5-1-2016

Estimating Ancestry and Genetic Diversity in Admixed Populations.

Anthony Koehl

Anthony Joseph Koehl

Candidate

Anthropology

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Jeffrey Long PhD, Chair

Keith Hunley PhD, Member

Osbjorn Pearson PhD, Member

Lindsay Smith PhD, Member

Mark Shriver PhD, Member

Estimating Ancestry and Genetic Diversity in Admixed Populations

BY

Anthony Joseph Koehl

B.S. Anthropology, Northern Kentucky University, 2003
M.S. Human Biology, University of Indianapolis, 2009
M.A. Anthropology, University of New Mexico, 2013

Dissertation

Submitted in Partial Fulfillment of the
Requirements of the Degree of
Doctor of Philosophy

Anthropology

The University of New Mexico
Albuquerque, New Mexico

January, 2016

ACKNOWLEDGMENTS

I wish to wholeheartedly acknowledge Dr. Jeffrey Long, my doctoral advisor and dissertation chair, who worked tirelessly to advance me through this stage of my career. Dr. Long's commitment in the classroom provided me with the knowledge to undertake this research. His door was always open, which allowed me to advance my research and my understanding of genetics and for that I am eternally grateful. He is the greatest teacher I have had and the best student I have ever seen. Dr. Long has been committed to me through this process as my mentor, and I hope that as I advance my career he will maintain that commitment as a colleague and as a friend.

In addition, I wish to thank my committee members, Dr. Keith Hunley, Dr. Ozzie Pearson, Dr. Mark Shriver, and Dr. Lindsay Smith, for their insight in improving this work and for advancing me as a professional in the field of anthropology. I hope to have the opportunity to collaborate with all of them in the future.

Finally, to my friends and family, without you I could not have succeeded in this endeavor. I am forever thankful to you for your support, and camaraderie.

Estimating Ancestry and Genetic Diversity in Admixed Populations

by

Anthony Joseph Koehl

B.S. Anthropology, Northern Kentucky University, 2003
M.S. Human Biology, University of Indianapolis, 2009
M.A. Anthropology, University of New Mexico, 2013
Ph.D. Anthropology, University of New Mexico

ABSTRACT

Admixture is a form of gene flow that occurs when long separated populations come into contact and exchange mates. Admixture has been a primary mechanism in the formation of many modern human populations. The genetic characteristics of an admixed population are intermediate to, yet distinct from, those of its ancestors. In this dissertation, I investigate biological and statistical factors that enter into the analysis of admixed populations using genetic marker data. In chapters one and two, I use genotype data from published sources that contain 618 microsatellite loci. In chapter three, I simulate genotypes of 500 microsatellite loci.

In chapter two, I present an analysis of genetic diversity within and among 17 populations in the Americas that were formed by admixture among continental Indigenous Americans, Africans and Europeans. This is the first application of a new method to partition the genetic distance between pairs of populations into components related to ancestry and genetic drift. I show that the genetic relationships among the continental sources and genetic drift occurring after population formation strongly influence the genetic structure of these populations.

In chapter three, I investigate a new strategy to find modern populations to serve as models for ancestors in admixture events that occurred in the past. This is a long-standing

challenge to admixture studies. This chapter focuses on the Cape Coloured people of South Africa, a population that formed by mixture of indigenous Africans, Europeans, and Asians. I propose a series of models for their ancestry and use the Akaike Information Criterion to choose the best model. This method from information theory identifies a simple model that proposes only African and Asian ancestors. I interpret this result in terms of both the principle of parsimony and the evolutionary recent common ancestor of the human species.

In chapter four, I use computer simulations to assess bias in ancestry fractions estimated by using maximum likelihood. These novel simulations were designed to produce data sets that mimic actual patterns of variation in human populations. I have found sampling strategies that produce reasonably unbiased results, despite the potential for maximum likelihood to produce biased estimates.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this dissertation, I use a common statistical approach for my admixture analyses. Tang et al. (2005) developed the statistical method using maximum likelihood to estimate ancestry in admixed populations from genotype data obtained from contemporary populations.

$$lnL(\theta) = \sum_{s=1}^{S} \sum_{i=1}^{N_S} \sum_{l=1}^{L} \sum_{j=1}^{J_l} [g_{silj} \times ln(y_{silj})] \tag{1.1}$$

where

$$y_{silj} = \sum_{k=1}^{K} p_{jlk} m_{ik}$$

is the predicted allele for the j^{th} allele at the l^{th} locus in the i^{th} individual in the s^{th} sample.

The genotype data g_{silj} are the counts of the j^{th} allele ($j = 1...J_l$), observed at the l^{th} locus ($l = 1...L$) from the i^{th} person ($i...N_s$), belonging to the s^{th} sample ($s = 1...S$). The parameters ($\theta = [\mathbf{p}, \mathbf{m}]$) are p_{jlk} the frequency of the j^{th} allele, from the l^{th} locus, from the k^{th} source population ($k = 1...K$), and m_{ik} the fraction of ancestry from the k^{th} source population contriubted to the i^{th} individual.

Researchers previously collected the genotype data from many contemporary populations found throughout the world (Cann et al., 2002; Rosenberg et al., 2002, 2006; Wang et al., 2007, 2008; Tishkoff et al., 2009). These genotypes consist of microsatellite loci. Pemberton et al. (2013) worked to centralize the data collected from other researchers into a single data set by calibrating the loci of more than 2,500 individuals from 248 worldwide populations. My research uses a subset of these samples and loci for all individuals included in my analyses.

There are several assumptions associated with maximum likelihood, which relate to biological and statistical factors in admixture analyses. First, each allele in a genotype of an individual represents an independent draw from one of the source populations. Second, contemporary individuals derive from populations that are in Hardy-Weinberg equilibrium, when conditioned on ancestry fractions. Third, marker loci are in linkage equilibrium, when conditioned on ancestry fractions. Fourth, ancestry is estimated from the true ancestral source populations. Finally, gene flow in the form of admixture is the only evolutionary process operating in this system.

I use maximum likelihood to address and overcome challenges in admixture analyses. The challenges involve proper identification of ancestral source populations that contributed to the admixture event. There are three primary challenges confronting the proper identification of ancestral source populations. (1) Admixture events that formed many contemporary populations began or occurred entirely in the past. (2) Ancestral source populations may no longer exist, or they have evolved since the time of the admixture event. (3) A sparse historical record prevents us from fully knowing the source populations. These challenges are ubiquitous in my research as well as in all other admixture analyses. I present ways to overcome these challenges, and address particular model assumptions in each of my dissertation chapters.

In chapter two, I analyze the genetic diversity within and among 13 Latin American and four African-American populations in the Americas. The admixture of continental Indige-

nous Americans, Africans, and Europeans formed the mixed populations. My analysis uses genotype data at 618 microsatellite loci from 949 individuals from 49 genetically sampled populations to estimate the ancestry and ancestral allele frequencies that contributed to the formation of these admixed groups. This analysis partitions Nei's minimum genetic distance into components of ancestry and genetic drift among the admixed populations (Nei, 1973, 1987). I partition genetic distance through a series of matrix calculations. First, I calculate Nei's minimum genetic distance. Then I calculate the expected minimum genetic distances from the expected allele frequencies of the sampled admixed populations and contemporary populatoin samples that serve as pseudo-ancestors, which are obtained using the likelihood method of Tang and colleagues (2005). The expected minimum distance values are the partitioned ancestry distances of the admixed populations. I obtain the genetic drift partition from matrix subtraction, which is simply the difference of the ancestry partition from Nei's minimum genetic distance. Recall an assumption of the likelihood model; admixture is the only evolutionary process operating in the model. I show, from this research, that genetic drift plays a prominent role in shaping the genetic diversity of the admixed populations, which provides a fuller perspective of the effect of the evolutionary process in these populations.

In chapter three, I investigate how to choose contemporary populations to serve as ancestral sources in admixture analyses of ancestry. This work addresses the challenge of using what Tang and colleagues (2005) call pseudo-ancestors. Pseudo-ancestors are populations that are closely related to the ancestral sources that contributed to the formation of an admixed population, but who did not aid directly in the formation of the admixed population (Tang et al., 2005). The use of pseudo-ancestors is necessary because of the challenges in admixture analyses. First, source populations may no longer exist or have evolved since the time of the admixture event. Second, a sparse historical record prevents us from fully knowing who the true ancestors of an admixed population were. I construct 26 models of proposed ancestry for the Cape Coloured population of South Africa who

serve as a focal admixed population to test this method. The method I use is the Akaike Information Criterion (AIC) to choose the best model from the 26 that estimate ancestry for the Cape Coloured population (Akaike, 1973, 1974). These models contain between two and five ancestral source populations, which include the Khoesan, Bantu speakers, European, South Asian, and East Asian. Each ancestral source population is comprised of genotype data containing 618 microsatellite loci of individuals from two contemporary population samples.

In chapter four, I investigate the concept of pseudo-ancestors further. I use coalescent simulations to examine ancestry proportion estimates of an admixed population (Excoffier and Foll, 2011). In knowing the relationships of the pseudo-ancestors to the true ancestors, I will determine if genotype data from pseudo ancestral sources in lieu of the true ancestors biases estimates of ancestry. This research addresses several challenges inherent in admixture analyses. Primarily, these challenges are admixture events that formed many contemporary populations began or occurred entirely in the past; and ancestral source populations may no longer exist, or they have evolved since the time of the admixture event. I begin by constructing a simulated consensus tree of pseudo-ancestors to serve as ancestral sources used to estimate ancestry proportions in an admixed population. The pseudo-ancestors in the tree mimic observed levels of genetic diversity from contemporary samples of actual African, European, and Indigenous American populations.

I then simulate the formation of an admixed population from a single admixture event between two of the pseudo-ancestral populations. I construct a series of eight models whereby ancestry proportions are estimated from varying pseudo-ancestors in the tree. The first four models estimate ancestry from the continental sources of Africa and Europe in the formation of an African-American population. The last four models estimate ancestry from European and American continental sources in the formation of a Latin American population. I estimate ancestry proportions from simulated genotype data, which contains 500 loci from the individuals of each pseudo-ancestral population, as well as the focal admixed

population. For each model, I vary the sample sizes for the ancestral source populations, as well as for the admixed population. The first sampling scenario samples 100 individuals among all populations, for each of the two ancestral sources and for the admixed population. The second sampling scenario samples 100 individuals from the admixed population, and 20 individuals from each of the ancestral source populations. The third sampling scenario estimates ancestry from a sample of 20 individuals from the admixed population, and 100 individuals from each of the ancestral source populations.

Chapter 2

The Contributions of Admixture and Genetic Drift to Diversity Among Post-Contact Populations in the Americas

---

2.1 Overview

Objective: We present a partition of Nei's minimum genetic distance in admixed populations into components of admixture and genetic drift. We applied this technique to 17 admixed populations in the Americas to examine how admixture and drift have contributed to the patterns of genetic diversity.

Materials and Methods: We analyzed 618 short tandem repeat loci in 949 individuals from 49 population samples. Thirty-two samples serve as proxies for continental ancestors. Seventeen samples represent admixed populations: (4) African-American and (13)

Latin American. We estimate ancestry fractions and allele frequencies for all populations. We partition genetic distance and calculate fixation indices and principal coordinates to interpret our results.

Results: The partition of genetic distance shows that both admixture and genetic drift contribute to patterns of genetic diversity. The admixture component of genetic distance provides evidence for two distinct axes of continental ancestry. However, the genetic distances show that ancestry contributes to only one axis of genetic differentiation. The drift component of genetic distance indicates that modest founder effects accompanied admixture in the formation of these populations.

Discussion: Our results show that the genetic structure of admixed populations in the Americas reflects more than admixture. We show that the evolution of the source populations influenced the genetic structure of the admixed populations. Notably, the history of serial founder effects constrains the impact of admixture on allele frequencies to a single dimension. Founder effects in the admixed populations imposed a new level of genetic structure onto that created by admixture.

2.2 Introduction

European colonization of the Americas beginning in the late 15th century had a major impact on the human species by bringing people living in Europe and Africa to the Americas. Populations on these continents had been isolated from each other for thousands of years before this. The result of re-contact was the formation of new genetically mixed populations that trace their recent ancestry to two or more continental regions. Many mixed populations formed and each one constituted a unique gene pool. The mixed populations resided in geographic locations dispersed throughout the Americas, and to varying degrees, the populations were isolated from each other. Each newly formed population had its genetic diversity structured by factors such as the composition of African, European, and Indige-

nous American ancestors, and their specific degree of isolation. Many population genetic studies have compared the fractions of continental ancestry across admixed populations in the Americas (Shriver et al., 2003; Bonilla et al., 2004, 2005; Galanter et al., 2012). Recent efforts have linked admixed populations to subpopulations of the continental ancestral groups (Wang et al., 2008; Moreno-Estrada et al., 2014). However, ancestry fractions, no matter how fine-grained, do not fully account for the patterns of genetic diversity in the admixed populations. A full account requires consideration of other evolutionary processes, notably genetic drift. The Indigenous American population experienced dramatic decline and rebound in the relatively short timeframe since European contact (Livi-Bacci, 2006; O'Fallon and Fehren-Schmitz, 2011). Many of the admixed populations formed during the time of Indigenous American decline, and necessarily grew from small to large size in their early generations. In essence, the foundation of these populations would have encompassed both mixing of continental ancestry and founder effects. No study heretofore has connected the genetic diversity within and among the mixed populations to the combination of continental ancestry and genetic drift. To this end, we developed a model and methods of analysis to partition genetic distance into ancestry and drift components. We analyzed a diverse set of admixed populations in North and South America. We found that demographic forces that go beyond admixture, which are related to genetic drift, have played a prominent role in shaping patterns of diversity within and among admixed populations in the Americas.

2.3 Population Genetic Model

Figure 2.1 presents the basics of our model and its main parameters. For the purpose of explanation, we consider a pair of recently formed populations that we will label A and B. Both A and B have ancestry from two continental sources that we will label 1 and 2. We assume that the continental source populations have been isolated from each other for

enough time to allow their allele frequencies to diverge by genetic drift.

Nei's minimum genetic distance is a principal quantity in the formulation of our model (Nei, 1973, 1987). For a single genetic locus, this genetic distance is a function of allele frequencies computed according to the formula,

$$D_{hi}^2 = \frac{1}{2}\sum_j (p_{hj} - p_{ij})^2 \tag{2.1}$$

where, p_{hj} and p_{ij} represent the frequency of the j^{th} allele in the h^{th} and i^{th} populations, respectively. The summation is taken over all alleles at the locus. Minimum genetic distances are typically reported as averages over many genetic loci.

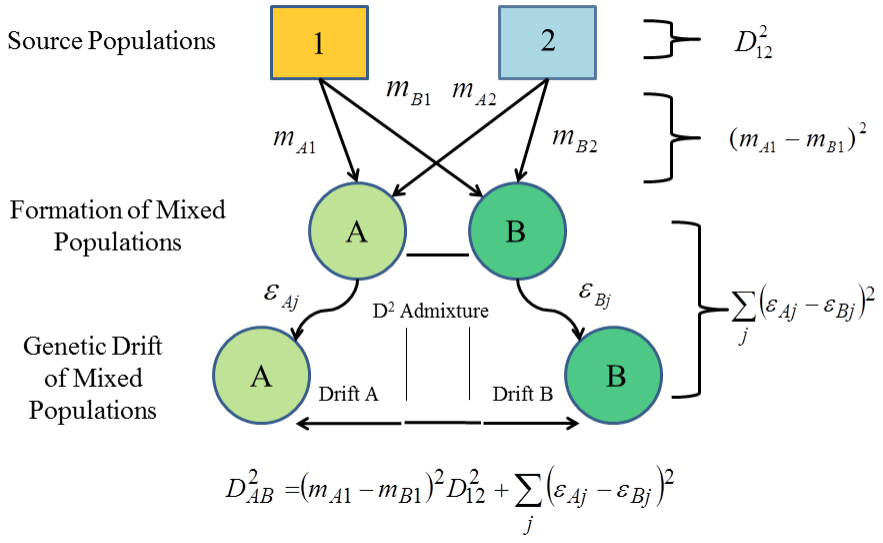

Figure 2.1: Schematic showing the independent contributions of admixture and genetic drift to genetic distance.

Our goal for the populations A and B is to partition their minimum genetic distance into two components, one representing ancestry and the other representing genetic drift. To accomplish this, we construct allele frequencies in A and B using allele frequencies in the source populations, ancestry fractions, and a contribution by genetic drift (Long, 1991). For the purpose of exposition, we will assume that 1 and 2 are the only ancestral populations

of A and B. This restriction can be relaxed and our methods generalize to any number of source populations that contributed ancestors to any number of mixed populations. We construct allele frequencies in the admixed populations according to the formulas

$$\begin{aligned} p_{Aj} &= p_{2j} + m_{A1}(p_{1j} - p_{2j}) + \epsilon_{Aj} \\ p_{Bj} &= p_{2j} + m_{B1}(p_{1j} - p_{2j}) + \epsilon_{Bj} \end{aligned} \tag{2.2}$$

where, m_{A1} and m_{B1} are the proportions of ancestry in populations A and B that were contributed by parental population 1. Since all ancestry in A and B must trace back to 1 or 2, we construct $m_{A2} = 1 - m_{A1}$ and $m_{B2} = 1 - m_{B1}$. The frequencies of the j^{th} allele in the source populations are represented by p_{1j} and p_{2j}, respectively. The final terms, ϵ_{Aj} and ϵ_{Bj} represent the deviations of the allele frequency from that which a pure admixture process would produce. We assume that these terms represent genetic drift occurring in the mixed populations during, or after, the admixture process.

To obtain the genetic distance between A and B in terms of our admixture and drift model, we substitute the allele frequency formulas from Eqs. 2.2 into Eq. 2.1

$$D_{AB}^2 = \sum_j (p_{2j} + m_{A1}(p_{1j} - p_{2j}) + \epsilon_{Aj} - p_{2j} - m_{B1}(p_{1j} - p_{2j}) - \epsilon_{Bj})^2 \tag{2.3}$$

After collecting terms and simplifying,

$$\begin{aligned} D_{AB}^2 &= (m_{A1} - m_{B1})^2 D_{12}^2 + \sum_j (\epsilon_{Aj} - \epsilon_{Bj})^2 \\ &= \Delta_{AB} + E_{AB} \end{aligned} \tag{2.4}$$

The component Δ_{AB} represents the portion of genetic distance related to admixture, while the component E_{AB} represents the portion of genetic distance related to drift, following admixture. It is clear from the admixture component of genetic distance that the impact of admixture depends on the level of differentiation of the source populations.

This model for genetic distance requires two assumptions. First, genetic drift and admixture are the only processes that have influenced allele frequencies in the admixed populations. Second, the effects of genetic drift and admixture have operated independently.

Neis minimum genetic distance is one of the simplest measures of genetic distance (Nei, 1987). We have chosen it as our primary metric because it is easy to partition into additive components related to the distinct processes of admixture and genetic drift. Moreover, this distance makes it easy to relate population differentiation to genetic phenomena such as homozygosity and heterozygosity. Some other measures of genetic distance (Shriver et al., 1995; Goldstein et al., 1995; Nei, 1973) utilize the mutation rate to measure divergence times in phylogenetic models. At best, these genetic distance measures provide indirect information about admixture. They are unsuited to the populations in this analysis because admixture produces genetic outcomes that are distinct from the outcomes of population fissions and phylogenetic radiation. We feel mutation is unlikely to influence the results for recently founded populations. In this light, we favor a method that is simple to interpret and likely to produce accurate results.

2.4 Materials and Methods

The focus of our analyses is a set of 17 populations of mixed ancestry in the Americas. This set includes 13 populations labeled in original sources as Mestizo (Wang et al., 2008) and four populations labeled in original sources as African-American (Tishkoff et al., 2009). Various investigators collected these samples in North and South America. To guide our analyses of mixed populations we include four populations labeled in original sources as

European, two populations labeled in original sources as Sub-Saharan African, and 26 populations labeled in original sources as Indigenous American (Cann et al., 2002; Rosenberg et al., 2002; Wang et al., 2007). The original investigators collected the African, European, and Indigenous American samples on their respective continents of origin. The primary sources for our data are (Cann et al., 2002; Rosenberg et al., 2002; Wang et al., 2007, 2008; Tishkoff et al., 2009). Tables 2.1 and 2.2 give the population names, geographic coordinates, sample sizes, and primary references for all 49 populations.

We analyze genotypes at 618 autosomal short tandem repeat (STR) loci. The genotyping service at the Marshfield Clinic performed the laboratory analyses for all of the original studies. The Marshfield Clinic selected these loci for linkage mapping in other studies. The loci are spaced on the genetic map approximately 5 cM to 10 cM apart. We use data from the set that Pemberton et al. (2013) created by calibrating allele sizes and combining across the original studies.

To test the statistical significance of genetic distance estimates between samples, we constructed confidence intervals using the jackknife method (Efron and Tibshirani, 1993). We rejected the null hypothesis of zero genetic distance if the confidence interval for an estimate did not span zero. One-sided confidence intervals are appropriate for these tests because genetic distance cannot be negative.

Table 2.1: Sampled contemporary populations that serve as ancestral proxies in our analyses, along with their associated sample sizes, global locations, and primary references.

Population Name	Sample Size	GPS Coordinates	Primary Reference
Orcadian	16	59°N −3°E	Cann et al. (2002); Rosenberg et al. (2002)
French	29	46°N 2°E	Cann et al. (2002); Rosenberg et al. (2002)
Italian	13	46°N 10°E	Cann et al. (2002); Rosenberg et al. (2002)
Russian	25	61°N 40°E	Cann et al. (2002); Rosenberg et al. (2002)
Mandenka	24	12°N −12°E	Cann et al. (2002); Rosenberg et al. (2002)
Yoruba	25	8°N 4°E	Cann et al. (2002); Rosenberg et al. (2002)
Yoruba	25	7.9°N 5°E	Tishkoff et al. (2009)
Pima	25	29°N −108°E	Cann et al. (2002); Rosenberg et al. (2002)
Mixtec	19	17°N −97°E	Wang et al. (2007)
Zapotec	17	16°N −97°E	Wang et al. (2007)
Mixe	20	17°N −96°E	Wang et al. (2007)
Maya	25	19°N −91°E	Cann et al. (2002); Rosenberg et al. (2002)
Kaqchikel	12	15°N −91°E	Wang et al. (2007)
Cabecar	20	9.5°N −84°E	Wang et al. (2007)
Guaymi	16	8.5°N −82°E	Wang et al. (2007)
Kogi	16	11oN −74°E	Wang et al. (2007)
Arhuaco	16	11°N −73.8°E	Wang et al. (2007)
Waunana	20	5°N −77°E	Wang et al. (2007)
Embera	11	7°N −76°E	Wang et al. (2007)
Zenu	18	9°N−75°E	Wang et al. (2007)
Inga	16	1°N −77°E	Wang et al. (2007)
Quechua	20	−14°−74°E	Wang et al. (2007)
Aymara	18	−22°N −70°E	Wang et al. (2007)
Huilliche	19	−41°N −73°E	Wang et al. (2007)
Kaingang	5	−24°N −52.5°E	Wang et al. (2007)
Guarani	10	−23°N −54°E	Wang et al. (2007)
Wayuu	17	11°N −73°E	Wang et al. (2007)
Piapoco-Curripaco	13	3°N −68°E	Cann et al. (2002); Rosenberg et al. (2002)
Ticuna Tarapaca	18	−4°N −70°E	Wang et al. (2007)
Ticuna Arara	15	−4°N −70°E	Wang et al. (2007)
Karitiana	24	−10°N −63°E	Cann et al. (2002); Rosenberg et al. (2002)
Surui	21	−11°N −62°E	Cann et al. (2002); Rosenberg et al. (2002)
Ache	17	−24°N −56°E	Wang et al. (2007)

Table 2.2: Sampled admixed populations used in our analyses, along with their associated sample sizes, global locations, and primary references.

