AN APPORTIONMENT OF AFRICAN GENETIC DIVERSITY BASED ON MITOCHONDRIAL, Y CHROMOSOMAL, AND X CHROMOSOMAL DATA

by

Maya Metni Pilkington

_____

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ANTHROPOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2008

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation

prepared by Maya Metni Pilkington

entitled An Apportionment of African Genetic Diversity Based on Mitochondrial, Y Chromosomal and X Chromosomal Data

and recommend that it be accepted as fulfilling the dissertation requirement for the

Degree of Doctor of Philosophy

_____ Date: May 11,2007
Michael F. Hammer

_____ Date: May 11,2007
Michael W. Nachman

_____ Date: May 11,2007
Ivy L. Pike

_____ Date: May 11,2007
Michael Worobey

_____ Date: May 11,2007
Stephen L. Zegura

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: May 11,2007
Dissertation Director: Dr. Michael F. Hammer

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

SIGNED: Maya Metni Pilkington

ACKNOWLEDGEMENTS

This work is the culmination of the efforts of many people, and I owe a debt of gratitude to them all. I thank the members of my committee (both present and past): Michael Hammer, Stephen Zegura, Ivy Pike, Michael Worobey, Michael Nachman, Mary Ellen Morbeck, and Mary Stiner. I want to especially thank Stephen Zegura, who has been essential for both his academic integrity, attention to detail, and the support he offered throughout this endeavor. He has read every page of this work from front to back and has given thoughtful input on all parts of it. My advisor, Michael Hammer, brought me into his lab and taught me about population genetics from the ground up, no easy task. His assistance has made me a better scientist and writer, and he provided me with the tools to create this body of work. I thank Michael Nachman for inviting me to his lab meetings and many thoughtful discussions on this work. Ivy Pike has offered her insight into the peoples of Africa as well as emotional support. Michael Worobey introduced me to many interesting evolutionary questions concerning disease and primates and the wonderful world of phylogenetics. I thank M.E. Morbeck for her guidance and advice, and for many thoughtful discussions about early human evolution. Mary Stiner brought me to the University of Arizona and although my interests diverged from hers, I am grateful for all I learned from her. Other faculty members who also deserve a special thank you include Rhonda Gillett-Netting, Steve Kuhn, William Stini, and Joe Watkins, all of whom have added greatly to my education. Importantly, John Olsen has never waivered in his support of me as a student and as a colleague, and always had time to listen to my concerns.

The Hammer lab has been a motivating and fun learning environment for me over the years. It is there that I found both wonderful friends as well as great sounding boards for ideas—so thank you to Matthew Kaplan, Jason Wilder, Heather Norton, Tanya Karafet, Tamar Erez, Murray Cox, Fernando Mendez, August Woerner, Amy Russell, and Tesa Severson. I also thank the postdocs, graduate students, technicians, undergraduates and other members of the lab, especially: Thiep Angoui, Abby Bigham, Latifa Borgelin, Veronica Chamberlain, Taylor Edwards, Elizabeth Galdi, Dan Garrigan, Veronica Kearney, Sarah Kingan, Zahra Mobasher, David Morales, Olga Savina, Micala Rider, and Elizabeth Wood. I am indebted to Nirav Merchant, Susan Miller, Gavin Nelson, and Dirk Harris for their fanatastic computer support. The U of A sequencing facility has done a remarkable job, especially Brian Coulihan, Janet Cooley, and Ryan Sprissler. The Anthropology and EEB graduate students have made my whole graduate school experience much more enjoyable, and I value their friendship and support over the years, especially: Michelle Gamber, Meigan Goodyer, Amy Margaris, Matthew Saunders, Jon Scholnick, and Gabriella Wlasiuk. I learned a great deal from discussions and courses with all of you.

I greatly appreciate the hard work and dedication of my collaborators who have donated samples for this study, especially: Thiep Angoui, Giovanni Destro-Bisol, Himla

Soodyall, and Beverly Strassmann. I thank them all for taking the time to read through many drafts of this work and the manuscripts which emerged from it.

The funding for this project came from the Department of Anthropology (Emil Haury, Mary Alice Sherry Helm, and Rieker grants), Social and Behavioral Sciences, Women in Science and Engineering, The National Science Foundation Dissertation Improvement Grant, the National Science Foundation IGERT in Functional, Evolutionary and Computational Genomics, and the Michael A. Cusonovich Fellowship.

This work would not have been possible without the love and support of my husband Guy, who has accompanied me on this long journey while accomplishing his own goals, and who has assisted me in fulfilling my dreams while juggling the responsibilities of family and school himself. My daughter Madeleine deserves her own thank you for being a terrific baby, keeping me grounded, and taking wonderfully long naps that allowed me to get a lot of this work done. My parents, Fouad and Mona always encouraged my academic pursuits and gave me an appreciation for science as well as hard work. I am lucky to have the support of a truly unique family (no matter how far away they may be) and I thank Nayla and Erik Della Penna, Najy and Meryl Metni, Jawad Metni, Nicole and Robert Pilkington, Carmen, Greig, and Mel Pilkington for always lending an ear and making me smile.

Finally, I want to express my sincerest thank you to the native peoples of Africa. I hope that this work will prove to be beneficial for you.

DEDICATION

This work is dedicated to my husband Guy, who has given me endless emotional support, and to the two most important women in my life, my past and my future—Mona Metni who showed me that I could be whatever I chose to be, and Madeleine Sloane Pilkington who inspires me to be it.

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" (I found it) but "That's funny...".

--Isaac Asimov

TABLE OF CONTENTS

LIST OF TABLES…… 10

LIST OF FIGURES……...13

ABSTRACT……..…15

CHAPTER 1: INTRODUCTION……..……..……..…17

AN OVERVIEW OF THIS STUDY……...18

THE CONNECTION BETWEEN DEMOGRAPHY AND ANTHROPOLOGICAL GENETICS……...20
Definition of Basic Terms…….20
Population Size…….21
Population Structure…….23

AFRICAN DEMOGRAPHY: POPULATION SIZE CHANGE AND STRUCTURE……25
Population Size Change……25
Population Size Change: The Bantu Expansions……..27
Population Structure: Sex-Biased Migration……30
Ancient Population Structure: DNA Sequence Evidence ……32
Ancient Population Structure: Fossil Evidence……34

GENOMIC REGIONS OF INTEREST……38
Rationale for the Choice of Loci…….……..38
The Mitochondrial DNA (mtDNA)……..39
The Y Chromosome (NRY)…….….42
The X Chromosome……..45

BACKGROUND ON POPULATIONS USED IN THIS STUDY ……....48
The Khoisan of Southern Africa……...49
The Southeast Bantu Speakers of Southern Africa……...51
The Dinka of the Sudan……...53
The Dogon of Mali……....54
The Bakola of Cameroon……..56
Comparative Non-Human Primates……....…….…58

IMPROVEMENTS ……...……..59
Multilocus Comparisons across Populations ……..……..…60
Direct Re-Sequencing Data…….…..……...60

Population-Based Sampling ……..…61

TABLE OF CONTENTS- *CONTINUED*

CHAPTER 2: MATERIALS AND METHODS ……..……..…64

OVERALL RESEARCH DESIGN…….65
Samples Used in This Study…….65
Loci Used in This Study……..…….……....80

METHODS……...84
Laboratory Procedures……..84
Analytical Procedures……...94

CHAPTER 3: CONTRASTING SIGNATURES OF POPULATION GROWTH FOR MITOCHONDRIAL DNA AND Y CHROMOSOMES AMONG HUMAN POPULATIONS IN AFRICA…….……102

ABSTRACT……104

INTRODUCTION…….105

MATERIALS AND METHODS……...…107
Populations and Loci Surveyed…..……..107
Population Genetic Analyses and Tests of Population Growth……..108

RESULTS……...…111
Patterns of Nucleotide Diversity…….……111
GENETREE Simulations……..……..112
Mismatch Distributions……...……113

DISCUSSION…….113
Variation in Mutation Rate and Mode……..……..114
Natural Selection…….……115
Sex-Specific Demographic Processes…….……116

CONCLUSIONS……...118

CHAPTER 4: SEX-SPECIFIC DEMOGRAPHIC PROCESSES IN SUB-SAHARAN AFRICA: EVIDENCE FOR WIDESPREAD MALE MIGRATION WITH REPLACEMENT…….....131

ABSTRACT……132

INTRODUCTION……133

TABLE OF CONTENTS- *CONTINUED*

MATERIALS AND METHODS……..135
Subjects and Molecular Loci……..…135
Isolation and Migration (IM)……....136
Population Genetic Analyses……..…138

RESULTS……..……...139
IM Analyses-- Migration and Effective Population Size.…...……..…139
Shared Haplotypes……140
Population Differentiation—F_{ST} and AMOVA.....……..…141

DISCUSSION……..…....143
The Bantu Expansions—Male Migration with "Replacement" Model…….....…....144

