times 0.993^{\text{age}} \times [1 + 0.159 \times \mathbb{1}(\text{black})], & \text{if female and scr} \leq 0.7 \\ 144 \times \left(\frac{\text{scr}}{0.7}\right)^{-1.209} \times 0.993^{\text{age}} \times [1 + 0.159 \times \mathbb{1}(\text{black})], & \text{if female and scr} > 0.7 \\ 141 \times \left(\frac{\text{scr}}{0.9}\right)^{-0.411} \times 0.993^{\text{age}} \times [1 + 0.159 \times \mathbb{1}(\text{black})], & \text{if male and scr} \leq 0.9 \\ 141 \times \left(\frac{\text{scr}}{0.9}\right)^{-1.209} \times 0.993^{\text{age}} \times [1 + 0.159 \times \mathbb{1}(\text{black})], & \text{if male and scr} > 0.9. \end{cases}$$

A binary indicator of chronic kidney disease (CKD) was then generated based on the calculated eGFR:

$$\text{CKD} = \begin{cases} 1 & \text{if eGFR} < 60 \\ 0 & \text{if eGFR} \geq 60. \end{cases}$$

Some individuals appeared in more than one contributing study. We removed duplicates and kept just one entry per subject. See Table C.1 for details.

Table C.1: Processing of TOPMed kidney phenotype data.
Calibration of serum creatinine levels across studies and removal of duplicate samples. No changes were made to the data from the GeneSTAR and JHS studies.

Study	Serum creatinine calibration	Duplicate removal
ARIC	n/a	Remove subjects also in GENOA, HyperGEN, WHI, JHS, or GeneSTAR
GENOA	n/a	Remove subjects also in JHS
HyperGEN	Add 0.0338 mg/dL	n/a
MESA	Multiply by 0.95	Remove subjects also in HyperGEN, ARIC, WHI, and GeneSTAR
WHI	n/a	Remove subjects also in HyperGEN

C.2 Local Ancestry Inference

C.2.1 SGDP Reference Panel

The reference panel used for local ancestry inference was drawn from the Simons Genome Diversity Project (SGDP) [153]. We considered only those individuals from the SGDP Regions *Africa*, *America*, and *West Eurasia*. Populations were excluded if they appeared to contain admixed individuals, based on examining the estimated admixture proportions in Extended Data Figure 3 in Mallick et al. [153]. We also restricted our attention to the European populations in the West Eurasia region, excluding Middle Eastern populations. Table C.2 lists all populations that were included in our reference panel.

C.2.2 RFMix Commands

We used the following `RFMix` commands to perform local ancestry inference in the TOPMed African American, African Barbados, and Hispanic/Latino samples:

```
python RunRFMix.py PopPhased chr22_AA.alleles AA.classes chr22.snploc
```

Table C.2: SGDP populations included in reference panel for TOPMed local ancestry inference.

Ancestral Population	SGDP Populations
AFR	BantuHerero, BantuKenya, BantuTwsana, Biaka, Dinka, Esan, Gambian, Igbo, Ju hoan North, Khomani San, Kongo, Lemande, Luhya, Luo, Mandenka, Mbuti, Mende, Yoruba
EUR	Albanian, Basque, Bergamo, Bulgarian, Czech, English, Estonian, Finnish, French, Hungarian, Icelandic, Norwegian, Orcadian, Polish, Russian, Sardinian, Spanish, Tuscan
NAM	Chane, Karitiana, Mayan, Mixe, Nahua, Piapoco, Pima, Quechua, Surui, Zapotec

```
  --forward-backward -o chr22_AA --disable-parallel -n 5 -G 6 --num-threads 1 -w 0.1

python RunRFMix.py PopPhased chr22_AB.alleles AB.classes chr22.snploc
  --forward-backward -o chr22_AB --disable-parallel -n 5 -G 8 --num-threads 1 -w 0.1

python RunRFMix.py PopPhased chr22_HL.alleles HL.classes chr22.snploc
  --forward-backward -o chr22_HL --disable-parallel -n 5 -G 10 --num-threads 1 -w 0.1
```

We found that running `RFMix` on all samples led to memory issues during the re-phasing step. To get `RFMix` to run without crashing, we had to split our samples into smaller subsets (size $n = 250$ for chromosomes 1 and 4 and $n = 500$ for the remaining chromosomes) and run `RFMix` separately on each subset. This does not affect the accuracy of our local ancestry inference, since `RFMix` performs inference on each haplotype independently when the EM option is not being used.

We used a smaller window size (0.1 cM) than the `RFMix` default (0.2 cM). Using this smaller window size helped with memory issues and allowed us to capture changes in local