Population Name	Sample Size	GPS Coordinates	Primary Reference
Oriente	19	$14.63°\text{N}\ -89.7°\text{E}$	Wang et al. (2008)
Mexico City	19	$19.4°\text{N}\ -99.2°\text{E}$	Wang et al. (2008)
CVCR	20	$11.5°\text{N}\ -84.1°\text{E}$	Wang et al. (2008)
Quetalmahue	20	$-42.4°\text{N}\ -73.5°\text{E}$	Wang et al. (2008)
Paposo	20	$24°\text{N}\ -70°\text{E}$	Wang et al. (2008)
Catamarca	12	$-29.3°\text{N}\ -65.8°\text{E}$	Wang et al. (2008)
Salta	19	$-24.8°\text{N}\ -65.4°\text{E}$	Wang et al. (2008)
Tucuman	19	$-27°\text{N}\ -65.2°\text{E}$	Wang et al. (2008)
RGS	20	$-31°\text{N}\ -54°\text{E}$	Wang et al. (2008)
Pasto	19	$1°\text{N}\ -78.5°\text{E}$	Wang et al. (2008)
Peque	20	$7.6°\text{N}\ -73°\text{E}$	Wang et al. (2008)
Medellin	20	$5.4°\text{N}\ -74.4°\text{E}$	Wang et al. (2008)
Cundinamarca	19	$3.2°\text{N}\ -74.1°\text{E}$	Wang et al. (2008)
Chicago	15	$42°\text{N}\ -87.9°\text{E}$	Tishkoff et al. (2009)
Pittsburgh	21	$40.5°\text{N}\ -80.2°\text{E}$	Tishkoff et al. (2009)
Baltimore	44	$39.2°\text{N}\ -76.7°\text{E}$	Tishkoff et al. (2009)
North Carolina	18	$35.9°\text{N}\ -78.8°\text{E}$	Tishkoff et al. (2009)

Fitting our population genetic model requires us to estimate each component of allele frequency given by Eq. 2.2. The following estimation steps underlie our analysis. (1) We identify source populations. (2) We estimate allele frequencies for the sources. (3) We estimate for the mixed populations the fraction of their ancestry attributable to each source population. (4) We estimate expected allele frequencies for each mixed ancestry population. The expected allele frequencies for a mixed population are the averages of source population allele frequencies weighted by the fractions of ancestry in mixed populations that are attributable to the sources. (5) We estimate the drift deviations for each allele frequency, in each mixed population, as the difference between the observed and expected allele frequencies.

We use the maximum likelihood approach of Tang and colleagues to make the estimates described in the previous paragraph (Tang et al., 2005). This method assumes a population model in which the ancestry in a mixed group traces back to a pre-specified number K of

ancestral sources. The method assumes that each allele in each genotype of an individual with mixed ancestry represents an independent draw from one of the source populations. This is equivalent to assuming that genotypes in mixed populations are in Hardy-Weinberg equilibrium when conditioned on the ancestry fractions. The method requires us to assume that the STR marker loci are also in linkage equilibrium when conditioned on ancestry fractions.

We have written new software for the method to accommodate STR data. We wrote this software using the Bloodshed Development Environment (http//www.bloodshed.net) in the C++ language. Prior implementations of Tang's method are restricted to single nucleotide polymorphism data (Alexander et al., 2009; Tang et al., 2005). The likelihood function is of extremely high dimension when applied to genomic scale data. Maximizing this function requires estimating thousands of parameters, consisting of allele frequencies and ancestry fractions. Our program uses the EM algorithm described by Tang and colleagues (2005) as a numerical method to obtain asymptotic results from the likelihood equation. Alexander and colleagues (2009) note that a stringent convergence criterion is necessary to obtain precise results.

Determining the number of source populations is a special case of determining the number of clusters in a mixture. This is a long-standing problem in statistics and population genetics. Following Tang et al. (2005), we intend that our source populations represent populations that were isolated on different continents in pre-Columbian times, but we investigate the possibility that alternative models with more source populations per continent may fit the data better than a model with one source per continent. To distinguish models, we apply the standard approach of tracking the increase in model likelihood that occurs with increasing the number of source populations, $i.e.$, increasing K. However, we take some additional steps too. We perform multiple runs and of the program and check for consistency in the maximized likelihood across runs. Then, we check the individual mixed populations to be certain that they make the same overall contribution to the overall like-

lihood across runs. Finally, we check the individual mixed populations to make sure the contributions of the source populations remain constant across replicate runs of the same model. In light of the complexity in identifying actual source populations and estimating their allele frequencies, we follow the precedence of recognizing these putative source populations as *pseudo-ancestors* (Tang et al., 2005).

We take the following steps to partition unbiased estimates of genetic distance into admixture and drift components. These equations allow any number of K ancestral source populations. First, we calculate Nei's unbiased estimate of minimum genetic distance between admixed populations A and B. Second, we create expected allele frequencies for each admixed sample according to

$$\hat{p}_{Aj} = \sum_{s=1}^{K} \hat{m}_{As}\hat{p}_{sj} \tag{2.5}$$

where, $\hat{p}_{Aj}$ is the expected frequency of the j^{th} allele in the A^{th} admixed population, and $\hat{m}_{As}$ is the estimated contribution of the s^{th} ancestral source population to the A^{th} admixed population. Third, we compute the admixture portion of the estimated genetic distance between admixed populations A and B as

$$\hat{\Delta}_{AB} = \sum_{j} (\hat{p}_{Aj} - \hat{p}_{Bj})^2 \tag{2.6}$$

Fourth, we compute the drift portion of the estimated genetic distance as

$$\hat{E}_{AB} = \hat{D}_{AB} - \hat{\Delta}_{AB} \tag{2.7}$$

where, $\hat{D}_{AB}$ is the estimate of Nei's minimum genetic distance.

We use original scripts written for the R statistical computing environment to manipulate allele frequency output from our likelihood program, to compute genetic distance matrices and their partitions, and to produce graphs (R Core Team, 2014).

To facilitate interpretation of our results, we use two supplemental approaches. First,

we compute the fixation index F_{ST} to help assess the extent of genetic drift in admixed populations (Wright, 1951; Nei, 1987; Long, 1991). We use the general formula,

$$\hat{F}_{ST} = \frac{\hat{H}_T - \hat{H}_O}{\hat{H}_T} \tag{2.8}$$

for estimation, where $\hat{H}_T$ is the estimated heterozygosity in a base population and $\hat{H}_O$ is the estimated heterozygosity in an observed sample. For the *pseudo-ancestors*, we compute $\hat{H}_T$ from the allele frequencies estimated for the specific continental source population, and for the admixed populations we compute $\hat{H}_T$ from the allele frequencies expected from the admixture process. Second, we use principal coordinates to represent distance matrices in lower dimension (Gower, 1966). We use the multidimensional scaling function in R to compute principal coordinates.

2.5 Results

We estimated genetic ancestry and allele frequencies twice. First, we assumed that K=3 ancestral source populations contributed to the 49 contemporary samples, and second we expanded to K=4 ancestral sources. We constructed these analyses in a partially supervised fashion. We constrained individuals from the four European samples to have 100% ancestry from one source population, and individuals from the two African samples to have 100% ancestry from a second source population. This construction obligated source one to represent European ancestors, and source two to represent African ancestors, and by default, sources three and four represented Indigenous American ancestors. We estimated ancestry in the populations labeled Indigenous American because prior research shows mixed continental ancestry in some of these samples (Hunley and Healy, 2011). All models necessitated estimating 6,333 independent allele frequencies per ancestral source population, and ancestry fractions for 792 individuals. In total, K=3 required estimating 20,583 parameters, and K=4 required estimating 26,916 parameters. To fit models, we used random starting

values for all parameters, and iterated the EM procedure until the likelihood changed by less than 10^{-6} between successive steps.

With K=3, we were able to replicate the highest likelihood in several runs of the ancestry estimation program using different starting values. Importantly, our ancestry estimates for individuals and populations were consistent across runs, generally not differing by more than 0.001. By contrast, our results for models with K=4 were less successful. Although, running the program with K=4 always yielded higher likelihoods than running it with K=3, we could not replicate the best likelihood on independent runs. Moreover, we found with K=4 that parameter estimates could be quite different from runs of the program that produced similar likelihoods. In light of our limited success with K=4, we preformed all subsequent analyses of ancestry and drift contributions to genetic distance using maximum likelihood estimates with K=3.

Table 2.3 gives sample size and estimates of continental ancestry for the 17 post-contact populations. The African-American populations have ancestry proportions similar to each other (approximately, 80% African and 20% European). By contrast, the Latin American populations vary widely in their ancestry; average African ancestry varies from 0% to 9%, average European ancestry varies from 33% to 73%, and average Indigenous American ancestry varies from 18% to 64%. The wide variation in ancestry within Latin American populations, and between Latin American and African-American populations, makes our questions about the contribution of variation in ancestry to genetic distance particularly salient.

Table 2.3: The post-contact populations included in our analyses with their associated sample sizes, inferred average continental ancestry, F_{ST}, and log likelihood estimates.

Sample	n	African	European	Indigenous American	F_{ST}	$lnL(i)$
Chicago	15	0.788	0.2	0.012	0.0006	-29,968
Pittsburgh	21	0.79	0.196	0.014	0	-43,142
Baltimore	44	0.828	0.158	0.014	0	-89,668
North Carolina	18	0.775	0.204	0.021	0	-36,181
Mexico City	19	0.035	0.621	0.344	0.0001	-35,594
Oriente	19	0.069	0.456	0.474	0.0002	-35,687
CVCR	20	0.044	0.711	0.245	0.0007	-38,385
Peque	20	0.051	0.437	0.512	0.0199	-37,423
Medellin	20	0.093	0.697	0.211	0.0042	-38,857
Cundinamarca	19	0.02	0.529	0.451	0.0051	-35,443
Pasto	19	0.035	0.457	0.508	0.0053	-34,909
Salta	19	0.024	0.332	0.644	0.0054	-33,878
Paposo	20	0.018	0.499	0.483	0.0176	-35,948
Tucuman	19	0.044	0.698	0.258	0	-35,164
Catamarca	12	0.027	0.594	0.379	0.0082	-22,480
RGS	20	0.094	0.731	0.175	0	-38,371
Quetalmahue	20	0.004	0.564	0.432	0.0293	-36,803

Table 2.3 also gives estimates of F_{ST}, which measures the drift of allele frequencies in each post-contact population from the expectations set by admixture of intercontinental sources. The four African-American populations independently show minimal influence from genetic drift based on F_{ST}. The Latin American populations show varying impact of genetic drift. F_{ST} is less than 0.001 for five populations, and greater than 0.01 for three populations. The remaining five Latin American populations show intermediate impact of drift, $0.001 \leq F_{ST} \leq 0.01$. While F_{ST} in this intermediate range seems low, it is typical of populations on the European continent.

We calculated the matrix of Nei's minimum genetic distances among pairs of the 49 populations analyzed (17 post-contact populations and 32 indigenous continental populations). The 17 post-contact populations (4 African-American and 13 Latin American) yield 136 pairs. The genetic distance was statistically significant with *p-values* below 0.05 for 135 of these pairs. The highest *p-value* was 0.06 between the African-American popula-

tions in Pittsburgh and Baltimore. The *p-value* was below 0.0001 for 116 of the pairwise comparisons. A *p-value* of 0.0004 is required for a conservative Bonferroni correction for multiple comparisons (Sokal and Rohlf, 2012). To visualize patterns, we extracted the Eigen vectors produced by multidimensional scaling to summarize patterns of genetic distance among these 49 populations.

Figures 2.2 and 2.3 show the outcomes of our ancestry and genetic distance analyses. Figure 2.2 displays ancestry estimates for the 49 populations in a triangle plot. As expected, the continental populations are concentrated on the vertices. The 17 post-contact populations occupy intermediate locations. The four African-American samples cluster tightly on the axis between continental African and continental Europeans. The 13 Latin American populations disperse along the axis between European and Indigenous American populations.

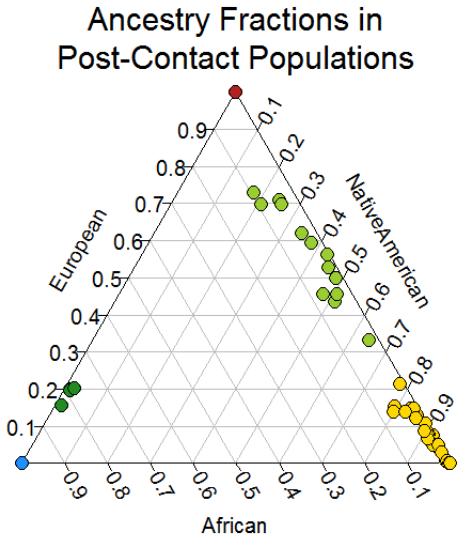

Figure 2.2: Proportions of continental ancestry fill a two-dimensional space defined by the constraint that ancestry fractions sum to 1.0. Ancestry estimates are presented for 49 populations. African (blue) and European (red) samples were constrained to 100% ancestry from their respective continental sources. The ancestry of contemporary Indigenous Americans (gold) was estimated from a three-way admixture model to account for recently introduced European and African ancestry. Samples from African-American populations are shaded dark green. Samples from Latin American populations shaded light green shading.

Figure 2.3 plots the first two principal coordinates of the matrix of genetic distances among the 49 populations. Population positions along the first axis, which accounts for 50% of the dispersion, correlate with continental ancestry. African populations occupy one extreme and Indigenous American populations occupy the other extreme. European populations lie intermediate to the other two continental groups. African-Americans lie between the African and European populations. Latin Americans lie between European and Indigenous American populations. Unexpectedly, the second principal coordinate separates a pair of Indigenous American population. The next several axes primarily differentiate Indigenous Americans.

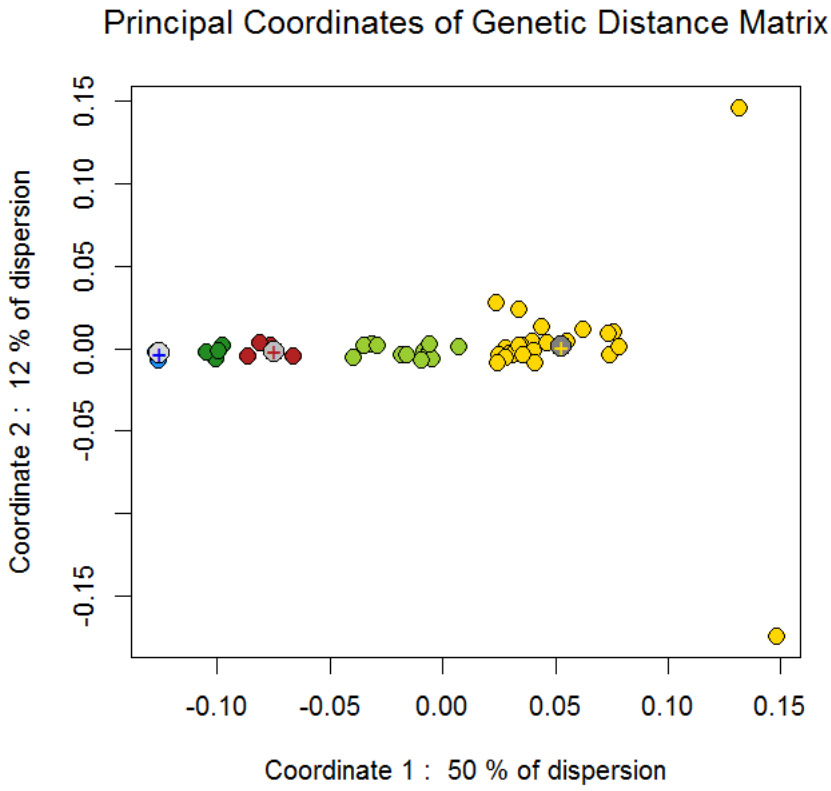

Figure 2.3: Only the first principal coordinate of the genetic distance matrix shows the ancestry pattern of continental populations and their post-contact descendants formed by admixture. The second, and subsequent coordinates, primarily reveal the extreme divergence of Indigenous Americans from a continental gene pool. The color conventions are those established in Figure 2. The dots containing crosses represent the putative ancestral populations for the continental sources.

We present Nei's minimum genetic distances for the 17 post-contact population in the Americas (Table 2.4). We partitioned the matrix of Nei's minimum genetic distances among the 17 post-contact populations into a matrix of admixture distances and a matrix of drift distances (Tables 2.5, and 2.6). Then we used the Eigen vectors produced by multidimensional scaling to summarize patterns within the distance matrices (fig. 2.4).

Table 2.4: Nei's minimum genetic distance for all the admixed populations included in our analyses.

	MC	OR	CR	PQ	MD	CN	PS	SL	PP	TC	CT	RGS	QT	BL	CH	NC	PT
MC	0	0.0041	0.0048	0.0235	0.0066	0.0042	0.0082	0.0156	0.0149	0.0049	0.006	0.0047	0.02	0.0523	0.0509	0.0465	0.0466
OR	0.0041	0	0.0096	0.0175	0.0125	0.0028	0.0056	0.0106	0.0144	0.0085	0.0058	0.013	0.0225	0.0573	0.0532	0.0503	0.0528
CR	0.0048	0.0096	0	0.0223	0.0043	0.007	0.0152	0.0248	0.02	0.0044	0.0112	0.0039	0.0247	0.0463	0.0464	0.0416	0.0416
PQ	0.0235	0.0175	0.0223	0	0.0237	0.0168	0.0212	0.0256	0.0303	0.0241	0.0214	0.0307	0.0353	0.0764	0.0753	0.0691	0.0711
MD	0.0066	0.0125	0.0043	0.0237	0	0.0094	0.0155	0.0281	0.0226	0.0052	0.0093	0.0026	0.0296	0.0424	0.0437	0.0394	0.0402
CN	0.0042	0.0028	0.007	0.0168	0.0094	0	0.0046	0.0116	0.0153	0.0078	0.0067	0.0131	0.021	0.0613	0.059	0.0567	0.0577
PS	0.0082	0.0056	0.0152	0.0212	0.0155	0.0046	0	0.0095	0.0151	0.0124	0.009	0.0183	0.0196	0.0647	0.0623	0.0592	0.0587
SL	0.0156	0.0106	0.0248	0.0256	0.0281	0.0116	0.0095	0	0.0181	0.0209	0.0135	0.0305	0.0256	0.0801	0.0783	0.0757	0.0747
PP	0.0149	0.0144	0.02	0.0303	0.0226	0.0153	0.0151	0.0181	0	0.0195	0.0166	0.0252	0.0261	0.0738	0.0721	0.0699	0.0698
TC	0.0049	0.0085	0.0044	0.0241	0.0052	0.0078	0.0124	0.0209	0.0195	0	0.0059	0.005	0.0246	0.0458	0.0439	0.0413	0.041
CT	0.006	0.0058	0.0112	0.0214	0.0093	0.0067	0.009	0.0135	0.0166	0.0059	0	0.0119	0.0196	0.0564	0.0548	0.0535	0.0521
RGS	0.0047	0.013	0.0039	0.0307	0.0026	0.0131	0.0183	0.0305	0.0252	0.005	0.0119	0	0.0287	0.0387	0.0371	0.0334	0.0354
QT	0.02	0.0225	0.0247	0.0353	0.0296	0.021	0.0196	0.0256	0.0261	0.0246	0.0196	0.0287	0	0.0787	0.0796	0.0761	0.0748
BL	0.0523	0.0573	0.0463	0.0764	0.0424	0.0613	0.0647	0.0801	0.0738	0.0458	0.0564	0.0387	0.0787	0	0.0048	0.003	0.0016
CH	0.0509	0.0532	0.0464	0.0753	0.0437	0.059	0.0623	0.0783	0.0721	0.0439	0.0548	0.0371	0.0796	0.0048	0	0.0043	0.0048
NC	0.0465	0.0503	0.0416	0.0691	0.0394	0.0567	0.0592	0.0757	0.0699	0.0413	0.0535	0.0334	0.0761	0.003	0.0043	0	0.004
PT[a]	0.0466	0.0528	0.0416	0.0711	0.0402	0.0577	0.0587	0.0747	0.0698	0.041	0.0521	0.0354	0.0748	0.0016	0.0048	0.004	0

[a]MC=Mexico City, Mexico; OR=Oriente, Guatemala; CR=Central Valley, Costa Rica; PQ=Peque, Colombia; MD=Medellin, Colombia; CN=Cundinamarca, Colombia; PS=Pasto, Colombia; SL=Salta, Argentina; PP=Paposo, Chile; TC=Tucuman, Argentina; CT=Catamarca, Argentina; RGS=Rio Grande do Sul, Brazil; QT=Quetalmahue, Chile; BL=Baltimore, United States; CH=Chicago, United States, NC=North Carolina, United States; PT=Pittsburgh, United States

Table 2.5: Ancestry partition of Nei's minimum genetic distance for the admixed populations included in our analyses.

	MC	OR	CR	PQ	MD	CN	PS	SL	PP	TC	CT	RGS	QT	BL	CH	NC	PT
MC	0	0.0021	0.0011	0.0033	0.002	0.0013	0.003	0.01	0.0021	0.0008	0.0001	0.0031	0.0009	0.0516	0.0474	0.0456	0.0475
OR	0.0021	0	0.0061	0.0002	0.0076	0.0003	0.0002	0.0031	0.0002	0.0055	0.0013	0.0098	0.0006	0.0561	0.0523	0.0504	0.0524
CR	0.0011	0.0061	0	0.008	0.0003	0.0046	0.0076	0.0175	0.0061	0	0.0019	0.0006	0.0038	0.0471	0.0427	0.0411	0.0429
PQ	0.0033	0.0002	0.008	0	0.0098	0.0005	0	0.0019	0.0002	0.0072	0.0021	0.0122	0.001	0.061	0.0572	0.0551	0.0572
MD	0.002	0.0076	0.0003	0.0098	0	0.0063	0.0095	0.0203	0.008	0.0004	0.0031	0.0001	0.0054	0.041	0.0368	0.0354	0.037
CN	0.0013	0.0003	0.0046	0.0005	0.0063	0	0.0004	0.0042	0.0001	0.0041	0.0006	0.0082	0.0001	0.0596	0.0555	0.0535	0.0556
PS	0.003	0.0002	0.0076	0	0.0095	0.0004	0	0.002	0.0001	0.0069	0.0019	0.0119	0.0008	0.0624	0.0585	0.0564	0.0585
SL	0.01	0.0031	0.0175	0.0019	0.0203	0.0042	0.002	0	0.0029	0.0164	0.0078	0.0238	0.0052	0.077	0.0733	0.0709	0.0732
PP	0.0021	0.0002	0.0061	0.0002	0.008	0.0001	0.0001	0.0029	0	0.0055	0.0012	0.0102	0.0003	0.0622	0.0581	0.0561	0.0582
TC	0.0008	0.0055	0	0.0072	0.0004	0.0041	0.0069	0.0164	0.0055	0	0.0016	0.0008	0.0033	0.0475	0.0431	0.0415	0.0433
CT	0.0001	0.0013	0.0019	0.0021	0.0031	0.0006	0.0019	0.0078	0.0012	0.0016	0	0.0045	0.0003	0.0543	0.0501	0.0482	0.0502
RGS	0.0031	0.0098	0.0006	0.0122	0.0001	0.0082	0.0119	0.0238	0.0102	0.0008	0.0045	0	0.0072	0.0403	0.0361	0.0347	0.0363
QT	0.0009	0.0006	0.0038	0.001	0.0054	0.0001	0.0008	0.0052	0.0003	0.0033	0.0003	0.0072	0	0.0601	0.0558	0.0538	0.0559
BL	0.0516	0.0561	0.0471	0.061	0.041	0.0596	0.0624	0.077	0.0622	0.0475	0.0543	0.0403	0.0601	0	0.0001	0.0002	0.0001
CH	0.0474	0.0523	0.0427	0.0572	0.0368	0.0555	0.0585	0.0733	0.0581	0.0431	0.0501	0.0361	0.0558	0.0001	0	0	0
NC	0.0456	0.0504	0.0411	0.0551	0.0354	0.0535	0.0564	0.0709	0.0561	0.0415	0.0482	0.0347	0.0538	0.0002	0	0	0
PT[a]	0.0475	0.0524	0.0429	0.0572	0.037	0.0556	0.0585	0.0732	0.0582	0.0433	0.0502	0.0363	0.0559	0.0001	0	0	0

[a]MC=Mexico City, Mexico; OR=Oriente, Guatemala; CR=Central Valley, Costa Rica; PQ=Peque, Colombia; MD=Medellin, Colombia; CN=Cundinamarca, Colombia; PS=Pasto, Colombia; SL=Salta, Argentina; PP=Paposo, Chile; TC=Tucuman, Argentina; CT=Catamarca, Argentina; RGS=Rio Grande do Sul, Brazil; QT=Quetalmahue, Chile; BL=Baltimore, United States; CH=Chicago, United States, NC=North Carolina, United States; PT=Pittsburgh, United States

Table 2.6: Drift partition of Nei's minimum genetic distance for the admixed populations included in our analyses.