CONCLUSIONS……...…147

CHAPTER 5: MULTILOCUS ESTIMATES OF ANCIENT POPULATION STRUCTURE AND THE ORIGIN OF MODERN HUMANS……..…….....164

ABSTRACT……165

INTRODUCTION……166

MATERIALS AND METHODS……...……169
Samples and Loci……..…….....……169
Population Genetic Analyses……...…….…...……170
IM Model and Computations……..……...…...…170

RESULTS……...…171
Population Divergence……..……..……172
Current and Ancestral Effective Population Size……173
Migration…….....175

DISCUSSION AND CONCLUSIONS…..……..……175

CHAPTER 6: SUMMARY……...……202

REFERENCES……206

LIST OF TABLES

CHAPTER 1

TABLE 1.1 *Alu* names, families and gene regions used in this study……62

CHAPTER 2

TABLE 2.1A Samples, population name, samples donor and haplotype for the four loci under consideration...……...66

TABLE 2.1B List of the haplotypes for mtDNA, related to **TABLE 2.1A**……..74

TABLE 2.1C List of the haplotypes for NRY, related to **TABLE 2.1A**……76

TABLE 2.1D List of the haplotypes for *PDHA1*, related to **TABLE 2.1A**…….…...77

TABLE 2.1E List of the haplotypes for *RRM2P4*, related to **TABLE 2.1A**…….…78

TABLE 2.2 Loci used in this study, alignment lengths and reference sequences for each population……..82

TABLE 2.3 NRY *Alu* nomenclature, family, Blat location (start and stop) and gene region.……..83

TABLE 2.4 Primers used for DNA amplification……..…….88

TABLE 2.5 Loci conditions for PCR amplification……..…….…89

TABLE 2.6 Primers used to sequence loci……..92

CHAPTER 3

TABLE 3.1 MtDNA and NRY polymorphism data for each sub-Saharan African population……..…121

TABLE 3.2 Comparison of observed and simulated Fu's *Fs* values for mtDNA and the NRY……..…123

LIST OF TABLES-*CONTINUED*

TABLE 3.3 Population parameters estimated using GENETREE for constant size (upper row for each population) and exponential growth (lower row) ...……..124

SUPPLEMENTARY TABLE 3.1 MtDNA haplotype frequencies……128

SUPPLEMENTARY TABLE 3.2 NRY haplotype frequencies…….130

CHAPTER 4

TABLE 4.1 IM estimates of migration rates from mtDNA and the NRY (with 90% HPD confidence intervals in parentheses)…….……...149

TABLE 4.2 MtDNA and NRY migration rate ratios and effective population size estimates from GENETREE and IM for hunter-gatherers (HG) and food-producers (FP).……...……...150

TABLE 4.3 Estimates of NRY N_{e1} (effective population size of the first population listed), N_{e2} (effective population size of the second population listed), and N_{eA} (ancestral effective population size) for the NRY…... 151

TABLE 4.4 Number of haplotypes and haplotype diversity data for mtDNA and the NRY, with the mean below…….……..152

TABLE 4.5 Analysis of molecular variance for mtDNA and the NRY (number of comparisons in parentheses)……...…….……..153

LIST OF TABLES-*CONTINUED*

CHAPTER 5

TABLE 5.1 Summary of descriptive statistics describing nucleotide polymorphism for the mtDNA *COIII*, NRY-*Alu*, and X-linked *RRM2P4*, and *PDHA1* loci……...……..…181

TABLE 5.2 Multilocus estimates of scaled effective population sizes, migration rates, and divergence times using the IM model…….....…..…183

SUPPLEMENTARY TABLE 5.1 Multilocus estimates of scaled effective population sizes and divergence times using the IM model (90% HDP confidence intervals in parentheses)……189

SUPPLEMENTARY TABLE 5.2 Multilocus estimates of effective population sizes and migration rates using the IM model (90% HDP confidence intervals in parentheses)……...190

LIST OF FIGURES

CHAPTER 1

FIGURE 1.1 Map of populations examined in this study……...…63

CHAPTER 2

FIGURE 2.1 Schematic representations of loci amplified for this work……..…79

FIGURE 2.2 Schematic representation of *PDHA1*……..…80

FIGURE 2.3 Schematic representation of *RRM2P4* pseudogene……81

FIGURE 2.4 Schematic of the Isolation with Migration model, with associated parameters……..……101

CHAPTER 3

FIGURE 3.1 Map showing the country of origin, the population name, the population name, abbreviation and the subsistence strategy (FP= food-producer, HG= hunter-gatherer) of the five sub-Saharan African populations…...……..……...126

FIGURE 3.2 Mismatch distributions for mtDNA and NRY……127

CHAPTER 4

FIGURE 4.1 Map showing the country of origin, the population name, the language family and the subsistence strategy of the five sub-Saharan African populations. (Abbreviations: FP= food-producer, HG= hunter-gatherer, W= west, E= east, and S= south)……...……154

FIGURE 4.2 Distribution of mtDNA and NRY **(A)** Population migration rates (only migration rates whose 90% highest posterior density interval did not include zero are shown), and **(B)** shared haplotypes (unique haplotypes shown in grey), with NRY Haplogroups A, B, and E denoted……..…….155

LIST OF FIGURES-CONTINUED

FIGURE 4.3 F_{ST} estimates for mtDNA and NRY, with focus on the Bantu-speaking populations denoted by black circles and unbroken lines (Bakola, Dogon, and SE Bantu)……... 156

FIGURE 4.4 Un-rooted UPGMA trees for F_{ST} based on **(A)** mtDNA and **(B)** NRY for the five sub-Saharan African populations. ……..….157

FIGURE 4.5 The estimated route of the Bantu expansions, based on linguistic and genetic data ……...……..158

FIGURE 4.6 Models of (A) population contraction with replacement, and (B) population expansion……..159

SUPPLEMENTARY FIGURE 4.1 Estimated mtDNA and NRY TMRCAs in thousands of years (KYR) for the five sub-Saharan African populations (with 95% confidence intervals) from GENETREE…….....162

SUPPLEMENTARY FIGURE 4.2 MtDNA and NRY TMRCAs in thousands of years (KYR) for the five sub-Saharan African populations simulated from ms using the mtDNA and NRY migration matrices generated from IM…….....163

CHAPTER 5

FIGURE 5.1 UPGMA trees based on F_{ST} for (**A**) mtDNA, (**B**) NRY, (**C**) *RRM2P4*, and (**D**) *PDHA1*……..185

FIGURE 5.2 Shared haplotypes for (**A**) mtDNA, (**B**) the NRY, (**C**) *RRM2P4*, and (**D**) *PDHA1*…….……...……..…186

FIGURE 5.3 Estimates of migration rates and divergence times (KYR) for five sub-Saharan African populations (migration rates with 90% HDP intervals not including zero shown)……..……187

FIGURE 5.4 UPGMA tree of the split times (in thousands of years) for the five sub-Saharan populations…….…….....188

SUPPLEMENTARY FIGURES 5.1-5.10 Posterior probability distributions……191

SUPPLEMENTARY FIGURE 5.11 The gene tree estimate for *PDHA1*…….……201

ABSTRACT

In an effort to better understand patterns of genetic variation in modern African populations, I surveyed nucleotide variability at four loci in five diverse sub-Saharan African populations. First, I analyzed the mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY), asking specifically if similar models of population size change could be fit to re-sequencing data from these two loci when examined in the same populations. Four tests of population growth were employed and results indicated that food-producing populations best fit a model of exponential growth for the mtDNA but not the NRY, and hunter-gathering populations best fit a model of constant population size for both mtDNA and the NRY. These results are likely due to sex-specific migration or differences in the effective population sizes of males and females.

Next, I examined mtDNA and NRY population structure in these same populations, to assess the relative effects of migration and effective population size on patterns of mtDNA and NRY nucleotide variability. I used an Isolation with Migration (IM) model to disentangle estimates of effective population size and migration. Results indicated that levels of mtDNA population structure are higher than those of the NRY, and female migration tends to be unidirectional while that of males is largely bidirectional. I found that in food-producing populations, male migration rate estimates are in fact higher, not lower, than those of females, while estimates of male effective population size are strikingly small. I infered that males have experienced a period of

population size reduction due to replacement, and that this most likely occurred during the Bantu expansions, approximately 5,000 years ago.

Finally, I assessed population structure in these populations using a multilocus approach which estimated current and ancestral effective population sizes, migration rates, split times and fraction of the ancestral population that contributed to current populations. Current and ancestral effective population sizes ranged from ~5,000-8,000 individuals. Most populations showed an increase in size relative to the ancestral population. Population split times ranged from 17-142 thousand years (KYR); the Khoisan split times were the oldest and the Niger-Congo speaking populations' split times the most recent. Since the oldest population split times precede the dates for the earliest modern humans outside of Africa, I posited that modern humans likely evolved at a time when structured populations already existed in Africa.

CHAPTER 1: INTRODUCTION

AN OVERVIEW OF THIS STUDY

This research provides the first nucleotide-level estimate of modern African population subdivision based on re-sequencing data from multiple loci, and should lead to deeper insights into the processes that have shaped patterns of genomic variation in African populations, such as major migration events and population size expansions. It is striking that relatively few genetic studies have concentrated on African populations, given that the majority of genetic and paleoanthropological data suggest that anatomically modern humans (AMH) originated in Africa. If we are to understand how this major evolutionary transition occurred, it is imperative that we develop better databases that allow us to characterize patterns of genetic variation within and between extant African populations. This work should help to expand public knowledge of African population history and indirectly benefit populations that currently are underrepresented in medical research.

