ANALYSES OF GENETIC DATA WITHIN AN INTERDISCIPLINARY FRAMEWORK TO INVESTIGATE RECENT HUMAN EVOLUTIONARY HISTORY AND COMPLEX DISEASE

By

AMY NON

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2010

To my parents

ACKNOWLEDGMENTS

I would like to first thank Professor Connie Mulligan, my Ph.D. advisor for her excellent guidance, support, and mentorship throughout my graduate career, and her sincere dedication to my academic success. I would also like to thank Professor Lance Gravlee for inspiring me to pursue research on health disparities, and sharing his data, statistical expertise, and the passion of his research with me. I would also like to thank Professors Steve Brandt and Marta Wayne for their early guidance as committee members. I thank current and former postdoctorate fellows and graduate students in the Mulligan lab, including Dr. Ryan Raaum, Dr. Viktor Cerny, Dr. Rachel Dvoskin, Dr. Marina Ascunce, Dr. Becca Gray, Dr. Drew Kitchen, Aida Miró, and Tamar Carter for their helpful insights, feedback, and support over the past 6 years. I also thank all of the tireless undergraduate students, especially Luisa Sanchez, Travis La Joie, Laurie Krass, and Sydney Stelmaszek, who have given much time and effort to assist with my dissertation projects. I profess deep gratitude to the Ethiopian and Yemenite individuals, Puerto Ricans, and members of the FBPP who all participated in studies that are part of this dissertation. I thank all my friends who have provided emotional support, including Cara Kimberg, Cristi Morrell, Deborah Herbstman, Nikki D'Errico, Jamie Waggoner, and Scott Payne. I also thank my family, and in particular, my parents and my sister, Melissa, for all of their support and encouragement of my academic pursuits. Finally, I thank my fiancé, Jorge Román, for his emotional (and statistical) support, and endless patience and understanding.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ANALYSES OF GENETIC DATA WITHIN AN INTERDISCIPLINARY FRAMEWORK TO INVESTIGATE RECENT HUMAN EVOLUTIONARY HISTORY AND COMPLEX DISEASE

By

Amy Non

August 2010

Chair: Connie Mulligan
Major: Anthropology

In this dissertation, I have integrated diverse interdisciplinary data with two aims: 1) to explore both evolutionary history in Eastern Africa and Yemen, and 2) to investigate health disparities in the complex disease of hypertension. In the first study, I generated mitochondrial DNA (mtDNA), and analyzed it in combination with historical and linguistic data to test various population histories in Yemenite and Ethiopian Jewish populations. I discovered that Yemenite Jewish mtDNA diversity reflects potential descent from ancient Israeli exiles as well as evidence of African and Middle Eastern ancestry, with little evidence for large-scale conversion of local Yemeni, while the mtDNA of the Ethiopian Jewish population suggests descent primarily through conversion of local Ethiopian women. In the second study, I analyzed a large dataset of mtDNA from Yemen alongside comparative samples from Eastern Africa, the Middle East, and India to gain insight into migrations of modern humans out of Africa and settlement patterns throughout Yemen. In the third study, I integrated ethnographic, biological, and genetic data on a southeastern Puerto Rican population to explore biological and cultural contributors to hypertension. I discovered that the sociocultural

variable of *color* classification (in an interaction with socioeconomic status), was more predictive of blood pressure (BP) than an estimate of genetic ancestry. Through integration of sociocultural data, I also revealed a new and significant association between a candidate gene polymorphism for hypertension and blood pressure. Finally, I examined a large epidemiological dataset to test hypotheses about biological and social dimensions of race that contribute to disparities in hypertension. I discovered that education level was significantly associated with decreased BP overall, and that education associates with BP differently for each racial (and gender) group.

This dissertation has integrated data from many diverse fields, including archaeology, history, linguistics, geography, medical anthropology, and public health. This broad synthesis of interdisciplinary data enabled a comprehensive investigation of the history of each population and of the causal factors contributing to racial disparities in disease. Thus, this dissertation demonstrates the value of a multidisciplinary perspective in permitting comprehensive analyses of complex research questions, and generating new avenues for future research.

CHAPTER 1
INTRODUCTION

Genetic variation in human populations throughout the world reflects a complex evolutionary process of hundreds of thousands of years of mutation, genetic drift, gene flow, and natural selection. Analysis of contemporary genetic diversity provides insight into both human history and human disease, two primary goals of molecular anthropology. These goals are complementary, as exposure of human populations to varied environments, including diverse pathogens, climates, and diets, among other selective pressures, has influenced migration patterns and fluctuations in population size, which ultimately affect the global distribution and diversity of human populations as well as the current distribution of human diseases (Balaresque et al. 2007). Thus, a simultaneous focus on both population history and disease enables a deeper understanding of the many processes that contribute to the genetic diversity of human populations. This diversity has typically been analyzed from two complementary perspectives: neutral and non-neutral variation. Neutral variation relates to genetic markers that do not encode proteins, and thus are free from selective pressures. Non-neutral variation applies to genetic markers subject to selection, such that they may be influenced by the surrounding environment. Adaptive variants may be beneficial to human health if they contribute to a better fit with the local environment, as in the case of lactose tolerance (Tishkoff et al. 2007b) or immune-defense against pathogens (Huang and Yang 2010). Alternatively, these variants may act in maladaptive ways that contribute to disease, such as salt retention in the thrifty gene hypothesis (Young et al. 2005).

The neutral markers typically used to infer human demographic history in early anthropological genetics included haploid uniparentally inherited systems of mitochondrial DNA (mtDNA) (e.g. Watson et al. 1996), which reflects maternal history, and the Y-chromosome, which is passed only through the male lineage (e.g. Hammer et al. 2001; Underhill et al. 2000). With the advance of genetic technology, researchers have increasingly typed larger datasets utilizing nuclear markers that capture the history of multiple genes from both male and female lineages. These include Alu elements (e.g. Watkins et al. 2003), microsatellites such as autosomal short tandem repeats (STRs) (e.g. Rosenberg et al. 2002), large regions of non-coding sequence data selected from throughout the nuclear genome (Gutenkunst et al. 2009; Patin et al. 2009), and extremely large sets of SNPs genotyped throughout the entire genome (Biswas et al. 2009; Novembre et al. 2008). In general, all of these neutral markers are useful for studies of evolutionary history, as any genetic variation within them can be attributed to demographic processes such as genetic drift, gene flow, bottlenecks, or founder effects, and are typically assumed to be free from selection. Non-neutral variation, conversely, is usually investigated with markers located within genes that code for proteins, which can be located throughout the nuclear genome, or in the mitochondrial genome as well (e.g. climatic studies of adaptation in mitochondrial genes (Elson et al. 2007; Wallace et al. 2003), or mtDNA implicated in disease (Chinnery et al. 2010). These “genic” markers are most often used in studies of adaptation and disease.

The divide between the application of neutral and non-neutral markers, however, is not as sharp as it first appears. Sometimes, neutral variation can be useful for disease studies, as historical scenarios might have important genetic consequences

that can influence distribution of disease or neutral alleles, e.g. contact between Europeans and Native Americans can cause high levels of admixture in Hispanic populations, which can serve as a potential confounder if not controlled for within analyses of disease (Choudhry et al. 2006; Hoggart et al. 2003); or a history of severe bottlenecks and rapid expansions in Jewish populations, such as the Ashkenzi, has led to an accumulation of many Jewish-specific disease alleles (Atzmon et al. 2010). On the other hand, non-neutral variation can also inform us about the evolutionary history of a population. With the recent availability of large datasets of nuclear SNPs from throughout the genome, such as those collected in the HapMap project, genetic anthropologists have begun to take advantage of this wealth of data collected for disease research and apply a new set of anthropological questions to investigate evolutionary history (e.g. Novembre and Di Rienzo 2009). These large datasets (>300k snps) enable better estimates of ancestry, as well as the ability to test more refined questions of geographic structure. Even genes under the influence of selection may be useful to inform us about our evolutionary past, as adaptation to selective pressures may change the demographic trajectory of a population, enabling population expansions and colonizations if well-adapted, or to contract or relocate if subject to harmful environmental exposures.

In this dissertation, I analyze a combination of neutral and non-neutral genetic variation to answer questions related to two key areas of anthropological investigation: human evolutionary history and health disparities in human disease. The historical projects in Chapters 2 and 3 investigate evolutionary processes and relationships among human populations in a key region of human migration and dispersals, the Horn

of Africa and Arabian Peninsula. These analyses are primarily conducted on neutral genetic variants in the control region of the mitochondria, which does not encode for proteins and thus is free from selective pressures. The two disease projects utilize a combination of neutral and non-neutral markers to investigate complex disease, using hypertension as a model complex disease influenced by both genetic and environmental/sociocultural exposures. Specifically, I investigate non-neutral variants located in candidate genes for hypertension in stress-signaling pathways. I also utilize neutral variation in SNPs located throughout the nuclear genome in an admixture analyses to control for population stratification, and to investigate the role of African ancestry in contributing to health disparities in hypertension.

Throughout all of my analyses, human genetic diversity is the primary unit of analysis. However, genetic diversity does not exist in a vacuum, but is clearly a dynamic and ever-changing product of hundreds of thousands of years of evolutionary history in combination with exposure to all levels of environmental influence. For this reason, in each of my projects I integrate data from numerous interdisciplinary fields including archaeology, history, linguistics, geography, medical anthropology, and public health. This synthesis of data from across such diverse disciplines enables a more comprehensive investigation than any single data type can provide, both of the history of each population and of the true causal factors contributing to racial disparities in disease. For evolutionary history studies in particular, any type of data analyzed in isolation is subject to misinterpretation, e.g. the history of a genetic locus does not necessarily correlate with the history of a population, i.e. “most human alleles are older than the populations in which they occur” (Barbujani 2000). Thus incorporation of

external data either directly into simulations (e.g. Kitchen et al. 2008), or as complementary data in empirical analyses (e.g. Chapters 2 and 3), can improve our ability to test competing hypotheses about evolutionary history. Interdisciplinary data are equally important in studies of disease, where environmental exposures can: a) alternatively explain disease associations (e.g. Chapters 4 AND 5), b) confound genetic analyses if not adjusted for with rigorous environmental measures (e.g. Chapter 4), c) interact with genes, i.e. certain alleles may only contribute to disease under specific environmental conditions, or d) alter gene expression through epigenetic mechanisms that are only beginning to be understood. In the remaining portions of this introduction, I will explain how genetic data can take on new meanings and interpretations when investigated within the context of these interdisciplinary frameworks.

Human Evolutionary History

The two evolutionary history projects in my dissertation integrate genetic, historical, linguistic, archaeological, and geographic data to explore the history of populations in Eastern Africa and the Arabian Peninsula at two different population scales. The first project focuses on small, regional, and relatively isolated Jewish populations in Ethiopia and Yemen (Chapter 2). In this study, mtDNA were generated for Yemenite (n=45) and Ethiopian (n=41) Jewish individuals, and compared with neighboring mtDNA of non-Jewish Yemeni (n=50) and Ethiopian (previously published data, n=186) populations. These genetic data were analyzed through comparison of haplogroup distributions, phylogenetic networks and simulations. The mitochondrial diversity of the Yemenite Jewish population suggests potential descent from ancient Israeli exiles and shared African and Middle Eastern ancestry with little evidence for large-scale conversion of local Yemeni. In contrast, the Ethiopian Jewish population

appears to be a subset of the larger Ethiopian population suggesting descent primarily through conversion of local women. However, the genetic analyses alone provide an incomplete picture without a context of historical and linguistic data, which were necessary to reconstruct a comprehensive population history for each Jewish population. For example, oral histories, Biblical texts, and historians present alternative scenarios of the founding of each Jewish population, either concluding descent from ancient Israelites or conversion from local non-Jewish neighboring populations. In some cases, the mtDNA data alone could not confidently reject certain scenarios, as the control region of the mitochondrial genome evolves too slowly to resolve very recent timescales, i.e. on average one mutation in the HVRI every ~15,000 years, assuming an HVRI mutation rate of 1.64×10^{-7} substitutions/nucleotide/year (Soares et al. 2009). In my study, however, I was able to compare genetic data of the Jewish populations to the diversity found in non-Jewish neighbors, and thus show that the genetic data supported one out of two possible scenarios for the origin of Yemenite Jews. For Ethiopian Jews, the genetic data were consistent with two out of three historical scenarios, but I was unable to choose between the two even after incorporating linguistic data, e.g. Hebrew words in the Ge'ez version of the Ethiopic Old Testament. Historical data was also drawn upon to refine the potential time frames that could be consistent with the genetic data, e.g. 13th Century accounts by Marco Polo of an ancient Jewish presence in Ethiopia.

My second evolutionary history project examines a larger continental-level scale across wider geographic space throughout Eastern Africa, the Arabian Peninsula, the Middle East, and India (Chapter 3). This study analyzes mitochondrial genetic diversity

of Yemen and surrounding regions to search for patterns of migration or settlement in different regions of Yemen, as well as larger-scale patterns of migration out of Africa through the Greater Red Sea Region. These analyses would not have been possible without the fine-scale geographic sampling of the Yemeni population, or the detailed Geographic Information System (GIS) data (i.e. latitude and longitude coordinates on each sample) needed to test for geographic structure in the genetic ancestry of Yemen and surrounding countries. Specifically, in this project, analyses were conducted on a large sample of 521 Yemeni from throughout every governorate in the country, along with ~500 additional mtDNA sequences from Eastern Africa, the Middle East, and India, collected from published databases. Additional analyses were conducted on 50 individuals with data from 27 nuclear Alu markers and 19 Y chromosome SNPs. On all datasets, I tested for genetic and geographic correlation using the spatial auto-correlation analysis (i.e. correlograms and Moran's I index). I also used hierarchical AMOVA to distinguish distinct Hg clusters and a geostatistical analysis of interpolated contour maps with inverse distance weighting analyses. The incorporation of a broad geographic context enabled me to address issues related to: 1) the place of Yemeni genetic diversity in migrations out of Africa, 2) ancient and/or more recent settlement patterns throughout regions of Yemen, 3) slave trade in Yemen from Eastern Africa, Europe, or SE Asia, and 4) the Hadrami Diaspora to Asia. These results have significant implications for other evolutionary history research throughout the world, by demonstrating the value of examining genetic data within a detailed geographic framework.

Health Disparities

The second half of my dissertation addresses one of the most pressing health issues in biomedicine, medical anthropology and public health today: racial inequalities in health. The concept of race has had a long complex history in anthropology. The stain of eugenics has tainted anthropological studies of race since the early 20th Century, when physical anthropologists maintained essentialist and hierarchical definitions of race. In recent decades, data from genetic anthropologists and other human geneticists have begun to weaken support for a biological basis of race, i.e. more genetic variation was found within populations than between, based on classical genetic markers (Lewontin 1972) and recently confirmed with microsatellite data from 377 loci throughout the human genome in 52 populations (Rosenberg et al. 2002). The debate over race has been revived in recent years in the biomedical community due to the large number of diseases that disproportionately afflict minority populations. In fear of reifying the race concept, many anthropologists have shied away from discussions of race. However, the growing problem of racial inequalities in health requires an explanation. Biological anthropologists who study human variation are well positioned to engage in this debate, and perhaps even have a moral obligation to do so, particularly because they are equipped with the tools and evolutionary perspective to counter racist arguments or racialized science. Interdisciplinary work is complicated by the fact that researchers from across disciplines, including geneticists, epidemiologists, clinicians, and anthropologists, do not conceptualize and use race in a congruent manner.

In the study of racial disparities in complex phenotypes (e.g. skin color, or immune system diversity), and complex disease in particular (e.g. asthma, diabetes,

cancer, cardiovascular disease [CVD]), the meaning and use of race has direct implications not only for interpreting results, but also for research design. These implications vary according to the field of study and the goals of research. For example, consideration of race is very important in epidemiological studies, where the goal is traditionally to track and monitor racial inequalities in health and in access to education and economic opportunities that affect health. Epidemiology and public health research thus depends on nation-wide census data on self-identified race. Alternatively, in biomedical studies, where the goal is more about understanding underlying causes of disease, racial categories are often used as a shortcut to summarize genetic background until more affordable individualized medicine is available. This shortcut can be dangerous, as it is based on the as-yet unvalidated assumption of genome-wide racial genetic differences that predict complex disease. This assumption often leads to research designs that confounds genetic and environmental factors, and often underestimates the role of the social environment (for discussion see Cooper et al. 2005; Cooper and Zhu 2001; Kaufman and Cooper 2001). There is also no evidence-based support for this assumption, as genetic variants associated with ancestry, or even with genes in functional pathways, have yet to explain “only a fraction of the inherited contribution to risk” in complex disease (McCarthy et al. 2008), and for CVD in particular, less than 10% of risk is yet explained by genetic variants (Mohlke et al. 2008). As for use of global estimates of genetic ancestry, there is yet no evidence that regional selection or isolation on a continental scale accounts for racial differences in susceptibility to complex disease in the same way we understand for Mendelian traits like sickle cell anemia (which is also not restricted to continentally-defined racial groups)

(Kaufman and Cooper 2001). Unfortunately, many biomedical researchers continue to design their research with these inherent racial assumptions without testing any specific functional loci or even testing other types of environmental data (see Non et al. 2010b in Appendix) or sometimes even in the absence of any genetic data (see Gravlee and Mulligan 2010). With too narrow a focus on genetic factors, it is impossible to distinguish the potential contributions of the sociocultural environment, or any interactions between genes and environment.

I address this limitation through two studies that examine both genetic and sociocultural contributors to racial health disparities, both in a Puerto Rican population and among African American participants of the Familial Blood Pressure Program (FBBP). Both of these projects empirically test and provide support for the central tenant of this dissertation: analyses based on a single data-type can have dramatically different conclusions when examined in the context of a broader interdisciplinary framework. By integrating the sociocultural data with genetic data in a single analysis, both of these studies are also better able to address the long-standing debate over the relative role of genetic and sociocultural variables in racial inequalities in health. These results have important implications for biomedical research, particularly in the post-human genome age where the rapid advancement of technologies have led to an abundance of genetic data often not well balanced by sociocultural or environmental data of equal rigor or sophistication.

The first health disparities project is a biocultural analysis of blood pressure variation in Puerto Rico (Chapter 4). This research extends upon previous research of Dr. Gravlee, where he established a link between blood pressure (BP) and the variable

color, an aspect of social classification, analogous to race in the U.S., which shapes everyday exposure to stressors. I conducted a genetic analysis on the same population to identify the relative contributions of genetic and sociocultural factors to disparities in BP. I generated genetic data on 89 individuals by assaying 78 ancestry informative markers and six BP candidate gene polymorphisms. I combined these genetic data with the sociocultural data, such as socioeconomic status (SES) and social classification of *color* to investigate three primary questions: 1) How is social classification based on *color* associated with individual genetic admixture? 2) Does genetic ancestry or *color* better predict BP? and 3) Does the inclusion of sociocultural data alter the association between BP and candidate gene polymorphisms for hypertension? This third question specifically tests the hypothesis that conclusions based on genetic data may change when examined in the context of data from other disciplines, as genetic associations may be confounded by important, and typically unmeasured, environmental variables. The conclusions show that in the absence of sociocultural data, genetic ancestry appears to be an important predictor of systolic blood pressure. However, when cultural classification of *color* is added to the regression model, in an interaction with socioeconomic status, it becomes clear that ancestry is less important, while the cultural classification is a strong and significant contributor to blood pressure variation in this Puerto Rican population.

The final project assesses the impact of education, self-identified race, and genetic ancestry on blood pressure in black and white Americans (Chapter 5). In this study, I analyze various blood pressure measures and BMI in African American populations from a large epidemiological database called the Familial Blood Pressure Program.

These data were previously analyzed in a more focused analysis (Tang et al. 2006), and their conclusions epitomize the problem of many health disparities studies which fail to incorporate environmental or sociocultural data into their analysis of complex phenotypes. Their study reports a marginal association between African genetic ancestry and BP, yet neglected to include any sociocultural variables, even though education data were available in the public database. In my analysis of these data I discovered that education is significantly associated with BP particularly in black Americans, and that the addition of education to the model reduces the association between ancestry and BP. Thus without any measure of the sociocultural environment, the genetic data suggested a relationship between ancestry and BP, but the addition of a measure as simple as education level has demonstrated that the sociocultural environment may be more statistically significant than genetic ancestry in contributing to racial disparities in hypertension. These results have important implications for research on racial inequalities, as they challenge previous reports of associations between genetic ancestry and health, which may in fact be due to sociocultural, and not genetic, differences between racial groups.

In sum, this dissertation demonstrates the importance of integrating data from multiple disciplines to gain a more comprehensive picture of both evolutionary history and to explain contemporary racial inequalities in complex disease. The inherently cross-fields nature of anthropology makes it an ideal and leading discipline to take on the type of interdisciplinary research represented by my dissertation. I have tried to show through these four projects the benefits gained from stretching beyond a single discipline to work with new types of data. Incorporation of multiple types of evidence

not only improves the quality of anthropological research, but in many cases is fundamental to accurately test or generate new hypotheses about evolutionary history and disease. Though we cannot all become experts in everything, the benefits gained from speaking the language of multiple disciplines can open the doors to collaborations, ideas, and research possibilities that far outweigh the costs in time and effort.

CHAPTER 2
MITOCHONDRIAL DNA REVEALS DISTINCT EVOLUTIONARY HISTORIES FOR JEWISH POPULATIONS IN YEMEN AND ETHIOPIA[1]

Introduction

It is generally believed that anatomically modern humans first emerged out of Africa through the Horn of Africa (HOA) (Lahr and Foley 1998; Stringer 2000). Human populations have since migrated back and forth across the Red Sea making Arabia and the HOA important regions for the study of human evolutionary history. As far back as the 7th millennium before the common era (BCE), archaeological evidence attests to a complex history of population movements in the Red Sea region, such as the existence of obsidian exchange networks between populations across the Red Sea (Fattovich 1997a). By the first millennium BCE, westward movement of Arabians to Africa led to an Ethio-Sabean state on the Tigrean Plateau, followed by the eastward expansion of the Ethiopian Aksumite kingdom along the northern Indian Ocean in the 1st millennium common era (CE) (Fattovich 1997a; Munro-Hay 1991). Over the past 1000 years, the Islamic commercial and political expansion brought migrants from Arabia back to East Africa (Fattovich 1997a). Linguistic evidence also supports the high mobility of populations throughout this region, as Semitic is the only language within the Afro-Asiatic language family that is spoken both within and outside of Africa (Greenberg 1971). Analysis of Semitic-speaking Jewish groups provides a glimpse into the more recent history, i.e. past several thousand years, of populations in this area. In the Red

---

[1] Non, AL, Al-Meeri A., Raaum RL, Sanchez LF, Mulligan CJ. Mitochondrial DNA reveals distinct evolutionary histories for Jewish populations in Yemen and Ethiopia. *Am Journal Phys Anthro* (in press).

Sea region, as throughout the Jewish Diaspora, this history is often characterized by marginalization, exile, and migration, which may leave distinct population genetic signatures. A wealth of oral and Biblical histories on Jewish populations in the Red Sea area, along with archaeological and linguistic data, provide a rich context within which genetic data can be analyzed to construct a multi-faceted history for Jewish populations in both Yemen and Ethiopia.

Numerous competing theories exist to explain the origins of both Yemenite and Ethiopian Jewish populations. The presence of Jewish populations in Yemen is thought to extend back at least 2500 years before present (YBP), though historical and archaeological evidence of their origin is scarce (Ahroni 1986; Blady 2000).Two of the main alternative histories for Yemenite Jews are: 1) descent from ancient exiles from the Kingdom of Israel, followed by additional waves of migration over the past 2500 years, and 2) historic conversion of many Yemeni people to the Jewish religion during the reign of the Himyarite Kingdom in the 5th and 6th centuries CE. The conversion theory is supported by some historians who claim that Jews in Yemen were actively proselytizing in the region throughout the first half of the first millennium CE, successfully bringing entire polytheistic tribes into the Jewish faith (Blady 2000).

Controversy also surrounds the origin of the Ethiopian Jewish population, who self-identify as the Beta Israel. The numerous perspectives of modern scholars, rabbis, and the Beta Israel themselves can be roughly summarized by three primary historical scenarios: 1) descent from ancient Jewish exiles from the Kingdom of Israel between the 8th and 6th centuries BCE, 2) descent from Ethiopian converts to Judaism in the 4th century CE, and 3) descent from Ethiopian Orthodox Christians who adopted Jewish

Biblical practices in 14th-16th centuries CE. In support of the first scenario, the antiquity of the Jewish presence (and isolation) in Ethiopia is indicated by their strict pre-Talmudic Biblical Judaism, uninfluenced by Talmudic laws of the past 2000 years (Blady 2000). In contrast to the direct Israeli descent theory, the second scenario views Ethiopian Jews as descendents of Ethiopian (possibly Agaw) converts to Judaism who refused to accept Christianity in the 4th century (Blady 2000). The third theory states that Ethiopian Jews descend from Ethiopian Christians who adopted Jewish Biblical practices as late as the 14th to 16th centuries (Kaplan 1992; Quirin 1992). This view is supported by the similarity of Ethiopian Jewish religious practices (e.g. observance of the Sabbath, circumcision eight days after birth, food prescriptions, etc. (Quirin 1992) with those of Ethiopian Orthodox Christianity.

In this study, we perform a regionally focused genetic analysis of Yemenite (n=45) and Ethiopian (n=41) Jewish populations as well as neighboring, non-Jewish populations (n=50 new Yemeni and n=186 previously published Ethiopian Semitic-speakers). We generate mitochondrial hypervariable region I (HVR-I) sequence data as well as several diagnostic coding variant haplotypes and compare each population's genetic diversity using haplogroup distributions and phylogenetic networks. Identification of a rare HVR-I variant at high frequency in the Ethiopian Jews was further investigated by generating coding genome sequence data (n=9) and development of a unique simulation method to estimate the origin of this variant and the female effective population size of Ethiopian Jews. Finally, we integrate these genetic data with other anthropological data to investigate competing origin theories for Yemenite and Ethiopian Jewish populations.

Materials and Methods

Samples and Genotyping

DNA from the Jewish populations (n=41 Ethiopian Jews and n=45 Yemenite Jews) was obtained from blood samples donated to the National Laboratory for the Genetics of Israeli Populations at Tel-Aviv University (http://nlgip.tau.ac.il/). Each of the donors reported the birthplace of their four grandparents as either Yemen or Ethiopia and the research protocol was approved by the University of Florida Institutional Review Board. Samples from 552 non-Jewish Yemeni were collected in Spring 2008. A subset of 50 samples was chosen for sequence analysis in the current study, of which all reported the birthplace of their four grandparents to be within Yemen. All major geographic areas and primary ethnic groups of mainland Yemen are represented in the 50 samples. Yemeni samples were collected with written informed consent under Western IRB approval, Olympia, WA, and the Jewish samples were collected with written informed consent under protocols approved by the Human Subjects Review Committee of Tel-Aviv University. Previously published mtDNA data on Semitic (n=186) and Cushitic-speaking (n=46) Ethiopian non-Jews (Kivisild et al. 2004) were also used for comparison.

Sequence data for hypervariable region I (HVRI; bps 16024-16381) were generated for the Yemenite Jewish individuals (GenBank accession numbers GQ863581-GQ863625), the Ethiopian Jewish individuals (GQ863635 -GQ863666; not including those individuals who were sequenced for the mitochondrial genome and are listed below) and the Yemeni individuals (GQ863667- GQ863716; includes HVRII sequence data that are not discussed here). HVRI segments were amplified and sequenced using the primer pair L15997/H16401. Variation in the cytosine

homopolymer associated with mutation 16189C (bps 16182 to 16193) was removed prior to all analyses. A subset of nine Ethiopian Jews who shared identical HVRI regions were further sequenced for the rest of the control region (bps 1-576) and ~6800bps of coding region DNA (bps 577-2632; 11337-16023), comprising ~41% of the complete mitochondrial genome (GenBank accession numbers GQ863626-GQ863634). The nine DNAs were amplified with 26 primers to create 13 overlapping fragments, which were then sequenced with the same 26 primers, achieving on average ~1.5x coverage. Diagnostic coding region restriction fragment length polymorphisms (RFLPs) were assayed for a subset of samples in order to further refine haplotype classification, following the nomenclature of (Behar et al. 2008a; Santos et al. 2004; Torroni et al. 1996); *Alu*I -7025 (defines H), *Alu*I +10397 (defines M), *Hpa*I +3592 (defines L3/4/7 vs. other L), *Hinf*I +12308 (defines U), *Mnl*I +10871 (defines N), *Hae*II -9052 (defines K), +15606 *Alu*I (defines T), +10394 *Dde*I (distinguishes J/T), along with two additional coding region sites (guanines at bps 769 and 1018, which generally distinguish L3 versus L4) that were assayed via direct sequencing in two individuals (see Table 2-1).

Phylogenetic Networks

Phylogenetic networks were created using the HVRI data from both Jewish and non-Jewish populations. Networks were generated using the median-joining algorithm in the Network v.4.5.1 program (Bandelt et al. 1999). The maximum parsimony (MP) algorithm in Network v.4.5.1 was applied to the Ethiopian Jewish and non-Jewish network to reduce reticulations and find the shortest weighted tree (Steiner network).

Probability of Zero Mutations

The probability of accumulating zero coding region mutations across nine individuals was calculated using the formula $e^{-9\mu t}$. This formula was derived from the

probability of a single mutation arising by time t in a single sequence, which is represented by $1-e^{-\mu t}$. Thus, the probability of a single mutation not occurring is $e^{-\mu t}$ and the probability of zero mutations in nine individuals is $e^{-9\mu t}$. Using two coding region mutation rate estimates of 1.7×10^{-8} mutations/site/year (Ingman et al. 2000) and 1.26×10^{-8} mutations/site/year (e.g. Endicott et al. 2009; Soares et al. 2009) across ~6800 coding base pairs, the probability of zero mutations occurring in any of the nine sequences was calculated for 10 different time periods ranging from 0 to 7500 YBP. In light of recent investigations of mutation rates (e.g. Endicott et al. 2009; Soares et al. 2009) that independently estimate rates that are both substantially faster than Mishmar et al's (2003) rate, we believe that the faster rate of Ingman et al. (2000) is likely to be more accurate.

Simulations to Estimate the Age of the 16305(T) Variant

In order to estimate an age for the 16305(T) variant in the Ethiopian Jewish population, we developed a novel simulation analysis based on the unique pattern of nine identical mitochondrial genome sequences in a sample of 41 individuals. The first step was to obtain an accurate estimate of the female effective population size, N_{ef}, of Ethiopian Jews. Simcoal2 (Laval and Excoffier 2004) was used to generate DNA sequence data comparable to our dataset of control and coding region DNA sequence. Fifteen N_{ef} values were tested, ranging from 50 to 7,000 individuals. At each value of N_{ef}, the frequency of simulations that created a dataset of nine identical sequences in a sample of 41 was fit to a gamma distribution. The mode of the distribution was chosen as the most representative estimate of N_{ef}, while the upper and lower 95% confidence intervals provided a range of likely N_{ef} values.

These estimates of N_{ef} were then used to calculate the age of the variant in a simulation involving a three-part chromosome structure, in which the first two parts represented the empirical control and coding region and the third part was a virtual DNA segment assigned a fast mutation rate in order to generate sufficient variation for our estimates. The segments were: 1) 1000 bps of control region mtDNA with a standard control region mutation rate of $4.7x10^{7}$ mut/site/year (Howell et al. 2003), 2) 6800bps of coding region mtDNA with a standard coding region rate of $1.7x10^{-8}$ mut/site/year (Ingman et al. 2000), and 3) 2200bps of 'fast-evolving' DNA with a mutation rate 100x that of the control region (as explained above, we have chosen to focus on the mutation rates estimated by Howell et al. 2003 and Ingman et al. 2000 since faster rates such as these have been supported by recent, independent investigations of mutation rates, e.g. Endicott et al. 2009, Soares et al. 2009). A sample of 41 sequences was simulated 10,000 times, again using the program Simcoal2 (Laval and Excoffier 2004). Simulated sequence sets were selected that matched the observed pattern of control region mutations, i.e. nine identical individuals in both control and coding region sections. Of those selected sequence sets, maximum pairwise differences between individuals were calculated on the third section of fast-evolving sequence data and these values were used to estimate the age of the variant in the Jewish population. We chose maximum pairwise differences to reflect the deepest nodes in the phylogeny, which are most representative of the origin of variants in the population. Simulations using several N_{ef} values (mode, mean, and upper and lower 95% confidence interval) were used to calculate the age of the variant, based on the following formula, $\frac{1}{2}(n/\mu*l)$, where n is the maximum number of pairwise differences, l is the length of the DNA sequence, and u is

the mutation rate. Slower mutation rates were also tested for the third "fast-evolving" segment of DNA (again, using 10,000 simulations), which produced similar age estimates, but with wider confidence intervals as fewer datasets of exactly nine identical individuals were generated.

Results

Comparison Between Jewish Populations

A network of mitochondrial haplotypes was constructed to investigate the evolutionary relationship of the Yemenite and Ethiopian Jewish populations to each other (Figure 2-1). The network reveals a striking lack of identical haplotypes shared between the two populations. Complete lack of shared haplotypes has not been found before among other Jewish populations (with the exception of a previous study that found no haplotype sharing between Ethiopian Jews and other Jewish populations (Behar et al. 2008a), e.g. at least one haplotype is shared between Ashkenazi, Sephardic, North African, and Oriental Jewish populations in different regions (Behar et al. 2008a; Picornell et al. 2006). Thus, our result is surprising, especially considering the geographic proximity of the Yemenite and Ethiopian Jewish populations, and suggests unique maternal population histories for each group.

Mitochondrial DNA analysis also revealed a high diversity of sub-Saharan African and Eurasian haplotypes in both the Yemenite and Ethiopian Jewish populations (Figure 2-2). Specifically, common haplotypes (haplotypes present at >5%) in Yemenite Jews include the African haplogroup L3x1 and Eurasian haplogroups R0a (renamed from (preHV)1 (Torroni et al. 2006), HV1, J2a1a (renamed from J1b (Palanichamy et al. 2004)) K, R2, U, and U1, and in Ethiopian Jews include African haplogroups L2a1b2 and L5a1 and Eurasian haplogroups R0a and M1a1 (Figure 2-2). Overall, sub-Saharan

African L haplotypes (hereafter referred to as L(xM,N), i.e. all African haplotypes except M and N, following the nomenclature of Behar et al. (2008a), comprise a large proportion of the genetic variation in both Jewish populations, representing ~20% in the Yemenite Jews and ~50% in Ethiopian Jews. This high frequency contrasts with other Jewish populations, such as Near Eastern and Ashkenazi Jews, who almost entirely lack L(xM,N) haplogroups (Richards et al. 2003; Thomas et al. 2002).

Along with many L haplotypes, a very high frequency of Eurasian haplogroup R0a was also identified in both Yemenite Jews (11%) and Ethiopian Jews (22%) (Figs. 1-2). In the Yemenite Jewish population, five individuals share four different R0a haplotypes demonstrating fairly high R0a diversity in this population. In contrast, all of the nine R0a individuals in the Ethiopian Jewish population share an identical haplotype (the nine samples likely represent at least three different families). In addition to the HVRI defining variants at 16126 and 16362, this high frequency haplotype contains a unique transversion at base pair 16305 that has been found only in 10 Ethiopian individuals (Kivisild et al. 2004; Watson et al. 1996) out of >20,000 HVR1 sequences worldwide (as compiled in Kivisild et al. 2004); thus, the 16305(T) transversion may provide insight into the history of the Ethiopian Jewish population.

Comparison between Jewish Populations and non-Jewish Neighbors

Haplogroup distributions and haplotype networks: Yemenite Jews and non-Jews. The mitochondrial haplogroup distributions of Yemenite Jews were compared to those of their non-Jewish neighbors (Figures 2-3 and 2-4). The frequency of sub-Saharan African L(xM,N) haplotypes is higher in the non-Jewish population (30%) than in the Jews (20%) (Figure 2-2). A previous study of 115 non-Jewish Yemeni found even higher levels of L(xM,N) haplotypes (48%, Kivisild et al. 2004) although those samples

were collected in Kuwait, which likely produced an over-sampling of individuals of African ancestry as Yemeni living in Kuwait are predominantly Akhdam (a low status Yemeni population of African ancestry) or from Hadrawmat (a region known to have higher African ancestry (Cerny et al. 2008). We believe that our expansive sampling of Yemen provides a better representation of non-Jewish Yemeni genetic diversity than any previously collected sample and, thus, we focus subsequent analyses on our Yemeni samples. The diversity of L(xM,N) haplotypes is also reduced in the Yemenite Jewish population as compared with non-Jewish Yemeni, with a complete lack of deeper L0-L2 haplotypes that are common in the non-Jewish Yemeni population. A previous study of Yemenite Jews found a similar frequency of L(xM,N) (18%) and similar lack of L0-L2 haplotypes (Behar et al. 2008a). Though reduced in comparison to Yemenis, the frequency and diversity of L(xM,N) haplotypes in Yemenite Jews is still much higher than typically seen in other Jewish populations.

The Yemenite Jewish population also contains a high frequency of west Eurasian haplogroups, including most of the major subclades of R (~67%) as well as non-R haplogroups N and W (~9%), in addition to haplogroup M (~4%) that is found both in and out of Africa. Some of the Eurasian haplogroups in the Yemenite Jews are entirely absent in the non-Jewish Yemeni population, i.e. HV1, I1, J2a1a, U, U1 and W, suggesting some non-Yemeni Eurasian maternal contribution to the Jewish population. The R subclades, e.g. HV, J, U, are found at high frequencies further north in the Middle East/Caucasus region; e.g. Saudi Arabia (70.5%), Jordan (77%), Palestine (74%), Syria (87%), Bedouin (69%), Iraq (87%) and Iran (80%) (compiled in (Abu-Amero et al. 2007), indicating a potential source for R variation in the Yemenite Jews.

A network analysis was used to represent the exact sequence differences between haplotypes and illustrate the evolutionary relationship of Jewish and non-Jewish Yemeni (Figure 2-3). The two populations are clearly different, as only two of the haplotypes in the Jewish population are identical to haplotypes in the non-Jewish population (2/30, 6.7%). On average, each Jewish individual's haplotype is ~2.1 mutational steps away from a non-Jewish Yemeni haplotype, with a few individuals as many as four or six mutations away. When comparisons are made that include the larger Yemeni dataset (n=115) collected by Kivisild et al. (2004) in Kuwait and the larger Yemenite Jewish sample (n=119) collected by Behar et al. (2008), 19.3% (11/57) of Jewish haplotypes are identical to a haplotype in the non-Jewish population.

Haplogroup distributions and haplotype networks: Ethiopian Jews and non-Jews. Next, Ethiopian Jewish individuals were compared with non-Jewish Semitic-speaking Ethiopians (Kivisild et al. 2004) (Figure 2-2, Figure 2-4). Examination of the haplogroup distribution reveals Ethiopian Jews to be a potential subset of the greater Ethiopian population (Figure 2-2). With one exception (L3b2), no unique haplogroup was identified in the Ethiopian Jewish population that was not also present in the non-Jewish population. The Ethiopian Jewish proportion of L(xM,N) haplotypes is nearly identical to that found in Semitic-speaking Ethiopians, i.e. just over 50% in both populations. Ethiopian Jewish and non-Jewish haplotypes show a great deal of overlap, with a high frequency of Jewish haplotypes found in the non-Jewish population (9/27, 33.3%). Furthermore, each Ethiopian Jewish haplotype is on average only 0.88 mutations away from a non-Jewish Ethiopian haplotype, suggesting a close and recent relationship between the two populations. In contrast, the Ethiopian Jewish population

shows very low mitochondrial similarity to other Jewish populations, with no haplotypes identical to any found in other Jewish populations (Behar et al. 2008a; Thomas et al. 2002).

Detailed Investigation of Rare 16305(T) Variant in Ethiopian Jews

Coding region sequence analysis. The nine individuals with identical HVRI sequences who carry the rare transversion were drawn from at least three independent sample collections, and thus are unlikely to share very recent (i.e. within the past 50 years) descent, particularly given that all males contained unique Y chromosome haplotypes across nine typed microsatellite markers (data not shown). In the event that these individuals are more related, our analyses would underestimate the age of the origin of the haplogroup. To more fully investigate these individuals, we generated HVRII and coding region sequences (bps 1-2632 and 11337-16023) for all nine individuals. Previous research indicates that approximately half of the mitochondrial genome, including the HVR, carries sufficient variation for accurate coalescent and phylogenetic analyses (Non et al. 2007). All nine individuals were found to share an identical mitochondrial genome sequence, classified as R0a1b that is defined by coding region mutations 2355 and 15674 and HVRII mutation 57i (Abu-Amero et al. 2007). The nine individuals represent a unique R0a1b haplotype based on the 16305(T) variant that has not been identified previously except in one Ethiopian Jewish individual (Behar et al. 2008a); ten non-Jewish Ethiopian individuals with 16305(T) have identical HVRI sequences to the nine Ethiopian Jews, but were not assayed for the additional coding region variants (Kivisild et al. 2004; Watson et al. 1996). Other R0a1b haplotypes are represented by single individuals broadly distributed across the HOA, Arabian Peninsula, Iberian Peninsula, and South Asia (Abu-Amero et al. 2007; Behar et al.

2006). In sum, these results indicate a recent and local origin for this new haplotype of R0a1b that carries 16305(T).

Age estimation of 16305(T) in Ethiopian Jews. In comparison to the surrounding non-Jewish population, Ethiopian Jews carry an elevated frequency of the 16305(T) variant. In order to identify a plausible timeframe for the origin of this variant in the Ethiopian Jewish population, we perform a simple analysis to determine the range of years for which it is reasonable to expect zero mutations to accumulate in nine independent mitochondrial sequences. The probability of zero mutations is represented in Figure 2-5. Although the highest likelihood for zero mutations is at zero years, our analysis demonstrates that the origin of the R0a1b variant in the Ethiopian Jewish population could range anywhere from 0-2800 years ago (based on 1.7×10^{-8} mutations/site/year, Ingman et al. 2000) or 0-3800 years ago (1.26×10^{-8} mutations/site/year, Mishmar et al. 2003), at which point the likelihood of experiencing zero mutations in nine sequences becomes very small, i.e. <5%. A previous study of this haplogroup estimated its age between 0-7091 years (Behar et al. 2008a), consistent with our calculations but with a fairly broad range.

We also used a Bayesian technique to more specifically estimate the age of the 16305(T) variant in the Ethiopian Jewish population. Tavaré et al.(1997) devised the equation $\mathrm{E}(T_n/D) = \sum_{j=2}^{n} 2/j(j+\theta-1)$, which provides an exact solution for the time to the most recent common ancestor in cases where no genetic variation is present in the assayed sample. Combining data from the control and coding regions of the nine identical 16305(T) sequences, we calculate a theta of 6.56, which indicates an age of 2160 years for the 16305(T) variant in the Ethiopian Jewish population.

Next, we developed a novel simulation method to refine our estimation of the age of the 16305(T) variant in Ethiopian Jews. To generate the most accurate simulation, we first needed an estimate of the female effective population size, N_{ef}, of Ethiopian Jews. Simulations using Simcoal2 (Laval and Excoffier 2004) were run over 15 N_{ef} values, ranging from 50 to 7000. Results were fit to a gamma distribution that showed the frequency with which a simulated sample of 41 individuals carried nine identical sequences (Figure 2-6). The resultant N_{ef} plot demonstrated a mode of 224 effective females, with a mean of 364 and 95% quartiles of 64 and 939. Though small, these values are similar to estimates of other founding populations, such as the Etruscans with an N_{ef} of 770 women (Belle et al. 2006). Using this empirically-derived range of estimates for N_{ef}, we implemented our simulation technique that was designed to provide sufficient variation in order to obtain a more precise estimate of the origin of the 16305(T) variant. Specifically, we created an artificial three-part mitochondrial genome, in which the first two parts represent the empirical control and coding regions and the third part is a virtual DNA segment assigned a fast mutation rate 100x that of the control region. The N_{ef} estimates generated above were then used in simulations with this three-part chromosome structure to estimate the age of the 16305(T) variant (Figure 2-7). The age estimates increase with population size as larger populations require longer times for sequences to coalesce. The peak of the N_{ef} distribution at 224 (Figure 2-6) indicates the largest number of simulations that matched our dataset and thus presents the best estimate of effective population size; simulations using this value estimate an age of 1004+/-27 YBP for the origin of 16305(T) (Figure 2-7). The lower and upper 95% quartile values of N_{ef} provide age estimates ranging from ~420 to ~1690 YBP. These

ages are consistent with our calculation of the likelihood of observing nine identical sequences (see above).

Discussion

Jewish Diaspora populations are some of the most studied populations in the world due to their successful migration and colonization of multiple global regions, their unique history of population isolation, and the wealth of archaeological, linguistic, and genetic data available. We focus on two Jewish groups at the geographic crossroads of some of the most significant migrations in human history, including the first migration of modern humans out of Africa. We compare mitochondrial diversity between these Jewish populations and surrounding groups in order to better understand their maternal population histories and the broader history of the region itself. We integrate our data with other anthropological data to investigate alternate origin and migration histories.

Our data show that Yemenite and Ethiopian Jewish populations differ from other Jewish Diaspora populations due to a high frequency of sub-Saharan African L(xM,N) mitochondrial haplotypes, i.e. ~20% in Yemenite Jews and ~50% in Ethiopian Jews. A higher frequency of L(xM,N) haplogroups could be explained by proximity to Africa except for the fact that all North African Jewish populations (n= 33 Moroccan, Tunisian, and Libyan) carry only a single L haplotype and other Middle Eastern Jewish populations (n=23 Iranian and Iraqi) have no L haplotypes (Picornell et al. 2006). This high frequency of L(xM,N) haplogroups in both Yemenite and Ethiopian Jews suggests a relatively large number of African founders and/or a fair amount of admixture with African populations. A very high frequency of the Eurasian haplogroup R0a was also found in both Yemenite Jews (11%) and Ethiopian Jews (22%). However, despite the high frequency of L(xM,N) and R0a haplogroups in both Yemenite and Ethiopian Jewish

populations, no exact haplotype matches were identified between the two Jewish populations. This low level of haplotype sharing implies very little gene flow between the two populations, despite their geographic proximity, and potentially unique population histories for each population.

Yemenite Jewish Population History

In light of our data, we investigate support for the various origin theories for the formation of each Jewish population. The primary alternative theories for Yemenite Jewish origins include descent from ancient Israeli exiles (6^{th} and 8^{th} centuries BCE) followed by additional waves of migration over the past 2500 years, or large-scale conversions during the reign of the Himyarite Kingdom (5^{th} and 6^{th} centuries CE). The fact that Yemenite Jews are characterized by a combination of shared and unshared haplotypes with Yemeni non-Jews suggests that the Jewish population was formed by both local Yemeni and external sources. We expect to see mostly exact sequence matches between Yemenite Jews and any contributing source, as 2500 years is typically insufficient time to generate many new control region haplotypes in a small population, given a mutation rate of $4.7x10^{7}$ mut/site/year (Howell et al. 2003). Thus, the low number of Jewish haplotypes found in the non-Jewish population (Figures 2-2, 2-3) suggests a minor contribution of Yemeni haplotypes to Yemenite Jews, indicating that mass conversions during the reign of the Himyarite Kingdom are likely to have played a minor role in forming the Jewish population. However, if many of the converted individuals relinquished their Jewish identity during the rise of Christianity or Islam in the region, the genetic data would be unable to detect the original mass conversion event. Furthermore, if extensive population subdivision characterized the ancient Yemeni population, it's possible that a subpopulation of Yemeni was converted to Judaism,

containing variation that no longer exists in contemporary Yemeni or is not represented in our samples.

The presence of some Eurasian haplogroups in the Jewish population but not in the surrounding Yemeni, e.g. HV1, J2a1, and W, suggests an external contribution to the maternal founders of the Yemenite Jews. Exact matches for HV1 are found in non-Jewish Ethiopians (Kivisild et al. 2004), Iraqi (Arabic) (Richards et al. 2000), and Ashkenazi Jewish (UK) populations (Thomas et al. 2002), for J2a1 in Palestinians, Iraqis (Arabic) (Al-Zahery et al. 2003; Richards et al. 2000) and Iraqi Jews (Richards et al. 2000) and for W in Syrians (Richards et al. 2000). The most parsimonious explanation for the presence of these haplotypes in the Jewish population is some degree of gene flow from the Middle East (and possibly Europe or HOA for HV1) into the Yemenite Jewish population, which is consistent with the first theory that migrants from Israel may have contributed to the formation of this population. This theory is also supported by oral and Biblical traditions, which attest to emigrations as far back as the time of Moses (2^{nd} millennium BCE) (Ben-Zvi 1957) to the first and second exiles of Jewish people from the Kingdom of Israel (722 BCE, 586BCE) (Ahroni 1986), and more recently in response to persecution by Seljuk (Turkish) rulers in 10^{th} and 11^{th} centuries CE (Ben-Zvi 1957).

Y chromosome data from previous studies largely support our mtDNA conclusions on the wide-ranging sources of Yemenite Jewish genetic diversity with evidence for minimal contributions from surrounding Yemeni (e.g. Y haplogroup J1) and longer-range gene flow from the Middle East (e.g. J2e) and the HOA (e.g. A3b3, E3b1) (Cadenas et al. 2008; Shen et al. 2004). In contrast to the mtDNA, however, the Y data

on Yemenite Jews show greater evidence of shared Jewish ancestry. In particular, four Y haplogroups (A3b2, E3b3a, E3b1, and J2e) are shared between Yemenite and Ethiopian Jewish populations, whereas no exact mitochondrial haplotypes are shared between these two populations. Additionally, four Yemenite Jewish Y haplogroups (E3b1, E3b1b, J1, and R1b10) are also shared with other Jewish populations (including Ashkenazi, Iraqi, Libyan, and Moroccan Jews), as well as Druze and Palestinians. This paternal similarity across Jewish populations is consistent with the theory that most Jewish Diaspora populations share more paternal ancestry than maternal ancestry (Thomas et al. 2002).

In sum, neither Yemenite Jewish mtDNA nor Y data support the origin theory of large-scale conversions of Yemeni to Judaism during the 5^{th}-6^{th} centuries CE, based on minimal contribution from the neighboring non-Jewish Yemeni population. In contrast, molecular genetic data support descent from ancient Israeli exiles due to haplotypes shared with other Jewish populations (as seen in the Y chromosome) in addition to shared East African and more generalized Middle Eastern ancestry (supported by both mtDNA and Y).

Ethiopian Jewish Population History

In contrast to Yemenite Jews, our analysis of Ethiopian Jewish mitochondrial diversity indicates strong similarity with the neighboring non-Jewish population, e.g. both populations carry ~50% L(xM,N) haplotypes and 38% of the Jewish haplotypes have exact matches with non-Jewish Ethiopian haplotypes. Ethiopian Jewish mitochondrial diversity is, effectively, a subset of Ethiopian non-Jewish diversity. This result supports previous studies of Ethiopian Jewish mtDNA, which also show a high

similarity between Ethiopian Jewish and non-Jewish sequences (Behar et al. 2008a; Ritte et al. 1993; Thomas et al. 2002; Zoossmann-Diskin et al. 1991).

The main alternative origin theories for the Ethiopian Jewish population consist of 1) descent from ancient Israeli exiles ~2500-2700 years ago, 2) descent from Ethiopian converts to Judaism in the 4th century, and 3) descent from Ethiopian Christians who adopted Jewish Biblical practices in 14th-16th centuries. The fact that Ethiopian Jewish mitochondrial diversity represents a subset of non-Jewish Ethiopian diversity suggests that the Jewish community may have formed from a local sub-sampling of the Ethiopian population. Additionally, the lack of shared haplotypes with other Jewish populations (Behar et al. 2006; Thomas et al. 2002) further weakens support for descent from ancient (female) Israeli exiles. Finally, the age of ~1000-2160 years for the 16305(T) variant implies a minimal age for the Ethiopian Jewish population that is too recent to be consistent with the first theory.

The second and third origin theories both suggest that the Jewish population derived directly from converted Ethiopians, either from Agaw-speakers in the 4th century CE or Ethiopian Orthodox Christians in the 14th - 15th centuries CE. Our simulation analyses suggest that the number of converted female individuals, i.e. the founding maternal population, might have been as small as 224 effective individuals. The timeframes for both origin theories are within the range of our simulation estimates for the minimum age of the Jewish population, i.e. ~420 - 1690YBP (although the age estimated using the method of Tavaré et al. (1997) is older than both origin theories). Linguistic and historic data support the idea that Judaism was present in Ethiopia prior to Ethiopian Orthodox Christianity, thus accounting for the similarity between the two

religions, i.e. contra the third origin theory. Specifically, there are numerous Jewish-Aramaic or Hebrew loanwords for Jewish religious concepts (meswat ["commandment" or "charity"], tabot ["Noah's Ark" or "Ark of the Covenant], ta'ot [idol]) in the Ge'ez version of the Ethiopic Old Testament, attesting to a Jewish presence in Ethiopia at or prior to the time of the translation of the Christian Bible from Greek (Kaplan 1992; Polotsky 1964). Additionally, an historical account by Venetian explorer Marco Polo describes Jews in Abyssinia (Ethiopia) as far back as the 13th century CE (Quirin 1992), thus supporting an early presence of Jews in Ethiopia. However, if the second origin theory is correct, Ethiopian Jews would have spoken Agaw dialects prior to the adoption of Amharic (Henze 1977; Kessler 1996; Quirin 1992) leading to the prediction that Ethiopian Jews might be genetically closer to Cushitic-speakers than Semitic-speakers (Henze 1977). This prediction is not supported by our data, i.e. there are fewer haplotypes in the Ethiopian Jewish population found in Cushitic Ethiopians than Semitic Ethiopians (15.4% vs 34.6%). Thus, our genetic data provide support for a theory of large-scale conversion of local Ethiopians, but cannot distinguish between the two primary conversion theories.

Maternal and paternal histories may vary for Ethiopian Jews as for Yemenite Jews. For instance, we assayed nine Y microsatellite markers in the male R0a1b individuals and found all males exhibited a different Y haplotype (data not shown), in contrast to their identical mitochondrial genomes. A previous study of Y variation showed a uniquely close relationship between Ethiopian Jews and Africans, similar to our mitochondrial data, but also demonstrated a high level of shared Jewish ancestry in Ethiopian Jews in contrast to our mitochondrial data (Shen et al. 2004). Shared

paternal, but not maternal, Jewish ancestry would be consistent with the popular Ethiopian legend of the Queen of Sheba, in which Prince Menelik, the purported son of Israeli King Solomon and Ethiopian Queen of Sheba, may have introduced as many as 10,000 Israeli men into the local Ethiopian population, converting many Ethiopian women to the Jewish religion (Quirin 1992).