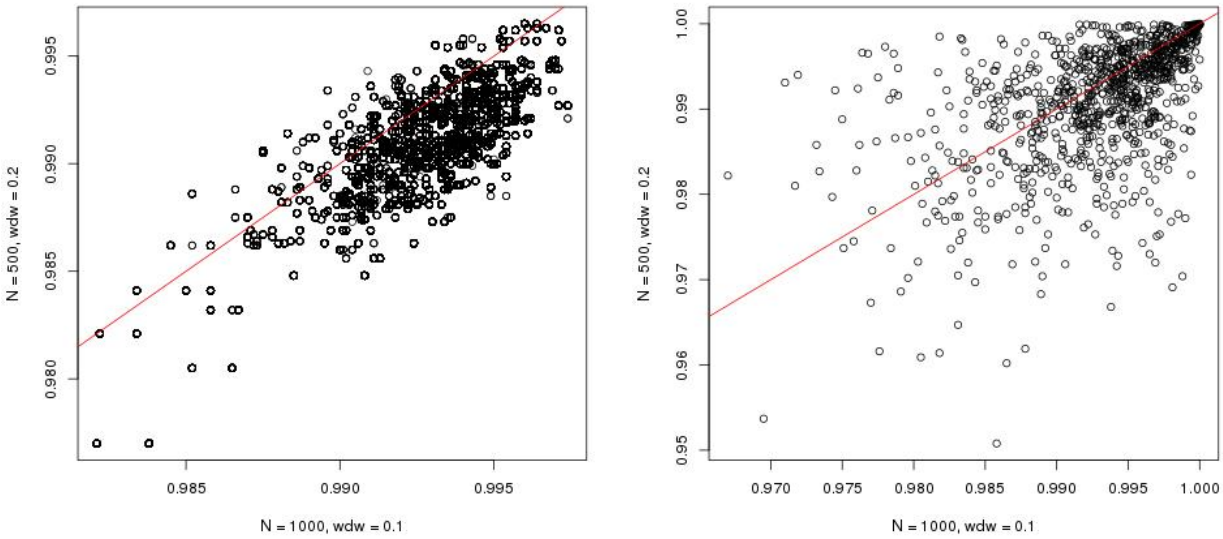

Figure C.1: Comparison of `RFMix` posterior probabilities using different window sizes. Average maximum posterior probabilities for each SNP (left panel) or haplotype (right panel) using `RFMix` to call local ancestry in TOPMed with a window size of 0.2 cM (the default) or 0.1 cM. The red line is the $y = x$ line.

ancestry segments at a finer resolution. In addition, we found that `RFMix` tended to have larger posterior probabilities when we used this smaller window size. Comparing the posterior probabilities generated by `RFMix` on chromosome 22 after using a subset of size 500 individuals and window size of 0.2 cM, versus a subset of size 1000 individuals and window size of 0.1cM, we saw that 87% of SNVs and 62% of haplotypes had a higher average maximum posterior probability using the 0.1 cM window size option (see Figure C.1). Running `RFMix` with a window size of 0.2 cM and subsets of size 1000 individuals crashed due to memory issues on chromosome 22.

C.2.3 Local Ancestry Quality Checks

Comparison to University of Michigan Local Ancestry Calls

A group at the University of Michigan previously performed local ancestry inference on the TOPMed *freeze 5b* samples. Although they used `RFMix` as we did, their reference panel differed considerably from ours. Most notably, they considered seven reference populations—Africa, Europe, America, Central/South Asia, East Asia, Middle East, and Oceania—while we only considered three. In addition, their reference panel was based on genotype data rather than sequence data. The number of samples and SNVs included in each reference panel is summarized in Table C.3.

Table C.3: Description of reference panels used for TOPMed local ancestry inference by two groups.
Source of reference panel data and number of ancestral populations, number of samples, and number of SNVs included in the reference panel.

	Source	No. Populations	No. Samples	No. SNVs
Michigan	HGDP [111]	7	938	639,958
Us (Chap. 5)	SGDP [153]	3	92	49,791,567

Local ancestry inference can only be performed on the overlapping set of SNVs between the reference panel and TOPMed data. As mentioned in Chapter 2, we found in previous analyses that the accuracy of local ancestry inference improved when we consider a larger number of SNVs, even if fewer samples are included in the reference panel (see also: [50]). This motivated our decision to use sequence data from SGDP, even though a smaller number of samples were available for our reference panel (92 African, European, and Native American individuals versus 321 from those same populations in HGDP).

The Michigan group chose to include four additional ancestral populations (Oceania, East Asian, Central/South Asia, and Middle East) in their reference panel. While this decision is useful for various population genetic analyses, for our admixture mapping analyses

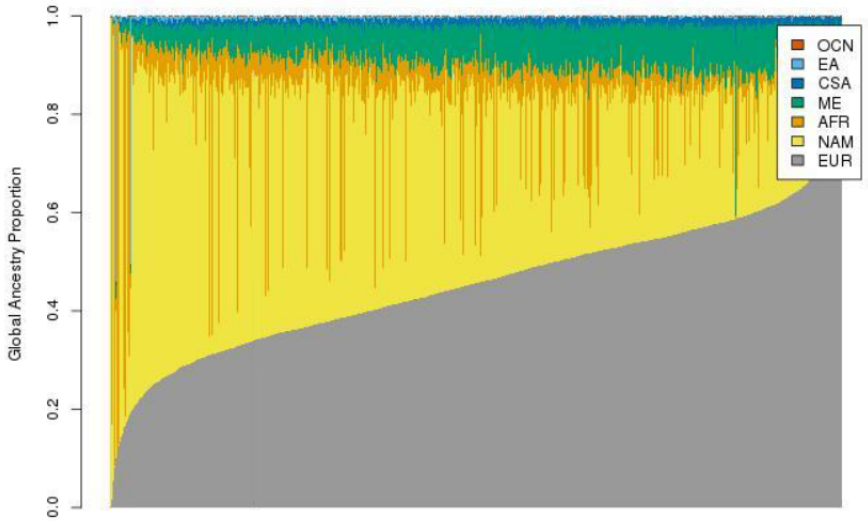

Figure C.2: Barplot of estimated admixture proportions for 4,695 TOPMed Hispanic/Latino samples based on Michigan local ancestry calls.
Seven ancestral populations were considered: Oceania (OCN), East Asian (EA), Central/South Asia (CSA), Middle East (ME), Africa (AFR), America (NAM), and Europe (EUR).

we preferred a more parsimonious analysis looking at just the three ancestral populations (African, European, and Native American) most widely considered to contribute to the ancestry of Hispanics/Latinos and African Americans. In addition, looking at the Michigan local ancestry calls, we found a larger amount of inferred Middle Eastern ancestry than we would expect, particularly among Hispanic/Latino (Figure C.2) and European American (Figure C.3) subjects. This could be explained, at least in part, by previous work that suggests that `RFMix` struggles to distinguish between Middle Eastern and European ancestry [165].