	MC	OR	CR	PQ	MD	CN	PS	SL	PP	TC	CT	RGS	QT	BL	CH	NC	PT
MC	0	0.0019	0.0037	0.0202	0.0046	0.0029	0.0052	0.0057	0.0128	0.0041	0.0058	0.0016	0.0191	0.0007	0.0035	0.0009	0
OR	0.0019	0	0.0035	0.0174	0.0049	0.0025	0.0054	0.0075	0.0142	0.003	0.0045	0.0032	0.0219	0.0012	0.0009	0	0.0004
CR	0.0037	0.0035	0	0.0143	0.004	0.0024	0.0075	0.0072	0.0138	0.0044	0.0093	0.0033	0.021	0	0.0037	0.0005	0
PQ	0.0202	0.0174	0.0143	0	0.0139	0.0162	0.0212	0.0237	0.0301	0.0169	0.0192	0.0185	0.0343	0.0154	0.0181	0.014	0.0139
MD	0.0046	0.0049	0.004	0.0139	0	0.0032	0.006	0.0078	0.0146	0.0048	0.0061	0.0025	0.0242	0.0014	0.0069	0.0041	0.0032
CN	0.0029	0.0025	0.0024	0.0162	0.0032	0	0.0042	0.0074	0.0152	0.0037	0.0062	0.0049	0.021	0.0017	0.0035	0.0032	0.0022
PS	0.0052	0.0054	0.0075	0.0212	0.006	0.0042	0	0.0075	0.015	0.0055	0.0071	0.0064	0.0188	0.0023	0.0039	0.0028	0.0002
SL	0.0057	0.0075	0.0072	0.0237	0.0078	0.0074	0.0075	0	0.0152	0.0045	0.0057	0.0067	0.0204	0.0031	0.005	0.0048	0.0015
PP	0.0128	0.0142	0.0138	0.0301	0.0146	0.0152	0.015	0.0152	0	0.014	0.0154	0.015	0.0258	0.0116	0.014	0.0138	0.0116
TC	0.0041	0.003	0.0044	0.0169	0.0048	0.0037	0.0055	0.0045	0.014	0	0.0044	0.0042	0.0214	0	0.0008	0	0
CT	0.0058	0.0045	0.0093	0.0192	0.0061	0.0062	0.0071	0.0057	0.0154	0.0044	0	0.0074	0.0193	0.0021	0.0047	0.0053	0.0019
RGS	0.0016	0.0032	0.0033	0.0185	0.0025	0.0049	0.0064	0.0067	0.015	0.0042	0.0074	0	0.0215	0	0.001	0	0
QT	0.0191	0.0219	0.021	0.0343	0.0242	0.021	0.0188	0.0204	0.0258	0.0214	0.0193	0.0215	0	0.0186	0.0238	0.0223	0.0189
BL	0.0007	0.0012	0	0.0154	0.0014	0.0017	0.0023	0.0031	0.0116	0	0.0021	0	0.0186	0	0.0046	0.0028	0.0015
CH	0.0035	0.0009	0.0037	0.0181	0.0069	0.0035	0.0039	0.005	0.014	0.0008	0.0047	0.001	0.0238	0.0046	0	0.0043	0.0048
NC	0.0009	0	0.0005	0.014	0.0041	0.0032	0.0028	0.0048	0.0138	0	0.0053	0	0.0223	0.0028	0.0043	0	0.004
PT[a]	0	0.0004	0	0.0139	0.0032	0.0022	0.0002	0.0015	0.0116	0	0.0019	0	0.0189	0.0015	0.0048	0.004	0

[a]MC=Mexico City, Mexico; OR=Oriente, Guatemala; CR=Central Valley, Costa Rica; PQ=Peque, Colombia; MD=Medellin, Colombia; CN=Cundinamarca, Colombia; PS=Pasto, Colombia; SL=Salta, Argentina; PP=Paposo, Chile; TC=Tucuman, Argentina; CT=Catamarca, Argentina; RGS=Rio Grande do Sul, Brazil; QT=Quetalmahue, Chile; BL=Baltimore, United States; CH=Chicago, United States, NC=North Carolina, United States; PT=Pittsburgh, United States

We partitioned the matrix of Nei's minimum genetic distances among the 17 post-contact populations into a matrix of admixture distances and a matrix of drift distances. Then we used the Eigen vectors produced by multidimensional scaling to summarize patterns within the distance matrices.

The admixture distance matrix produced one Eigen vector that explained 98.6% of the dispersion (Figure 2.4-top). The positions of the 17 post-contact populations on this axis correlate strongly with ancestry fractions. R^2 = 0.91 between position and either African ancestry or Indigenous American ancestry. R^2 = 0.37 with European ancestry; however, R^2 = 0.96 when computed between European ancestry and the absolute value of axis position. Because our model includes three sources of continental ancestry - African, European, and Indigenous American - we expected to find two principal axes of ancestry. However, the genetic diversity among continental sources forms a linear gradient that projects into the mixtures among sources.

Ten Eigen vectors explained the drift distance matrix. However, most of the dispersion was concentrated in the first three Eigen vectors (Figure 2.4-bottom). It is easy to see how these Eigen vectors relate to drift by comparing the positions of populations to their population specific estimates of F_{ST} (Table 2.3). The populations with the highest values of F_{ST} occupy the terminal positions on the first axis. The second axis draws a contrast between the population with the third highest estimate of F_{ST} and the two populations with higher F_{ST}. The next three axes contrast populations with middle levels of F_{ST}. Axes seven through ten explain small amounts of dispersion, and do not reflect a coherent pattern.

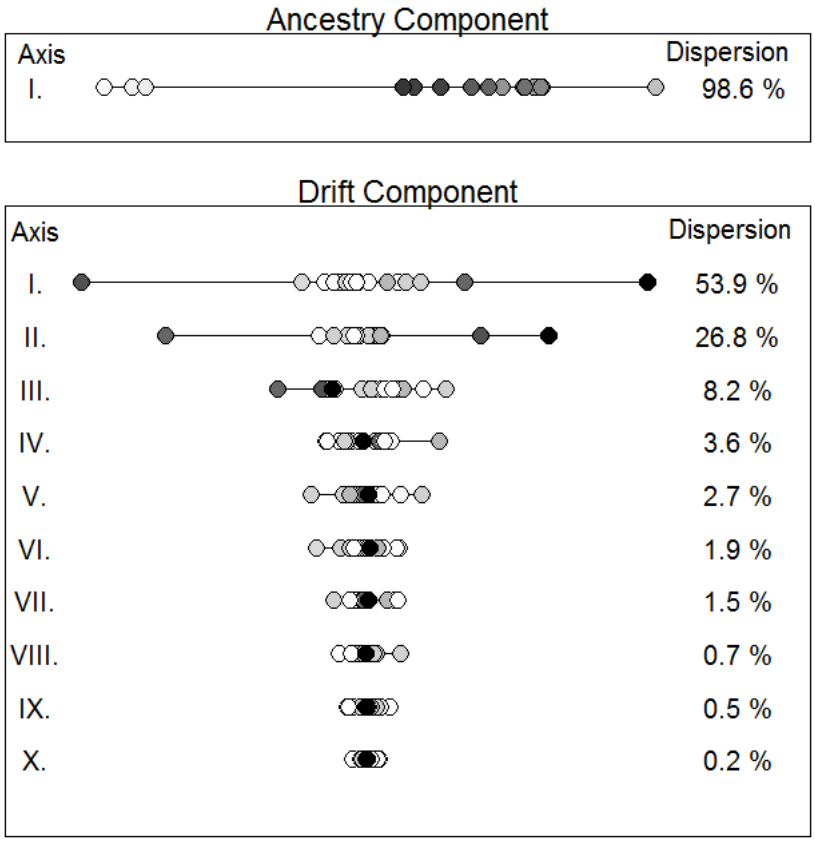

Figure 2.4: (Top) Positions of the 17 post-contact populations along the principal Eigen vector of the ancestry component of the genetic distance matrix. The shading represents increased European ancestry. (Bottom) Positions of the 17 populations along ten principal Eigen vectors of the drift component of the genetic distance matrix. The shading is proportional to F_{ST}.

Finally, we can relate the patterns found in our decomposition of genetic distances back to the patterns evident in the total distance matrix. The positions of populations on the first Eigen vector of the total distance matrix correlate highly with their positions on the principal Eigen vector of the admixture distance matrix (R^2 = 0.97). Moreover, the first Eigen vector of the total distance shows little correlation with any of the ten Eigen vectors of the drift distance matrix ($0.00 < R^2 < 0.07$). The positions of populations on the second Eigen vector of the total distance matrix are uncorrelated with the Eigen vector of the admixture distance matrix. However, they show strong correlation with positions of populations on the first Eigen vector of the drift genetic distance matrix (R^2= 0.88) and little correlation with positions on the remaining Eigen vectors of genetic drift ($0.00 <$

$R^2 < 0.03$). In a similar vein, the third Eigen vector of the total distance matrix shows high correlation with the second Eigen vector of the drift distances (R^2 = 0.74) and little correlation with positions on the remaining Eigen vectors of genetic drift ($0.00 < R^2 < 0.14$).

Figure 2.5 shows some interesting unexpected patterns involving the role of genetic drift in the differentiation of post-contact populations in the Americas. Overall, the trend is negative - the greater the genetic distance the less genetic drift has contributed to it. However, this negative trend is absent in all three groupings of populations, when viewed individually. African-American - by - African-American comparisons show a strong positive relationship (R^2 = 0.69), although the number of comparisons is small, and the trend is not statistically significant. When comparing pairs of Latin American populations, there is no relationship between the total genetic distance and the percent that genetic drift accounts for (R^2 = 0.00). For example, among Latin American populations showing the least differentiation, between 20% and 100% of the total differentiation owes to drift. Similarly, among Latin American populations showing the most differentiation, between 20% and 100% of the total differentiation owes to drift. Finally, genetic drift can be important to differentiation, even when comparing a Latin American population with an African-American population.

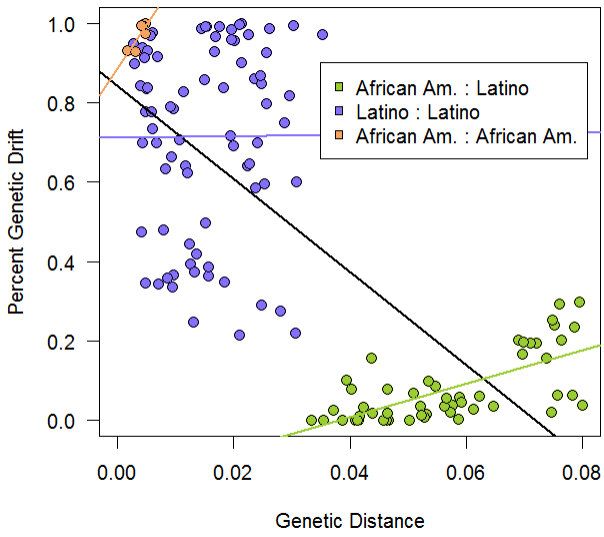

Figure 2.5: Percent genetic distance owing to genetic drift plotted against total genetic distance. Points are color-coded to identify three levels of comparison: African-American by African-American (orange), Latin American by Latin American (blue), and Latin American by African-American (green).

2.6 Discussion

Estimating the allele frequencies of source populations has been a significant issue throughout the history of genetic admixture studies (Reed, 1969; Cavalli-Sforza and Bodmer, 1971; Adams and Ward, 1973). There are two significant problems. First, the populations that mixed may be unidentified, or no longer exist (Chakraborty, 1986). Second, the source populations may have evolved since the time of mixing. Modern statistical methods partially ameliorate both these problems. The likelihood method from Tang et al. (2005) apportions allele frequencies from the mixed samples back to the ancestral sources. However, allele frequencies from modern proxies guide the apportionment, and poor choices for the proxies can bias the ancestry estimation. Allele frequency drift in the proxies may also skew the apportionment. Despite this potential problem, we found an admixture model that fits this large data set well. F_{ST} is below 0.001 in all four African-American samples, and five of the thirteen Latin American samples. F_{ST} exceeds 0.01 in only three Latin American

samples. Nonetheless, the three populations that show the most drift enter into about a third of the pairwise distance comparisons.

There are different ways to characterize the genetic structure of admixed populations. One approach is in the space defined by proportions of ancestry from different continental source populations. Another approach is in the space defined by allele frequencies in the admixed populations. These vantage points are connected, both evolutionarily and methodologically. From the perspective of evolution, a population receives its alleles from its ancestors. From the methodological perspective, we estimate ancestry from the alleles contained in samples from populations. The results of this study seem paradoxical in light of the fundamental connection between ancestry and allele frequencies in admixed populations. Notably, the plot of ancestry proportions in Figure 2.2 looks distinct from the plot of genetic distance coordinates in Figure 2.3. The triangle plot of ancestry fractions in Figure 2.2 fills a two-dimensional space, whereas ancestry correlates with only one major axis of allele frequencies summarized as genetic distances. We can graphically resolve the two plots by projecting the apex of the triangle, which represents European ancestry, onto the axis between African and Indigenous American ancestry. We can also resolve the apparent disparity analytically and evolutionarily.

The ancestry component of genetic distance (Fig. 2.1 and Eq. 2.4) is complicated by the fact that differences in ancestry between the admixed populations are modulated by genetic distances between the sources. In other words, the degree to which differences in ancestry contribute to the genetic differentiation of mixed populations depends on the levels of differentiation among the ancestral source populations (Cavalli-Sforza et al., 1994). The single axis of ancestry that we see in the principal coordinate plot reflects the recent evolution of human diversity. Genetic differentiation on the intercontinental scale has been driven by a series of founder effects (Ramachandran et al., 2005; Hunley et al., 2009). The entire species traces back to a population that lived in Africa approximately 200,000 years ago. A founder effect led to the habitation of Eurasia more recently, less than 100,000

years ago. The peopling of the Americas resulted from a subsequent founder effect from a population residing in Eurasia. A consequence of this history is that populations living in Africa have the greatest diversity, in terms of both the kinds of alleles and heterozygosity. Eurasian populations have a subset of the allelic types found in Africans and lower heterozygosity. Indigenous American populations have a subset of the allelic types found in Eurasians and lower yet heterozygosity (Li et al., 2008; Long et al., 2009). Ultimately, loss of variation via the founder effects created a single trajectory of genetic distances among populations on different continents. The ancestry of admixed populations will determine their placement on the axis, but it cannot introduce new axes of variation.

A full account of the genetic structure of admixed populations requires us to look at the effects of genetic drift, in addition to admixture. The principal coordinates of the drift distance matrix display two principal findings for the 13 Latin American populations. First, these Latin American populations have drifted independently. There is no evidence for a concerted pattern that a phylogenetic radiation from a single founding event would produce. It is likely that Latin American populations were founded independently by admixture in many locations. The proportions of continental ancestry differed among the populations. In a few populations, such as the Peque, Paposo, and Quetalmahue, high values of F_{ST} indicate that modest founder effects after, or during, the formation of populations (Table 2.3). These founder effects superimposed a new level of genetic structure on that created by admixture. Second, the drift and ancestry fractions contribute about equally to the pattern of genetic differentiation among the Latin Americans (Fig. 2.5). We do not observe a correlation between genetic distance and the impact of drift. Drift may predominate the distance between either closely related, or distantly related, populations. These two findings lend further support to the position expressed by Tishkoff and Kidd (2004) that anthropologists and geneticists cannot conceive of Latin Americans as a homogeneous genetic population. Our analysis shows that the genetic structure of Latin Americans involves more than varying proportions of continental ancestry.

Drift accounts for over 90% of the genetic distance among the four African-American populations. However, it should be noted that small differences characterize the populations analyzed here. Broader coverage of African-American populations, perhaps including the Gullah of South Carolina (Parra et al., 2001), and African-Americans living on the West Coast (Reed, 1969), could increase genetic distances and show instances where both admixture and drift drive patterns of differentiation.

Genetic drift plays a less dominant role in the genetic differentiation between African-American populations and Latin American populations (Fig. 2.5). A large role for continental ancestry is unsurprising because all of the 13 Latin American populations have below 10% African ancestry, while the African-American populations have above 75% African ancestry. However, it is notable that genetic drift accounts for up to a third of the distances between the most divergent populations.

In conclusion, this research introduces a new method to assess genetic diversity in admixed populations. Specifically, we show how to partition the minimum genetic distance between a pair of admixed populations into two components, one related to differences in continental ancestry and the other related to genetic drift in the admixed population. This partition allows greater precision in identifying how the recent evolutionary process has shaped modern human diversity. Our work paves the way for future investigations of geographic regions such as the Caribbean where many populations were formed by a complex combination of admixture and founder effects.

Chapter 3

Identifying the Number of Source Populations and Their Identities in Genetic Ancestry Analyses

---

3.1 Overview

Objective: We investigate the ancestry of a mixed population whose ancestry is uncertain. We propose multiple ancestry models that differ in the number of populations that contribute to the mixed population, and use the Akaike Information Criterion to choose the best model. Our focal admixed population is the Cape Coloured of South Africa. The Cape Coloured exemplify the challenges associated with estimating genetic ancestry.

Materials and Methods: We provide a history of South Africa to describe the development of the Cape Coloured. We analyzed the genotypes of 207 individuals from 11 contemporary populations at 618 autosomal microsatellite loci. Using maximum likelihood, we estimate allele frequencies for the ancestral sources, ancestry proportions and expected

allele frequencies among the Cape Coloured. We construct 26 models, ranging from two to five ancestral sources, and use AIC to determine the best fitting model.

Results: The ancestry estimates of the Cape Coloured fluctuate based on which ancestral sources are included in each model. AIC indicates the best fitting model consists of two ancestral sources, the San and East Asians, and estimated 9,712 parameters. The fit for each model decreased as the number of parameters increased. All models have high R^2 values for the observed and predicted Cape Coloured allele frequencies, ranging from 0.930 to 0.951.

Discussion: We demonstrate the utility of AIC in multi-model hypothesis testing for admixture research. Our analyses support the concept of parsimony. The best fitting models have a minimal number of parameters, and contain two ancestral sources.

3.2 Introduction

A goal in admixture analyses is to estimate the contributions of ancestors to admixed individuals and populations. This is typically achieved by constructing allele frequencies in a mixed sample as a linear combination of allele frequencies in populations that contributed ancestors to the mixed sample. Estimating the allele frequencies of the ancestors is a challenging problem because the true sources of ancestry may no longer exist, or may not have been genetically sampled, or are otherwise unavailable for study. Tang et al. (2005) recommend a solution to this problem which consists of constructing pseudo-ancestors, who are descendants from close relatives of the true ancestors. However, for populations such as the Cape Coloured, it can be difficult to choose pseudo-ancestors because the number and identities of the true ancestral sources is unknown.

Here we describe an approach to investigate the ancestry of a contemporary mixed population when there is uncertainty about the sources of ancestry. In this approach, we propose multiple models that differ in the number of populations that contribute ancestry to the

mixed population and we use the Akaike Infromation Criterion (AIC) to choose the best model. Our focal admixed population to which we apply the AIC is the Cape Coloured of South Africa. Coloured is a nationally recognized ethnic group in South Africa. The Cape Coloured exemplify the challenges in analysis of genetic ancestry in an admixed population. Population geneticists have designated the Cape Coloured as a population of mixed ancestry (Tishkoff et al., 2009), and have also classified them as Afro-Europeans of mixed ancestry (Pemberton et al., 2013). However, Cape Coloured history suggests that such labels are too simple because the Cape Coloured people are likely to have ancestors from as many as five ethno-geographic populations, including non-Africans and non-Europeans. Two recent studies of Cape Coloured ancestry have postulated different ethno-geographic sources of ancestry, and, as should be expected, produced differing results (Patterson et al., 2010; de Wit et al., 2010).

Research design and statistical methods play an important role in the identification of sources of ancestry. All designs and methods have a similar recognition of the population genetic process. Allele frequencies in an admixed individual, or population, are a linear combination of allele frequencies in the ancestral sources, and the coefficients of the linear combination represent ancestry fractions. Two distinct strategies emerge from this common starting point.

Strategy #1: First, assemble a meta-sample that combines individuals from a focal mixed population with samples from regional populations throughout the world. Second, fit cluster models that treat the meta-sample as a mixture of a predefined number of ancestral source populations. Third, run a sequence of cluster analyses that increase the number of ancestral sources for the meta-sample, until the regional populations appear as having approximately homogeneous ancestry, while the individuals from the focal mixed population have varying degrees of ancestry from the ancestral sources (Pritchard et al., 2000; de Wit et al., 2010).

Strategy #2: First, predetermine the sources of ancestry for the admixed population

through ethno-historical analysis, or a screening method such as principal components. Second, apply a regression-like method to estimate the allele frequencies in the ancestral source populations and the ancestry fractions in the mixed sample (Patterson et al., 2010).

Both of the above strategies run a risk of producing over-determined ancestry models for populations with complex histories. We show that the AIC can guide researchers to models that fit the data well and are parsimonious. We conclude the paper with recommendations for sampling designs for ancestry analyses when there is uncertainty about the sources of ancestry.

3.3 Founding of the Cape Coloured People

The Cape Coloured people formed by the genetic mixing of individuals from African indigenous populations with European colonists and their slaves, and Asian people who migrated to South Africa. Figure 3.1 presents a summary of the major dates of the Cape Colony that impacted the founding of the Cape Coloured people.