Conclusions

Our mitochondrial DNA study of Jewish populations in the HOA and Arabia sheds light on the history of two religiously-defined groups formed relatively recently in these geographic regions. We found no identical haplotypes shared between Yemenite and Ethiopian Jewish populations despite their geographic proximity. Our data on Yemenite Jews suggest possible maternal descent from ancient Israeli exiles, and also demonstrate shared African and Middle Eastern ancestry with little evidence for large-scale conversion of local Yemeni. In contrast, our data on Ethiopian Jews suggest maternal descent primarily from the local Ethiopian population.

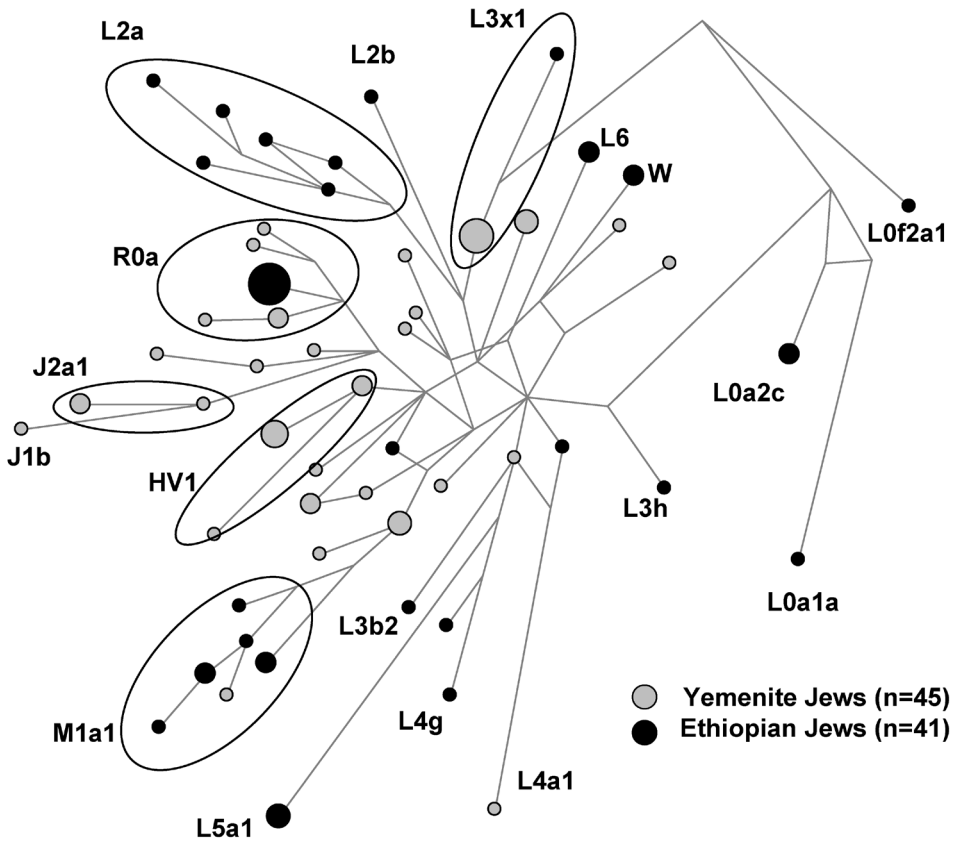


Figure 2-1. Mitochondrial DNA network of Yemenite and Ethiopian Jewish populations. Median-joining network based on HVRI sequence data. Size of circle reflects haplotype frequency and length of line is proportional to the number of mutations. Major haplogroups are indicated and were defined based on diagnostic mutations as described in Tables 2-1 and 2-2.

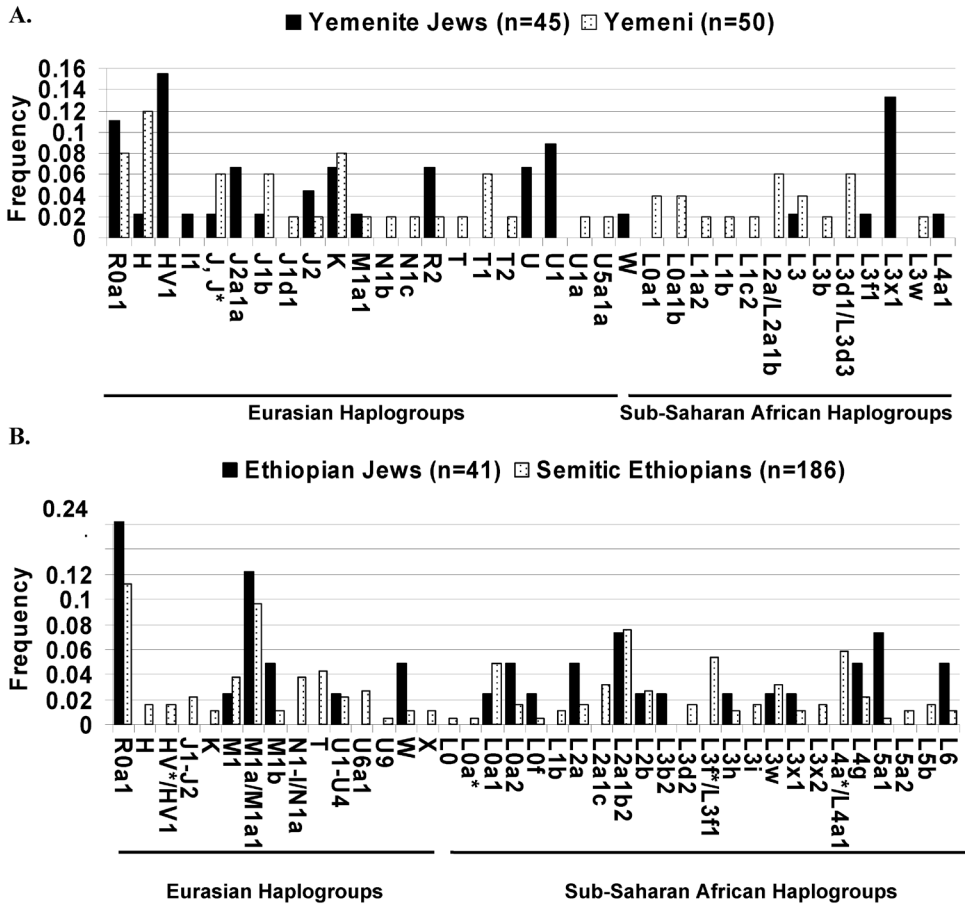


Figure 2-2. Mitochondrial haplogroup distributions of Yemenite and Ethiopian Jewish and non-Jewish populations. Histograms represent haplogroup frequencies based on mitochondrial HVRI and diagnostic coding region mutations. Panel A) Yemenite Jewish and non-Jewish Yemeni. Panel B) Ethiopian Jewish and Semitic non-Jewish Ethiopians. For detailed haplotype information on each individual, see Tables 2-1, 2-2, and 2-3. Frequencies for haplotypes H, J2a1, K, R0a include all derivatives within them.

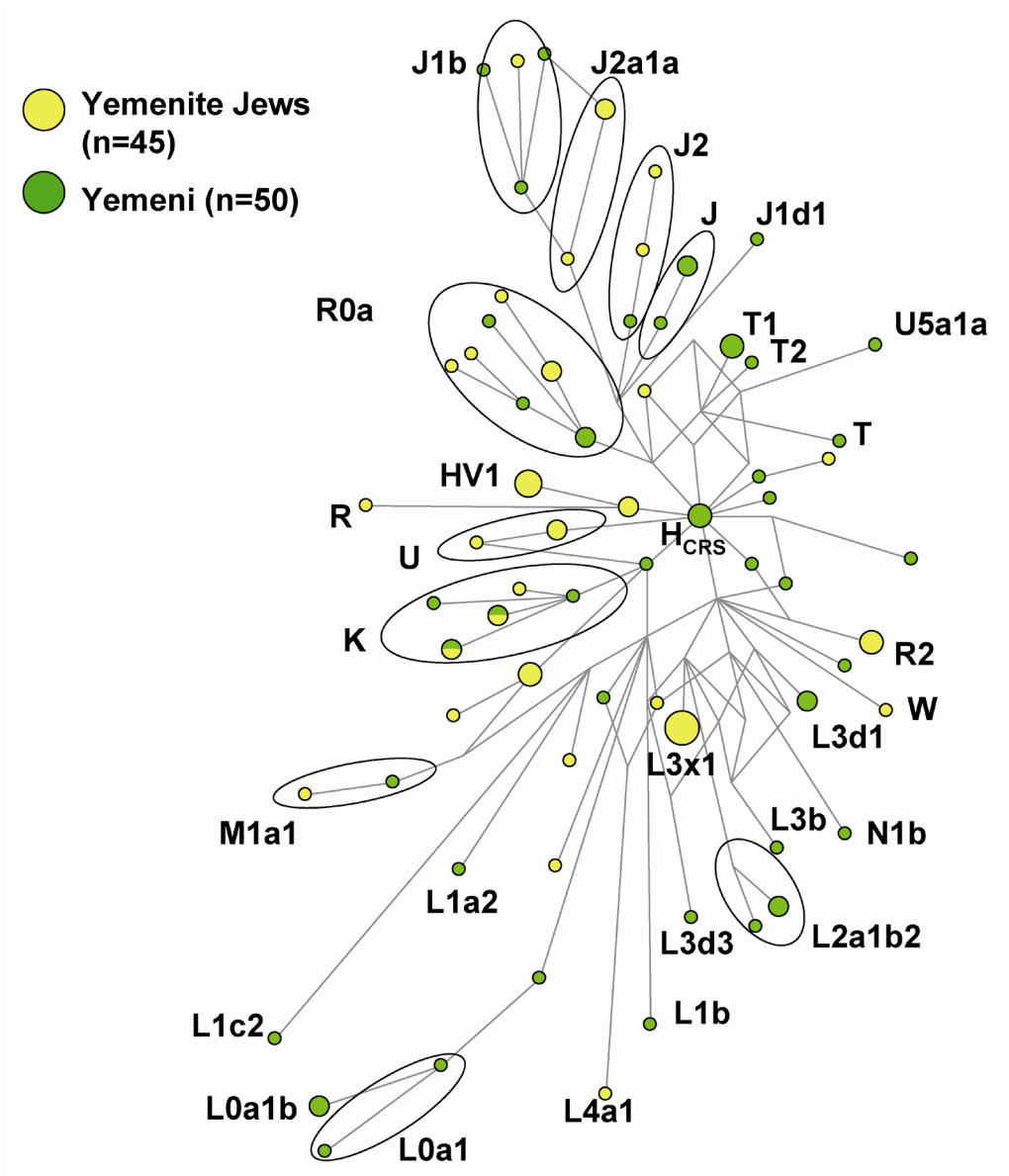


Figure 2-3. Mitochondrial DNA network of Yemenite Jewish and non-Jewish haplogroups. Median-joining network based on HVRI sequence data. Size of circle reflects haplotype frequency and length of line is proportional to the number of mutations. Major haplogroups are indicated and were defined based on diagnostic mutations as described in Table 2-2 and 2-3.

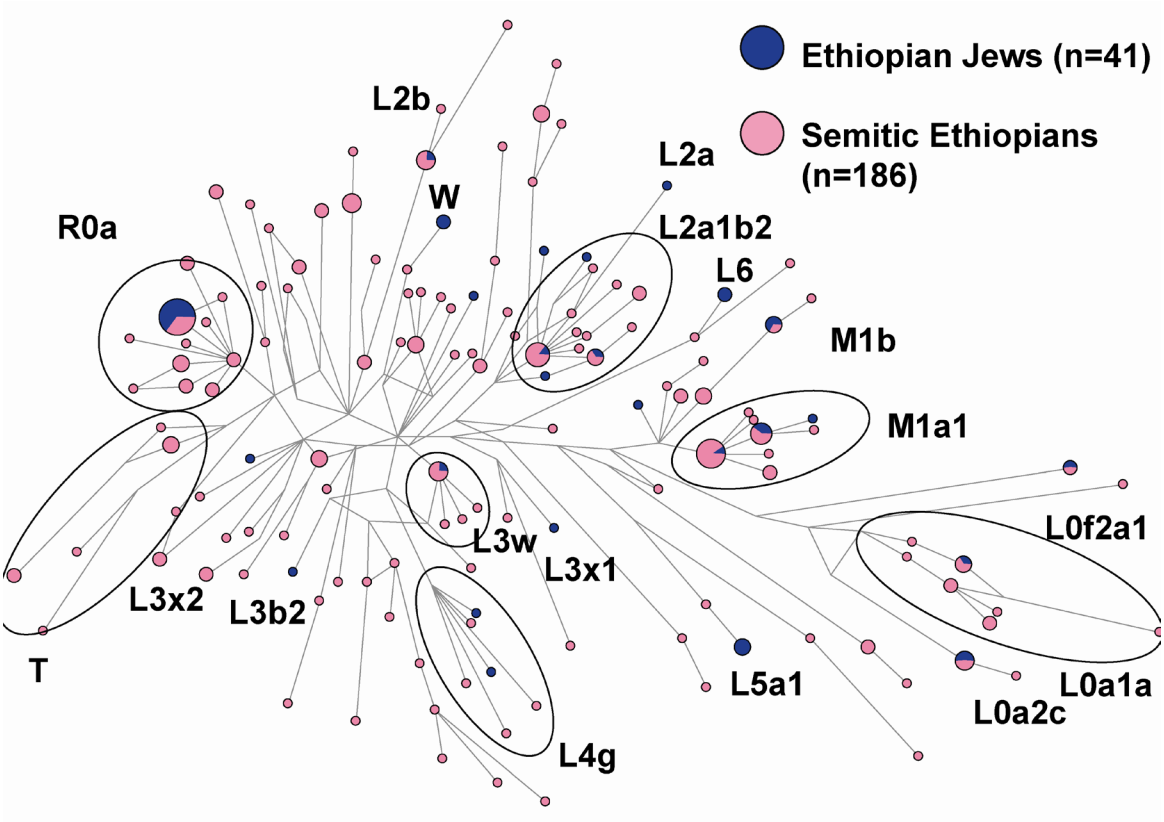


Figure 2-4. Median-joining network based on HVRI sequence data. Maximum Parsimony algorithm in Network v.4.5 was applied to reduce reticulations and find the shortest weighted tree. Size of circle reflects haplotype frequency and length of line is proportional to the number of mutations. Major haplogroups are indicated and were defined based on diagnostic mutations as described in Table 2-1.

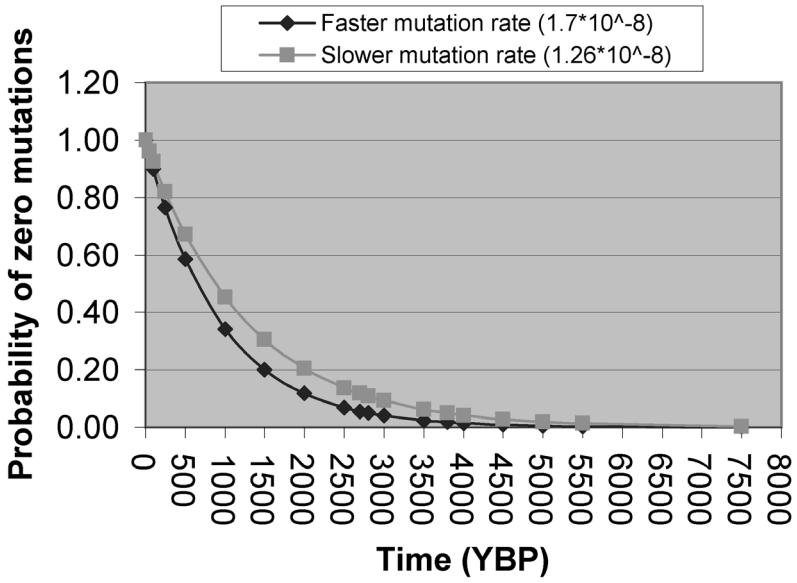


Figure 2-5. Distribution of the probabilities of zero mutations. This plot represents the probability of zero mutations occurring in 6800 bps of coding region sequence in a dataset of nine individuals over different time periods.

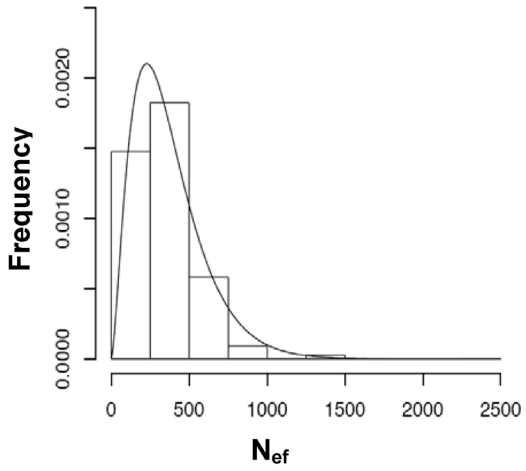


Figure 2-6. Frequency of simulations that produced nine identical sequences at varying Nef values. A gamma distribution was fit to the results representing the frequency of simulated samples of 41 individuals that carried nine identical sequences. Fifteen different Nef values were tested, ranging from 50 to 7000, although only values less than 2500 are shown as all larger values produced zero sets of nine identical sequences. The mode of the distribution (Nef =224) represents the highest frequency of simulations with nine identical sequences, and thus provides the best estimate of Nef for the Ethiopian Jewish population.

Gamma Distribution	N_{ef} value	Age of 16305(T) (YBP)
lower 95%	65	421 (+/- 70)
mode	224	1004 (+/- 27)
mean	365	1261 (+/- 74)
upper 95%	939	1689 (+/- 88)

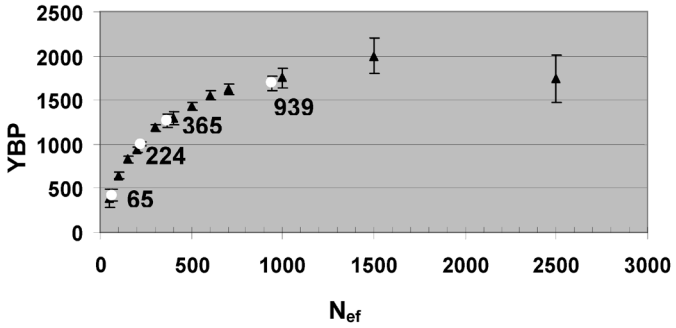


Figure 2-7. Simulation results to estimate age of 16305(T) variant. The results of all simulation analyses are plotted across all tested values of Nef. Black bars represent 95% confidence intervals around each age estimate. The black triangles represent the age of the variant as calculated from maximum pairwise differences at each value of Nef. The white circles represent the age estimates from the mode, mean, and upper and lower 95% confidence intervals of Nef values determined from the gamma distribution in Figure 2-4 (the inset table reports these Nef values and estimated ages based on each Nef value).

Table 2-1. Haplotypes for Ethiopian Jewish individuals (n=41)

Ethiopian Jewish Sample #	HVR-I (without 16000) 16024-16383	#	Frequency	mtDNA Haplotype	+10397 *Alu*I (M)	-3592 *Hpa*I (L3, L4, L7)	-7025 *Alu*I (H)	+12308 *Hinf*I (U, K)	+15606 *Alu* I (T)	+10394 *Dde*I (J/T,+=J)	+10871 *Mnl*I (N)	-9052 *Hae*II (K)	bp 769 (G=L3, A=L4)	bp 1018 (G=L3, A=L4)
1818	129 148 168 172 187 188G 189 223 230 293 311	1	0.024	L0a1a										
2232, 4568	148 172 187 188A 189 214 223 230 234 311	2	0.049	L0a2c										
1831	148 192 223 234 311	1	0.024	L3h	-	-								
1749	129 169 172 173 187 189 223 230 239 278 311 327 368	1	0.024	L0f2a1										
2238	129 223 242A 278 294 309 380	1	0.024	L2a										
1729	189 192 223 278 294	1	0.024	L2a		+								
2231	189 192 223 278 294 309	1	0.024	L2a1b2										
2234	145 150 189 223 278 294 309	1	0.024	L2a1b2										
4626	189 223 278 294 309	1	0.024	L2a1b2										
1354	114A 129 145 213 223 278	1	0.024	L2b										
1804	93 124 223 278 311 362	1	0.024	L3b2										
2229	223 260 311	1	0.024	L3w										
1613	169 231 278 311	1	0.024	L3x1										
1427	170 223 293T 311 355 362	1	0.024	L4g										

Table 2-1. Continued

Ethiopian Jewish Sample #	HVR-I (without 16000) 16024-16383	#	Frequency	mtDNA Haplotype	+10397 *Alu*I (M)	-3592 *Hpa*I (L3, L4, L7)	-7025 *Alu*I (H)	+12308 *Hinf*I (U, K)	+15606 *Alu* I (T)	+10394 *Dde*I (J/T,+=J)	+10871 *Mnl*I (N)	-9052 *Hae*II (K)	bp 769 (G=L3, A=L4)	bp 1018 (G=L3, A=L4)
2224	172 223 287 293T 311 355 362	1	0.024	L4g										
4566, 4635, 4694	129 148 166 187 189 223 278 311 355 362	3	0.073	L5a1										
2227, 2237	48 69 223 224 278 311	2	0.049	L6										
2241	129 154 183C 189 223 249 311	1	0.024	M1										
1859	93 129 189 223 278 294 309	1	0.024	M1a	+									
1137	129 189 223 249 311 359	1	0.024	M1a1										
2228, 2245	93 129 189 223 249 311 359	2	0.049	M1a1							+			
2246	93 129 183C 189 213 223 249 311 359	1	0.024	M1a1										
1610	182C 183C 189 223 249 260 311 320	1	0.024	M1b										
1683	189 223 249 260 311 320	1	0.024	M1b										
1599, 1822, 2225, 2233, 2249, 2250, 2253, 2256, 4524	126 305T 362	9	0.220	R0a1b										
1570	183C 249	1	0.024	U1	-	-		+						
1574, 2537	166 192 223 292 343	2	0.049	W										
Total		41												

Table 2-2. Haplotypes for Yemenite Jewish individuals (n=45)

Yemenite Jewish Sample #	HVR-I (without 16000) 16024-16383	#	Frequency	mtDNA Haplotype	+10397 *Alu*I (M)	-3592 *Hpa*I (L3, L4, L7)	-7025 *Alu*I (H)	+12308 *Hinf*I (U, K)	+15606 *Alu* I (T)	+10394 *Dde*I (J/T,+=J)	+10871 *Mnl*I (N)	-9052 *Hae*II (K)	bp 769 (G=L3, A=L4)	bp 1018 (G=L3, A=L4)
1269	223 311 362	1	0.023	L3	-	-					-		G	G
1357	93 209 223 292 311	1	0.023	L3f1										
1446, 1351, 1508, 1382, 1330, 1334	169 223 278	6	0.136	L3x1										
1412	207T 220 223 234 260 261 311 360 362	1	0.023	L4a1									A	A
1463	129 354	1	0.023	H2a1	-		-							
1420	93 126 355 362	1	0.023	R0a										
1348	126 162 355 362	1	0.023	R0a										
1336	126 260 304 362	1	0.023	R0a2c										
1341, 1350	126 304 362	2	0.045	R0a2c										
1352, 1519, 1597, 1335	67 274	4	0.091	HV1										
1373, 1443	67	2	0.045	HV1										
1356	126 214 231	1	0.023	J2										
1426	69 126 214 231	1	0.023	J2										
1349	126 192	1	0.023	J*	-				-	+				
1548	69 126 145 261	1	0.023	J2a1a†††										
1416, 1344	69 126 136 145 221 261	2	0.045	J2a1a†††										
1345	69 126 145 168 222 235 261	1	0.023	J1b										
1656	93 224 311	1	0.023	K										

Table 2-2. Continued.

Yemenite Jewish Sample #	HVR-I (without 16000) 16024-16383	#	Frequency	mtDNA Haplotype	+10397 *Alu*I (M)	-3592 *Hpa*I (L3, L4, L7)	-7025 *Alu*I (H)	+12308 *Hinf*I (U, K)	+15606 *Alu* I (T)	+10394 *Dde*I (J/T, +=J)	+10871 *Mnl*I (N)	-9052 *Hae*II (K)	bp 769 (G=L3, A=L4)	bp 1018 (G=L3, A=L4)
1435	192 210 224 311	1	0.023	K2c										
1440	129 224 311	1	0.023	K	-		+	+				-		
1327	129 223 293C 311	1	0.023	I1										
1329	129 183C 189 223 249 311 359	1	0.023	M1a1										
1325, 1381, 1339	71 188 223	3	0.068	R2	-						+			
1340	67 93G 189	1	0.023	HV1			+							
1375	86 119 311	1	0.023	U			+	+						
1156, 1328	86 119	2	0.045	U			+	+						
1343	166d 189 249 311	1	0.023	U1				+			+	+		
1333, 1609	166d 182C 183C 189 249 311	2	0.045	U1										
1368	182C 183C 189 193 249 311	1	0.023	U1										
1338	183C 189 223 292 295	1	0.023	W										
Total		45	1.000											

Table 2-3. Haplotypes for Yemeni individuals (n=50).

Yemeni Sample	HVR-I (without 16000) 16024-16383	#	Frequency	mtDNA Haplotype	+10397 *Alu*I (M)	-3592 *Hpa*I (L3, L4, L7)	-7025 *Alu*I (H)	+12308 *Hinf*I (U, K)	+15606 *Alu* I (T)	+10394 *Dde*I (J/T,+=J)	+10871 *Mnl*I (N)	-9052 *Hae*II (K)	bp 769 (G=L3, A=L4)	bp 1018 (G=L3, A=L4)
Y305	129, 148, 168, 172, 190G, 223, 230, 311, 320,	1	0.020	L0a1										
Y475	093, 129, 148, 168, 172, 190G, 223, 230, 271, 311, 320	1	0.020	L0a1										
Y245	129, 148, 168, 172, 187 188G 189, 223, 230, 263, 278, 292, 293, 320	1	0.020	L0a1b										
Y334	129, 148, 168, 172, 187, 188G, 189, 223, 230, 278, 293, 311, 320	1	0.020	L0a1b										
Y179	148, 172, 190G, 223, 230, 311, 320	1	0.020	L1a2										
Y140	126, 223, 264, 270, 278, 289, 293, 311	1	0.020	L1b										
Y327	078, 129, 187, 189, 223, 265C, 286A, 294, 311, 320, 360	1	0.020	L1c2										
Y269	093, 189, 223, 278, 294, 309	1	0.020	L2a1b2										
Y001	189, 192, 223, 278, 294, 309	1	0.020	L2a1b2										
Y196	189, 192, 223, 278, 294, 309	1	0.020	L2a1b2										
Y094	223, 355	1	0.020	L3[†]	-	-	+				-		G	G
Y160	093, 129, 223, 270, 311, 327	1	0.020	L3	-	-					-		G	G
Y229	093, 124, 223, 278, 362	1	0.020	L3b										
Y271, Y294	124, 223, 319	2	0.040	L3d1										
Y058	124, 189, 223, 278, 304, 311	1	0.020	L3d3										

Table 2-3. Continued

Yemeni Sample	HVR-I (without 16000) 16024-16383	#	Frequency	mtDNA Haplotype	+10397 *Alu*I (M)	-3592 *Hpa*I (L3, L4, L7)	-7025 *Alu*I (H)	+12308 *Hinf*I (U, K)	+15606 *Alu* I (T)	+10394 *Dde*I (J/T,+=J)	+10871 *Mnl*I (N)	-9052 *Hae*II (K)	bp 769 (G=L3, A=L4)	bp 1018 (G=L3, A=L4)
Y502	223, 260, 311	1	0.020	L3w										
Y101, Y230, Y042	CRS	3	0.060	H										
Y134	218	1	0.020	H20										
Y222	354	1	0.020	H2a1										
Y381	311	1	0.020	H										
Y063, Y144	069, 084, 126, 241	2	0.040	J										
Y390	069, 126, 241	1	0.020	J										
Y005	069, 126, 136, 145, 221, 222, 261	1	0.020	J1b										
Y085	069, 126, 145, 222, 261	1	0.020	J1b										
Y215	069, 126, 145, 222, 261	1	0.020	J1b										
Y506	069, 126, 193, 300, 309	1	0.020	J1d1										
Y175	069, 126, 231	1	0.020	J2										
Y046	176, 193, 224, 311	1	0.020	K								-		
Y459	192, 210, 224, 311	1	0.020	K2c										
Y463	224, 311	1	0.020	K										
Y464	129, 224, 311	1	0.020	K								-		
Y197	129, 189, 223, 249, 311, 359	1	0.020	M1a1										
Y503	145, 176G, 223, 295, 362	1	0.020	N1b							+			
Y374	201, 223, 265	1	0.020	N1c							+			

Table 2-3. Continued

Yemeni Sample	HVR-I (without 16000) 16024-16383	#	Frequency	mtDNA Haplotype	+10397 *Alu*I (M)	-3592 *Hpa*I (L3, L4, L7)	-7025 *Alu*I (H)	+12308 *Hinf*I (U, K)	+15606 *Alu* I (T)	+10394 *Dde*I (J/T,+=J)	+10871 *Mnl*I (N)	-9052 *Hae*II (K)	bp 769 (G=L3, A=L4)	bp 1018 (G=L3, A=L4)
Y059	126, 355, 362	1	0.020	R0a										
Y369, Y376	126, 362	2	0.040	R0a										
Y508	092, 126, 278, 362	1	0.020	R0a										
Y366	071	1	0.020	R2										
Y493	126, 294, 296	1	0.020	T										
Y143	126, 163, 294	1	0.020	T1										
Y217	126, 163, 294	1	0.020	T1										
Y396	126, 163, 186, 189, 294	1	0.020	T1										
Y051	126, 292, 294	1	0.020	T2										
Y073	184A, 249, 355	1	0.020	U1a			+							
Y035	192, 256, 270, 294	1	0.020	U5a1a										
Total		50	1.000											

Haplotypes are defined by HVRI variants and select coding region variants based on previously published definitions (Behar et al. 2008a; Kivisild et al. 1999; Kivisild et al. 2004; Metspalu et al. 2004; Richards et al. 2000; Salas et al. 2004; Santos et al. 2004; Torroni et al. 1996). In addition, Ethiopian R0a1b haplotypes (n=9) include HVRII and 6800 bps of coding region sequence. †Individual Y94 is classified as L3 based on lack of the diagnostic marker for H (-7025), lack of the diagnostic marker for M (+10397), presence of the 3592 mutation (which generally defines L3/L4/L7), and presence of G's at bp 769 and bp 1018 (which generally define L3). †††J2a1 is renamed from J1b (Palanichamy et al. 2004). Deletions are represented by a "d" following the nucleotide position, and all mutations are transitions unless the nucleotide change is indicated. CRS indicates the sequence is identical to the Cambridge Reference Sequence. Frequencies for haplotypes H, J2a1, K, R0a include all derivatives within them.

CHAPTER 3
INTEGRATION OF MTDNA, GIS, AND ARCHAEOLOGICAL DATA TO ADDRESS REGIONAL AND CONTINENTAL/BROADER SCALE GEOGRAPHIC PATTERNS IN YEMEN AND SURROUNDING REGIONS

Introduction

The geographically central location of Yemen directly between Africa, Europe, and Asia, has facilitated complex interactions/connections between human populations in these regions for tens of thousands of years (Rose and Petraglia 2009). Yemen's key position just across the Red Sea from East Africa places it as one of the first stops in the Southern Dispersal Route out of Africa, positioning Yemen as a critical area for genetic investigations of human evolutionary history. In historical times, Yemen has served as a particularly active crossroads of great trade and slave networks among the world's first three complex civilizations in Ancient Egypt, Mesopotamia, and the Indus Valley. In addition to its central geography, Yemen also has an extremely diverse set of microclimates, ranging from the hot and arid coastal Tihama region to the temperate Central Highlands. This diverse climate has led to the development of distinct subsistence strategies and cultures throughout the country (Wenner 1991). The Yemeni population is so de-centralized that its people are dispersed amongst approximately 50,000 distinct towns and villages within 135,000 square kilometers of land (Wenner 1991). Specifically, 53% of its population live in villages of less than 50 people, 78% in villages of less than 100 (Dresch 1989), and 86% in villages of less than 1,000 (Wenner 1991). These features of the Yemeni landscape have led to a unique diversity within the Yemeni people that can be explored through analyses of genetic data.

The material record of ancient human occupations in Yemen is sparse, as archaeological research has been impeded by war and by the harsh climate, described

today as “one of the hottest and most inhospitable deserts in the world” (Rose and Petraglia 2009). Researchers have instead traditionally focused on more accessible regions of the Levant and East Africa, leaving the history of Yemen and the Arabian Peninsula in relative obscurity. Though limited, the archaeological record attests to a complex history of migrations and trade routes into and out of Yemen (and Arabia) since at least the emergence of anatomically modern humans. Evidence for obsidian trade networks extends back to the 7th Millenium BCE, with lithic material from Early and Mid-Holocene archaeological sites found in Western Yemen which likely originates in Ethiopia and Eritrea (Francaviglia 1995; Ridl et al. 2009). Agricultural evidence supports contact with both the Levant and East Africa, through the discovery of domesticated crops of both Levantine origin (3-3.5 BCE), e.g. wheat, barley, lentils, and chickpeas (Edens 2005), and East African origin (2.5-1 BCE), e.g. sorghum and millets (de Moulins et al. 2003). The material record attests to a strong Arabian presence in the Ethiopian highlands by 1,000 BCE., as well as an Ethiopian Aksumite occupation of Yemen during the 1st millennium CE (Fattovich 1997b; Munro-Hay 1991). The earliest non-African migrants into Yemen are likely the Sabeans, who came from northeastern Arabia and established the first Arabian kingdom of Saba (i.e., Sheba). During the first millennium BCE and the first few centuries CE, Saba’s capital of Marib became a key trading center along the camel caravan route (i.e. “The Incense Route”) allowing for frankincense to be transported from Eastern Yemen to India and Eastern Africa, and throughout the Mediterranean (Muller 1988). By the 9th century CE, migrants were entering Yemen from distant sites including Iran, while military invasions introduced Persian communities, among others, which were eventually absorbed into the Yemeni

population. Most recently, in the past 1000 years, expansion of the Islamic commercial and political spheres has brought migrants from Arabia back to East Africa and to other surrounding regions (Fattovich 1997b), thereby creating a complex and nuanced genetic landscape around the Red Sea.

The recent increase in the availability of genetic data from Eastern Africa and the Arabian Peninsula has permitted more detailed analyses of human population history in the greater Red Sea Region (Abu-Amero et al. 2007; Abu-Amero et al. 2008; Cerny et al. 2010; Cerny et al. 2008; Cerny et al. 2009; Kivisild et al. 2004; Non et al. 2010a). At the same time, new advances in GIS technology have made geographic data increasingly available for genetic spatial analyses. Despite the widespread use of spatial variation methods across other disciplines, including ecology (Infascelli et al. 2009), conservation biology, and most recently parasitology and epidemiology (Goovaerts 2010; Saxena et al. 2009; Schröder 2006), most studies of human evolutionary history do not incorporate geographic data into genetic analyses. Yet these data are essential for understanding the history of population migrations and settlement patterns. Over the past few years, top journals have called for greater rigor among geneticists to collect and share spatial data linked with genetic data in public databases such as GenBank (2008; Field 2008). Unfortunately, these data have not typically been collected by most human geneticists or biological anthropologists (as discussed in Relethford 2008), despite the fact that humans are among the most mobile of all species.

In this study, I analyze a new large dataset of mitochondrial DNA (mtDNA) from Yemen linked to the most accurate set of geographic coordinate information to date in

this region of the world. I analyze these data first within Yemen, to determine regional geographic structure within the mitochondrial haplotype data that may indicate signatures of past colonizations or settlement patterns. I also compare the Yemeni population to those from surrounding regions of Eastern Africa, the Middle East, and India, to provide a larger-scale understanding of genetic diversity of this region. I present mtDNA haplotype frequencies on the largest and most geographically widespread set of Yemeni samples to date, and also apply spatial analysis techniques such as interpolation contour maps and correlograms of spatial auto-correlation along with traditional techniques to measure genetic variation, e.g. principal component analyses and AMOVA. These techniques explicitly incorporate the geographic information into the analyses to test for significant patterns of geographic structure within the genetic data. Individual mtDNA haplotypes are analyzed under the assumption that each haplotypes could potentially reflect a unique (maternal) migration event. I discovered that Yemen's geographically central location is reflected by its complex genetic diversity, which contains a mixture of African, Middle Eastern, and Eurasian haplotypes throughout its various governorates. Only by taking a wider-range view of the Red Sea Region was I able to visualize that, despite a lack of regional genetic structure, Yemeni genetic diversity fits into a larger pattern of migrations out of (and perhaps back into) Africa across the Red Sea. The interpolation maps in particular provide compelling visual displays of unique frequency patterns for different haplotypes, which may reflect, in part, past population movements into and out of Yemen and throughout the Red Sea region. In addition to demonstrating the benefits of explicitly incorporating geographic information into analyses of human population history, these

genetic analyses also address questions related to 1) prehistoric migrations out of Africa along the Southern Dispersal Route 2) ancient and/or more recent settlement patterns throughout regions of Yemen, 3) slave trade in Yemen from Eastern Africa, Europe, or SE Asia, and 4) the Hadrami Diaspora to Asia.

Materials and Methods

Samples

Saliva samples were obtained from 550 unrelated individuals collected throughout Yemen in the Spring of 2007. Of these samples, 546 were amplified and sequenced for the HVRI in the control region of the mitochondria. Geographic location on all samples was collected for birthplace, maternal birthplace, maternal grandmother birthplace, as well as paternal, paternal grandmother, and paternal grandmother birthplace. In the current study, only samples with maternal grandmother birthplaces from governorates with very small sample sizes (Ibb, n=6; Raymah, n=2; Al Malwhit, n=1; and Sana'a, n=1) were excluded, leaving a final sample size of 536 sequenced Yemeni for geographic analyses. Of these, 37 could not be unambiguously assigned to a single macrohaplogroup (L, M, or N), but only 15/536 (2.8%) were excluded from analyses, as the rest were classifiable into a few potential haplotypes that were not analyzed in detail in the current study, and thus retained for purpose of maintaining sample size. In sum, 521 Yemeni samples were used in all analyses. Preliminary Y chromosome haplotypes were assigned to 50 Yemeni samples based on typings of up to 19 SNP markers (including YAP, P29, P2, 12f2, M9, M11, M12, M33, M67, M74, M75, M78, M81, M89, M123, M172, M175, M201, and M267). Data from 27 nuclear Alus insertion polymorphisms were also generated on the same 50 samples (see Raaum et al. 2010 for description).

An additional 1420 samples were collected from published databases (Table 3-1), including Genbank (http://www.ncbi.nlm.nih.gov/sites/entrez) Mitomap, (http://www.mitomap.org), and mtDB (http://www.genpat.uu.se/mtDB), and unpublished data shared by Martin Richards (2004), to compare Yemen to surrounding regions of Eastern Africa, the Middle East, and Asia (Table 3-1, Figure 3-1). With the exception of Yemen, where exact maternal grandmother birthplace locations were known, the locations of sample collections were all estimated from maps or locations described in the original publications (Abu-Amero et al. 2007; Tishkoff et al. 2007a; Vernesi et al. 2001). In cases where no specific geographic location was specified, a central position in the country was estimated, e.g. Israel and Iraq.

Sequencing

DNA was isolated from saliva samples using the Oragene saliva kit from DNA Genotek (Ontario, Canada) and the standard protocol for manual purification of DNA from 0.5mL of Oragene/saliva (dnagenotek.com). For some Yemeni samples, the HVRI segment of DNA was sequenced at the Interdisciplinary Center for Biotechnology Research (ICBR) at the University of Florida using primer pairs L15811 and H5 (3' positions relative to Cambridge Reference Sequence (Gonder et al. 2007). Sequence data were generated by Non, Cerny, Sanchez, and La Joie. Comparative sequences were collected by Non and Stelmaszek. All analyses described in this dissertation were conducted by Non.

Haplotype Assignment

Sequences were assigned an initial haplotype designation using the program mtDNAresource (http://www.mtdnaresource.com/), which assigns haplotypes based on previously published sequences (Kivisild et al. 1999; Kivisild et al. 2004; Metspalu et al.

2004; Richards et al. 2000a; Salas et al. 2002). Each classification was refined by two independent readers by comparison to the most updated mtDNA tree available in Phylotree, compiled from 98 updated references (http://www.phylotree.org). For the majority of sequences, only HVRI variation was considered when assigning haplotypes, because for most sequences, this was the only sequence data available. For sequences from India (Eaaswarkhanth et al. 2010), previously published haplotypes classifications were used, based on HVRI, HVRII, and selected coding region sites, because for Indian haplotypes, the additional sequence information was required to obtain clear haplotype classifications. Indian sequences classified as R01 by Eaaswarkhanth et al. (2010) were renamed R0a for comparison to the rest of the total sample. For sequences from Tanzania (Tishkoff et al. 2007a) and Saudi Arabia (Abu-Amero et al. 2008), previously published haplotypes classifications based on HVRII and coding region information were also used only for the13 Tanzanian and 7 Saudi Arabian samples in which HVRI data was not sufficient to obtain a macrohaplogroup classification. The individual haplotype frequencies were used as the unit of analysis in all following analyses, except AMOVA, where full HVRI sequence information was utilized.

Genetic Analyses

Haplotype frequencies for each geographic region were calculated to create Principal Component plots using SAS software v9.2 (SAS Institute Inc, 2009). In Yemen, frequencies were calculated for each governorate with a minimum of 10 maternal grandmother birthplaces, or in one case, frequencies were calculated for governorates with small sample sizes by pooling samples together, e.g. a Southwest (SW) Yemen site was created by pooling samples with maternal grandmother birthplaces in Lahij, Aden, and Taizz.

Principal Component Analysis (PCA) is a model-free summary technique to detect key dimensions of structure within the data (McVean 2009). The goal of PCA is to transform a large number of variables into a smaller number of "principal components," which explain the main patterns in the data. A two-dimensional summary of the structure of the data is produced when the samples are projected along the first few axes, or components, the first of which represents the greatest amount of variability in the data, and each succeeding component accounts for as much of the remaining variability as possible. The most frequent haplotypes/groups in Yemen, i.e. (L(xM,N) (all L except L2a1), L2a1, J1b, H (including all H derivatives), R2, T1a, K (including all K derivatives), I3/I5a (previously named N* by Cerny et. al. (2008)), and R0a, were selected to produce the PCA plots of mtDNA frequencies for all samples. .

Analysis of molecular variance (AMOVA) was also used to compare genetic variation in different regions of Yemen. AMOVA analyses are generally used to determine degree of variation within and between populations in a hierarchical framework. Population subdivisions were created based on natural geographic features of the landscape, i.e. Highlands-Coastal, and Coastal-inland and the East-West division was based on a pattern detected in preliminary contour map analyses using kriging based on a subset of 93 HVRI sequences from Yemen (data not shown).

Spatial Analyses

Interpolation maps were created using the geostatistical analyst extension of ArcMap software version 9.3.1, to visually display the geographic distribution of each of the most common mtDNA haplotypes in Yemen, i.e. L(xM,N), J1b, H, R2, T1a, L2a1, K, I3/I5a, R0a. The maps were created using the inverse distance weighting (IDW) algorithm, which predicts values for all unmeasured points by placing greater weight on

points closer to the prediction location than more distant points. An IDW^2 interpolation was used, along with a circular shape for selection of neighbors, which does not assume any *a priori* directional influence on the weight of the neighboring points. A small set of five neighbors (with a minimum of at least 1 neighbor) was considered in interpolating frequencies for each area on the map, because low autocorrelation values suggested that haplotypes frequencies at greater distances were not statistically correlated. All presented maps show intervals in the contour lines (as seen in each map's legend) based on a calculated geometric intervals, i.e. "smart quantiles," which are an alternative to standard quantile values designed for data that are not normally distributed. Geometric intervals are derived by dividing the difference in the first interval by a coefficient that is computed to best fit the data for each consecutive intervals (see Arcmap Manual for further discussion). Alternatively, maps were also created with equal sized intervals, but less resolution was produced than with the geometric intervals, presumably because the data are not normally distributed.

Semivariogram/covariance clouds were also created for each of these haplotypes, to provide a visual display of the spatial autocorrelation between pairs of samples. Each point in the semivariogram cloud represents the difference squared between the haplotypes frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. As closer geographic points are expected to appear more alike than more distant points, the semivariogram is expected to show increasing frequency distances with increasing geographic distances along the x-axis.

Moran's I indices using the inverse distance method were calculated for each haplotype to generate a global measure of spatial autocorrelation. Euclidean (straight-line) distances ($dij=\sqrt{[(x_i-x_j)^2(y_i-y_j)^2]}$) were used in Moran's I calculations because the geographic distances were not great enough that the curvature of the Earth significantly changes the values. Inferences on the significance of Moran's I are based on the test of the null hypothesis of no significant geographic clustering of genetic data. Thus, a positive value of Moran's I indicates significant clustering of haplotype frequencies, while negative values indicate significant dispersion of haplotype frequencies, and values close to zero indicate randomness in spatial clustering of haplotypes.

Correlograms were created to visually display Moran's I indices plotted against geographic distance using the software PaSSAGE v2.0. For all presented correlograms, the geographic data were binned into 10 distance classes with an uneven number of samples per distance class. Again, Euclidean geographic distances were used and inverse distances for haplotypes frequencies, with an $\alpha=0.05$ to determine significance in each distance class.

Results

Yemen Haplotypes

The frequency of each HVRI-based mitochondrial haplotype is reported in Figure 3-2. Of the total Yemeni sample, the majority of individuals were assigned to one of three major macrohaplogroups, L, M, or N (484/521, 92.90%), based on HVRI sequence data; the remaining 7.10% could not be classified based on HVRI data alone. In general, Yemen is characterized by a large number of low frequency haplotypes, i.e. 33.4% of the 521 samples; 89 distinct haplotypes are present at frequencies less than 1% (Figure 3-2). Of the major non-sub-Saharan African macrohaplogroups found in

Yemen, macrohaplogroup R represented the largest proportion of haplotypes in Yemen with 305/521 (59.11%) of the total sample. Among these, R0a was the most frequent haplotype in Yemen at 17.7%, followed by J* (7.10%), J1b (4.99%), H (4.41%), and R2 (3.45%), K (2.50%), and T1a (2.50%). Macrohaplogroup M was the next largest Eurasian macrohaplogroup at 5.75%. Fewer Yemeni sequences were classified within macrohaplogroup N (8.4%), including I3/I5a (could not be classified further based on HVRI, 3.45%), and a few samples of W, X, and Y, each at less than 1%.

Overall, sub-Saharan African haplotypes, hereafter referred to as L(xM,N) following the nomenclature of Behar et al. (2008b), were found at 22.07%, a relatively high proportion for regions outside of Africa. Among the L haplotypes, L2a1 was the most frequent in Yemen at 4.2%. The proportion of L(xM,N) haplotypes in East Yemen (as defined by all governorates east of the highlands, see Figure 3-3) is significantly higher than in West Yemen (30.30% versus 22.3%, respectively, p=0.02). A chi-square test indicates significant differences in frequency of L(xM,N) haplotypes between at least two of the tested governorates (p=0.041). Multiple proportion comparison tests were also conducted to compare the frequency of L(xM,N) haplotypes in each governorate with the frequency in every other governorate. Significant differences were identified only between the governorates of Hadramout and Dhamar, and between Hadramout and Amran (Figure 3-3).

AMOVA analyses were also conducted to test for significant differences between different regional subgroupings. Yemen was first divided into East and West groups (as designated in Fig 3), to test for a significant difference in variation as suggested by the two sample test of proportions. Second, a mainland-Socotra distinction was also tested.

The AMOVA test is more rigorous than the two sample test of proportions as it is based on all variation in the full mtDNA sequences, rather than based on a summary statistic such as haplotype frequency. The results of the AMOVA analyses are shown in Table 3-2. Of the two comparisons, only one the Socotra versus Mainland division showed significant apportioning of variance between regions (5.13%, $p<0.0001$).

The frequency of African L(xM,N) haplotypes within Yemeni governorates was also compared to frequencies in other populations on a broader geographic scale from Eastern Africa to India (Figure 3-4). The frequency of African haplotypes decreases dramatically from the highest frequencies in Eastern Africa (e.g. Tanzania-all ethnic groups greater than 89%, Kenya: 79%, Sudan Dinka 72.1%) to the lowest frequencies in Israeli Druze (0%), Marib Yemen (0%), and the Indian populations of Western Dawhoodi Bohra, Sunni, Shia, and Mappla (0%). African frequencies are not perfectly predicted by longitude, however, as the lowest frequencies of African haplotypes were found in centrally located Middle Eastern regions of Israel (Druze, 0%), Syria (4.6%), Northern and Southern Saudi Arabia (5.6% and 4.6%) and Marib, Yemen (0%), while the most eastern populations in Yemen had higher frequencies, e.g. Hadramout (44.8%) (Figure 3-4, Panels A, B, Table 3-3).

Overall Geographic Structure Using Principal Components

I next used mitochondrial haplotype frequency data from the most frequent mtDNA haplotypes in Yemen to construct principal component plots of all populations included in the study (Figure 3-5). These plots provide an overall visual display of the relationship between all sampled populations, combining the information from the nine most frequent Yemeni haplotypes and haplogroups. The first principal component (PC1) along the x-axis captures the largest proportion of variation in the data, 31.38%,

followed by the second principal component (PC2) plotted on the Y axis, which captures 16.30%, contributing to a cumulative proportion of 47.68% in the first two components (Figure 3-5a). PC1 primarily captures the largest distinction in this sample: the difference between samples of African and non-African ancestry. This is supported by examination of the component pattern profile (Figure 3-6), which illustrates the high contribution of haplotypes L(xM,N) and L2a1 to PC1. However PC1 is also influenced by high frequencies of certain Eurasian haplotypes including J1b, H, T1a, K, and R0a. Those with high frequencies of these haplotypes are pushed more to the right on the x-axis, which may explain why India appears more centrally located than expected given its very low frequency of African haplotypes. It is interesting to note that governorates in eastern Yemen (i.e. Hadramout and Hajjah) cluster more with African samples, while western Yemeni governorates cluster with other Middle Eastern and Eurasian samples reflecting the higher frequency of L(xM,N) haplotypes in the eastern Yemen populations.

PC2 separates Socotra and Mahra from the rest of the global dataset (Figure 3-5a and 3-5b). This difference is driven mainly by frequencies of haplotypes R2 and I3/I5a (Figure 3-6), both of which have highest frequencies in Socotra and Mahra, and very low frequencies in the other populations. PC3 mainly separates Marib and Shabwah (at the top of the plot), and Israeli Druze and Syria (at the bottom of the plot) from the remaining populations (Figure 3-5b), based mainly on high frequencies of J1b and R0a for Marib and Shabwah, and high frequencies of K and T1a for Druze and Syria (Figure 3-6, Table 3-4). PC4 (Figure 3-5c) primarily separates Socotra and Mahra from the remaining populations based on high frequencies of I3/I5 (22.2%) a in Socotra, and high frequencies of H (13.9%) and R2 (13.9%) in Mahra relative to the other populations.

Principal component analyses were also conducted on 27 nuclear Alu polymorphic sites from a subset of 50 Yemeni individuals (current study) and a geographically widespread sample of 31 populations from throughout the world, previously published by Watkins et al. (2003). The Alu dataset provides a nuclear genome-based estimate of the relationships of Yemen to a global dataset of populations to complement the uniparental data from the mtDNA. The first PC captures 43.21% of the variation, and the 2nd PC, 15.24%, and the third PC, 7.69% (Figure 3-7). Similar to the mtDNA PC plots, the first PC primarily separates African (on the left) from non-African populations (on the right), with Middle Eastern samples in the center, including Yemeni, Bedouins, and Palestineans, due to their mixed African and Asian ancestry. It is interesting to note that the HOA samples also have central PC1 values, indicating they have experienced a great deal of admixture with Middle Eastern or Eurasian individuals. Component two generally separates all non-Africans into two groups: 1) Europeans, Central Asians, Middle Eastern, and Native Americans and 2) Indians, and Asians. Component two also separates the African pygmy populations from Zaire and the Central African Republic from all other African populations (Figure 3-7a). It is unclear what the third principal component represents (Figure 3-7b, c), but it appears to be separating Indians and Middle Eastern groups (at the bottom of plot in Figure 3-7b) from other Asian and European populations (top of plot in Figure 3-7b). The component pattern profile for the Alu data (Figure 3-8) shows PC1 to capture the widest variation of PC values with strong contributions from throughout all 27 Alu markers. B65, Col3A1, Sb19_12, Ya5NBC347, and Yb8NBC405 are the only five Alu markers with negative eigenvalues

in component one, indicating that these five markers are capturing the African component of ancestry in PC1.

Geographic Structure of Each Haplogroup in Inverse Distance Weighted (IDW) Interpolation Maps

Sub-Saharan African haplogroups

In order to further explore the geographic patterning of each of the most common Yemeni haplogroups (i.e. L(xM,N), J1b, H, R2, T1a, L2a1, K, I3/I5a, R0a) within a larger-scale inter-continental context, I created interpolation maps to visually display the geographic ranges of each haplotype/group. I was first interested in the distribution of the African L(xM,N) macrohaplogroup, as it may provide insight into the role of Yemen and the Arabian Peninsula in migration paths out of Africa and early settlement patterns within Yemen. The IDW map of macrohaplogroup L(xM,N) shows a pattern of high frequency in Eastern Africa, reducing eastwards through the Arabian Peninsula and into Eurasia (Figure 3-9). All sampled populations have a frequency of L(xM,N) greater than 1% except for the Israeli Druze, three of the four India populations (all except DB South), and the governorate of Marib in Yemen. Within Yemen, the frequency of L(xM,N) is highest in the eastern governorate of Hadramount at 44%, but is also high (>25%) in Abyan, Al Hudayda, Al Jawf, Hajjah, and Southwest Yemen. A correlogram of Moran's I versus geographic distance classes shows a strongly significant pattern of decreasing clustering of haplotypes frequencies with increasing geographic distance, indicating a significant cline in the frequency of L(xM,N) in the sampled regions (Figure 3-9b). A semivariogram cloud confirms this result, by showing increasing genetic differentiation with increasing geographic distance (Figure 3-9c). An overall structuring

of the genetic data is further supported by a global Moran's I index of 0.7172 ($p<0.0001$) for the total sample.