Comparing the admixture proportions generated from the Michigan local ancestry calls to our own, we see a nice correspondence between the two (Figure C.4). In particular, the African admixture proportions are nearly perfectly correlated. There is slightly lower correlation across the sets of admixture proportions for the European and Native American components, but we can improve this correlation by combining the European, Middle Eastern

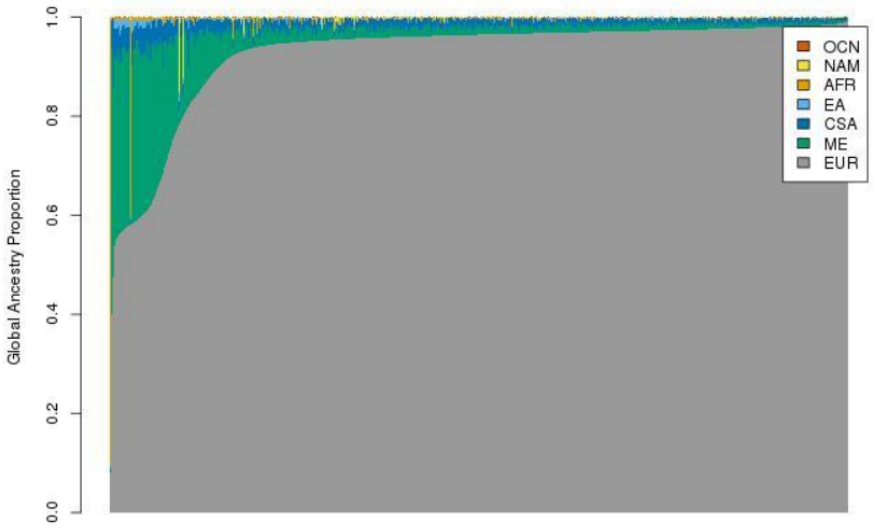

Figure C.3: Barplot of estimated admixture proportions for 28,602 TOPMed European American samples based on Michigan local ancestry calls.
Seven ancestral populations were considered: Oceania (OCN), America (NAM), Africa (AFR), East Asian (EA), Central/South Asia (CSA), Middle East (ME), and Europe (EUR).

and Asian local ancestry calls from Michigan into a single category.

Distribution of Local Ancestry Assignments

At each locus, we calculated the proportion of local ancestry calls assigned to the African, European, and Native American ancestral populations in the TOPMed admixed samples. We plotted the distribution of these proportions across the genome in Figure C.5. The proportion of local ancestry calls assigned to each ancestral population stays relatively constant across the genome, although there are a few regions (e.g., parts of chromosomes 6, 8, and 21) that deviate slightly from the genome-wide average. Deviations could be due to artifacts in the local ancestry caller, or due to processes such as selection [166, 167]. A list of these regions is available to TOPMed investigators using our local ancestry calls upon request.

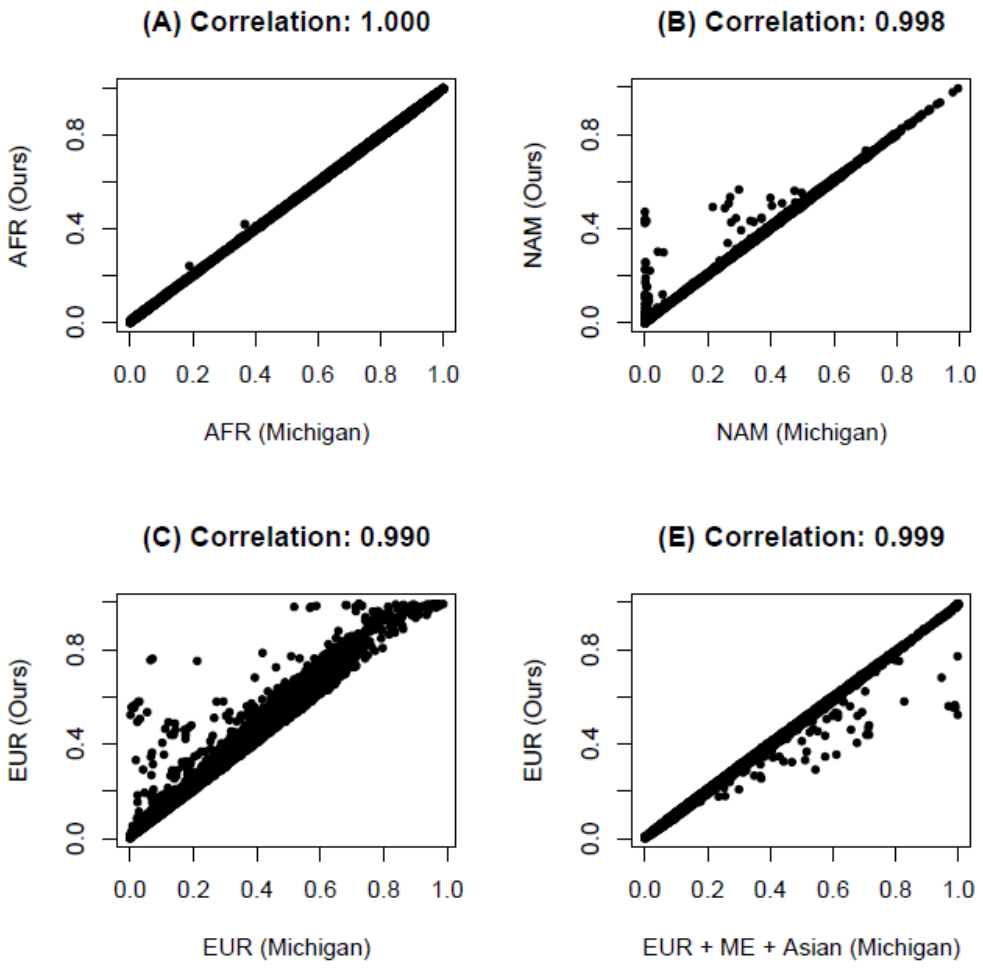

Figure C.4: Comparison of estimated admixture proportions, based on our local ancestry calls or the calls generated by a group at the University of Michigan, for 20,050 TOPMed admixed individuals.
Panel (A) compares the two sets of African admixture proportions, (B) compares the Native American proportions, (C) compares the European proportions, and (D) compares our European proportion to the sum of Michigan's European, Middle Eastern, and Asian admixture proportions.