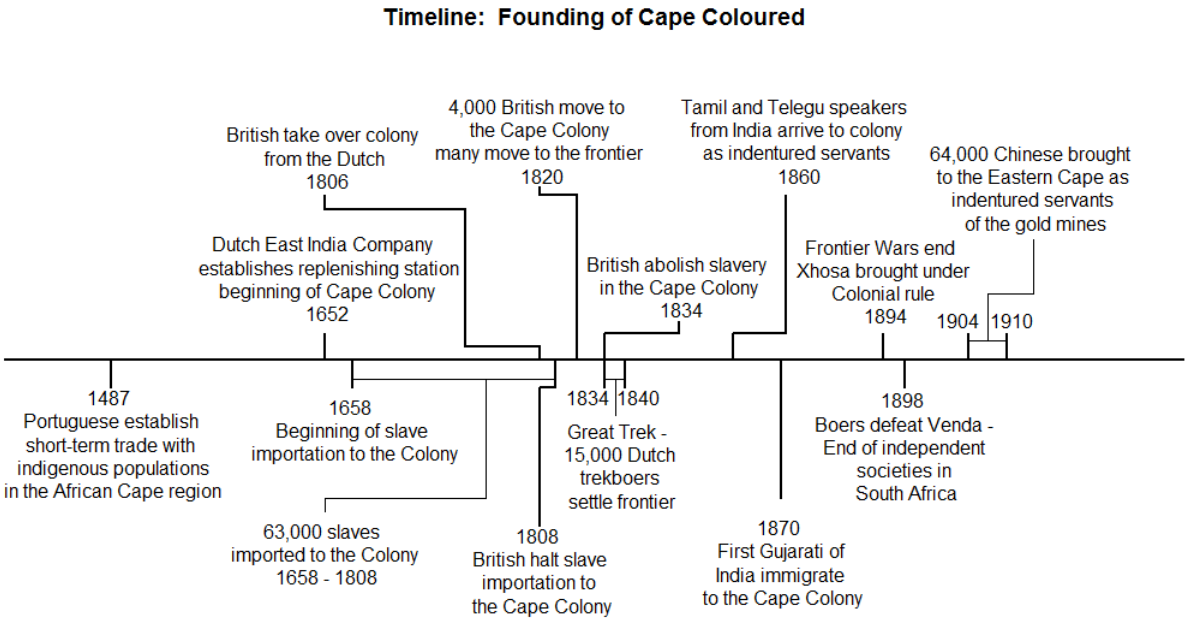

Figure 3.1: Timeline of the major historical events of the Cape Colony that contributed to the formation of the Cape Coloured people.

Three prominent indigenous groups of Southern Africa that may be involved in the formation of the Cape Coloured (Thompson, 2014). The first are the San foragers. Historically, the San were a heterogeneous group that spoke a variety of languages (Elphick, 1977). The second was the Khoekhoe who practiced pastoralism, and spoke a single language throughout Southern Africa (Elphick, 1977; Thompson, 2014). Researchers believe the Khoekhoe and San are distant genetic relatives who share similar languages; they are often collectively referred to as Khoesan (Elphick, 1977; Meyer, 2014). A third potential source are Bantu-speaking populations found throughout Southern Africa (Thompson, 2014). These Bantu speakers practiced pastoralism, and agriculture (Thompson, 2014). The indigenous Khoesan and Bantu speaking peoples potentially intermixed prior to the entry of Europeans to the region (Elphick, 1977; Meyer, 2014; Thompson, 2014).

European contact with South Africa began when the Portuguese arrived at Mossel Bay in 1487 (Elphick, 1977; de Villiers, 2014a; Thompson, 2014). At this early stage, Europeans undertook short-term trade with the indigenous groups to resupply their ships and continue to further East. The early Portuguese visitors never established a permanent colony (Elphick, 1977). In 1652, the Dutch East India Company (VOC) sent 80 employees to Cape Town, South Africa to establish a replenishing station for ships traveling to and from India and Southeast Asia (Davenport and Saunders, 2000; de Villiers, 2014a; Thompson, 2014). The VOC paid passage for Dutch men and French Huguenots, to assist with the way station (Keegan, 1996; de Villiers, 2014a).

European colonists imported 63,000 slaves to the Cape colony between 1658 and 1808, 26.3% were from Africa, 25.1% were from Madagascar, 25.9% were from India, and 22.7% were from Indonesia (Shell, 2014). The slaves of local origin often had mixed ancestry (Keegan, 1996). European admixture began when European men fathered a high proportion of children born to slave women (Keegan, 1996). The British took over the colony in 1806 (Davenport and Saunders, 2000). In the early 1800s, British Parliament passed several laws, which improved the conditions for slaves and labor classes in the colony. In 1808,

the British stopped the importation of slaves to the colony (Davenport and Saunders, 2000; Keegan, 1996; Shell, 2014; Thompson, 2014). In 1828, the British passed a law giving free blacks and indigenous groups equal standing to whites in the colony, and released them from enforced labor (Thompson, 2014; Visagie, 2014a). The British abolished slavery in the Cape colony in 1834 (Davenport and Saunders, 2000; de Villiers, 2014b; Thompson, 2014).

The VOC granted some employees free burgher status these people played a pivotal role in the formation of the Cape Coloured People. The free burghers became known as the Boers. The Boers are the ancestors of contemporary Afrikaners (Davenport and Saunders, 2000; Keegan, 1996; Thompson, 2014; Visagie, 2014a). As the number of free burghers increased they moved ever further into the frontier. Trekboers encountered the Bantu-speaking Xhosa in 1813 (Keegan, 1996). These interactions set into motion a series of nine Frontier wars with the Xhosa that lasted for nearly a century (Davenport and Saunders, 2000; Thompson, 2014). These wars were perpetuated by increasing tensions due to colonial encroachment along the frontier. In 1820, an additional 4,000 British moved to the Cape colony, many of whom moved to the frontier (de Villiers, 2014b; Thompson, 2014). The largest migration of colonists to move to the frontier occurred during the Great Trek (Visagie, 2014b). The Great Trek consisted of over 15,000 trekboers who, due to grievances with the colony, left between 1834 and 1840 to settle the frontier as Voortrekkers (Davenport and Saunders, 2000; Visagie, 2014b). Finally, in 1894, the last of the Frontier wars brought the Xhosa under colonial control (Grobler, 2014). The Bantu-speaking Venda was affected by Voortrekker expansion beginning in 1848 (Grobler, 2014). The Voortrekker and Venda interactions were friendly for a long period, until 1898, when the Boers attacked and defeated the Venda, marking the end of independent indigenous societies in South Africa (Grobler, 2014).

Dutch colonists originally brought about 16,300 slaves from Bengal and southern India to the Cape between 1657 and 1808 (Shell, 2014; Vahed, 2014). The majority of Indi-

ans in South Africa today descend from nearly 150,000 Indians who migrated between 1860 and 1911 (Vahed, 2014). The first Indians who migrated during this period came to South Africa as indentured laborers and spoke Tamil and Telegu of South India (Vahed, 2014). Later, in the 1870s, many Gujarati immigrated independently to South Africa (Vahed, 2014). The Chinese had a minimal presence in South Africa before 1900, but between 1904 and 1910, nearly 64,000 Chinese served as indentured laborers in the gold mines of the Eastern Cape (Joubert, 2014). The South African government, during the apartheid era, annexed the Chinese to a subgroup of the Coloured class (Joubert, 2014). Indonesians, numbering 14,300, were brought to the Cape Colony as slaves between 1652 and 1808 (Shell, 2014). Indonesians are a population that feature prominently in South Africa to this day.

Class structure and demography provided ample potential for genetic mixing to occur in the Cape colony. Our review of the history of the Cape Coloured population shows many mixtures and proportions were possible, but the history does not provide all the details and combinations that did in fact occur. Early on, the majority of colonists were men, who sought partners outside their cultural and class groups. This practice was more prominent along the frontier, which formed new cultural groups derived from a mix of indigenous and European ancestry (Davenport and Saunders, 2000; Keegan, 1996; Thomas, 2014; Thompson, 2014; Visagie, 2014b). As early as the 1850s mixed ancestral groups began to identify as Afrikanders (Thomas, 2014). By 1880, they came to be known as the Coloured people, which is the current identification of people derived from mixed ancestry in South Africa (Thomas, 2014).

3.4 Materials and Methods

We employ the maximum likelihood approach for admixture analysis devised by Tang et al. (2005). This method assumes a population model in which the ancestry of a mixed pop-

ulation derives from a specified number of ancestral sources. The genotype data consist of samples of individuals from the admixed group, as well as from contemporary populations who serve as proxies for the ancestors of the admixed group. Tang et al. (2005) identify these proxies as pseudo-ancestors. Ideally, the pseudo-ancestors have descended from close relatives of the true ancestors, but for populations such as the Cape Coloured, the appropriate samples to serve as pseudo-ancestors are uncertain.

We write the likelihood equation as

$$lnL(\theta) = \sum_{s=1}^{S} \sum_{i=1}^{N_S} \sum_{l=1}^{L} \sum_{j=1}^{J_l} [g_{silj} \times ln(y_{silj})] \tag{3.1}$$

where

$$y_{silj} = \sum_{k=1}^{K} p_{jlk} m_{ik} \tag{3.2}$$

is the predicted allele for the j^{th} allele at the l^{th} locus in the i^{th} individual in the s^{th} sample.

The data g_{silj} are the counts of the j^{th} allele ($j = 1...J_l$), observed at the l^{th} locus ($l = 1...L$) from the i^{th} person ($i...N_s$), belonging to the s^{th} sample ($s = 1...S$). The parameters ($\theta = [\mathbf{p}, \mathbf{m}]$) are p_{jlk} the frequency of the j^{th} allele, from the l^{th} locus, from the k^{th} source population ($k = 1...K$), and m_{ik} the fraction of ancestry from the k^{th} source population contributed to the i^{th} individual.

The likelihood function is of extremely high dimension when genomic scale data are used. Maximizing this function requires estimating thousands of parameters, consisting of allele frequencies and ancestry fractions. Our program uses the Expectation Maximization algorithm described by Tang et al. (2005) as a numerical method to obtain asymptotic results from the likelihood equation. We follow the recommendation of Alexander and colleagues (2009) and use a strict convergence criterion of 10^{-6} to ensure convergence to the maximum. We used the Bloodshed Development Environment (http://www.bloodshed.net)

in the C++ language to write new software for the method to accommodate microsatellite data. Other implementations of this likelihood method are restricted to single nucleotide polymorphism data (Alexander et al., 2009; Tang et al., 2005).

The model in Equation 3.1 is implicitly a regression analysis because it constructs the admixed allele frequencies as a linear combination of source population allele frequencies. Plotting the actual allele frequencies in the admixed population against the predictions from the admixture model is a simple way to assess the goodness of fit of the admixture model. R^2 serves as a familiar way to describe and compare such plots.

Correlation in allele frequency among pseudo-ancestors is an important factor in fitting the admixture model. We assess the extent of such correlations by calculating Pearson correlation coefficients of allele frequencies within and between all pairs of pseudo-ancestors. While other statistics such as F_{ST} typically serve this purpose in population structure analysis, we feel that the Pearson correlation coefficient is appropriate here because of the connection between admixture analysis and linear regression.

Based on our historical outline of the Cape Coloured, previously published genetic analyses, and available samples (de Wit et al., 2010; Patterson et al., 2010; Pemberton et al., 2013), we consider five populations as potential sources of ancestry for the Cape Coloured people. The five source populations are Khoesan, Bantu speaking peoples of South Africa, European, South Asian, and East Asian. Each of the five potential source populations is represented by two samples (Table 3.1). Each sample consists of individuals from which we analyzed genotypes at 618 autosomal microsatellite loci. The genotyping service at the Marshfield Clinic performed the laboratory analyses for the original studies. The loci were selected for linkage mapping and are spaced on the genetic map approximately 5 cM to 10 cM apart. We use the data set that Pemberton et al. (2013) created by calibrating allele sizes across three original studies. The original studies from which these samples came are listed in Table 3.1. We developed 26 models, in total, as hypotheses for the ancestry of the Cape Coloured population 3.2. The models consist of all combinations of two, three, four,

and five of the potential sources for the ancestors of the Cape Coloured people.

Table 3.1: Populations used in our analyses, samples obtained from Pemberton et al. (2013).

Population Name	Sample Size	Ancestral Source	Primary Reference
Cape Coloured	33	Mixed Ancestry	Tishkoff et al. (2009)
San	7	Khoesan	Cann et al. (2002); Rosenberg et al. (2002)
!Xun Kxoe	6	Khoesan	Tishkoff et al. (2009)
Xhosa	27	Bantu	Tishkoff et al. (2009)
Venda	11	Bantu	Tishkoff et al. (2009)
French	29	European	Cann et al. (2002); Rosenberg et al. (2002)
Orcadian	16	European	Cann et al. (2002); Rosenberg et al. (2002)
Hindi	28	S. Asian	Cann et al. (2002); Rosenberg et al. (2006)
Tamil	29	S. Asian	Cann et al. (2002); Rosenberg et al. (2006)
Han North China	10	E. Asian	Cann et al. (2002); Rosenberg et al. (2002)
Cambodian	11	E. Asian	Cann et al. (2002); Rosenberg et al. (2002)

We introduce a novel research design here. First, we start with data from individuals in a focal mixed population, and samples from a set of regional populations throughout the world. The regional samples will serve as pseudo-ancestors for sources of ancestry in the mixed sample. Second, we postulate a series of ancestry models that specify a pre-defined number of ancestral sources for the mixed sample. We construct a separate data set for each model that contains data from only the mixed sample and the appropriate pseudo-ancestors for the ancestral sources in the model. For example, we would not include samples from Europe, East Asia, and Africa, if our model specifies only two sources of ancestry, because three continental populations would imply three ancestry sources. Third, we fit a cluster model with the appropriate number of sources for the data set. Fourth, we test the fit of each cluster model to the mixed sample by comparing the observed allele frequencies to the allele frequencies predicted from inferred source populations. Fifth, we use the Akaike Information Criterion (AIC) to decide on which ancestry model is the most appropriate representation of the data amongst the 26 models evaluated. We provide the equation for AIC below.

$$AIC = -2lnL(\hat{\theta}) + 2P \tag{3.3}$$

The AIC is a measure of information contained within a fitted model, and is calculated from the natural log likelihood penalized by the number of parameters estimated in the model (Akaike, 1973, 1974, 1981a,b, 1983; Anderson, 2008; Burnham et al., 2011). The best ranking model has the lowest AIC value. We order the AIC values for the 26 models from lowest to highest, and then calculate the difference between each model and the model with the lowest AIC.

$$\Delta_i = AIC_i - AIC_{min} \tag{3.4}$$

where, AIC_i is the Akaike Information Criterion for the i^{th} model, and AIC_{min} is the minimum AIC. From the Δ values, we assess the level of support of each mode in relation to the best-supported model following established guidelines. A Δ value less than two means that a model carries as much information as the highest-ranking model (Burnham et al., 2011). If a Δ value ranges between nine and 11, then it provides low support for the highest-ranking model relative to the alternative. A model carries no additional support relative to the highest-ranking model if the Δ value is greater than 20 (Burnham et al., 2011).

Model	Khoesan	Bantu	European	S. Asian	E. Asian	Allele Fr.	Ancestry	Total
1	■	■				9712	33	9745
2	■		■			9522	33	9555
3	■			■		9766	33	9799
4	■				■	9244	33	9277
5		■	■			9866	33	9899
6		■		■		10076	33	10109
7		■			■	9668	33	9701
8			■	■		9670	33	9703
9			■		■	9278	33	9311
10				■	■	9514	33	9547
11	■	■	■			15303	66	15369
12	■	■		■		15576	66	15642
13	■	■			■	15033	66	15099
14	■		■	■		15192	66	15258
15	■		■		■	14679	66	14745
16	■			■	■	14979	66	15045
17		■	■	■		15594	66	15660
18		■	■		■	15147	66	15213
19		■		■	■	15408	66	15474
20			■	■	■	14820	66	14886
21	■	■	■	■		21368	99	21467
22	■	■	■		■	20824	99	20923
23	■	■		■	■	21128	99	21227
24	■		■	■	■	20628	99	20727
25		■	■	■	■	21120	99	21219
26	■	■	■	■	■	27090	132	27222

Figure 3.2: Twenty-six models, which serve as hypotheses in testing ancestry among the Cape Coloured population of South Africa. The number of allele frequency and ancestry fraction parameters estimated per model are shown. Red denotes the ancestral source populations included in each model. White denotes the ancestral source populations omitted from each model.

3.5 Results

Table 3.2 provides the log likelihood, R^2, AIC, Δ, and ancestry coefficients for our 26 models proposed for the ancestry of the Cape Coloured population. The log likelihood, AIC, and Δ values are reported from the portion of the analysis involving only the Cape Coloured population because we are primarily concerned with the evaluation of the model

in regards to the admixed group. If we consider these statistics for the entire model, then we may be evaluating the models in terms of their fit to their pseudo-ancestors. We will return to the concept of pseudo-ancestors in the Discussion section.

Table 3.2: Model rankings for the putative ancestry for the Cape Coloured people, which includes all possible models with two or more, and as many as five ancestral populations. [$1 < ancestralpopulations \leq 5$]

Rank	Khoesan	Bantu	European	S. Asian	E. Asian	lnL	AIC	Δ	$Par.$	R^2
1	0.447	-	-	-	0.553	-62629	143812	0	9277	0.951
2	0.756	0.244	-	-	-	-62273	144036	224	9745	0.949
3	-	-	0.243	-	0.757	-62850	144322	510	9311	0.943
4	0.479	-	0.512	-	-	-62622	144354	542	9555	0.948
5	-	-	-	0.219	0.781	-62857	144808	996	9547	0.946
6	-	0.531	-	-	0.469	-62716	144834	1022	9701	0.948
7	0.427	-	-	0.573	-	-62778	145154	1342	9799	0.948
8	-	0.531	0.469	-	-	-62790	145378	1566	9899	0.946
9	-	0.498	-	0.502	-	-62834	145886	2074	10109	0.951
10	-	-	0.834	0.166	-	-64040	147486	3674	9703	0.930
11	0.409	-	0.342	-	0.249	-62572	154634	10822	14745	0.950
12	0.088	0.447	-	-	0.465	-62580	155358	11546	15099	0.947
13	-	-	0.186	0.103	0.711	-62812	155396	11584	14886	0.943
14	0.407	-	-	0.418	0.175	-62731	155552	11740	15045	0.948
15	-	0.474	0.286	-	0.240	-62566	155558	11746	15213	0.947
16	0.423	-	0.221	0.356	-	-62641	155798	11986	15258	0.950
17	0.136	0.404	0.461	-	-	-62609	155956	12144	15369	0.944
18	-	0.469	-	0.338	0.193	-62683	156314	12502	15474	0.949
19	0.124	0.374	-	0.502	-	-62628	156540	12728	15642	0.947
20	-	0.495	0.199	0.306	-	-62713	156746	12934	15660	0.951
21	0.396	-	0.227	0.208	0.168	-62571	166596	22784	20727	0.948
22	0.101	0.383	0.301	-	0.216	-62409	166664	22852	20923	0.949
23	0.107	0.371	-	0.362	0.16	-62551	167556	23744	21227	0.947
24	-	0.474	0.191	0.176	0.159	-62597	167632	23820	21219	0.950
25	0.117	0.378	0.194	0.311	-	-62516	167966	24154	21467	0.948
26	0.100	0.379	0.198	0.176	0.147	-62430	179304	35492	27222	0.951

The R^2 values between expected (model-based) and observed allele freqquencies are high for all models. They fall into a narrow range, from 0.930 to 0.951. Figure 3.3 presents the scatter plots from three models, which include two models with the highest R^2 of 0.951, and the model with the lowest R^2 = 0.930. In these plots the predicted allele frequencies are plotted along the abscissa, and the observed allele frequencies are plotted on the ordinate. The predictions in Figures 3.3a and 3.3b are made from two ancestral source populations.

The predictions in Figure 3.3c derive from five ancestral source populations. The broader distribution coupled with a few highly dispersed points explains the lower R^2 value (fig. 3.3b). The R^2 values do not provide the resolution to discriminate between models.

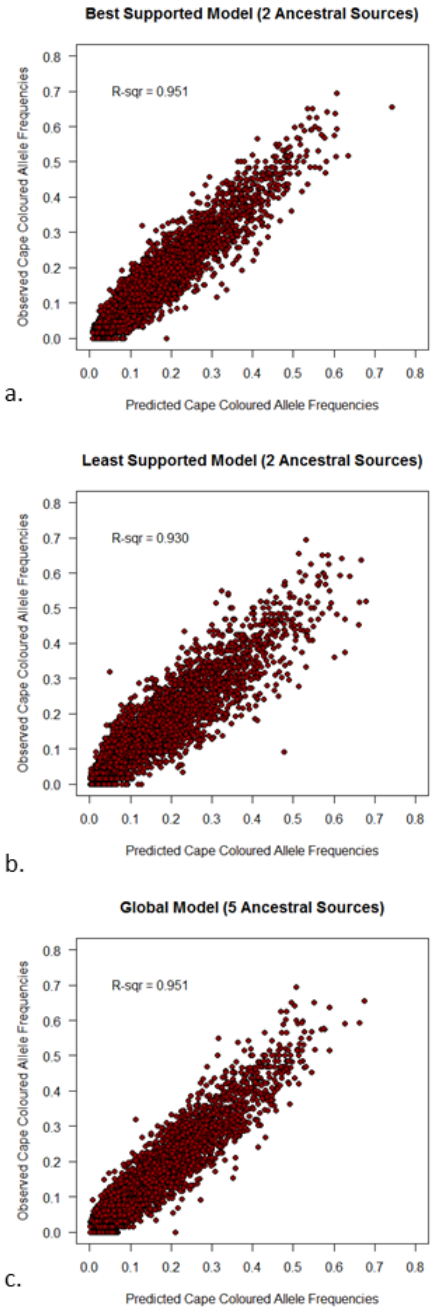

Figure 3.3: Three scatter plots and their associated R^2 values. a) The scatter plot contains the Khoesan and East Asians as the pseudo-ancestral populations and has an R^2 of 0.951. b) The scatter plot for the model with the lowest R^2 of 0.930, which contains Europe and South Asia as proxies for two ancestral source populations. c) The scatter plot containing all five ancestral source populations, and has an R^2 of 0.951.

These high R^2 values are explained by multicollinearity, which is a common phenomenon in regression analyses. Multicollinearity occurs when the predictor variables

in a regression model are correlated (Kutner et al., 2005). In our case, the allele frequencies of the ancestral source populations serve as the predictor variables in the regression model. Table 3.3 shows the Pearson correlations of the observed allele frequencies within and among the five ancestral source populations constructed in our analyses. We see that a correlation exists between each pair of ancestral sources, and high correlation is present among the non-African ancestral source populations.

Table 3.3: The correlation values of the observed allele frequencies among the pseudo-ancestral sources.

	Khoesan	Bantu	European	S.Asian	E.Asian
Khoesan	1	0.732	0.584	0.609	0.555
Bantu	0.732	1	0.697	0.721	0.661
European	0.584	0.697	1	0.896	0.785
S.Asian	0.609	0.721	0.896	1	0.844
E.Asian	0.555	0.661	0.785	0.844	1

By contrast to R^2, the Δ values discriminate among the 26 models. The log likelihood is the first factor of AIC in determining the rank of each model in the proposed set of models. The log likelihood values fall into a narrow range, from -62,273 to -64,040. It is notable that the model with the lowest R^2 also had the lowest log likelihood. The number of parameters estimated for a model is the second factor in determining the model's AIC. The AIC accounts for the number of parameters to serve as a penalty to reduce the risk of a model overfitting the data (Anderson, 2008; Burnham et al., 2011). The number of parameters range widely, from 9,277 to 27,222. In examining Table 3.2, we see that model rank falls out according to the number of ancestral source populations. All two ancestral source models rank higher than three ancestral source models. All models containing three ancestral source populations rank higher than the four ancestral source models. All other models rank higher than the global model, which contains five ancestral source populations. Recall that a Δ value greater than 20 carries no additional support for the highest-ranking model (Burnham et al., 2011). The next closest Δ value to the highest-ranking model in our model set is 224. Therefore, none of the models provide additional information to the

highest-ranking model.