The goal of this research is to use DNA sequence data from mitochondrial, Y chromosomal, and X chromosomal loci from multiple African populations to further our understanding of African population structure and history. First I begin by examining haploid genetic data to (1) determine the demographic model of population size change that is most appropriate for the populations in this study, and (2) resolve whether maternally inherited mitochondrial and paternally inherited Y chromosomal data fit the same model of population size change. I show that differences in models of population size change for the mitochondrial DNA (mtDNA) and non-recombining portion of the Y chromosome (NRY) are likely due to differences in effective population size and/or

migration rates. Therefore, I next turn to these sex-specific demographic processes by examining the haploid data to (3) assess the relative effects of differential mtDNA and NRY effective population sizes and migration rates on patterns of diversity in sub-Saharan Africans. These data are then placed in a larger genomic context by combining the haploid data with data from two unlinked X loci (Ribonucleotide Reductase M2 Polypeptide Pseudogene 4, *RRM2P4*, and Pyruvate Dehydrogenase Alpha 1, *PDHA1*) to (4) obtain a multilocus estimate of the time the populations split, current and ancestral effective population sizes, and migration rates.

This introductory chapter gives a brief background on studies of human demography, including an explanation of the relationship between the demographer's and the anthropological geneticist's definition of the term "demography". I then define the terms "population" and "structure" and the critically important concepts of effective population size (N_e) and migration rate (m). Next, I present what is already known about these demographic factors based on African populations. I take a closer look at the data used to estimate population size change in Africa and levels of population structure (both modern and ancient). The chapter concludes with a discussion of the loci chosen for this work, the populations examined, and the improvements this dissertation makes upon available work.

THE CONNECTION BETWEEN DEMOGRAPHY AND ANTHROPOLOGICAL GENETICS

Definition of Basic Terms

"Demography" is a multi-faceted term that varies in specificity when used by demographers and anthropological geneticists. In the simplest sense, demography is the statistical study of human populations. In the past, demography has been mostly concerned with the population characteristics of countries (Keyfitz and Flieger 1968), and less concerned with behavior. The classic definition of demography is "the study of the size, territorial distribution, and composition of a population, changes therein, and the components of such change" (Hauser and Duncan 1959: 2). The "size" generally refers to fertility and mortality, measured in terms of births and deaths, while the "composition" refers to migration, measured in terms of in-migration and out-migration. These factors provide a strong quantitative approach to studying populations. Anthropological geneticists describe and measure populations; they combine quantitative approaches (where processes and events are examined) with the study of behavioral and social processes (Mielke and Fix 2007).

The "population" is one of the most basic terms in demography, and generally refers specifically to the people living in a particular geographic location at a specific point in time (Murdock and Ellis 1991). The term is further refined when used by anthropological geneticists because genetics is concerned with the "transmission of genes among individuals across generations" (Mielke and Fix 2007, p.114). Therefore a population must persist through time, and should not be merely a transient grouping of people. It is the subdivision of these populations that is of interest here, and this can be

caused by many factors, both socio-cultural (e.g. different linguistic affiliation or specific marriage practices) and environmental (e.g., geographic barriers).

Population Size

In the past, demography has been concerned with census sizes; however, problems arise when attempting to (1) define the population and (2) actually record every person in the population. Therefore, the population effective size (N_e) was introduced by Sewall Wright (1931) to take into account all of the factors that cause a population to deviate from the idealized randomly mating population. These factors include age structure, unequal sex ratios, overlapping generations, differences in fertility and mortality rates, and inbreeding. Anthropological geneticists are concerned with the population effective size, which can be roughly one third of the census population once all of these factors are taken into account. The difference between census size and effective population size is significant, since the amount of genetic drift in a population is directly dependent on this factor--that is, the expected variance in gene frequencies in a population is critically related to the effective population size, not the census size.

Now that we have a way to quantify population size, we can examine population size change through time. In anthropological genetics, the direction of inference is generally from genetics to demography; that is, the estimate of genetic diversity in humans is so low that we infer a low number of founding ancestors, based on population genetic theory. The most commonly referred to estimate of effective population size in humans is ~10,000 (Takahata 1993; Hammer 1995; Harding et al. 1997; Harpending and Rogers

2000; Jorde, Watkins, and Bamshad 2001). It is likely that on a local scale there have been population crashes or bottlenecks as well as expansions, and that globally populations did not increase dramatically in size until relatively recently. This size increase (N_{t+1}) can be determined by the formula $N_{t+1}=N_te^{rt}$, where N is population size, *r* is the annual growth rate and *t* is time, and e is the base of the natural logarithms. Using such a formula, it is evident that population size can quickly become enormous, if left unchecked.

The estimate of effective size is not without its caveats, though. Both population structure and extinction and recolonization events can dramatically affect estimates of effective population size. It has been noted that Africans have higher levels of genetic heterozygosity than non-African populations—possibly due to higher rates of migration or larger effective size of the population (Stoneking et al. 1997). However, Africans also tend to have higher levels of differentiation between populations, which argues against high migration rates (this idea is explored further in **CHAPTER 4** of this work). Eller's work (2002) showed that if a model of recurrent extinction and recolonization events is used for early humans, the total number of founding breeding individuals (N_b) could have exceeded 300,000. Under Wright's island model of migration, $N_e = N_b/1\text{-}F_{ST}$, where N_b is the total breeding size of the population. Through simulations, Eller showed that as the extinction rate increases in a population, so does F_{ST} and the breeding size of a population. Therefore, the results of this work support a model whereby the effective population size may have been relatively low (~10,000) while the breeding size of the

population could have potentially been much higher (~300,000), given the right circumstances.

Population Structure

Though some terms are interchangeable between demography and anthropological genetics, the term "population structure" is not one of them. A brief examination of the history of the study of demography is useful here. "Formal demography" is mostly concerned with fertility, mortality, age structure, and composition of populations, and can be traced back to John Graunt in 1662 (who created the first life table). Another type of demography, called "population studies" or "social demography" has a more recent origin and can be traced back to John Malthus in 1798, who was deeply concerned with population growth when left unchecked. Population studies focus on "population compositions and changes from substantive viewpoints anchored in another discipline, be [they] sociological, economic, biological, or anthropological" and have been more interdisciplinary (Xie 2000: 670). The field of population genetics integrated natural selection with Mendelian genetics, and has foundations in the works of Fisher (1930), Wright (1931), and Haldane (1932). Shortly thereafter, a unified theory of evolution was presented and the "Modern Synthesis" was born (Huxley 1942). Over about a decade, the works of Dobzhansky (1937), Mayr (1942), and Simpson (1944) combined facts from studies of natural selection, genetics, and paleontology. However, early demographers still questioned the relative influence of random genetic drift over natural selection. Since genetic drift is dependent on population breeding size, early work such as that of

Birdsell (1951) attempted to estimate the population breeding size and examine the effects of depopulation. Work on relative fertility and mortality in relation to natural selection did not come until slightly later (Crow 1958). Presently, demographers generally use the term "population structure" to refer to the age and sex composition of a population, which is often presented in life tables containing mortality rates by age and sex in a population. Although an ideal demographic study is longitudinal, demographic data are typically a cross-section of a particular cohort at a particular moment (and as such, can often be incomplete with respect to birth, death, and migration data). This stands in stark contrast to the forces of biological evolution, which transcend generations. Anthropological geneticists use the term "population structure" very broadly to describe the many forces of evolution that can influence patterns of genetic variation within a population (Cavalli-Sforza 1959; Gage 2000).

Likewise, the term "migration" means two different things between the fields as well. In demography, the term "migration**"** generally refers to the movement of people to a new location for at least a year with the intent of remaining in the new location (Newell, 1988). However, for anthropological geneticists "migration" is defined as one-way movement to a new population (Fix 1999), where the importance of the term lies in the *consequences* of migration—gene flow. Migration can serve to spread newly arisen mutations through populations, homogenize populations genetically, or, in the case of kin-structured migration, make populations look more genetically distinct (Fix 1978). Socio-cultural factors have profound influence on human tendencies to migrate; factors such as kinship (patrilineal and matrilineal descent), subsistence strategies (technology

can negate distance, and economy can necessitate it), and culture (shared language, religion, and ethnicity influence mate choice) can largely shape migration patterns. This is discussed further in **CHAPTER 3** of this work.

AFRICAN DEMOGRAPHY: POPULATION SIZE CHANGE AND STRUCTURE

Many population genetic analyses rely on the assumption that human populations are in genetic equilibrium. However, some populations clearly show the signs of population expansion, migration or subdivision. Useful techniques for understanding the processes which have shaped modern African genetic variation include obtaining estimates of the population effective size and population structure (including in-migration and out-migration, and population subdivision). Below is a discussion of how our understanding of extant African genetic variation has been affected by the study of these factors.