Because no samples in this dataset fall in the regions between the Arabian Peninsula and India, all frequency estimates (colors) in these regions are based on interpolation values from neighboring regions. In order to test for geographic patterns only in regions with high density sampling, IDW analyses were also conducted without India included in the analysis (Figure 3-10). This map did not show any change in the pattern of contour lines in Eastern Africa, the Middle East, and Arabian Peninsula. It is also possible that the very dense sampling in Yemen relative to the other countries may have biased the analysis by inflating the geographic structure within Yemen relative to the surrounding regions. Thus I also conducted an IDW analysis on L(xM,N) frequencies where all the governorates of Western Yemen (Abyan, Al Baydah, Al Hudaydah, Amran, Dhamar, SW Yemen, Al Jawf, Marib, and Hajjah) are pooled into one Western sample point and all eastern governorates (Shabwah, Hadramout, and Al Mahra) are pooled into one eastern sample point (Figure 3-11). The pattern of L(xM,N) haplotypes emerging out of Africa generally remains unchanged, as can be seen in the map and in the significant cline in the autocorrelation analysis. However, the fine-scale resolution of geographic clustering within Yemen, is no longer apparent. The higher frequency of L(xM,N) haplotypes in the East (30.3%) relative to the West (22.3%) is also not visible, as both frequencies fall within the fifth color category (0.2-.03).

I also conducted an IDW analysis on L2a1, the highest-frequency individual sub-Saharan African haplotype in Yemen (Figure 3-12). This analysis includes all derivatives of L2a1, e.g. L2a1a, and L2a1b. The IDW map of L2a1 haplotypes does not show as

clearly the West-East clinal pattern of L(xM,N) frequencies in total (Figure 3-9). This lack of a West-East cline was expected, considering the different pattern for L2a1 versus L(xM,N) in Figure 3-4, Panel B, where highest frequencies of L2a1 are found across Sudan, Amhara Ethiopians, and in Hajjah, Yemen. Unlike for the L(xM,N) macrohaplogroup, the correlogram of L2a1 does not support a significant clinal pattern, though the global Moran's I index indicates significant geographic clustering of the haplotype (I: 0.779, $p<0.0001$) suggesting that L2a1 is not distributed at random. When India is included in the map (data not shown), the Moran's I index is no longer significant (I=0.03, p=0.479), indicating that the significant clustering is concentrated in the Western portion of the map, but washed out by the inclusion of India in the analysis, where no L2a1 haplotypes are found.

In addition to L(xM,N) haplotypes, M haplotypes are also thought to potentially originate in Africa, and spread rapidly along the Southern Dispersal route to Southeast Asia, and particularly in India where they are found at very high frequencies (ranging from 44-77% (Eaaswarkhanth et al. 2010)). In Figure 3-13a, the distribution of all M haplotypes is shown, with highest frequencies in India (e.g. Mappla: 77%), and lowest in Tanzania, (<2.7%), Iraq (0%), and Syria (0%), while Yemen shows a mixture of M frequencies, with lowest frequencies in the East, and higher frequencies in the West. A statistically significant cline of decreasing clustering of haplotype frequencies with increasing geographic distance is shown in the autocorrelation plot after crossing a threshold distance of 30, which represents distances between India and all other locations (Figure 3-13b). This same threshold is evident in the semivariogram cloud in

which increasing frequency differences with increasing geographic distances is clearly evident past the same geographic distance threshold.

M1 (including all derivatives) is the most common M haplotype in Yemen (3.1%), and its distribution is shown in the IDW map in Figure 3-14a. In stark contrast to the pattern of all M haplotypes, M1 haplotypes are completely absent from India, but are frequent in the Western Arabian Peninsula and in Eastern Africa. In Yemen, M1 is found at highest frequencies in Western Yemen, particularly in Hajjah (12.5%), Al Hudaydah (7.8%), and Abyan (7.7%). Autocorrelations show no significant clinal pattern for M1, though the Global Moran's I index is significant (I=0.16, p=0.0078) (Figure 3-14b). M1a1 is the most common M1 subtype in Yemen, and in fact 12 of the 16 M1 lineages in Yemen are M1a1. Thus, the IDW map for M1a1 (Figure 3-15a) shows a similar pattern to the M1 map within Yemen, as expected, considering the near overlap of the samples. The only noticeable differences between the M1 and M1a1 maps, in fact, are located in central Saudi Arabia, which has a very high frequency of M1 (10%) but M1a1 is absent, and in Ethiopia, where M1a1 frequencies are much lower than M1 frequencies. The autocorrelation for M1a1 (Figure 3-15b) does not show a significant clinal pattern, and the semivariogram (Figure 3-15c) does not show a pattern different from random. The Global Moran's I value for the total map is not significant, (I=0.08, p-value=0.1339) indicating no geographic clustering for this haplotype different from random.

Eurasian haplogroups

Haplotypes K, T1a, and J1b all show similar trends to each other in IDW maps, with highest frequencies in western Saudi Arabia, western Yemen, Iraq, and Israel, indicating a possible migration route for these Eurasian haplotypes down the western coast of the Arabian Peninsula from Europe (Figure 3-16a). Though all three haplotypes

appear to be frequent on the coast of eastern Africa, much of the contour patterns are actually due to interpolation from neighboring regions in Arabia, and not any samples in this dataset. For example, within Africa, K is only found at varying frequencies ranging from 3 to 10% throughout Ethiopia, while T1a is only found at low frequencies (<3%) in Tigrais-speaking Ethiopians and Nubians in Sudan, and J1b is completely absent from Africa. The low frequencies for K and T1a in eastern Africa may serve as signals of a small back-migration(s) from Eurasia within the last 20,000 years. Correlograms do not support any significant clinal patterns for K, T1a, or J1b (Figure 3-16b). However, global Moran's I indices for maps K (I= 0.155, p=0.008) and J1b (I=0.366, p=0.0000) indicate significant spatial clustering for both of these haplotypes. The distribution of the T1a haplotype does not quite produce a significant global Moran's I index (I=0.093, p=0.0833), indicating a pattern that is not significantly different from random. However, when India is included in the analysis, Moran's I values increase for all three haplotypes, and for T1a, it becomes significant (I=0.1245, p=0.0329). Based on the semivariograms (Figure 3-16c), the relationship between haplotypes frequency differences and geographic distances appears random for Haplotype K, but the haplotypes frequency differences actually decrease with increasing geographic distances for haplotypes T1a and J1b.

Geographic patterning of haplogroup H (including all derivatives) shows a slightly different pattern in the Arabian Peninsula than Eurasian haplotypes K, T1a, or J1b (Figure 3-17a) as H is most concentrated in Southeast Yemen and Central Saudi Arabia (e.g. 13.9% in Mahra, and 10% in Shabwah), while K, T1a, and J1b are at highest frequency in Southwest Yemen and Western Saudi Arabia, indicating potentially

different migration paths for these haplogroups from Europe into Southern Arabia. Haplogroup H is one of the most common haplogroups in Europe (estimated to range from 44.5-48.2%, (Richards et al. 2000a)), and also is found at high frequencies throughout the Middle East: Israeli Palestineans (14%), Druze (10%), Syrians (13.6%), Central Saudi Arabians (10%), and Iraqis (9.5%). A correlogram of H haplotypes supports a statistically significant clinal pattern for H with descending significant Moran's I values particularly at larger distance classes (Figure 3-17b), and the global Moran's I index was highly significant (I: 0.301, $p<0.0001$). When India was removed from the analysis, the patterns on the map look relatively the same (data not shown) and the Moran's I was nearly identical (I: 0.287, $p<0.0001$).

Socotra haplotypes

Socotra is a large and isolated island 380km off the southeast coast of Yemen. Its long history of isolation, in part due to the strong winds of the south-west monsoon and in part due to a lack of sea-faring boats until the early 20th century, has effectively isolated the island from contact with surrounding regions leading to evolution of some of the earth's most unique flora and fauna. Genetic anthropologists have just begun to explore the human genetic variation on this unique island (Cerny et al. 2009) . In our sample of 54 Socotri individuals, haplotypes R0a, I3/I5a, J*, and R2 are the four most common haplotypes, representing 40.74%, 22.22%, 16.67%, and 5.56% of the 54 Socotri individuals, respectively. In contrast to the high frequencies in Socotra, all four haplotypes are found at lower frequencies in mainland Yemen (i.e. R0a: 17/467, 15.0%, I3/I5a: 1/467, 0.21%, and J*: 28/467, 6.00%, and R2: 15/467, 3.2%), indicating the likely role of isolation and genetic drift driving these haplotypes to higher frequency on the island. The haplotype classified as I3/I5a is especially unique to Socotra, as it was

found in no other sampled locations with the exception of one individual in Central Saudi Arabia and one individual in Hadramout (Figure 3-18a). R2 is found at its highest frequency outside Socotra in the closest governorate of Mahra (13.9%), and a few other scattered locations throughout Yemen (all <5% frequency), but nowhere outside of Yemen except Syria (2.3%) and (Shia) India (0.77%) (Table 3-3). Though R0a is found at highest frequencies in Socotra (40.7%), it is also frequent throughout all Yemeni governorates (>8% in every governorates except Hadramout where it is 5.1%), and its highest frequency outside Socotra is in the nearest governorate of Mahra (27.7%) (Figure 3-14, Table 3-3). As I3/I5a, R2, and J* are mostly found at high frequencies only in Socotra, none of the these show a significant clinal pattern in the correlograms (Figure 3-18b), and similarly the global Moran's I indices are not significant (I3/I5a: I=0.02, p=0.223, R2: I=0.06,p=0.1273, all p>=0.28). R0a, in contrast, is found at high frequencies throughout the Red Sea region, including the Saudi Arabia, Eastern Africa, and India (Figure 3-19A), and thus shows a significant Moran's I index of 0.307, p<0.0001. It has a complex correlogram, however, indicating evidence for a cline at low distance classes, but after distances of ~26, Moran's I values increase (i.e. increasing geographic clustering) (Figure 3-19b).

Discussion

This study represents a regional and inter-continental geographic analysis of mtDNA haplotypes in the Greater Red Sea Region. This area is of particular interest for human population history considering its geographic proximity to the putative birthplace of anatomically modern humans and the first migration of humans out of Africa, as well as its central location amongst the world's first civilizations. Examining Yemeni genetic data within this broader geographic context permits exploration of the genetic evidence

for 1) migration routes out of Africa, 2) settlement patterns of human populations throughout Yemen 3) slave trade from Eastern Africa, the Middle East, Europe, and SE Asia, and 4) the Hadrami Diaspora to Asia.

Out of Africa Migrations

Archaeologists and geneticists have proposed both a northern and southern migration route for the primary successful spread of modern humans out of Africa and into Eurasia. The northern route is thought to originate in Eastern Africa, and spread up through the Nile Valley of Egypt, cross over the Sinai Peninsula, and spread through the Levant (Bar-Yosef and Belfer-Cohen 2001). This hypothesis is supported by similarities in archaeological assemblages from the Nile region and the Levant (Van Peer 1998), as well as genetic evidence for dispersals from Africa bringing mtDNA lineages A and B into Central Asia (Maca-Meyer et al. 2001). The southern dispersal route is also thought to originate in Eastern Africa and cross the Red Sea into Arabia, likely via the Bab al Mandab strait at the southern tip of the Red Sea (Field and Lahr 2005; Lahr and Foley 1998). There is scarce archaeological evidence in Arabia to support this route, e.g. the few discovered lithic remains in Hadramawt, Yemen have not been sufficiently analyzed (Crassard 2009). A very recent study reports archaeological and genetic evidence of an early human presence in Arabia and South Asia earlier than recently believed, i.e. 130-70kya (Petraglia et al. 2010). Genetic data in general suggest a rapid expansion of African mtDNA along the coastal route through southern and southeast Asia to Australasia (Kivisild et al. 2004; Macaulay et al. 1999; Mellars 2006; Quintana-Murci et al. 1999; Thangaraj et al. 2006). Regardless of the route, there is general consensus that the first emergence of modern humans outside of Africa into Eurasia occurred around 54 +/-8 kya (Forster 2004).

In this geographically widespread sample of Yemen, mitochondrial haplotypes provide further support for both southern and northern dispersal routes. The sub-Saharan African L(xM,N) macrohaplogroup represents one of the largest frequency macrohaplogroups in Yemen at 22.7%, second only to macrohaplogroup R. Analyses focused only within Yemen show no evidence for a clinal pattern of African haplotypes increasing from West to East, as expected under the assumption of a southern dispersal route out of Africa. Instead, more African haplotypes are found in eastern Yemen than central/western Yemen, though with not quite as high a frequency in Hadramawt as previously reported (44.1% in current study versus 60% (Cerny et al. 2008)). When the view is broadened to include geographic regions surrounding Yemen, a pattern of decreasing frequency of African haplotypes emerging from Africa and across Arabia towards India becomes apparent, with a mixture of African and non-African haplotypes in Yemen, which is positioned in the center of the out-of-Africa cline. Assuming the haplotype frequency patterns represent potential past migration routes, the IDW map of L(xM,N) haplotypes suggests two migration routes of Africa, one northward through the Levant, and another eastward through Yemen and Southern Arabia (Figure 3-7a). However it is also possible that traces of L lineages in Yemen are alternatively due to the slave trade from Eastern Africa, as discussed below.

It is difficult to determine if the Yemeni individuals with African maternal ancestry are descendants of those individuals who arrived as part of the initial out of Africa dispersal(s) or from slave trade or other recent migrations from Eastern Africa to the Arabian Peninsula. Without other genetic or non-genetic markers (i.e. historic, archaeological, or linguistic data), it is also difficult to determine if any of the migrations

from Africa merely passed through Yemen along a coastal expansion route (as suggested by Abu-Amero et al. (2008), or settled in Yemen and later re-expanded throughout Eurasia in successive migrations. Both of these questions are particularly complex, as more recent gene flow could have introduced both early and more recently derived African lineages into Yemen. Additionally, the age of a lineage can be much older (or younger) than the age of a population, depending on stochastic evolutionary processes such as genetic drift or bottlenecks. In spite of these limitations, I draw on a collective set of refined phylogenetic analyses from previous studies in subsequent sections to provide some indication of the age of the most frequent lineages in Yemen.

Among the L(xM,N) lineages in Yemen, the most frequent are L0 lineages (22.2% of Ls), L2 lineages (24.8% of Ls), and L3 lineages (32.5% of Ls). L0 lineages are considered the most ancient L lineages, but some of the specific clades in Yemen (L0a1, L0a1b, and L0a2) are thought to be more recently evolved: e.g. L0a1 is estimated to coalesce in East Africa at 33.4 (SE 16.6) kya and L0a2 is estimated at 8.3 (SE 3.7) kya, based on rho-based estimates of time to most recent common ancestor (TMRCA) (Salas et al. 2002). Thus, these lineages could not have migrated to Yemen as early as the initial ~50kya estimated out-of-Africa migration, but could be remnants of later migrations, including the very recent slave trade from Eastern Africa. L6 lineages, which were the most frequent haplogroup in Yemeni samples collected from Kuwait, were proposed by Kivisild et al. (2004) to be a remnant of the original expansion out of Africa. It is interesting to note that L6 lineages are completely absent in the regional Yemeni samples of Cerny et al. (2008), and absent from the larger and more geographically widespread Yemeni sample of the current study. However some

lineages found in low frequencies in Yemen, such as L2a*, L2d, and L3e are estimated to coalesce (using rho-based estimates) as far back as 55,150 (SE 19.35), 121,900 (SE 34,200), and 49,250 (SE 11,750), respectively (Salas et al. 2002), and could theoretically have been introduced at the time of the initial out-of-Africa expansion. Alternatively, these ancient lineages could have been introduced to Yemen through migrations from Africa at any later time. In sum, there is little evidence of more ancient L(xM,N) lineages that may represent the initial out-of-Africa migration, and the L lineages in Africa may simply be a product of recent African slave trade. This hypothesis is further supported by the sex bias of higher L(xM,N) lineages in the mtDNA of Yemen than in male Y chromosomes, suggesting that a large portion of African haplotypes in Yemen is more likely a result of offspring of African female slaves, and not due to the initial expansion out of Africa.

Previous studies have suggested that macrohaplogroup M may also carry traces of the initial southern dispersal route out of Africa, due to the high frequency of diverse and autochthonous M haplotypes along the coasts of India and southeast Asia (Forster and Matsumura 2005; Macaulay et al. 1999; Quintana-Murci et al. 1999). In contrast, other studies argue that M lineages originate in India, and their presence in the Middle East and Africa are the result of back migrations (Gonzalez et al. 2007; Olivieri et al. 2006) or trade and military links with remote regions of Asia after British rule (Cabrera et al. 2009). In this study only 5.8% M sequences are found in Yemen overall, which are fairly evenly distributed throughout the country, with the highest frequency in Marib (2/12, 16.7%), followed by Dhamar (11/105, 10.48%), Abyan (3/39, 7.7%), Al Hudaydah (4/51, 7.80%), and Hadramawt (3/57, 5.3%). These M haplotypes (including only those

which could be definitively classified based on HVRI data alone) were compared with all M lineages in the collected dataset of previously published sequences (as listed in Table 3-1). Of the 20 unique M haplotypes in Yemen, only 3 haplotypes showed an exact match with M haplotypes outside Yemen (e.g. with Ethiopians, Israelis, Saudi Arabians, and Sudanese) and none matched exactly with haplotypes in India. The number of individuals outside of Yemen carrying exact matches with Yemeni M lineages was highest in Ethiopia (2 haplotypes shared by 11 Yemenis and 13 Ethiopians), though the most frequent M haplotype in Yemen, M1a1, (7/521, 1.3%) represented by mutations 16129-16189-16223-16249-16311-16359, is widespread throughout Eastern Africa and the Middle East: an exact match is found in 15 African individuals (13 Ethiopians including 6 Amharic-speakers, 5 Tigrais-speakers, 1 Gurage-speaker, and 1 Oromo-speaker, and 2 Sudanese including 1 Dinkan and 1 Nubian) and also in 3 Israelis (2 Palestineans and 1 Druze) and 2 West Saudi Arabians. Of the remaining Yemeni M lineages, M1a1b (defined by 16093-16129-16189-16223-16249-16311-16359) carried by 3 Yemenis is shared by 3 Ethiopians (2 Amharic, 1 Tigrais); M5 (defined by 16129) found in one Yemeni is shared by one Palestinian. Additionally, 8 haplotypes that could potentially be classified as M (but are not yet fully classified based on too few HVRI mutations) matched exactly with HVRI sequences classified as M (based on more coding region data) in Africa, the Middle East, and India. Of these 8 'potential' M lineages, 4 were found throughout Ethiopia, Iraq, Israel, Sudan, Syria, and Kenya and 7 of the 8 'potential' M Yemeni lineages matched exactly with HVRI variants found in 29 individuals in India. Thus, when considering all firmly classified and potential M lineages together, those in Yemen are nearly equally affiliated with Indian M lineages

(7 haplotype matches) as with African M lineages (6 haplotype matches), though more African M individuals carry these matching haplotypes than anywhere else.

Cerny et al. (2008) also find generally low frequencies of M in Yemen, with the highest frequencies in Hadramawt at 7.5%. Their Hadrami M samples show closest affinity with Indian M sequences (they found only 2 M1 lineages), suggesting an Asian rather than African origin of Yemeni M individuals. They interpret this as evidence supporting a back-migration from India to Yemen. The samples in the current study, however, carry many M lineages that are relatively widespread throughout Eastern Africa, the Middle East, and India, but with the most common M haplotypes in Yemen (M1) found at higher frequencies in Eastern Africa. These data suggest a closer affinity of Yemeni M lineages with African M lineages than Indian M lineages. However, as only three of the seven M lineages in Yemen that match with M lineages in India are also found in Eastern Africa, these data can not be used to definitively support or refute a back migration from India through Yemen into Africa. Though it may seem more parsimonious to assume the M1 lineages in Yemen came directly from Africa, it is also possible that the M1 lineages found in both Yemen and Africa evolved from Indian M lineages sometime after leaving India during the westward expansion through the Arabian Peninsula and into Eastern Africa. Coalescence estimates from a limited number of whole genome mtDNA sequences (n=10 M1, n=3 M1a) from Asia, Ethiopia, Senegal, Morocco, Spain, India, and the Middle East (Jordan, Georgia, and Saudi Arabia) suggest an origin of M1 at 20,774 to 42,322 YBP, and an origin of M1a at 10,759 to 30,698, using the widest range of confidence intervals for two sets of estimates based on the faster and slower mutation rates of Ingman and Mishmar,

respectively (Gonzalez et al. 2007). Based on these relatively young coalescence dates, along with the high frequency of older M lineages in India, and the similarity in distributions and timing with Eurasian U6 lineages, Gonzalez et al. (2007) and Olivieri et al. (2006) both assume that the presence of M1 in Eastern Africa and Yemen are a result of a back migration of M lineages from India. However, as M1 lineages are completely absent from India, and the coalescence dates are based on very few full genome sequences, it is still likely that the Yemeni M1 lineages could have entered Yemen directly from Eastern Africa. Though the direction and timing of the dispersal is still very much disputed, the one firm consensus in the literature is that M1 lineages spread across the southern expanse of the HOA rather than through the Levant, as M1 lineages are absent or at very low frequencies in Egypt (0.025) (Rowold et al. 2007), Syria (0%) and Israel (0.02%) (Table 3-3).

Archaeological discoveries indicate that lithic technologies in Arabia may have originated both in the Levant and in the Horn of Africa. Excavations in Oman have discovered blade arrowhead technologies (e.g. Levallois points) similar to those found in the Levant (Crassard 2009) (though no dates are available) along with the presence of domesticated sheep, goat, and/or cattle in eastern Arabian excavations suggests some of the earliest occupation of eastern Arabia may have come from the southern Levant (Uerpmann et al. 2009). These migrations may have been motivated by severely increased aridity in the northern Arabian Peninsula circa 6200 BCE, though this is considered unlikely given the equally severe conditions documented in Yemeni lakes (Amirkhanov 1996). Alternatively, migrations could have been fueled by population expansion due to favorable conditions in eastern Arabia in the early

Holocene (Amirkhanov 1996; Uerpmann et al. 2009). However, the Arabian "Bifacial Tradition" also includes some elements that do not originate in the Levant but are found only locally in South Arabia (Amirkhanov 1996) and in late Paleolithic sites of the Horn of Africa (Rose 2006). Uerpmann et al. (2009) suggest that the presence of such diverse types of artifacts could indicate two independent settlements of southeastern Arabia from both the Levant and the HOA.

Settlement Patterns Within Yemen

Though all individuals are certainly not still living in the same geographic regions as their ancestors, the ecology of Yemen has likely remained unchanged for at least the past 1000 years (Dresch 1989). Thus, subsistence strategies linked to these ecological niches may retain people in certain geographic regions across generations (Wenner 1991). In fact, it is believed that tribes have not moved much since the start of the Islamic era (Dresch 1989) and that the names of tribes and governorates have remained unchanged for nearly 3000 years (Muller 1988). Migration estimates based on locations of individual birthplace, maternal birthplace, and maternal grandmother birthplace from the Yemeni sample in this study suggests that 19.1% of the population has migrated within the last generation, but only 5.09% in the previous generation. Thus, the extreme decentralization of Yemen and its tribal composition may have established some regional genetic differences in the past that may still be detectable in contemporary genomes – though recent migrations to urban areas may have reduced this signal. It is important to remember that the migration patterns detected in this study are based solely on mitochondrial DNA inherited through the maternal line, and locations of maternal grandmother birthplaces, and thus I am only examining maternal migration history. As the genetic data are plotted on the maps in the locations of the

maternal grandmother birthplaces, the patterns detected are likely traces of more ancestral geographic patterns, rather than the most recent population movements that have occurred more frequently in the past two generations. The most interesting settlement patterns revealed by this dataset are described below.

Geographic structure in regional analyses. Regional analyses within Yemen reveal a complex mixture of haplotypes throughout all governorates. I did not find any significant linear relationship between overall geographic and genetic distances within Yemen (i.e. non-significant relationship in Mantel test based on full HVRI sequence data [data not shown]). Yemen also does not appear to fit into the global cline of haplotypes spreading eastwards out of Africa. The complex structure of Yemeni mitochondrial data is likely due to a lot of recent admixture with migratory populations passing through Yemen and/or too little time since the divergence of Yemen from its neighboring populations. However, some geographic clusters of haplotypes did not appear randomly distributed throughout the country. For example, a statistically significant higher frequency of the L(xM,N) sub-Saharan African haplotypes was found in Eastern Yemen relative to Western Yemen (Figures 3-3, 3-7), confirming the previous finding of high frequency African haplotypes in the Hadramout population of Yemen in a study with less dense sampling (Cerny et al. 2008). I also found other regional clustering: e.g. R2 appears only in the southeast corner of Yemen and Socotra (Figure 3-8), and Eurasian haplotypes K and T1a appear mainly in Western Yemen (Figure 3-9). These Eurasian haplotypes may have entered Yemen around the time of the last glacial maximum (LGM), during which individuals likely spread southwards from Europe into the southern Arabian Peninsula (and perhaps Eastern Africa) to escape the harsh northern climate

(Rowold et al. 2007). Coalescence estimates for haplotype K in Oman (17.9+/-14.2 kya), United Arab Emirates (20.2+/-11.1 kya) and Egypt (16.9+/-7.5 kya), as well as haplotype T1 (26.3+/-8.3 kya) (from global networks), and J1b (15.5+/-5 kya) (Rowold et al. 2007) are consistent with the time frame of the last glacial maximum (LGM, estimated at 25-15kya (Fagan 1990)).

The clear pattern of higher African ancestry in Eastern Yemen is surprising given its geographic distance from Africa relative to other parts of Yemen. This pattern cannot yet be explained by any significant historic or climatic event, and may in fact be a product of random drift. However, it may also reflect historical relationships between eastern Africa and the individuals who settled in eastern Yemen. Amos and Manica (2006) recognizes that genetic drift tends to operate very slowly in comparison to the timescale of human migrations, and thus "current genetic relationships likely reflect historical rather than current demography." Thus the high frequency of African haplotypes in eastern Yemen may be a relic of a past migration out of Africa that settled in eastern Yemen in the initial migrations out of Africa, but could also reflect the importation of slaves into eastern Yemen in the past few thousand years.

Considering that only one uniparentally-inherited locus is represented in these analyses, it is striking that I was able to detect this degree of regional geographic structure. Previous studies using model-based clustering methods have found strong evidence for some discrete genetic clusters among major continental regions, or strong world-wide gradients of isolation by distance, but have failed to find more regional structuring, even when using thousands of genetic markers (Prugnolle et al. 2005; Rosenberg et al. 2002; Serre and Paabo 2004). Perhaps the extremely dense regional

sampling provided more power to detect regional patterns than previous studies. Future analyses that can combine this dense regional sampling with a greater number of markers would likely provide even more informative patterns of regional geographic structuring of genetic data.

Settlement of Socotra. The ancestry of Socotri individuals has become a topic of recent interest (Cerny et al. 2009), due in part to its unique geographic position, which may have facilitated the isolation and survival of descendants from some of the first individuals to leave Africa along the southern dispersal coastal route. The mtDNA results in Socotra are in general agreement with the findings of Cerny et al. (2009), in that all Socotri individuals were classified within macrohaplogroups L, R, and N, but no M individuals were identified. Of the I3/I5a individuals, all 12 in this sample share the unique haplotype of 16086-16129-16148-16223-16391-16519 (1/12 I3/I5a individuals also carries the additional mutation of 16325). This haplotype was also found at high frequency (23%) in the Socotri sample of Cerny et al. (2009) (previously named N*) but found in no surrounding regions when compared to a database of 6,641 samples from SW Asia or Africa, except for one Hadrami Yemeni and one Central Saudi Arabian individual (Table 3-1). Based on this frequency distribution, it is likely that I3/I5a originated in Socotra and spread at very low frequencies to SE Yemen and Saudi Arabia.

Examination of the IDW maps for R2 and R0a (Figures 3-13A and 3-14A) show a clear connection between Mahra and Socotra in the elevated frequency of these haplotypes in both regions. This close genetic relationship is predicted by their geographic proximity and also supported by a large amount of historical and linguistic

evidence; e.g. Sultans of the Al Afrariya Sultante of the Mahara ruled Socotra in 1480, and 400 years later in the 1890s, many Mahra tribesmen settled in Socotra when the seat of the Mahra sultanate transferred to the island (Serjeant 1963). Linguistically, Socotri and Mahri are two of six extant Modern South Arabian languages still spoken in Yemen and Oman (others include Shehri, Bathari in Yemen, and Harsusi, and Hobyot in Oman). This language group was once thought to be ancestral to modern Arabic, but is now typically classified as a separate branch of Eastern South Semitic languages more closely related to the extinct Old South Arabian languages (i.e. Sabaean, Minaean, Qatabanian, and Hadramautic) and to Ethiopian Semitic languages (e.g. Ge'ez, Amharic, Guarage, etc.), than to the Central Semitic languages of Arabic and Hebrew (Lewis 2009). Thus, the close ancestral tie between Socotri and Mahri seems to be reflected by the shared high frequency of the rare R2 haplotype, as well as the more common R0a haplotype. The R2 haplotype (defined by HVRI position 16071) is very rare globally, and out of the samples collected in the current study, it is only found in 18 Yemeni, two Syrians, one Shia Indian, and in a search of >2100 sequences in the mtDB (http://www.genpat.uu.se/mtDB), it is only found in one other Indian individual. As R2 is found at such high frequency in Yemen, (particularly in Mahra and Socotra), it is possible that Southeast Yemen is the birthplace of this haplotype, which then spread at low frequency to Syria or other Middle Eastern regions where it is not yet detected in current samples.

Linguistic evidence places the divergence of Modern South Arabian languages from Ethiosemitic languages at ~4525 (CI=2700-6825) ybp (Kitchen et al. 2009), and the divergence of Soqotri and Jibbali from Mehri (spoken in Mahra) and Harsusi at

~1300 (CI=475-2550) ybp, placing a clear upper limit on the origin of the language group of the Socotri speakers, and perhaps the Socotri people themselves. Though too few R2 individuals are available to date the origin of this haplotype, the origin of a unique R0a1a1 haplotype recently discovered in Socotra has been estimated to 3,336+/- 2,378 ybp, and the R0a2f1 haplotype estimated to 5,881+/-1,372 ybp, with an upper limit based on the origin of R0a1 of 11,418+/-4,198 ybp (Cerny et al. 2009). Given the relatively large errors associated with these estimates, the dates for the origin of the language group and for the R0a1a1 haplotype are in relative agreement, suggesting that the settlement of Socotra and divergence of the Socotri language group may have been concurrent and relatively recent, i.e. within the past few thousand years.

Expansion of R0a. R0a is an interesting Yemeni haplotype to explore, as it is not only the highest frequency haplotype in Socotra (40.7%), but also the highest frequency haplotype in Yemen (17.7%). R0a is also common throughout Ethiopia, at frequencies >10% among all Ethiopian ethnic groups/locations, and in Saudi Arabia, where it is found >19% in Southern and Northern Saudi Arabia. In a study of Saudi Arabian full genome sequences, R0a was estimated to 19+/-7 kya (Abu-Amero et al. 2007). These Saudi Arabian samples, however, appear to represent more recent R0a derivatives, leading Abu Amero et al to propose that the Arabian Peninsula may have played a minor role in the Pre-Holocene diversification of R0a haplotypes. However, the very high frequency of R0a in Yemen, both in the current study, and that of Cerny et al. (2008), supports an older history of R0a in Yemen, and suggests that Yemen may even be the origin of the expansion of this haplotypes throughout the Middle East and into Africa.

The IDW map and correlogram for R0a in Eastern Africa and the Arabian Peninsula produced in Cerny et al. (2010) mirrors the IDW map and correlogram in this study, with the exception of the greater sampling farther East to India. The addition of Indian samples reveals a low frequency of R0a along the coast of India – reflecting a potential eastward coastal expansion out of Arabia. Based on coalescence estimates from full genome R0a sequences, R0a is thought to originate at 22,996+/-910 ybp, and R0a1 at 22,559+/-1,472 ybp. These dates are older than those estimated by Abu Amero et al. (2007), though they fall within their larger confidence interval. Based on these coalescence dates, and the current distribution of R0a haplotypes, R0a likely did not expand out of Africa with the initial ~50,000 ybp expansion, but probably evolved from R lineages in the Near East, where the oldest R haplotypes are found (Richards et al. 2000a). R0a then likely expanded in the Arabian Peninsula, (perhaps in Yemen where it shows highest frequency) between 20-25,000 years ago, and then spread throughout the Middle East and back into Ethiopia, where coalescence dates for the most common African subtypes are relatively recent, e.g. R0a2a: 7,219+/-2156 ybp and R0a2b: 11,296+/-2733 ybp (Cerny et al. 2010).

Slave Trade

The geographic patterns of African ancestry in this dataset may in part be due to slave trade throughout the Islamic world, which was common from as far back as 2500 years ago until the abolition of slavery in 1962 in both Yemen and Saudi Arabia. This pervasive practice forced migrations of people mainly from northern, eastern, and Sub-Saharan Africa, but also from Central Asia, Iran, India, and Spain into Yemen and other areas of the Arabian Peninsula (Lewis and Lewis 1990). Africa was the primary continental source of slaves by the 19th Century, but was also a popular source

throughout the past 1700 years, in which the majority of the slaves were Abyssinians (aka Ethiopians), sold through important slave trading centers in Zabid, Aden, and Tihama in the 4th, 11th, and 17th centuries CE, respectively (Abdallah al-Amri 1988). I did not expect to find genetic evidence of largely African groups in these areas, as our sampling strategy targeted only tribal Yemeni people, and not the marginalized Akhdam community ("servant" in Arabic), which are considered to be the offspring of African slaves (2005). Even without considering the Akhdam, however, the rise of Islam created the first inter-racial society through conversion and manumission of African women and children, which may have left some trace of their African genetic origin within the larger Yemeni population.

Higher traces of African ancestry in the mtDNA of the Yemeni more than among the Y chromosomes, is expected, as male slaves were often used as eunuchs or soldiers, and left few descendants relative to women among the Arab populations (Lewis and Lewis 1990; Richards et al. 2003). Islamic rule on miscegenation, in fact, legally permitted racial intermarriage, though in practice racial/religious taboos still limited marriages outside Muslim faith (Lewis and Lewis 1990). This rule was more supportive of Muslim men marrying or mating with female concubines of non-Muslim faith than the reverse, because the Muslim faith would be upheld through the male lineage and passed onto the children. The practice of taking black women as concubines was particularly acceptable with Nubian, Nilotic, and Ethiopian women (often referred to as Habash, who were the most highly esteemed), compared with Zanj women from Bantu-speaking tribes of southeast Africa (Lewis and Lewis 1990). Ancient literature attests to the frequency of this practice, considering the number of Arab poetry

and legends with famous figures born to Arabic fathers and Ethiopian mothers (Lewis and Lewis 1990).

The popularity of black concubines and wives among Arab men may explain the sex bias previously reported in the genetic ancestry of the Yemeni population. Specifically, the maternally-inherited mtDNA has more African ancestry than the Y chromosome in Arabia (Richards et al. 2003). Preliminary Y SNP data in the current study also support this sex bias, with very little paternal African ancestry in Yemen, as essentially all Y chromosomes share a common Middle Eastern type, 41/50, J1/M267. The remaining Y samples in our dataset are classified as a single J* haplotype and four individuals each with different E haplotypes (E3b1c*/M123E2/M75, E3b1a/M78, E/M35). These results are comparable to another Y chromosome study, where 72.6% (45/62) of Yemeni Y chromosomes were classified as J1/M267, and the remaining samples were classified as J2 (9.6%), E3b subtypes (16.1%), or G2 (1.6%) (Cadenas et al. 2008). The common J1/M267 haplotype in Yemen is also found at high frequencies in the Arabian Peninsula (58.3% in Qatar, 34.8% in United Arab Emirates, and >33% in Oman (Cadenas et al. 2008), at varying frequencies throughout Central Asia: Turkey (8.41%) (Cinnioglu et al. 2004), Northern and Southern Iran (9.09%, 11.97%, respectively), (Regueiro et al. 2006), Egypt (Luis et al. 2004), and at low frequencies in both Pakistan (Sengupta et al. 2006) and Tajikistan (Regueiro et al. unpublished, cited in Cadenas et al. 2008). J1/M267 is virtually absent from the HOA, including from Somalia (Sanchez et al. 2005), Kenya (Luis et al. 2004), but M267 was not typed Oromo and Amharic speaking Ethiopians (Semino et al. 2002; Underhill et al. 2001; Underhill et al. 2000). The E3b1c-M123 haplotype is only found in 2% of our Yemeni sample, but in Cadenas

et al. (2008) Yemeni sample it is found at 8.1%. This haplotype is found at low frequencies throughout the Near East (3.1% in UAE, 1.4% in Qatar) and Egypt, but is completely lacking in Kenya (Luis et al. 2004), Tanzania (Underhill et al. 2000), and Sudan (Cruciani et al. 2004).

Researchers have proposed that the African Y chromosomes found in Yemen are descendents of more ancient African migrations than the African mtDNA haplotypes in Yemen, which may derive from more recent slave trade (Cadenas et al. 2008; Richards et al. 2003). The distribution of E3b haplotypes in the Arabian Peninsula are thought to support a northern dispersal route of male lineages through North Africa and then out of Africa through the Levant. E3b1 subtype M35 in particular is not believed to have arrived in Arabia along the southern dispersal route because of the complete absence of E3b1*-M35 from Oman and Qatar, and very low frequency in Yemen (3.2%), but high frequency in Eastern Africa. Additionally, the near absence of M78 from Yemen (0%) and Oman (1.7%), but high frequency in Ethiopia (22.7%), (Underhill et al. 2000) and Somalia (77.6%) (Sanchez et al. 2005) also supports a lack of connection between HOA and the nearest Arabian countries across the Red Sea (as discussed in Cadenas et al. (2008)). These E3b haplotypes are also likely to represent a more ancient dispersal than the slave trade of the past 2500 years due to the estimated coalescence dates: e.g. E3b1c-M123 is estimated at 10.6+/-4.1 kya in Yemen, and at 11.1kya in the UAE, based on 15 YSTR loci and the SNP-STR coalescence methods (Cadenas et al. 2008).

Estimates of African ancestry were also generated on the Yemeni data in the current study (by Raaum) from preliminary nuclear data based on ~370,000 SNP

genotypes from a geographically widespread subset of 62 Yemeni samples using the program HAPMIX (Price et al. 2009). Ancestral populations were selected from HapMap samples of two Bantu-speaking groups: Yorubans in Idaban Nigeria (YRI), and Luhya in Webuye Kenya (LWK). One Nilotic-speaking group of Maasai in Kinyawa, Kenya (MKK) was also tested, along with a group of white Americans from the CEPH panel with northern and western European ancestry (CEU). The proportion of African ancestry was estimated based on three separate analyses using the European reference population paired with each of the African populations. A low degree of African ancestry was estimated from calculations with the Bantu-speaking populations (CEU-YRI mean 0.062, sd 0.044) and CEU-LWK (mean 0.076, sd 0.049). No significant difference was found between these two estimates ($p=0.13$, two tailed Welch two-sample t test). Proportion of African ancestry using the Maasai population however, produced estimates more than twice as high (CEU-MKK mean 0.148, sd 0.060), suggesting closer affinity with contemporary Maasai than with the Yoruba or Luhya (unpublished data by Raaum). Even considering this wide range of estimates of African ancestry (from 6-14%), the nuclear DNA support a lower frequency of African ancestry than the 22.07% of L(xM,N) mitochondrial haplotypes. This low frequency of African ancestry in the nuclear genome suggests that a maternal contribution from Africa likely arrived in Yemen mostly at one initial time point, as continual admixture would be expected to produce higher African ancestry in the nuclear genome than observed in the STR data. Additionally, the lack of African Y chromosomes in Yemen further supports the sex bias in Arabian populations which all shower higher female African ancestry.

Hadrami Diaspora

The earliest recorded reports of Hadrami migrations from Yemen to India come from those participating in Islamic conquests in the 7th Century CE (Ho 2006). Through military invasions, Muslim Kingdoms were established in India, followed by waves of mercenaries, businessmen, and political emissaries from Iran and the Arabian Peninsula, who were thought to admix with the local Indian population (Eaaswarkhanth et al. 2010). Historians believe that the height of the Hadrami Diaspora brought migrants from Hadramout to Asia in the early 1500s. Thus, one would expect to see some Yemeni haplotypes in India as a result of this very recent migration. In this study, the only Yemeni haplotypes found in India were J1b (4 individuals), T (2 T1a and 3 T2), H (only 1 individual, with the Cambridge Reference sequence), R2 (only 1 individual), R0a (13 individuals), M derivatives (245 individuals, though none with the frequent Yemeni haplotypes of M1 or M1a1), and L0a2a2. Among the J1b lineages, none are exact matches with Yemeni J1b, though one individual is one mutation away from a Yemeni haplotype. Among the T lineages, one Indian individual is an exact match (16126-16163-16186-16189-16294) with four Yemeni individuals. Among the 13 R0a lineages, there is one haplotype shared by nine Indians and 24 Yemeni individuals (16126-16362). Among all the 245 Indians with M lineages, only 29 individuals (seven haplotypes) show exact matches with M lineages in Yemen, and these were only among the Yemeni M lineages that could not be definitively classified as M based only on HVRI sequence variants, as discussed above. Finally, the L0a2a2 haplotype represented by three individuals in the Dawoodi Bohra South population show an exact match to five Yemeni individuals (three in Al Hudaydah, one in Dhamar, and one in Amran). Of all these lineages, only haplotypes R0a is found primarily in Eastern Yemen,

and also at a frequency greater than 1 individual in India (Figure 3-14). Thus, a very low level of identical individuals, or relatedness in general, is seen between Yemen and India, suggesting that very few Yemeni migrants moved and settled in India in the past 500 years. The PCA plots (Figure 3-5) suggest a closer relationship between the Indian populations with East Africans and Hadramout, than to other Middle Eastern populations. This relationship with Africa is slightly misleading, and is probably due to the fact that India does not share any of the most frequent Yemeni haplotypes used to make the PCA plots – much like Hadramout and Africa, which also have low frequencies of the Middle Eastern haplotypes.

These results are consistent with the results of Eaaswarkhanth et al. 2010, who purposefully sampled Muslim Indian groups, in search of the origin of Indian Muslims, which was hypothesized to be a result of migrations of people from the Middle East or simply the migration of cultural practices, in which a large population of Indians may have adopted Islam (Eaaswarkhanth et al. 2010). They do not find any genetic affinity in the mtDNA between the five Muslim Indian populations (Dawoodi Bohra West, Dawoodi Bohra South, Shia, Sunni, or Mappla) with Middle Eastern populations, including Iraq, Jordan, Syria, Iran, Arabia, Qatar, Bedouins, Turkey, or the United Arab Emirates, but instead closer affinity to Pakistan and other non-Muslim Indian populations. They expected to see higher affinity for the Y chromosome, due to reports of marriages between Muslim men and Hindu women, but again do not find higher affinity of Indian Muslim populations with Middle Eastern Y chromosomes, but instead see closer affinity to local non-Muslim Indian populations. Only one Muslim Indian population of recent Shia immigrants from Iran (excluded from the current analysis) showed genetic affinity

to Middle Eastern populations. According to oral history, these Iranian Shia were thought to have emigrated from Iran only three generations back "for business" purposes, though historical records do attest to Shia communities emigrating from Iran from the 16th-18th centuries (Eaaswarkhanth et al. 2010). In comparing the same Muslim Indians to our large Yemeni population, and other Middle Eastern populations, we see a similar pattern of very few shared mtDNA haplotypes. One exception is the L0a2a2 individuals that show exact matches between individuals in the Dawoodi Bohra South population and those in Western Yemen. This shared haplotypes is particularly interesting, as Eaaswarkhanth et al. (2010) had hypothesized L0a2a2 individuals in Dawoodi Bohra South to be evidence of a migration of a Shia sect of Islam from Yemen, where L0a2 is found at ~3% (based on Kivisild et al. (2004) samples). Using our larger and more widespread Yemeni sample, this study confirms an exact haplotype match between Yemeni and Indian L0a2a2 lineages, indicating a relatively recent (though likely small) migration from Yemen to India. This migration may have been motivated by severe persecution by the Sunni majority of the Shia in Yemen sometime after the 12th Century CE, and is unlikely to be a remnant of the initial out-of-Africa migrations, due its young age estimate (10,300 ybp) and its absence from other non-Muslim Indian populations (Eaaswarkhanth et al. 2010).

Conclusion

These analyses of the geographic structure of the mitochondrial genetic diversity in Yemen and surrounding regions provides new insights into human population history in this complex and highly admixed region of the world. This history is particularly interesting as Yemen occupies the first stop in the purported southern dispersal route out of Africa, and is a central location among the world's first large civilizations. In

general, the Yemeni mtDNA shows ancestry from Eastern Africa primarily in the East, and from a mixture of the Middle East and Eurasia in the West, attesting to complex history of settlement patterns due in part to the initial migrations out of Africa and more likely to later African slave trade, and in part to later migrations down the Arabian Peninsula from Europe. The high frequency of African mtDNA haplotypes (but not Y chromosomes) in Yemen is likely due to African slave trade over the last 2500 years in which children from unions of African women and Yemeni men were absorbed into the Yemeni population. In regional analyses, the data support close connections between the island of Socotra and the closest governorate of Mahra. There is also evidence that Yemen may be the origin of the expansion of R0a haplotypes throughout East Africa and the Middle East and India. Finally, there is limited genetic evidence for the Hadrami Diaspora to India, but a few haplotypes (L0a2a2 and R0a) attest to some migration from throughout Yemen into the Dawoodi Bohra populations of South and West India. In sum, only by interpreting genetic data within a framework of broad-scale geographic sampling, detailed GIS data, and inclusion of historical, archaeological, and linguistic information, can we achieve a comprehensive picture of the evolutionary history of Yemen and surrounding regions around the Red Sea.

Table 3-1. Populations

Continent Country	Region/Language Group	N	Reference
Middle East			
Saudi Arabia	Total	73	Abu-Amero 2008
	Central	20	
	Northern	18	
	Western	19	
	Southern	16	
Syria (Northern)	Total	44	Vernesi et al 2001

Table 3-1. Continued

Iraq (Arabic)	Total	42	Richards et al 2000
Israel	Total	146	Richards et al 2000
	Palestinian	100	
	Druze	46	
Yemen	Total	521	Current Study
	Abyan	39	
	Al Bayda	47	
	Al Huday	51	
	Al Jawf	12	
	Amran	32	
	Dhamar	114	
	Hadramout	58	
	Hajjah	9	
	Mahra	65	
	Marib	16	
	SW Yemen*	14	
	Shabwah	10	
	Socotra	54	
Middle East Total		826	
Africa			
Ethiopia	Total	223	Kivisild et al. 2004
	Amhara	113	
	Gurage	20	
	Tigrais	44	
	Oromo	30	
	Afar	16	
Sudan	Total	129	Krings et al 1999
	Nubia	75	
	Nuer	11	
	Dinka	43	
Tanzania	Total	271	Tishkoff et al. 2007
	Hadza	76	
	Sandawe	79	
	Datog	37	
	Turu	31	
	Burunge	37	
	Sukuma	11	
Kenya	Total	68	Watson et al 1996
	Kikuyu	19	
	Turkana	34	
	Somail	15	

Table 3-1. Continued

Africa Total		691	
Asia			
India	Total	424	Eaaswarkhanth et al. 2010
	Dawoodi Bohra - West	50	
	Dawoodi Bohra - South	62	
	Indian Shia	120	
	Indian Sunni	131	
	Mappla	61	
Asia Total		424	
Global Total		1941	

Table of all samples analyzed in the current study. Numbers in bold represent the total sample size for each country and those underlined represent totals for each continental region. *SW Yemen includes 7 Lahij, 4 Taizz, and 3 Aden. Sample sizes (n) represent the number of samples from each region for which a single macrohaplogroup could be classified.

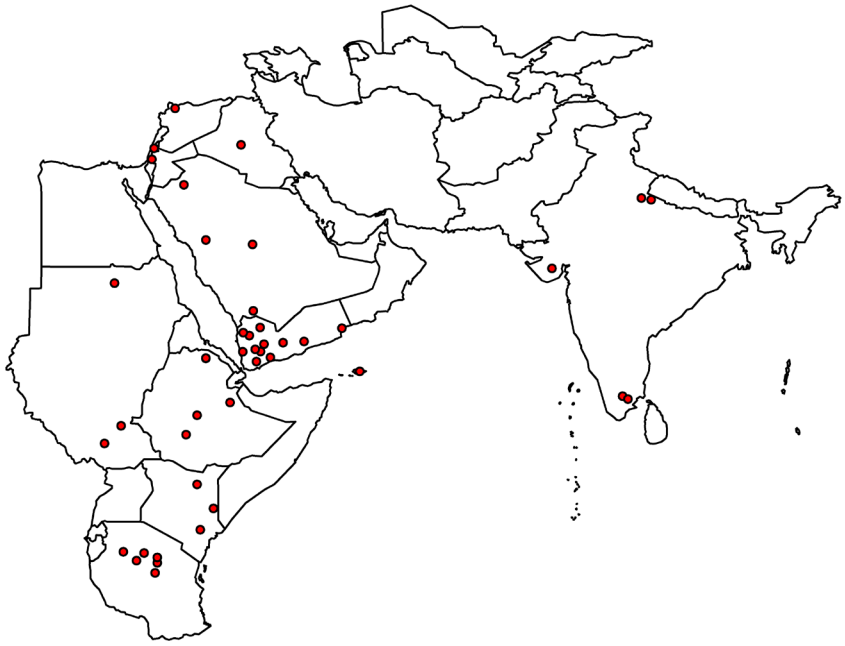


Figure 3-1. Sample locations. Map of the geographic locations of the maternal grandmother birthplaces of all the samples collected in Yemen and the sample locations of all the samples collected from previously published data in surrounding regions, as estimated from maps or locations described in the original publications (e.g. Burchard et al. 2003; Haga and Venter 2003; Kaufman and Cooper 2001; Lee et al. 2008; Lee et al. 2001; Sankar et al. 2007; Stevens 2003 among many others).

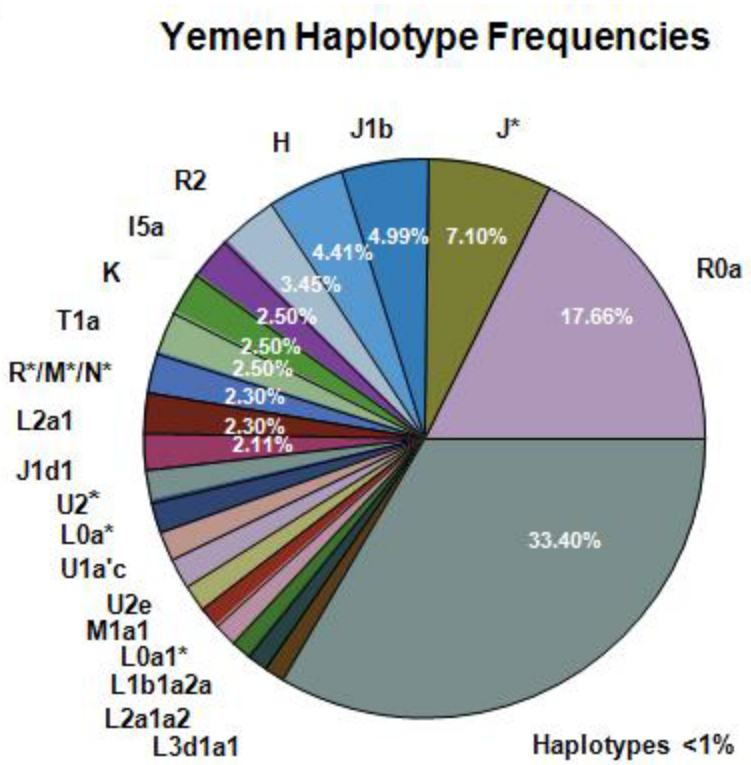


Figure 3-2. Frequency of Yemen haplotypes. Pie chart of frequencies of all haplotypes found in Yemen, listed at the finest resolution at which they could be classified, based on HVRI sequence data. All haplotypes present below 1% are pooled in the bottom right quadrant of the chart.

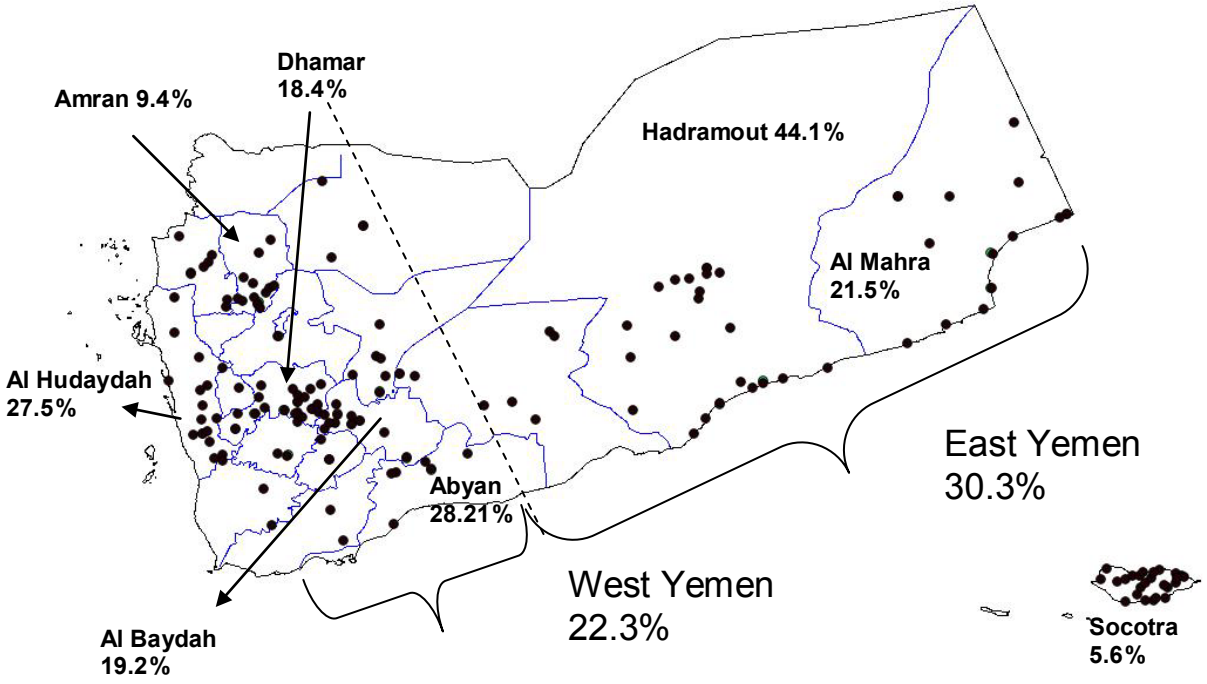


Figure 3-3. African haplotypes throughout Yemen. Map of sample locations for maternal grandmother birthplaces of all Yemeni samples. Only those governorates with sample sizes large enough to conduct multiple tests of proportions are labeled, with associated L(xM,N) haplotypes frequencies. The dotted line indicates where the division between West and East Yemen was placed for the two sample test of proportions; Socotra was not included in either East or West Yemen groupings.

Table 3-2. AMOVA Results: Proportion of variance assigned to each sub-grouping.