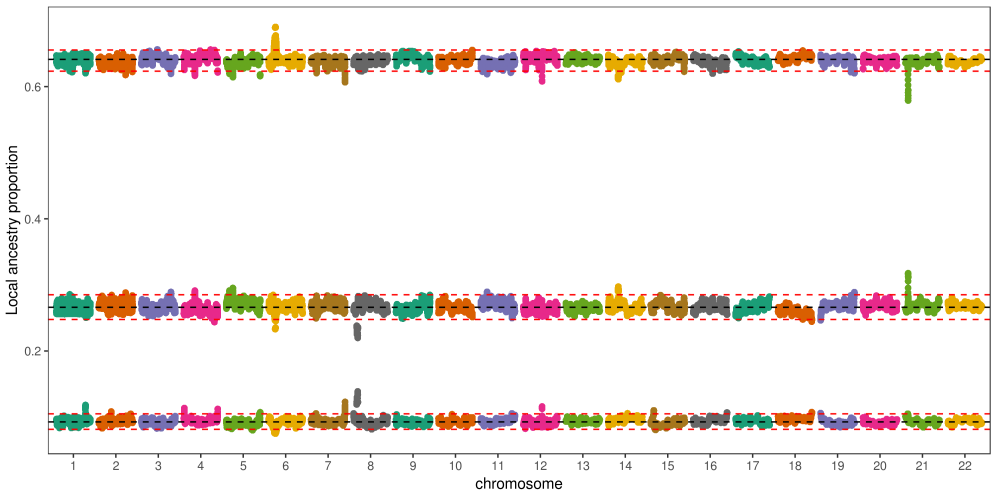

Figure C.5: Proportion of TOPMed local ancestry calls assigned to each ancestral population at each locus.

Proportion of calls assigned to Native American (bottom), European (middle) or African (top) ancestry across all 20,048 African American, African Caribbean, and Hispanic/Latino TOPMed samples. Points falling outside the red dashed lines (NAM: 0.081–0.105, EUR: 0.247–0.285, AFR: 0.624–0.658) represent the 1% of loci with proportions furthest from the genome-wide averages (black dashed lines; NAM: 0.093, EUR: 0.266, AFR: 0.641).

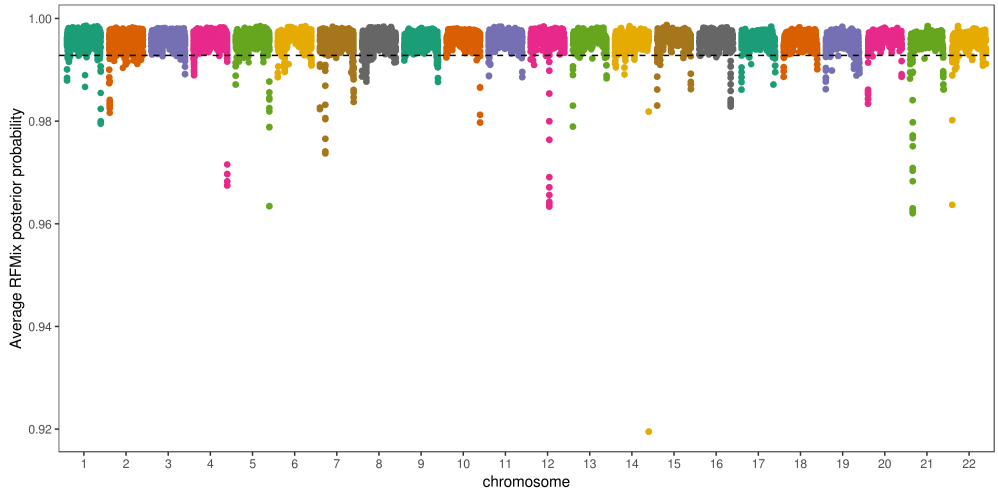

Figure C.6: Average maximum `RFMix` posterior probabilities at each locus in TOPMed. The average, across all admixed TOPMed haplotypes, of the maximum posterior probability at each locus. Points falling below the black dashed line ($y = 0.993$) represent the 1% of loci with the smallest posterior probabilities.

RFMix Posterior Probabilities

As discussed in Chapter 2, at each locus `RFMix` produces posterior probabilities of each ancestry given the observed haplotype, and the ancestry with the largest posterior probability is the one that is ultimately called at that position. In our analysis of the TOPMed data, three posterior probabilities (for African, European, and Native American ancestry) were generated for each haplotype at each position. We recorded the largest of these three probabilities, corresponding to the posterior probability for the called ancestry, and compared the average (across haplotypes) of these maximum posterior probabilities at each position along the genome (Figure C.6). In general, these posterior probabilities were quite high, but there are some regions—often at the ends of chromosomes—where `RFMix` was somewhat less confident in the calls it made. Again, a list of these regions is available upon request.

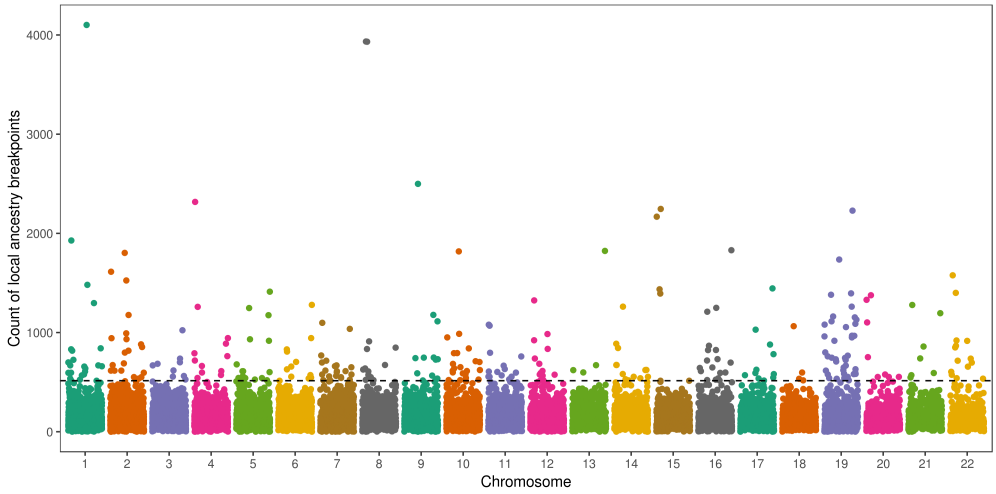

Figure C.7: Breakpoints in inferred local ancestry segments in TOPMed.
Number of local ancestry segments (out of ≈ 40k haplotypes) that terminate in each 0.1 cM window across the genome. Points falling above the black dashed line ($y = 514$) represent the 1% of windows containing the largest number of breakpoints.