Population Size Change

We know from archaeological remains that human populations expanded in size relatively recently, as first evidenced by a Late Pleistocene shift in hunting technology and later by the advent of agricultural subsistence during the Early Neolithic (Stiner, Munro, and Surovell 2000). The effective population size of humans is consistently estimated to be around 10,000 (Li and Sadler 1991; Rogers and Harpending 1992; Takahata 1993; Horai et al. 1995; Zietkiewicz et al. 1998), but is usually based on a pooling of African and non-African samples, which can result in a bias towards the sample size of non-Africans (Tishkoff and Williams 2002) When African populations

are analyzed separately from non-African populations, Africans have a larger N_e than non-Africans (some as high as 20,000) for autosomal (Harris and Hey 1999a; Jaruzelska, Zietkiewicz, and Labuda 1999) Y chromosome (Hammer et al. 1997; Hammer et al. 1998), and mtDNA sequence variation (Sherry et al. 1994). Since this holds across many loci, it is most likely indicative of a significant demographic event in human history. The migration of modern humans from Africa to colonize the rest of the world, including a bottleneck upon leaving Africa, has probably had the greatest influence on patterns of extant human genetic variation. Since it is my goal to accurately characterize African genetic diversity, in this study I estimate the N_e of Africans based on five African population samples, across four independent loci from different parts of the human genome.

Most studies of human mitochondrial DNA (mtDNA) have indicated that humans show strong evidence of population growth (Cann, Stoneking, and Wilson 1987; Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Sherry et al. 1994; Horai et al. 1995; Penny et al. 1995; Jorde et al. 1997; Harpending et al. 1998; Relethford 1998; Ingman et al. 2000; Maca-Meyer et al. 2001); however, patterns in Africans have been shown to be heterogeneous (Excoffier and Schneider, 1999). Data from the non-recombining region of the Y chromosome are no clearer. While some studies show evidence of recent population growth (Pritchard et al. 1999; Shen et al. 2000; Thomson et al. 2000; Underhill et al. 2000), others clearly do not (Jorde et al. 1995; Clark et al. 1998; Pereira et al. 2001; Dupanloup et al. 2003). These conflicting results are discussed in much further detail in **CHAPTER 3** of this work.

Bi-parental nuclear DNA datasets have not been consistent in indicating population expansions, and, more importantly, have revealed conflicting patterns of population expansion in comparison with other genomic regions (Jorde et al. 1995; Hey 1997). In fact, many nuclear DNA loci indicate an excess of intermediate frequency haplotypes, which is inconsistent with a model of population growth. For example, *PDHA1*, shows evidence of extensive polymorphism and ancient variation, like many other X chromosomal and autosomal loci (Harris and Hey, 1999). Based on frequency spectra, Frisse et al. (2001) claim that Africans appear to be consistent with an equilibrium model while non-Africans do not. For autosomal loci, such as *Dys44* (Zietkiewicz et al. 1998) and *β-globin* (Harding et al. 1997), there is a paucity of rare alleles outside of Africa. This may be the result of a deficiency in new polymorphisms, not a loss of pre-existing ones (Zietkiewicz et al. 1998). Direct comparison of mtDNA and NRY sequence data to sequence data from other genomic compartments has not been available until recently (Wilder et al. 2004). As a result, the question of whether multiple compartments of the human genome show signs of an historical expansion remains largely unresolved.

Population Size Change: The Bantu Expansions

African demography has undoubtedly been affected by the Bantu Expansions, which are best characterized as a number of major demographic events that are estimated to have occurred ~3,000-5,000 years ago (Cavalli-Sforza, Menozzi, and Piazza 1994; Ehret 2001). It is thought that populations developed agriculture and subsequently dispersed across Africa either intermarrying with or replacing extant hunter-gathering

populations. Evidence of these expansion events comes from three major areas of study: linguistics, archaeology, and genetics.

Perhaps the greatest support for the expansion of Bantu-speaking peoples across Africa comes from linguistic studies. A matter of great curiosity has been how such a vast number of people, nearly one quarter of all Africans, speak one of the many Bantu languages. The Bantu subgroup is categorized within the Niger-Congo language family, and includes about 500 languages and 100 million speakers (Cavalli-Sforza, Menozzi, and Piazza 1994). The origin of this subgroup appears to be quite recent (Greenberg 1963; Hiernaux 1968). Based on linguistic evidence, Bantu-speaking people are likely to have spread from the area of Nigeria and Cameroon to the east and south prior to the Iron Age, most likely during the Neolithic, around 1000 B.C. (Cavalli-Sforza, Menozzi, and Piazza 1994, see Figure 4.8 for a map). Though there is clear evidence of an "Eastern" Bantu subgroup that is distinct from proto-Bantu, the evidence of a "Western" Bantu counterpart has been the source of some debate (Guthrie 1963; Oliver 1966; Ehret 1972; Vansina 1984). The "Eastern" Bantus occupy present-day Uganda and Kenya to South Africa, excluding the Cape region of South Africa (Ehret 2001), and the work presented here is concerned with this group of Bantu-speaking peoples.

Archaeological evidence, such as the Dimple-based and Channelled wares, suggests an expansion of agricultural peoples with dates and geographical ranges that reinforce those suggested by the linguistic evidence (Hiernaux 1968). It is estimated that the origins of agriculture in Africa took place during the first millennia B.C., at approximately the same time as the Bantu Expansions (Holden 2002a). It is unclear

whether the earliest Bantu speakers were able to work iron, and thus spread this technology with the Bantu Expansions, and archaeologists and linguists alike have attempted to address this question. Some of the earliest iron smelting sites in Africa date to around 840-420 B.C., and are found in Senegal, Niger, Chad, Nigeria, Central African Republic, Rwanda, Barundi, and Cameroon (Vansina 2006). Recent linguistic work suggests that the technology for ironworking was obtained from western Africa, and was not created independently elsewhere in Africa, adding more support to the idea that the Bantu Expansions occurred rather rapidly, that is, within 400 years (Vansina 2006).

Beyond the linguistic and archaeological evidence for a major recent demographic event in Africa, the work on multiple genetic loci has demonstrated a relatively homogenous Bantu grouping existing within only two clusters on a genetic tree of all Africans, lending strong support to the idea of a rapid expansion of Bantu-speaking peoples (Cavalli-Sforza, Menozzi, and Piazza 1994, Figure 3.5.1) (incidentally, the Bantus cluster closely with Nilotics, who are also included in this dissertation work). In fact, Cavalli- Sforza, Menozzi, and Piazza (1994) attempt to illustrate the path of the Bantu Expansions in their Figure 3.9.3; however, they were only able to infer the directionality of the migration events based on *archaeological* information and not genetic data, using the methods available at the time. **CHAPTER 4** of this work explores this point further, and uses newer methods to directly estimate the directionality of gene flow. Other studies have focused specifically on the eastern and western Bantu expansions, and have provided similar evidence of a rapidly occurring homogenization of the gene pool (Vansina 1984; Nabulsi et al. 1993; Lane et al. 2002; Salas et al. 2002;

Destro-Bisol et al. 2004a; Beleza et al. 2005). Studies of the Y chromosome, in particular, show that Bantu-speaking peoples cluster tightly on phylogenetic trees of all Africans (Cruciani et al. 2002; Wood et al. 2005). The pattern of expansion proposed by both linguists and archaeologists is, in a general sense, confirmed by the genetic data.

Population Structure: Sex-Biased Migration

Wood et al. (2005) support differential patterns of male and female genetic differentiation and gene flow in Africa. The results for the study of the NRY clearly show no correlation between genetic and geographic distance but do show a correlation between genetic and linguistic distances. Conversely, mtDNA shows a weak correlation between genetics and both linguistics and geography. A model of sex-biased migration is invoked to explain the differences in the degree of association between genetic and linguistic variation for the mtDNA and NRY, though these results were not based on direct sequencing data for the NRY and compared different samples for mtDNA and the NRY. Adding another level of complexity, Destro-Bisol et al. (2004) find evidence for clear differences in the population structure of males and females in food-producers and hunter-gatherers in Africa. These differences are attributed to asymmetric gene flow resulting from social inequalities as well as differing levels of polygyny and patrilocality between the two groups. In food-producing populations, where land is passed down for generations, groups tend towards patrilocality more often than hunting-gathering populations, thereby resulting in higher rates of female than male migration (Cavalli-Sforza, Menozzi, and Piazza 1994). Likewise, food-producers tend to practice polygyny

to a greater degree than hunter-gatherers (Murdock 1967; Destro-Bisol et al. 2004b). The practice of polygyny can serve to decrease the effective population size of males compared with females since fewer males are contributing to the next generation.

Previously, work based on Y chromosomal data by Hammer et al. (2001) and Seielstad et al. (1998) offered conflicting results concerning sex-mediated gene flow on the structure of modern African populations. Whereas Hammer et al. (2001) concluded that subdivision in modern Africans is quite low ($\Phi_{st} = 0.222$), Seielstad et al. (1998) pointed to higher levels of subdivision on the Y chromosome than the mtDNA, supposedly the result of female-mediated gene flow due to patrilocality. Hammer et al. (2001) proposed increased male migration rates on an intercontinental scale, though on a regional scale females may in fact have higher migration rates than males based on elevated among-groups and among-populations-within-groups variances for Y chromosome compared to mtDNA data.