	East vs. West	Island vs. Mainland
Between Population	0.88% (p=0.000)	5.13% (p=0.000)
Within Population	99.12%	94.87%

AMOVA analyses reveal that the Island Mainland division was the only significant apportioning of variance between geographic regions. The East-West division was based on the line shown in Figure 3-3, with Shabwah, Hadramout, and Mahra grouped in the East, and all other governorates grouped in the West.

A.

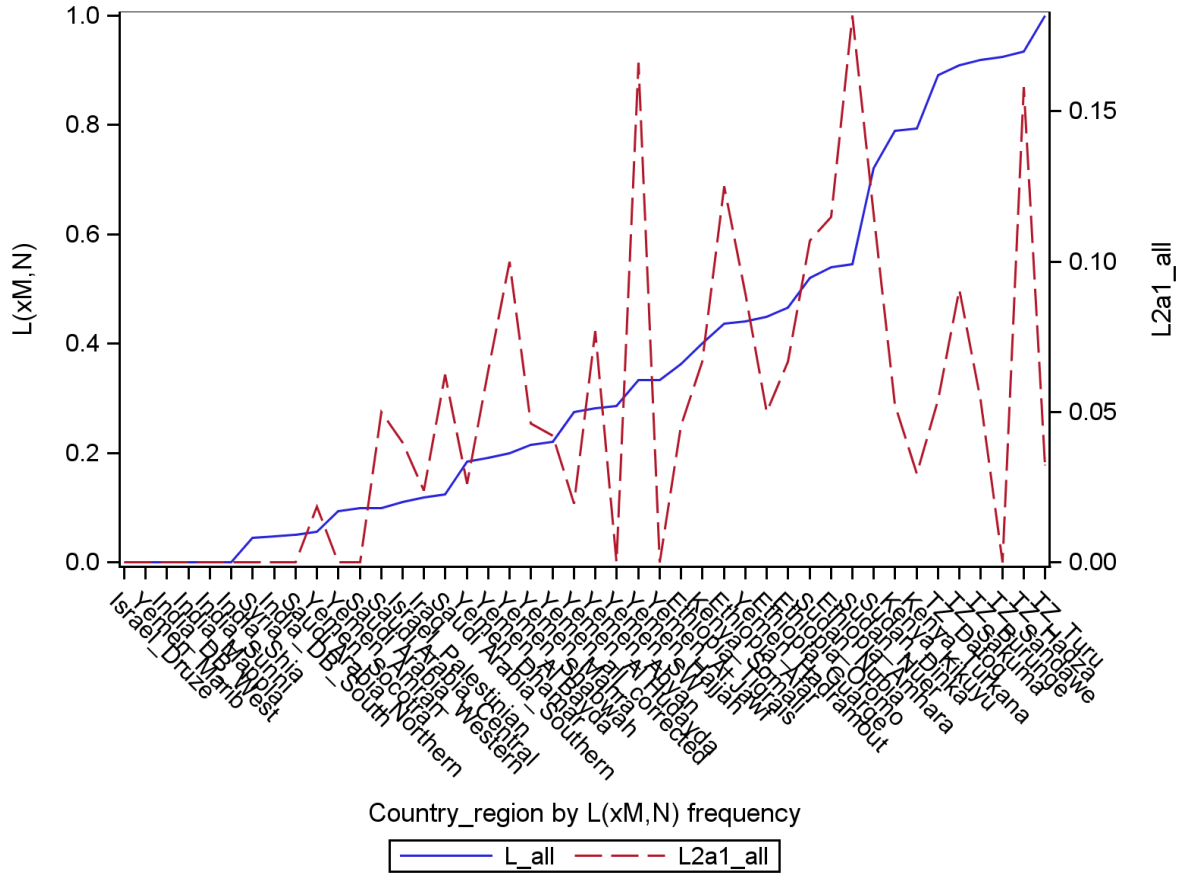


Figure 3-4. Frequency of African L(xM,N) haplotypes by longitude (A) and in increasing frequency (B). Plot A shows the frequency of L(xM,N) haplotypes in order of longitude from the most western population on the far left to the most eastern population on the far right. In Plot B, the blue line represents the frequency of L(xM,N) haplotypes in order of increasing frequency, while the red line represents the frequency of the L2a1 haplotype (including all derivatives of L2a1).

B.

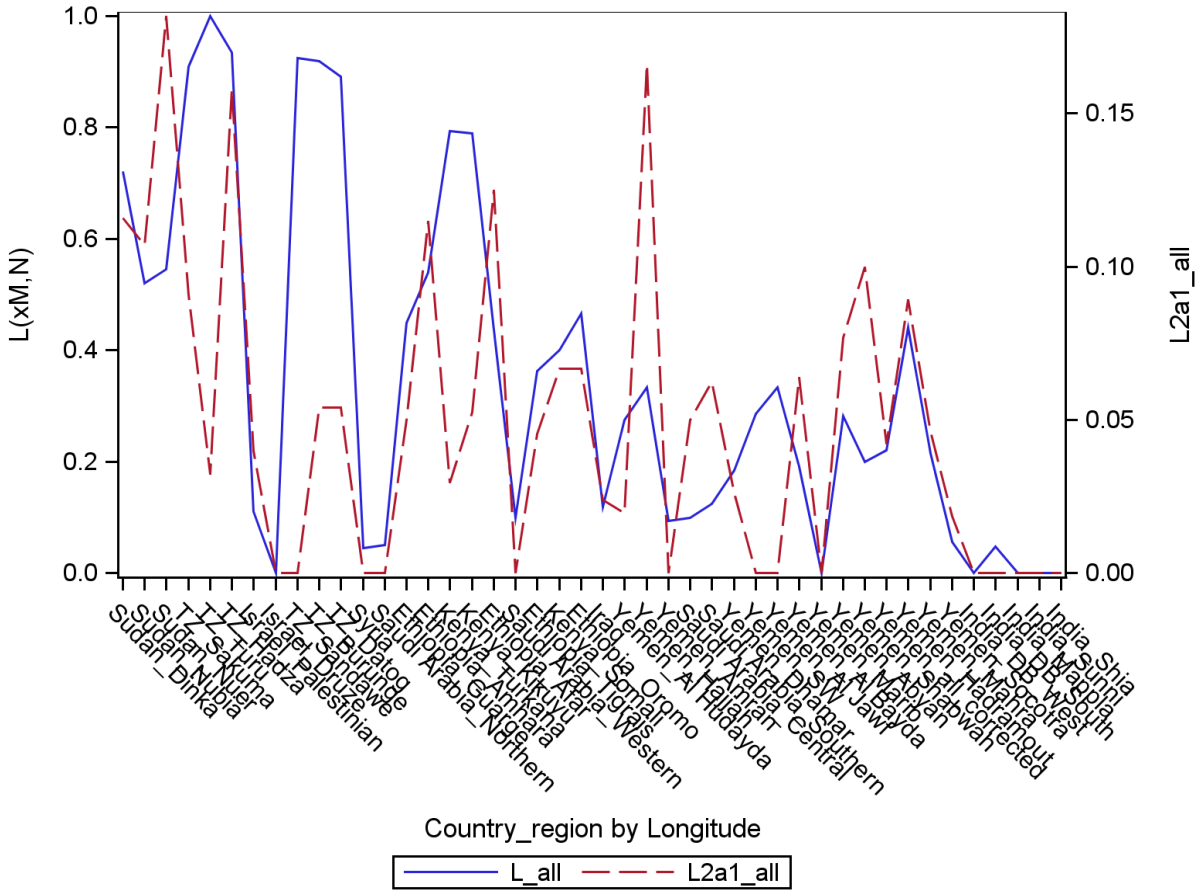


Figure 3-4. Continued.

A

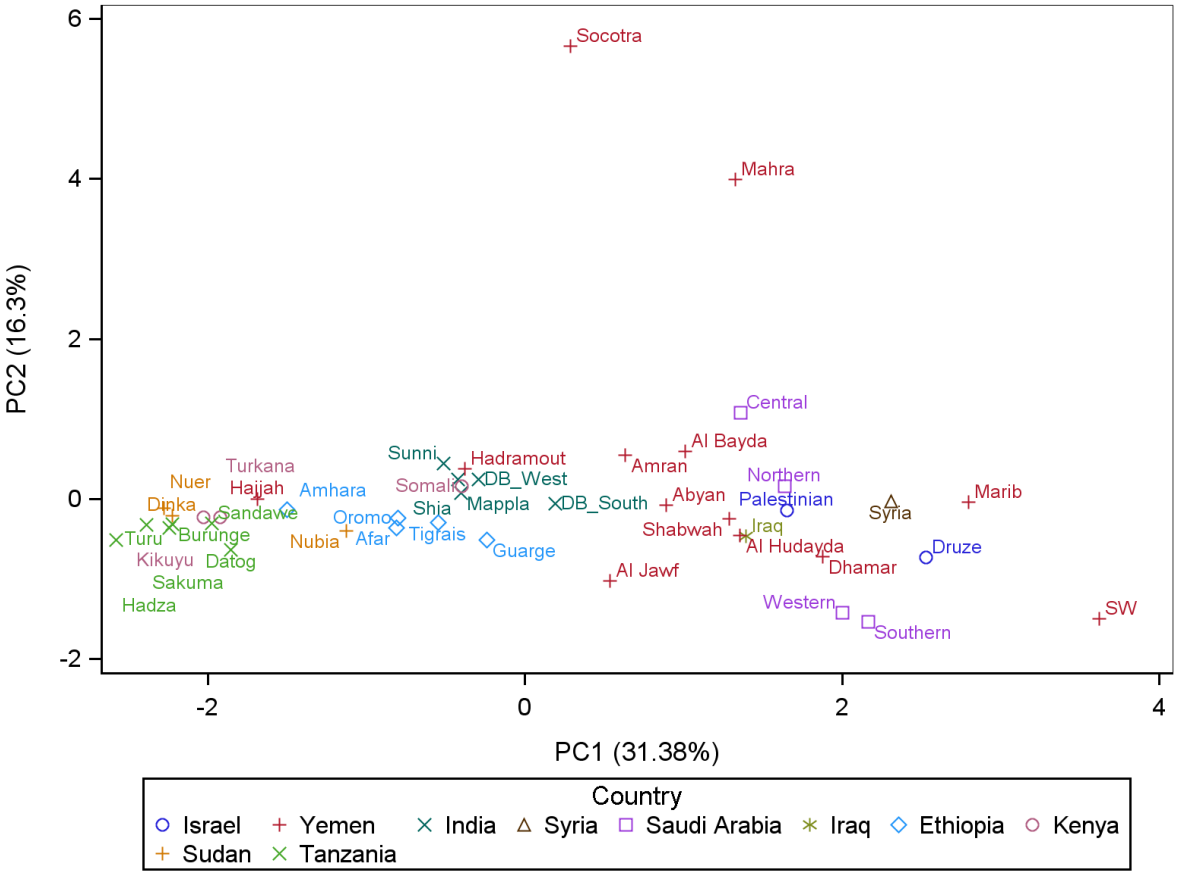


Figure 3-5 Principal component plots based on mtDNA haplotype frequencies. PCA plot of A) PC 1 against PC2, B) PC2 against PC3, and C) PC1 against PC4, based on haplotype frequencies of the most frequent Yemeni mtDNA haplogroups: L(xM,N), J1b, H, R2, T1a, L2a1, K, I5a, and R0a. Labels refer to regions within each country. SW=Southwest Yemen (Lahij, Taizz, and Aden pooled), DB_West=Dawoodi Bohra West India, DB_South=Dawoodi Bohra South India, and Sunni and Shia are both Muslim populations within India.

B

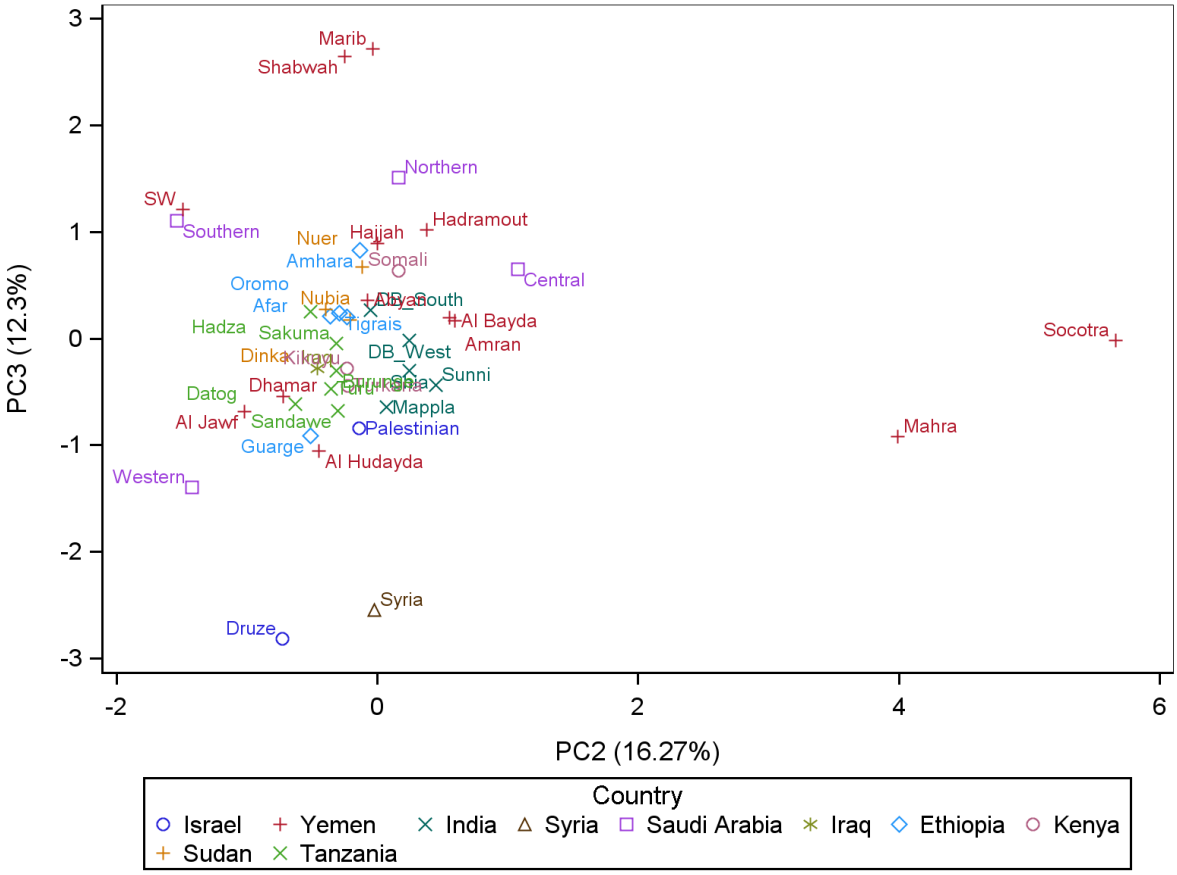


Figure 3-5. Continued

C.

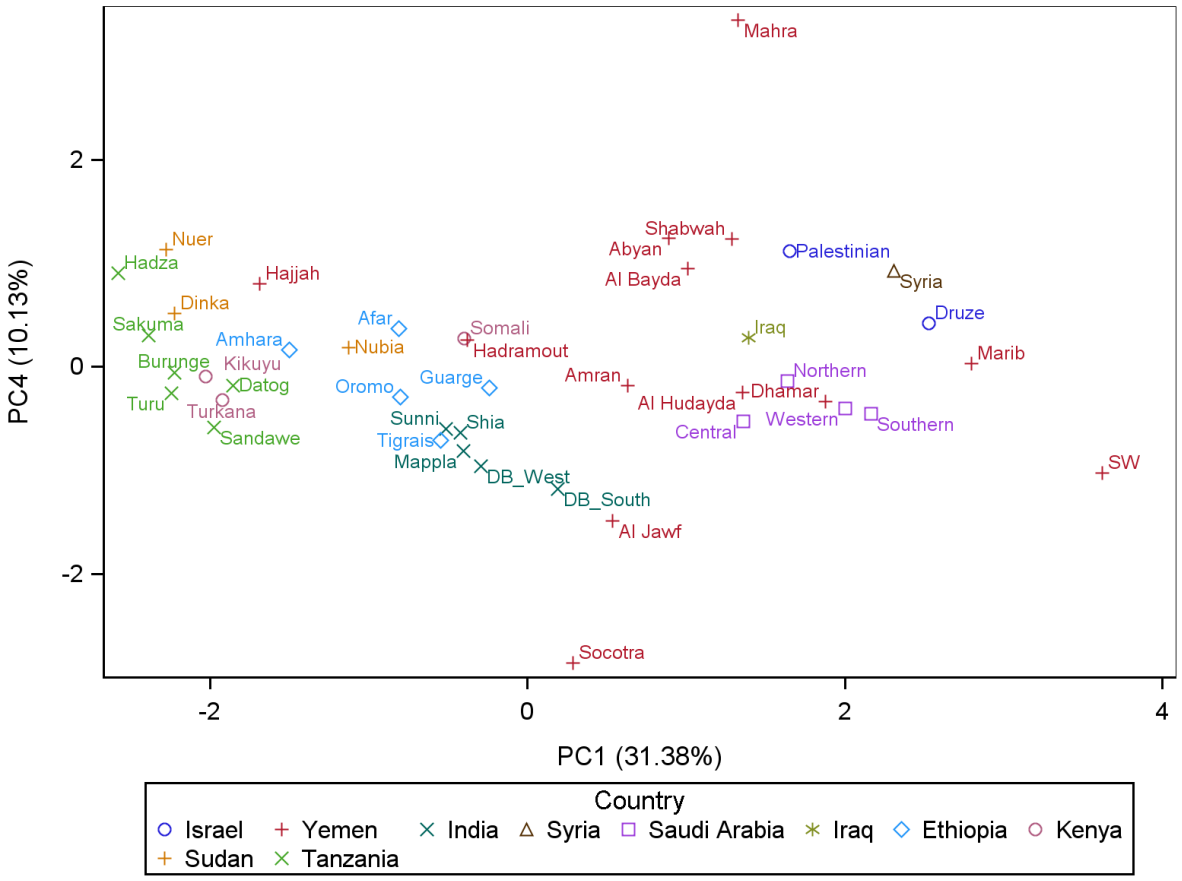


Figure 3-5 Continued

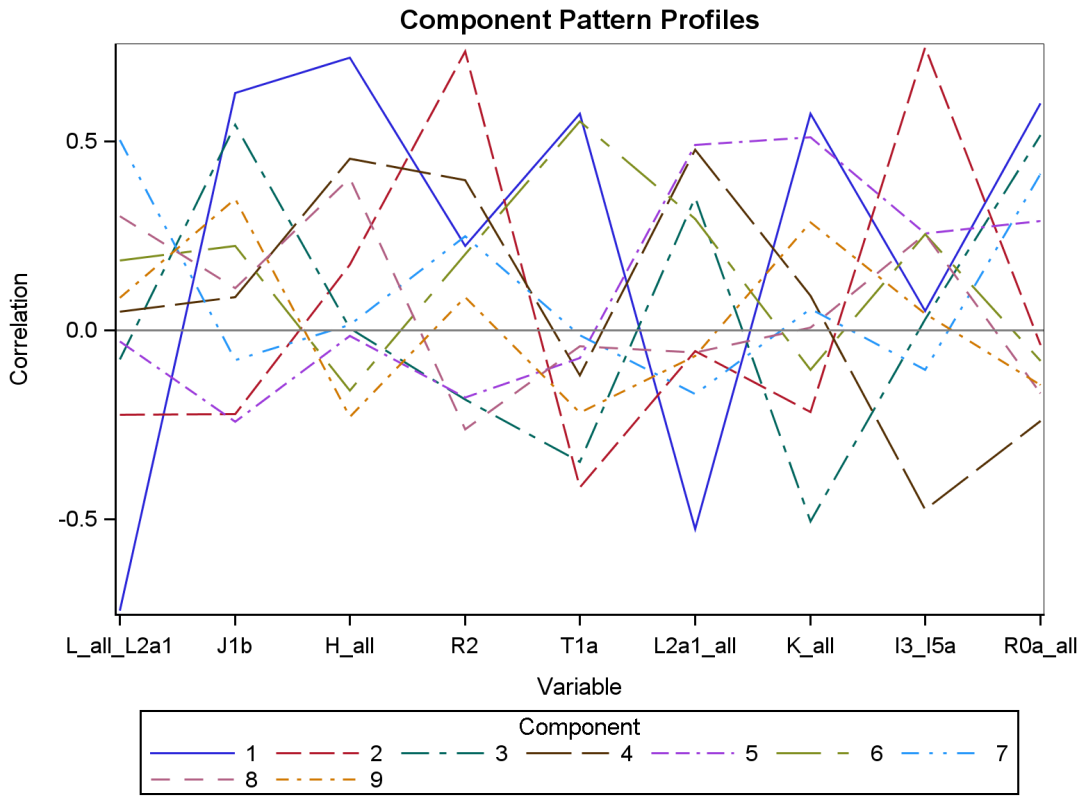


Figure 3-6. Component pattern profiles for mtDNA data. Plot of the contribution of each haplotype to each principle component displayed in Figure 3-5. Haplotypes with positive correlation values push samples to the righ (or top) of the PC axis while samples with negative correlation values push samples to the left (or bottom). L_all_L2a1 indicates all L(xM,N) samples with the L2a1 samples removed. I3_I5a indicates samples that could not be classified finer than I3/I5a.

A

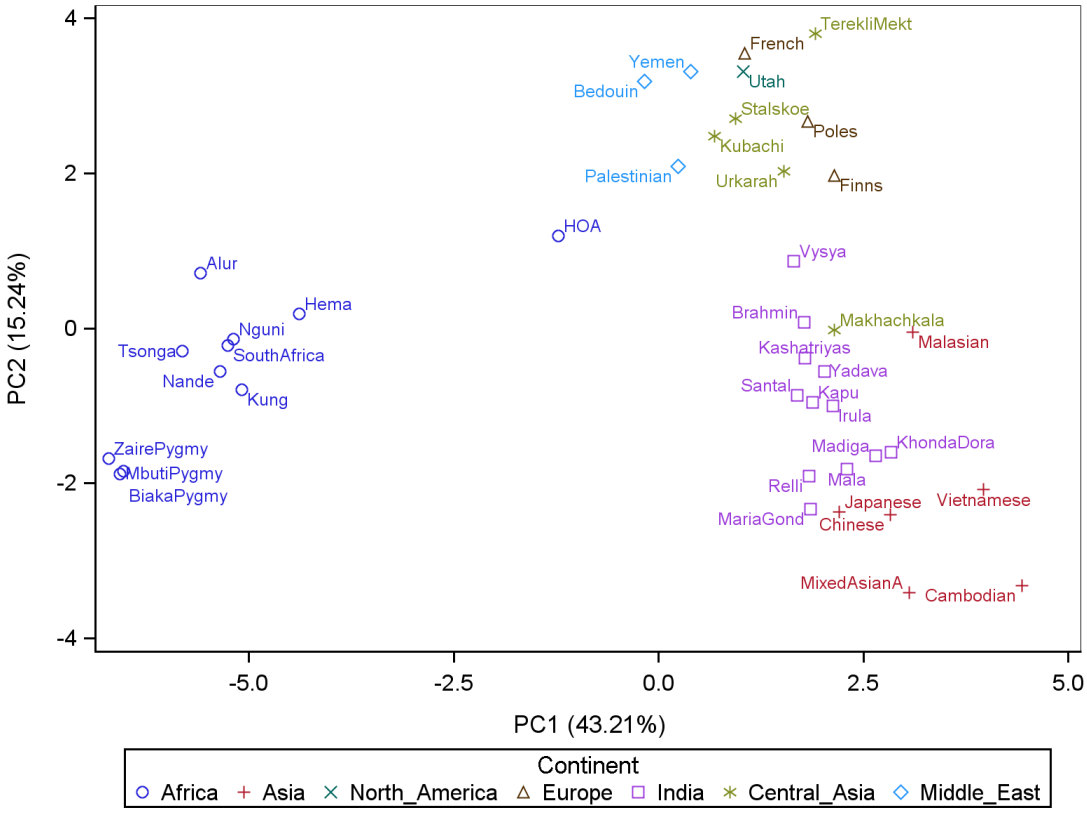


Figure 3-7. Principal component analyses of Alu polymorphism data. PCA plot of A) PC 1 against PC2, B) PC1 against PC3, and C) PC1 against PC3, based on Alu marker frequencies for 27 Alus. Labels refer to regions within each country while the legend lists continental groups.

B

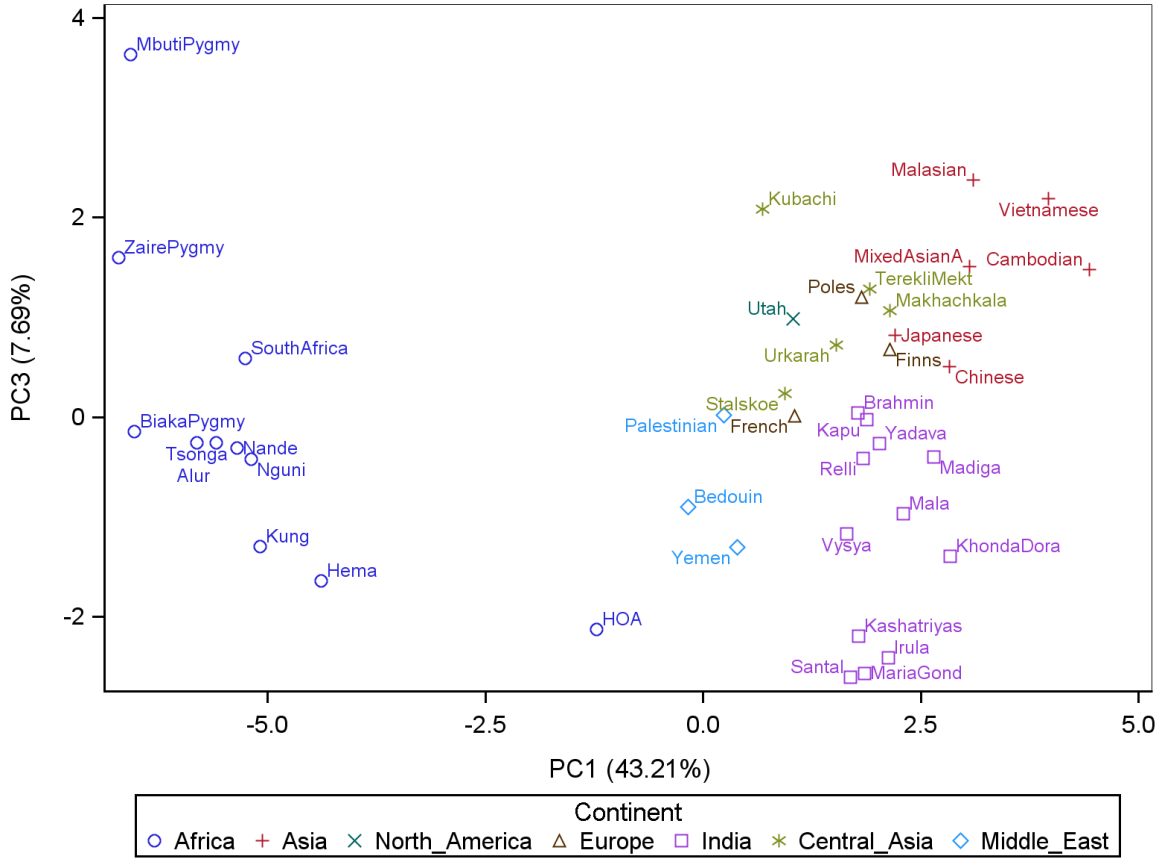


Figure 3-7. Continued

C

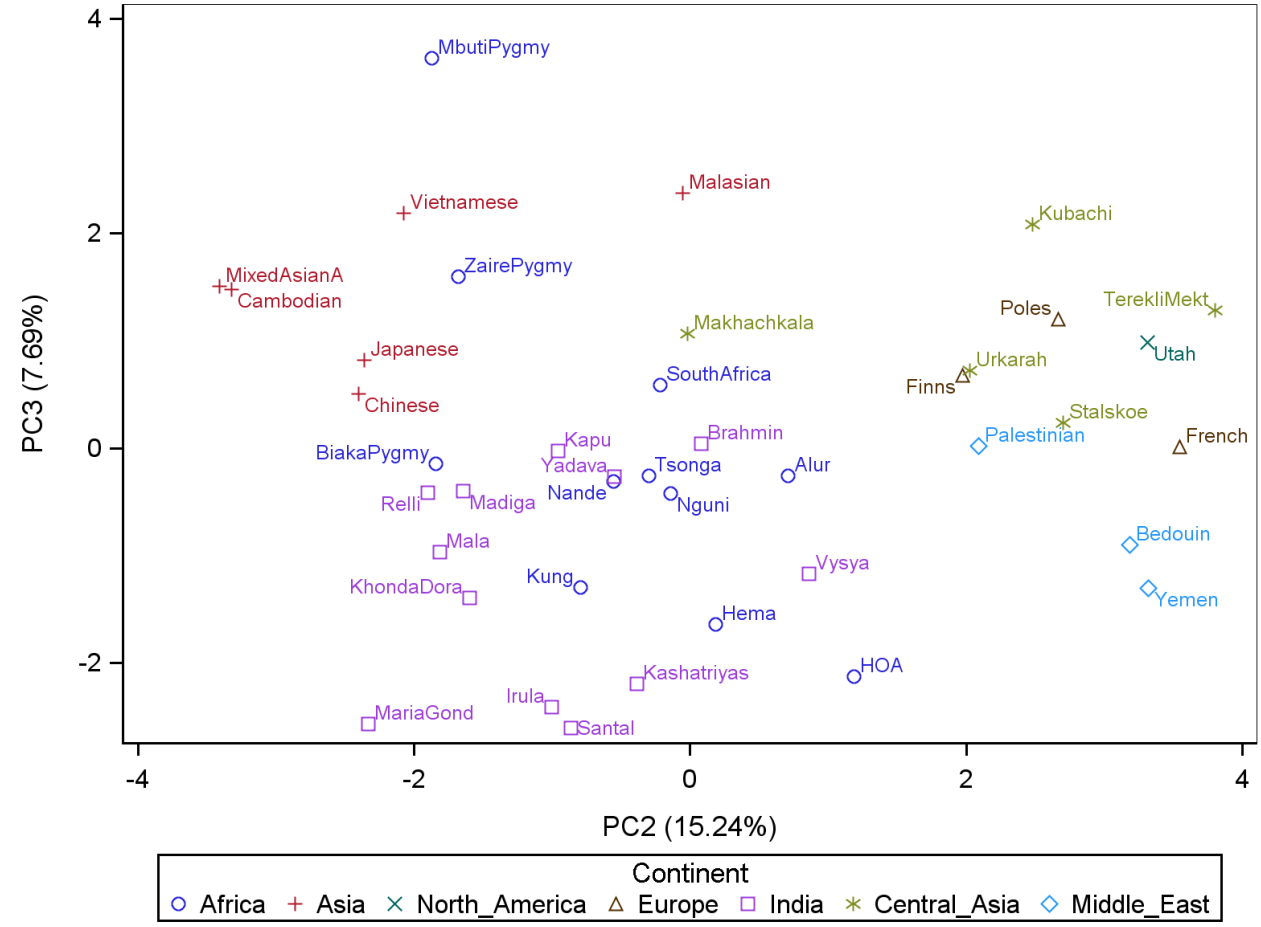


Figure 3-7. Continued

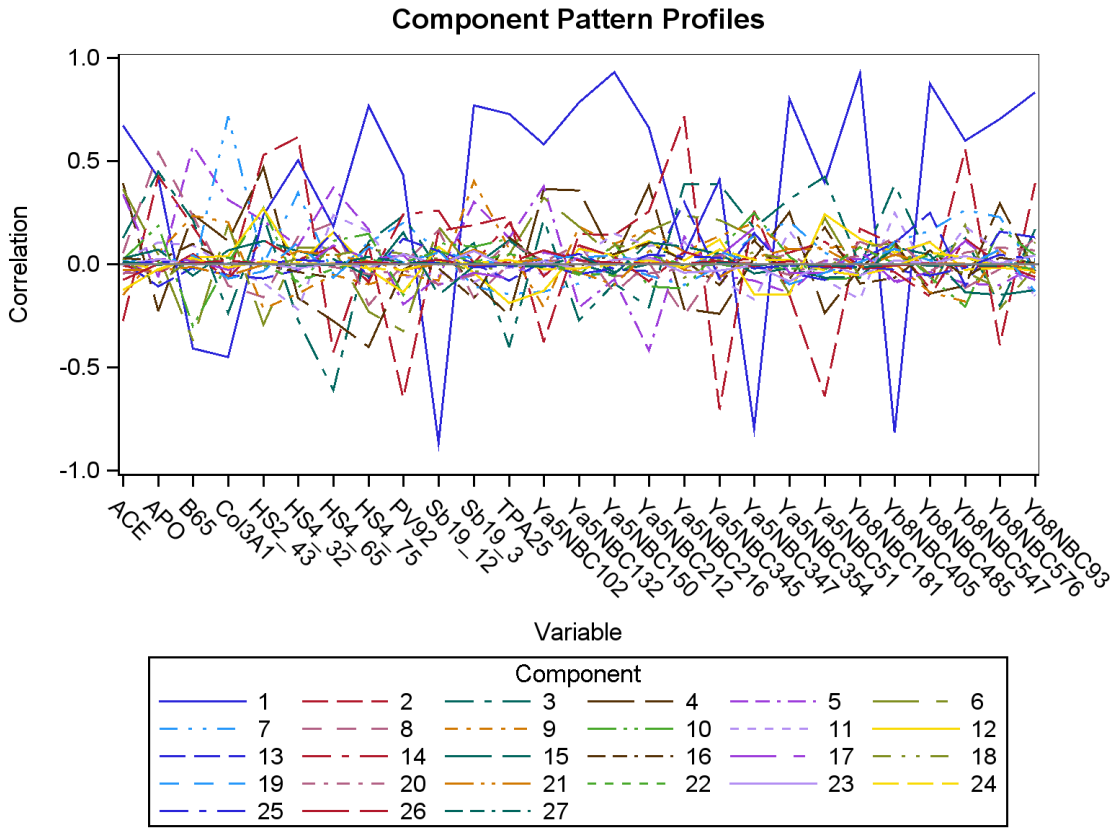


Figure 3-8. Component pattern profile for Alu markers data. Plot of the contribution of each Alu marker to each principle component displayed in Figure 3-7. Haplotypes with positive correlation values push samples to the righ (or top) of the PC axis while samples with negative correlation values push samples to the left (or bottom)

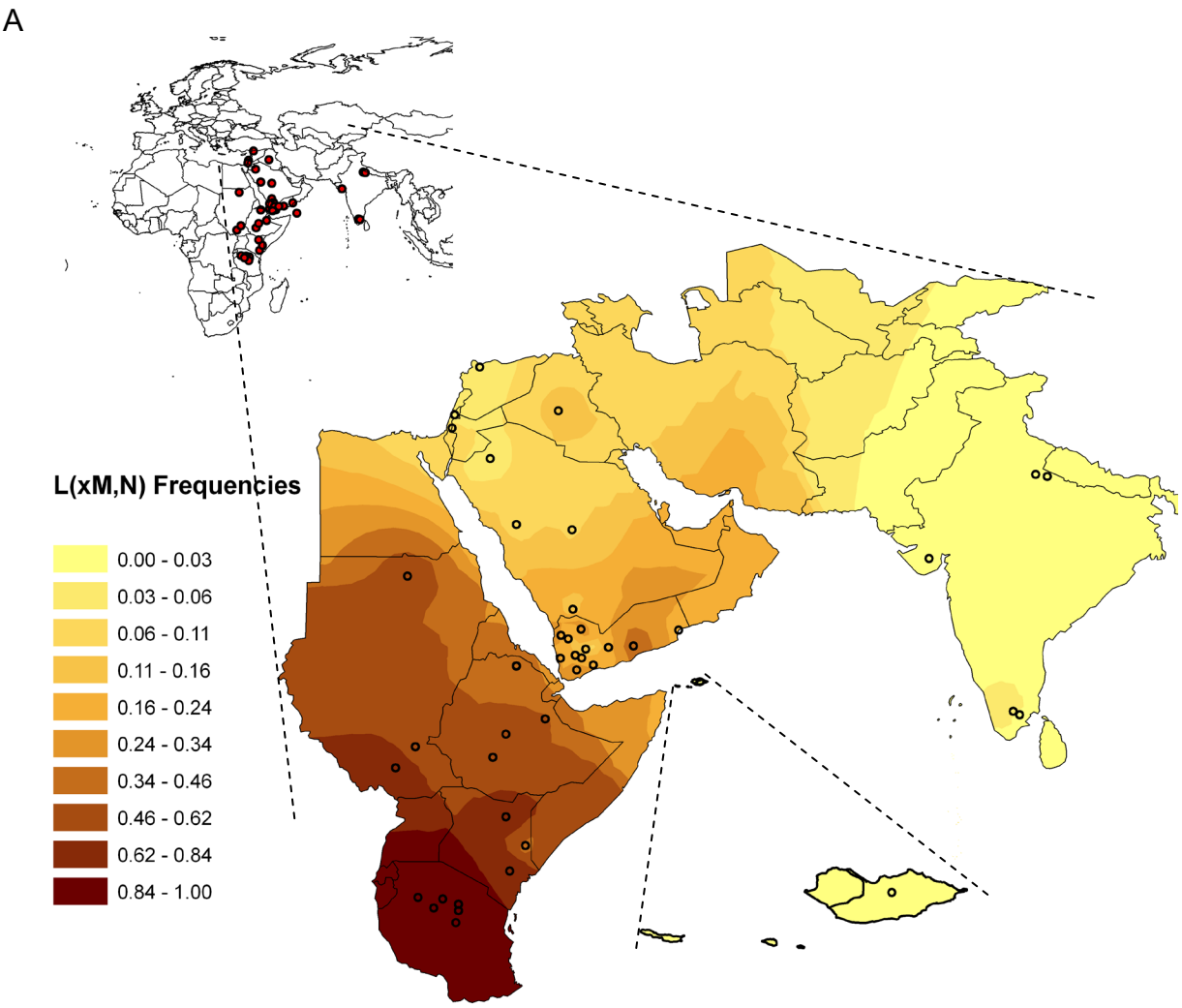
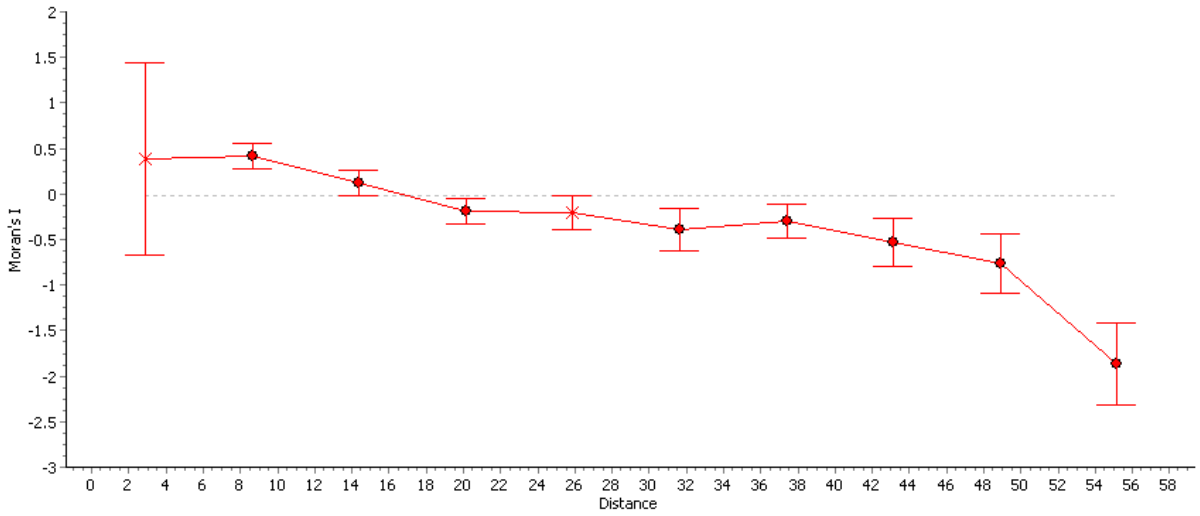


Figure 3-9. Distribution of haplotype L(xM,N). Panel A shows the IDW interpolated map of L(xM,N) haplotype frequencies throughout all sampled regions. Dashed lines indicate areas in the map that are magnified. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. A statistically significant cline is indicated by a pattern of decreasing spatial autocorrelation with increasing geographic distance (closed circles indicate significant Moran's I and asterisks indicate lack of significance. The Global Moran's I value for the total map is 0.7172, with a z-score of 10.49 and p-value <0.0001. Panel C shows the semivariogram of haplotypes frequency differences at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. This semivariogram shows the pattern expected under the assumption of isolation by distance, in which haplotypes frequencies increase with increasing geographic distances on the x-axis.

B



C.

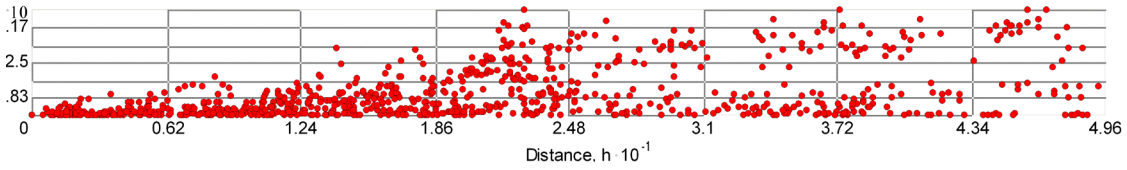


Figure 3-9. Continued

A

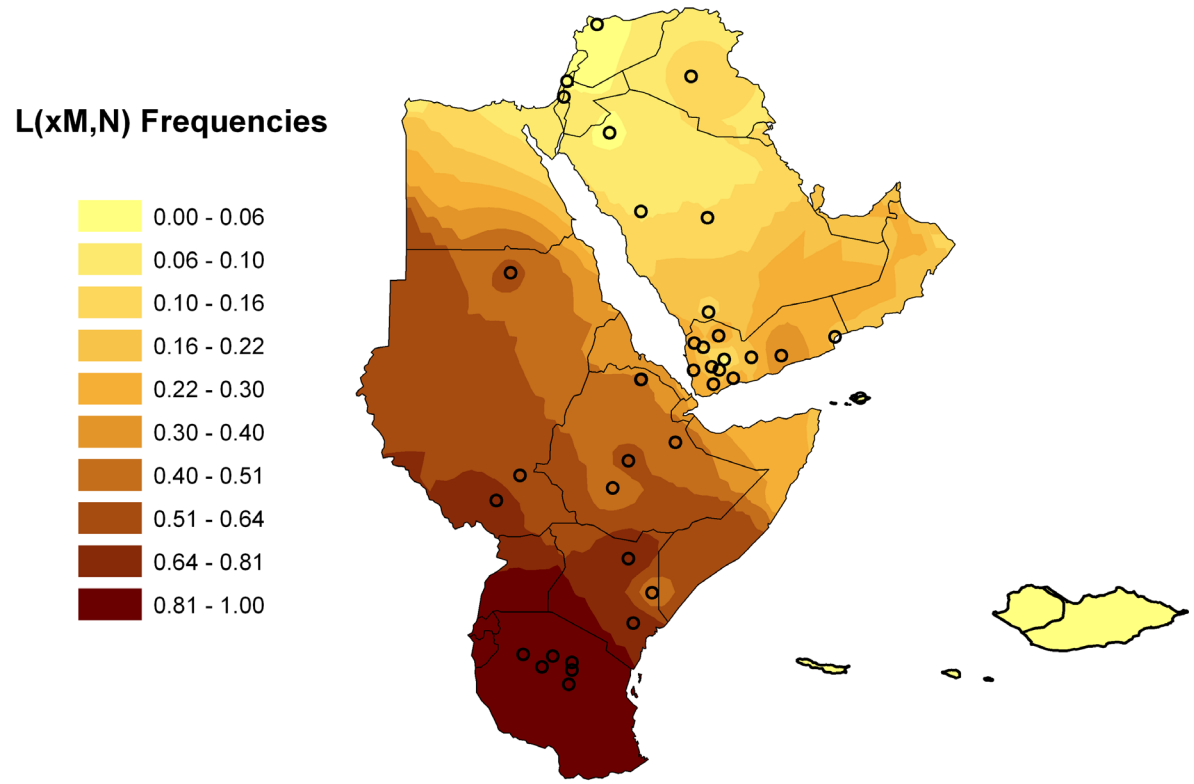
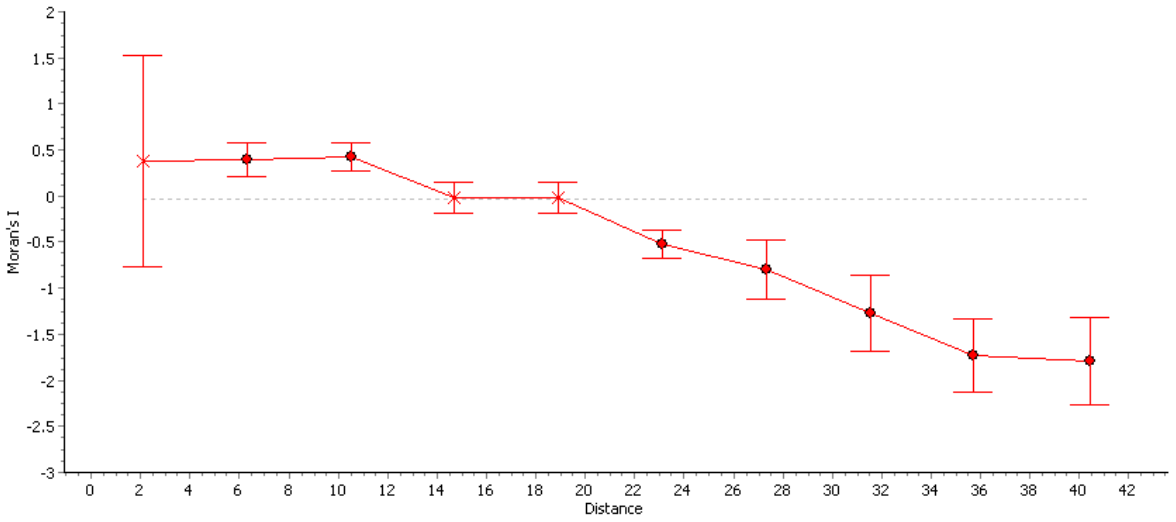


Figure 3-10. Distribution of haplotype L(xM,N) (without India). Panel A shows the IDW interpolated map of L(xM,N) haplotype frequencies in areas just surrounding the Red Sea (all regions used in Figure 7 except India). Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. A statistically significant cline is indicated by a pattern of decreasing spatial autocorrelation with increasing geographic distance (closed circles indicate significant Moran's I and asterisks indicate lack of significance. The Global Moran's I value for the total map is 0.7534, with a z-score of 10.97 and p-value<0.0001. Panel C shows the semivariogram of haplotypes frequency differences at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis.

B



C

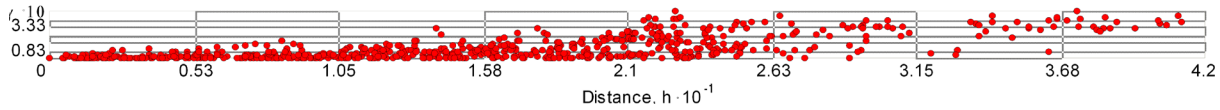


Figure 3-10. Continued.

A

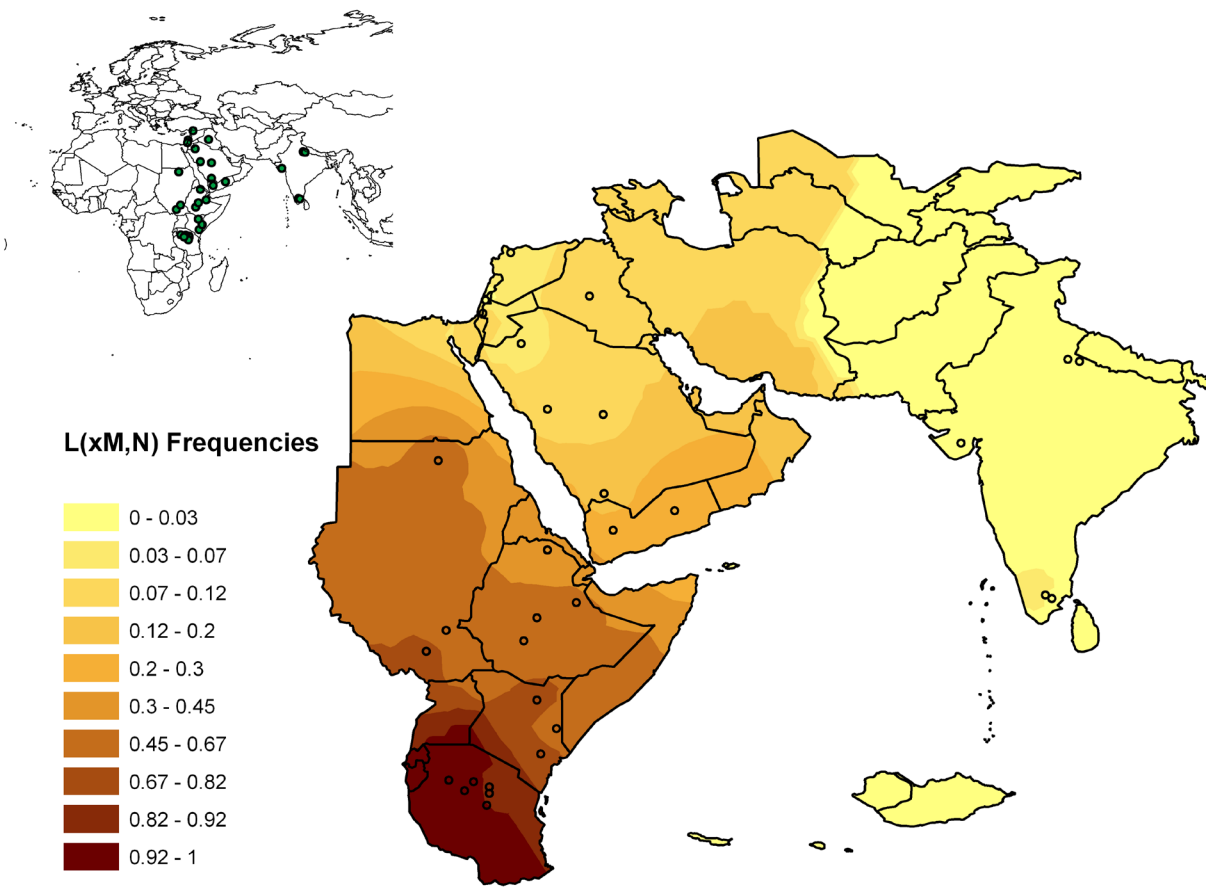
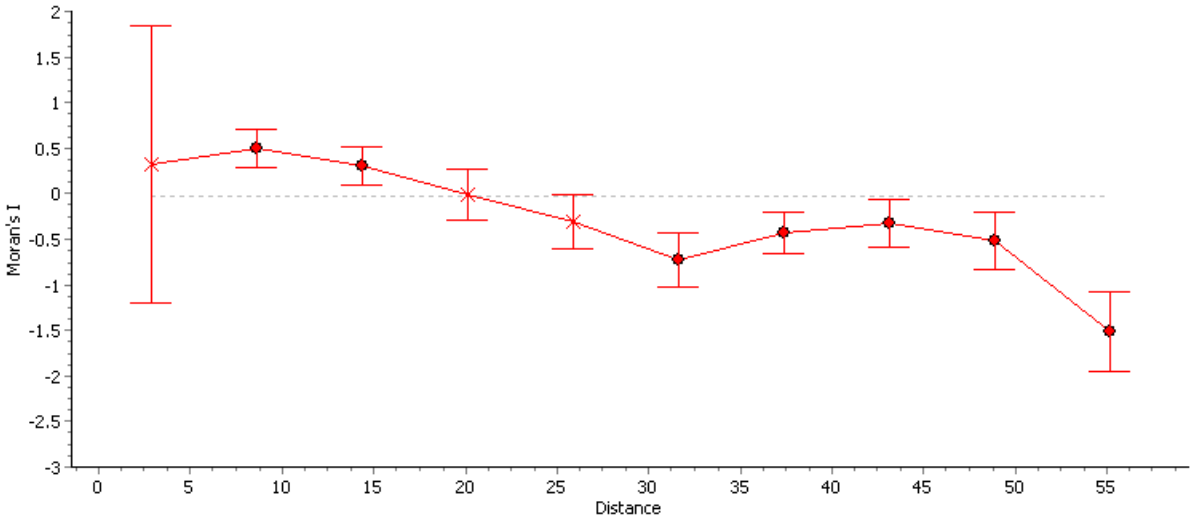


Figure 3-11. Distribution of haplotype L(xM,N) with Yemen pooled into East/West sample groups. Panel A shows the IDW interpolated map of haplotype L(xM,N) (including all derivatives) frequencies throughout the Greater Red Sea Region, with all Yemeni samples pooled into one eastern and one western Yemeni datapoint. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. A unique pattern is seen, similar to that in R0a Map, in which there is evidence for a cline at low distance classes, but after distances of ~30, Moran's I values increase (i.e. increasing geographic clustering). The Global Moran's I value for the total map is 1.203, with a z score of 9.99 and p-value of <0.0001. Panel C shows the semivariogram of haplotypes frequency differences at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. Haplotypes frequency differences appear to increase with geographic distances, as expected under isolation by distance.