Distribution of Local Ancestry Breakpoints

As a final quality check, we investigated the distribution of the breaks in inferred local ancestry segments across the genome (Figure C.7). Variability in the number of breaks occurring in each region of the genome could be due to biases in the local ancestry caller, or due to underlying biological causes such as variation in recombination rates across the genome. A list of regions with unusually high numbers of local ancestry breakpoints is available upon request.

C.3 Estimating the Number of Generations Since Admixture

According to our work in Chapter 3 (see also: [59]), we can use the observed pattern of local ancestry correlation in our sample to estimate the number of generations since admixture. This estimate is used by our program `STEAM`, along with theoretical results regarding the expected pattern of local ancestry correlation in an admixed population with population structure as observed in our sample, to determine the appropriate genome-wide significance

threshold for our admixture mapping study. Although this approach was developed in the context of unrelated samples, it still seems to perform well here, at least in the sense that the expected local ancestry correlation curves derived under the assumption of unrelated samples are consistent with the observed patterns of local ancestry correlation in the TOPMed data (Figures C.8, C.9, C.10).

Our estimate of the number of generations since admixture varied depending on whether we used all admixed samples ($\hat{g} = 6.67$), African Americans only ($\hat{g} = 6.69$), or Hispanics/Latinos only ($\hat{g} = 9.39$). In all three cases, we observed a nice correspondence between the observed and expected patterns of local ancestry correlation (Figures C.8, C.9, C.10), perhaps with the exception of the correlation of the Native American ancestry component at pairs of loci across all (bottom right panel of Figure C.8) or African American (bottom right panel of Figure C.9) samples. Furthermore, the estimates $\hat{g}$ are consistent with previous studies [59, 48, 115, 116, 9, 23] (see also: Chapter 3).

C.4 Quantile-Quantile Plots for Association Studies

Quantile-quantile (QQ) plots are often used as diagnostic tools in genetic association studies. To create a QQ plot, we sort the *observed* p-values from our association study versus the *expected* p-values under the null hypothesis (assuming independence across loci; i.e., p-values $\sim_{iid} \text{Uniform}(0, 1)$). These plots, along with the corresponding inflation factors λ, are often used to diagnose inadequate control of population structure or other artifacts [168, 169]. It is worth noting that QQ plots may not be the most appropriate diagnostic tool for admixture mapping studies, given the increased amount of correlation among admixture mapping test statistics relative to the genome-wide association studies in which these plots are typically used. Nonetheless, we provide the QQ plots corresponding to our primary (Figure C.11) and secondary (Figures C.12–C.15) admixture mapping analyses and association mapping analysis (Figure C.16) here.

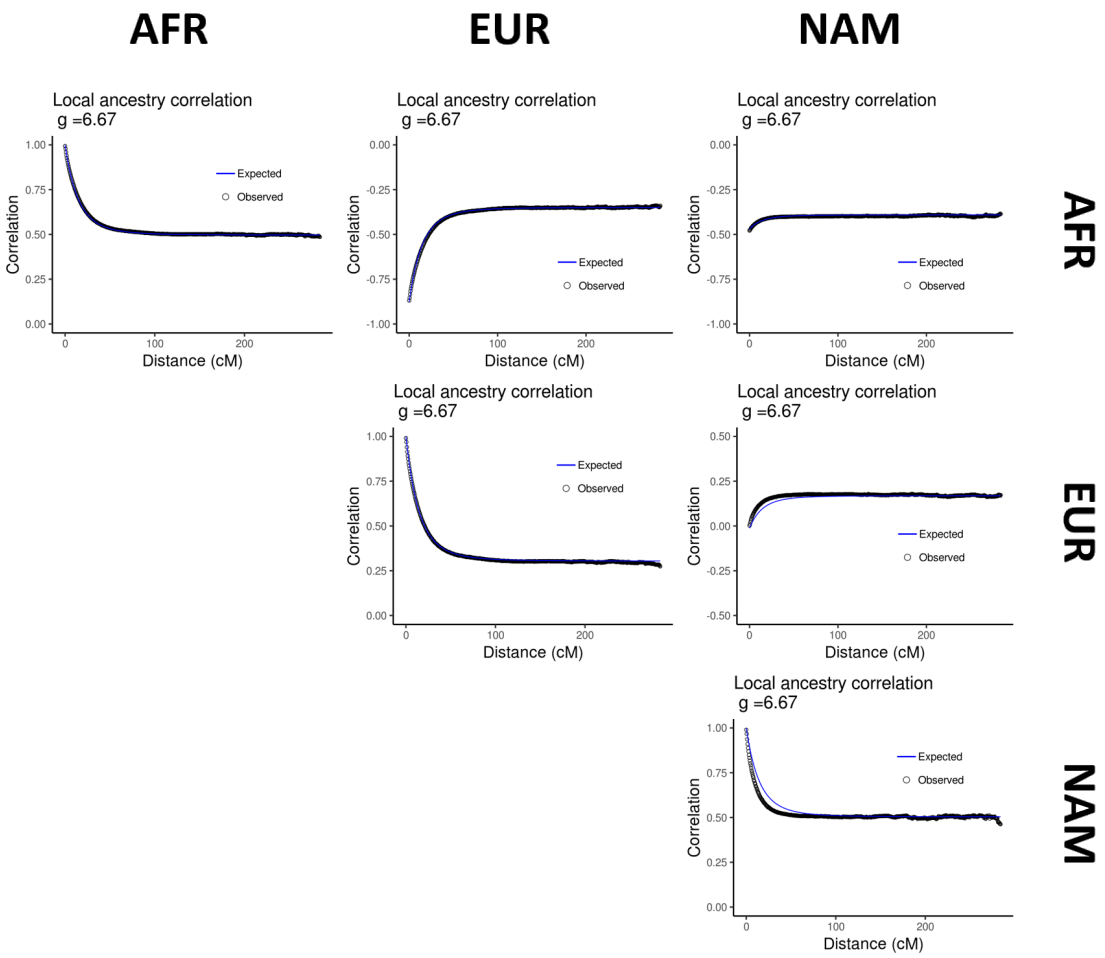

Figure C.8: Observed and expected local ancestry correlation curves for all TOPMed samples.
We plot the correlation in local ancestry components at pairs of SNVs versus the distance between those loci. Expected values are based on the estimate $\hat{g} = 6.67$ from running our non-linear least squares regression approach (proposed in Chapter 3; implemented in `STEAM`) on all admixed samples included in our admixture mapping analysis.