The resulting differences between the Hammer and Seielstad studies most likely reflect the differences in population sampling strategies employed, which emphasize the significant impact of socio-cultural factors on human migration and population structure. While the Hammer study includes hunter-gatherers and food producers (agriculturalists and pastoralists), the Seielstad study only includes food-producers. This is noteworthy, because factors such as subsistence strategy affect not only the economy of the population, but the location of marital residence (patrilocal or matrilocal), and the marriage practices (monogamy or polygyny). In fact, subsequent work by Wilder et al.

(2004) with an improved sampling technique has indicated relatively similar levels of male and female genetic differentiation on a global level.

Ancient Population Structure: DNA Sequence Evidence

The structure of the earliest AMH population is a topic that rarely has been addressed (Harris and Hey, 1999; Yu et al. 2001). Yet the breadth of morphological variation associated with Middle Pleistocene hominin fossils leaves open the possibility that early AMH evolved in a subdivided population. Indeed, the premise of this work is that genetic data can be used to address questions of ancient population structure.

There are very few systematic studies of African populations at the nucleotide level, other than those focusing on mtDNA. Most of what we know about African population structure comes from work on classical markers which were the first to indicate that Africans harbor more genetic variation than populations from other continents (Cavalli-Sforza 1966). This suggested that African populations are either ancestral to all AMH (i.e., they have been accumulating mutations for a longer time), and/or the effective population size of Africans is larger than that of any other region in the world (Relethford 2001a). DNA-level studies over the past two decades have continued to support the hypothesis of greater African diversity and have revealed additional patterns of genetic variation. For example, when gene trees could be reconstructed for a given DNA sequence, Africans tended to fall on both sides of the most ancestral node, while non-Africans were usually limited to a single side of the tree (Cann, Stoneking, and Wilson 1987; Takahata, Lee, and Satta 2001). African populations also tend to harbor more rare

alleles than non-African populations and appear to have lower levels of linkage disequilibrium (LD). These patterns have also been interpreted to support an older (ancestral) or larger African population (Przeworski, Hudson, and Di Rienzo 2000; Reich et al. 2001).

There is accumulating evidence regarding population structure among African populations based on DNA sequences from mtDNA (Chen et al. 1995; Watson et al. 1997; Chen et al. 2000; Destro-Bisol et al. 2004a; Destro-Bisol et al. 2004b; Wilder et al. 2004), the Y chromosome (Spurdle, Hammer, and Jenkins 1994; Underhill et al. 2000; Hammer et al. 2001; Underhill et al. 2001; Cruciani et al. 2002; Ke et al. 2002; Semino et al. 2002; Hammer et al. 2003; Knight et al. 2003; Wood et al. 2005), the X chromosome (Harding et al. 1997; Harris and Hey 1999b; Labuda, Zietkiewicz, and Yotova 2000; Garrigan et al. 2005a; Garrigan et al. 2005b), and autosomes (Nickerson et al. 1998; Zietkiewicz et al. 1998; Alonso and Armour 2001; Stephens et al. 2001; Yu et al. 2001; Gilad et al. 2002; Bamshad et al. 2003; Watkins et al. 2003). Some studies provide evidence of highly divergent lineages that could be the result of long-term population subdivision in humans (Harris and Hey 1999b; Labuda, Zietkiewicz, and Yotova 2000; Yu et al. 2001; Garrigan et al. 2005a; Shimada et al. 2007). For example, data from the X-linked *PDHA1* locus were the first to present clear evidence for ancient population structure that appears to have occurred before the dates for the earliest anatomically modern human fossils. Based on these data, Harris and Hey (1999) posited that the transition from early "archaics" to later modern humans occurred in a geographically subdivided ancestral population. Yu et al. (2001) also find evidence of ancient

geographic subdivision based on their study of a 10-kb region of chromosome 1; however, their African sample was small (10 individuals), and not directly comparable to that of Harris and Hey (1999). Though based on only one locus, Harding et al. (1997) and Hamblin et al. (2002) found possible evidence for ancient population structure (at ß-globin and the Duffy blood group locus, respectively), using much larger sample sizes; however, both of these loci are likely to be targets of natural selection in Africans. Thus, there may be a suggestion of ancient population subdivision in modern Africans, but this has not yet been confirmed by comparisons of data across multiple, neutrally evolving loci. No studies to date have directly compared patterns of DNA sequence variation within and among individual African populations, nor have they examined subdivision in modern Africans using multiple loci from different compartment of the human genome (e.g. the haploid regions *versus* autosomal and X chromosomal loci).

Ancient African Population Structure: Fossil Evidence

Until relatively recently, humans and their extinct ancestors have been placed in the family Hominidae and called "hominids". Here, I place humans in the Subfamily Homininae along with chimpanzees, based on recent molecular data suggesting that humans and chimps are more similar to each other than they are to other great apes (Patterson et al. 2006). I further place humans within the Tribe Hominini and refer to this group as "hominins", as is the modern genetics-based convention (chimpanzees are placed in the Tribe Panini). As there is no way to classify fossil hominins according to reproductive isolation (per Mayr 1963), one must operate under the Phylogenetic

(Cracraft 1983), Evolutionary (Simpson 1961), or Cohesive species concepts (Templeton, Otte, and Endler 1989). The concept of "species" requires some attention here, as I am interested in characterizing the genetic structure of early modern human populations in Africa, and understanding how populations were connected through gene flow (**CHAPTER 5** of this work). These three species concepts allow for the possibility of *interspecific* hybridization (a process that may have shaped the hominin fossil record), while Mayr's Biological Species Concept does not. In fact, according to the Evolutionary Species Concept, even though an evolutionary species is defined as "a lineage (an ancestral-descendent sequence of populations) evolving separately from others and with its own unitary evolutionary role and tendencies", there is still a possibility for interspecific hybridization, and "two species may interbreed to some extent without losing their distinction in evolutionary roles" (Simpson 1961, p.153). Early in hominin evolution, it is possible that a group of hominins (or morphospecies) whose range of variation does not exceed the range of variation of a closely related living species joined another such group in a hybrid zone and back-crossed (i.e., took novel gene combinations back to the parent populations). These novel gene complexes may have been deemed better or worse adapted than their predecessors (Barton 2001), and may have lead to splitting within a lineage, or cladogenesis. This scenario for early hominids is directly analogous to the study of the morphologically distinct *Papio* and *Therapithecus* genera (Jolly 2001). It is possible there were multiple demes of palaeospecies inhabiting the African landscape early in hominin evolution.

Whether the genus *Homo* is monophyletic or paraphyletic has yet to be resolved from the fossil data (Stringer 1987; Wood and Collard 1999; Leakey et al. 2001; Asfaw et al. 2002; Cameron 2003; White 2003). In particular, the taxonomic affinities of *Homo habilis* are highly contentious, as some researchers claim that representatives of this species can instead be divided into *Homo rudolfensis* and *Homo habilis* based on morphological characteristics (Wood and Collard 1999; but see Lee and Wolpoff 2005). They claim that these two roughly contemporary species should be incorporated into the late australopithecines, most likely making the genus *Australopithecus* paraphyletic. A similar debate centers around *Homo erectus* (Andrews 1984; Walker and Leakey 1993). Researchers have suggested that the *Homo erectus* hypodigm can be differentiated into early African *Homo erectus*, African *Homo ergaster*, and Asian *Homo erectus* (Wood and Collard 1999). It appears that *Homo erectus* may be represented by demes of individuals who are geographically widespread and morphologically variable. Importantly, interbreeding between these demes is a possibility that has not been eliminated. Moreover, recent work has made it unclear whether the species *Homo ergaster* existed in Africa at all (Spoor et al. 2007). All that remains clear is that there is a large amount of morphological variation in African hominins during the Late Pliocene.

As we move closer in time to the present, the picture becomes no more lucid. Both archaeological evidence from South Africa (Henshilwood et al. 2002) and paleontological evidence from Ethiopia (Haile-Selassie, Asfaw, and White 2004) implicate Africa as the birthplace of AMH. However, there is evidence for an earlier lineage of transitional hominins called "archaic" *Homo sapiens* just prior to the earliest

AMH fossils from the Levant and Africa (100-200 kya), (Grun, Stringer, and Schwarcz 1991; Clark et al. 2003; McDougall, Brown, and Fleagle 2005). These hominins are widely overlooked by geneticists studying the origin of anatomically modern humans. The remains are described as being mosaic forms of the more robust *Homo erectus* and the more gracile modern *Homo sapiens*, and they are often characterized as having a broad array of anatomical features. Species that are often included in the "archaic" *Homo sapiens* group in Africa include late *Homo ergaster*, late *Homo erectus*, *Homo heidelbergensis*, and *Homo rhodesiensis*. They are represented by fossils from numerous sites in South, East, and North Africa as well as Europe and Asia (a testament to the wide geographic range occupied by these transitional hominins), and are generally dated to be between 125,000 and 600,000 years old (Santa Luca 1978; Clark et al. 1994). The archaeological evidence associated with these remains further reinforces the idea that they are truly intermediary forms between the earlier *Homo erectus* and later completely modern *Homo sapiens*. Remnants of the African Early, Middle, and Late Stone Age, Acheulean, and Aterian industries are associated with these remains, as well as some tools that have been described as "levallois-like" from Jebel Irhoud. Thus, the fossil record leaves ample room for the possibility that the earliest AMH populations were highly subdivided. Moreover, it is possible that modern humans were evolving in an environment where morphologically variable, contemporaneous hominins were potentially exchanging genes.