B



C

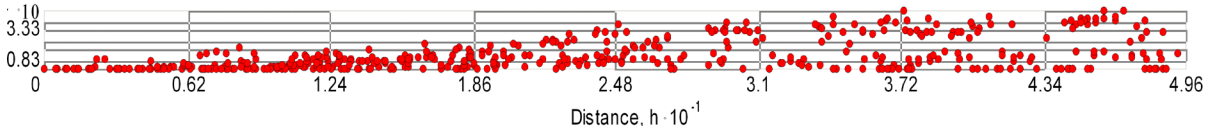


Figure 3-11. Continued

A

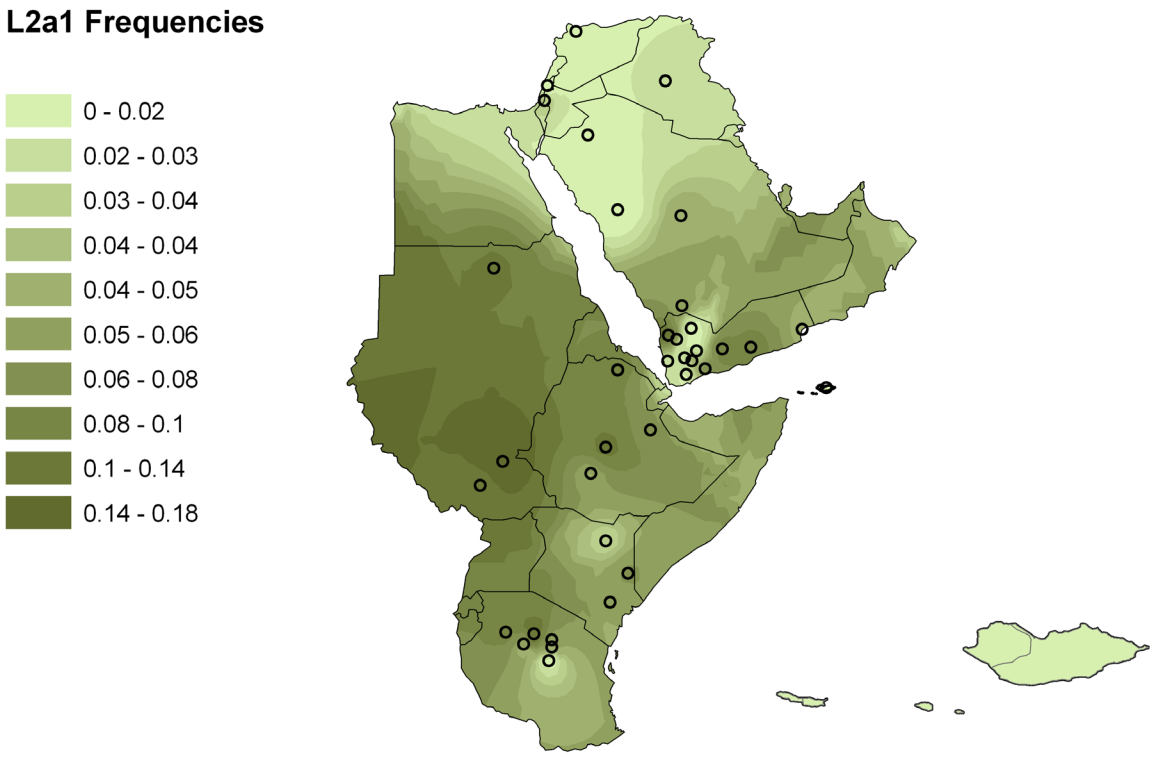
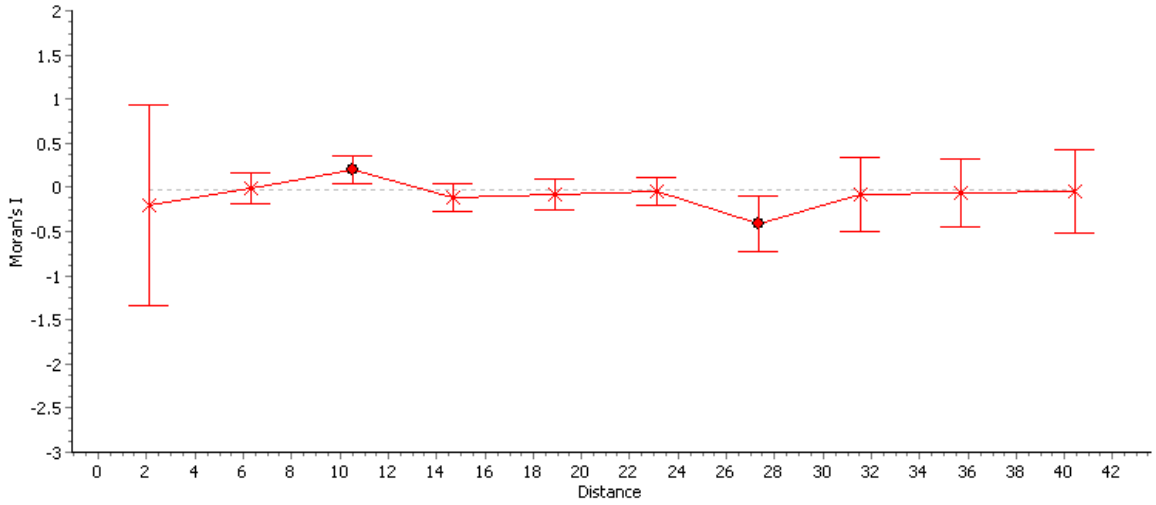


Figure 3-12. Distribution of haplotype L2a1. Panel A shows the IDW interpolated map of haplotype L2a1 (including all derivatives, e.g. L2a1a1, etc) frequencies in regions around the Red Sea. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. No statistically significant cline is evident. The Global Moran's I value for the total map is 0.779, with a z score of 11.36 and p-value of <0.0001. When India is included in the analysis, the Moran's I index is no longer significant (I=0.03, p=0.479). Panel C shows the semivariogram of haplotypes frequency differences at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The relationship between haplotypes frequency differences and geographic distances appears random.

B



C

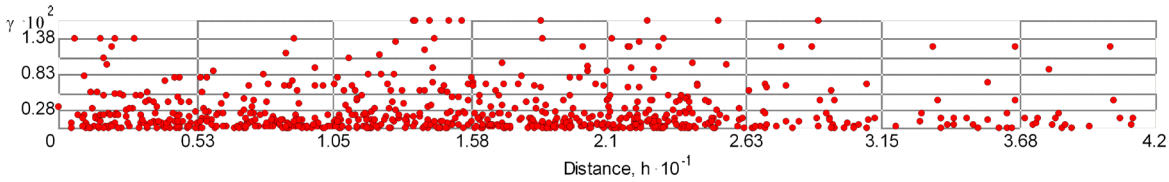


Figure 3-12. Continued.

A

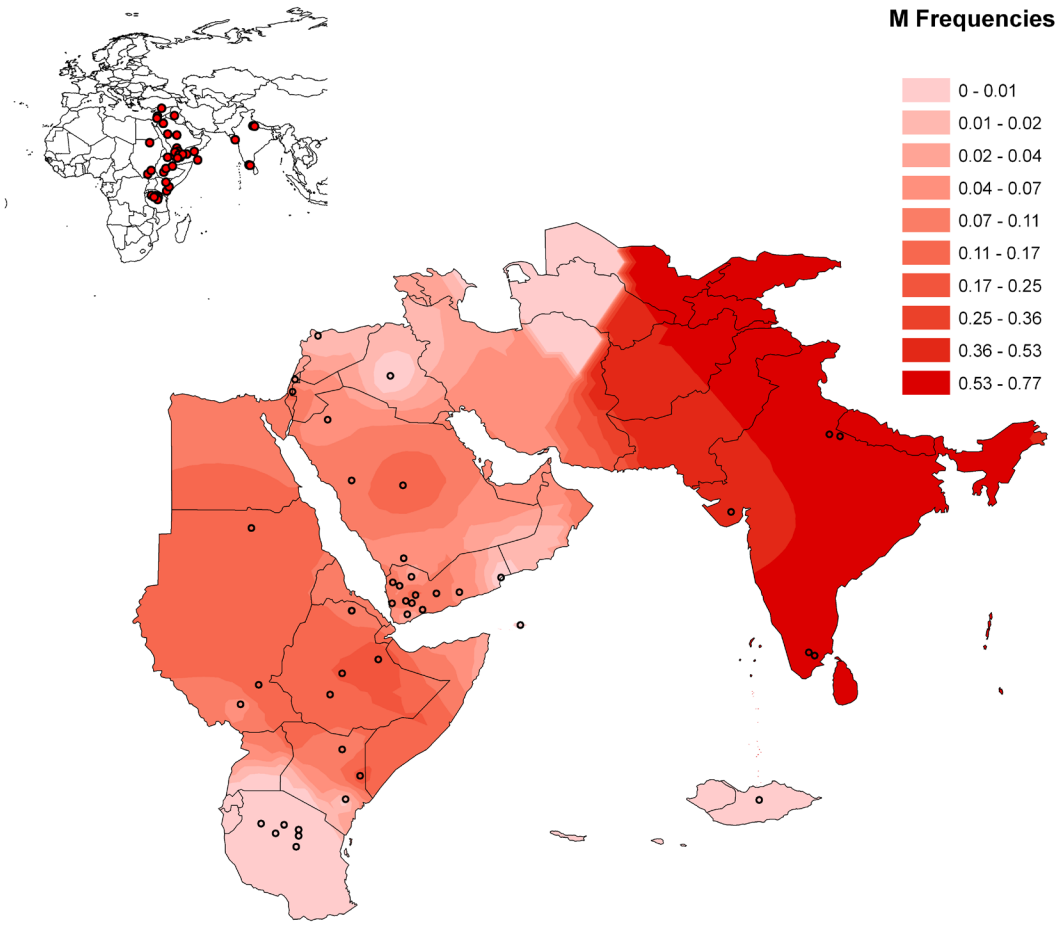
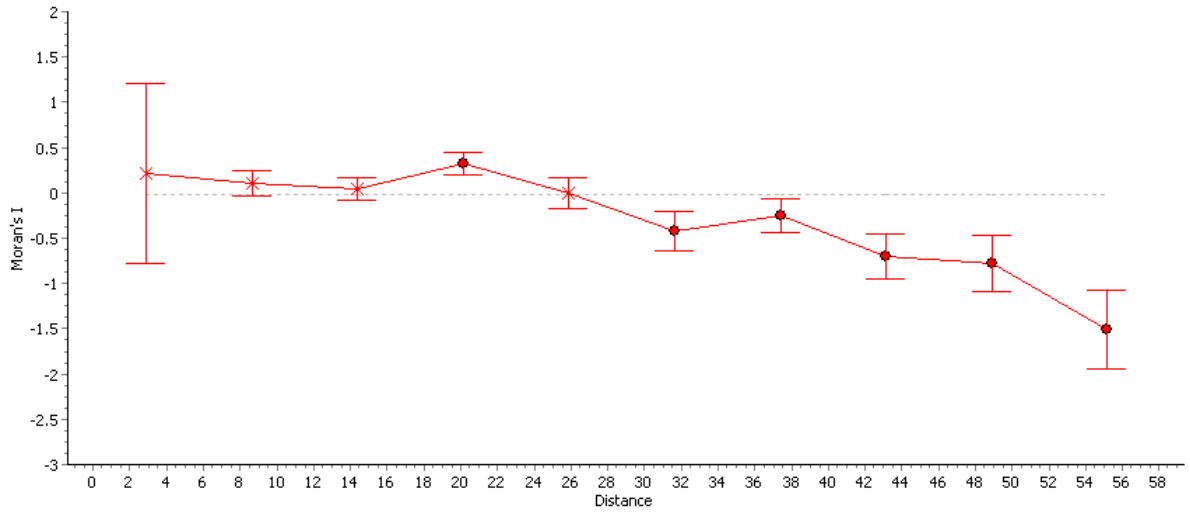


Figure 3-13. Distribution of haplotype M. Panel A shows the IDW interpolated map of M (including all derivatives) haplotype frequencies throughout the sampled regions. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. The Global Moran's I value for the total map is 0.33, with a z score of 5.32 and p-value of <0.0001. Panel C shows the semivariogram of haplotypes frequency differences at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The expected pattern of increasing frequency differences with increasing geographic distances is clearly evident past a certain geographic distance threshold.

B



C

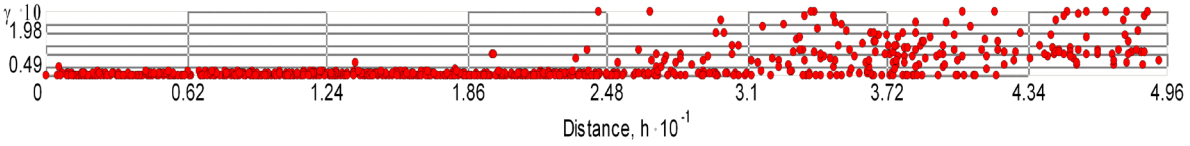


Figure 3-13. Continued.

A

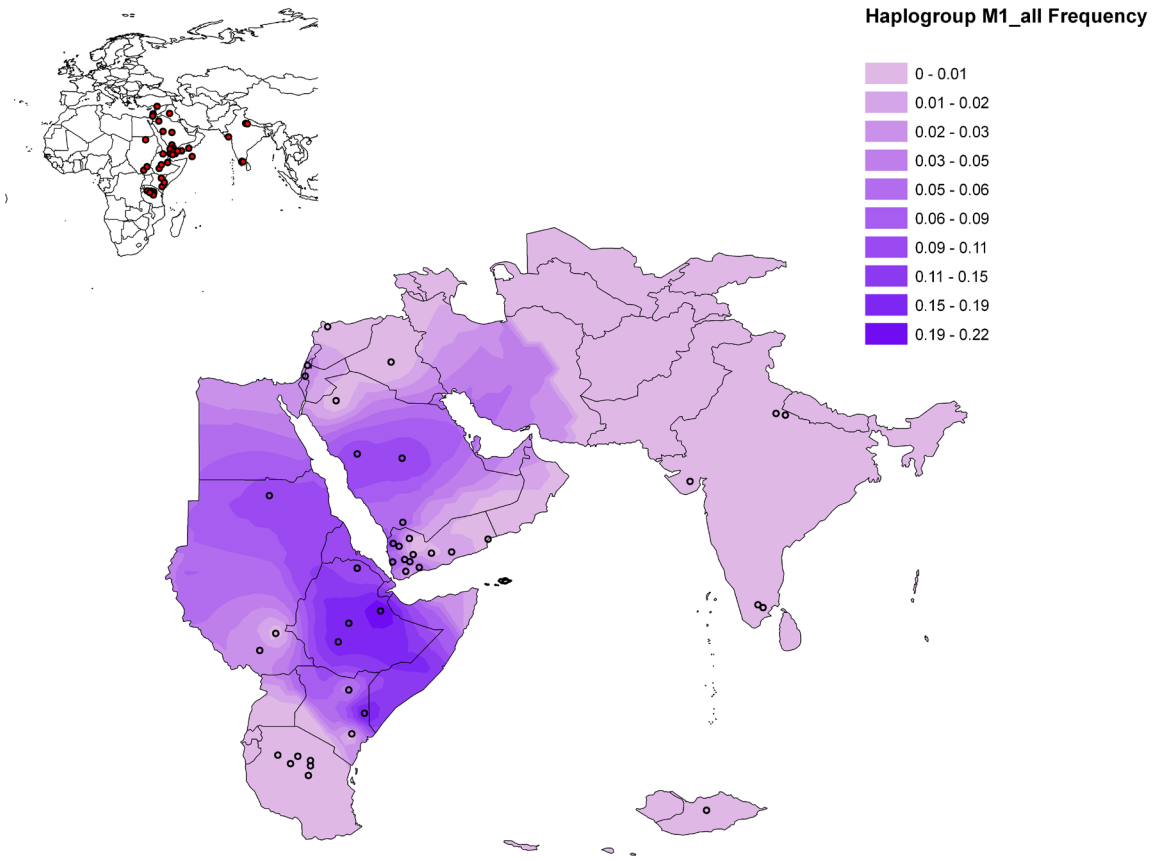
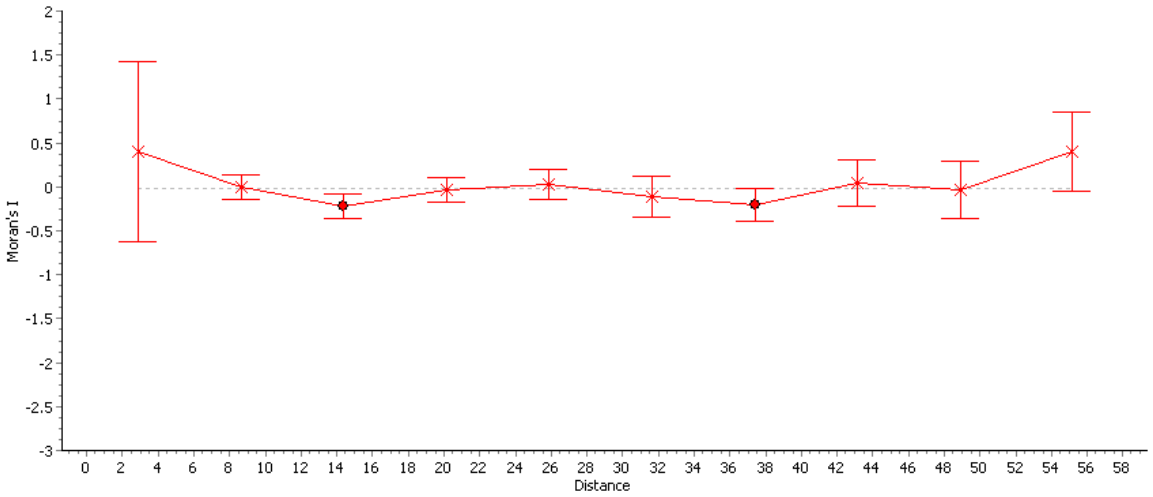


Figure 3-14. Distribution of haplotype M1. Panel A shows the IDW interpolated map of M1 (including all derivatives) haplotype frequencies throughout the sampled regions. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. The Global Moran's I value for the total map is 0.16, with a z score of 2.66 and p-value of 0.0078. Panel C shows the semivariogram of haplotypes frequency differences at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The expected pattern of increasing frequency differences with increasing geographic distances is not evident.

B



C

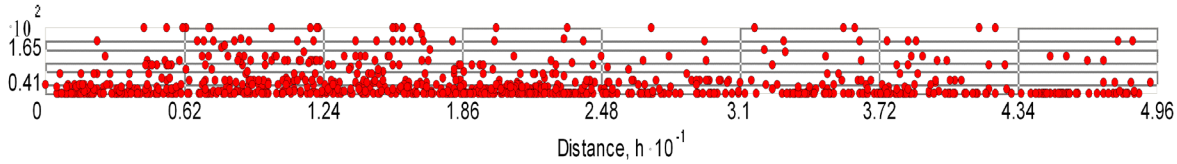


Figure 3-14. Continued

A

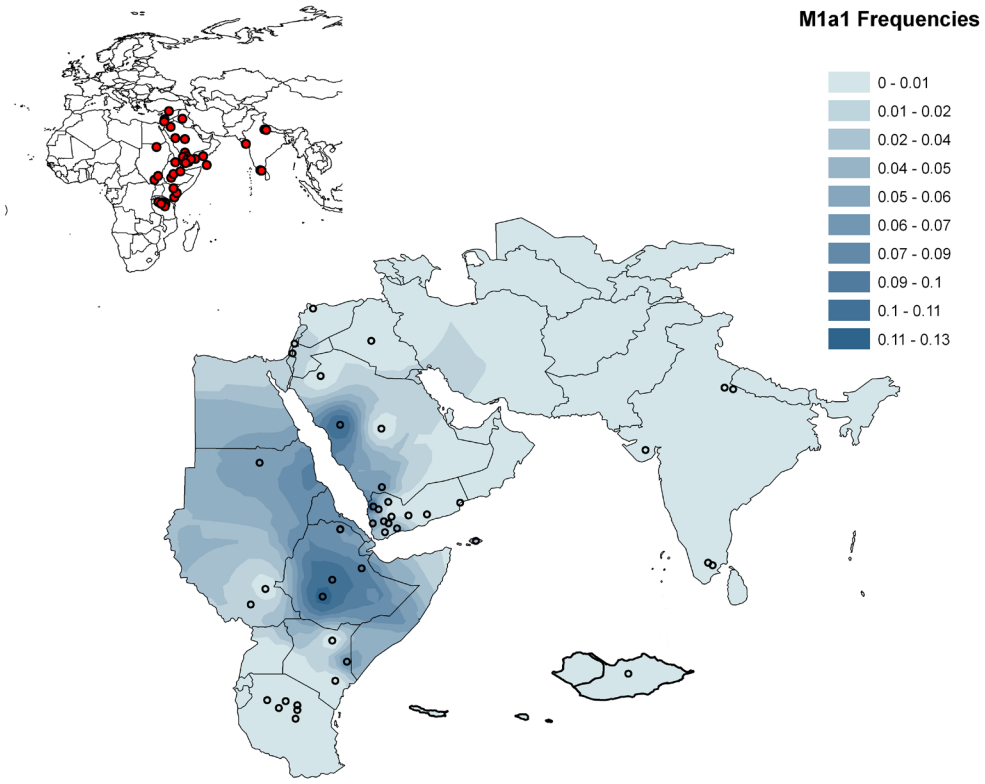
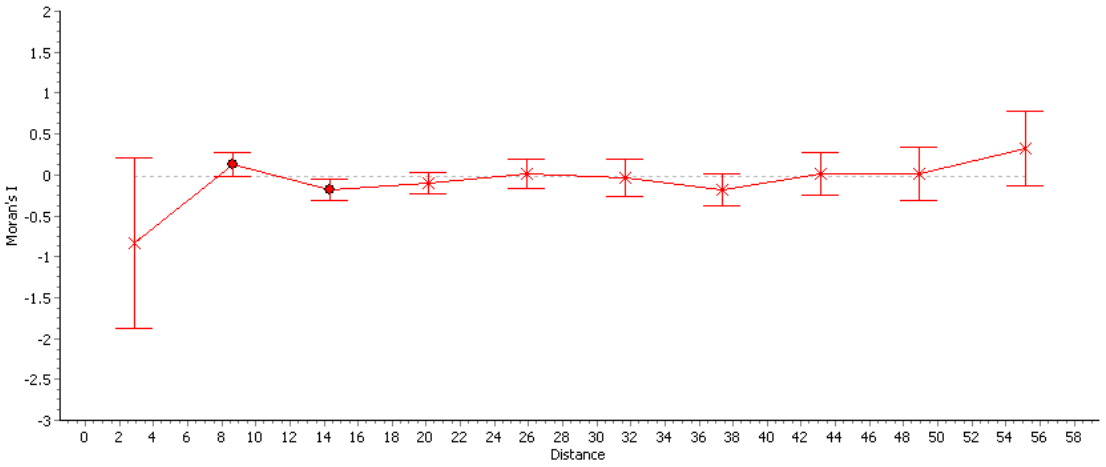


Figure 3-15. Distribution of haplotype M1a1. Panel A shows the IDW interpolated map of M1a1 (including all derivatives) haplotype frequencies throughout the sampled regions. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. The Global Moran's I value for the total map is 0.08, with a z score of 1.50 and p-value of 0.1339. Panel C shows the semivariogram of haplotypes frequency differences at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The expected pattern of increasing frequency differences with increasing geographic distances is not evident.

B



C

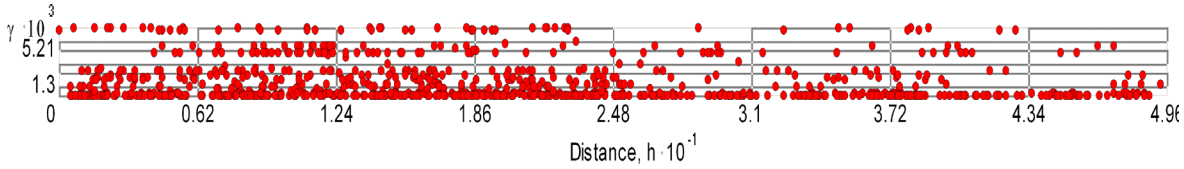


Figure 3-15. Continued

A

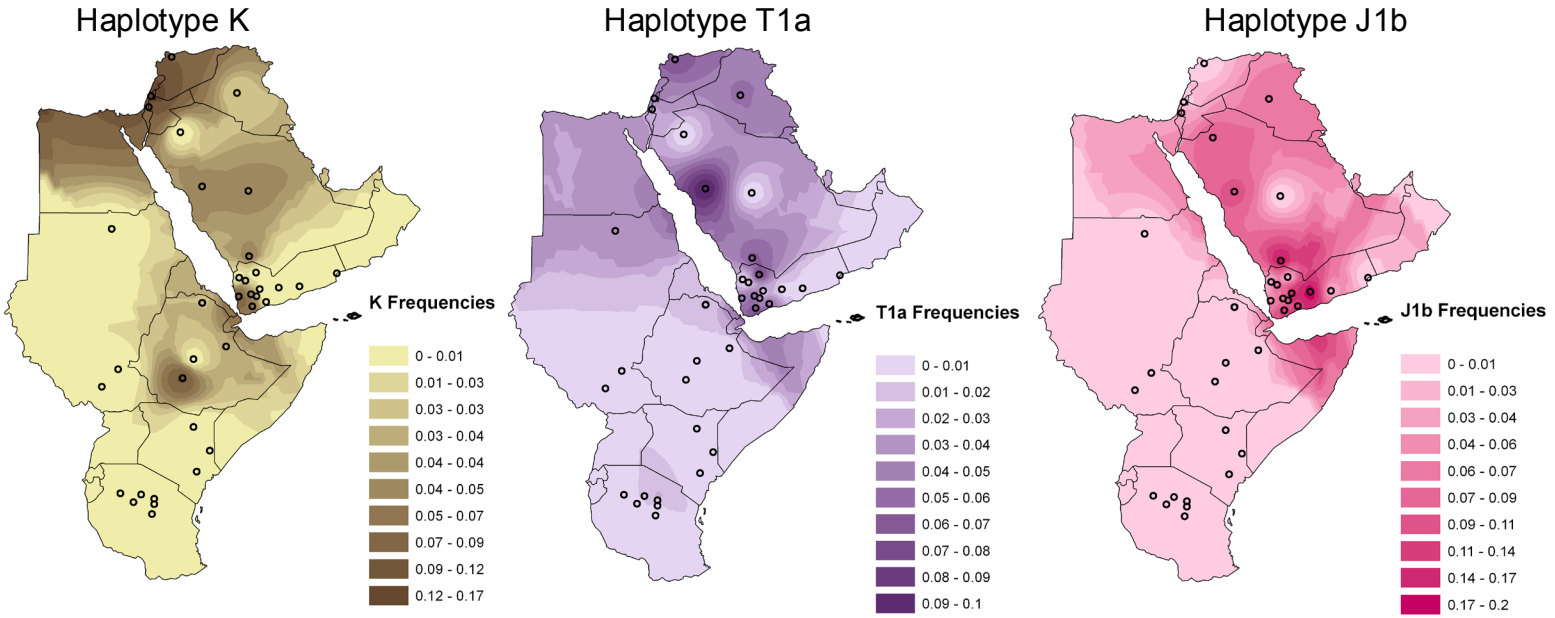
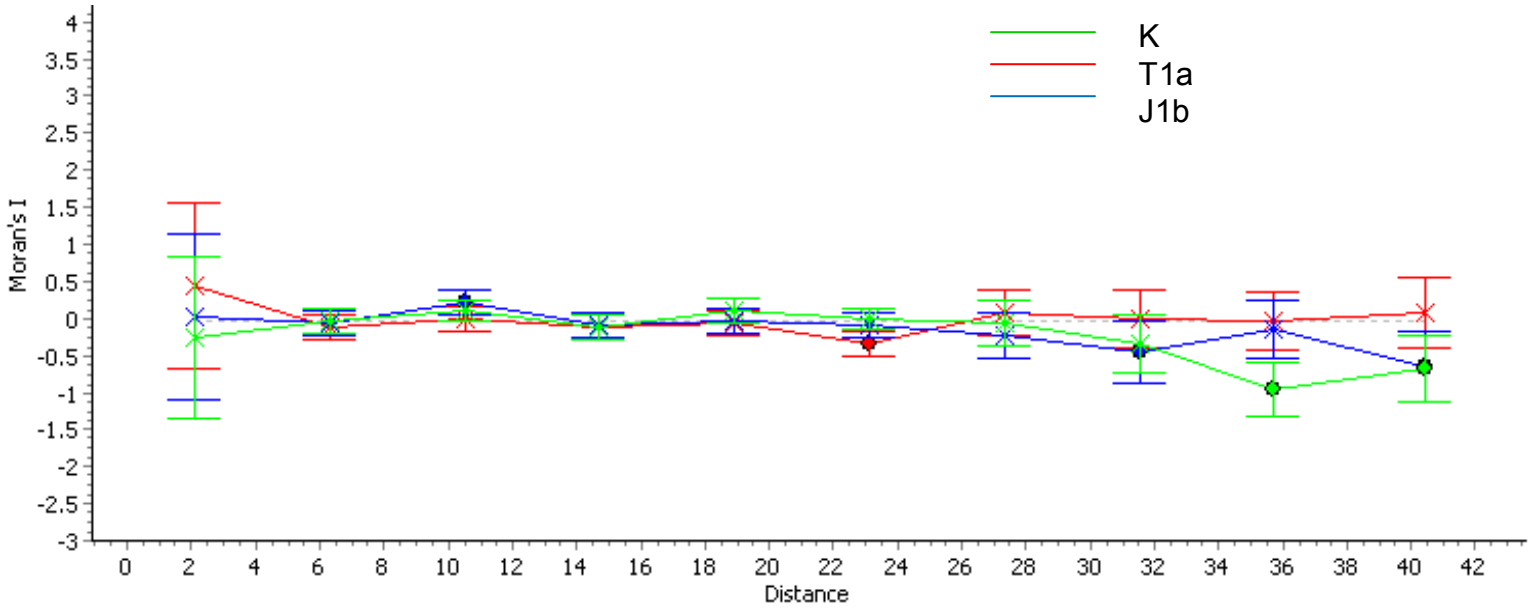


Figure 3-16. Distribution of haplotypes K, T1a, and J1b. Panel A shows the IDW interpolated maps of haplotypes K (including all derivatives), T1a, and J1b frequencies in regions around the Red Sea. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis. No statistically significant clines are evident in the autocorrelations. The maps of K and J1b show significant clustering, with global Moran's I values of 0.155, (z score= 2.65, p-value=0.008) for map K; I=0.366, (z=5.64, p=0.000) for map J1b. Map T1a has a Moran's I value that is not quite significant, (I=0.093, z=1.70, p=0.0883). When India is included in the analysis, Moran's I values increase for all three haplotypes, and for T1a, becomes significant (I=0.12, z=2.13 p=0.0329). Panel C shows the semivariograms of haplotypes frequency differences at each geographic distance class for each map. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The relationship between haplotypes frequency differences and geographic distances appears random for Haplotype K semivariogram, but the haplotypes frequency differences actually appear to decrease with increasing geographic distances for haplotypes T1a and J1b.

B



C

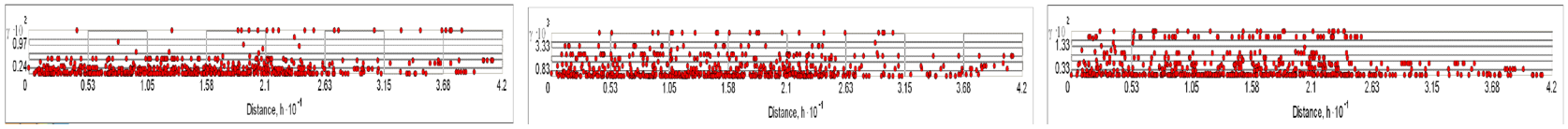


Figure 3-16. Continued

A

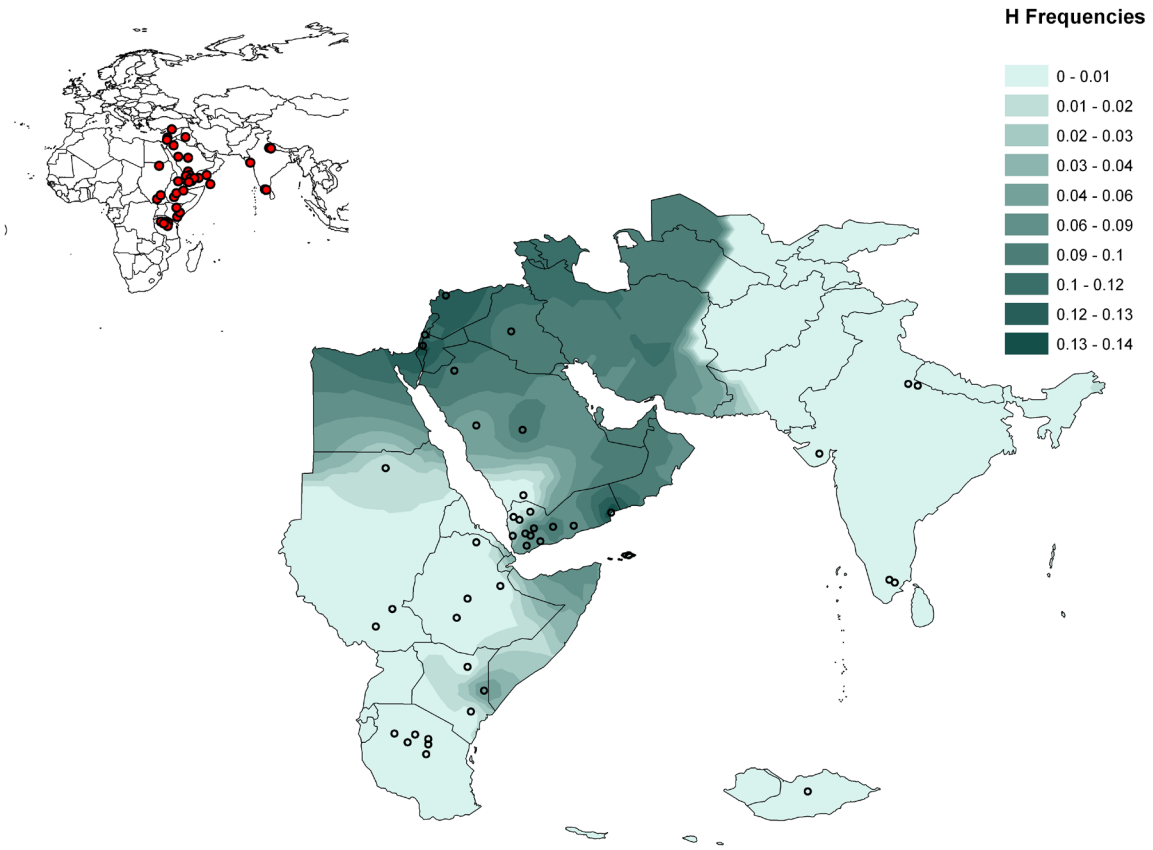
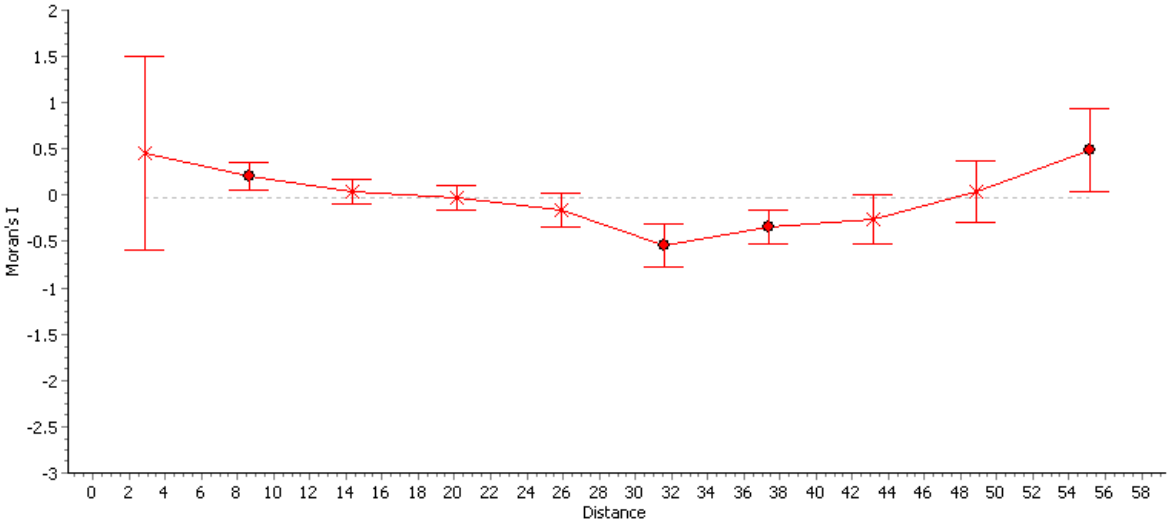


Figure 3-17. Distribution of haplotype H. Panel A shows the IDW interpolated map of haplotype H (including all derivatives) frequencies throughout the sampled regions. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis for H. The Global Moran's I value is 0.287, $p<0.0001$. Panel C shows the semivariogram of the H haplotype at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The relationship between haplotypes frequency differences and geographic distances appears random.

B



C

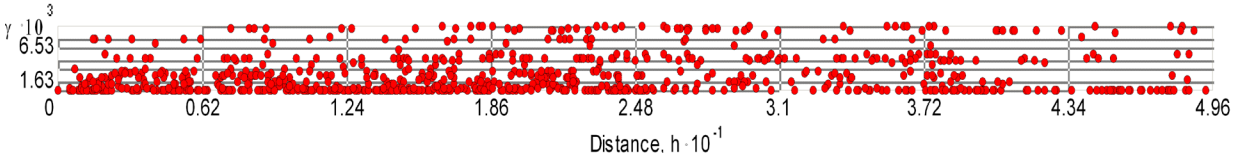


Figure 3-17. Continued

A

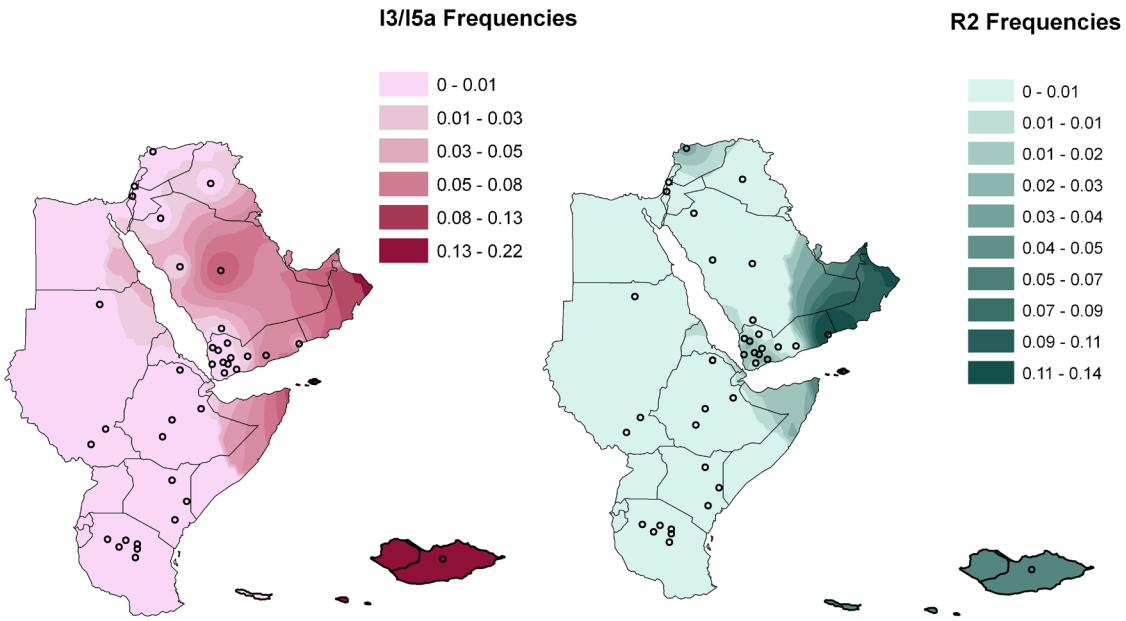
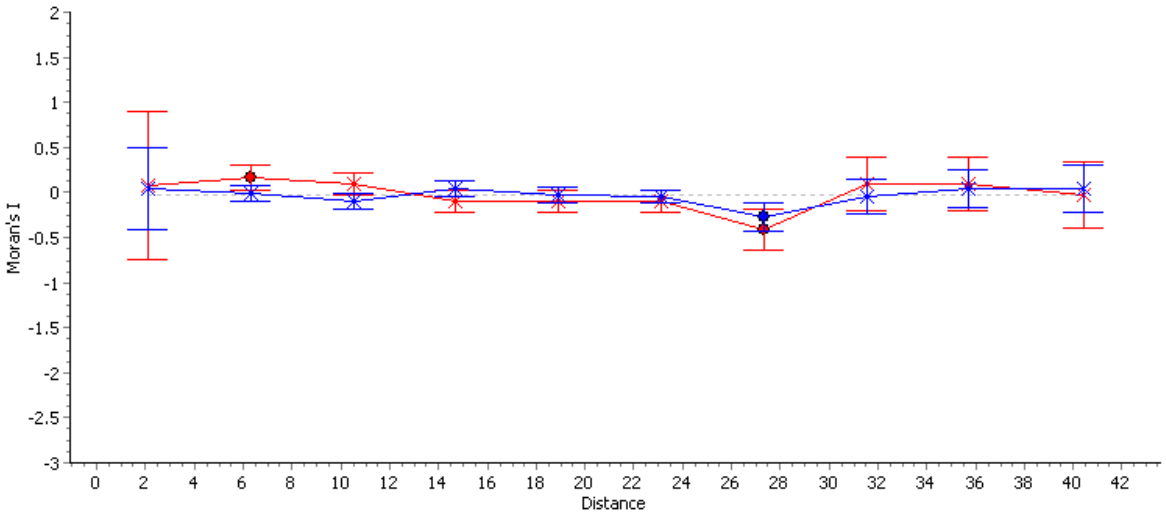


Figure 3-18. Distribution of haplotypes I3/I5a and R2. Panel A shows the IDW interpolated map of I3/I5a and R2 haplotype frequencies in regions around the Red Sea. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis for I3/I5a, and R2, respectively. No significant cline is evident for either haplotype. The Global Moran's I value for the total I3/I5a map is 0.024 (z=1.22, p=0.223), and for haplotypes R2 is 0.058 (z=1.52, p=0.1273). Panel C shows the semivariogram of the I3/I5a and R2 haplotypes at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The expected pattern (under isolation by distance) of increasing frequency differences with increasing geographic distances is not evident; instead there is a pattern of some high frequency differences and some low frequency differences across all geographic distances.

B



C

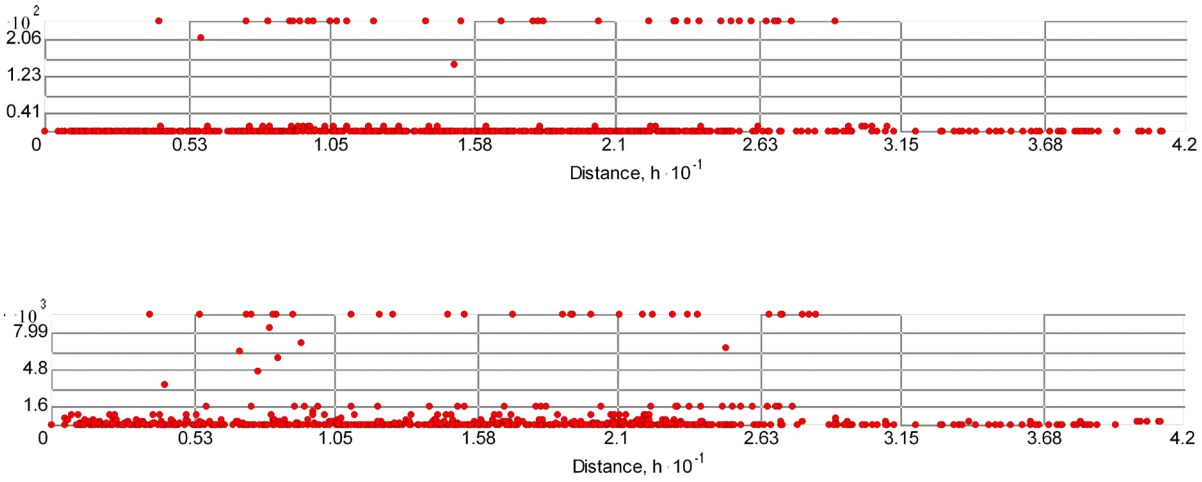


Figure 3-18. Continued

A

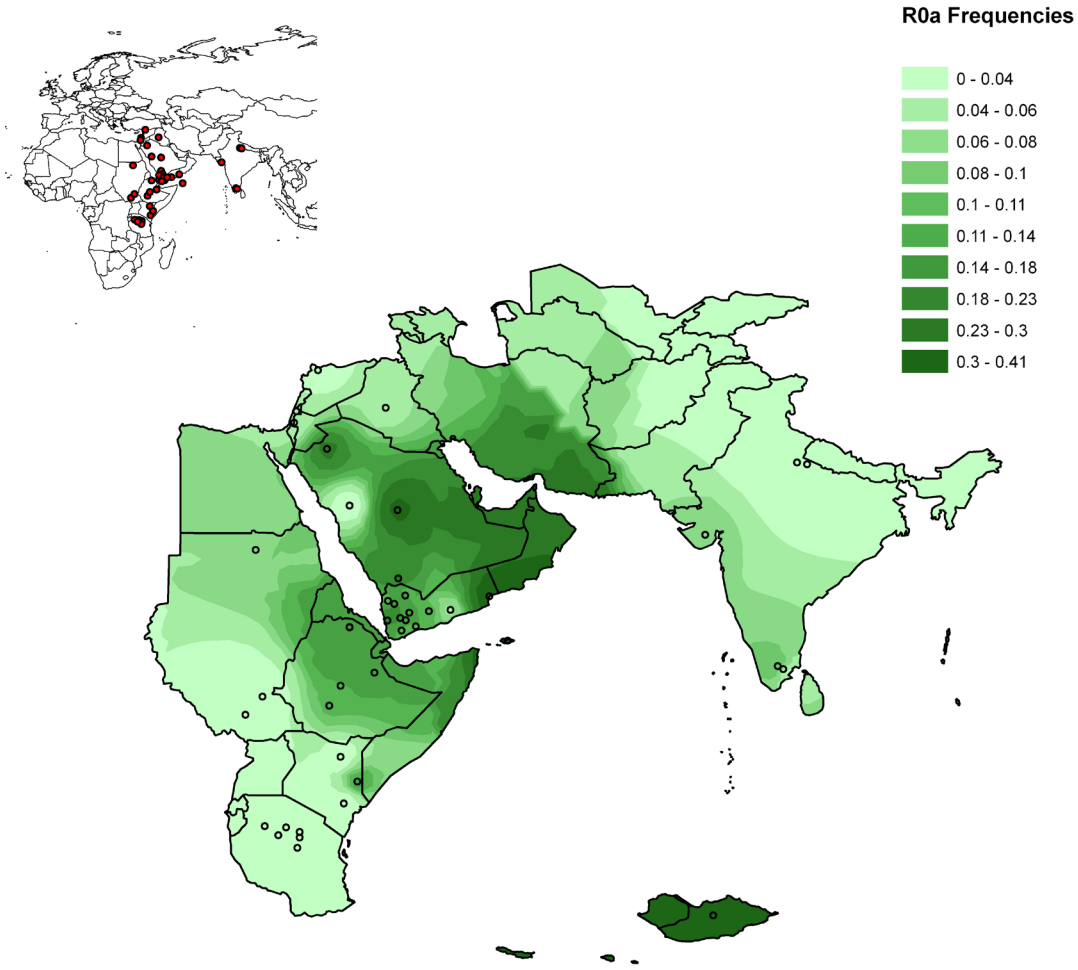
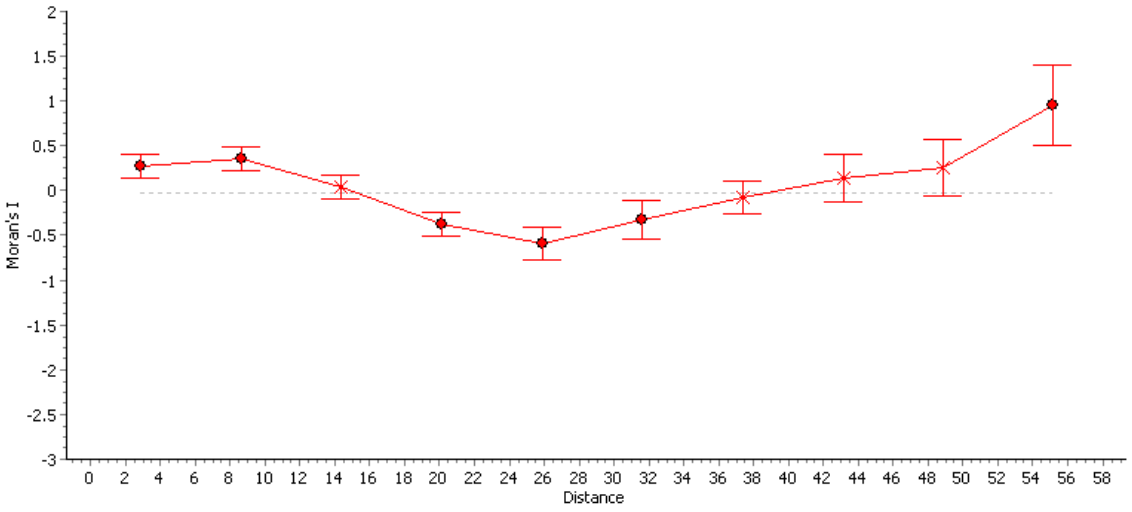


Figure 3-19. Distribution of haplotype R0a. Panel A shows the IDW interpolated map of R0a haplotype frequencies throughout the Greater Red Sea Region. Panel B shows the spatial autocorrelation (Moran's I) on the Y axis against plotted against geographic distance on the X axis for R0a. A unique pattern is seen, in which there is evidence for a cline at low distance classes, but after distances of ~26, Moran's I values increase (i.e. increasing geographic clustering). The Global Moran's I value for R0a is highly significant (I=0.307, z=4.82, p<0.0001). Panel C shows the semivariogram of the R0a haplotypes at each geographic distance class. Each point in the semivariogram cloud represents the difference squared between the haplotype frequencies for each pair of locations on the y-axis relative to the geographic distance between the points on the x-axis. The expected pattern (under isolation by distance) of increasing frequency differences with increasing geographic distances is evident only in the smaller distance classes; at higher distance classes there are low frequency differences.

B



C

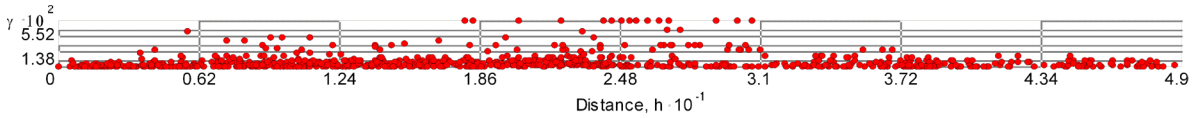


Figure 3-19. Continued

Table 3-3. Haplotype frequencies in African, Middle Eastern, and Asian populations for most common haplotypes in Yemen

Country Region or Ethnicity	Lat.	Long.	L(xM,N)	L2a1	J1b	H	R2	T1a	K	I3/I5a	R0a	M	M1	M1a1
Ethiopia														
Gurage	6.80	38.2	0.450	0.050	0.000	0.000	0.000	0.000	0.100	0.000	0.100	0.158	0.1579	0.1063
Amhara	8.54	39.2	0.540	0.115	0.000	0.000	0.000	0.000	0.000	0.000	0.115	0.167	0.1696	0.0982
Afar	13.70	40	0.438	0.125	0.000	0.000	0.000	0.000	0.063	0.000	0.125	0.063	0.0625	0
Oromo	9.70	42.2	0.467	0.067	0.000	0.000	0.000	0.000	0.033	0.000	0.133	0.222	0.2222	0.0741
Tigrais	13.71	40.01	0.364	0.046	0.000	0.000	0.000	0.023	0.000	0.000	0.136	0.146	0.1463	0.122
Kenya														
Kikuyu	-1.80	39.5	0.790	0.053	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0	0
Somali	0.13	40.7	0.400	0.067	0.000	0.067	0.000	0.000	0.000	0.000	0.133	0.200	0.2	0.0667
Turkana	2.30	39.2	0.794	0.029	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.0333	0
Sudan														
Dinka	6.00	30.8	0.721	0.116	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.025	0.025
Nubia	20.50	31.7	0.520	0.107	0.000	0.013	0.000	0.027	0.000	0.000	0.080	0.160	0.0933	0.0533
Nuer	7.60	32.3	0.546	0.182	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.182	0	0
Tanzania														
Burunge	-4.8	35.6	0.9189	0.0541	0	0	0	0	0	0	0	0	0	0
Datog	-4.3	35.6	0.8919	0.0541	0	0	0	0.027	0	0	0	0.027	0	0

Table 3-3. Continued

Hadza	-3.9	34.4	0.9342	0.1579	0	0	0	0.0132	0	0	0	0	0	0
Sakuma	-3.8	32.5	0.9091	0.0909	0	0	0	0	0	0	0	0	0	0
Sandawe	-5.7	35.4	0.9241	0	0	0	0	0	0	0	0	0.013	0.0127	0.0127
Turu	-4.6	33.7	1.00	0.0323	0	0	0	0	0	0	0	0	0	0
Israel														
Druze	32.70	35.3	0.000	0.000	0.000	0.106	0.000	0.043	0.170	0.000	0.043	0.022	0.0222	0.0222
Palestinian	31.70	35.1	0.110	0.040	0.040	0.140	0.000	0.020	0.080	0.000	0.040	0.13	0.02	0.02
Saudi Arabia														
Northern	29.40	38	0.056	0.000	0.111	0.056	0.000	0.000	0.000	0.000	0.20	0.05	0	0
Western	24.40	40	0.105	0.000	0.105	0.053	0.000	0.105	0.105	0.000	0.000	0.1	0.1	0.1
Southern	18.00	44.3	0.063	0.063	0.188	0.000	0.000	0.063	0.063	0.000	0.188	0.0625	0.0625	0.0625
Central	24.00	44.24	0.100	0.050	0.000	0.100	0.000	0.000	0.050	.050	0.25	0.15	0.1	0
Syria														
Northern	36.30	37.2	0.046	0.000	0.000	0.136	0.023	0.068	0.091	0.000	0.000	0	0	0
Iraq														
unknown	33.00	43.2	0.119	0.024	0.071	0.095	0.000	0.048	0.024	0.000	0.048	0	0	0
Yemen														
Abyan	13.79	45.86	0.282	0.077	0.128	0.077	0.026	0.051	0.000	0.000	0.103	0.0769	0.0769	0.0769
Al Bayda	14.32	44.98	0.192	0.064	0.085	0.043	0.043	0.021	0.043	0.000	0.149	0.0233	0	0

Table 3-3. Continued

Al Hudayda	14.30	43.35	0.275	0.020	0.039	0.020	0.020	0.039	0.098	0.000	0.157	0.0784	0.0784	0.0784
Al Jawf	16.49	44.94	0.333	0.000	0.000	0.000	0.000	0.083	0.000	0.000	0.083	0	0	0
Amran	15.75	43.96	0.094	0.000	0.094	0.000	0.031	0.000	0.031	0.000	0.188	0.0323	0	0
Dhamar	14.51	44.49	0.184	0.026	0.035	0.061	0.009	0.070	0.053	0.000	0.148	0.1048	0.0667	0.0476
Hadramout	15.21	48.92	0.441	0.086	0.017	0.069	0.000	0.000	0.000	0.017	0.051	0.0526	0.0175	0
Hajjah	16.03	43.4	0.333	0.166	0.000	0.000	0.000	0.000	0.000	0.000	0.083	0.125	0.125	0.125
Mahra	16.42	52.35	0.215	0.046	0.000	0.139	0.139	0.000	0.015	0.000	0.277	0	0	0
Marib	14.98	45.28	0.000	0.000	0.188	0.125	0.000	0.000	0.000	0.000	0.125	0.1667	0	0
Shabwah	15.12	47.02	0.200	0.100	0.200	0.100	0.000	0.000	0.000	0.000	0.1	0.1	0	0
Socotra	12.53	53.96	0.056	0.019	0.000	0.019	0.056	0.000	0.000	0.222	0.407	0	0	0
Southwest	13.40	44.6	0.286	0.000	0.143	0.071	0.000	0.071	0.071	0.000	0.08	0	0	0
DB West	21.82	71.41	0	0	0	0.1	0	0	0	0	0.08	0.44	0	0
DB South	10.27	77.83	0.0484	0	0.0161	0.03226	0	0.0161	0	0	0.1452	0.4839	0	0
Shia	28.03	80.42	0	0	0.0077	0	0.0077	0	0	0	0	0.525	0	0
Sunni	28.19	79.54	0	0	0.01176	0.01765	0	0	0	0	0	0.6336	0	0
Mappla	10.012	78.31	0	0	0	0	0	0.0161	0	0	0	0.7705	0	0

**Yemen is represented by each governorate containing at least 10 samples; Lahij, Taizz, and Aden were pooled into a single SW site. ***India is comprised of Muslim communities from throughout India who show genetic affinity towards non-Muslim Indians (excludes Iranian Shias thought to derive from Iran). DB =Dawoodi Bohra. Note the total sample size used for calculating M frequencies was smaller than that used for calculating other haplotypes frequencies because more haplotypes could not be distinguished between M and other haplotypes, and were thus excluded from analyses of M lineages.