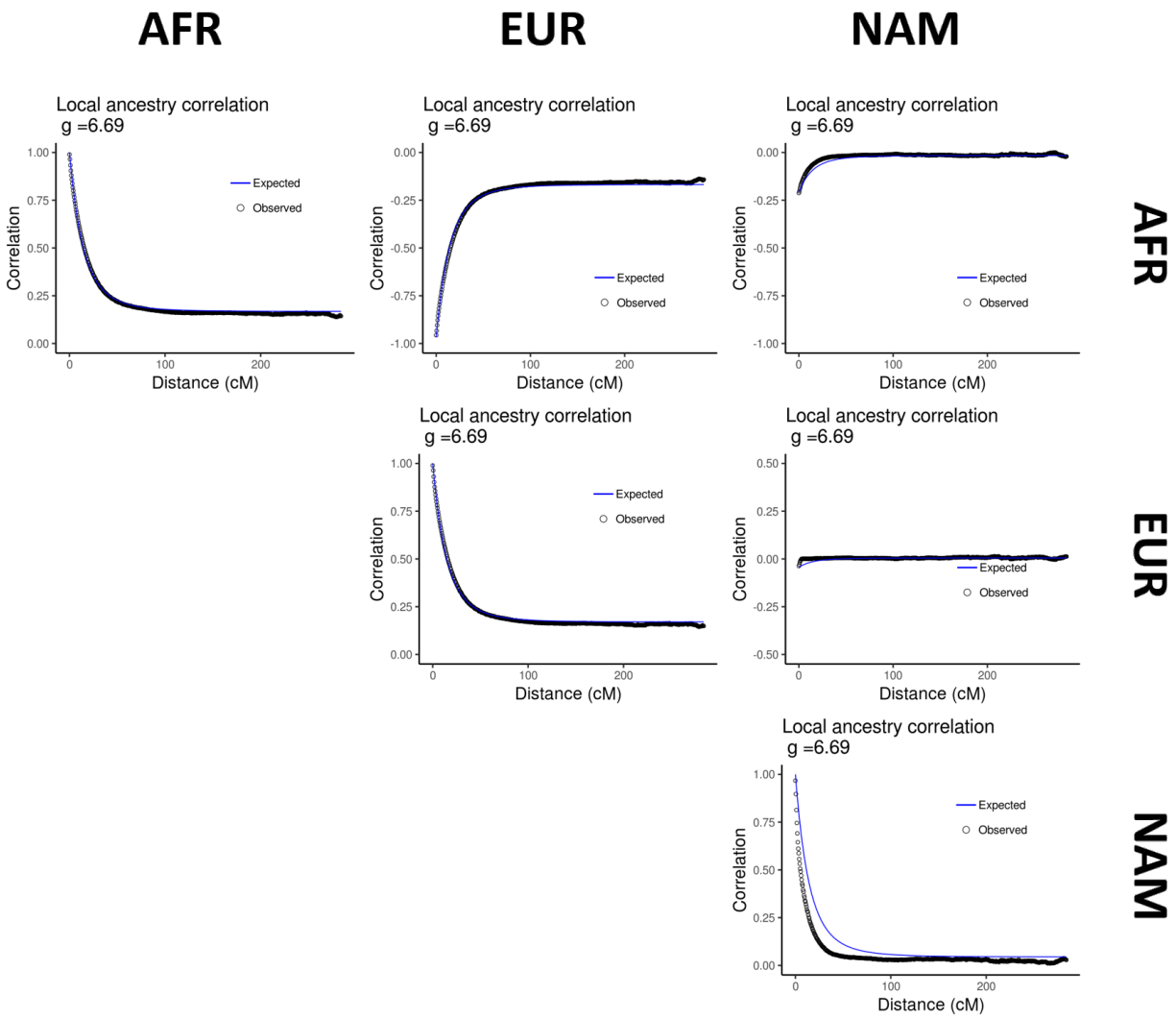

Figure C.9: Observed and expected local ancestry correlation curves for African American TOPMed samples.

We plot the correlation in local ancestry components at pairs of SNVs versus the distance between those loci. Expected values are based on the estimate $\hat{g} = 6.69$ from running our non-linear least squares regression approach (proposed in Chapter 3; implemented in `STEAM`) on African American samples included in our admixture mapping analysis.