In addition, Table 3.2 includes the ancestry estimated for each source in the models. The ancestry estimated from a source varies highly based on the other sources of ancestry in a model. Figure 3.4 shows the distribution of ancestry fraction estimates among four of the ancestral source populations: Khoesan, Bantu, Europe, and East Asia. Each ancestral source appears in 15 of the 26 models analyzed. Thus, the frequency (ordinate) of each ancestral source sums to 15 (fig.3.4). For any source, the wide distribution of ancestry estimates is alarming because all models have a similar goodness of fit for the predicted allele frequencies. The Khoesan ancestry estimates vary widely, and range from 0.088 to 0.756. The ancestry fraction estimates of the Bantu are the most centralized, however, they still range from 0.244 to 0.531.

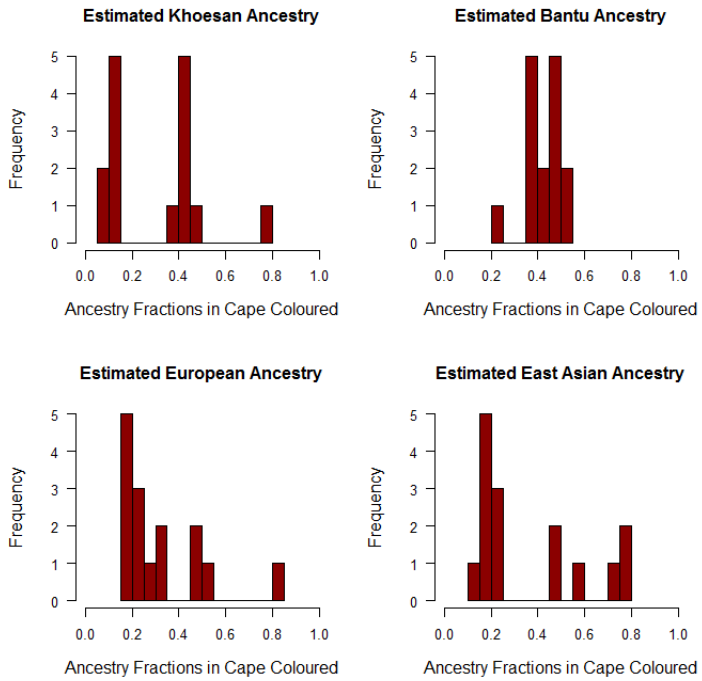

Figure 3.4: The distribution of ancestry fraction estimates among the source regions, across all models. Each ancestral source was used in 15 of the 26 models.

3.6 Discussion

All 26 of our proposed models fit the Cape Coloured data well (Table 3.2). The R^2 values fell into the narrow range $[0.930 \leq R^2 \leq 0.951]$. Moreover, the four different models shared the highest R^2 value. Despite the uniformly high R^2 values, the ancestry coefficients were inconsistent. For example, the estimated fraction of Khoesan ancestry varied between 10.0% and 75.6% among the 15 models that postulated Khoesan as a source of ancestry (Figure 3.4). By contrast to R^2, the AIC criterion clearly separates the 26 models. There is one unequivocal best model. This model is surprisingly sparse, consisting of only two sources of ancestry, represented by Khoesan and East Asians as pseudo-ancestors. It is clear that AIC separates these models by the number of parameters they estimate (Table 3.2). The model most favored by AIC required the estimation of 9,277 parameters while the model least favored by AIC required the estimation of 27,222 parameters. Ironically, these two models were tied for $R^2 = 0.951$.

It is important to resolve why models with such different ancestral populations do equally well in predicting allele frequencies in the Cape Coloured. The solution to this problem comes from recognizing that genetic ancestry analysis is a generalized regression problem (see Equation 3.2) where allele frequencies in the ancestral sources serve as the predictor variables, and the allele frequencies in the admixed sample serve as the response variables. In the case of the human species, allele frequencies among the pseudo-ancestors are highly correlated, even among the most diverged populations such as Khoesan and East Asian. This phenomenon is due to the evolutionary history of our species, whereby genetic diversity was shaped by a series of founder effects (Ramachandran et al., 2005). Populations located in Africa have the highest levels of gene diversity, both in allelic types and heterozygosity in the world. Eurasian populations contain a subset of the alleles found in Africa, and have lower heterozygosity. Indigenous American populations contain a subset of alleles found in Eurasians with reduced heterozygosity. The evolutionary pattern of human diversity manifests statistically because the correlation of allele frequencies result in

the multicollinearity of regression models in admixture analyses.

We should point out that the ancestry fractions estimated from two earlier studies of the Cape Coloured do not agree well with each other (de Wit et al., 2010; Patterson et al., 2010). The results from our study do not match either of the previous two. The estimates of African ancestry (combined San and Bantu speaking peoples) illustrate this. de Wit et al. (2010) estimated 61.9% African ancestry, Patterson et al. (2010) estimated 36.9% African ancestry, and we estimate 47.9% African ancestry. It is difficult, if not impossible, to pinpoint what accounts for the differences between the three studies. For example, the studies differ in (1) the genetic markers analyzed, (2) the choice of populations to fill the role of pseudo-ancestors, (3) sample sizes for the Cape Coloured and the pseudo-ancestors, and (4) and the statistical methods employed for ancestry estimation. However, we believe that the issue of choosing pseudo-ancestors deserves some scrutiny. de Wit et al. (2010) use the first research design that we identified in the introduction. With four ancestral sources, samples from East Africa, Europeans living in the United States, and Melanesia serve as pseudo-ancestors that contribute to their estimated component of 'European' ancestry in the Cape Coloured. If the genes in these diverse populations represent the same source of ancestry, then this ancestor must have existed deep in the past, and would have little relevance to the recent founders of the Cape Coloured people. Similar to Patterson et al. (2010), we recommend using known history to guide the choice of pseudo-ancestors.

The key to identifying sources of ancestry lies in the formulation of AIC, Equation 3.3. The ultimate goal should be to design models that will increase the negative log likelihood without a concomitant increase in the number of parameters. Model discrimination will be achieved by increasing the number of individuals sampled for a fixed number of genetic markers, and it will be thwarted by increasing the number of genetic markers for a fixed number of individuals. The challenge for future studies will be to find the appropriate balance between the correct number of markers and individuals. Large numbers of individuals for both the admixed sample and the pseudo-ancestors will be desirable, but large

samples may be impossible to collect given the accessibility to the groups in question. One possibility to manage the number of markers for a given sample size is to use Ancestry Informative Markers ($AIMs$). However, this strategy has two limitations. First, the choice of which $AIMs$ to use requires some knowledge of the true sources of ancestry, which is the main question for groups such as the Cape Coloured. Second, there may not be $AIMs$ that distinguish the sources of ancestry, depending on how closely they are related, $e.g.$, Europeans and South Asians.

Recent increases in our ability to collect molecular data and to evaluate computationally intensive models have taken studies of genetic ancestry and admixture to a new level. Our analyses of the Cape Coloured population of South Africa show that these advances have not removed the danger of producing ambiguous results. Methods such as AIC for measuring information, and for comparing competing models, are a useful addition to existing methods of model fitting and goodness-of-fit tests.

Chapter 4

Using Contemporary Populations as Pseudo-Ancestors to Estimate Ancestry Fractions

---

4.1 Overview

Objective: We investigate bias associated with using maximum likelihood to estimate genetic ancestry proportions. We use coalescent simulation to simulate eight models of a single admixture event. Models one through four simulate admixture of an African-American population. Models five through eight simulate admixture of a Latin American population. **Materials and Methods:** We recapitulate the gene identities of African, European, and Indigenous American populations in these simulations. For each model, we vary sample sizes among the admixed population and proxies for their ancestral sources at (i) 100 individuals for each population; (ii) 100 admixed individuals and 20 individuals for each source; and (iii) 20 admixed individuals and 100 individuals for each source. We make an-

cestry estimates directly from the populations that contributed to the admixture event. We also estimate ancestry from pseudo-ancestral sources. Pseudo-ancestors are related to the parental population but did not contribute to the admixed population. We assess the bias of ancestry estimates, by comparing the observed estimates to their parameter values.

Results: Our results show low bias among all models that estimate ancestry from the direct descendants in all sampling scenarios. We observe varying levels of bias under sampling scenario (i) when estimating ancestry from ancestral proxies. Bias exists when we estimate ancestry from sampling scenario (ii). Sampling scenario (iii) presents overestimation of minor ancestry contributions and an underestimation of major ancestry contributions across all models.

Discussion: These findings show that using pseudo-ancestors in admixture analyses causes biased ancestry estimates. To minimize biases in ancestry estimation, we recommend using large sample sizes that represent ancestral source populations.

4.2 Introduction

Admixture is a type of gene flow in which populations that have been isolated for a long period come into contact and exchange mates (Cavalli-Sforza and Bodmer, 1971; Cavalli-Sforza et al., 1994). A common goal in admixture analyses is to estimate the ancestral contributions of mixed populations, and estimate the timing of mixing events (Adams and Ward, 1973; Chakraborty, 1986).

The allele frequencies in an admixed population are a linear combination of allele frequencies in the ancestral source populations (Long, 1991). The coefficients of the linear combination reflect ancestry proportions in the mixed population contributed by the respective ancestral source populations. We employ the maximum likelihood method developed by Tang et al. (2005) to estimate the allele frequencies of ancestral source populations, and the fractions of ancestry in admixed populations using microsatellite genotype data from

contemporary populations. Tang et al. (2005) developed the concept of pseudo-ancestors. In this concept, contemporary genotype samples are used to estimate ancestry in an admixed population. Thus, the samples are consturcted as pseudo-ancestors. These samples are closely related to the ancestors that contributed to the founding of the mixed population (Tang et al., 2005).

In admixture analyses, the statistical method of maximum likelihood requires specific properties from the data. The sample sizes included in the analyses must be large. The samples must contain genotype data from both the admixed population, and its ancestral source populations. The evolutionary model is constrained in that admixture is assumed to be the only evolutionary force that has shaped allele frequencies in the mixed population. Here we investigate if failing to meet the requirements of these data biases ancestry estimation. A statistic is an unbiased estimator of a parameter when the long-term average of that statistic computed over replicate data sets is equal to the parameter (Hogg et al., 2013).

The factors we consider of the data used in maximum likelihood estimation of ancestry fractions include: (i) How large of a sample is considered sufficiently large? (ii) How estimates are affected if we sample from populations that are not the true ancestors. This consideration is pertinent because the true ancestral populations may no longer exist. In addition, a sparse historical record prevents us from fully knowing the true ancestral populations. (iii) How ancestry estimates are affected by genetic drift since the founding of the mixed population. Even if the ancestral populations still exist, allele frequency drift has occurred since the admixture event (Chakraborty, 1986).

We use the maximum likelihood method of Tang and colleagues (2005). We first estimate the gene identities of nine contemporary populations from genotype samples using 618 microsatellite loci. Next, we simulate a population tree to match patterns of genetic diversity from contemporary populations using FASTSIMCOAL. The simulated tree is constructed using genotype data composed of 500 microsatellite loci from contemporary genotype samples. The contemporary samples serve as pseudo-ancestors. We then simu-

late a single admixture event between two pseudo-ancestral groups from distinct regions. Finally, we estimate ancestry in the admixed populations from varying pseudo-ancestral contributors. In using simulated data, we know the relationships of the pseudo-ancestors to the true ancestors. From this knowledge, we are able to determine if genotype data from pseudo-ancestral populations in lieu of the true ancestors biases estimates of ancestry.

4.3 Materials and Methods

We developed eight genetic models using simulated data to estimate the ancestry fractions in admixed populations. The simulated data was generated using FASTSIMCOAL (Excoffier and Foll, 2011). We use FASTSIMCOAL to simualte parameters for (i) the dempgraphic history of the ancestral sources, (ii) the admixture event, and (iii) the ancestry fractions in the admixed population. Each model uses 500 microsatellite loci to estimate the ancestry fractions of an admixed population that formed from a single admixture event between two ancestral source populations. Models one through four simulate the formation of admixed African-American population. Models five through eight simulate the formation of an admixed Latin American population.

The first step in our analysis is to establish parameters to simulate data that resembles those of pseudo-ancestors that will be used in analyses of actual African-American and Latin American populations. To do this, we chose nine reference populations, three from Africa, three from Europe, and three from the Americas. The African populations we included are the Yoruba, Brong, and Mandenka. We chose these populations because they are all located in West Africa, and are related to the people who were brought to the Americas as slaves. The European populations we included are the French, Orcadians, and Italians. The European populations we chose are dispersed across central, southern and western Europe. The Indigenous Americans we included in the tree are the Pima of Mexico, Karitiana, and Guaymi. We chose these Indigenous American populations because they are

broadly dispersed throughout Central and South America, and they possess greater than 95% Indigenous American ancestry (Hunley and Healy, 2011).

We calculated a gene identity matrix from 618 loci for the nine actual population samples (Cann et al., 2002; Rosenberg et al., 2002; Wang et al., 2007; Tishkoff et al., 2009). We used the data set that combined these data from their original studies and were calibrated according to allele size (Pemberton et al., 2013). We used the neighbor joining method to estimate a tree from the observed gene identity matrix. We used generalized hierarchical modeling (GHM) to estimate branch lengths and root position within the tree (fig.4.1) (Long et al., 2009). We used a step-wise model in FASTSIMCOAL with a microsatellite mutation rate of 9×10^{-5} to find appropriate effective population sizes, and separation times that would recreate the observed tree, and its branch lengths.

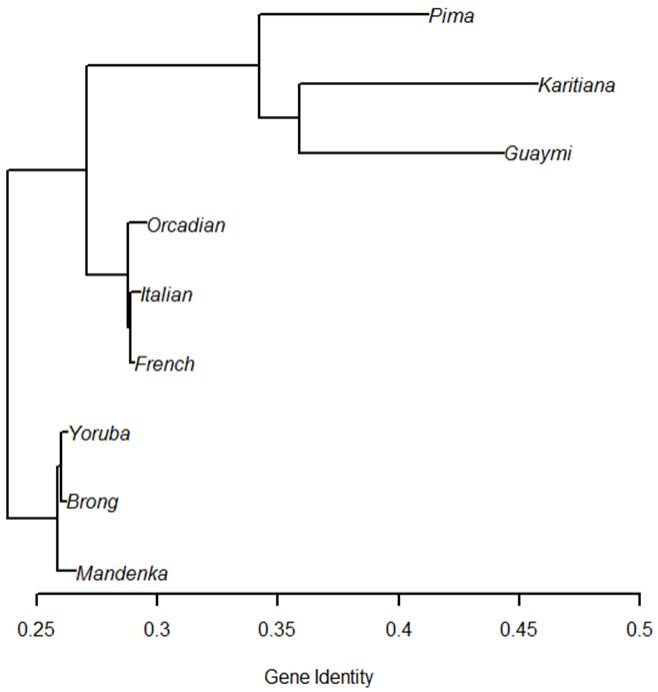

Figure 4.1: A population tree that serves as a reference for our simulations. Using FASTSIMCOAL, we simulate the population histories of this respective tree to recapitulate the gene identities in our simulated coalescent tree. We then simulated an admixture event from specific populations from this simulated tree to determine the potential bias associated with ancestry fraction estimates in varying sampling strategies.

We have written new software for the likelihood method developed by Tang et al. (2005)

to accommodate microsatellite data. We wrote this software using the Bloodshed Development Environment (http//www.bloodshed.net) in the C++ language. Other available implementations of Tang et al. (2005) method are restricted to single nucleotide polymorphism data (Alexander et al., 2009; Tang et al., 2005). The likelihood function is of extremely high dimension when applied to genomic scale data. Maximizing this function requires estimating thousands of parameters, consisting of allele frequencies and ancestry fractions. Our program uses the EM algorithm described by Tang and colleagues (2005) as a numerical method to obtain asymptotic results from the likelihood equation. Alexander and colleagues note that a stringent convergence criterion is necessary to obtain precise results (2009). In our simulation, we used a convergence criterion of 10^{-4}.

For each model, we examined ten levels of admixture with 100 replicate data sets each. The ancestry fraction estimates range from 0.05 to 0.95 in intervals of 0.10. We calculated the bias associated for each ancestry fraction estimate, which is simply the mean of 100 replicates for each ancestry estimate minus its parameter value. We repeated this process for each model so that each model consists of 10 ancestry fraction estimates, and the bias associated with each estimate.

In models one through four, "Simulated African 1" and "Simulated European 1" are descended from the true ancestral source populations that formed the admixed African-American population. In this instance "Simulated African 1" mimics the gene identity of the contemporary Yoruba, and "Simulated European 1" mirrors the gene identity of the contemporary French. The populations denoted the blue boxes represent the pseudo-ancestors from whom we make ancestry estimates. Model one estimates ancestry from the psuedo-ancestors who are the descendants of the true ancestral source populations (Figure 4.2, left). In model two, we estimate ancestry fractions from pseudo-ancestors that are closely related to the actual source contributors, which are "Simulated African 2" and "Simulated European 2" (Figure 4.2, right). These simulated populations mirror the gene identities of the Brong and Italians, respectively.

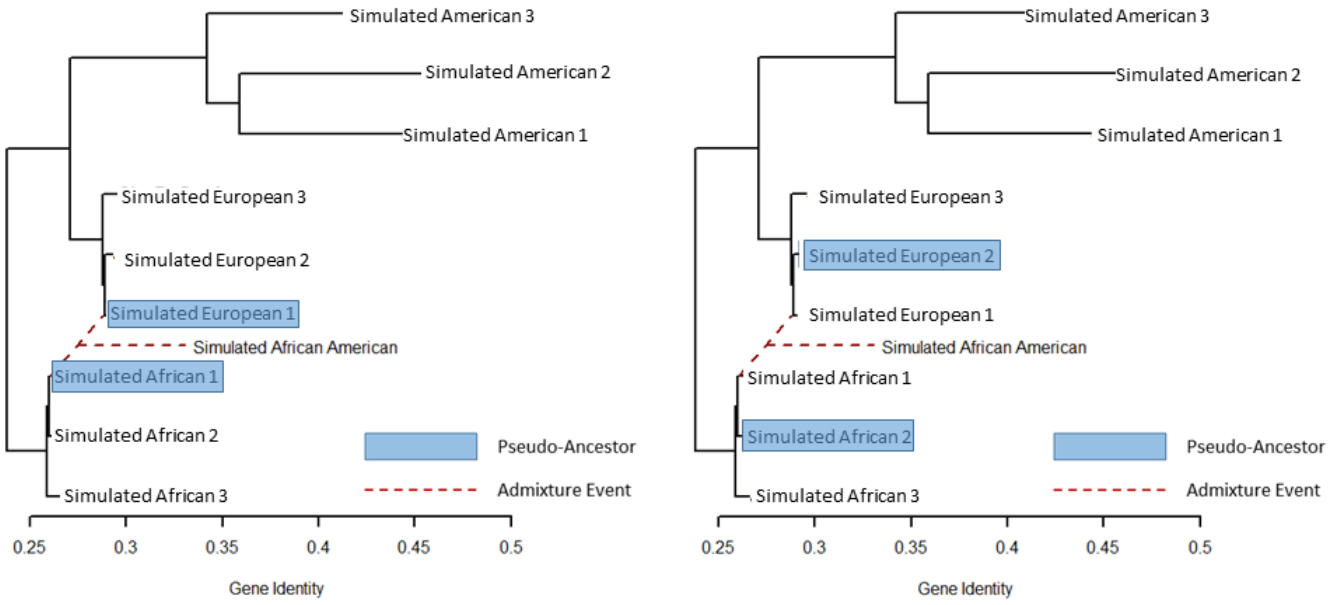

Figure 4.2: Model one (left) estimates ancestry fractions from pseudo-ancestors who are the descendants of the actual sources that form a simulated admixed African-American population. Model two (right) estimates ancestry fractions from closely related pseudo-ancestral sources that formed a simulated admixed African-American population.

In model three, we estimate ancestry fractions from more distantly related populations from the actual ancestry source contributor in each region (Figure 4.3, left). Therefore, we will estimate ancestry fractions of the admixed population from "Simulated African 3" and "Simulated European 3", whose gene identities mirror the Mandenka and Orcadian contemporary samples. In model four, we estimate the ancestry fractions of the admixed population from continental regional proxies (Figure 4.3, right). In this instance, we estimate the ancestry fractions by pooling the simulated populations that mimic the gene identities of the Brong and Mandenka, and the Italian and Orcadian contemporary samples. Our ancestral sources are composed of four simulated contemporary populations. The African ancestral source contains "Simulated African 2" and "Simulated African 3". The European ancestral source contains "Simulated European 2" and "Simulated European 3".

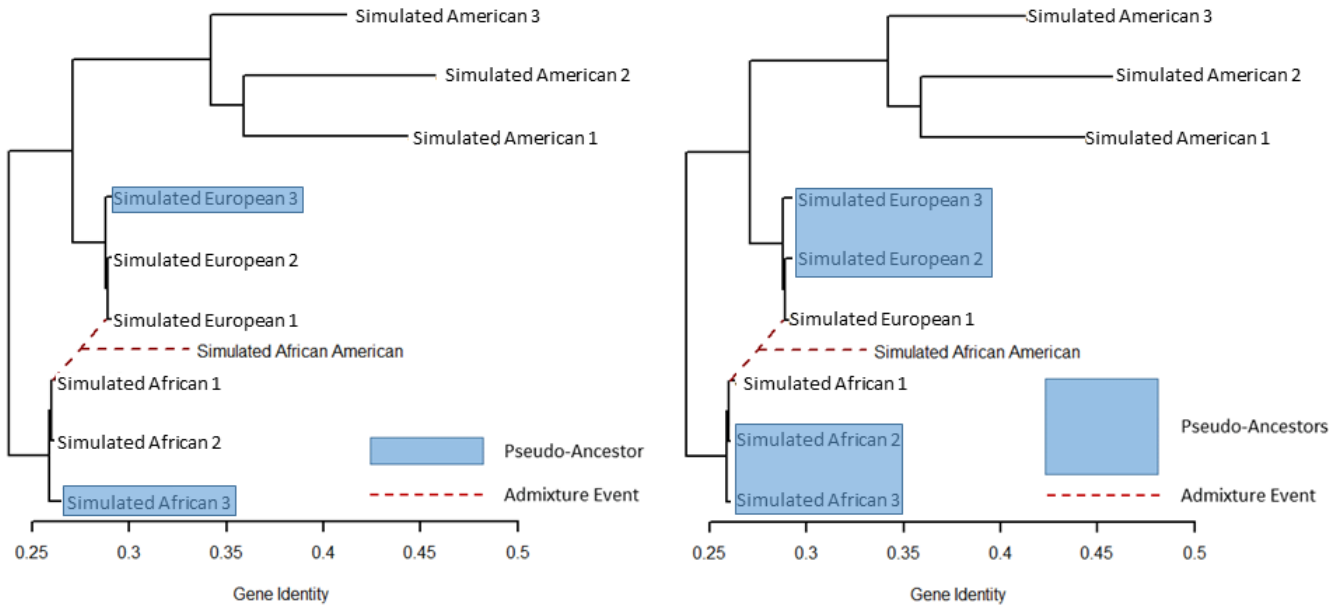

Figure 4.3: Model three (left) estimates ancestry fractions of the simulated African-American population from the most distantly related source proxy of the true source contributor within Africa and Europe. Model four (right) estimates ancestry fractions of the admixed African-American population from multiple pseudo-ancestors within Africa and Europe.

In models five through eight, the true ancestral sources of the admixed population are ancestors of the "Simulated European 1" and "Simulated American 1" populations. In model five, we estimate the ancestry of the admixed population from the descendant pseudo-ancestors, which are the contemporary "Simulated European 1" and "Simulated American 1" samples (Figure 4.4, left). These simulated populations mirror the gene identities of the French and Guaymi. In model six, we estimate the ancestry of the admixed population from the most closely related populations of the true ancestors in the tree (Figure 4.4, right). In the case of the European cluster, we make this estimate from the "Simulated European 2" sample. The closest relation to the "Simulated American 1" in the Americas from our tree is the "Simulated American 2".