GENOMIC REGIONS OF INTEREST

Rationale for Loci Choice

A systematic analysis of genetic diversity and estimates of population subdivision in modern African populations can serve as a window into the past. Here, I conducted a systematic survey of DNA sequence data from the mitochondrial DNA *COIII* (*cytochrome c oxidase subunit 3*) gene, the non-recombining portion of the Y chromosome, and two genes on the X chromosome (*PDHA1* and *RRM2P4*). These loci have been demonstrated to be evolving neutrally in African populations, i.e., are apparently unlinked to sites under selection (Harris and Hey 1999b; Hammer et al. 2004c). Single copy genes were chosen to avoid spurious PCR amplification and sequencing errors. Differences in effective population sizes mean that the expected time depth of X chromosomal diversity will be three times deeper than that for the mtDNA and the NRY. The average mtDNA mutation rate is estimated to be much faster than those estimated from bi-parental nuclear loci, thus making it more suitable for tracing more recent population history, while the X-linked loci allow a deeper window for the examination of more ancient demographic signals. Below I give background information for each locus including its genomic context which is followed by specific information from previous studies of the locus concerning human evolution. A schematic of each locus is included in **CHAPTER 2** of this work.

Mitochondrial DNA (mtDNA)

Mitochondrial DNA is probably the most widely studied region of the human genome. MtDNA is found in the mitochondria, differing from autosomal, X, and Y chromosomal DNA which are found within the nucleus of the cell. Its key role is in the generation of energy in the form of adenosine triphosphate (ATP) (Ingman and Gyllensten 2001). The circular mtDNA genome is made up of roughly 16,568 basepairs, encoding 13 polypeptides, 22 transfer RNAs (tRNAs), 2 ribsomal RNAs (rRNAs), and a hypervariable control region (the D-loop) that regulates replication and transcription. Each mitochondrial compartment within the cell has approximately 0-15 copies of the mtDNA genome, meaning that the entire cell contains around 100-1000 copies of the mtDNA genome (Cavelier, Johannisson, and Gyllensten 2000). MtDNA is found mainly within the cytoplasm of the egg and in fewer numbers in the midpiece of the sperm (5-10 mitochondria). Although heteroplasmy (different nucleotides at the same site in a single individual) does occur, it happens relatively rarely (Parr et al. 2006).

MtDNA has several notable features: it exists in high copy numbers (Brown, George, and Wilson 1979), for the most part it does not undergo recombination (Olivio et al. 1983), and it is usually inherited solely through the mother (Gyllensten, Wharton, and Wilson 1985). The high copy number facilitates the discovery and amplification of mtDNA in both extant and fossil specimens, and the uniparental inheritance allows the evolutionary history of maternal lineages to be more easily constructed. The lack of recombination means that changes in the mitochondrial DNA take place only through mutation and not through recombination.

It has been argued that the lack of recombination and uniparental mode of inheritance of the mtDNA makes it more susceptible to the effects of natural selection since many slightly deleterious mutations can still exist within the population under purifying selection (Ingman and Gyllensten 2001). It has also been hypothesized that mtDNA variation might be due to adaptations to cold climatological conditions (Mishmar et al. 2003; Ruiz-Pesini et al. 2004; Mishmar et al. 2006; but see Sun, Kong, and Zhang 2007). Such a process might result in an excess of non-synonymous mutations as compared to synonymous mutations. However, it has been demonstrated that such an excess also occurs in African populations found in warm climates as well (Kivisild et al. 2006). There is support for background selection or selective sweeps on the mtDNA acting to increase the frequency of rare alleles (Nachman, Boyer, and Aquadro 1994; Rand, Dorfsman, and Kann 1994; Nachman et al. 1996; Hey 1997; Harpending et al. 1998; Elson, Turnbull, and Howell 2004).

The substitution rate of the mtDNA is approximately five times higher than that of nuclear DNA (Brown et al. 1982). The substitution rate for the entire mtDNA molecule (that is, excluding the D-loop) is estimated to be 1.70×10^{-8} substitutions per site per year, based on a divergence from chimpanzee of 5 Myr (Ingman et al. 2000). The mutation rate of the D-loop is highly variable, but is thought to be higher than the rest of the mtDNA. For this reason, nearly 30% of the polymorphic sites found in a large global sample were found to be located in the D-loop even though it comprises only 7% of the mtDNA genome (Ingman and Gyllensten 2001).

Our understanding of human genetic diversity and history has been greatly influenced by early studies of human nucleotide variation focused on mtDNA. Almost all of these studies were based on the hypervariable regions. However, these hypervariable regions may not be the most ideal for addressing questions concerning human demographic history because the increased mutation rate most likely has led to an abundance of sites where recurrent mutation has taken place, thereby obscuring phylogenetic analyses (Ingman et al. 2000).

One of the earliest papers by Cann, Stoneking, and Wilson (1987) used restriction mapping which yielded one sub-Saharan African cluster and one cluster of non-sub-Saharan individuals. Based on these data, Cann et al. (1987) concluded that Africa was the most likely source of ancestor of all female lineages, a claim that was further supported by subsequent work by Vigilant et al. (1991). The authors also estimated the Time of the Most Recent Common Ancestor (TMRCA) by sequence divergence from a reconstructed ancestor to be roughly 214,000 years (Cann, Stoneking, and Wilson 1987). In fact, most mtDNA trees coalesce between 150,000 and 250,000 years ago (Vigilant et al. 1991; Rogers and Harpending 1992; Goldstein et al. 1995; Horai et al. 1995; Reich and Goldstein 1998; Ingman et al. 2000; Ingman and Gyllensten 2001). Studies of mtDNA have also greatly affected our views of population size change in humans, most of which support a population expansion that occurred sometime during the Pleistocene (Rogers and Harpending 1992; Harpending 1994; Sherry et al. 1994; Relethford 1998; Excoffier and Schneider 1999; Schneider and Excoffier 1999).

Here I chose to examine the mtDNA *COIII* locus because it harbors a large number of polymorphisms in humans, while it is not highly divergent from chimpanzee. The *COIII* region contains less homoplasy than the D-loop region of the mtDNA because of its lower mutation rate, thereby allowing us to more accurately reconstruct past demographic events. The mutation rate of the *COIII* region was estimated to be 1.25×10^{-8} substitutions per site per year by Ingman and Gyllensten (2001), and to be 1.58×10^{-8} substitutions per site per year by Wilder, Mobasher and Hammer (2004).

The Y Chromosome (NRY)

Like mtDNA, the Y chromosome (for the most part) also does not undergo recombination. The NRY (the non-recombining portion of the Y chromosome, also called the MSRY, or Male Specific Region of the Y chromosome) is inherited solely from the father, and is thus useful for tracing paternal descent. The human X and Y chromosomes co-evolved from a pair of autosomes in a mammalian ancestor approximately 300 million years ago (Lahn and Page 1999). The Y chromosome is very small, depauperate of diversity and contains very few genes in comparison to the sexually recombining autosomes from which it evolved. For this reason it is often referred to as "degenerate", and it has been hypothesized that it will cease to contain functional genes within 10 million years (Aitken and Marshall Graves 2002). However, about one half of the Y chromosome consists of tandemly repeated microsatellites, and there may be some evidence to suggest that gene conversion in these palindromic regions delays the decay of the Y chromosome which occurs because of a lack of recombination (Graves 2004). It is

also possible that selective sweeps on the NRY could dramatically influence patterns of variation (Malaspina et al. 1990; Dorit, Akashi, and Gilbert 1995; Haussler et al. 1995; Jaruzelska, Zietkiewicz, and Labuda 1999; Pritchard et al. 1999). In particular, selective sweeps can act to reduce variation on the NRY.

Like mtDNA, early work on the Y chromosome has also greatly shaped our understanding of human genetic diversity and history, though to a somewhat lesser degree because these studies were preceded by those on mtDNA. Past studies of the NRY in humans have yielded results similar to those of mtDNA, but have also added new and interesting information as well. Like mtDNA data, NRY data support an African origin for modern humans, but they also provide support for a back migration to Africa from Asia (Templeton 2002). Like mtDNA, many studies of the NRY provide support for a population size increase within the Pleistocene (Pritchard et al. 1999; Thomson et al. 2000). However, it is interesting to note that the estimates of the Time to the Most Recent Common Ancestor (TMRCA) for the Y chromosome are generally half those of mtDNA (Wilder, Mobasher, and Hammer 2004) It has been hypothesized that this may be due to differences in the effective population sizes of males and females (Dupanloup et al. 2003; Wilder, Mobasher, and Hammer 2004).