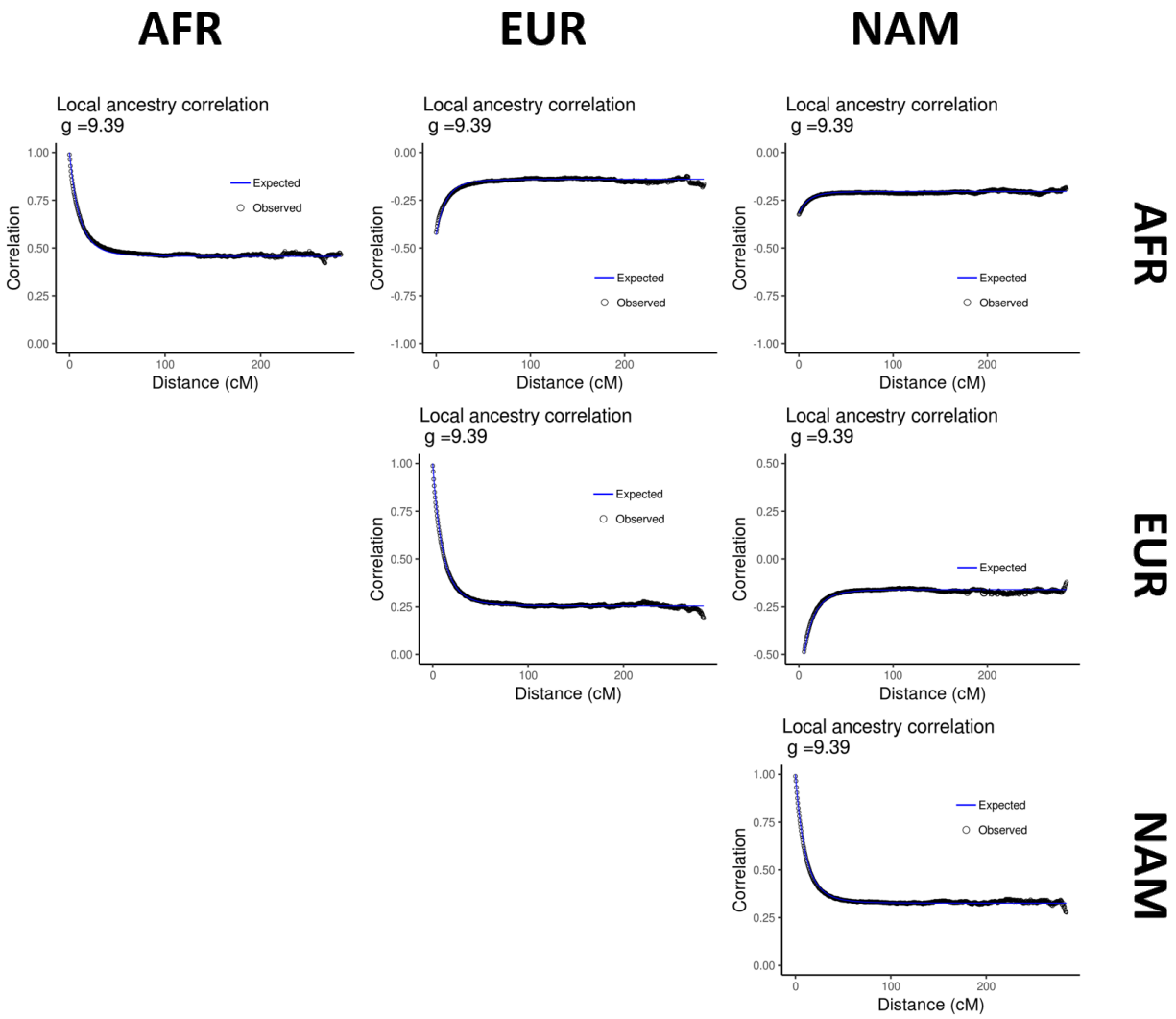

Figure C.10: Observed and expected local ancestry correlation curves for Hispanic/Latino TOPMed samples.

We plot the correlation in local ancestry components at pairs of SNVs versus the distance between those loci. Expected values are based on the estimate $\hat{g} = 9.39$ from running our non-linear least squares regression approach (proposed in Chapter 3; implemented in `STEAM`) on Hispanic/Latino samples included in our admixture mapping analysis.

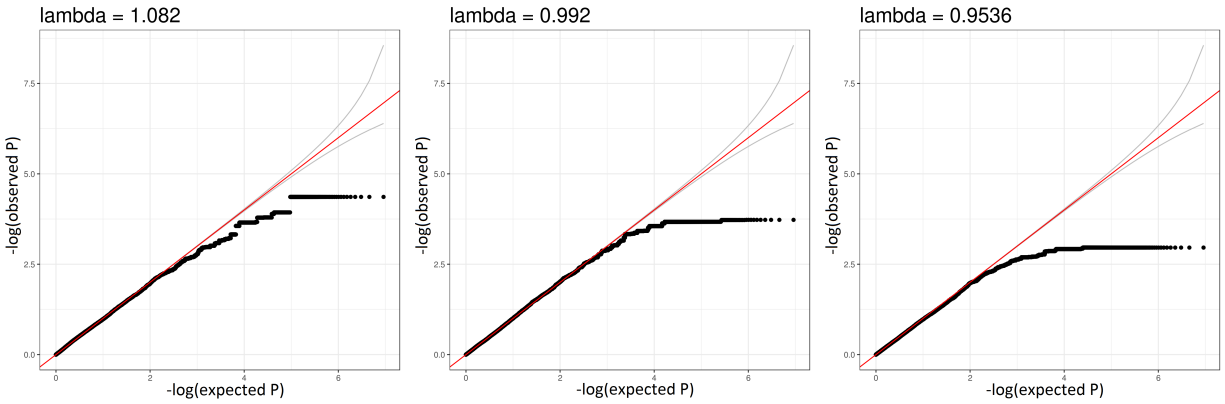

Figure C.11: QQ plots for eGFR admixture mapping study in 9,479 admixed samples.
Expected versus observed p-values from testing the association between eGFR and African (left panel), European (middle panel), and Native American (right panel) local ancestry.

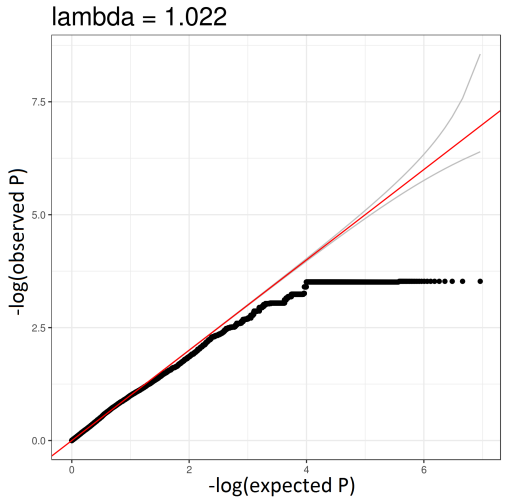

Figure C.12: QQ plot for eGFR admixture mapping study in 8,303 African American samples.
Expected versus observed p-values from testing the association between eGFR and African local ancestry.