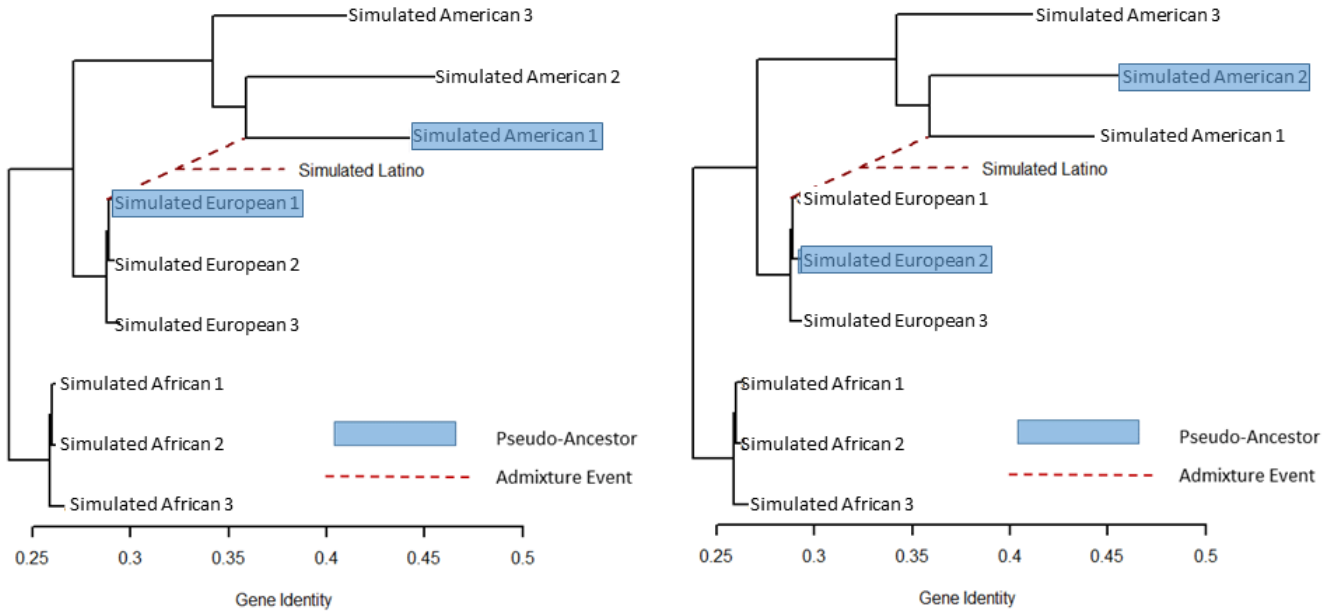

Figure 4.4: Model five (left) estimates ancestry of a simulated Latin American population from pseudo-ancestors that are the descendants of the true ancestors. Model six (right) estimates ancestry fractions from closely related pseudo-ancestors sources that formed an admixed Latin American population.

In model seven, we make ancestry estimates based on the most distantly related populations for each region from the tree (Figure 4.5, left). In this instance, we estimate ancestry from the "Simulated European 3" and "Simulated American 3" even though the known ancestors of the admixed population are those of the "Simulated European 1" and "Simulated American 1". In model eight, we estimate the ancestry of the admixed population by combining the pseudo-ancestral populations from both Europe and the Americas (Figure 4.5, right). In this instance, we use the "Simulated European 2", "Simulated European 3", "Simulated American 2", and "Simulated American 3" contemporary samples as pseudo-ancestors to estimate the ancestry fractions in the admixed population.

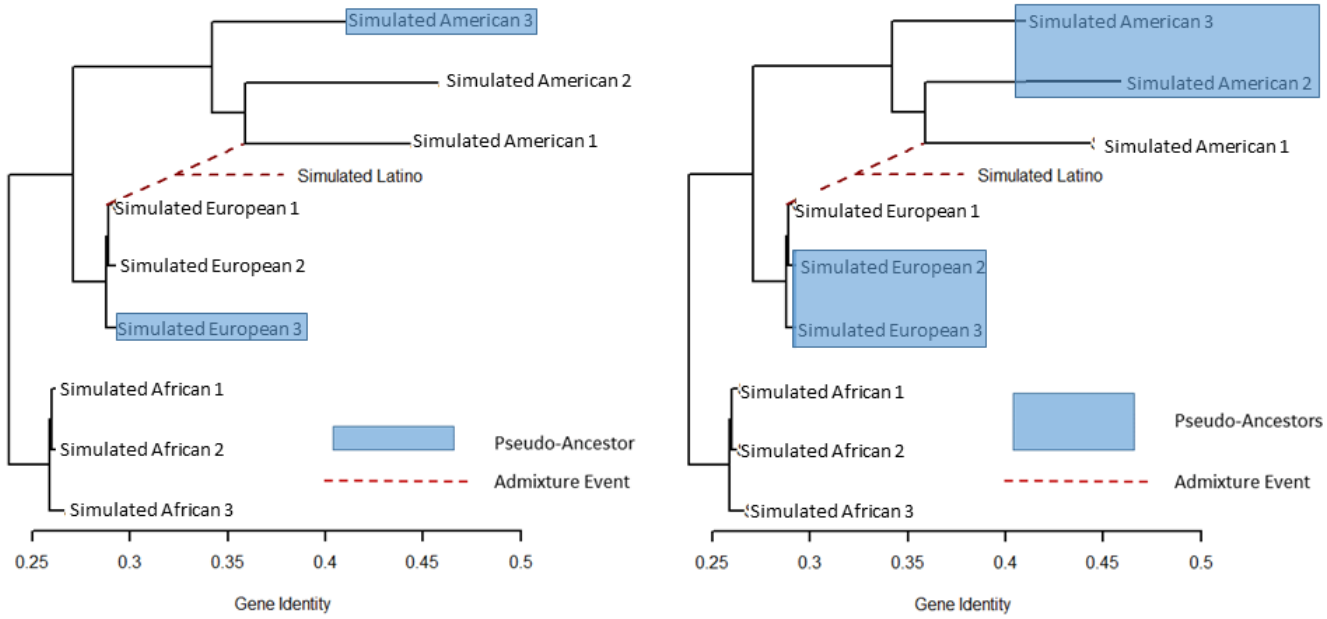

Figure 4.5: Model seven (left) estimates ancestry fractions of an admixed Latin American population from the most distantly related pseudo-ancestral source of the true source contributor within Europe and the Americas. Model eight (right) estimates ancestry fractions of an admixed Latin American population from multiple pseudo-ancestors sources within Europe and the Americas.

We also simulate other scenarios within each of these models. We vary the sample sizes among both the admixed population and the source proxies. Therefore, each model we will present three scenarios for these sampling scenarios. The sampling scenarios that we present include: (1) equal sampling of 100 individuals for the admixed population, as well as the pseudo-ancestral sources; (2) 100 individuals sampled in the admixed population, and 20 individuals from each of the two pseudo-ancestral sources; and (3) 20 individuals sampled from the admixed population, and 100 individuals from each of the two pseudo-ancestors. In addition to these sampling scenarios, we change the timing of the admixture event to account for the effect of genetic drift among the ancestral source proxies. We vary the timing of the admixture event at one generation, ten generations, and twenty generations in the past. We present our findings according to the sample size scenarios for each model, with the timing of the admixture event plotted according to the ancestry fraction estimate. Thus, there are 9,000 simulations per model, which consist of 100 simulations per ancestry

fraction estimate.

4.4 Results

Model one adheres to all but one of the assumptions of our model. The violation occurs when we estimate ancestry fractions within the admixed African-American population from the descendants of the true ancestral sources, which are contemporary samples from the "Simulated African 1" and "Simulated European 1". The upper left panel of Figure 4.6 displays model one as explained in our methods section. Specifically, Figure 4.6a and Figure 4.6 c, we see a lack of bias, because the estimated ancestry fraction values nearly equal their parameter values. The ancestry fraction estimates are measured from the first ancestral source population in all of our simulations. Thus, in models one through four we present the ancestry estimates for the simulated African ancestral source. Genetic drift is more prominent in Figure 4.6a and b than in Figure4.6c. The effect of genetic drift is apparent when the timing of the admixture event occurs further back in time. Figure 4.6b, samples 100 admixed individuals and only 20 individuals from the ancestral source proxy populations. Through this sampling scenario, we see a broad distribution of bias based on the timing of the admixture event. The ancestry fraction estimates for an admixture event one generation in the past are relatively unbiased. We see the bias is more prevalent in smaller ancestry fraction estimates, those less than 0.45. The bias for all ancestry fraction estimates increases as the admixture event occurs further in the past. However, all estimates are nearly unbiased with nearly equal contributions from the ancestral sources, which we see with estimates of 0.45-0.55.

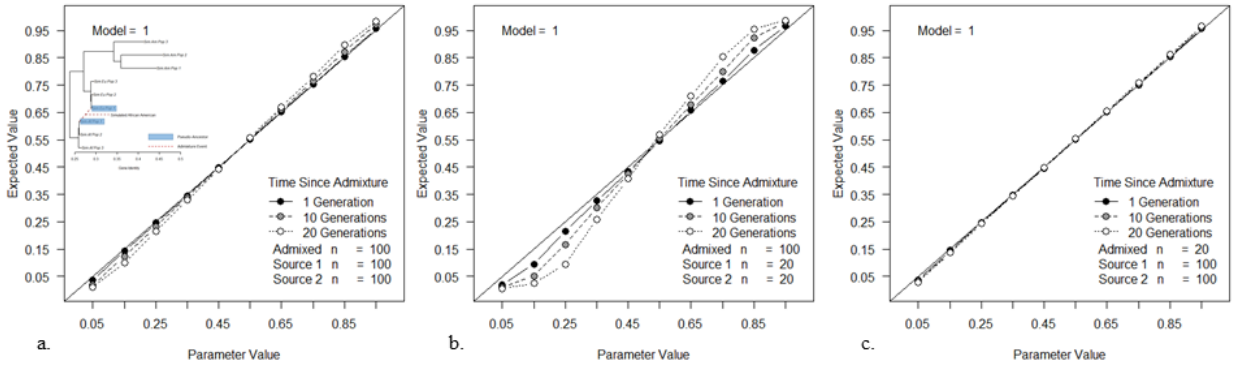

Figure 4.6: Results from model one, which estimates ancestry from the pseudo-ancestors that are the descendants of the the ancestral sources. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed African-American population, and each of the contemporary populations serving as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed African-American population, and 20 individuals from each of the contemporary populations serving as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed African-American population, and 100 individuals from each of the contemporary populations serving as ancestral sources.

Figure 4.7 presents the results from model 2. Figure 4.7a, contains an inset of the model 2, which we presented in our methods section. Model 2 violates the assumption of our model, whereby it makes ancestry fraction estimates from pseudo-ancestors. In the case of model 2, the pseudo-ancestral sources are “Simulated African 2” and “Simulated European 2”. In violating this assumption, model 2 has more biased ancestry fraction estimates compared to model 1, although both models use the same sampling parameters. Figure 4.7a, presents the results of model 2 with equal sample sizes of 100 individuals for both the admixed population and the ancestral source proxies. The results from Figure 4.7a show us that the ancestry fractions underestimate their parameter values. In addition, the effect of genetic drift is apparent as bias increases in the simulations that have the admixture event occurring further in the past. Figure 4.7b shows the ancestry fraction estimates when we sampled 100 individuals from the admixed population, and 20 individuals from each of the ancestral source proxies. Under this sampling scenario, we see a high level of bias for all estimates at every time for the occurrence of the admixture event. There is a clear underestimation of ancestry fractions less than 0.65. The ancestry fractions above the

parameter value of 0.65 are overestimated. Figure 4.7c samples 20 individuals from the admixed African-American population and 100 individuals from each of the source proxies. In addition, we see that the lowest ancestry fractions, 0.05-0.15, are an overestimate of their parameter values. The ancestry fraction estimates at 0.25 are unbiased. The ancestry fractions are increasingly biased and underestimating their parameter values from 0.35 to 0.95. The effect of genetic drift is nominal when we alter the timing of the timing of the admixture event from one generation to 20 generations in the past.

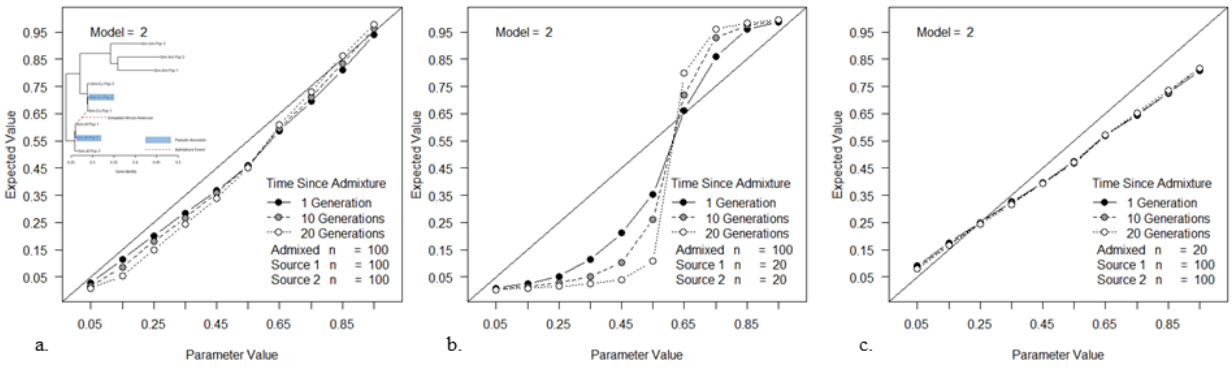

Figure 4.7: Results from model two, which estimates ancestry fractions from ancestral proxies that are closely related contemporary populations to the actual ancestral sources. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed African-American population, and each of the contemporary populations, which serve as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed African-American population, and 20 individuals from each of the contemporary populations, which serve as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed African-American population, and 100 individuals from each of the contemporary populations, which serve as ancestral sources.

Figure 4.8 shows the results from model 3, which makes ancestry fraction estimates from the pseudo-ancestral samples of the “Simulated African 3” and “Simulated European 3”. Recall that the true source contributors are the ”Simulated African 1” and “Simulated European 3”. The details of the model are displayed in the top left of Figure 4.8a. Figure 4.8a samples 100 individuals from each of the admixed population, as well as both source proxies. In this sampling scenario, we see little bias associated with the ancestry estimates occurring one generation in the past. The bias of the ancestry fraction estimates increase as the admixture event occurs further in the past. Figure 4.8b samples 100 individuals from

the admixed population and only 20 individuals from each of the two source proxies. We see from this sampling scenario that the ancestry fraction estimates are extremely biased. Essentially none of the ancestry fraction estimates are close to their parameter values. In addition, under this scenario, we see much dispersion of the ancestry fraction estimates pertaining to the change in timing of the admixture event. The only estimates that do not vary greatly with the timing of the admixture event occur at 0.05, 0.15, 0.85, and 0.95. Figure 4.8c shows us that the ancestry fractions of overestimated from 0.05 to 0.25. The ancestry fraction estimates at 0.35 are unbiased. We see increasing bias as the ancestral source contributions from source population one increase from 0.35 to 0.95. In this instance, the ancestry fractions are underestimated compared to their parameter values.

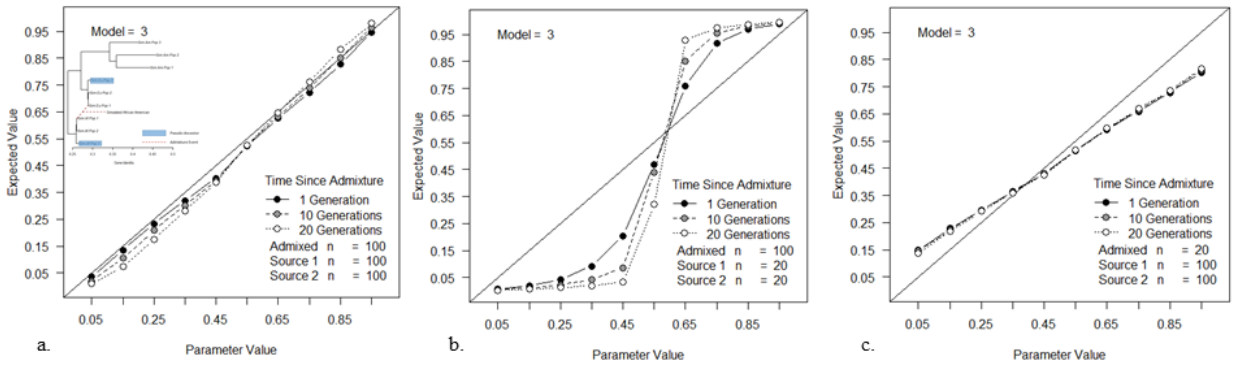

Figure 4.8: Results from model three, which estimates ancestry from distantly related pseudo-ancestors to the actual ancestral sources in their continental regions. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed African-American population, and each of the contemporary populations, which serve as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed African-American population, and 20 individuals from each of the contemporary populations, which serve as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed African-American population, and 100 individuals from each of the contemporary populations, which serve as ancestral sources.

Figure 4.9 displays the results for model four. Here we sample pseudo-ancestral sources from each continental region to estimate the ancestry fractions among the admixed population. The upper left panel of Figure 4.9 a depicts model four, whereby the African pseudo-ancestors consist of samples simulated to resemble the gene identity of the Brong and Mandenka. The simulated African samples are “Simulated African 2” and “Simulated

African 3". The European pseudo-ancestors consist of samples simulated to recapitulate the gene identity of the Italians and Orcadians. The simulated samples we use were "Simulated European 2" and "Simulated European 3". Recall that our simulations mirror the gene identities of the Yoruba and the French, "Simulated African 1" and "Simulated European 1", who serve as the true ancestral sources that provided the allele frequencies to the admixed group. In making ancestry fraction estimates from two ancestral source populations per continental region, we are doubling our sample size for the ancestral sources. Figure 4.9a depicts 100 admixed individuals sampled from the admixed African-American population. In addition, we sampled 200 individuals that serve as ancestral source proxies in each continental region. Figure 4.9a shows that the underestimates the ancestry fractions in the admixed population, as in prior models. However, this model appears more stable, lacking a larger underestimation around 0.45 as in models two and three. In these prior models, we consider only a single ancestral proxy per continental region to estimate ancestry fractions. Figure 4.9b shows the same pattern as prior models under this sampling scheme, even though there are twice as many individuals sampled among each ancestral source proxy. The ancestry fraction estimates vary widely from their parameter values. The effect of genetic drift, based on the timing of the admixture event, also varies widely under this sampling scenario. Figure 4.9c shows us that the ancestry fraction estimates are unbiased at 0.25 and 0.35 in the African source. There is a consistent overestimation of ancestry when the African source is the minor contributor in the formation of the admixed population across all models thus far. In addition, there is a consistent underestimation of ancestry when the African source is the major contributor.

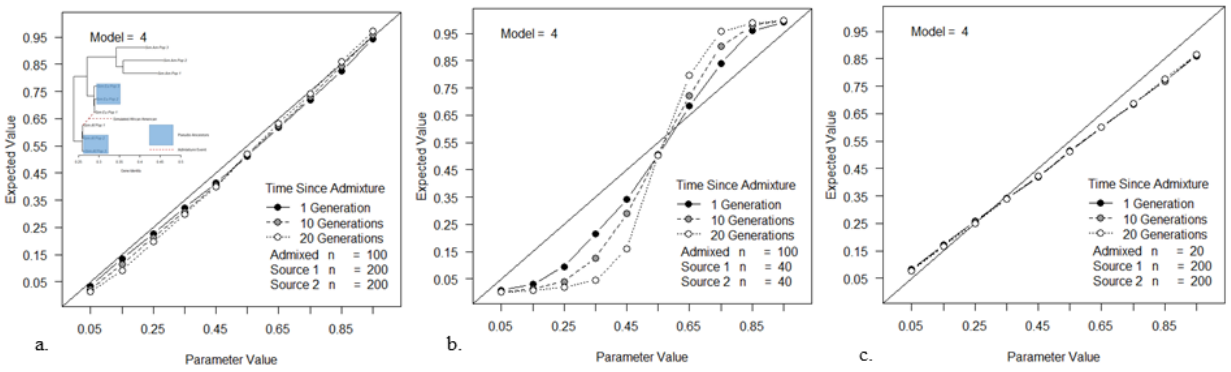

Figure 4.9: Results from model four, which estimates ancestry in a simulated African-American population using multiple related pseudeo-ancestors from their continental regions. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed African-American population, and 200 individuals from each of the contemporary populations, which serve as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed African-American population, and 40 individuals from each of the contemporary populations, which serve as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed African-American population, and 200 individuals from each of the contemporary populations, which serve as ancestral sources.

Figure 4.10 shows the results of model 5 in which the "Simulated European 1" and "Simulated American 2" are the descendants of the ancestral sources that contributed to the admixture event. This scenario conforms to the assumptions of our model in that the pseudo-ancestral sources are the descendants of the known ancestors that formed the admixed population in the Americas. Figure 4.10a samples 100 individuals in the admixed population and both source populations. We see in this scenario that the ancestry fraction estimates are relatively unbiased. We see the most biased estimates in Figure 4.10a are when the admixture event occurred 20 generations in the past. At 20 generations there is an underestimation of the parameter value when the when the first ancestral source is the minor contributor. We see this in the parameter values ranging between 0.05 and 0.25. There is an overestimation in ancestry fractions when the first ancestral source proxy is the major contributor. We see this in the parameter values ranging from 0.65 to 0.95. Figure 4.10b samples 100 individuals from the admixed population and 20 individuals from each of the ancestral source proxies. Figure 4.10b shows us that the ancestry fraction estimates are biased when the first ancestral source proxy is the minor contributor. The ancestry

fraction estimates less biased when the ancestral source proxy is the major contributor to the admixture event. Genetic drift has a larger effect on bias in this scenario as we see from changes in the timing of the admixture event. The change in the timing of the admixture event is notable with the minor ancestral contributions when the parameter values are less than 0.35. Figure 4.10c samples 20 individuals from the admixed population and 100 individuals from each of the ancestral source proxies. We see from this scenario that all ancestry fraction estimates are unbiased. The changes in the timing of the admixture event have no effect on the bias related to the ancestry fraction estimates.

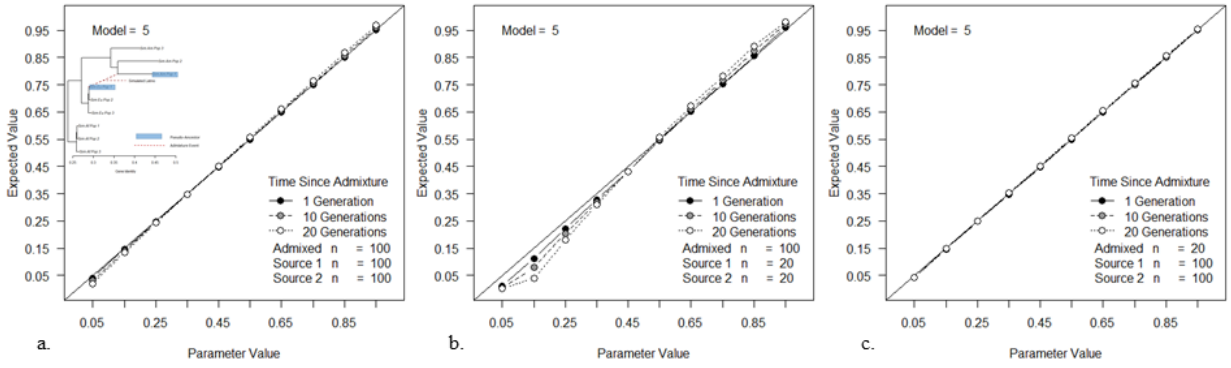

Figure 4.10: Results from model five, which estimates ancestry in a simulated Latin American population from the contemporary descendants of the ancestral sources. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed Latin American population, and each of the contemporary populations serving as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed Latin American population, and 20 individuals from each of the contemporary populations serving as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed Latin American population, and 100 individuals from each of the contemporary populations serving as ancestral sources.