CHAPTER 4
SOCIAL CLASSIFICATION, SKIN COLOR, AND GENETIC ANCESTRY: A BIO-CULTURAL ANALYSIS OF HEALTH DISPARITIES[2]

Introduction

The use and meaning of race in health research is a topic of considerable and continuing debate across scientific disciplines over the past decade (e.g. Burchard et al. 2003; Haga and Venter 2003; Kaufman and Cooper 2001; Lee et al. 2008; Lee et al. 2001; Sankar et al. 2007; Stevens 2003 among many others). Biological anthropologists are especially positioned to contribute to this debate, as they specialize in the study of human diversity both in contemporary populations and throughout human evolutionary history. Most bioanthropologists have dismissed a biological basis of race, but at the same time continue to use racial categories in their research, often assuming an underlying genetic difference between races (e.g. forensic and genetic anthropologists often rely on morphological and genetic differences to assign individuals to racial groups). In biomedical studies, the same confusion exists, as some researchers continue to rely on race as a useful variable for detecting underlying genetic susceptibility to disease (Risch et al. 2002) while others argue that race is purely a cultural concept that does not capture the diversity of human genetic variation (2001; Schwartz 2001), or at least any variation relevant for development of complex disease (Cooper et al. 2003; Haga and Venter 2003). Very few studies integrate genetic and sociocultural data to investigate complex disease, despite the fact that complex disease, by definition, is a product of genetic and environmental factors. As a result, most studies

[2] This chapter is based largely on on a study by Gravlee, Non, and Mulligan (2009) recently published in PLoS One.
Gravlee CC, Non AL, and Mulligan CJ. 2009. Genetic ancestry, social classification, and racial inequalities in blood pressure, *Public Library of Science One* 4(9): e6821.

are unable to test competing hypotheses, control for potential confounders, or to test for potential interactions between genes and environment. I directly address this problem by integrating measures of genetic ancestry and cultural classification of race to test competing genetic and sociocultural explanations for health disparities in the complex disease of hypertension.

Hypertension is an ideal case study for understanding racial inequalities in health. First, hypertension is responsible for the largest portion of racial disparities in life expectancy in the U.S. (Wong et al. 2002). Hypertension is also over-represented in those of darker skin color throughout the African Diaspora, including Latin America and the Caribbean (Cooper et al. 1997; Halberstein 1999). Finally, hypertension is a model disease for testing competing hypotheses, because blood pressure (BP) is directly influenced by many potential environmental factors, including psychosocial stress, often associated with low social status (Adler and Snibbe 2003; Sweet et al. 2007), behavioral traits, such as diet and exercise (Kvaavik et al. 2009; Maruthur et al. 2009), as well as many potential genes, such as those implicated in the rennin-angiotensin-aldosterone system, the adrenergic system, salt homeostasis, and Type II diabetes mellitus (Deo et al. 2007).

In the current study, I integrate genetic and sociocultural factors together to investigate their relative contributions to the development of high blood pressure in Guayama, a population in Southeast Puerto Rico. The sociocultural factors in this study were developed by previous ethnographic research in this Puerto Rican population (Gravlee 2005; Gravlee et al. 2005). They include standard variables such as socioeconomic status (SES), as derived from a combination of income and education,

as well as a unique measure of culturally defined *color* (ko-lór), a form of social classification, analogous to "race" in the United States, which estimates how participants are perceived by other Puerto Ricans in everyday social interactions. Previous research has shown this ethnographically-derived measure has direct implications for health, as social classification can potentially shape exposure to stressors (Gravlee et al. 2005). For the genetic portion of the study, I generated an estimate of African ancestry based on 78 ancestry informative markers (AIMs; markers that show large frequency differences across ancestral populations). All individuals were also genotyped for six polymorphisms in three candidate genes for hypertension in the family of adrenergic receptors. These variants are located within three genes that encode for members of the GTP-protein-coupled receptor (GPCR) family – a large super-family of membrane receptors that are important for a wide variety of physiological and metabolic processes, with health implications ranging from hypertension to obesity, depression, and schizophrenia (Tang and Insel 2005). Specifically, the α_2AR plays an important role in control of neurotransmitter release in the central and peripheral sympathetic nervous system, especially in the heart and central neurons (Hein et al. 1999). The β-Arenergic receptors (β-ARs) control heart rate and contractility, smooth muscle relaxation, and multiple metabolic events (Kirstein and Insel 2004). The particular variants within the genes encoding these receptors were chosen, as they show varying frequencies across racial groups and are implicated in hypertension. The six loci include one insertion/deletion polymorphism within a gene that codes for the α_{2C}-Adrenergic Receptor (α_{2C}AR, 12bp deletion at amino acid positions 322-325), the two most common SNPs in a gene that encodes the β_1-

Adregeneric Receptor (β_1AR, Ser49Gly(145A→G) and Arg389Gly(1165G→C)), and two SNPs in a gene that encodes the β_2-Adrenergic Receptor (β_2AR, Gly16Arg andGln27Glu) that are the most common across racial groups of Caucasians, Africans, and Asians. This combination of data allows isolation of the genetic and sociocultural dimensions of race that are typically conflated in studies of health disparities. Additionally, the inclusion of candidate genes enables an investigation of their contribution to hypertension, and also provides an opportunity to investigate how the addition of sociocultural data to a standard genetic analysis might alter associations between candidate disease genes and a complex disease.

Puerto Rico is an ideal site to investigate questions about race due to the high degree of genetic admixture between Europeans, Africans, and Native Americans, leading to wide variability in skin color, hair types, and facial features, along with the use of dozens of cultural terms to classify *color* (Gravlee 2005). Racial classifications in Puerto Rico are considered to be more fluid than those in the U.S., such that certain social factors like income and occupation can influence or change one's classification throughout one's lifetime (Godreau 2000). This high degree of fluidity and admixture has generated a significant amount of genetic diversity across cultural groups, especially relative to cultural groups in the U.S., where the antiquated "one-drop rule" still classifies a person as "black" based on any degree of African ancestry (Peery and Bodenhausen 2008). Thus, the considerable variation across all dimensions of race in Puerto Rico provides a unique opportunity to isolate the contribution of each dimension to explore the causes behind the racial inequality in hypertension.

In the current study, I integrate sociocultural and genetic data on the same population to investigate the relative role of each type of data in contributing to blood pressure variation in Puerto Rico. I first investigate how social classification of *color* is associated with individual genetic ancestry. Next, I explore through multiple linear regressions whether genetic ancestry or *color* better predict BP, and test for any interactions between ancestry, social classification, and socioeconomic status. Finally, I test if the inclusion of sociocultural data alters the association between BP and candidate gene polymorphisms for hypertension? The results show that genetic ancestry measures vary widely across levels of cultural classification, indicating the need for independent measures of both components of race. Second, there is no evidence to support an association between African genetic ancestry and BP, but instead there is strong support for an interaction between SES and social classification that better explains BP variation. Finally, the analyses show that unmeasured cultural heterogeneity can mask a significant association between a genetic variant (α_{2C} adrenergic receptor deletion) and a complex phenotype (blood pressure variation). These results are significant as they provide the first direct comparison of genetic and sociocultural dimensions of race in contributing to complex disease, and because they show how greater attention to the measurement of sociocultural phenomena can empower the next generation of genetic association studies.

Materials and Methods

Samples

Participants were selected from Guayama Puerto Rico through a random sampling procedure in four residential areas. These areas were selected by ethnographic research to represent a broad range of people from varying

socioeconomic status and *color* categories, as described previously (Gravlee and Dressler 2005). One adult aged 25-55 was randomly selected from each of 25 randomly selected households within each of the four residential areas for a total of 100 individuals, 96 of which donated buccal swabs and gave informed consent for future genetic analyses.

Genotyping.

The 96 samples were genotyped for 100 biallelelic autosomal SNP markers at Prevention Genetics (Marshfield, Wisconsin). This panel of AIMs were selected for large frequency differences between West African, European, and Native American populations, derived from a GeneMapping 10K Affymetrix array screening (Shriver et al. 2005). After removing samples and markers with greater than 25% missing data, the dataset included 87 individuals typed for 78 AIMs. Of these, 84 individuals were successfully assayed via pyrosequencing at the University of Florida Center for Pharmacogenomics for six polymorphisms within three different genes of the adrenergic receptor family: Ser49Gly and Arg389Gly in β_1AR, Gly16Arg, Gln27Glu, and Arg523Arg in β_2AR, and Del322-325 in α_{2C}AR.

Individual Ancestry Estimation.

Three independent methods were used to estimate individual ancestry: two Bayesian approaches utilizing an MCMC algorithm implemented in *Structure* 2.2 (Falush et al. 2003) and ADMIXMAP (McKeigue et al. 2000) and a maximum likelihood estimation approach (MLE), kindly provided by Xianyun Mao. Genotypes and allele frequencies from unadmixed populations of West Africans, Europeans, and Native Americans (needed for *Structure* and ADMIXMAP/MLE, respectively) were kindly provided by Mark Shriver. A value of K=3 ancestral populations was selected for all

three programs because of the historical documentation of ancestral contribution from three distinct populations in Puerto Rico (Bonilla et al. 2004). The three programs did not produce significantly different ancestry estimates and yielded similar results in all analyses; in this dissertation, I present maximum likelihood estimates.

Sociocultural Measures.

The sociocultural measure of ascribed *color* used in this analysis was derived and assigned to each survey respondent, as described in a previous study (Gravlee et al. 2005). In brief, ascribed *color* assignments were based on ethnographic data from the same community where the survey was conducted (Gravlee et al. 2005). Ethnographic informants were asked to identify the *color* of 72 standardized facial portraits that varied systematically in skin tone, hair texture, and facial features. Cultural consensus analysis was then used to determine the culturally appropriate categorization of each portrait (Romney 1986). In the survey, two observers independently selected the portrait that best resembled each respondent (for detailed discussion see Gravlee et al. 2005). In the current paper, the cultural consensus estimate of *color* is used for the matched portrait as an estimate of the culturally appropriate categorization of respondents' *color*. This measure approximates how the respondents' *color* would be perceived by others in everyday social interaction.

Socioeconomic status (SES) was also measured in a previous study, as a combination of self-reported education (years) and household income (total from all sources, before taxes, during the last 12 months) (Gravlee et al. 2005). Multiple ways of modeling SES were tested, none of which significantly altered the main results; in this dissertation, as in previous studies, I use scores on the first principal component of education and household income (88% common variance explained) to represent SES.

Physiological Measures

Physiological data analyzed in this study included blood pressure, height, and weight, as collected in Gravlee et al. (2005). Blood pressure was measured with an automatic oscillometric blood pressure monitor (Omron HEM-737AC; Omron Healthcare, Inc., Vernon Hills, IL) that has been validated for population-based studies. Three measurements were taken at standardized intervals at the beginning, middle, and end of the hour-long interview. Respondents had been seated for at least 10 minutes and had not ingested caffeine or tobacco for at least 30 minutes before each measurement. Measurements were taken with the left arm supported at heart level. Mean systolic (SBP) and diastolic blood pressure (DBP) from the three measurements are treated as dependent variables. The other physiological measures were collected using a digital scale and a portable stadiometer for weight and height, respectively. These measures were used to calculate body mass index (BMI, weight in kg/height in m^2). Additional standard covariates included sex (0=female, 1=male), age (years), and current use of antihypertensive medication (0=no, 1=yes).

Multiple Linear Regression Analyses.

In this study, regression analyses were performed separately for SBP and DBP, using age, sex, antihypertensive medication, BMI, SES, genetic ancestry, and ascribed *color* as predictors. Ascribed *color* was entered as two categorical variables using a reverse Helmert coding scheme. The first variable tested for differences between *trigueño* and *blanco*; the second for differences between *negro* and the average of *trigueño* and *blanco*. This scheme was based on ethnographic research suggesting the largest differences would be found between those classified as *negro* from the mean of those classified as *trigueño/blanco*. I tested for an interaction between SES and

ascribed *color* using cross-product interaction terms. Similar results were found with log-transformed SBP and DBP; here I show results using non-log-transformed data for ease of interpretation. Age, BMI, and African ancestry estimates were mean-centered to reduce multicollinearity and aid in interpretation of interaction terms. In models with the α_{2C}AR gene included, only 84 individuals who genotyped successfully for this variant were included.

Results

Through this study, I investigated the association of diverse genetic and sociocultural data with blood pressure phenotypes in southeastern Puerto Rico (n=87 individuals). Measures of cultural classification placed respondents into three primary categories of *color: negro (*black), *trigueño* (literally wheat-colored), and *blanco* (white). Statistically significant differences in genetic ancestry and education, but no other variables, were found between these three categories (Table 4-1). Although mean levels of ancestry differed across cultural categories, the overlap between categories was the key finding, (e.g. at least two *color* categories overlap across 88% of the observed range of estimated ancestry (Figure 4-1). This result confirms the hypothesis that genetic ancestry and social classification represent distinct concepts, and neither can be used to predict the other.

I next performed multiple linear regressions to test if ancestry or social classification better predicts blood pressure variation using both SBP and DBP as response variables, respectively (Tables 4-2 and 4-3). Standard covariates included in each analysis were age, gender, body mass index, use of antihypertensive medication, and an estimate of genetic ancestry based on the AIM genotypes. I found that African genetic ancestry was positively associated with SBP (p=0.037) but not DBP (p=0.326)

after adjusting only for these standard covariates (Model A). Next, ascribed color was added to the model, along with SES and the *color**SES interaction term (Model B), based on previous research that identified *color**SES as a significant interaction (Gravlee et al. 2005). The interaction was again significantly associated with SBP (p=0.003). At the same time, the association between genetic ancestry and SBP was attenuated (p=0.176), implying that *color* and SES better contribute to BP variation than does genetic ancestry. Other interaction terms were tested, including SES*genetic ancestry, but none were significantly associated with blood pressure.

The direction of the interactions between *color* and SES on SBP and DBP are illustrated in Figures 4-2 and 4-3. It is interesting to note that SES does not have the same effect on blood pressure across different social classification categories. Specifically, as SES increases, those in the *trigueño-blanco* category show reduced SBP and DBP. In contrast, those classified as *negro* show the reverse pattern: higher SES is associated with higher SBP and DBP. These interaction plots demonstrate the complexity of the health effects of SES, which do not appear to be consistent across cultural groups.

Finally, I investigated the role of each of the six assayed polymorphisms by adding each separately to the models of SBP and DBP. (Tables 4-2 and 4-3, Models C and D). I first ran the analyses without sociocultural data (i.e. *color* and SES) and later added these variables to test their effects on the allele-phenotype associations. Without the sociocultural data, there was no evidence of a significant association between any of the candidate gene alleles and SBP or DBP (Tables 4-2 and 4-3; only SBP results for α_{2C}Del322-325 allele are presented). However, when *color* and SES were included as

main effects and as an interaction term, the Del322-325 polymorphism in the α_{2C} adrenergic receptor gene showed a significant association with SBP (a similar decrease in p-value was seen when sociocultural data were added to the model for DBP, although none of the DBP results were significant). Specifically, α_{2C}Del322-325 homozygotes showed a negative association with SBP (Table 4-1; p=0.0089). The statistical significance of this finding is especially compelling, given the relatively small sample size.

No consensus exists regarding the best method to correct for multiple testing in association studies. However, even a conservative Bonferroni correction (corrected p-value = 0.05/6 assayed polymorphisms = 0.0083) yielded a marginally significant association (p=0.0089). More importantly, however, the association between the α_{2C}Del322-325 polymorphism and SBP (or DBP) increased with the inclusion of sociocultural data in the model. Overall, the adjusted model accounted for 52.3% of the variability in SBP. The addition of the α_{2C} variant to the model accounted for significantly more variation than did models without the gene (adjusted R^2 =.475, F-test p=.0251) or models without the gene and the sociocultural data (adjusted R^2 = 0.3687, F-test p=.0047).

I further explored the group of α_{2C}Del322-325 homozygotes to determine why the addition of sociocultural data altered the association for this genotype. Within the homozygotes, I identified three categories of people; *negro*/low SES, *negro*/high SES and *trigueño* plus *blanco*/low SES. Within each category, the α_{2C}Del322-325 homozygotes had lower SBP relative to non-deletion homozygotes by an average of 14 mmH: (*negro*/high-SES: SBP of α_{2C}Del322-325 homozygotes =136.0 mmHg vs SBP of

non-deletion homozygotes = 152.0 mmHg; *negro*/low-SES: 111.3 mmHg vs 129.9 mmHg; *trigueño* plus *blanco*/low-SES: 117.6mm Hg vs 125.5 mmHg). Thus, the sociocultural information was needed to partition the culturally heterogeneous categories within the α_{2C}Del322-325 homozygotes and reveal a genetic association that was otherwise hidden.

The physiology of adrenergic receptors is not yet completely understood. In general, studies have shown that adrenergic receptors are involved in a complex feedback loop to regulate the release of epinephrine and norepinphrine, thereby mediating the effects of the sympathetic nervous system throughout the body (Philipp et al. 2002). Recently, SNPs in the α_{2A} gene have shown significant associations with metabolic diseases, including Type II diabetes and hypertension i.e. SNP rs553668 may confer a protective effect against high blood pressure (Rosengren et al. 2010). However, for the α_{2C}Del322-325 polymorphism assayed in this study, I identified a protective effect against high blood pressure that has not been shown before. Previous allele-phenotype association studies in largely African American populations have shown conflicting results for this variant. The largest study of this SNP (n=3398) found no association with the α_{2C}Del322-325 polymorphism and hypertension (Li et al. 2006b). The results of this study suggest that undetected sociocultural heterogeneity within previous samples of the deletion variant may have masked significant allele-phenotype associations.

Discussion

A major challenge of health disparities research is to disentangle the effects of social and genetic aspects of racial groups to determine the root causes of the disparity. This study is significant in providing the first example of independently measured

dimensions of race tested simultaneously to determine their relative contributions to variation in blood pressure. The results suggest: (1) social classification based on *color* better predicts blood pressure than does a genetic-based estimate of ancestry, and (2) adding *color* and SES reveals a significant association between a candidate gene polymorphism and blood pressure. These preliminary findings have important implications for future research on racial inequalities in health.

The Relationship between Ancestry, Sociocultural Environment, and Hypertension

A large literature supports the influence of sociocultural factors on health, which may mediate the relationship between African ancestry and hypertension seen in this study (Dressler 1991; Gravlee et al. 2009; Sweet et al. 2007). Social classification can affect environmental exposures linked to increased blood pressure, including the chronic stress of racism and discrimination (Brondolo et al. 2003), SES, poor neighborhood environment, or decreased access to healthy food, exercise opportunities, or healthcare (Gravlee 2009). As those with darker skin are often subject to more of these stressors than lighter skinned people (Klonoff and Landrine 2000), their worse health outcomes may be due to exposure to more of these adverse environmental conditions, rather than greater levels of African ancestry. The measure of *color* in this study appears to capture these meaningful differences in life experiences and exposure to social stressors that better explains blood pressure variation than genetic ancestry.

The significant interaction between SES and *color* in this study demonstrates different effects of SES across cultural categories. The pattern for *blanco/trigueño* respondents is in agreement with most studies which show increased SES to be

associated with increased health (or lower blood pressure). In contrast, those classified as *negro* in this study show the opposite pattern: higher SES is associated with higher blood pressure. This result is consistent with other recent findings, such as a larger study of blood pressure among African Americans, where higher SES in black individuals is associated with higher blood pressure (Sweet et al. 2007). These results may appear surprising, but support an interesting hypothesis suggested by Gravlee et al (2005): Puerto Ricans classified as *negro* may encounter more frequent frustrating social interactions at higher levels of SES than their lighter skinned counterparts due to increased institutional and interpersonal racism encountered in higher SES-settings. This hypothesis could be explored by further ethnographic analysis across the SES spectrum in different cultural settings.

Implications for Genetic Studies

In the recent age of high-throughput genomic studies, researchers are increasingly generating large amounts of genetic data, often for the purpose of generating estimates of continental ancestry. This ancestry data is typically included in genetic studies to correct for population stratification (Choudhry et al. 2006; Hoggart et al. 2003; Reiner et al. 2007; Shriver et al. 2003). However, geneticists also commonly test genetic estimates of continental ancestry directly as predictors for complex diseases, under the assumption of a racially-specific genetic predisposition to disease. For example, Tang et al. (Tang et al. 2006) tested for association between African genetic ancestry and hypertension in a large cohort of African Americans from the Familial Blood Pressure Program study. They report a positive but insignificant association between African genetic ancestry and various blood pressure measures, and conclude that genetic differences between blacks and whites must account for disparities in hypertension.

However, they do not test any potential confounders, such as sociocultural measures that may be linked with ancestry and alternatively explain the health disparity.

In this study, the addition of *color* and SES to the model revealed a significant genetic association that otherwise would have remained undetected. This result demonstrates another important reason to identify and include critical sociocultural variables into genetic analyses, as they may stratify the dataset to reveal the significance of candidate gene polymorphisms that may otherwise go undetected. Despite the publication of hundreds of gene association studies over the past decade, only a small proportion of heritable variation responsible for complex diseases has been identified (McCarthy et al. 2008), and less than 10% of risk for cardiovascular diseases in particular can be attributed to genetic variants (Mohlke et al. 2008). Improved measurement of sociocultural variables may be one factor that can improve our ability to detect associations between candidate genes and complex disease.

Future Directions

Despite a small sample size, this study significantly demonstrates significant associations (and interactions) between race and SES with BP. The next step is to replicate these findings with a larger sample size and in other geographic settings. Further ethnographic research may be needed to explore the social effects of darker skin color across the SES spectrum, in order to test the hypothesis that more pernicious racism is experienced at higher levels of SES. One limitation of this study is that the estimates of African genetic ancestry are based on a limited number of AIMs, and have 95% confidence intervals on average of 27%. Genotyping of additional markers could improve the estimates of ancestry and potentially alter the comparison of genetic and sociocultural data. Finally, the addition of data on health behaviors, including diet and

exercise, as well as additional measures of perceived racism and discrimination, could provide a more comprehensive picture of all the factors contributing to hypertension in future studies.

Table 4-1. Descriptive statistics for study variables.

	Blanco (n=37)	Trigueño (n=31)	Negro (n=19)	Total (N=87)
Systolic blood pressure	124.79 (15.96)	124.44 (20.59)	128.45 (16.46)	125.46 (17.71)
Diastolic blood pressure	79.74 (10.37)	80.37 (11.05)	80.21 (12.25)	80.07 (10.91)
Age, y	40.08 (7.42)	37.52 (8.06)	39.05 (9.48)	38.94 (8.11)
Sex, % female	56.76	74.19	42.11	59.77
Body mass index	28.75 (5.30)	28.96 (6.49)	27.52 (7.11)	28.56 (6.11)
Education, y*	15.59 (2.63)	13.06 (4.55)	12.47 (2.74)	14.01 (3.69)
Household income	6.05 (2.58)	3.39 (3.07)	3.31 (2.19)	4.51 (2.98)
Antihypertensive medication, %	10.81	12.9	0.0	9.2
African ancestry*	0.19 (0.10)	0.28 (0.12)	0.44 (0.11)	0.28 (0.15)

Means (SD) are reported for continuous variables, percentages for categorical variables. *ANOVA shows statistically significant differences across categories of *color* for education and African ancestry.

Table 4-2. Multiple linear regression of systolic blood pressure (SBP) on genetic ancestry, ascribed color, socioeconomic status, α2C adrenergic receptor genotype, and standard covariates.

	Model A		Model B		Model C		Model D	
	B	P	B	P	B	P	B	P
Constant	120.84	0.000	120.96	0.000	128.44	0.000	130.92	0.000
Age (years)	1.02	0.000	1.15	0.000	1.07	0.000	1.23	0.000
Sex (1=male, 0=female)	7.02	0.026	8.55	0.006	6.44	0.048	7.46	0.000
Body mass index	-0.02	0.944	0.01	0.966	-0.04	0.879	0.00	0.998
Antihypertensive use	19.53	0.001	21.87	0.000	19.90	0.001	22.42	0.000
African ancestry	22.22	0.037	18.18	0.176	27.55	0.017	19.96	0.139
Socioeconomic status			-2.00	0.294			-1.50	0.446
Ascribed *color*								
Trigueño versus Blanco			-3.44	0.348			-2.86	0.435
Negro versus Trigueño/Blanco			3.94	0.444			8.39	0.136
SES * ascribed *color*								
SES * Trigueño versus Blanco			-3.95	0.253			-4.23	0.228
SES * Negro versus Trigueño/Blanco			13.92	0.003			17.14	0.001
α_{2C}Del322-325 homozygotes					-12.60	0.107	-19.47	0.007
α_{2C}Del322-325 heterozygotes					-1.56	0.665	-1.58	0.619
Adjusted R^2	0.378	0.000	0.495	0.000	0.379	0.000	0.530	0.000

Body mass index (BMI) = weight kg / (height m)^2; SES = socioeconomic status; antihypertensive = self-reported use of antihypertensive medication. B = unstandardized regression coefficients.

Table 4-3. Multiple linear regression coefficients (B) for diastolic blood pressure (DBP) on genetic ancestry, ascribed *color*, socioeconomic status, α_{2C} adrenergic receptor genotype, and standard covariates.

	Model A		Model B		Model C		Model D	
	B	P	B	P	B	P	B	P
Constant	79.10	0.000	79.04	0.000	79.15	0.000	81.06	0.000
Age (years)	0.50	0.000	0.58	0.000	0.51	0.001	0.59	0.000
Sex (1=male, 0=female)	0.53	0.811	1.91	0.407	0.40	0.862	1.59	0.505
Body mass index	0.12	0.516	0.16	0.376	0.10	0.592	0.15	0.431
Antihypertensive use	8.17	0.037	9.98	0.010	8.05	0.048	10.04	0.013
African ancestry	7.37	0.326	4.62	0.645	8.03	0.325	5.41	0.608
Socioeconomic status			-2.06	0.150			-2.01	0.195
Ascribed *color*								
Trigueño versus Blanco			-2.01	0.466			-1.93	0.502
Negro versus Trigueño/Blanco			0.87	0.822			1.79	0.684
SES * ascribed *color*								
SES * Trigueño versus Blanco			0.04	0.988			0.12	0.967
SES * Negro versus Trigueño/Blanco			7.06	0.040			7.66	0.048
α_{2C}Del322-325 homozygotes					-2.99	0.593	-5.58	0.315
α_{2C}Del322-325 heterozygotes					0.43	0.869	0.15	0.952
Adjusted R^2	0.172	0.001	0.249	0.000	0.150	0.007	0.232	0.002

Sex (1=male, 0=female); Body mass index (BMI = weight kg / (height m)^2); SES = socioeconomic status; antihypertensive = self-reported use of antihypertensive medication.

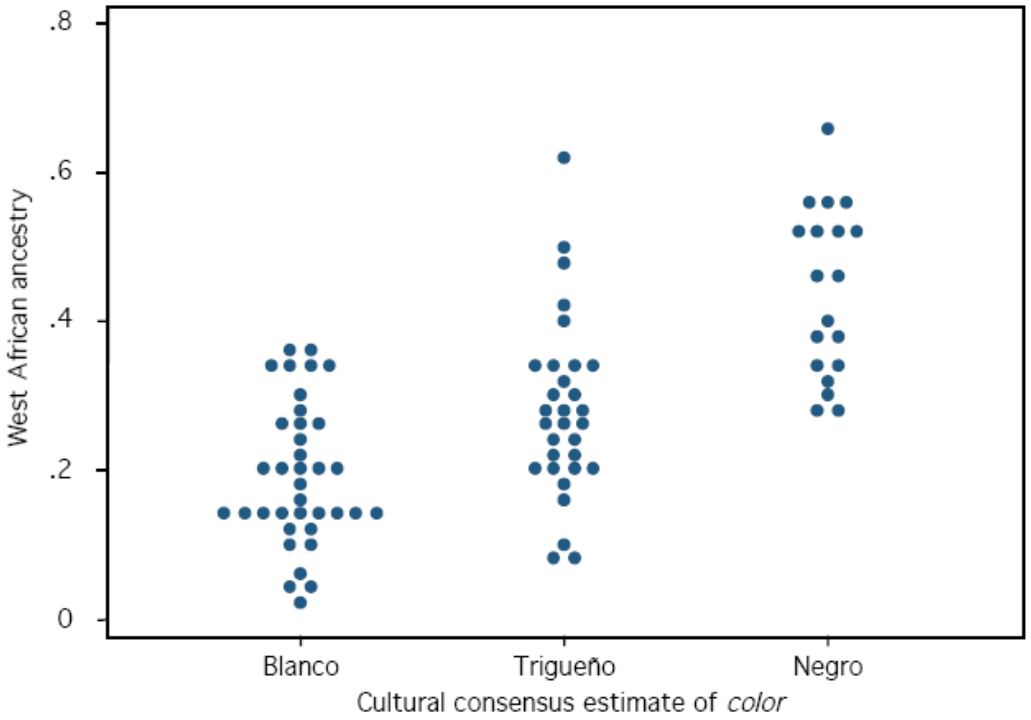


Figure 4-1. Relationship between individual level of African genetic ancestry and culturally ascribed *color*. Mean genetic ancestry varies across categories of *color* (*blanco* versus *trigueño:* $F[1,84]=12.04$, $p=.003$; *blanco* versus *negro*: $F[1,84]=73.65$, $p<.001$; *trigueño* versus *negro*: $F[1,84]=29.31$, $p<.001$; Šidák-correction for multiple comparisons). Individual African ancestry ranges from 0.08 to 0.69 in our sample. The distribution of African ancestry overlaps for *blanco* and *trigueño* across 40 percent of the observed range (0.08–0.35); for *trigueño* and *negro* across 48 percent (0.27–0.59); and for all three *color* categories across 12 percent (0.27–0.35).

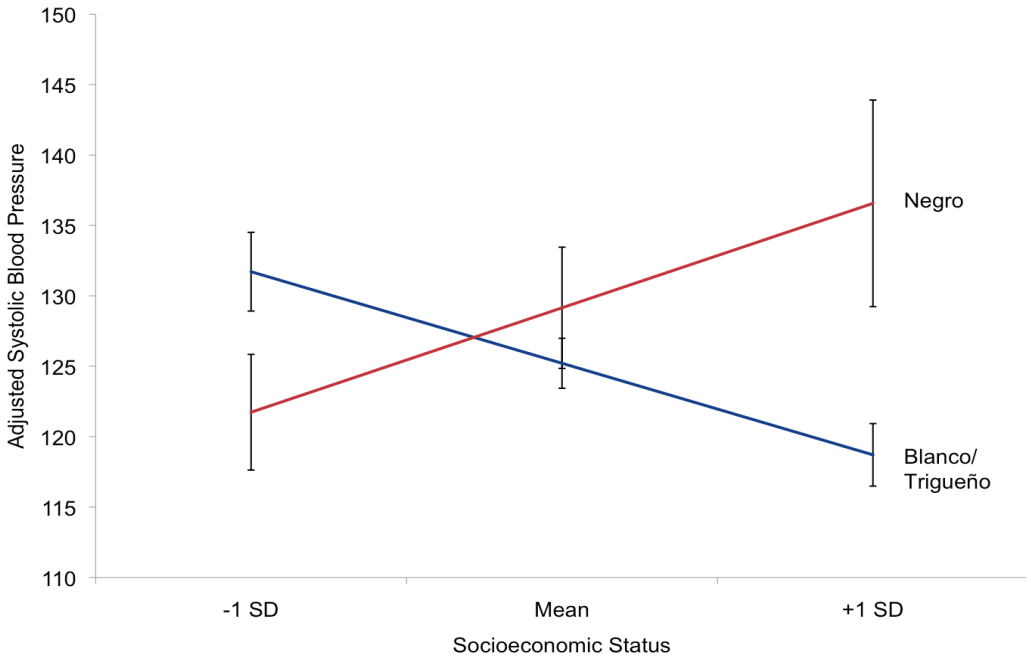


Figure 4-2. Interaction effect between culturally ascribed color and socioeconomic status for systolic blood pressure adjusted for age, sex, body mass, use of antihypertensive medications, and genetic ancestry.

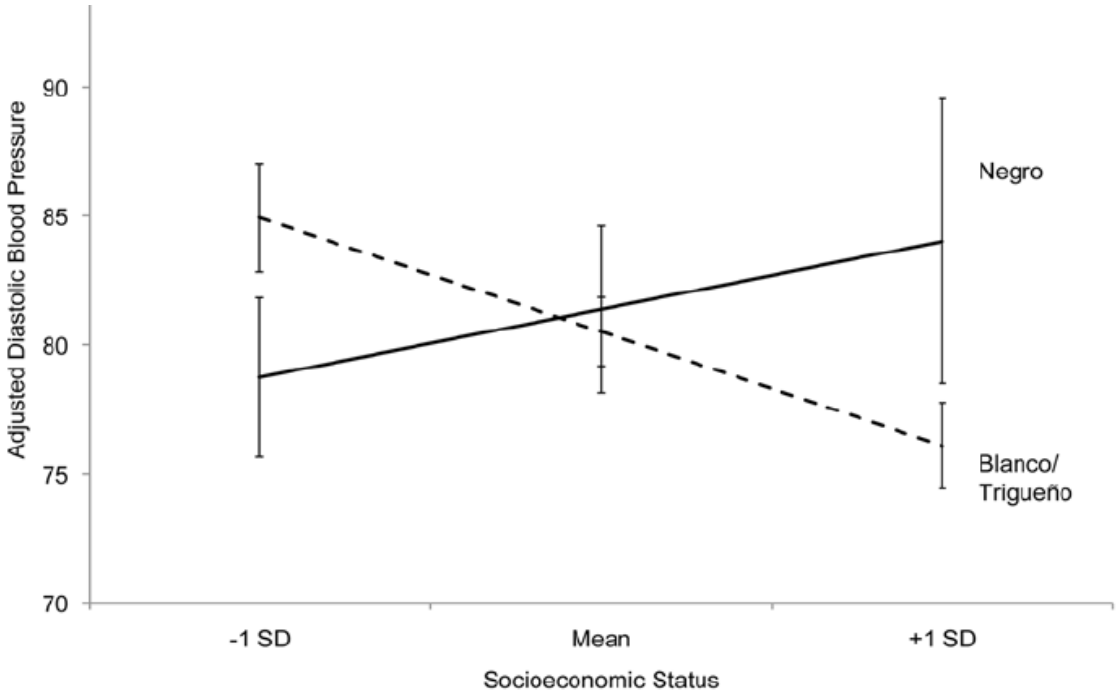


Figure 4-3. Interaction effect between culturally ascribed color and socioeconomic status for diastolic blood pressure adjusted for age, sex, body mass, use of antihypertensive medications, and genetic ancestry.

CHAPTER 5
EXPLAINING HEALTH DISPARITIES IN COMPLEX DISEASE: THE IMPACT OF EDUCATION AND GENETIC ANCESTRY ON BLOOD PRESSURE IN BLACK AND WHITE AMERICANS

Introduction

One of the great public health challenges in the U.S. today is to determine the origin of health disparities in complex diseases between Black and White Americans. Reduction in racial disparities in complex diseases would largely fulfill one of the two leading goals of 10-year national objective of Healthy People 2010, to reduce health disparities in general. The debate over the relative role of environmental and genetic factors in contributing to health disparities has remained unresolved, in part, because researchers continue to investigate each risk factor individually, without considering that both genetic and environmental components are important as independent and interacting contributors to racial disparities in complex diseases (Cooper et al. 2003; Gravlee et al. 2009; Martinez 2008; Williams et al. 2010). In this paper, I refer to racial disparities as differences in health between those self-identified as Black or as White in the U.S. Minority groups in the U.S, on average, experience different socioeconomic environments, which may account for a large component of racial disparities in health, e.g. the median household income for Blacks is ~$21,300 less than for Whites (U.S. Bureau of the Census 2008). Research from throughout the last decade has indicated that socioeconomic status (SES), acting either directly through resource deprivation, or indirectly through behavioral and psychosocial consequences, is an important predictor of health (Adler and Snibbe 2003; Cabrera et al. 2001; Dressler and Santos 2000; Gravlee 2009; Kim et al. 2005; Steenland et al. 2004a; Steenland et al. 2002; Steenland et al. 2004b; Sweet et al. 2007; Williams et al. 2010). One important aspect of the

socioeconomic environment is education, which has shown to independently influence risk of complex diseases and mortality in general, in Black and White populations (Albano et al. 2007; Franks et al. 2006; Kim et al. 2005; Steenland et al. 2002), Education may be an important contributor to health, not only for the associated increase in material resources and information about health behaviors, but also for the social status it confers, which may be linked to psychosocial factors that affect health (Adler and Snibbe 2003).

Another question in studies of racial disparities in health is the possibility of genetic differences between racial groups that contribute to complex disease. While genetic differences between continentally-defined racial groups are relatively small (between 2-3% of the variation in 377 microsatellite loci (Rosenberg et al. 2002)), and mainly located in non-coding regions of the genome, many researchers continue to hypothesize that genetic variants unique to Blacks may predispose them to develop certain complex diseases more often (Burchard et al. 2003; Risch et al. 2002). However, studies have yet to show any strong evidence for racial differences in any functional alleles predisposing to complex disease (Cooper et al. 2003). With the recent availability of multiple high-throughput genotyping platforms, a new wealth of genomic information is now available. One consequence of this new technology is that more and more researchers are estimating genetic ancestry to capture a presumed genetic basis to racial disparities in health (Reiner et al. 2007; Tang et al. 2006; Tsai et al. 2009). However, any associations found between genetic ancestry and disease are potentially confounded by crudely measured or unmeasured environmental factors such as socioeconomic status (SES), neighborhood environment, or psychosocial factors like

perceived stress or discrimination (Non et al. 2010b). We must be cautious that this new wealth of genomic data does not overwhelm the balance between genetic and environmental factors.

Research to understand and reduce health disparities in hypertension can make an especially important public health impact because hypertension is the leading contributor to racial disparities in life expectancy in the U.S. (Wong *et al.* 2002). Furthermore, because blood pressure is a complex phenotype, it illustrates the need to test both environmental and genetic hypotheses. A number of studies have independently shown the effects of socioeconomic factors, like education, (Gold et al. 2006; Kim et al. 2005; Knox et al. 1996), in addition to genetic factors, such as those implicated in the rennin-angiotensin-aldosterone system and the adrenergic system, among others, as contributors to higher blood pressure (Deo et al. 2007; Johnson and Terra 2002; Li et al. 2006a). Thus, analyses that jointly explore the root causes of inequalities in hypertension can also contribute to the larger debate over the relative roles of socioeconomic and genetic factors in complex diseases.

In this study, I examine a dataset from the Family Blood Pressure Program (FBPP), a large multi-center genetic study of blood pressure and other related conditions across multiple ethnic groups, funded by the National Heart, Lung, and Blood Institute (NHLBI). This dataset is valuable for testing hypotheses about the biological and social dimensions of race that contribute to disparities in hypertension as it contains genetic, biological and sociocultural data, specifically genetic markers for estimating ancestry, multiple measures of blood pressure, as well as education level and self-identified race and ethnicity (SIRE). A focused analysis of this dataset was previously

conducted to investigate the relationship between genetic ancestry and hypertension, but did not find significant associations between ancestry and blood pressure (Tang *et al.* 2006). I examine this dataset from both inter- and intra-population perspectives to address three new goals. First, using a large and geographically widespread subset of unrelated Black and White Americans, I test the hypothesis that education may be an important predictor of blood pressure (BP) both within the total sample, and within each racial group individually. I discover that education greater than a High School (HS) degree is significantly associated with decreased BP only in the Black population. Using the full dataset of all related samples, I next explore how education may be acting within each racial group through a series of regression analyses that incorporate important covariates and investigate potential interactions between race and education. I find that education is a significant predictor of various BP measures and associates with BP differently in each racial group. Finally, I further explore blood pressure variation just within the Black population by incorporating a measure of genetic ancestry to determine the relative roles of environmental and genetic factors in explaining the high rates of BP in Black Americans. The results show that education, and not genetic ancestry, is significantly associated with BP in this dataset.

Materials and Methods

Participants

The participants in the current study are members of the large FBPP database of 9,896 White, Black, and Hispanic individuals, as described in detail by Tang et al. (2006; 2005). Race was self-reported using the census category choices of African American, White, and Hispanic. As the current study is focused on Black-White health disparities, only African American and White individuals were included. Though self-

identified as African American, throughout the remainder of the paper these individuals will be referred to as Black for compatibility of terms with White. The White individuals included subjects from field sites located in Tecumseh, Missouri, Rochester, Minnesota, Forsyth County, North Carolina, Minneapolis, Minnesota, Framingham, Massachusetts, and Salt Lake City, Utah, while the Black individuals were located in field sites in Birmingham, Alabama, Forsyth County, North Carolina, Jackson, Mississippi, and Maywood, Illinois. IRB approval was obtained from the University of Florida to analyze the publicly available FBPP dataset.

Four datasets were created from the full FBPP database for three different sets of analyses. The first dataset was designed for categorical analyses of unrelated Blacks and Whites (UBW_categ), such as chi square tests and two sample tests of proportion. This dataset was comprised of unrelated Black (n=1614) and White (n=1185) individuals who were selected from all available individuals classified as hypertensive or normotensive, and not missing data on education. As these tests require unrelated individuals, and the FBPP dataset was composed of large pedigrees, a single individual was chosen from each pedigree in the manner described by Tang et al. 2005. Specifically, one hypertensive (HT) individual was selected from each family with a single HT subject and no other relatives and one HT individual was randomly selected from families with multiple HT subjects and no more than one normotensive (NT) subject. For NT subjects, one individual was selected from each family with a single NT subject and no relatives and one NT individual was randomly selected from families with multiple NT subjects and no more than one HT subject. After eliminating those missing data on education, this selection resulted in 1614 Black individuals (1283 HTs and 331

NTs), and 1185 White individuals (833 HTs and 352 NTs). For analyses of SBP, DBP, and MAP, (t-tests and ANOVAs) a second dataset of unrelated Blacks and Whites was created (UBW_cont.) for analyses of continuous BP variables. This dataset includes all the unrelated individuals not missing data on the continuous BP measures (including those excluded from the categorical analyses because they were classified as hypotensive) (n=1604 Blacks and n=1283 Whites). The third dataset was created for analyses that used more complex methods that enabled use of the related Black and White participants, but excluded those who were not taking hypertensive medication or missing data on the continous BP measures or education (n_{Blacks}=2034, n_{Whites}=1656). For this dataset, related individuals were included because relatedness could be accounted for in the modeling of the various BP measures. Finally, a fourth dataset of related Blacks (RB) (with associated genetic ancestry estimates) was created to allow for comparison of genetic and environmental data within one population. This dataset was comprised of all related Black individuals who did not have any missing phenotypes or more than 20% missing genetic markers (n=1077).

Blood Pressure and Covariates

Phenotypic data on three measures of blood pressure (SBP, DBP, and MAP) were utilized from the FBPP database. These biological measures were collected using standard procedures (FBPP 2002). Hypertensive status was also used as a composite variable, as classified in the FBPP dataset based on clinic blood pressure measurements, anti-HT medication status, and inclusion/exclusion criteria. Additional variables utilized as covariates from the database in the current study include age (in years), sex, SIRE (as chosen from census categories), field site, and education level (coded into 21 categories and also 6 categories). Specifically, education was coded into

a 21 category variable, ranging form no school (0) to 1+ years of grad school (20) (see Figure 5-1) and into six categories of no school, less than HS, HS, some vocational-technical school (vo-tech), some college, some grad school. The lowest categories of "no school" (0) and 1st Grade (1) were included in Figure 5-1, but excluded from all subsequent analyses as too few members fall into these category for statistical comparison.

Estimation of Ancestry

The methods for estimating ancestry in this study closely follow those of Tang et al. (2006; 2005) for the purpose of comparing results across studies. Genotype data were analyzed with the program *Structure* 2.2 (Pritchard *et al.* 2000) in order to estimate genetic ancestry of the Black individuals. To select parental populations in the analysis of the admixed Blacks, genotypes were included from a set of randomly selected unrelated self-identified Whites from across all FBPP networks (n=1300), and all unrelated sub-Saharan Africans from the World Diversity Panel (n=119), including Yorubans from Nigeria, Mandenka from Senegal, Biaka pygmies from Central African Republic, Mbuti pygmies from the Congo, Bantu from Kenya, and San from Namibia (Rosenberg *et al.* 2002), as chosen by Tang et al. (2006; 2005). Estimates were based on genotype data from a matching set of 294 autosomal microsatellite markers that were genotyped in both the FBPP dataset and the World Diversity Panel (from Marshfield Screening Set 8 and 10). These markers are spaced throughout the genome an average of 10cM apart. As ancestry estimates can become biased when using too many related individuals, all Black individuals were randomly divided into four subsets of mostly unrelated individuals, and paired with the same European American and African parental populations. Although I follow the methods of Tang et al. (2006)

closely, the average values of African ancestry in hypertensives (HT: 80.68%) and normotensives (NT: 79.1%) in the dataset differ from those previously reported (HT: 86.4% vs NT: 85.1%), due to differences in study participants available in the online public database versus the internally pooled databases of Tang et al. 2006.

Statistical Analyses

Analyses of education on BP (using UBW dataset)

Chi-square tests were used to test for differences in education categories among HTs and NTs in the combined sample of Black and White Americans, and also within each racial group independently. Two sample tests of proportions were also used to more specifically test the null hypothesis that the proportion of Blacks of lower education is the same among HTs as among NTs. Analyses were performed using education as both a six-category and two-category variable. T-tests were used to compare mean differences in SBP between the two education categories in the total sample of Blacks and Whites and in both unrelated Black and White populations separately.

Inter-population analyses of BP (in RBW dataset)

Regression models were used to test for associations between education and each response variable (SBP, DBP, MAP) in the total sample of combined Black and White populations. Because many of the FBPP participants were related, regressions were conducted using the program SOLAR, which allows for familial relatedness, and calculates a variable of residual heritability, (i.e. the residual proportion of variance in BP due to familial (genetic) effects after the effects of significant environmental covariates were removed. The first model adjusted for standard covariates, including age, sex, SIRE, residual heritability, and field site. Cross product interaction terms were sequentially included to test for interactions between education and SIRE. The

covariates age^2 and an age*sex cross product interaction term were also both tested in each regression analysis, as they were found significant in some previous analyses of this dataset (Tang et al. 2006).

Analysis of BP and genetic ancestry (using RB dataset)

In the final analyses, T-tests were used to test for significant differences in African genetic ancestry between HTs and NTs among the unrelated Black individuals within each field site and across all centers using t-tests. Dot plots, box plots, and loess plots were also created to test for an association between education and African ancestry in the Black population. The Loess curve fits simple models to localized subsets of the data, creating a smooth curve through the data utilizing weighted least squares methods. Ancestry was compared with education as a 2-category, 6-category, and 20-category variable.

Multiple linear regression models were also used to test for associations between African genetic ancestry and each response variables of blood pressure (SBP, DBP, MAP) in all untreated Black participants (n=1466), adjusting for relatedness in SOLAR as I did for the RBW regressions. The first model adjusted for the same standard covariates as used in the RBW analyses, including age, sex, residual heritability, and field site. I then sequentially added education, treated either as a continuous 20-level variable or as a 6-level or 2-level categorical variable. Next I added, education*ancestry interaction terms, along with other interactions suggested by previous studies. Only the age*sex interaction term was adjusted for in final models of SBP, where it was a significant predictor (but not in models of DBP, or MAP).

Sensitivity analyses

For the regression analyses, the 20 available categories of education were treated as a continuous variable, as well as grouped into a two-category variable (<=HS degree, >HS degree) and into six categories (no school, completed 2 to 11 years of secondary school, completed 1-3 years of vo-tech school, completed 1+ years of college, and completed 1+years of graduate or professional school). The final presented models treat education as a 20 category variable, as the wider measurement scale provides the most narrow confidence intervals around the Beta estimate for education. All models were also tested with age^2 and sex^2, and age*sex following Tang, though only the age*sex interaction was significant in any of the regressions and included in the reported models.

As all regression models used related individuals, they were modeled in three different ways to adjust for relatedness. The presented data were modeled in SAS using generalized estimating equations in proc genmod, to account for correlated observations within families, as described by (Horton and Lipsitz 1999). Two structures of the working correlation matrix were tested, including the exchangeable (equal correlation estimated for each pair of individuals within a family) and the unstructured (unique correlation estimated for each pair of individuals within a family). The third method used a random effects model in SOLAR, which uses pedigree information to calculate residual heritability, a random effect included in the modeling of the BP outcome. The regression estimates based on the GEE using the exchangeable correlation matrix are presented in the current paper, because more flexibility in modeling was permitted by SAS, and they produced regression estimates similar to those produced by the random effects model in SOLAR. The exchangeable correlation

matrix was chosen over the unstrutured because the sample size was not large enough to support the large number of regression parameters required by the unstructured matrix.

Regression case diagnostics were examined with plots of residuals against predictors in each model. Imperfectly centered plots of residuals were corrected by log transforming the BP measures. In all models tested, residual kurtosis was high, and thus BP outcomes were also tested following a log-transformation (data not shown). The transformed models showed the same substantive results: African ancestry was not significant in any of the models (with or without education or BMI), but education was highly significant (p=0.002). The non-log-transformed results are reported in all tables for ease of interpretation. Finally, correlation between the predictors of African ancestry and education (as a 20-category continuous variable) was also tested to avoid multicollinearity and were not found to be highly correlated (r=0.12).

Results

Is Education an Important Predictor of HT?

Comparing HT status by education level in Blacks and Whites. The distribution of education level for all unrelated individuals within each of the Black and White populations are shown in histograms in Figure 5-1. The highest frequency category of education within both racial groups was High School (HS) graduate (Black: 31.58%, White: 39.18%). Education level was generally lower in the Black population, with 28.42% of the Black population achieving less than a HS degree versus only 6.49% of the White population.

The first set of analyses was conducted to determine if education is associated with hypertension status within the large and geographically diverse sample of Black

and White Americans across the U.S. Using a Chi-square test, I found that within the total sample, HT status differs significantly by education level ($\mathbf{X^2}$=24.87, p<0.0001, Table 5-1). I also tested for significant differences in various measures of BP across education levels (with education grouped as less than or equal to HS degree or more than a HS degree) using t-tests and tested for significant differences across six education categories using ANOVA. Mean SBP and MAP (but not DBP) each differed significantly among those of lower versus higher education (Tables 5-2). Education was then divided into six categories, to further explore where the significant differences lie in SBP and MAP. ANOVAs showed significant differences in SBP (F=18.06, p<0.0001) and in MAP (F=7.6, p<0.0001) between at least 2 education categories (Table 5-3). Tukey tests revealed significant differences in SBP between the category of less than HS degree and every other category of education, and significant differences in MAP between the category of less than HS degree compared to HS degree, some college, or some grad school.

In the next set of analyses, the total UBW sample was divided into racial groups, to test where the association between education and BP was strongest. I discovered a significant difference in hypertensive status across education levels for Blacks in each field site and for all field sites combined (Table 5-4). To determine the direction of the association between education level and hypertensive status, I used a two sample test of proportions (e.g. tested the null hypothesis that the proportion of Blacks of lower education is the same among HTs as among NTs). This hypothesis was rejected, as the data showed a significantly higher frequency of Blacks with lower education among HTs at each individual field site (all p<0.05) and at all field sites combined (p=0.0004) (Table

5-5). The same pattern was not seen in the White population. Neither Chi square tests nor two sample tests of proportion showed any significant difference in hypertensive status by education level in any of the field sites of White participants or in all field sites of White participants combined (Tables 5-4 and 5-5). In fact at one field site (Tecumseh), the proportion of NTs at lower education was higher than the proportion of HTs at lower education, which is the reverse pattern seen in each of the field sites with Black participants, though it is important to note that very few samples fall into this category.

Mean SBP, DBP, and MAP were also compared across education levels within both unrelated Black and White individuals using t-tests and ANOVA (Tables 5-6 and 5-7). There were significant differences in mean SBP and MAP between Black individuals of low versus high education (Table 5-6). In all cases, those of low education showed higher mean SBP and MAP than those of high education, though the difference in was most pronounced in measures of SBP ($p=<0.001$). Conversely in the White population, no significant differences were seen in SBP, DBP, or MAP for those of low versus high education (though SBP and MAP were nearly significant, $p=0.056$, $p=0.085$). The effect size of higher education in both SBP and DBP was in fact more than twice as large in the Black population than in the White population (4.60 mmHg vs. 2.24 mmHg difference in SBP, and 2.02 mmHg vs 0.028 mmHg difference in DBP, for those with education greater than HS degree. When education was divided into six categories, ANOVA results showed significant differences across education levels for SBP ($F=8.87$, $p<0.0001$) and MAP ($F=3.95$, $p=0.0034$) for Blacks but no significant difference in SBP ($F=1.56$, $p=0.184$) or MAP ($F=0.57$, $p=0.68$) for Whites between any education

categories (Table 5-7). In the Black population, significant differences in SBP were seen specifically between those with less than HS degree ($p<0.05$) versus those from each of the other education categories (except vo-tech), and for MAP only those with less than High School Degree were significantly different from those with some grad school ($p<0.05$). In sum, significant differences by education status were found in the Black population, but not among the White population, when testing either the categorical measure of hypertensive status or the continuous variables of SBP or MAP.