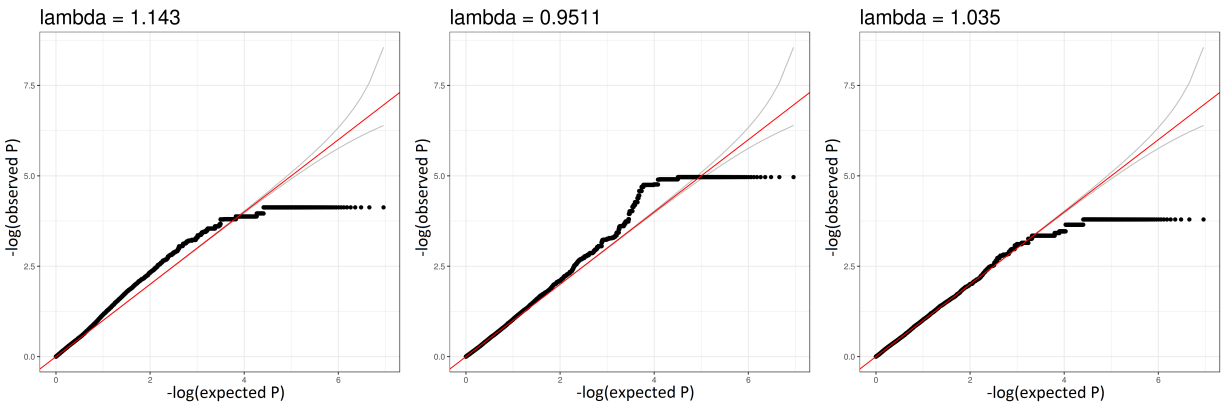

Figure C.13: QQ plot for eGFR admixture mapping study in 1,176 Hispanic/Latino samples. Expected versus observed p-values from testing the association between eGFR and African (left panel), European (middle panel), and Native American (right panel) local ancestry.

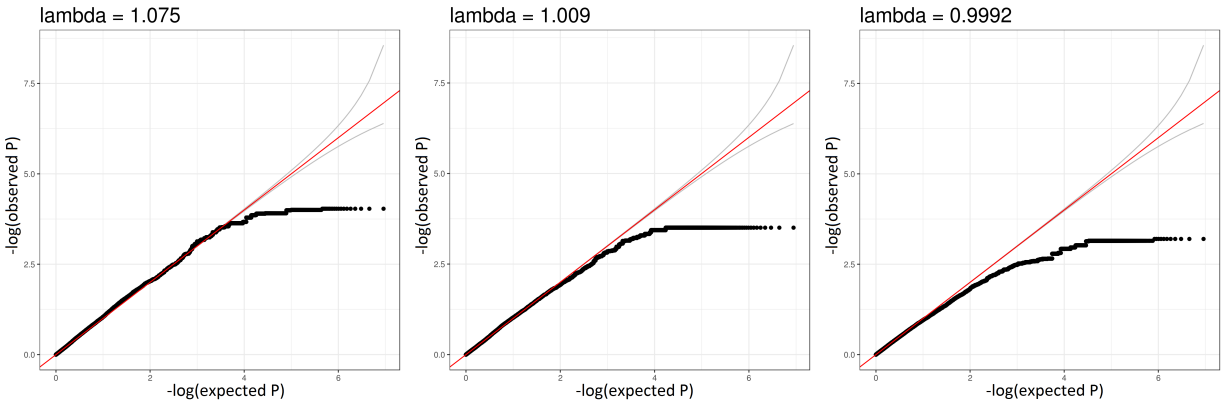

Figure C.14: QQ plot for serum creatinine admixture mapping study in 9,479 admixed samples.
Expected versus observed p-values from testing the association between serum creatinine and African (left panel), European (middle panel), and Native American (right panel) local ancestry.

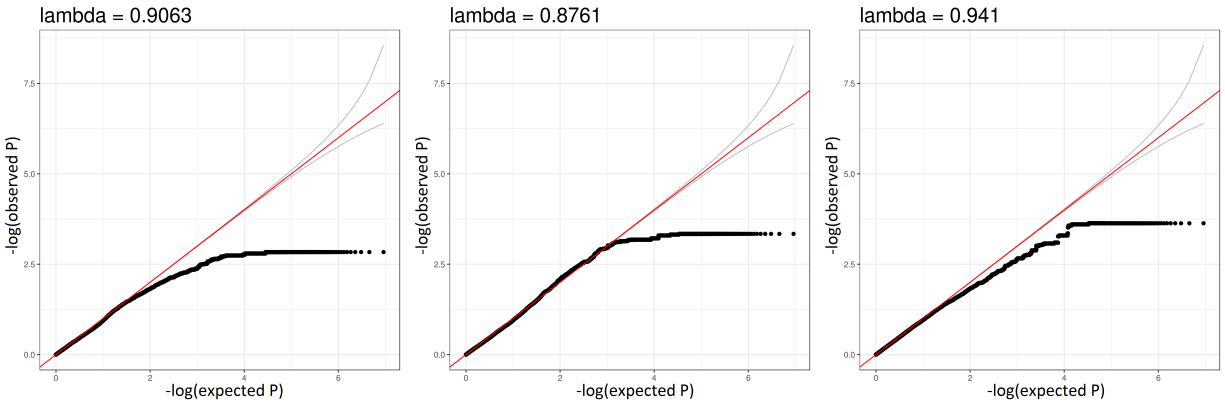

Figure C.15: QQ plot for CKD admixture mapping study in 9,479 admixed samples. Expected versus observed p-values from testing the association between chronic kidney disease and African (left panel), European (middle panel), and Native American (right panel) local ancestry.

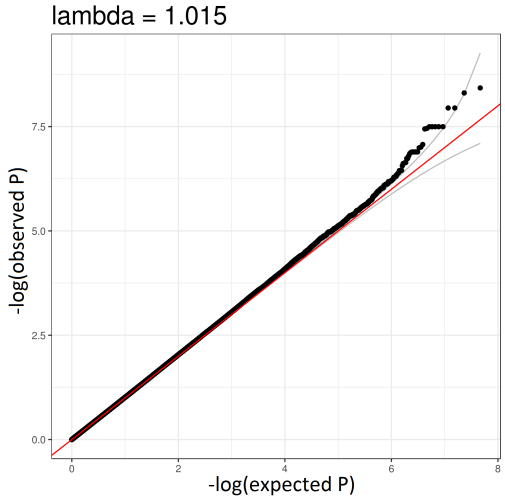

Figure C.16: QQ plot for eGFR association mapping study in 9,479 admixed samples. Expected versus observed p-values from testing the association between eGFR and genotype.