Figure 4.11 presents the results for model six. The ancestry fraction estimates come from simulated pseudo-ancestral sources, "Simulated European 2" and "Simulated American 2", which mirror the gene identities of Italians and Karitiana. The actual source contributors have gene identities that mirror the French and Guaymi. Our simulation samples are identified as "Simulated European 1" and "Simulated American 1". The ancestry estimates present a distinct picture from our previous simulations. Figure 4.11a samples 100 individuals from the admixed population and each of the ancestral source proxies. Figure

4.11a shows that the polar estimates are relatively unbiased. Here for the first time in our simulations, we see an overestimation of ancestry fractions toward the middle parameter values. We see this overestimation in the parameter values ranging from 0.15 to 0.85 for first ancestral source. Figure 4.11b samples 100 individuals from the admixed population and 20 individuals from each of the ancestral source proxies. In this scenario, we see a poor fit of the ancestry fraction estimates to their parameter values. We see that the ancestry fractions are underestimated when the first ancestral source proxy is the minor contributor with parameter values less than 0.25. The ancestry fraction estimates for the parameter values greater than 0.25 are greatly overestimated. Figure 4.11c uses sample sizes of 100 individuals from each ancestral proxy, and 20 individuals from the admixed population. The ancestry fraction estimates are biased except for the highest parameter value. All other ancestry fractions in this scenario are overestimated.

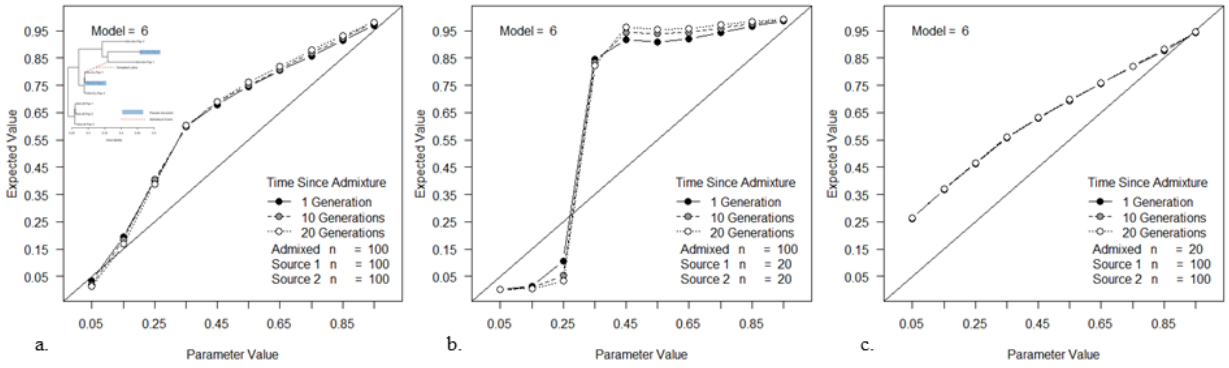

Figure 4.11: Results from model six, which estimates ancestry in a simulated Latin American population from contemporary samples that serve a pseudo-ancestors that are closely related to the actual ancestral sources. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed Latin American population, and each of the contemporary populations, which serve as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed Latin American population, and 20 individuals from each of the contemporary populations, which serve as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed Latin American population, and 100 individuals from each of the contemporary populations, which serve as ancestral sources.

Figure 4.12 shows us the ancestry fraction estimates from model seven. Recall that this sampling scenario assumes the true ancestral sources are those simulated populations that mirror the gene identities of the French and Guaymi. The simulated samples, in this

instance are “Simulated European 1” and “Simulated American 1”. In this model, we estimate ancestry from simulated populations whose gene identities mirror the Orcadians and Pima, which are “Simulated European 3” and “Simulated American 3”, repectively. Figure 4.12a samples 100 individuals from the admixed population, and each of the pseudo-ancestral sources. This figure shows us an initial underestimation of ancestry fractions at parameter values 0.05 and 0.15. The other parameter values, 0.25 to 0.95, show an overestimation of the ancestry fractions. We see the greatest bias in the estimates ranging from 0.35 to 0.55. Figure 4.12b samples 100 admixed individuals and 20 individuals from each of the pseudo-ancestral sources. Figure 4.12b shows us that using this sampling scenario with the following pseudo-ancestral sources produces highly biased results. The only estimates that are near their parameter values are seen at 0.35. However, we see that genetic drift has a large effect when we change the timing of the admixture event at this parameter value. Figure 4.12c samples 20 individuals from the admixed population and 100 individuals from each of the two ancestral proxies. Figure 4.12c shows a consistent overestimation of ancestry fractions ranging from 0.05 to 0.45. The bias of ancestry fractions, although decreasing, is still overestimated from 0.55 to 0.75. The ancestry fraction estimates are the least biased at the parameter value 0.85, and are slightly underestimated. The ancestry fraction that is clearly underestimated occurs at 0.95.

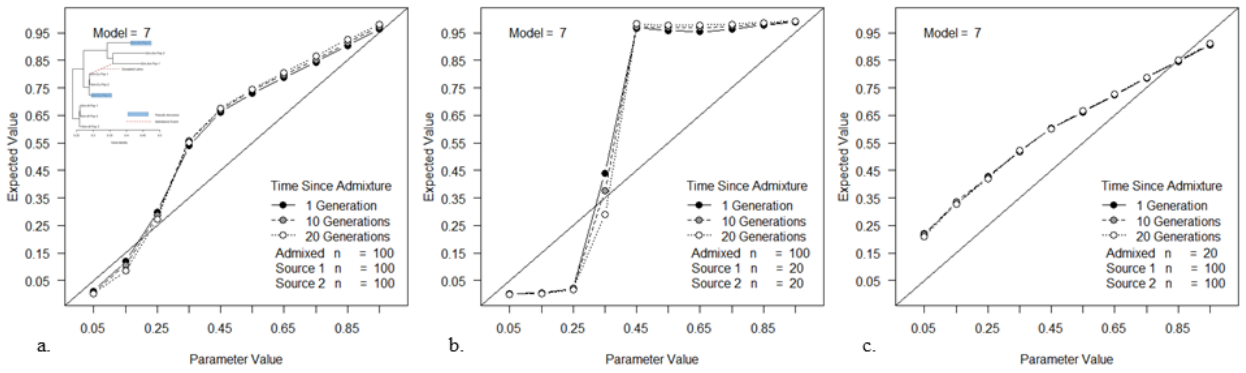

Figure 4.12: Results from model seven, which estimates ancestry in a simulated Latin American population from pseudo-ancestors that are distantly related to the actual ancestral sources. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed Latin American population, and each of the contemporary populations, which serve as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed Latin American population, and 20 individuals from each of the contemporary populations, which serve as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed Latin American population, and 100 individuals from each of the contemporary populations, which serve as ancestral sources.

Figure 4.13 presents the results from model eight. Recall that model eight samples two pseudo-ancestral sources from each continental region, Europe and the Americas, respectively. We simulated the gene identities of these ancestral proxies to recapitulate the genetic qualities of the Italians, Orcadians, Karitiana, and Pima. Our simulations mirror these samples and are designated as "Simulated European 2", "Simulated European 3", "Simulated American 2", and "Simulated American 3". We estimate ancestry fractions in a simulated Latin American population from these pseudo-ancestral sources. For Figure 4.13a, we estimated the ancestry fractions by sampling 100 individuals from the simulated Latin American population, and 200 individuals from the pseudo-ancestors form each continental region. Here, we see a similar pattern of ancestry fraction estimates as models six and seven. However, the bias in model 8 is not as pronounced as models six and seven. In all three simulations (Fig. 4.13a), we see the most biased estimate at 0.45. The most biased of these occurred with the admixture event 10 generations in the past. The estimate of the parameter is 0.5505, thus the bias of the estimate is 0.1005. In Figure 4.13b, we sampled 100 individuals from the simulated Latin American population, and 40 individuals

from each of the ancestral proxies from Europe and the Americas. The ancestry estimates at 0.05 and 0.15 have as much bias as the estimates themselves. The estimate at 0.05 is 0.0008, and the associated bias is -0.0492. The ancestry estimate at 0.15 is 0.0093, and bias of this estimate is -0.01407. The negative values show that these are underestimates of the parameter values. This sampling scenario demonstrates high bias is ancestry estimation as none of the estimates is close to their parameter values. In Figure 4.13c, we sampled 20 individuals from the simulated Latin American population, and 200 individuals from each of the ancestral proxies from Europe and the Americas. The ancestry fraction estimates are biased when the first ancestral source is the minor ancestral contributor. For the parameter values from 0.05 to 0.35, the bias of the estimates is near 0.10 for all. The least biased estimates in this scenario occur at the 0.85 and 0.95 parameter values. The bias of these estimates is approximately 0.01 and -0.01, respectively.

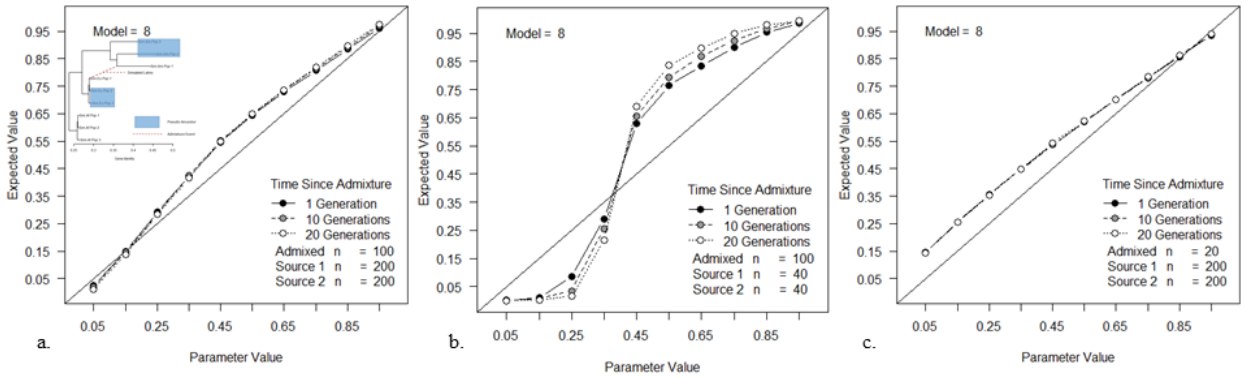

Figure 4.13: Results from model eight, which estimates ancestry in a simulated Latin American population from multiple pseudo-ancestors that from each continental region. Each panel contains three series of simulations that vary the timing of the admixture event, 1 generation in the past (black), 10 generations in the past (gray), and 20 generations in the past (white). (a) We sample 100 individuals from the contemporary admixed Latin American population, and 200 individuals from each of the contemporary populations, which serve as ancestral proxies. (b) We sample 100 individuals from the contemporary admixed Latin American population, and 40 individuals from each of the contemporary populations, which serve as ancestral proxies. (c) We sample 20 individuals from the contemporary admixed Latin American population, and 200 individuals from each of the contemporary populations, which serve as ancestral sources.

There are consistent patterns of bias seen among our models based on the results. These patterns of bias are related to the selection of pseudo-ancestral sources from which we estimate ancestry fractions in the admixed population, the timing of the admixture event, and

the sample size of the populations from which we make these estimates. In models one and five, we estimated ancestry fractions from contemporary populations that descended from the true ancestral sources. We see that all ancestry fraction estimates are close to the parameter values, thus they are the least biased. This is because the assumptions of our model are met, and the ancestry fraction estimates come from the populations that contributed to the admixture event. The other models, two through four and six through eight, violate the assumption in our model, in that the ancestral sources did not contribute to the formation of the admixed population. Models two through four simulate the formation of an admixed African-American population. The ancestral proxies we used come from have the lowest simulated gene identities in our analyses, which reflect African and European origins. Recall that Figure 4.7a and Figure 4.8a sample 100 individuals from both pseudo-ancestors and the contemporary admixed population, and the ancestry fractions are underestimated in all of these models. Figure 4.12a and Figure 4.13a simulate the formation of an admixed Latin American population from European and Indigenous American pseudo-ancestors. These Latin American simulations overestimate ancestry fractions with the same sample sizes as Figure 4.7a and Figure 4.8a. The only difference in these models is the simulated gene identities of the source proxies used to estimate ancestry. In all of these simulations, we see the effect of genetic drift increases as the timing of the admixture event occurs further in the past. The results show us that sampling more admixed individuals than those that serve as the ancestral proxies produces highly biased ancestry fraction estimates. These biased estimates are only exacerbated the further back in time the admixture event occurred. In the simulations that use pseudo-ancestors that are not descended from the populations that contributed to the admixture event and mirror African-American admixture. We see ancestry fractions in this case are overestimated when ancestral source one is the minor contributor (fig. 4.7c, fig. 4.8c, and fig. 4.8c). In the simulations that recapitulate the formation of an admixed Latin American population, the majority of ancestry fractions are overestimated (fig 4.11c, fig 4.12c, and fig. 4.13c). The models using ancestral proxies,

and sample a large number of individuals from each proxy, have the least amount of bias. In other words, all models that combined ancestral proxies from their respective continents fared better than those models that used only one.

4.5 Discussion

The allele frequencies of an admixed population are a linear combination of allele frequencies found in the ancestral source populations (Long, 1991). The coefficients of the linear combination reflect ancestry fractions in the mixed population attributed to each ancestral source population. For this reason, maximum likelihood is a common statistic used in the estimation of ancestry fractions in genetically mixed populations (Tang et al., 2005). Here, we have addressed several factors relevant to the structure of data used in maximum likelihood estimation of genetic ancestry. Specifically, we investigated how the structure of the data may bias ancestry estimation.

The first factor of the data that we considered in estimating ancestry of admixed populations was sample size. We worked to determine how large a sample size must be to sufficiently obtain unbiased estimates of ancestry. The sampling scenarios presented compelling and consistent results. In sampling 100 individuals from each of the populations that serve as ancestral sources and for the admixed group, we saw that bias was present when the ancestral contributions were nearly equal from each ancestral source. This pattern is absent in evolutionary models one and five because these models sample individuals that are the descendants of the actual ancestral source populations. We examined a second sampling scenario in which we included a large sample of admixed individuals (n=100) and a small number of individuals from each population that serve as ancestral sources (n=20). In our evolutionary models four and eight we sampled the regional averages of 40 individuals from each region that serve as ancestral sources. We were surprised by the findings of this sampling scheme. We expected the sampling of these evolutionary models to perform

well with the maximum likelihood method. We reasoned that a smaller number of individuals sampled as the ancestral sources would merely guide the iterative likelihood method. We anticipated the majority of information to estimate the ancestral source allele frequencies and ancestry fractions came from the admixed sample. Simply put, we were wrong in this expectation, as this sampling scenario in all eight of our evolutionary models produced the most biased results. Our final scenario sampled 20 individuals from the admixed population and 100 individuals from each of the populations that serve as the ancestral source populations. Our evolutionary models four and eight sampled the regional averages for each region that serve as ancestral sources (n=200). The The minor ancestral contributions in these scenarios show the most bias, while equal ancestral contributions from the samples we selected as ancestral source populations were unbiased. Our evolutionary models four and eight sample a regional average in the estimation of ancestry produce less biased results compared to other models that only sample a single population to serve as an ancestral source per region. The reduction of bias occurs because of two possibilities. The first is that taking a regional average of allele frequencies to estimate ancestry fractions provides more information from which the likelihood method can draw. The second possibility is due simply to sample size whereby more individuals are included into the evolutionary model. These possibilities are confounding factors in our analyses, and will be the focus of future investigation.

Next, we considered the misspecification of ancestral source populations. Hence, we estimated ancestry from samples of populations in our analyses that are closely related to the actual ancestral populations, but did not contribute to the admixture event. The misspecification of ancestral source populations must be considered in all admixture analyses because it is possible that ancestral source populations no longer exist. In reality no ancestral source population exists unchanged from the time they contributed to an admixture event. Ancestral source populations have been affected by allele frequency drift and have evolved since the time of the admixture event. This point is relevant to models one and four

in which the genetic samples we chose to serve as pseudo-ancestors are the descendants of the true ancestors that formed the mixed group. It is for that reason that these evolutionary models produced unbiased estimates of ancestry. Additionally, a sparse historical record prevents us from fully knowing who the true ancestors of a mixed population were. For this we must use contemporary genetic samples of populations to serve as proxies, i.e., pseudo-ancestors, in the estimation of ancestry in admixed populations. Thus, all contemporary samples serving as ancestral source populations must be considered as pseudo-ancestors in admixture analyses.

Finally, we examined how ancestry estimates are affected by genetic drift since the founding of the mixed population. In our evolutionary models we see that bias due to genetic drift occurs when we sampled a large number of individuals from the mixed population and the samples that serve as ancestral sources. The ancestry estimates are unbiased when we simulate the time of the admixture event at one generation in the past, but bias increases as the admixture event occurs further in the past. The second scenario, sampling a large number of admixed individuals and fewer individuals that serve as an ancestral source population, is highly biased. The estimates are not close to their parameter value and only deviate from the parameter value as the admixture event occurred further back in time. The last sampling scenario, using a small number of individuals from the admixed population and a large sample of individuals as an ancestral source population, produced no deviation from the parameter value as we set the admixture event to occur further in the past. The effect of genetic drift and the timing of the admixture event are factors we must take into consideration in our research design.

In converting unknown historical and genetic variables into known parameters, we were able to assess the magnitude and direction of bias related to ancestry fraction estimates. Through this research, we demonstrate the impact of using pseudo-ancestors in estimating ancestry. The use of a single pseudo-ancestor per continental region creates bias in ancestry fraction estimates. Increasing the number of pseudo-ancestors per continental region alle-

viates the bias associated with the ancestry estimates. The various sample sizes we used for each model shows that the best scenario is to incorporate more individuals to serve as pseudo-ancestors relative to the number of admixed individuals in an analysis. This study validates the research conducted heretofore in that multiple contemporary populations are used to serve as proxies in the estimation of ancestry fractions among admixed populations.

This research allows us to make recommendations regarding the sampling of future admixture studies. If possible, sample contemporary populations that are known to have descended from the true ancestors who contributed to the admixture event. These scenarios show the least biased ancestry estimates in our results. However, as we have demonstrated, the true descendants are not known. This being said, we must sample a large number of individuals from multiple contemporary populations, which serve as pseudo-ancestors. Our simulations show that this sampling scenario reduced the biased associated with ancestry estimation.

At first glance, it may seem that these simulations are an oversimplified approach to address long-standing problems in genetic admixture studies. However, to examine the few parameters we did took hundreds of computational hours. Each simulation series consisted of 1000 simulations with 100 simulations per ancestry fraction estimate and took 4 hours on average using FASTSIMCOAL. Running these simulations through our programs took between 12 to 72 hours depending on the sample sizes used for each simulation. We ran 72,000 simulations to conduct this research. The average computational time for 1000 simulations took 36 hours and the total computational time spanned nearly 2,600 hours.

Given the nature of this research, we could still do much more. For instance, we could examine other parameters within this model. These parameters could include, but are not limited to, increasing the number of loci, further varying sample sizes, investigating changes of effective population size in the admixed population, estimating ancestry from three or more ancestral sources, and including more pseudo-ancestors per continental region. In addition, we could use a model of continuous gene flow to examine how pseudo-

ancestors to effect estimating ancestry. The benefit might be that we could improve model selection, and the choice of pseudo-ancestors from the use of simulated data in admixture analyses.

Chapter 5

Conclusion

In this dissertation, I have addressed assumptions inherent in maximum likelihood estimation, and challenges related to admixture analyses. The primary assumption of maximum likelihood estimation that I examined is that gene flow in the form of admixture is the only evolutionary process operating in this system. A difficulty involves proper identification of ancestral source populations that contributed to the admixture event. There are three primary challenges confronting the identification of ancestral source populations. (1) Admixture events that formed many contemporary populations began or occurred entirely in the past. (2) Ancestral source populations may no longer exist, or they have evolved since the time of the admixture event. (3) A sparse historical record prevents us from fully knowing the source populations. These challenges require the use of pseudo-ancestors. Tang and colleagues (2005) describe pseudo-ancestors as contemporary populations that are descended from close relatives of the true ancestors of admixed populations.

In chapter two, I introduced a new method, which assesses more fully the genetic diversity of admixed populations. I partitioned Nei's minimum genetic distance into components of ancestry and genetic drift in admixed populations of the Americas. In partitioning genetic distances, I showed there are significant contributions due to the processes of admix-

ture and genetic drift that shape the pattern of genetic diversity within and between admixed populations. The genetic structure of admixed populations in the Americas reflects more than continental ancestry. Allele frequencies are another way to view the genetic structure of populations. The fact that genetic drift is active in these populations violates the assumption that admixture is the only evolutionary process operating in the system.

The results show that both ancestry and genetic drift contribute to the genetic structure in admixed populations of the Americas, yet they manifest in varying ways. Ancestry proportions as shown in Figure 2.2, are one way to view genetic structure. The ancestral sources in this figure are located in the corners of the triangle plot with European ancestors plotted at the apex. When I use principal coordinates to examine the genetic structure according to the genetic distance of allele frequencies, we see ancestry correlates to only one primary axis (fig. 2.3). The second axis on the principal coordinate plot shows the differentiation of populations due to F_{ST}. On the primary axis, the European ancestors are in line with, and intermediate to the African and Indigenous American sources. In addition, the admixed populations fall on a line between their ancestral sources. For example, African-American populations are located between their African and European ancestral sources. The admixed populations are genetically intermediate to, yet distinct from their ancestral sources (Cavalli-Sforza et al., 1994).

Figure 2.5 depicts the outcome of partitioning genetic distance into components of ancestry and drift. The pairwise comparisons of Figure 2.5 demonstrate how drift is a prominent force among the admixed populations. For the African-American by Latin American comparisons, we see that they have the greatest genetic distances between them. Genetic drift has the least effect between African-American and Latin American populations, yet genetic drift explains up to 30% of the genetic distance between some of these populations. Thus, the genetic distances are due predominantly to differences in their ancestry fractions. The pairwise comparisons among the Latin American populations shows high variability in both genetic distance and the effect of drift. This dispersion is due to a lack of shared

common ancestry. Much differentiation exists in the proportion of ancestry each founding group contributed to the admixed populations. In addition, independent demographic circumstances occurred among these populations, which may be explained by small effective population sizes and relative isolation from one another. This isolation is reflected in the F_{ST} values of both the samples that serve as pseudo-ancestors, and in the admixed populations themselves. These findings demonstrate that Latin American populations are not a homogeneous meta-population. They formed by a discrete independent process across the landscape of the Americas. Latin Americans are a much more complex population, or rather, suite of populations, than the common label of "Latin American" suggests. The pairwise comparisons of the African-American populations show the smallest genetic distances among these populations. This fact is due in large part to a shared ancestry in both the populations that contributed to their founding, as well as the proportion of ancestry each ancestral contributed. The vast majority their genetic differentiation is due to genetic drift.

The pattern of this result reflects the genetic diversity of populations on a global scale is driven by a series of founder effects (Ramachandran et al., 2005). Populations living in Africa have the highest levels of genetic diversity, in terms of both the kinds of alleles and heterozygosity. Eurasian populations have lower heterozygosity and possess a subset of the alleles found in Africans. Indigenous Americans have a subset of allelic types found in Eurasians and ever lower heterozygosity. The reduction of genetic diversity due to founder effects created the pattern of genetic distances of populations found on different continents. Therefore, the ancestry of admixed populations determines their placement on the axis according to genetic distance but does not introduce new axes of variation.