How is Education Important?

Modeling blood pressure among combined Black and White populations. After establishing the association of education with blood pressure through simple analyses of the unrelated samples, I next explored *how* education is behaving within the Black and White populations. This question was examined in a larger and more complex dataset of related individuals, using regression analyses that can account for relatedness, as well as important covariates that may alter the associations. On this dataset, I first conducted linear regressions on BP in the combined Black-White population (RBW) in order to confirm the role of education, among other factors, in contributing to the racial disparity in blood pressure. Specifically, I examined the impact of self-identified race (SIRE) and education on the various BP measures of SBP, DBP, and MAP (Table 5-8). All regressions presented here were modeled using the GEE to account for relatedness within each family cluster. As BMI was found to be a significant predictor of BP in previous studies (Tang et al. 2006), it was included in all models, along with standard covariates of age, sex, and SIRE, as well as an interaction term for age*sex. All of the variables in this model were significant predictors of SBP with p-values <0.0001. Education shows an important impact on SBP, even after adjusting for

SIRE, as it estimates a reduction of 0.57+/-0.09 mmHg for each increased year of education. A large health disparity is also evident, as those self-identified as Black are predicted to have 4.33+/-0.65 mmHg higher SBP than those self-identified as White, even after adjusting for education level. An interaction term between SIRE and education was also tested, but was not found to be significant (p=0.3448) and thus not presented in the final model. Similar results were also found for models of DBP and MAP, such that education was highly significant (p<=0.002) even after adjusting for SIRE, and SIRE was significant after adjusting for education (p<0.0001) in both models (Table 5-8). The interaction between education and SIRE was again not significant in either of these models (p>0.290). To test if education affects BP in a non-linear pattern, a quadratic term ($education^2$) was also tested in the SBP model but not found to be significant (B=-0.0014, p=0.9577) and not included in the final model.

In order to explore the independent effects of each level of education further, models of SBP, DBP, and MAP were tested with education coded into two or six nominal categories, and each category was tested in an interaction with SIRE (Tables 5-9 and 5-10). Each category of education was compared to the sixth category of 'some graduate school' which was chosen as the reference category for comparing to the effect of each level of education. Treating education as a nominal variable allows for independent effects of each level of education without assuming a predetermined order or evenly spaced distances between levels. Education was found to be a significant predictor of blood pressure in both the Black and White populations at almost all levels. Unlike the models in which education was treated as a continuous variable, in these models the effect of education differs between the racial groups when an interaction

between education and SIRE was included. Specifically for the model with two categories of education, those with education lower or equal to a HS degree are predicted to have an increase in SBP by 4.68+/-0.83 mmHg ($p<0.0001$) for Blacks and 1.48 +/-0.83 ($p=0.0768$) mmHg for Whites relative to those with more than a HS degree. The interaction term between education and race was significant at $p=0.0064$. The direction of this interaction can be seen in the interaction plot in Figure 5-2, Panel A, where it is clear that higher education causes a larger decrease in blood pressure for Blacks than for Whites. A similar pattern was seen for DBP, in that education less than or equal to a HS degree also caused a significant increase in DBP for Blacks by 1.88+/-0.54 ($p=0.0005$) but not for Whites, who actually showed a (non-significant) decrease by -0.22+/- 0.51 mmHg ($p=0.6629$). Again the interaction term was significant (β =2.11+/-0.75, $p=0.0047$). Finally, the same pattern was seen for MAP as for SBP, in that education less than or equal to HS caused an increase in MAP by 2.82+/-0.60 mmHg for Blacks ($p<0.0001$), but no significant increase for Whites (β =0.34+/-0.58, $p=0.5541$), and again the interaction term was significant (β =2.48+/-0.83, $p=0.0028$).

For the model with six categories of education, the effect of each level of education on each BP measure also differed between the SIRE groups (Table 5-10) though none of the interaction terms between the education level and SIRE were significant (all $p>0.10$). Specifically, in the Black population, SBP steadily decreased with increasing education with the exception of the vo-tech category, which also caused a decrease, but did not fit into the linear pattern: those in the Black population with less than a HS degree, a HS degree, some vo-tech school, and some college were predicted to have a change of SBP relative to those with some grad school by 6.70 +/-2.01

(p=0.0008), 5.04+/-1.84 (p=0.0063), 0.49+/-2.77 (p=0.8589), and 1.25+/-1.81 (p=0.4894) mmHg, respectively. In the White population, SBP decreased with each increased level of education relative to graduate school: SBP increased by 6.53 +/-2.31 (p=0.0047) mmHg for those with less than HS degree, 4.36 +/-1.53 mmHg (p=0.0045) for those with a HS degree, 5.74 +/-1.72 mmHg (p=0.0009) for those with some vo-tech school, and 3.05 +/-1.51 mmHg (p=0.0430) for those with some college, respectively. The main effect of SIRE was significant even after adjusting for education in this model (p=0.0188), indicating that the education categories were not completely explaining all SBP differences between the racial groups. The different effects of each level of education on SBP in Black and White populations are illustrated in Figure 5-2, Panel B.

Finally, I also examined the effects of education (6 categories) and SIRE separately within each gender to test how all three factors of education, race, and gender contribute to BP (Figure 5-3). The effects of gender appear even larger than the effects of race, consistent with the highly significant effect of sex throughout all regression analyses on the combined Black-White population (Tables 5-8 through 5-10). Specifically, Black males are predicted to have the highest level of SBP, followed by White males, Black Females, and White females, respectively, across all 6 levels of education (with the exception of vo-tech school, in which White males are predicted to have higher SBP than Black males). The education-SBP gradient across the education levels also differed significantly by race and gender. Within Black males, the SBP-education gradient (Figure 5-3) largely mirrors the gradient in the total Black population (Figure 5-2), with highest SBP at lowest levels of education, but very little difference between some college or some graduate school. In White males, (Figure 5-3) SBP

largely mirrors the pattern in the total White population (Figure 5-2), except all SBP levels are higher. In Black females, there is a general decline of SBP across levels of <HS degree, to HS degree, to some vo-tech, and then SBP levels off among higher education categories, with a slight decrease at some grad school relative to some college. In White females, SBP is fairly consistent and low across the entire education spectrum, and only drops at the highest level of education, some graduate school. These results suggest that gender, race, and education all have unique effects on BP, and should all be considered simultaneously to understand the complete effects of education on health disparities.

Is Genetic Ancestry Necessary?

Many previous studies have argued for the importance of testing for genetic ancestry in admixed populations to explain racial disparities in complex disease. Blacks in the U.S. are typically described as an admixed population comprised of a mixture of African and European ancestry. In this final set of analyses, I examined the association of genetic ancestry with blood pressure within the Black population relative to a socioeconomic factor, e.g. education, in contributing to higher rates of hypertension and higher blood pressure measures. First, t-tests were performed to test for differences in levels of African ancestry by HT status among all unrelated Blacks (dataset UB). Unlike education, average levels of African ancestry did not differ significantly between unrelated HT or NT Black individuals in the total sample ($p=0.103$), or at each individual field site (all $p>0.2$, Table 5-11). However, an analysis of variance combining all four field sites together shows African ancestry to differ significantly by field site ($p=0.0003$), (data not shown). These results are comparable to those reported by Tang et al. 2006,

in showing a non-significant difference in African ancestry between HTs and NTs, though their estimates of ancestry were slightly different from those reported here.

Modeling blood pressure within the Black population. A final series of regression analyses was performed to test the relative role of African genetic ancestry and education on blood pressure variation just within the Black population (including all related individuals not on hypertensive medication, n=1466, (dataset RB)). All analyses account for relatedness among the samples in the same manner as described in the previous regressions. Additionally, field sites were included as covariates in these models to test for significant differences between locations (field sites were not included in regression analyses with Black and White individuals because SIRE was essentially synonymous with field site, thus causing a problem of multicollinearity). The first model (Table 5-12, Model A) shows the effect of education (as a continuous variable) along with age, sex, BMI, and field site in predicting SBP. In this model, education is highly significant (p=0.0005) in predicting BP among Blacks. The next model (Table 5-12, Model B) adds genetic ancestry. In this model, the effect of African ancestry is low and not significant (β =5.23+/-4.55, p=0.2507), while the effect of education, though small (β =-0.51+/-0.15), is highly significant (P<0.001). Specifically, the β coefficient for education indicates that each increasing year of education is associated with a 0.51 mmHg decrease in SBP. In the final model (C), an education by ancestry interaction term is tested, but is not found to be significant (p=0.2081). Thus, none of the models show African ancestry to be a significant predictor of SBP either as a main effect or in an interaction with education, whereas education was significant in models with and without ancestry included. The same three models were also used to test the effects of

ancestry and education on DBP and MAP (Tables 5-13 and 5-14). The results for DBP and MAP are similar to those found with SBP: African ancestry is not significantly associated with either BP measure, but education shows a significant effect for MAP (p=0.0076 in model with just education), but not for DBP.

To further explore the associations between education, ancestry, and SBP, all of the models in Table 5-12 were also tested with education coded as a nominal variable of 2 categories and as 6 categories. These models showed similar results, in that most categories of education were significantly associated with lower BP, while African ancestry was not (Data not shown). Specifically, when education was split into two categories, education less than or equal to a HS degree predicted a significant increase in SBP by 3.77+/-1.38 mmHg (p=0.0063) relative to those with greater than a HS degree, while African ancestry had no statistically significant effect on SBP (β =6.29+/-4.57, p=0.1679), and the interaction term between ancestry and education was not significant (p=). In the model with education divided into six nominal categories each level of education (non-significantly) changed BP relative to the highest level of some grad school (e.g. less than HS degree (β = 4.20+/-2.81, p=0.1351), HS degree (B=2.05 +/- 2.56, p=0.4228), some vo-tech school (B=-3.08 +/-3.45, p=0.3719), and some college (B=-1.02+/-2.54, p=0.6891). This model basically reproduces the same analysis of SBP as was conducted in the RBW dataset, with the addition of ancestry, and results in a similar pattern of SBP across education levels as seen in the interaction plot of Figure 5-2. Again, when this model was tested with an interaction term between ancestry and each level of education, none of the cross-product terms were significant.

Although education and ancestry did not show a significant interaction in any of the tested models, it is important to test if genetic ancestry is associated in any way with education, under the hypothesis that higher African genetic ancestry, and thus darker skin color, may serve as a marker for environmental exposures that limit educational opportunities. If this were true, and if education is causally associated with blood pressure, education may serve as a confounding factor in the relationship between genetic ancestry and SBP in any previous studies that neglected to include this variable. To test for association, a Pearson correlation was conducted between African ancestry and education, but no linear association was evident (r^2=0.014). When African ancestry was plotted against the six categories of education, however a categorical trend of increasing African genetic ancestry with increasing education level is evident (Figure 5-4). An ANOVA test shows a significant difference between at least 2 categories (p=0.0001), and Tukey tests indicate significant differences in African ancestry between HS degree or less education versus college or more education, and also between some college and some graduate school (data not shown). The individuals with some vo-tech training have the widest range of variation, and do not appear to have any significant difference in African ancestry from those with a HS degree or those with some college, which is unsurprising given that vocational training can provide a wide range of job opportunities, and may not serve as a clear status marker in society.

Discussion

Intra versus Inter-Population Analyses

One of the most controversial and pressing issues in public health in the U.S. is how to explain racial inequalities in complex disease. The results of this study address racial disparities in a complex phenotype of hypertension -- the phenotype that causes

the largest differences in Black and White life expectancy in the U.S. (Wong et al. 2002). A recent critique of typical epidemiological approaches has claimed that too few studies consider both within and between group variance when addressing health disparities (Schwartz and Meyer 2010). Specifically, studies that focus only on inter-population differences enable identification of causal exposures (e.g. education), but cannot explain how the exposure is leading to the outcome (e.g. hypertension). Conversely, many studies too narrowly focus on within-group processes that can reveal the mechanisms of the exposure but cannot show its differential effect between racial/cultural groups. One strength of this study is the inclusion of both intra and inter-population analyses, as I examine the role of education both across and within racial groups. The analyses of the combined Black and White Americans provide an opportunity to demonstrate the magnitude of the racial disparity in hypertension, and the differential effect of education on hypertension between racial groups. The within group analyses further explore the effects of education within each racial group and allow direct comparison of education with African ancestry in the Black population. In sum, the results across all levels of analyses show a large racial disparity in BP measures, and consistently suggest that education represents one important socioeconomic factor that acts differently in Black and White populations, and may contribute to racial differences in hypertension.

The Magnitude of the Racial Disparity

Demonstration of significant racial disparities in blood pressure in this large and geographically widespread dataset contributes to a broad body of literature supporting the persistence of racial disparities in the U.S. The first regression analysis (Table 5-8) shows that on average, Blacks are predicted to have 4.33 +/- 0.65 mmHg higher SBP

than Whites, after adjusting for education as well as standard covariates, e.g. age, sex, and BMI. This statistically and clinically significant difference in blood pressure is in part responsible for the large racial disparities in related diseases in the U.S., as uncontrolled blood pressure can lead to heart attack, heart failure, stroke, and other related cardiovascular symptoms – all of which are more common in Black than in White Americans (http://www.americanheart.org./presenter.jhtml?identifier=2129). Thus, the magnitude of this disparity across the U.S. has significant public health implications, and demonstrates that the goals to reduce health disparities in Healthy People 2010 are from complete (Soares et al. 2009).

Is Education an Important Predictor of BP?

One important goal of this study was to determine if education is an important contributor to higher blood pressure in the total sample of Black and White Americans. In all analyses of unrelated Blacks and Whites, (e.g. Chi square tests, two-sample tests of proportion, t-tests, and ANOVAs), a significant association was found between education and hypertension status across all BP measures (Tables 5-1 to 5-3, 5-8). Specifically, each year of education predicted a decrease of -0.55 +/- 0.10 mmHg in SBP, after adjusting for the effect of SIRE (Table 5-8). When education was coded as two categories, education less than or equal to a HS degree caused an increase in SBP of 1.59 mmHg ($p<0.0001$) relative to education greater than a HS degree (Table 5-9). These results are consistent with the general findings in the literature that education is associated with risk of complex diseases and mortality (e.g. Albano et al. 2007; Farmer and Ferraro 2005; Gold et al. 2006; Kim et al. 2005; Knox et al. 1996).

Despite its significant effect, education does not completely explain the disparity in BP between racial groups. The effect of education in this study was in fact smaller in

magnitude than the effect of SIRE, where self-identification as Black caused an increase in BP of 4.33+/-0.65 mmHg, even after adjusting for education effects (when education was coded as a continuous variable of 20 categories, Table 5-8). The persistence of racial disparities, even after accounting for SES variables, is a common finding in health disparities research (Albano et al. 2007; Crimmins et al. 2007; Franks et al. 2006; Kaufman et al. 1997). This residual confounding may be due, in part, to the difficulty in accurately capturing SES with a single variable of education (Kaufman et al. 1997), and the fact that measures of SES can have different meanings between racial group, i.e. college-educated blacks are more likely to be unemployed and have lower incomes than college-educated whites (Williams and Collins 1995). However, education still proves to be a useful predictor of BP, and in fact other studies have shown it to be a better predictor of all-cause mortality, including all-vascular disease mortality, than other area-level indicators of SES (Steenland et al. 2004b). The significance of education in these analyses confirms that when multiple measures of the socioeconomic environment are not readily available, simple proxies for SES, like education, still prove useful for capturing some aspect of the social environment. Other cultural and psychosocial factors should also be considered, such as psychosocial stress and racism/discrimination, which are likely also contributing to the persistence of racial disparities (Adler and Snibbe 2003; Gravlee 2009).

Why is Education an Important Predictor of BP?

Many studies have posited various explanations as to how educational attainment confers health benefits, either as a direct causal factor or as an indirect marker of associated factors, ranging from 1) increased knowledge of protective health behaviors (i.e. healthy diet and exercise), 2) increased income, which may permit easier access to

healthier foods, exercise facilities, and/or healthcare, 3) increased engagement with people in different social networks that potentially expose them to wider social norms, 4) education may serve as a marker for individual personality traits that are associated with both educational achievement and better health, such as conscientiousness, self-discipline, adaptability to new information, or future orientation (Adler 2009; Johnson et al. 2009), 5) education may affect early life experiences (Case and Paxson 2002; Wadsworth 1997), 6) education may decrease psychosocial stress (Adler and Snibbe 2003; McEwen 1998), or 7) unmeasured genetic factors may be linked with education and health (Johnson 2009, Rutter 2007). It is likely that many, if not all of these explanations may be capturing different aspects of education that affect blood pressure and general health, and should be tested through more ethnographic and other qualitative analyses.

Does Education Explain the Racial Disparity in BP?

Education may contribute largely to the racial disparity in blood pressure due to the significantly distinct effects for each racial group demonstrated across analyses in the current study. For example, when education was classified as less than/equal to versus greater than a HS degree, education was significantly associated with BP measures in the Black but not in the White population (Tables 5-2 to 5-7). The regression analyses also showed racial differences in BP across levels of education. The interaction plots in Figure 5-2 specifically show three consistent trends in the relationship between race, education, and BP: 1) SBP is higher in Black populations than White populations across all levels of education; 2) In both Black and White populations, SBP is generally higher in those of lower versus higher levels of education; and 3) Education has different effects on BP between racial groups. With respect to the

differential effect of education, in the White population, each increase in level of education reduced SBP significantly, as expected, with the exception of vo-tech school, which caused an increase in SBP. In contrast, for the Black population, the results suggest less of a benefit of higher education above the vo-tech level on SBP: some college education surprisingly showed an increase in SBP relative to vo-tech and some grad school showed a very minimal decrease in SBP relative to some college, and as compared with the clear benefits in the White population (Figure 5-2). The results for the vo-tech category of education were unique, in that relatively equivalent SBP measures were seen across the racial groups, but with extremely wide confidence intervals. These results suggest an ambiguous affect on health within the vo-tech education category, perhaps because vo-tech education does not confer a clear or consistent social status in U.S. society.

The racial difference in BP across education categories is intriguing, especially when considered within the context of the long-standing public health debate over the relationship between SES, race, and health disparities (Farmer and Ferraro 2005; Franks et al. 2006; Williams et al. 2010), Though race and SES are clearly important predictors of health, controversy remains over the direction of the interaction between SES and race, which often differs according to ethnic group, geographic location, or disease phenotype under study. For example, a clear dose-response education-health gradient has primarily been reported in the U.S. White population, but has not been consistently shown across all U.S. racial/ethnic groups (Acevedo-Garcia et al. 2007; Goldman et al. 2006; Kimbro et al. 2008). Two different models have been posited to explain the different effects of SES commonly seen between racial groups. The first,

termed the “diminishing returns” hypothesis, refers to those studies that find a greater gap in health at the higher end of the SES spectrum (Bowles and Gintis 1976; Farley 1984; Farmer and Ferraro 2005; Gravlee et al. 2005). This gap is explained by the idea that Blacks do not benefit as much as Whites from higher SES, or higher education in particular, perhaps due to fewer income benefits of higher education or greater awareness of social injustices and discrimination at higher levels of SES. The second model, termed the “minority poverty” hypothesis, posits that the largest gap in health between Black and White Americans is at the lower end of the SES spectrum, and this gap diminishes as health improves for Blacks at higher SES levels. This hypothesis is based on the compounded disadvantages faced by Black people living in poverty and experiencing discrimination, which exaggerate the differences in health at lower levels of SES. One study that supports this hypothesis has shown that Black Americans with low SES have higher lifetime morbidity odds than if either factor is tested independently (Gold *et al.* 2006). Waitzman and Smith (1994) similarly report that Black adults at the highest risk of hypertension are found in the bottom of the occupational hierarchy.

The data in the current study are unique in their consistence with both hypotheses of minority poverty and diminishing returns, in that large differences in SBP are seen both at the low and high end of the education spectrum, but similar values in the middle. Although none of the interaction effects between each of the six levels of education and SIRE are statistically significant, the data show a trend that is consistent with both hypotheses. The worst blood pressures are predicted for people who face the double burden of being less educated and self-defined as Black. Increasing education improves the situation—but only to a point. Above the vocational tech level, higher education

yields no further benefits in terms of BP in the Black population. The "diminishing returns" from increased education are likely due to the persistence of racism/discrimination limiting opportunities even among the more highly educated in the minority group. The diminishing returns hypothesis was first suggested, when Bowles and Gintis (1976) noted that minorities receive fewer payoffs for education than Whites, including less high-paying jobs at levels of high educational achievement. Though the racial gap in education (at least at the HS level) has considerably narrowed in the past 50 years, (e.g. 23.1% gap between Blacks and Whites earning HS degrees in 1960 versus a 3.9% in 2006), the income gap among college-educated Blacks versus Whites widened by 10% between 1979 and 1997 (Changing America 1998). The evidence for diminishing returns, in this study and others, suggests that despite improvements in educational attainment, other factors are still affecting the health of Black Americans at the higher end of the SES spectrum. Some explanations for racial differences in SBP across the SES spectrum have been proposed by recent studies. At high levels of SES, racism and discrimination are suspected to be even more pernicious than at the lower levels, as indicated by results of recent studies by Gravlee et al. (2010), and Sweet et al (2009). Neighborhood-level factors, such as residential segregation, may also affect the quality and location of education opportunities across racial groups. A recent study found that location of education (foreign versus native) was a significant predictor of health in Asians (Walton et al. 2009).

Gender is also an important factor interacting with race and SES, and may contribute to the different education-BP gradients across racial groups. The analyses separated by gender (Figure 5-3) revealed health benefits of education on SBP in Black

men and women only up to the vo-tech level, after which increased education had little effect on women's BP, and was in fact associated with higher SBP among Black males. In White men, education causes a significant decrease in BP across the education spectrum, (with the exception of vo-tech education). However, in white women education does not appear to have much effect on SBP. These data are consistent with other recent studies that show complex interactions between race, education, and gender. For example, education was associated with a higher risk for cancer mortality for Black versus White men across 5 of 6 education levels, but black women had a higher risk than White women only at the 3 highest education levels, and a lower risk at the lower 3 education categories (Albano et al. 2007). Similarly the CARDIA study found a complex interaction between race, SES, and gender, i.e. education was associated with higher risk of a poor lipid profile for black men versus white men, but a decreased risk for Black women and Whites (Knox et al. 1996).

Despite the benefits of education in certain race-gender combinations, it is important to note that in the current dataset, even at higher levels of education, large disparities in BP persist within both males and females. These data suggests that other factors, such as psychosocial risk factors associated with exposures to racism or discrimination, or unmeasured genetic factors, may still be important even after adjusting for differences in education.

Comprehensive Analysis of Genetic and Environmental Data

With the rise of new high-throughput sequencing platforms, genetic data have become increasingly available to test theories of genetic predisposition to complex disease in different racial groups. These genetic data are increasingly used to generate estimates of continental ancestry for each racial group, to be tested directly as a causal

factor in contributing to complex disease (Salari et al. 2005; Tsai et al. 2009). One such study by Tang et al. (2006) analyzed this same FBPP dataset but only incorporated measures of genetic ancestry and standard covariates of age, sex, and BMI (similar to Model B, Table 5-12). This previous study reported a similar trend of positive (but not significant) regression coefficients for the effect of genetic ancestry on SBP, DBP, and MAP within the black population. Ancestry was tested in the current study to compare its effect to education, and to determine if it may have a significant effect in an interaction with education. When genetic ancestry was added to the model with just education, it was not found to be significant as a main effect, or in an interaction with education, whereas education was a significant predictor of BP in models with and without ancestry included.

This result directly relates to many recently published and ongoing studies that claim genetic basis to disease often in the absence of genetic data from any functional locus or sometimes without any genetic data at all (Gravlee and Mulligan 2010). Studies that incorporate a continental measure of ancestry sometimes find genetic ancestry to be associated directly with disease. There is a serious problem of confounding, however, when these studies do not also include any measurements of environmental factors. Burchard et al., (2003) argue that in Latino populations, those of darker skin color are more likely to experience discrimination, receive less education, and hold less prestigious jobs. Thus, measures of socioeconomic status, acculturation, racial identity, etc., among other sociocultural variables, are likely interrelated with genetic ancestry. Thus any association between genetic ancestry and disease may be confounded by these often-unmeasured factors, which may be responsible for the ultimate cause of the

disease (Burchard et al. 2003). Without these measures, any conclusion of genetic causation is problematic. In our study, genetic ancestry seems to be moderately associated with education level (Figure 5-3). This association is likely due to differences in environmental exposures related to varying skin color: those of darker skin color are more likely exposed to harsher sociocultural environments that might be less conducive to achieving higher education and more likely to increase blood pressure (Gravlee and Dressler 2005). Thus it is important to adjust for relevant sociocultural factors, like education, in analysis where genetic ancestry is suspected as a causal factor. Even after adjusting for SES in many studies, researchers tend to point to the persistence of genetic associations as proof of underlying genetic causation (Non et al. 2010b). One must be cautious, however, that rigorous and appropriate measures of the sociocultural environment are included, to avoid the problem of residual confounding due to incomplete or poor measurements of SES or other important sociocultural variables (Kaufman et al. 1997).

Strengths and Limitations

In this study, education serves as the only available measure of the sociocultural environment. Admittedly, many other aspects of the sociocultural environment are also likely to affect BP, including social classification (Gravlee et al. 2009), perceived discrimination, or subjective social status (Adler 2009). Furthermore, education can be considered only a crude proxy of SES, when considering that higher education is not necessarily associated with higher income, especially in younger individuals, and that many other factors may contribute to SES including wealth, residential neighborhood environment, etc (LaVeist 2005). Thus it is particularly striking that even such a crude measure of the socioeconomic environment as education shows such a significant

association with SBP and at the same time attenuates the association between ancestry and BP. While education may not be the only important aspect of the sociocultural environment affecting health, it appears to be a useful predictor of SBP. Furthermore, these results may be the first to suggest that education, as one aspect of the sociocultural environment, is more important than genetic ancestry in explaining the large racial disparity in blood pressure in the total sample between Black and White Americans.

Conclusions/Future Directions

This study has confirmed the importance of education on BP in explaining the disparity in hypertension between Black and White Americans. One important remaining question is to determine which features of education are most responsible for increasing health. Though the FBPP dataset only provides information on education level, it is possible that the type of education may be important aspect that affects health. In fact, racial differences in average education quality or type of degree may contribute to the racial disparity in BP (e.g. Blacks and other minorities are still very underrepresented in the higher paying science, technology, engineering, and math (STEM) fields (Nelson 2009). Ross and Mirowsky (1999) investigated three aspects of education, including quantity (i.e. # of years), credential (i.e. college degree), and selectivity of schools, and discovered that years of education was the most significant factor in affecting perceived health and physical functioning, followed by selectivity, while credential had no effect at all. It would be interesting to pursue these questions further with datasets that have more detailed information about the type and quality of education. Additionally, ethnographic analyses are needed to explore in more detail how education confers health benefits differently in different sociocultural contexts.

In sum, the results of this study suggest that educational attainment has a protective effect against higher blood pressure, which may act differently in U.S. Blacks than in U.S. Whites. One unique finding is that the interaction between race and education support both the diminishing returns and minority poverty hypotheses, such that Blacks have higher BP both at the lower and higher ends of the socioeconomic spectrum. This pattern could help to reconcile some of the seemingly contradictory findings in other studies and could set out new directions for research on how race and education intersect. It is also important to note that education does not fully explain the health disparity between Black and White Americans. However, we must be cautious not to jump to a genetic explanation for the remaining disparity, without genetic data to support this claim, as dozens of other unmeasured environmental factors may also contribute. In this study, the analyses of blood pressure and genetic ancestry in the Black population, suggests that genetic ancestry, as a global continental measure, does not appear to directly contribute to higher BP in the Black population. Though future analyses may yet discover race-specific functional alleles, any conclusion of genetic causation will require proof of a specific gene effect that differs in frequency between racial groups. In the meantime, the reduction of health disparities may be better achieved through research and policy efforts to ensure greater access to education for minorities, and equal benefits of higher education across racial groups.

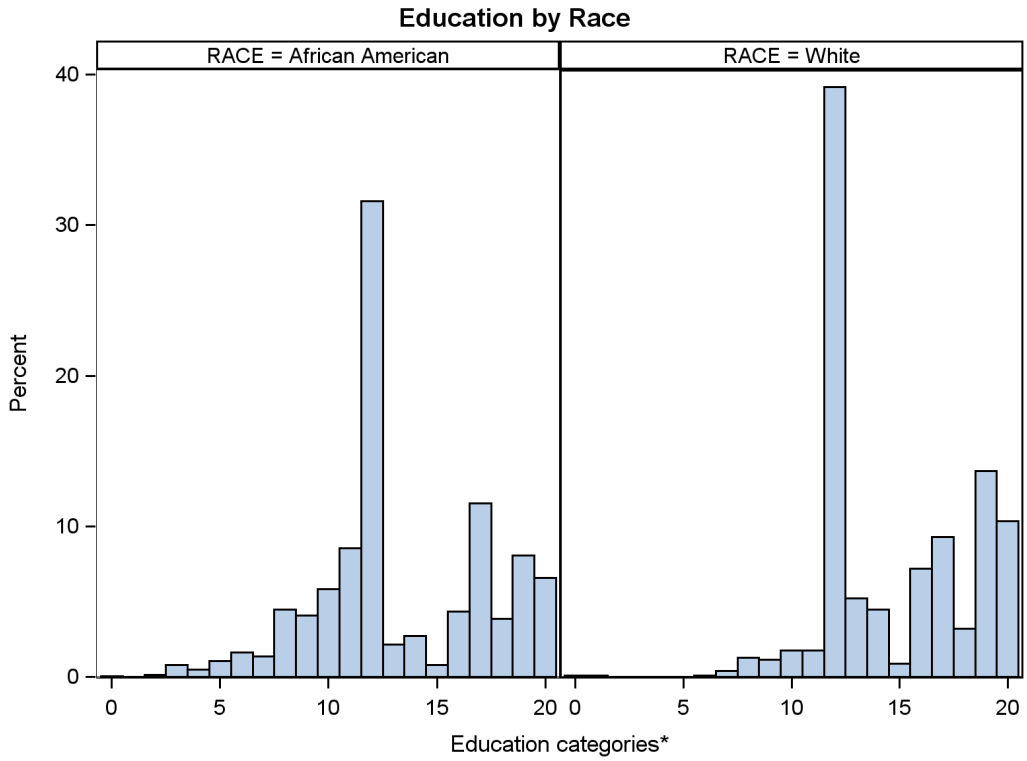


Figure 5-1. Distribution of education in Black and White populations. *Education is coded such that 0=no school, 1=1st grade, 2=2nd Grade, 3=3rd grade, 4=4th grade, 5=5th grade, 6=6th grade, 7=7th grade, 8=8th grade, 9=9th grade, 10=10th grade, 11=11th grade, 12=12th grade or GED, 13=1year Vo-tech, 14=2 years Vo-tech, 15=3 years Vo-tech, 16=1 year college, 17=2 years college, 18=3 years college, 19=4 years college, 20=1+years grad/professional school. Each level listed here represents the highest level of education achieved, as self-reported by participants.

Table 5-1. Chi Square test of HT status by education using UBW total sample.

Educ	HT	NT	Total
<=HS	1192	310	1502
>HS	924	373	1297
Total	2116	683	2799
X^2		24.8736	
p-value		<0.0001	

X2 tests indicate if the proportion of hypertensives and normotensives differs by education status. (Includes only individuals that don't have African ancestry and not missing any key phenotypes, n=2799).

Table 5-2. Comparison of SBP, DBP, and MAP in total sample across low and high education using t-tests.

BP	<=HS Degree	>HS Degree	Diff	t value (p-value)
N	1520	1310	--	--
Mean SBP (SD) (95% CI)	133.0 (22.77) (131.9, 134.2)	128.10 (20.12) (127.1, 129.2)	4.90	6.08† (<0.0001)*
Mean DBP (SD) (95% CI)	73.60 (11.92) (73.02, 74.22)	73.26 (11.28) (72.66, 73.88)	0.35	0.80† (0.4225)
Mean MAP (SD) (95% CI)	93.40 (14.16) (92.71, 94.14)	91.55 (12.91) (90.86, 92.26)	1.87	3.67† (0.0002)*

Using UBW dataset. *Indicates statistical significance at α=0.05.
† Satterthwaite method used as variances were not equal (equality of variance F test p=<0.05 in each case) n=2830 (less than Chi square sample size because not including those missing SBP, DBP, MAP data).

Table 5-3. Comparison of SBP and MAP in total sample by education level.

	n	Mean SBP +/-SE	Mean MAP+/-SE
Less than HS degree	535	137.31+/-1.04*	95.18+/-0.63*
HS Degree	983	130.72+/-0.69*	92.46+/-0.44*
Some vo-tech	221	130.16+/-1.40*	92.93+/-0.89
Some College	858	127.94+/-0.69*	91.42+/-0.45*
Some Grad school	231	126.93 +/-1.26*	90.75+/-0.78*
Total	2828	130.77+/-0.41	92.56+/-0.26

*ANOVA showed significant difference in SBP and in MAP between at least 2 categories. Tukey tests reveal sig differences in SBP between the category of less than HS degree with every other category of education, and sig differences in MAP between the category of less than HS degree with HS degree, some college, and some grad school.

Table 5-4. Chi Square tests of HT status by education at each field center and across all field centers of Black and White participants.

Black				White			
Field Center	N	$X^{2\dagger}$	p-value	Field Center	N	X^2	p-value
Maywood	222	4.43	0.0353*	Salt Lake City	212	1.98	0.159
Jackson	545	12.37	0.0004*	Framingham	238	0.86	0.353
Forsyth	275	7.31	0.0069*	Minneapolis	237	2.48	0.115
Birmingham	572	7.82	0.0052*	Rochester	404	0.61	0.436
--				Tecumseh	94††	3.89	0.049*
Total centers	1614	20.21	0.0001*	Total centers	1185	2.0947	0.1478

† X2 tests indicate if the proportion of hypertensives and normotensives differs by education status. (Includes all individuals, even those that don't have African ancestry, and not missing any key phenotypes, n=2799).
*Indicates statistical significance at α=0.05.
††44% of the data are missing, some cells with <5 counts; Fisher's exact test two sided p-value= 0.0887.

Table 5-5. Two sample tests of proportions comparing education status among HT and NT subjects.

Field Center	HT with low education (freq of HT)	NT with low education (freq of NT)	Z score	p-value
Blacks				
All centers	806/1283 (0.63)	163/331 (0.49)	4.50	0.0001*
Maywood	55/73 (0.75)	91/149 (0.61)	2.10	0.0353*
Jackson	356/509 (0.70)	15/36 (0.42)	3.52	0.0004*
Forsyth	163/276 (0.59)	24/61 (0.39)	2.80	0.0051*
Birmingham	269/488 (0.55)	33/85 (0.39)	2.78	0.0055*
Whites				
All centers	386/833 (0.4634)	147/352 (0.4176)	1.45	0.0739
Tecumseh	4/15 (0.2667)	43/79 (0.5443)	-1.97	0.9757
Rochester	172/349 (0.4928)	24/55 (0.4364)	0.78	0.218
Minneapolis	87/154 (0.5649)	38/83 (0.4578)	1.58	0.0576
Framingham	73/166 (0.4398)	27/72 (0.375)	0.93	0.1762
Salt Lake City	50/149 (0.3356)	15/63 (0.2381)	1.41	0.0798

*Indicates statistical significance at $\alpha=0.05$.

Table 5-6. Comparison of SBP, DBP, and MAP in Black and White individuals across low and high education using t-tests.

	Blacks				Whites			
	<=12th grade	>12th grade	Diff	t value (p-value)	<=12th grade	>12th grade	Diff	t value (p-value)
N	960	644	-	-	560	666	-	-
Mean SBP (SD) (95% CI)	136.2 (23.08) (134.8,137.7)	131.0 (20.17) (129.5, 132.6)	5.190	4.64 (<0.001)*	127.6 (21.13) (125.8,129.3)	125.3 (19.69) (123.8, 126.8)	2.24	1.91 (0.056)
Mean DBP (SD) (95% CI)	75.42 (12.26) (74.64,76.19)	74.98 (11.24) (74.11, 75.85)	0.436	0.73† (0.463)	70.54 (10.62) (69.66,71.42)	71.61 (11.09) (70.77, 72.46)	-1.08	-1.73 (0.085)
Mean MAP (SD) (95% CI)	95.69 (14.45) (94.78, 96.60)	93.67 (12.7452) (92.68, 94.66)	2.021	2.95† (0.003)*	89.55 (12.79) (88.49, 90.61)	89.52 (12.76) (88.55, 90.49)	0.028	0.04 (0.970)

*Indicates statistical significance at α=0.05. † Satterthwaite method used as variances were not equal (equality of variance F test p=<0.05 in each case). Includes different samples (n=1604 Blacks, n=1226 Whites total=2830) than chi square test because includes those listed as hypotensive or without HT status but not those missing sbp, dbp, map.

Table 5-7. Comparison of SBP and MAP in Blacks and Whites across all six levels of education.

	Blacks			Whites		
	n	Mean SBP +/-SE	Mean MAP+/-SE	n	Mean SBP +/-SE	Mean MAP +/- SE
less than HS degree	455	138.69 +/- 1.13*	96.15+/-0.68**	80	129.49+/- 2.46	89.66+/-1.47
HS Degree	504	133.98 +/- 0.97*	95.22+/-0.63	479	127.30+/- 0.96	89.57+/-0.58
Some vo-tech	91	134.22 +/- 2.53	96.30+/-1.52	130	127.32+/-1.54	90.58+/-1.04
Some College	447	130.91 +/- 0.93*	93.71+/-0.60	411	124.73+/-1.01	88.94 +/-0.64
Some Grad school	106	128.90+/- 1.83*	91.25+/-1.05**	125	125.25+/-1.72	90.33 +/-1.14
Total	1603	134.14 +/- 0.55*	94.86+/-0.34	1225	126.43+/-0.56	89.54+/-0.36

*ANOVA and Tukey tests show significant differences in SBP among those with less than HS degree versus those in all other categories except vo-tech school ($p<0.05$) in the Black population, but no significant differences between any categories for Whites. **For MAP, only those with less than High School Degree were significantly different from those with some grad school ($p<0.05$) in the Black population, but no significant differences seen in Whites. Slightly fewer samples (n=2828) than previous table because does not include those with no school at all (n=2).

Table 5-8. Multiple linear regression for BP in Blacks and Whites (RBW dataset) (using GEE with exchangeable correlation matrix, n=3170).

	SBP		DBP		MAP	
	β coeff +/- SE	p-value	β coeff +/- SE	p- value	β coeff +/- SE	P value
Intercept	99.45 (2.81)	<0.0001*	61.98 (1.72)	<0.0001*	74.52 (1.79)	<0.0001
SIRE						
Black vs. White	4.33 (0.65)	<0.0001*	2.72 (0.42)	<0.0001*	3.25 (0.46)	<0.0001*
Age	0.23 (0.03)	<0.0001*	0.12 (0.03)	<0.0001*	0.16 (0.02)	<0.0001*
Sex						
Female vs. Male	-18.56 (1.91)	<0.0001*	-3.25 (1.32)	0.0141*	-8.33 (1.39)	<0.0001*
age*sex	0.23 (0.04)	<0.0001*	-0.05 (0.03)	0.0517*	0.04 (0.03)	0.1874
BMI	0.80 (0.05)	<0.0001*	0.29 (0.03)	<0.0001*	0.46 (0.04)	<0.0001*
Education	-0.57 (0.09)	<0.0001*	-0.17 (0.06)	0.0020*	-0.31 (0.06)	<0.0001*
†QICu	3698	--	3698	--	3698	--

†Quasilikelihood under Independence Model Criterion (QICu) is a measure of goodness of fit analogous to AIC in likelihood-based methods. QICu adds a penalty to Q based o # of parameters. Smaller QIC is preferred. Field centers not included as covariates because field site was essentially synonymous with SIRE.

Table 5-9. Multiple linear regression coefficients for BP measures in RBW dataset (with education as 2 categories, using GEE).

	SBP		DBP		MAP	
Parameter	Beta (SE)	P value	Beta (SE)	P value	Beta (SE)	P value
Intercept	89.40 (2.02)	<0.0001*	59.15 (1.42)	<0.0001*	69.27 (1.47)	<0.0001*
AGE	0.25 (0.03)	<0.0001*	0.13 (0.02)	0.0001*	0.17 (0.02)	<0.0001*
sex						
Female versus Male	-18.13 (1.92)	<0.0001*	-3.09 (1.33)	0.0196*	-8.08 (1.39)	<0.0001*
AGE*sex						
Age* Female versus Male	0.22 (0.04)	<0.0001*	-0.06 (0.03)	0.0406*	0.04 (0.03)	0.2508
BMI	0.81 (0.05)	<0.0001*	0.29 (0.03)	<0.0001*	0.46 (0.04)	<0.0001*
EDUC						
<= HS degree versus > HS degree	1.48 (0.83)	<0.0768	-0.22 (0.51)	0.6629	0.34 (0.58)	0.5541
SIRE						
Black versus White	3.05 (0.83)	0.0002*	1.80 (0.54)	0.0008*	2.21 (0.59)	0.0002*
Educ*SIRE						
<= HS degree versus >HS degree * Black	3.20 (1.17)	0.0064*	2.11 (0.75)	0.0047*	2.48 (0.83)	0.0028*
†QICu	3698		3698		3698	

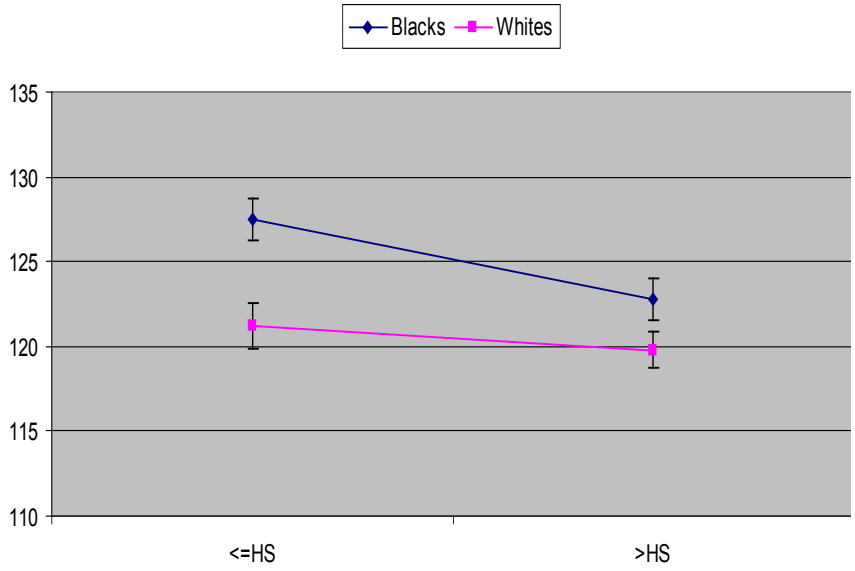
†Quasiliklihood under Independence Model Criterion (QICu) is a measure of goodness of fit analogous to AIC in likelihood-based methods. QICu adds a penalty to Q based o # of parameters. Smaller QIC is preferred.

Table 5-10. Multiple linear regression coefficients for BP in RBW dataset (with education as 6 categories, using GEE).

	SBP		DBP		MAP	
Parameter	Beta (SE)	P value	Beta (SE)	P value	Beta (SE)	P value
Intercept	86.44 (2.48)	<0.0001*	57.53 (1.70)	<0.0001*	67.19 (1.82)	<0.0001*
AGE	0.24 (0.03)	<0.0001*	0.14 (0.02)	0.0001*	0.17 (0.02)	<0.0001*
sex						
Female versus Male	-18.18 (1.92)	<0.0001*	-2.89 (1.33)	0.0300	-7.97 (1.40)	<0.0001*
AGE*sex						
Age* Female versus Male	0.22 (0.04)	<0.0001*	-0.06 (0.03)	0.0281*	0.03 (0.03)	0.2919
BMI	0.81 (0.05)	<0.0001*	0.28 (0.03)	<0.0001*	0.46 (0.04)	<0.0001*
EDUC						
Less than HS degree versus some Grad school	6.53 (2.31)	0.0047*	-0.76 (1.50)	0.6117	1.66 (1.66)	0.3155
HS Degree versus some Grad school	4.36 (1.53)	0.0045*	1.46 (0.97)	0.1323	2.43 (1.09)	0.0257*
Some Vo-tech versus some Grad School	5.74 (1.72)	0.0009*	3.54 (1.09)	0.0011*	4.30 (1.23)	0.0005*
Some College versus some Grad school	3.05 (1.51)	0.0430*	1.14 (0.94)	0.2227	1.77 (1.06)	0.0951
SIRE						
Black versus White	5.12 (2.18)	0.0188*	1.91 (1.31)	0.1454	2.95 (1.49)	0.0472*
Educ*SIRE						
Less than HS degree*Black vs White	0.17 (3.06)	0.9550	3.70 (1.91)	0.0525	2.55 (2.12)	0.2304
HS degree*Black versus White	0.68 (2.39)	0.7757	1.91 (1.47)	0.1947	1.54 (1.65)	0.3519
Some Vo-tech * Black versus White	-5.25 (3.26)	0.1071	-1.20 (1.97)	0.5418	-2.53 (2.27)	0.2647
Some College*Black versus White	-1.80 (2.34)	0.4427	0.31 (1.43)	0.8283	-0.36 (1.61)	0.8228
†QICu	3704		3704		3704	

†Quasilikelihood under Independence Model Criterion (QICu) is a measure of goodness of fit analogous to AIC in likelihood-based methods. QICu adds a penalty to Q based on # of parameters. Smaller QIC is preferred.

A



B

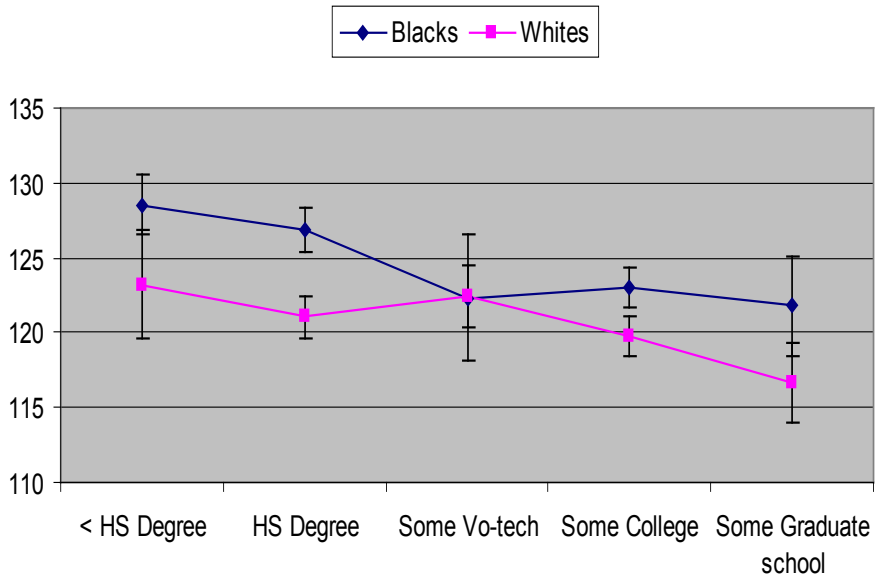


Figure 5-2. Interaction plots of SIRE and education. Interaction plots with 2 (panel A) and 6 (Panel B) categories of education, based on GEE analysis with exchangeable matrix in SAS.

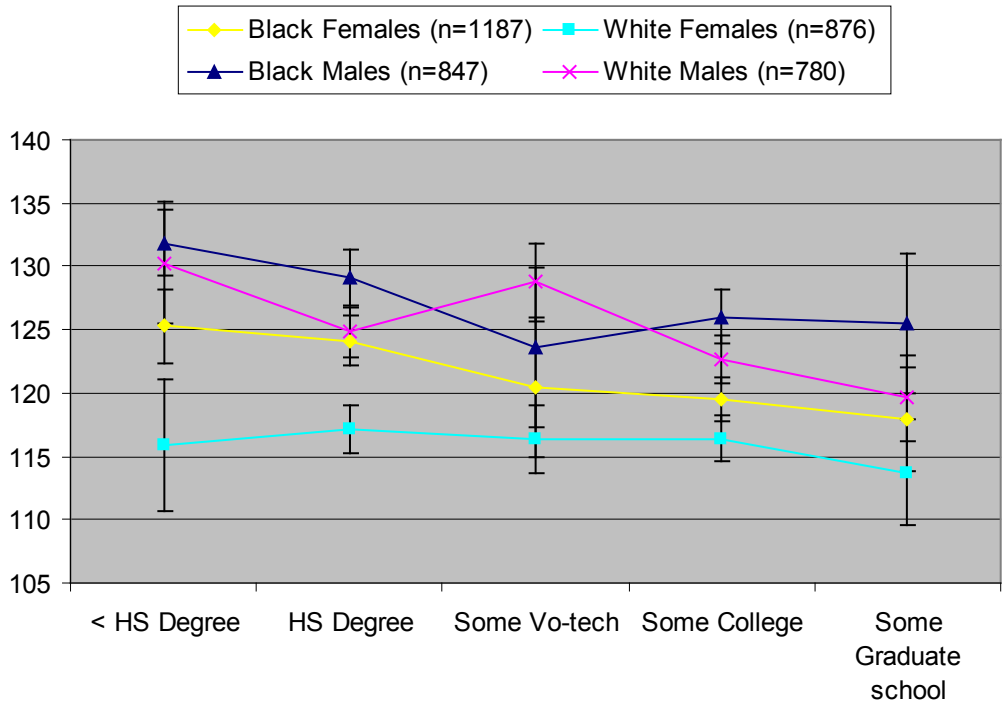


Figure 5-3. Interaction plot of SIRE and education on SBP within each gender. Interaction plot with 6 categories of education, separated by race and gender, based on GEE analysis with exchangeable matrix in SAS.

Table 5-11. Mean levels of African ancestry in unrelated Black individuals.

	Normotensive		Hypertensive		diff	p-value*
	N	Mean (SD) (95% CI)	N	Mean (SD) (95% CI)		
Maywood	135	0.799 (0.10) (0.782-0.818)	60	0.807 (0.11) (0.779-0.836)	0.007	0.654
Jackson	21	0.786 (0.12) (0.7312-0.8414)	262	0.793 (0.12) (0.778-0.809)	0.007	0.804
Forsyth	42	0.761 (0.13) (0.721-0.802)	150	0.785 (0.11) (0.766-0.803)	0.023	0.257
Birmingham	28	0.804 (0.09) (0.770-0.838)	379	0.821 (0.09) (0.812-0.830)	0.017	0.322
Total blacks	226	0.792 (0.11) (0.778-0.806)	851	0.805 (0.11) (0.798-0.812)	0.013	0.103

*p-value for t-tests of difference in levels of genetic ancestry between NT and HT at each individual site and at all sites combined. None are significant at alpha=0.05.

Table 5-12. Model fitting in linear regressions for SBP in RB dataset (n=1464) (using GEE with exchangeable matrix).

	Model A		Model B		Model C	
	β coeff +/- SE	p-value	β coeff +/- SE	p- value	β coeff +/- SE	P value
Intercept	85.14 (3.91)	<0.0001*	80.69 (5.29)	<0.0001*	99.98 (16.49)	<0.0001*
Age	0.67 (0.06)	<0.0001*	0.67 (0.06)	<0.0001*	0.68 (0.06)	<0.0001*
Sex						
Male versus Female	17.95 (3.05)	<0.0001*	17.94 (3.05)	<0.0001*	18.19 (3.04)	<0.0001*
age*sex	-0.27 (0.08)	0.0003*	-0.27 (0.08)	0.0003*	-0.28 (0.08)	0.0002
BMI	0.60 (0.07)	<0.0001*	0.60 (0.07)	<0.0001*	0.60 (0.07)	<0.0001*
Field Site						
Jackson vs. Maywood	-7.55 (1.62)	<0.0001*	-7.48 (1.62)	<0.0001*	-7.52 (1.62)	<0.0001*
Forsyth vs. Maywood	-3.55 (1.79)	0.0475*	-3.46 (1.80)	0.0541	-3.43 (1.79)	0.0558
Birmingham vs. Maywood	0.80 (1.29)	0.5319	0.69 (1.29)	0.5902	0.68 (1.29)	0.5954
Education	-0.53 (0.15)	0.0005*	-0.51 (0.15)	0.0008*	-1.87 (1.09)	0.0899
African ancestry			5.23 (4.55)	0.2507	-19.23 (20.44)	0.3469
Educ*Ancestry					1.69 (1.34)	0.2081
QICu	1473.0		1474.0		1475.0	

These models were tested on a larger sample size than the analyses in Table 5-6, which was limited only to individuals classified as hypertensive or normotensive. †Quasilikelihood under Independence Model Criterion (QICu) is a measure of goodness of fit analogous to AIC in likelihood-based methods. QICu adds a penalty to Q based on # of parameters. Smaller QIC is preferred.

Table 5-13. Model fitting in linear regressions for DBP in RB dataset (n=1464) (using GEE with exchangeable matrix).

	Model A		Model B		Model C	
	β coeff +/- SE	p-value	β coeff +/- SE	p- value	β coeff +/- SE	P value
Intercept	57.48+/-2.65	<0.0001	53.92 +/-3.57	<0.0001	64.78+/-9.70	<0.0001
Age	0.28+/-0.03	<0.0001	0.28 +/-0.03	<0.0001	0.28+/-0.03	0.0399
Sex						
Male versus Female	4.16+/-2.09	0.0465	4.15+/-2.09	0.0472	4.29+/-2.09	0.0399
age*sex	0.00+/-0.05	0.9591	0.00+/-0.05	0.9676	0.00+/-0.05	0.9054
BMI	0.26+/-0.06	<0.0001	0.26+/-0.06	<0.0001	0.26+/-0.06	<0.0001
Field Site						
Jackson vs. Maywood	-9.77+/-0.96	<0.0001	-9.74+/-0.97	<0.0001	-9.74 +/-0.96	<0.0001
Forsyth vs. Maywood	-6.28+/-1.18	<0.0001	-6.21+/-1.18	<0.0001	-6.19+/-1.18	<0.0001
Birmingham vs. Maywood	-0.99+/-0.86	0.2500	-1.08+/-0.86	0.2098	-1.09+/-0.86	0.2056
Education	-0.14+/-0.09	0.1190	-0.13+/-0.09	0.1693	-0.88+/-0.60	0.1417
African ancestry			4.18+/- 2.81	0.1369	-9.59+/-11.76	0.4149
Educ*Ancestry					0.95+/-0.75	0.2064
QICu	1473.0		1474.0		1475.0	

†Quasilikelihood under Independence Model Criterion (QICu) is a measure of goodness of fit analogous to AIC in likelihood-based methods. QICu adds a penalty to Q based on # of parameters. Smaller QIC is preferred.