In the past, studies of the Y chromosome have focused on the use of highly informative SNPs (Single Nucleotide Polymorphisms). There may be a large degree of bias associated with the use of SNPs that have been discovered in one population, and then examined in another. Direct sequencing of the Y chromosome has been difficult in

the past since there is a lack of diversity on this chromosome (Yu et al. 2001). This has made the study of a large number of sites on the Y chromosome excessively costly.

Alu element insertions have been extremely useful in finding large amounts of genetic variation on the Y chromosome because of their large numbers of CpG dinucleotides (Labuda and Striker 1989). *Alu* elements are short interspersed elements (SINEs) recognized by the restriction enzyme *Alu*I and are usually ~300 bp long. They are the most numerous mobile element in humans, found ubiquitously throughout the genome in 3' untranslated regions, introns, and intergenic regions (Batzer and Deininger 2002). *Alu* elements have been evolving in primates for approximately 65 Myr, and can be hierarchically ordered from oldest (J) to intermediate (S) to most recently integrated (Y) (Batzer et al. 1996). Insertion positions of these retrotransposable elements are ideal for studies of human phylogenetics because they are apparently homoplasy-free (identical by descent) because the probability of an element inserting in the same location in the genome is exceptionally small (Batzer and Deininger 2002). It is important to note that the age of an *Alu* subfamily is inversely proportional to the number of polymorphic elements it contains; therefore, more recently integrated elements tend to harbor more variable sites (Roy-Engel et al. 2001).

The regions of non-coding DNA on the NRY studied here were chosen because they have slightly higher mutation rates than other sequences on the NRY. I analyzed *Alu*s from the following sub-families: Yα5, Yc2, Yd2, and Y **(TABLE 1.1)**. These regions were chosen because they were young enough to harbor variation, but old enough to be present in the outgroups of interest (chimpanzee, gorilla, and orangutan), which at their earliest

stage diverged from humans over 18 MYA. However, two of the *Alu* insertions chosen (16e4 and 486) were not present in the chimpanzee outgroup (and thus were not present in gorilla or orangutan either), but were still studied since they harbored substantial variation in extant humans. Loci were also chosen to be in introns in single copy genes, and the insertion had to be fixed in the samples. The mutation rates for these regions were determined by comparison to the outgroup (chimpanzee) data, using a divergence time of 6 MYR. Here I used the substitution rate of 4.19×10^{-9} substitutions per site per year (Wilder and Hammer 2007a).

The X Chromosome

While the mtDNA and NRY do not undergo normal recombination, most of the X chromosome does in females. Here, I investigated two independent, X-linked loci, Ribonucleotide Reductase M2 Polypeptide Pseudogene 4 (*RRM2P4*), and Pyruvate Dehydrogenase Alpha 1 (*PDHA1*). Since males are hemizygous for the X chromosome, there is no need for cloning to ensure the correct phasing of haplotypes when males are investigated using these loci.

Ribonucleotide Reductase M2 Polypeptide Pseudogene 4 (RRM2P4)

The *RRM2P4* locus (Xq27.3) was chosen because it is a pseudogene, and as such should not be subject to selection. Moreover, it maps to a gene poor telomeric region on the X chromosome that experiences high rates of recombination; thus, it is not likely to be linked to sites under selection. Although it maps to a region of high recombination on

the X chromosome, the 2.5 kb *RRM2P4* sequence did not show any evidence for intragenic recombination based on a screen of direct sequence data from 41 globally sampled individuals (Hammer et al. 2004c).

The *RRM2P4* locus is of interest for many reasons, most importantly because it has an unusually deep TMRCA (~2 MYA) making it ideal for a study of ancient population structure. Also, unlike most loci that have been studied, it appears to have its root in East Asia and not Africa and diversity in non-Africans appears to be higher than that in Africans. A second study of the locus expanded the sampling of a single nucleotide site (site 2020, G/A) in 570 Africans and non-Africans, and supported the possibility of ancient admixture between archaic Asian lineages and modern *Homo sapiens* (Garrigan et al. 2005b). Using outgroup data, this study estimated the substitution rate for the locus to be 7.4×10^{-10} substitutions per site per year based on a 6 MYR divergence from chimpanzee. Here I directly sequenced approximately the same region as the Hammer et al. 2004 study, in total comprising ~2.5 kilobases of sequence data.

Pyruvate Dehydrogenase Alpha 1 (PDHA1*)*

The *PDHA1* locus is also located on the X chromosome (Xp22.1). The gene codes for a subunit of the Dehydrogenase enzyme, thus, unlike the *RRM2P4* locus, mutations in this locus can cause disruption of the function of the gene and can lead to disease (Robinson et al. 1996). Therefore, it is possible that selection at this locus can influence the portrait of demographic history observed. *PDHA1* was chosen because, like *RRM2P4*, previous work had shown that it exhibits an unusually deep TMRCA (1.86

MYR) and does not show much evidence of intragenic recombination, making it ideal for studies of ancient population structure (Harris and Hey 1999b).

Unlike *RRM2P4*, it has been demonstrated that Africa is the most likely place of the most recent common ancestor for this locus (Harris and Hey 1999b). Previous studies of *PDHA1* have suggested that nuclear data and mtDNA data do not support similar demographic histories concerning human origins; that is, the mtDNA *Hypervariable I* region data supported a model of population expansion while *PDHA1* (like many other X-linked datasets) did not (Hey 1997). Previous studies also indicated the possible effects of balancing selection, since the locus appears to harbor more diversity than most X-linked loci, with an excess of intermediate frequency mutations (Hey 1997).

A subsequent study by Harris and Hey (1999b) found a fixed nucleotide difference between Africans and non-Africans, and low nucleotide diversity in non-Africans. The authors estimated the time of onset of population subdivision between Africans and non-Africans to be 200 KYR. Because this pre-dated the appearance of the first anatomically modern humans in Africa (at the time the article was written), it was assumed that the transition from archaic *Homo sapiens* to anatomically modern humans took place in a population that was already geographically subdivided. This is a point that needs to be considered in concert with the authors' estimate of a very small effective population size for this locus.

In this study I directly sequenced approximately 4.2 kb of mostly non-coding DNA from the *PDHA1* locus. Using outgroup data, the substitution rate was estimated to be

9.7×10^{-10} substitutions per base per year based on a 5 MYR divergence from chimpanzee (Harris and Hey 1999a).

BACKGROUND ON POPULATIONS USED IN THIS STUDY

The sample consists of 25-50 male individuals from each of five African populations (samples numbers change depending on the locus surveyed). Samples come from the Dogon of central Mali, a group of Southeast Bantu speakers from southern Africa, the Khoisan of southern Africa, the Bakola of southwestern Cameroon, and the Dinka of southern Sudan (**FIGURE 1.1**). All samples were collected with the permission of and according to the protocol of the Human Subjects Review Board at the University of Arizona (Project #:A93.05). For outgroup comparisons and mutation rate estimates, orthologous regions in the common chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), and orangutan (*Pongo pygmaeus*) were also sequenced.

Populations were chosen based on their geographic locations, language affiliation, and cultural customs. The western, west-central, southern, and eastern African population samples represent three of the four major linguistic groups of Africa: Nilo-Saharan (the Dinka), Niger-Congo (the Dogon, Bakola, and Southeast Bantu), and Khoisan. Among the five African populations two groups are hunter-gatherers (the Khoisan and Bakola) and three groups are food-producers (the Dinka are pastoralists, the Dogon practice casual cultivation, and the Southeast Bantu speakers practice varying degrees of agricultural cultivation) (Murdock, 1967). Thus, the five populations represent a large portion of the linguistic and subsistence strategy diversity present in modern African

populations. The following section is meant to give a brief background sketch for each of these groups, including their linguistic affiliation, archaeological history, physical description, subsistence strategies, population census sizes (if known), and marital structures, so as to foster a greater understanding of factors affecting their patterns of genetic variation.

The Khoisan of Southern Africa

The Khoisan can be thought of as the "aboriginal" or early inhabitants of southern Africa (Lee 1976; Cavalli-Sforza, Menozzi, and Piazza 1994). The term Khoisan is actually a combination of two ethnic groups, the Khoi (also Khoekhoe or Hottentot) and the San (also *Sonqua* or Bushmen). It is known from archaeological sites containing pottery and domesticated animals that the Khoi were cattle-herding people (Klein 1986), while the San have traditionally been hunter-gathering people. There is archaeological evidence for the occupation of the western Cape of South Africa by the ancestors of the San as early as 4300 years B.P. (before present). These early archaeological sites are located along the beaches of the western Cape, and provide strong evidence that the occupants were hunter-gatherers (Parkington 1987). Upon the arrival of the Khoi peoples to the region around 2000 B.P. there is evidence of a shift in the San subsistence strategy from one largely based on hunting to one heavily invested in gathering, and rockshelters are also incorporated at this time (Manhire 1987).