In chapter three, I addressed a major goal in admixture analyses, which is to estimate the contributions of ancestors to admixed individuals and populations. Many statistical methods are available to estimate the number of sources and the contributions in an admixed population (Pritchard et al., 2000; Tang et al., 2005; Alexander et al., 2009). I used maximum likelihood to estimate the number of sources in a mixture (Tang et al., 2005).

The assessment of contributions to a mixture is typically achieved by constructing allele frequencies as a linear combination in populations that contributed ancestors to the mixed sample. In this line of research, I addressed many of the challenges in admixture analyses, which include, the true ancestral sources may no longer exist, or have not been genetically sampled, or are otherwise unavailable for study. These challenges require constructing pseudo-ancestors as a substitute for the true ancestors in admixture analyses.

Here, I advocated for the use of the Akaike Information Criterion (AIC) to rank multiple models of proposed ancestry for a focal mixed population, the Cape Coloured people of South Africa. AIC is the sum of twice the log likelihood and two times the number of parameters in a model (Akaike, 1974). I developed a novel strategy through this research. I started with genotype data from individuals of a focal mixed populations, and samples from many regional populations found throughout the world. The regional samples served as pseudo-ancestors for sources of ancestry in the mixed sample. I constructed a series of ancestry models that specified a predefined number of ancestral sources for the mixed sample. Each ancestry model was composed of a separate data set, and contained data from only the mixed sample and appropriate pseudo-ancestors that serve as ancestral sources in the model. Then, I constructed models to estimate ancestry based on the number of ancestral sources in each model. I tested the fit of each model to the mixed sample by comparing the observed allele frequencies to the allele frequencies predicted from the inferred source populations. I used the AIC to decide on which ancestry model is the most appropriate representation of the data.

The results showed that all 26 models of proposed ancestry fit the data of the Cape Coloured people extremely well. The R^2 values for all models fell into a narrow range $[0.930 \leq R^2 \leq 0.951]$. Despite the high R^2 values, the ancestry coefficients were highly inconsistent across models. By contrast, to R^2, the AIC criterion clearly separates the 26 models. There is one unequivocal best model, which consists of only two sources of ancestry, represented by Khoesan and East Asians as the ancestral source populations. AIC

ranks these models according to the number of parameters they estimate. The model most favored by AIC required the estimation of 9,277 parameters while the least favored model required the estimation of 27,222 parameters. It is important to understand why models with such different ancestral populations do equally well in predicting allele frequencies in the Cape Coloured people. This effect occurs because genetic ancestry analysis is a generalized regression problem in which allele frequencies in the ancestral sources serve as the predictor variables. In the case of the human species, allele frequencies among the pseudo-ancestors are highly correlated, even among the most diverged populations such as the Khoesan and populations from east Asia.

In chapter four, I used coalescent simulations to investigate the concept of pseudo-ancestors further. The use of data simulation allowed me to address many of the challenges related to admixture analyses. These challenges relate to using contemporary genetic samples to gain information of historical and evolutionary processes that began or occurred in the past. The populations that contributed to form the admixed group may no longer exist or have evolved since the time of the admixture event. A sparse historical record exists, which prevents us from fully knowing who the source populations were.

I addressed these challenges by using a simulated phylogenetic tree of populations that mirrored the genetic structure of samples from contemporary populations (fig. 4.1). I constructed three continental regions each of which was composed of three pseudo-ancestral populations. The simulated pseudo-ancestors in each continental region mirror the gene identities of the Yoruba, Brong, and Mandenka in Africa; the French, Italian, and Orcadian in Europe; and the Guaymi, Karitiana, and Pima of the Americas. These samples came from the literature (Cann et al., 2002; Rosenberg et al., 2002; Wang et al., 2007; Tishkoff et al., 2009). I used the data set that combined these data from original studies, which were calibrated according to allele size (Pemberton et al., 2013). I calculated a gene identity matrix from the above samples from 618 microsatellite loci, and used generalized hierarchical modeling (GHM) to estimate branch lengths and root the tree (Long et al., 2009). I then

used a step-wise model in FASTSIMCOAL (Excoffier and Foll, 2011) to find appropriate effective population sizes and separation times that recreate the observed tree (fig. 4.1).

I investigated eight models to determine how the implementation of pseudo-ancestors may bias ancestry fraction estimates. In models one through four, I simulated a single admixture event between two pseudo-ancestors in the African and European continental source regions. The admixture event occurred between 'Simulated African 1' and 'Simulated European 1' in the formation of an African-American population. The simulated populations mimic the gene identities of the Yoruba and French, respectively. In each model I estimated ancestry from different pseudo-ancestors in each continental region (fig. 4.2, fig. 4.3). In models five through eight, I simulated a single admixture event of two pseudo-ancestors in the European and American continental source regions. The admixture event occurred between 'Simulated European 1' and 'Simulated American 1' to form an admixed Latin American population. The gene identities of the simulated population mirror that of the French and Guaymi, respectively. In these models I estimated the ancestry coefficients from different pseudo-ancestors among the European and American continental regions (fig. 4.4, fig. 4.5).

In each model, I varied the sample sizes among the contributing ancestral source populations and the admixed population. First, I used equal sample sizes of 100 individuals from the admixed population, as well as each of the ancestral sources. Second, I sampled 100 individuals from the admixed population, and 20 individuals from each of the ancestral sources. Third, I sampled 20 individuals from the admixed population, and 100 individuals from each of the ancestral sources. The ancestry fraction estimates showed similar patterns of bias across all models based on the sampling scenario used.

Models one and five had the least biased estimates of ancestry (fig. 4.6, fig. 4.10). These models estimated ancestry from the descendants of the populations that actually contributed to the admixture event. The sampling of a single pseudo-ancestor that was not directly descended from the true ancestor from a continental region (models two, three,

six, and seven) yielded similar results (fig. 4.7, fig. 4.8, fig. 4.11, fig. 4.12). The sampling of multiple pseduo-ancestors (model four and eight) per continental region (fig. 4.9, fig. 4.13) alleviated much of the bias relative to a the models that sampled only a single pseudo-ancestor per region. The sampling scenario that yielded the most biased ancestry estimates sampled 100 admixed individuals and 20 individuals from each of the pseudo-ancestral population.

From this research, I can make some recommendations regarding sampling methods in future admixture studies. When possible, we must sample from contemporary populations that have descended from the true ancestors that contributed in the formation of the admixed population. The models that sampled from descendant pseudo-ancestors showed the least biased ancestry estimates in my analyses. In cases in which the descendant groups are unknown, we must sample from multiple pseudo-ancestors from each continental region of interest. My simulations suggest that it is best to sample a large number of individuals from multiple contemporary populations per continental region.

Bibliography

ADAMS, J. AND WARD, R. H. 1973. Admixture studies and the detection of selection. *Science* 180:1137–1143.

AKAIKE, H. 1973. Information theory and an extension of the maximum likelihood principle. *In* B. Petrov and F. Csaki (eds.), Proceedings of the 2nd International Symposium of Information Theory, pp. 267–281, Budapest. Akademiai Kiado.

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19:716–723.

AKAIKE, H. 1981a. Likelihood of a model and information criteria. *Journal of Econometrics* 16:3–14. Lp953 Times Cited:295 Cited References Count:18.

AKAIKE, H. 1981b. Modern development of statistical methods. Pergamon Press, Oxford.

AKAIKE, H. 1983. Statistical inference and measurement of entropy. Academic Press, New York.

ALEXANDER, D. H., NOVEMBRE, J., AND LANGE, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–64. Alexander, David H Novembre, John Lange, Kenneth eng GM53275/GM/NIGMS NIH HHS/ MH59490/MH/NIMH NIH HHS/ T32GM008185/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural 8/4/2009 9:00 Genome Res. 2009 Sep;19(9):1655-64. doi: 10.1101/gr.094052.109. Epub 2009 Jul 31.

ANDERSON, D. R. 2008. Model Based Inference in the Life Sciences: A Primer On Evidence. Springer, first edition.

BONILLA, C., GUTIERREZ, G., PARRA, E. J., KLINE, C., AND SHRIVER, M. D. 2005. Admixture analysis of a rural population of the state of guerrero, mexico. *Am J Phys Anthropol* 128:861–9. Bonilla, Carolina Gutierrez, Gerardo Parra, Esteban J Kline, Christopher Shriver, Mark D eng HG002154/HG/NHGRI NIH HHS/ Comparative Study Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 8/25/2005 9:00 Am J Phys Anthropol. 2005 Dec;128(4):861-9.

BONILLA, C., PARRA, E. J., PFAFF, C. L., DIOS, S., MARSHALL, J. A., HAMMAN, R. F., FERRELL, R. E., HOGGART, C. L., MCKEIGUE, P. M., AND SHRIVER, M. D. 2004. Admixture in the hispanics of the san luis valley, colorado, and its implications for complex trait gene mapping. *Annals of Human Genetics* 68:139–153.

BURNHAM, K. P., ANDERSON, D. R., AND HUYVAERT, K. P. 2011. Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology* 65:23–35. 701AF Times Cited:274 Cited References Count:72.

CANN, H. M., DE TOMA, C., CAZES, L., LEGRAND, M. F., MOREL, V., PIOUFFRE, L., BODMER, J., BODMER, W. F., BONNE-TAMIR, B., CAMBON-THOMSEN, A., CHEN, Z., CHU, J., CARCASSI, C., CONTU, L., DU, R., EXCOFFIER, L., FERRARA, G. B., FRIEDLAENDER, J. S., GROOT, H., GURWITZ, D., JENKINS, T., HERRERA, R. J., HUANG, X., KIDD, J., KIDD, K. K., LANGANEY, A., LIN, A. A., MEHDI, S. Q., PARHAM, P., PIAZZA, A., PISTILLO, M. P., QIAN, Y., SHU, Q., XU, J., ZHU, S., WEBER, J. L., GREELY, H. T., FELDMAN, M. W., THOMAS, G., DAUSSET, J., AND CAVALLI-SFORZA, L. L. 2002. A human genome diversity cell line panel. *Science* 296:261–2.

CAVALLI-SFORZA, L. L. AND BODMER, W. F. 1971. The Genetics of Human Populations. Dover Publications, Inc., Mineola, New York.

CAVALLI-SFORZA, L. L., MENOZZI, P., AND PIAZZA, A. 1994. The History and Geography of Human Genes. Princeton University Press, Princeton, New Jersey.

CHAKRABORTY, R. 1986. Gene admixture in human populations: Models and predictions. *Yearbook of Physical Anthropology* 29:1–43.

DAVENPORT, R. AND SAUNDERS, C. 2000. South Africa: A Modern History. St. Martin's Press Inc., New York, New York, fifth edition.

DE VILLIERS, J. 2014a. Cape colonial society under British rule, 1806-1834, book section 4. Protea Book House, Pretoria, South Africa, first edition.

DE VILLIERS, J. 2014b. The Dutch era at the Cape, 1652-1806, book section 2. Protea Book House, Pretoria, South Africa, first edition.

DE WIT, E., DELPORT, W., RUGAMIKA, C., MEINTJES, A., MOLLER, M., VAN HELDEN, P., AND SEOIGHE, C. 2010. Genome-wide analysis of the structure of the south african coloured population in the western cape. *Human Genetics* 128.

EFRON, B. AND TIBSHIRANI, R. 1993. An introduction to the bootstrap. Monographs on statistics and applied probability. Chapman and Hall, New York.

ELPHICK, R. 1977. Kraal and Castle. Yale University Press, New Haven and London, first edition.

EXCOFFIER, L. AND FOLL, M. 2011. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27:1332–4. Excoffier, Laurent Foll, Matthieu eng Research Support, Non-U.S. Gov't England Oxford, England 3/15/2011 6:00 Bioinformatics. 2011 May 1;27(9):1332-4. doi: 10.1093/bioinformatics/btr124. Epub 2011 Mar 12.

GALANTER, J. M., FERNANDEZ-LOPEZ, J. C., GIGNOUX, C. R., BARNHOLTZ-SLOAN, J., FERNANDEZ-ROZADILLA, C., VIA, M., HIDALGO-MIRANDA, A., CONTRERAS, A. V., FIGUEROA, L. U., RASKA, P., JIMENEZ-SANCHEZ, G., ZOLEZZI, I. S., TORRES, M., PONTE, C. R., RUIZ, Y., SALAS, A., NGUYEN, E., ENG, C., BORJAS, L., ZABALA, W., BARRETO, G., GONZALEZ, F. R., IBARRA, A., TABOADA, P., PORRAS, L., MORENO, F., BIGHAM, A., GUTIERREZ, G., BRUTSAERT, T., LEON-VELARDE, F., MOORE, L. G., VARGAS, E., CRUZ, M., ESCOBEDO, J., RODRIGUEZ-SANTANA, J., RODRIGUEZ-CINTRON, W., CHAPELA, R., FORD, J. G., BUSTAMANTE, C., SEMINARA, D., SHRIVER, M., ZIV, E., BURCHARD, E. G., HAILE, R., PARRA, E., CARRACEDO, A., AND CONSORTIUM, L. 2012. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the americas. *PLoS Genet* 8:e1002554.

GOLDSTEIN, D., RUIZ LINARES, A., CAVALLI-SFORZA, L. L., AND FELDMAN, M. W. 1995. An evalutation of genetic distances for use with microsatellite loci. *Genetics* 139:463–471.

GOWER, J. 1966. Some distance properties of latent root and vector methods usin in multivariate analysis. *Biometrika* 53:325–338.

GROBLER, J. 2014. State formation and strife, 1850-1900, book section 7. Protea Book House, Pretoria, South Africa, first edition.

HOGG, R., MCKEAN, J., AND CRAIG, A. 2013. Introduction to Mathematical Statistics. Pearson, Boston, seventh edition.

HUNLEY, K. AND HEALY, M. 2011. The impact of founder effects, gene flow, and european admixture on native american genetic diversity. *Am J Phys Anthropol* 146:530–8. Hunley, Keith Healy, Meghan eng Research Support, U.S. Gov't, Non-P.H.S. 9/14/2011

6:00 Am J Phys Anthropol. 2011 Dec;146(4):530-8. doi: 10.1002/ajpa.21506. Epub 2011 Sep 13.

HUNLEY, K. L., HEALY, M. E., AND LONG, J. C. 2009. The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. *Am J Phys Anthropol* 139:35–46.

JOUBERT, J.-J. 2014. The troubled teens: South African democracy, 2004-2013, book section 24. Protea Book House, Pretoria, South Africa, first edition.

KEEGAN, T. 1996. Colonial South Africa and the Origins of the Racial Order. University of Virginia Press, Charlottesville, Virginia, first edition.

KUTNER, M., NACHTSHEIM, C., NETER, J., AND LI, W. 2005. Applied Linear Statistical Models. McGraw-Hill Irwin, New York, fifth edition.

LI, J., ABSHER, D., TANG, H., AM, S., CASTO, A., RAMACHANDRAN, S., CANN, H. M., BARSH, G. S., FELDMAN, M. W., CAVALLI-SFORZA, L. L., AND MYERS, R. 2008. Worldwide human relationships iferred from genome-wide patterns of variation. *Science* 319.

LIVI-BACCI, M. 2006. the depopulation of hispanic america after the conquest. *Population and Development* 32:199–232.

LONG, J. C. 1991. The genetic structure of admixed populations. *Genetics* 127:417–28. Long, J C eng Research Support, U.S. Gov't, P.H.S. 2/1/1991 Genetics. 1991 Feb;127(2):417-28.

LONG, J. C., LI, J., AND HEALY, M. E. 2009. Human dna sequences: more variation and less race. *Am J Phys Anthropol* 139:23–34. Long, Jeffrey C Li, Jie Healy, Meghan E eng Research Support, Non-U.S. Gov't 2/20/2009 9:00 Am J Phys Anthropol. 2009 May;139(1):23-34. doi: 10.1002/ajpa.21011.

MEYER, A. 2014. South Africa's primeval past, book section 1. Protea Book House, Pretoria, South Africa, first edition.

MORENO-ESTRADA, A., GIGNOUX, C. R., FERNANDEZ-LOPEZ, J. C., ZAKHARIA, F., SIKORA, M., CONTRERAS, A. V., ACUNA-ALONZO, V., SANDOVAL, K., ENG, C., ROMERO-HIDALGO, S., ORTIZ-TELLO, P., ROBLES, V., KENNY, E. E., NUNO-ARANA, I., BARQUERA-LOZANO, R., MACIN-PEREZ, G., GRANADOS-ARRIOLA, J., HUNTSMAN, S., GALANTER, J. M., VIA, M., FORD, J. G., CHAPELA, R., RODRIGUEZ-CINTRON, W., RODRIGUEZ-SANTANA, J. R., ROMIEU, I., SIENRA-MONGE, J. J., DEL RIO NAVARRO, B., LONDON, S. J., RUIZ-LINARES, A., GARCIA-HERRERA, R., ESTRADA, K., HIDALGO-MIRANDA, A., JIMENEZ-SANCHEZ, G., CARNEVALE, A., SOBERON, X., CANIZALES-QUINTEROS, S., RANGEL-VILLALOBOS, H., SILVA-ZOLEZZI, I., BURCHARD, E. G., AND BUSTAMANTE, C. D. 2014. Human genetics. the genetics of mexico recapitulates native american substructure and affects biomedical traits. *Science* 344:1280–5.

NEI, M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* 70:3321–3323.

NEI, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

O'FALLON, B. AND FEHREN-SCHMITZ, L. 2011. Native americans experienced a strong population bottleneck coincident with european contact. *Proc Natl Acad Sci U S A* 108.

PARRA, E. J., KITTLES, R. A., ARGYROPOULOS, G., PFAFF, C. L., HIESTER, K., BONILLA, C., SYLVESTER, N., PARRISH-GAUSE, D., GARVEY, W. T., JIN, L., MCKEIGUE, P. M., KAMBOH, M. I., FERRELL, R. E., POLLITZER, W. S., AND SHRIVER, M. D. 2001. Ancestral proportions and admixture dynamics in geographically defined african americans living in south carolina. *Am J Phys Anthropol* 114:18–29.

PATTERSON, N., PETERSEN, D. C., VAN DER ROSS, R. E., SUDOYO, H., GLASHOFF, R. H., MARZUKI, S., REICH, D., AND HAYES, V. M. 2010. Genetic structure of a unique admixed population: implications for medical research. *Hum Mol Genet* 19:411–9.

PEMBERTON, T. J., DEGIORGIO, M., AND ROSENBERG, N. A. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. *G3 (Bethesda)* 3:891–907.

PRITCHARD, J. K., STEPHENS, M., AND DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959. 321VJ Times Cited:9542 Cited References Count:30.

R CORE TEAM 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RAMACHANDRAN, S., DESHPANDE, O., ROSEMAN, C. C., ROSENBERG, N. A., FELDMAN, M. W., AND CAVALLI-SFORZA, L. L. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proc Natl Acad Sci U S A* 102:15942–7.

REED, T. E. 1969. Caucasian genes in american negroes. *Science* 165:762–8.

ROSENBERG, N. A., MAHAJAN, S., GONZALEZ-QUEVEDO, C., BLUM, M. G., NINO-ROSALES, L., NINIS, V., DAS, P., HEGDE, M., MOLINARI, L., ZAPATA, G., WEBER, J. L., BELMONT, J. W., AND PATEL, P. I. 2006. Low levels of genetic divergence across geographically and linguistically diverse populations from india. *PLoS Genet* 2:e215.

ROSENBERG, N. A., PRITCHARD, J. K., WEBER, J. L., CANN, H. M., KIDD, K. K., ZHIVOTOVSKY, L. A., AND FELDMAN, M. W. 2002. Genetic structure of human populations. *Science* 298:2381–5.

SHELL, R. 2014. People of bondage, book section 3. Protea Book House, Pretoria, South Africa.

SHRIVER, M., BOERWINKLE, E., DEKA, R., AND FERRELL, R. E. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Molecular Biology and Evolution* 12:914–920.

SHRIVER, M. D., PARRA, E. J., DIOS, S., BONILLA, C., NORTON, H., JOVEL, C., PFAFF, C., JONES, C., MASSAC, A., CAMERON, N., BARON, A., JACKSON, T., ARGYROPOULOS, G., JIN, L., HOGGART, C. J., MCKEIGUE, P. M., AND KITTLES, R. A. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet* 112:387–99.

SOKAL, R. AND ROHLF, F. 2012. Biometry. W.H. Freeman and Company, New York, fourth edition.

TANG, H., PEN, J., WANG, P., AND RISCH, N. J. 2005. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology* 28:289–301.

THOMAS, C. 2014. Coloureds: a complex history, book section 25. Protea Book House, Pretoria, South Africa, first edition.

THOMPSON, L. 2014. A History of South Africa. Yale University Press, New Haven and London, fourth edition.

TISHKOFF, S. A. AND KIDD, K. K. 2004. Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* 36:S21–7.

TISHKOFF, S. A., REED, F. A., FRIEDLAENDER, F. R., EHRET, C., RANCIARO, A., FROMENT, A., HIRBO, J. B., AWOMOYI, A. A., BODO, J. M., DOUMBO, O., IBRAHIM, M., JUMA, A. T., KOTZE, M. J., LEMA, G., MOORE, J. H., MORTENSEN, H., NYAMBO, T. B., OMAR, S. A., POWELL, K., PRETORIUS, G. S., SMITH, M. W.,

THERA, M. A., WAMBEBE, C., WEBER, J. L., AND WILLIAMS, S. M. 2009. The genetic structure and history of africans and african americans. *Science* 324:1035–44.

VAHED, G. 2014. The Indians in South Africa, book section 26. Protea Book House, Pretoria, South Africa, first edition.

VISAGIE, J. 2014a. The emigration of the Voortrekkers into the interior, book section 6. Protea Book House, Pretoria, South Africa, first edition.

VISAGIE, J. 2014b. Migration of the societies north of the Gariep River, book section 5. Protea Book House, Pretoria, South Africa, first edition.

WANG, S., LEWIS, C. M., JAKOBSSON, M., RAMACHANDRAN, S., RAY, N., BEDOYA, G., ROJAS, W., PARRA, M. V., MOLINA, J. A., GALLO, C., MAZZOTTI, G., POLETTI, G., HILL, K., HURTADO, A. M., LABUDA, D., KLITZ, W., BARRANTES, R., BORTOLINI, M. C., SALZANO, F. M., PETZL-ERLER, M. L., TSUNETO, L. T., LLOP, E., ROTHHAMMER, F., EXCOFFIER, L., FELDMAN, M. W., ROSENBERG, N. A., AND RUIZ-LINARES, A. 2007. Genetic variation and population structure in native americans. *PLoS Genet* 3:e185.

WANG, S., RAY, N., ROJAS, W., PARRA, M. V., BEDOYA, G., GALLO, C., POLETTI, G., MAZZOTTI, G., HILL, K., HURTADO, A. M., CAMRENA, B., NICOLINI, H., KLITZ, W., BARRANTES, R., MOLINA, J. A., FREIMER, N. B., BORTOLINI, M. C., SALZANO, F. M., PETZL-ERLER, M. L., TSUNETO, L. T., DIPIERRI, J. E., ALFARO, E. L., BAILLIET, G., BIANCHI, N. O., LLOP, E., ROTHHAMMER, F., EXCOFFIER, L., AND RUIZ-LINARES, A. 2008. Geographic patterns of genome admixture in latin american mestizos. *PLoS Genet* 4:e1000037.

WRIGHT, S. 1951. The genetical structure of populations. *Annals of Eugenics* 15:323–354.