Table 5-14. Model fitting in linear regressions for MAP in RB dataset (n=1464) (using GEE with exchangeable matrix)

	Model A		Model B		Model C	
	β coeff +/- SE	p-value	β coeff +/- SE	p- value	β coeff +/- SE	P value
Intercept	66.82+/-2.82	<0.0001	62.95+/-3.79	<0.0001	76.78 +/- 11.08	<0.0001
Age	0.41+/-0.04	<0.0001	0.41+/-0.04	<0.0001	0.41 +/- 0.04	<0.0001
Sex						
Male versus Female	8.72+/-2.22	<0.0001	8.71+/-2.22	<0.0001	8.89+/-2.22	<0.0001
age*sex	-0.09+/-0.05	0.0787	-0.09+/-0.05	0.0804	-0.10 +/- 0.05	0.0654
BMI	0.37+/-0.06	<0.0001	0.37+/-0.06	<0.0001	0.37 +/-0.06	<0.0001
Field Site						
Jackson vs. Maywood	-9.01+/-1.08	<0.0001	-8.94+/-1.09	<0.0001	-8.97+/-1.09	<0.0001
Forsyth vs. Maywood	-5.33+/-1.28	<0.0001	-5.25+/-1.29	<0.0001	-5.24+/-1.29	<0.0001
Birmingham vs. Maywood	-0.40+/-0.93	0.6653	-0.50+/-0.93	0.5935	-0.51 +/-0.93	0.5850
Education	-0.27+/-0.10	0.0076	-0.26+/-0.10	0.0126*	-1.22 +/-0.71	0.0838
African ancestry			4.54+/-3.03	0.1348	-12.99+/- 13.60	0.3395
Educ*Ancestry					1.21+/-0.88	0.1670
QICu	1473.0		1474.0		1475.0	

†Quasilikelihood under Independence Model Criterion (QICu) is a measure of goodness of fit analogous to AIC in likelihood-based methods. QICu adds a penalty to Q based on # of parameters. Smaller QIC is preferred.

Education Level	Description
2	2nd-11th grade
3	HS graduate
4	1+ years Vo-tech
5	1+ years college
6	1+ years graduate school

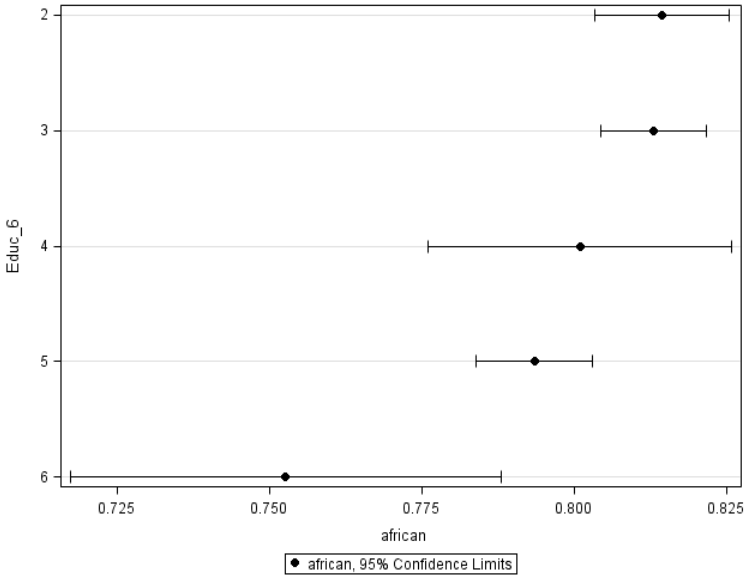


Figure 5-4. Relationship between individual level of African genetic ancestry and 6 education levels.The accompanying table shows the coding of education levels into 6 categories, and the chart shows how African genetic ancestry is distributed across these 6 education levels.

CHAPTER 6
CONCLUSION

The current global distribution of human genetic diversity is a product of evolutionary forces within a context of changing environments. This evolution occurs through random processes, such as genetic drift and mutation, along with more directed processes, such as migration and natural selection, which have allowed humans to enter and adapt to new environments. To understand the complexity of human genetic diversity, therefore, requires an exploration of neutral markers to trace historical scenarios, and non-neutral markers to understand how adaptation (or mal-adaptation) to new environments has altered the human genome. In this dissertation, I have demonstrated how the multidisciplinary nature of anthropology facilitates a comprehensive exploration of human genetic diversity. Throughout each chapter, I have integrated data, analytical techniques, and perspectives from diverse disciplines both within and outside of anthropology, including archaeology, linguistics, geography, history, medical and cultural anthropology, and public health. In each of the evolutionary history projects, I specifically examined human genetic diversity within a context of historical, linguistic, and archaeological data to choose between competing historical scenarios, while for each of the two disease projects I have integrated sociocultural data directly into genetic analyses to formally test competing genetic and sociocultural explanations for health disparities.

Across the fields of anthropology, public health, medicine, and many others, researchers are expanding the scope of their research to include multiple levels of analyses in order to resolve complex research questions. This requirement for more integrative research has led to higher degrees of collaboration across disciplines. For

example, Gehlert et al. (2010) just recently noted that traditional research paradigms have failed to identify and eliminate the causes behind persistent health disparities, and that new transdisciplinary collaborations will be needed to address all required levels of analyses, “from the molecular to the societal.” At the same time, new technological advances across disciplines, e.g. genetics, GIS data collection, and computational power, along with theoretical advances e.g. new statistical analyses tools and simulation techniques, are providing new opportunities to explore research questions with greater sophistication, and to develop new questions that may have been inconceivable only a decade ago. Through my dissertation projects, I have demonstrated that the exploration of anthropological questions can greatly benefit from the use of this collaborative interdisciplinary environment and the new range of available data. The incorporation of diverse data types and multidisciplinary perspectives in this dissertation has led to a number of new insights about human history and human disease, including the interrelatedness of these two topics. The results of these studies have sparked ideas for new avenues of research and new types of data that may be required to accurately reconstruct evolutionary history, and to advance our understanding of the ultimate causal factors behind health disparities.

With regard to evolutionary history of Jewish populations (Chapter 2), the genetic data in my dissertation helped to narrow the range of possible historical scenarios, by providing a timeframe within which the Jewish populations were founded, and a quantitative measure of the relatedness between each Jewish population with its neighboring Jewish and non-Jewish populations. For example, the simulation-based date for the origin of the R0a1 haplotype in the Ethiopian Jewish population, along with

the lack of shared maternal genetic ancestry with other Jewish populations, allowed elimination of the first historical scenario of direct descent of Ethiopian Jews from ancient (female) Israeli exiles ~2500 years before present (YBP). However, the genetic data were consistent with the two conversion theories, which were both estimated within 1000 years of each other. Linguistic and historical data were then considered as supporting evidence to narrow the choice between the two remaining scenarios of conversion during the 4th or 14th Centuries. These multiple lines of evidence implicated descent from an Ethiopian population that adopted Judaism prior to the 14th Century. Even after including linguistic and historic data, however, the exact origin of the Jewish population remains uncertain.

Future studies of the Ethiopian Jewish population could reach to integrate even more interdisciplinary data, such as archaeological evidence of Jewish artifacts to document the earliest material record of a Jewish presence in Ethiopia. Increased availability of full genome sequencing would also allow for more high-throughput genetic data from Jewish individuals that may provide insights beyond what can be gained by monoallelic markers, i.e. mitochondrial DNA (mtDNA). Continuing interest in the origin of Jewish populations and in the use of more advanced genetic techniques to address historical questions is attested to by the continual publication of studies exploring the genetic history of Jewish populations with ever more advanced genetic techniques (Atzmon et al. 2010; Behar et al. 2010). The first genome-wide analysis of Jewish groups was just recently published on seven groups of Jewish populations and comparative non-Jewish groups throughout Eurasia (Atzmon et al. 2010). With the use of nuclear SNP arrays, they discovered two major clusters of Jewish groups – Middle

Eastern and European/Syrian, which all share common Middle Eastern ancestry and varying degrees of admixture from Europeans and Africans. These findings are similar to those in my mtDNA-based analysis of Yemenite Jews, though they did not include Ethiopian Jews in their sample. The second genome-wide study assayed >225,000 SNPs in 121 samples from 14 Jewish communities, (including Ethiopian Jews) and 1166 individuals from 69 comparative non-Jewish populations (Behar et al. 2010). Again their results largely mirror ours for the Yemeni Jews (i.e. similarity to other Jewish populations and not local non-Jewish Yemeni, and for the Ethiopian Jews (i.e. similarity to local non-Jewish Ethiopians but not other Jewish populations). Though the question of ancestry in Jewish Diaspora populatoins has now largely been answered, perhaps future study of these large genome-wide datasets in Jewish populations can shed light on Jewish-specific diseases that arose due to their isolated histories.

In the second evolutionary history project (Chapter 3), mtDNA was analyzed within a geographic framework to test for geographic structuring of the genetic data, e.g clines or clusters of mtDNA frequencies. By taking a wide geographic view of countries surrounding Yemen, I was able to demonstrate that the complexity and diversity of Yemen's mtDNA is a reflection of its central position at the crossroads of great migration and trading routes across the Red Sea from Africa, up through the Arabian Peninsula to the rest of the Middle East, and across the Indian Ocean towards Asia, over at least the past seven millennia. Detailed geographic data of this sort (e.g. latitude and longitude coordinates) are rarely incorporated directly into genetic analyses of human populations, despite the high mobility of human populations. By integrating data on the exact location of each individual's maternal grandmother's birthplace, I was able to test

for geographic patterning in the maternal history of Yemen, and to demonstrate the utility of these types of analyses for other researchers with similar questions. I have also integrated archaeological data when considering the support for different dispersal routes out of Africa – though the insights from archaeology were limited as very few artifacts have been successfully classified or dated in this region of the world.

Future studies could take advantage of the new availability of full genome data to achieve more fine-scale resolution of the timing of historic migration events. One remaining question in my Yemen study is to establish the timeframe of the migration of individuals of African origin into Yemen. I have begun preliminary analyses of genome-wide data from >370,000 SNPs in a subset of 64 Yemeni individuals which may enable a more resolved history of this very complex and admixed region of the world. These larger genome-level datasets could potentially provide a more accurate reconstruction of the history of a population than previous analyses of an individual locus, as multiple loci can test for consistent genetic patterns throughout regions of the genome, while an individual locus is subject to the idiosyncrasies of random drift or selection. Specifically, I plan to analyze the full genome data through principal component analyses to capture the primary dimensions of variability within the data, which can then be placed directly on a geographic map to determine the geographic structure of the entire genome. Similar work has recently been conducted by Novembre et al. (2008) and Biswas et al. (2009), among others, who have discovered that genomic data of this sort can detect very finely resolved differences between closely related groups. Additionally, the interdisciplinary perspective I have gained from this study has led to new ideas about incorporating other types of geographic data into evolutionary history projects, such as

the location of potential geographic barriers to population movements (deserts or mountain ranges), or altitude and rainfall information into GIS map analyses in Yemen or other places. Finally, by reaching into the field of public health, I have gained new ideas for the use of geographic information in future epidemiological studies, such as determining the geographic distribution and expansion of disease vectors (e.g. malaria-bearing mosquitoes), or even analyses of the geography of chronic diseases, such as obesity or hypertension, which may be socially transmitted throughout geographic space in a manner similar to infectious diseases (Christakis 2004; Cohen-Cole and Fletcher 2008).

With regards to my research into health disparities in Puerto Rico (Chapter 4), the integration of multidisciplinary data permitted two unique analyses: a) the determination of relative contributions of genetic and environmental factors to blood pressure (BP) variation, and b) the stratification of data along sociocultural lines to reveal genetic effects that were otherwise undetectable. In the next steps of this research project, functional assays could be conducted to further investigate the genetic variant I identified and determine if and why it may have a protective effect against the development of high blood pressure. Also, other candidate genes could be explored in this sample, or others, using the same integrative manner to identify other important genes, and to confirm the value of the sociocultural information to reveal genetic effects. Finally, further ethnographic analyses would be useful to determine exactly why darker-skinned individuals at higher SES have higher blood pressure in Puerto Rico. The hypothesis of increased chronic stress due to more pervasive racism in higher

socioeconomic environments may be one of many explanations for the higher blood pressure in darker skinned Puerto Ricans at higher SES.

In my second health disparities project (Chapter 5), I have built upon lessons learned in the Puerto Rican study to test genetic and environmental data together in a large and geographically widespread epidemiological dataset of Black and White Americans. I determined that the environmental factor of education explained much of the disparity in BP between racial groups, and also explained more of the variation in BP in the Black population than genetic ancestry. A future direction for this study might be to explore different aspects of education in other datasets where more detailed data on education can be collected, such as type of education, location, or credentials, to determine which factor might be driving the impact of education on health. Ethnographic or other types of qualitative analyses would also be useful to determine exactly why education appears to have "diminishing returns" in the Black population, and why education might affect health differently for men and women. If other large scale datasets are available with more associated sociocultural information, other factors could also be explored at a similarly large epidemiological scale, such as racism/discrimination, neighborhood environment, and poverty.

The interdisciplinary nature of my health disparities studies, along with the geographic elements in my evolutionary studies, has inspired new ideas about the types of factors that might contribute to health disparities in different regions of the country. It would be interesting to explore the importance of geographic location in the FBPP dataset in affecting both genetic ancestry and environmental factors that impact health, e.g. ethnographic and historical research could explore how the different history of racial

relations in each city contributes to different cultural contexts that affect the level of admixture in each region, the current sociocultural environment, and their independent (and/or interacting) impacts on health. Although these more complex factors are not as easily measured as years of education, I believe through collaborative efforts with anthropologists, sociologists, and epidemiologists, among others, it is possible to raise the level of sophistication of environmental/sociocultural variables, even in large-scale epidemiological studies.

These types of integrative and collaborative analyses are critical to resolve the persisting problem of health disparities. In public health, this perspective, termed the "ecological model," has gained tremendous favor in recent years, due to the growing awareness of the complexity of all the interrelated factors that affect health. The ecological model argues for multiple levels of analyses, including individual (e.g. genetic, biological, psychological), interpersonal (relationships, social support), community (resources, neighborhoods, structural violence), societal (social policies, institutional racism), and national level (laws and policies) that all act together to affect health (Sallis and Owen 2002). In my projects on health disparities, I have tried to take a few small steps towards advancing this paradigm. Ultimately it will take a great collaborative effort, in which anthropology is well positioned to play a key central role.

Throughout my dissertation projects I have integrated multidisciplinary data and utilized multiple marker systems to address anthropological genetic questions that have relevance beyond anthropology to public health, medicine, archaeology, history, linguistics, and geography. This interdisciplinary perspective – which values the integration of genetic and environmental data – has led me to a new research question

relating to health disparities. In the next stage of my research, I have developed an interest in the biological consequences of different social environments during specific life stages of development. I believe that stressors in the fetal environment can increase the risk of adverse birth outcomes and also lifetime disease risk, as predicted by the developmental origins of health and disease theory (Barker 2004). Specifically, I am interested to explore how maternal stress during the critical windows of fetal development, and early life stages, impacts the fetus through epigenetic changes, i.e. changes in gene expression due to causes other than changes in underlying DNA sequence. I believe epigenetic mechanisms may provide a link between early environmental exposures and gene expression, by explaining how the effects of a mother's health or environment may be transmitted to her child in a heritable manner that is independent of her genetic code. In building upon the insights gained from the multidisciplinary nature of my dissertation, I also plan to integrate psycho-socio-behavioral data into these future genetic and epigenetic analyses to determine exactly which aspects of the social environment have the greatest impact on the health of the next generation.

Ten years following the completion of the first draft of the human genome sequence, it is clear that we still have much to learn about the genes behind complex diseases. Despite tremendous efforts, gene association studies have only identified a small proportion of heritable variation responsible for complex diseases (McCarthy et al. 2008). The interdisciplinary approach utilized in my dissertation has great potential for generating new directions to examine disease and health. This approach also has implications for the field of anthropology, and academia as a whole. The interdisciplinary

perspective within each project has demonstrated the benefits of integrating a wide range of data and analytical techniques to permit a comprehensive analysis of complex anthropological questions. This type of integrative research requires the willingness to expand beyond traditional disciplinary boundaries and tackle a new set of challenges related to cross-field collaborations. Though it may seem daunting to engage in dialogue with (sometimes) unfamiliar vocabulary from distant disciplines, and to adopt new analytical techniques, interdisciplinary research has the potential to lead to exponential gains in research progress.

APPENDIX
QUESTIONING THE IMPORTANCE OF GENETIC ANCESTRY AS A CONTRIBUTOR TO PRETERM DELIVERY AND RELATED TRAITS IN AFRICAN AMERICAN WOMEN[3]

To the Editors: We read with interest Tsai and colleagues' report of a significant association between African genetic ancestry in African Americans and preterm delivery (PTD) (Tsai et al. 2009). They conclude that "more intensive investigation of genetic admixture in African Americans" is needed to identify novel PTD-related susceptibility genes. Their conclusions, however, rely on inconsistent trends in ancestry levels, use potentially biased estimates of ancestry, and neglect important alternative risk factors for PTD.

First, their conclusion that subjects with higher levels of African ancestry have a higher risk of PTD is not supported by the inconsistent trends they report across the eight PTD traits. For some traits, the 2nd and 3rd quartiles of African ancestry are significantly associated, while for others the 4th quartile is significant. It is difficult to imagine what sort of biological function related to PTD could be associated with intermediate levels, but not high levels, of African ancestry. Furthermore, the marginally significant associations, with P-values ranging from 0.007 to 0.04, are likely to disappear after correction for multiple testing across the PTD-related traits.

Second, accurate estimation of ancestry is highly dependent upon choosing a sufficient number of markers and accurate reference populations (Barnholtz-Sloan et al. 2008). Tsai et al. rely on a small set of 57 markers and comparative populations of

[3] Non, AL, Gravlee, CC, Mulligan CJ. 2009. Questioning the importance of genetic ancestry as a contributor to preterm delivery and related traits in African Americans, Letter to the Editor, *American Journal of Obstetrics and Gynecology*.

Yorubans from Nigeria and Mormons from Utah, which are unlikely to represent all West African and European diversity. They report similar ancestry estimates when using fewer AIMs, but this approach does not test the need for more markers or alternative reference populations, which could potentially change their estimates and alter the association with PTD.

Third, Tsai et al. dismiss sociodemographic and environmental factors as unable to explain the racial disparity in PTD. Although they measure some environmental and behavioral variables (e.g. education, drug use), they do not test for significant differences between cases and controls, despite finding higher rates of smoking and illicit drug use in PTD cases. Furthermore, previous research has shown strong associations between gestational outcomes and sociocultural factors such as poverty, social support, residential segregation, and discrimination (Lu and Halfon 2003). Including these sociocultural factors in the Tsai study may cause the modest association between ancestry and PTD to disappear, as shown recently in a study of genetic ancestry and blood pressure in Puerto Rico (Gravlee et al. 2009).

In the absence of specific evidence for race-specific susceptibility genes, it may be more productive to investigate potentially modifiable sociocultural and behavioral risk factors that clinicians and policymakers can address to reduce the risk for PTD.

LIST OF REFERENCES

2001. Genes, drugs and race. Nat Genet 29:239-240 | doi:10.1038/ng1101-239. Published online November 2001.

2008. A place for everything. Nature 453:2-2 | doi:10.1038/453002a. Published online 30 April 2008.

Abu-Amero KK, Gonzalez AM, Larruga JM, Bosley TM, and Cabrera VM. 2007. Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BMC Evol Biol 7:32.

Abu-Amero KK, Larruga JM, Cabrera VM, and Gonzalez AM. 2008. Mitochondrial DNA structure in the Arabian Peninsula. BMC Evol Biol 8:45.

Acevedo-Garcia D, Soobader MJ, and Berkman LF. 2007. Low birthweight among US Hispanic/Latino subgroups: the effect of maternal foreign-born status and education. Soc Sci Med 65:2503-2516.

Adler NE. 2009. Health disparities through a psychological lens. Am Psychol 64:663-673.

Adler NE, and Snibbe AC. 2003. The role of psychosocial processes in explaining the gradient between socioeconomic status and health. Current Directions in Psychological Science 12:119-123.

Ahroni R. 1986. Yemenite Jewry : origins, culture, and literature. Bloomington: Indiana University Press.

Al-Zahery N, Semino O, Benuzzi G, Magri C, Passarino G, Torroni A, and Santachiara-Benerecetti AS. 2003. Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. Mol Phylogenet Evol 28:458-472.

Albano JD, Ward E, Jemal A, Anderson R, Cokkinides VE, Murray T, Henley J, Liff J, and Thun MJ. 2007. Cancer mortality in the United States by education level and race. J Natl Cancer Inst 99:1384-1394.

Amirkhanov H. 1996. Bilinear cultural parallelism in the Arabian Early Neolithic. In: Afanas'ev G, Cleuziou S, Lukacs R, and Tosi M, editors. The prehistory of Asia and Oceania. Forli: Abaco. p 135-140.

Amos W, and Manica A. 2006. Global genetic positioning: Evidence for early human population centers in coastal habitats. Proceedings of the National Academy of Sciences of the United States of America 103:820-824.

Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E, and Ostrer H. 2010. Abraham's Children in the Genome Era:

Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. The American Journal of Human Genetics 86:850-859.

Balaresque PL, Ballereau SJ, and Jobling MA. 2007. Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet 16 Spec No. 2:R134-139.

Bandelt HJ, Forster P, and Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37-48.

Bar-Yosef O, and Belfer-Cohen A. 2001. From Africa to Eurasia - early dispersals. Quaternary International 75:19-28.

Barbujani G. 2000. Geographic patterns: how to identify them and why. Hum Biol 72:133-153.

Barker DJ. 2004. The developmental origins of adult disease. J Am Coll Nutr 23:588S-595S.

Barnholtz-Sloan JS, McEvoy B, Shriver MD, and Rebbeck TR. 2008. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. Cancer Epidemiol Biomarkers Prev 17:471-477.

Behar DM, Metspalu E, Kivisild T, Achilli A, Hadid Y, Tzur S, Pereira L, Amorim A, Quintana-Murci L, Majamaa K, Herrnstadt C, Howell N, Balanovsky O, Kutuev I, Pshenichnov A, Gurwitz D, Bonne-Tamir B, Torroni A, Villems R, and Skorecki K. 2006. The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. Am J Hum Genet 78:487-497.

Behar DM, Metspalu E, Kivisild T, Rosset S, Tzur S, Hadid Y, Yudkovsky G, Rosengarten D, Pereira L, Amorim A, Kutuev I, Gurwitz D, Bonne-Tamir B, Villems R, and Skorecki K. 2008a. Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. PLoS ONE 3:e2062.

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, and Rosset S. 2008b. The dawn of human matrilineal diversity. Am J Hum Genet 82:1130-1140.

Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G, Khusnutdinova EK, Balanovsky O, Semino O, Pereira L, Comas D, Gurwitz D, Bonne-Tamir B, Parfitt T, Hammer MF, Skorecki K, and Villems R. 2010. The genome-wide structure of the Jewish people. Nature advance online publication.

Belle EM, Ramakrishnan U, Mountain JL, and Barbujani G. 2006. Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans. Proc Natl Acad Sci U S A 103:8012-8017.

Ben-Zvi I. 1957. The exiled and the redeemed. Philadelphia,: Jewish Publication Society of America.

Biswas S, Scheinfeldt LB, and Akey JM. 2009. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. Am J Hum Genet 84:641-650.

Blady K. 2000. Jewish communities in exotic places. Northvale, NJ: Jason Aronson.

Bonilla C, Shriver MD, Parra EJ, Jones A, and Fernandez JR. 2004. Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. Hum Genet 115:57-68.

Bowles S, and Gintis H. 1976. Schooling in capitalist America : educational reform and the contradictions of economic life. New York: Basic Books.

Brondolo E, Rieppi R, Kelly KP, and Gerin W. 2003. Perceived racism and blood pressure: a review of the literature and conceptual and methodological critique. Ann Behav Med 25:55-65.

Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D, and Risch N. 2003. The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 348:1170-1175.

Cabrera C, Helgesson O, Wedel H, Bjorkelund C, Bengtsson C, and Lissner L. 2001. Socioeconomic status and mortality in Swedish women: opposing trends for cardiovascular disease and cancer. Epidemiology 12:532-536.

Cabrera VM, Abu-Amero KK, Larruga JM, and Gonzalez AM. 2009. The Arabian peninsula: Gate for human migrations out of Africa or Cul-de-Sac? A mitochondrial DNA phylogeographic perspective. In: Rose JI, and Petraglia MD, editors. The Evolution of Human Populations in Arabia. New York City: Springer Science+Business Media. p 79-87.

Cadenas AM, Zhivotovsky LA, Cavalli-Sforza LL, Underhill PA, and Herrera RJ. 2008. Y-chromosome diversity characterizes the Gulf of Oman. Eur J Hum Genet 16:374-386.

Case A, and Paxson C. 2002. Parental behavior and child health. Health Aff (Millwood) 21:164-178.

Cerny V, Mulligan CJ, Fernandes V, Silva N, Alshamali F, Non AL, Harich N, Cherni L, El Gaaied ABA, Al-Meeri A, and Pereira L. 2010. Internal diversification of mitochondrial haplogroup R0a reveals post-Last Glacial Maximum demographic expansions in South Arabia. Mol Biol Evol (in review).

Cerny V, Mulligan CJ, Ridl J, Zaloudkova M, Edens CM, Hajek M, and Pereira L. 2008. Regional differences in the distribution of the sub-Saharan, West Eurasian, and South Asian mtDNA lineages in Yemen. Am J Phys Anthropol 136:128-137.

Cerny V, Pereira L, Kujanova M, Vasikova A, Hajek M, Morris M, and Mulligan CJ. 2009. Out of Arabia-the settlement of island Soqotra as revealed by mitochondrial and Y chromosome genetic diversity. Am J Phys Anthropol 138:439-447.

Chinnery PF, Elliott HR, Syed A, and Rothwell PM. 2010. Mitochondrial DNA haplogroups and risk of transient ischaemic attack and ischaemic stroke: a genetic association study. Lancet Neurol 9:498-503.

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N, Matallana H, Avila PC, Casal J, Torres A, Nazario S, Castro R, Battle NC, Perez-Stable EJ, Kwok PY, Sheppard D, Shriver MD, Rodriguez-Cintron W, Risch N, Ziv E, and Burchard EG. 2006. Population stratification confounds genetic association studies among Latinos. Hum Genet 118:652-664.

Christakis NA. 2004. Social networks and collateral health effects. Bmj 329:184-185.

Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, and Underhill PA. 2004. Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet 114:127-148.

Cohen-Cole E, and Fletcher JM. 2008. Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. J Health Econ 27:1382-1387.

Cooper R, Rotimi C, Ataman S, McGee D, Osotimehin B, Kadiri S, Muna W, Kingue S, Fraser H, Forrester T, Bennett F, and Wilks R. 1997. The prevalence of hypertension in seven populations of west African origin. Am J Public Health 87:160-168.

Cooper RS, Kaufman JS, and Ward R. 2003. Race and genomics. N Engl J Med 348:1166-1170.

Cooper RS, Wolf-Maier K, Luke A, Adeyemo A, Banegas JR, Forrester T, Giampaoli S, Joffres M, Kastarinen M, Primatesta P, Stegmayr B, and Thamm M. 2005. An international comparative study of blood pressure in populations of European vs. African descent. BMC Med 3:2.

Cooper RS, and Zhu X. 2001. Racial differences and the genetics of hypertension. Curr Hypertens Rep 3:19-24.

Crassard R. 2009. The middle Paleolithic of Arabia: The view from the Hadramawt Region, Yemen. In: Petraglia MD, and Rose JI, editors. The evolution of human populations in Arabia : paleoenvironments, prehistory and genetics. Dordrecht ; New York: Springer. p 151-168.

Crimmins EM, Kim JK, Alley DE, Karlamangla A, and Seeman T. 2007. Hispanic paradox in biological risk profiles. Am J Public Health 97:1305-1310.

Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, Moral P, Watson E, Guida V, Colomb EB, Zaharova B, Lavinha J, Vona G, Aman R, Cali F, Akar N, Richards M, Torroni A, Novelletto A, and Scozzari R. 2004. Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am J Hum Genet 74:1014-1022.

de Moulins D, Phillips C, and Durrani N. 2003. The archaeobotanical record of Yemen and the question of Afro-Asian contacts. In: Neumann K, Butler A, and Kahlheber S, editors. Food, fuel and fields, progress in African archaeobotany. Koln: Heinrich-Barth-Institut.

Deo RC, Patterson N, Tandon A, McDonald GJ, Haiman CA, Ardlie K, Henderson BE, Henderson SO, and Reich D. 2007. A high-density admixture scan in 1,670 African Americans with hypertension. PLoS Genet 3:e196.

Dresch P. 1989. Tribes, Government, and History in Yemen. Oxford: Clarendon Press.

Dressler WW. 1991. Social class, skin color, and arterial blood pressure in two societies. Ethn Dis 1:60-77.

Dressler WW, and Santos JE. 2000. Social and cultural dimensions of hypertension in Brazil: a review. Cad Saude Publica 16:303-315.

Eaaswarkhanth M, Haque I, Ravesh Z, Romero IG, Meganathan PR, Dubey B, Khan FA, Chaubey G, Kivisild T, Tyler-Smith C, Singh L, and Thangaraj K. 2010. Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. Eur J Hum Genet 18:354-363.

Edens C. 2005. Exploring early agriculture in the highlands of Yemen. In: Sholan A, Antonini S, and Arbach M, editors. Sabean studies, archaeological, epigraphical, and historical studies in honour of Yusuf M Abdallah, Alessandro de Maigret and Christain J Robin on the occasion of their 60th birthdays. Naples: Torcoliere. p 185-211.

Endicott P, Ho SY, Metspalu M, and Stringer C. 2009. Evaluating the mitochondrial timescale of human evolution. Trends Ecol Evol 24:515-521.

Fagan BM. 1990. The journey from Eden : the peopling of our world. New York, N.Y.: Thames and Hudson.

Falush D, Stephens M, and Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Farley R. 1984. Blacks and whites : narrowing the gap? Cambridge, Mass.: Harvard University Press.

Farmer MM, and Ferraro KF. 2005. Are racial disparities in health conditional on socioeconomic status? Soc Sci Med 60:191-204.

Fattovich R. 1997a. The Near East and eastern Africa: their interaction. In: Vogel J, editor. Encylopedia of Precolonial Africa. Walnut Creek: AltaMira Press. p 479-484.

Fattovich R. 1997b. The Near East and eastern Africa: their interaction. . In: JO V, editor. Encylopedia of Precolonial Africa. Walnut Creek: AltaMira Press. p 479-484.

FBPP Investigators. 2002. Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). Hypertension 39:3-9.

Field D. 2008. Working together to put molecules on the map. Nature 453:978-978.

Field JS, and Lahr MM. 2005. Assessment of the southern dispersal: GIS-based analyses of potential routes at oxygen isotopic stage. Journal of World Prehistory 19:1-45.

Forster P. 2004. Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. Philos Trans R Soc Lond B Biol Sci 359:255-264; discussion 264.

Forster P, and Matsumura S. 2005. Evolution - Did early humans go north or south? Science 308:965-966.

Francaviglia V. 1995. Il existait déja au Neolithique un commerce d'obsidienne a travers la mer Rouge. Revue d´ Archéometrie:17.

Franks P, Muennig P, Lubetkin E, and Jia H. 2006. The burden of disease associated with being African-American in the United States and the contribution of socio-economic status. Soc Sci Med 62:2469-2478.

Gehlert S, Murray A, Sohmer D, McClintock M, Conzen S, and Olopade O. 2010. The importance of transdisciplinary collaborations for understanding and resolving health disparities. Soc Work Public Health 25:408-422.

Godreau I. 2000. La semántica fugitiva: "raza", color y vida cotidiana en Puerto Rico. . Rev Cienc Soc Nueva Época 9:52-71.

Gold R, Michael YL, Whitlock EP, Hubbell FA, Mason ED, Rodriguez BL, Safford MM, and Sarto GE. 2006. Race/ethnicity, socioeconomic status, and lifetime morbidity burden in the women's health initiative: a cross-sectional analysis. J Womens Health (Larchmt) 15:1161-1173.

Goldman N, Kimbro RT, Turra CM, and Pebley AR. 2006. Socioeconomic gradients in health for white and Mexican-origin populations. Am J Public Health 96:2186-2193.

Gonder MK, Mortensen HM, Reed FA, de Sousa A, and Tishkoff SA. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24:757-768.

Gonzalez AM, Larruga JM, Abu-Amero KK, Shi Y, Pestano J, and Cabrera VM. 2007. Mitochondrial lineage M1 traces an early human backflow to Africa. BMC Genomics 8:223.

Goovaerts P. 2010. Geostatistical Analysis of County-Level Lung Cancer Mortality Rates in the Southeastern United States. Geogr Anal 42:32-52.

Gravlee CC. 2005. Ethnic classification in southeastern Puerto Rico: the cultural model of "color". Social Forces 83:949-970.

Gravlee CC. 2009. How race becomes biology: embodiment of social inequality. Am J Phys Anthropol 139:47-57.

Gravlee CC, and Dressler WW. 2005. Skin pigmentation, self-perceived color, and arterial blood pressure in Puerto Rico. Am J Hum Biol 17:195-206.

Gravlee CC, Dressler WW, and Bernard HR. 2005. Skin color, social classification, and blood pressure in southeastern Puerto Rico. Am J Public Health 95:2191-2197.

Gravlee CC, and Mulligan CJ. 2010. Re: Racial disparities in cancer survival among randomized clinical trials of the Southwest Oncology Group. J Natl Cancer Inst 102:280; author reply 280-282.

Gravlee CC, Non AL, and Mulligan CJ. 2009. Genetic ancestry, social classification, and racial inequalities in blood pressure in Southeastern Puerto Rico. PLoS One 4:e6821.

Greenberg JH. 1971. Language, culture, and communication. Stanford, Calif.,: Stanford University Press.

Gutenkunst RN, Hernandez RD, Williamson SH, and Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695.

Haga SB, and Venter JC. 2003. Genetics. FDA races in wrong direction. Science 301:466.

Halberstein RA. 1999. Blood pressure in the Caribbean. Hum Biol 71:659-684.

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, and Zegura SL. 2001. Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18:1189-1203.

Henze PB. 1977. Ethiopian journeys : travels in Ethiopia, 1969-72. London: E. Benn.

Ho E. 2006. The graves of Tarim: genealogy and mobility across the Indian Ocean. Berkeley: University of California Press.

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, and McKeigue PM. 2003. Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492-1504.

Horton NJ, and Lipsitz SR. 1999. Review of software to fit generalized estimating equation regression models. The American Statistician 53:160-169.

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, and Herrnstadt C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72:659-670.

Huang X, and Yang Y. 2010. Targeting the TLR9â€“MyD88 pathway in the regulation of adaptive immune responses. Expert Opinion on Therapeutic Targets 0:1-10.

Infascelli R, Pelorosso R, and Boccia L. 2009. Spatial assessment of animal manure spreading and groundwater nitrate pollution. Geospat Health 4:27-38.

Ingman M, Kaessmann H, Paabo S, and Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713.

Integrated Regional Information Networks (IRIN). 2005. Yemen: Akhdam people suffer history of discrimination. Sana: IRIN Middle East (the humanitarian news and analysis service of the UN Office for the Coordination of Humanitarian Affairs). http://www.irinnews.org/report.aspx?reportid=25634. Published online 1 November 2005.

Johnson JA, and Terra SG. 2002. Beta-adrenergic receptor polymorphisms: cardiovascular disease associations and pharmacogenetics. Pharm Res 19:1779-1787.

Johnson W, Kyvik KO, Mortensen EL, Skytthe A, Batty GD, and Deary IJ. 2009. Education reduces the effects of genetic susceptibilities to poor physical health. Int J Epidemiol 39:406-414.

Kaplan S. 1992. The Beta Israel (Falasha) in Ethiopia : from earliest times to the twentieth century. New York: New York University Press.

Kaufman JS, and Cooper RS. 2001. Commentary: considerations for use of racial/ethnic classification in etiologic research. Am J Epidemiol 154:291-298.

Kaufman JS, Cooper RS, and McGee DL. 1997. Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race. Epidemiology 8:621-628.

Kessler D. 1996. The Falashas : a short history of the Ethiopian Jews. London ; Portland, Or.: Frank Cass.

Kim C, Eby E, and Piette JD. 2005. Is education associated with mortality for breast cancer and cardiovascular disease among black and white women? Gend Med 2:13-18.

Kimbro RT, Bzostek S, Goldman N, and Rodriguez G. 2008. Race, ethnicity, and the education gradient in health. Health Aff (Millwood) 27:361-372.

Kitchen A, Ehret C, Assefa S, and Mulligan CJ. 2009. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. Proc Biol Sci 276:2703-2710.

Kitchen A, Miyamoto MM, and Mulligan CJ. 2008. A three-stage colonization model for the peopling of the Americas. PLoS ONE 3:e1596.

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, Laos S, Parik J, Watkins WS, Dixon ME, Papiha SS, Mastana SS, Mir MR, Ferak V, and Villems R. 1999. Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9:1331-1334.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, and Villems R. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752-770.

Klonoff EA, and Landrine H. 2000. Is skin color a marker for racial discrimination? Explaining the skin color-hypertension relationship. J Behav Med 23:329-338.

Knox SS, Jacobs DR, Jr., Chesney MA, Raczynski J, and McCreath H. 1996. Psychosocial factors and plasma lipids in black and white young adults: the Coronary Artery Risk Development in Young Adults Study data. Psychosom Med 58:365-373.

Kvaavik E, Klepp KI, Tell GS, Meyer HE, and Batty GD. 2009. Physical fitness and physical activity at age 13 years as predictors of cardiovascular disease risk factors at ages 15, 25, 33, and 40 years: extended follow-up of the Oslo Youth Study. Pediatrics 123:e80-86.

Lahr MM, and Foley RA. 1998. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. Am J Phys Anthropol Suppl 27:137-176.

Laval G, and Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20:2485-2487.

LaVeist TA. 2005. Minority populations and health : an introduction to health disparities in the United States. San Francisco: Jossey-Bass.

Lee SS, Mountain J, Koenig B, Altman R, Brown M, Camarillo A, Cavalli-Sforza L, Cho M, Eberhardt J, Feldman M, Ford R, Greely H, King R, Markus H, Satz D, Snipp M, Steele C, and Underhill P. 2008. The ethics of characterizing difference: guiding principles on using racial categories in human genetics. Genome Biol 9:404.

Lee SS, Mountain J, and Koenig BA. 2001. The meanings of "race" in the new genomics: implications for health disparities research. Yale J Health Policy Law Ethics 1:33-75.

Lewis B, and Lewis B. 1990. Race and slavery in the Middle East : an historical enquiry. New York: Oxford University Press.

Lewis MP. 2009. Languages of the world, 16th Edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com/.

Lewontin RC. 1972. The apportionment of human diversity. Evolutionary Biology 6:381-398.

Li JL, Canham RM, Vongpatanasin W, Leonard D, Auchus RJ, and Victor RG. 2006a. Do allelic variants in alpha2A and alpha2C adrenergic receptors predispose to hypertension in blacks? Hypertension 47:1140-1146.

Li Y, Jain S, Patil S, and Kumar A. 2006b. A haplotype of angiotensinogen gene that is associated with essential hypertension increases its promoter activity in adipocytes. Vascul Pharmacol 44:29-33.

Lu MC, and Halfon N. 2003. Racial and ethnic disparities in birth outcomes: a life-course perspective. Matern Child Health J 7:13-30.

Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, and Herrera RJ. 2004. The Levant versus the Horn of Africa:

evidence for bidirectional corridors of human migrations. Am J Hum Genet 74:532-544.

Maca-Meyer N, Gonzalez A, Larruga J, Flores C, and Cabrera V. 2001. Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2:13.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, and Torroni A. 1999. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64:232-249.

Martinez FD. 2008. Gene-environment interaction in complex diseases: asthma as an illustrative case. Novartis Found Symp 293:184-192; discussion 192-187.

Maruthur NM, Wang NY, and Appel LJ. 2009. Lifestyle interventions reduce coronary heart disease risk: results from the PREMIER Trial. Circulation 119:2026-2031.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, and Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369.

McEwen BS. 1998. Protective and damaging effects of stress mediators. N Engl J Med 338:171-179.

McKeigue PM, Carpenter JR, Parra EJ, and Shriver MD. 2000. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann Hum Genet 64:171-186.

McVean G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet 5:e1000686.

Mellars P. 2006. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. Science 313:796-800.

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MT, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, and Villems R. 2004. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet 5:26.

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, and Wallace DC. 2003. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100:171-176.

Mohlke KL, Boehnke M, and Abecasis GR. 2008. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. Hum Mol Genet 17:R102-108.

Muller W. 1988. Outline of the history of ancient southern Arabia. In: Daum W, editor. Yemen : 3000 years of art and civilisation in Arabia Felix

Frankfurt/Main: Umschau-Verlag. p 49-54.

Munro-Hay SC. 1991. Aksum : an African civilisation of late antiquity. Edinburgh: Edinburgh University Press.

Nelson L. 2009. Minority Students Needed in Math and Science to Combat 'Brain Drain,' Professors Say. The Chronicle of Higher Education. Washington, DC.

Non AL, Al-Meeri A, Raaum RL, Sanchez LF, and Mulligan CJ. 2010a. Mitochondrial DNA reveals distinct evolutionary histories for Jewish populations in Yemen and Ethiopia. Am J Phys Anthropol. (in press).

Non AL, Gravlee CC, and Mulligan CJ. 2010b. Questioning the importance of genetic ancestry as a contributor to preterm delivery and related traits in African American women. Am J Obstet Gynecol 202:e12.

Non AL, Kitchen A, Mulligan CJ. 2007. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. Mol Phylogenet Evol 44:1164-1171.

Novembre J, and Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. Nat Rev Genet 10:745-755.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, and Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456:98-101.

Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM, Coudray C, Santachiara-Benerecetti AS, Semino O, Bandelt HJ, and Torroni A. 2006. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 314:1767-1770.

Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, and Zhang YP. 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75:966-978.

Patin E, Laval G, Barreiro LB, Salas A, Semino O, Santachiara-Benerecetti S, Kidd KK, Kidd JR, Van der Veen L, Hombert JM, Gessain A, Froment A, Bahuchet S, Heyer E, and Quintana-Murci L. 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. PLoS Genet 5:e1000448.

Peery D, and Bodenhausen GV. 2008. Black + white = black: hypodescent in reflexive categorization of racially ambiguous faces. Psychol Sci 19:973-977.

Petraglia MD, Haslam M, Fuller DQ, Boivin N, and Clarkson C. 2010. Out of Africa: new hypotheses and evidence for the dispersal of Homo sapiens along the Indian Ocean rim. Annals of Human Biology 37:288-311.

Philipp M, Brede M, and Hein L. 2002. Physiological significance of alpha 2-adrenergic receptor subtype diversity: one receptor is not enough. p R287-295.

Picornell A, Gimenez P, Castro JA, and Ramon MM. 2006. Mitochondrial DNA sequence variation in Jewish populations. Int J Legal Med 120:271-281.

Polotsky HJ. 1964. Aramaic, Syriac, and Ge'ez. Journal of Semitic Studies 9:1-10.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, and Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519.

Pritchard JK, Stephens M, and Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Prugnolle F, Manica A, and Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol 15:R159-160.

Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, and Santachiara-Benerecetti AS. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through Eastern Africa. Nature Genetics 23:437 - 441.

Quirin JA. 1992. The evolution of the Ethiopian Jews : a history of the Beta Israel (Falasha) to 1920. Philadelphia: University of Pennsylvania Press.

Raaum RL, Wang AB, Al-Meeri AM, and Mulligan CJ. 2010. Efficient population assignment and outlier detection in human populations using biallelic markers chosen by principal component–based rankings. Biotechniques:in press.

Regueiro M, Cadenas AM, Gayden T, Underhill PA, and Herrera RJ. 2006. Iran: tricontinental nexus for Y-chromosome driven migration. Hum Hered 61:132-143.

Reiner AP, Carlson CS, Ziv E, Iribarren C, Jaquish CE, and Nickerson DA. 2007. Genetic ancestry, population sub-structure, and cardiovascular disease-related traits among African-American participants in the CARDIA Study. Hum Genet 121:565-575.

Relethford JH. 2008. Geostatistics and spatial analysis in biological anthropology. Am J Phys Anthropol 136:1-10.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A,

Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, and Bandelt HJ. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251-1276.

Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, Scozzari R, Macaulay V, and Torroni A. 2003. Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. Am J Hum Genet 72:1058-1064.

Ridl J, Edens CM, and Cerny V. 2009. Mitochondrial DNA Structure of Yemeni Population: Regional Differences and the Implications for Different Migratory Contributions. In: Rose J, I., and Petraglia M, D., editors. Tracking the Origin and Evolution of Human Populations in Arabia. New York City: Springer Science+Business Media. p 69-78.

Risch N, Burchard E, Ziv E, and Tang H. 2002. Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3:comment2007.

Ritte U, Neufeld E, Prager EM, Gross M, Hakim I, Khatib A, and Bonne-Tamir B. 1993. Mitochondrial DNA affinity of several Jewish communities. Hum Biol 65:359-385.

Romney. 1986. American Anthropology 88.

Rose J, I., and Petraglia M, D. 2009. Tracking the Origin and Evolution of Human Populations in Arabia. In: Rose JI, and Petraglia MD, editors. The Evolution of Human Populations in Arabia. New York City: Springer Science+Business Media. p 1-11.

Rose JI. 2006. Among Arabian Sands: defining the Paleolithic of Southern Arabia. Ph.D. dissertation. Dallas: Southern Methodist Universitiy.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, and Feldman MW. 2002. Genetic structure of human populations. Science 298:2381-2385.

Rosengren AH, Jokubka R, Tojjar D, Granhall C, Hansson O, Li D-Q, Nagaraj V, Reinbothe TM, Tuncel J, Eliasson L, Groop L, Rorsman P, Salehi A, Lyssenko V, Luthman H, and Renstrom E. 2010. Overexpression of Alpha2A-Adrenergic Receptors Contributes to Type 2 Diabetes. p 217-220.

Ross CE, and Mirowsky J. 1999. Refining the association between education and health: the effects of quantity, credential, and selectivity. Demography 36:445-460.

Rowold DJ, Luis JR, Terreros MC, and Herrera RJ. 2007. Mitochondrial DNA geneflow indicates preferred usage of the Levant Corridor over the Horn of Africa passageway. J Hum Genet 52:436-447.

Salari K, Choudhry S, Tang H, Naqvi M, Lind D, Avila PC, Coyle NE, Ung N, Nazario S, Casal J, Torres-Palacios A, Clark S, Phong A, Gomez I, Matallana H, Perez-Stable EJ, Shriver MD, Kwok PY, Sheppard D, Rodriguez-Cintron W, Risch NJ, Burchard EG, and Ziv E. 2005. Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. Genet Epidemiol 29:76-86.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, and Carracedo A. 2002. The making of the African mtDNA landscape. Am J Hum Genet 71:1082-1111.

Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, and Carracedo A. 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74:454-465.

Sallis JF, and Owen N. 2002. Ecological models of health behavior. In: Glanz K, Rimer BK, and Lewis FM, editors. Health behavior and health education: theory, research, and practice. 3rd ed. San Francisco: John Wiley & Sons, Inc.

Sanchez JJ, Hallenberg C, Borsting C, Hernandez A, and Morling N. 2005. High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. Eur J Hum Genet 13:856-866.

Sankar P, Cho MK, and Mountain J. 2007. Race and ethnicity in genetic research. Am J Med Genet A 143A:961-970.

Santos C, Montiel R, Angles N, Lima M, Francalacci P, Malgosa A, Abade A, and Aluja MP. 2004. Determination of human caucasian mitochondrial DNA haplogroups by means of a hierarchical approach. Hum Biol 76:431-453.

Saxena R, Nagpal BN, Srivastava A, Gupta SK, and Dash AP. 2009. Application of spatial technology in malaria research & control: some new insights. Indian J Med Res 130:125-132.

Schröder W. 2006. GIS, geostatistics, metadata banking, and tree-based models for data analysis and mapping in environmental monitoring and epidemiology. International Journal of Medical Microbiology 296:23-36.

Schwartz RS. 2001. Racial profiling in medical research. N Engl J Med 344:1392-1393.

Schwartz S, and Meyer IH. 2010. Mental health disparities research: the impact of within and between group analyses on tests of social stress hypotheses. Soc Sci Med 70:1111-1118.

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, and Underhill PA. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am J Hum Genet 70:265-268.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, and Underhill PA. 2006. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet 78:202-221.

Serjeant RB. 1963. The Portuguese off the South Arabian coast. Oxford,: Clarendon Press.

Serre D, and Paabo S. 2004. Evidence for gradients of human genetic diversity within and among continents. Genome Res 14:1679-1685.

Shen P, Lavi T, Kivisild T, Chou V, Sengun D, Gefel D, Shpirer I, Woolf E, Hillel J, Feldman MW, and Oefner PJ. 2004. Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. Hum Mutat 24:248-260.

Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD, Friedlaender J, Jorde LB, Watkins WS, Bamshad MJ, Gutierrez G, Loi H, Matsuzaki H, Kittles RA, Argyropoulos G, Fernandez JR, Akey JM, and Jones KW. 2005. Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2:81-89.

Shriver MD, Parra EJ, 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 CJ, McKeigue PM, and Kittles RA. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387-399.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, and Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740-759.

Steenland K, Henley J, Calle E, and Thun M. 2004a. Individual- and area-level socioeconomic status variables as predictors of mortality in a cohort of 179,383 persons. Am J Epidemiol 159:1047-1056.

Steenland K, Henley J, and Thun M. 2002. All-cause and cause-specific death rates by educational status for two million people in two American Cancer Society cohorts, 1959-1996. Am J Epidemiol 156:11-21.

Steenland K, Hu S, and Walker J. 2004b. All-cause and cause-specific mortality by socioeconomic status among employed persons in 27 US states, 1984-1997. Am J Public Health 94:1037-1042.

Stevens J. 2003. Racial meanings and scientific methods: changing policies for NIH-sponsored publications reporting human variation. J Health Polit Policy Law 28:1033-1087.

Stringer C. 2000. Palaeoanthropology. Coasting out of Africa. Nature 405:24-25, 27.

Sweet E, McDade TW, Kiefe CI, and Liu K. 2007. Relationships between skin color, income, and blood pressure among African Americans in the CARDIA Study. Am J Public Health 97:2253-2259.

Tang H, Jorgenson E, Gadde M, Kardia SL, Rao DC, Zhu X, Schork NJ, Hanis CL, and Risch N. 2006. Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. Hum Genet 119:624-633.

Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, and Risch NJ. 2005. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 76:268-275.

Tavare S, Balding DJ, Griffiths RC, and Donnelly P. 1997. Inferring coalescence times from DNA sequence data. Genetics 145:505-518.

Thangaraj K, Chaubey G, Singh VK, Vanniarajan A, Thanseem I, Reddy AG, and Singh L. 2006. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. BMC Genomics 7:151.

Thomas MG, Weale ME, Jones AL, Richards M, Smith A, Redhead N, Torroni A, Scozzari R, Gratrix F, Tarekegn A, Wilson JF, Capelli C, Bradman N, and Goldstein DB. 2002. Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. Am J Hum Genet 70:1411-1420.

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, and Mountain JL. 2007a. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24:2180-2195.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, and Deloukas P. 2007b. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31-40.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, and Wallace DC. 1996. Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835-1850.

Tsai HJ, Yu Y, Zhang S, Pearson C, Ortiz K, Xu X, Bauchner H, Zuckerman B, and Wang X. 2009. Association of genetic ancestry with preterm delivery and related traits among African American mothers. Am J Obstet Gynecol 201:94 e91-10.

Uerpmann H-P, Potts DT, and Uerpmann M. 2009. Holocene (Re)-Occupation of Eastern Arabia. In: Rose J, I., and Petraglia M, D., editors. Tracking the Origin and Evolution of Human Populations in Arabia. New York City: Springer Science+Business Media. p 205-214.

Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, and Cavalli-Sforza LL. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet 65:43-62.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, and Oefner PJ. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet 26:358-361.

Van Peer P. 1998. The Nile corridor and the out-of-Africa model - An examination of the archaeological record. Current Anthropology 39:S115-S140.

Vernesi C, Di Benedetto G, Caramelli D, Secchieri E, Simoni L, Katti E, Malaspina P, Novelletto A, Marin VT, and Barbujani G. 2001. Genetic characterization of the body attributed to the evangelist Luke. Proc Natl Acad Sci U S A 98:13460-13463.

Wadsworth ME. 1997. Health inequalities in the life course perspective. Soc Sci Med 44:859-869.

Waitzman NJ, and Smith KR. 1994. The effects of occupational class transitions on hypertension: racial disparities among working-age men. Am J Public Health 84:945-950.

Walton E, Takeuchi DT, Herting JR, and Alegria M. 2009. Does place of education matter? Contextualizing the education and health status association among Asian Americans. Biodemography Soc Biol 55:30-51.

Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, Reddy PG, Das PK, Batzer MA, and Jorde LB. 2003. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res 13:1607-1618.

Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, and Paabo S. 1996. mtDNA sequence diversity in Africa. Am J Hum Genet 59:437-444.

Wenner MW. 1991. The Yemen Arab Republic. Boulder, CO: Westview Press.

Williams DR, and Collins C. 1995. Us Socioeconomic and Racial-Differences in Health - Patterns and Explanations. Annual Review of Sociology 21:349-386.

Williams DR, Mohammed SA, Leavell J, and Collins C. 2010. Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. Ann N Y Acad Sci 1186:69-101.

Wong MD, Shapiro MF, Boscardin WJ, and Ettner SL. 2002. Contribution of major diseases to disparities in mortality. N Engl J Med 347:1585-1592.

Young JH, Chang YP, Kim JD, Chretien JP, Klag MJ, Levine MA, Ruff CB, Wang NY, and Chakravarti A. 2005. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. PLoS Genet 1:e82.

Zoossmann-Diskin A, Ticher A, Hakim I, Goldwitch Z, Rubinstein A, and Bonne-Tamir B. 1991. Genetic affinities of Ethiopian Jews. Isr J Med Sci 27:245-251.

BIOGRAPHICAL SKETCH

Amy Non graduated from Central Bucks West High School in Doylestown, PA in Spring, 2000. She then attended Brandeis University in Waltham, Massachusetts from Fall 2000 to Spring 2004 and graduated in May 2004 with a Bachelor of Science degree in biology and a Bachelor of Arts degree in anthropology. She next began graduate school at the University of Florida in August 2004, and received a Master of the Arts degree in anthropology in December 2005 and a Master of Public Health degree in December 2009. She received a Ph.D. in anthropology in August 2010, and began a postdoctorate fellowship at Harvard University through the Robert Wood Johnson Health and Society Scholars Program in September 2010.