The Khoi and the San are thought to be closely related since they share many biological features such as steatopygia (large fat deposits near the buttocks (Nurse, Weiner, and Jenkins 1985)) and "the tablier" (an elongation of the labia minora (Tobias 1978)). The two groups also share the use of clicks in their closely related languages. It has been speculated that the transfer of certain aspects of the click language from the Khoisan (specifically the Khoi) to the Nguni people (a Southeast Bantu group) was facilitated by the invasion of the Khoisan lands by the Nguni (Murdock 1959). The Nguni language has most likely incorporated some of the click sounds from the Khoisan language, probably due to intermarriage between the groups (Barnard 1992). Murdock hypothesizes that the invading Bantus married Khoisan women and thus the clicks were added to their Bantu language. The linguistic evidence for this is that many Khoi terms for herding and cattle have been incorporated into Nguni (Louw 1979). The genetic evidence also seems to support gene flow between the Khoisan and Southeast Bantu populations (see below).

Genetic differences do exist between the Khoi and the San, especially regarding lactase persistence. The Khoi have a greater ability to digest lactose and have much higher gene frequencies of lactase persistence than the San-- almost 30% *versus* virtually 0% (Cavalli-Sforza, Menozzi, and Piazza 1994). These gene frequencies most likely reflect a longer history of pastoralism in the Khoi than the San. In general, the Khoisan exhibit very high F_{ST} values in comparisons with other populations and are said to differ from other sub-Saharan Africans more than any sub-Saharan African group differs from another sub-Saharan African group (Cavalli-Sforza, Menozzi, and Piazza 1994, p.175).

The Khoisan are genetically most similar to the surrounding Southeast Bantus (and next most closely related to the Central-Eastern Bantus), and show evidence of gene flow in both directions with the Southeast Bantus, based on the GM marker. Since natural selection could be affecting this immunoglobulin, work in this area needs to be further substantiated. In this study we examine this relationship (as well as the relationships of the Khoisan to other African populations) using neutrally evolving, unlinked markers.

It is estimated that the Khoisan number 146,000 across all the countries they inhabit (mostly Namibia and South Africa) (Gordon 2005). They generally practice bilateral descent and are ambilocal (meaning they take up residence with both the maternal and the paternal sides of the family). The Khoisan practice exogamy and are mostly monogamous (Murdock 1959; Murdock 1967). If polygyny occurs, it is generally sororal (meaning that the wives are sisters) (Murdock, 1959).

The Southeast Bantu Speakers of Southern Africa

The Southeast Bantu language group is part of the Niger-Congo language family (Benue-Congo subgroup) and comprises the Nguni, Sotho, Tswana, Venda, Tsonga, and Tonga peoples. The Southeast Bantu languages differ considerably from Shona, which is considered to be South-Central Bantu (Smith 1973). My population sample is made up of individuals who are Swazi, Ndebele, Xhosa, Zulu, Tswana, Sotho, Pedi, and Tsonga. The Southeast Bantu language group is thought to have originated in the northeastern Transvaal of South Africa (Smith 1973). In comparison to the lighter skinned Khoisan people of southern Africa, many Bantu populations tend to have darker pigmentation,

possibly indicating the recent expansion (within the last 5000 years) from a more northern location. These movements (often referred to as the Bantu Expansions) are discussed in more detail earlier in this chapter. There is archaeological evidence supporting a Bantu occupation of the western part of Central Africa by 1000 B.C. (Vansina 1984). There appears to have been a western and an eastern stream of migration to the south, and each has archaeological evidence (mostly pottery) to support its existence (Philipson 1977).

The Southeast Bantu group can be divided into the Nguni peoples and the Sotho peoples. The Nguni group of South Africa, Zimbabwe, and Swaziland numbers over 18 million and consists of the Xhosa, Zulu, Ndebele, and Swazi (Marks and Atmore 1970). The isiXhosa and isiZulu languages are similar enough to be thought of as dialects. Nguni are generally exogamous and marriage within the group is strictly forbidden. The Nguni are also patrilineal and mostly monogamous-- although polygyny does rarely occur (Murdock 1959). Until relatively recently the Nguni have been agriculturalists and cattle herders (Murdock 1967).

The Sotho group consists of the Southern Sotho, the Northern Sotho (Pedi), and the Tswana. The seSotho languages, spoken by more than ten million people in southern Africa, do not contain any click sounds. There are about nine million Sotho and two million Tswana in South Africa, and another 1.3 million Sotho in Lesotho. Most of these peoples have traditionally been herders of cattle, goats, and sheep and cultivators of grain and tobacco, although many started working in the mines in South Africa starting in the 19th century. Like the Nguni, Sotho groups are patrilineal. However, the Sotho marriage

customs differ markedly from the Nguni in that they are generally endogamous, meaning that the marriage usually takes place within a patriline (Murdock 1959).

The Dinka of the Sudan

The Dinka are a Nilotic people who make up the largest ethnic group in the Sudan. They inhabit the south of the country, numbering over two million people, and speak a Nilo-Saharan language. The Dinka are renowned for their tall stature and very dark complexion. They have a long history of contact with Arabs (Deng 1972), and this relationship has not always been harmonious, as evidenced by the civil unrest which we are witnessing today in the Sudan.

Genetically, the Nilo-Saharans are quite similar to the Bantus—specifically the Central-Western and Southwestern Bantus (Cavalli-Sforza, Menozzi, and Piazza 1994). This is expected since the Bantu expansions had an eastern and southern route that would have allowed for genetic admixture to occur. Based on protein polymorphism data, Cavalli-Sforza, Menozzi, and Piazza (1994) note a considerable amount of genetic admixture between the Bantus and the Nilotics. They also have an unusually high frequency of lactase persistence for an African population, due to their apparently long history of pastoralism. The Dinka are not nomadic but rather live in fixed settlements, practicing pastoralism, agriculturalism, and fishing as well. The long-standing relationship with cattle is evident -- they are closely tied to their cattle, and many aspects of their lives reveal this association. Many of their cultural traditions revolve around their respect for their cattle.

The Dinka peoples are mostly monogamous although polygyny is not forbidden. Dinka groups are exogamous and marriage to first and second cousins is forbidden (Murdock 1967). During the first year of marriage groups are ambilocal (living at both maternal and paternal parents' residences), then become patrilocal after that (Murdock 1967). Kin groups are lineages whose core membership normally comprises residents of more than one community. There is no unilineal descent-- that is, they are not strictly patrilineal or matrilineal (Murdoch 1967).

The Dogon of Mali

The Dogon (also called Habe) are a population of West Africans found in what is present-day Mali, in the centrally located but relatively remote Bandiagara Cliffs. Their language (Dogon) is most definitely classified within the Niger-Congo language family, but its place within the family is much less clear. Some of the earliest work on this subject placed Dogon closest to Gur or Mande (Hochstetler and Durieux 2004), but more recent work has demonstrated that Dogon diverged from other members of the Niger-Congo family very early and positioned Dogon on its own branch of Niger-Congo (Williamson and Blench 2000). In fact, it was once thought that Dogon consisted of one single language; however, researchers have recently identified more than 17 varieties of Dogon, some of which are mutually unintelligible. Other languages in the region include Bambara and Fulfulde, though the two are held in somewhat lower esteem by the Dogon (Hochstetler and Durieux 2004).

It has been estimated that there are approximately 600,000 Dogon living in this area (Petit and Vandewalle 1991), though recent census data is not available. The Dogon are thought to have arrived here in the 18th century from Mande country in western Africa (Cazes 1986). They are surrounded by the Mossi and Bobo agriculturalists living to their south and the Bozo fishermen on the Niger River to their north. The Dogon practice casual cultivation in the sandstone cliffs where they dwell. They are predominantly agriculturalists, growing onions, millet, sorghum, rice, tobacco, and other vegetables, yet they also keep small numbers of livestock as well (Murdock 1967).

Their marriage customs are very similar to many other African groups. They are patrilocal, patrilineal, and like the Dinka, polygyny is accepted in their clans. Extremely high female fertility rates (~8.6 live births per female) are coupled with very high mortality rates (Strassmann 1992). Females use menstrual huts for five days during their menses, and it is apparent that the population has a very accurate understanding of a woman's reproductive availability (Strassmann and Warner 1998). The Dogon are endogamous and first cousin marriage is not forbidden (Cazes 1990). Each village, called a *ginna*, is made up of a patriline, generally inhabited by many brothers, their wives, and children. An early study by Cazes (1986) indicated that there is a strong degree of subdivision within the Dogon population, and he found evidence of four distinct "isolates" across the villages he studied. However at this time not much is known about Dogon demography based on population genetic work.

The Bakola of Cameroon

The Bakola (or Kola, Gyele, or Bagyele) are part of the Binga (Babinga, Babenga, or Yadinga) group of hunter-gatherers located in southwest Cameroon, and number approximately 3500 people (Loung 1981). The Bakola speak a Bantu dialect called the Kola/Mvumbo A80 dialect, which is also spoken by the Kwassio (which includes Ngoumba and Mabea). It is thought that all pygmies once spoke their own languages, but no relic of those languages exists (Murdock 1959; Blench and Spriggs 1999). If indeed the original language of pygmy groups was replaced, it is curious that such "substrates" such as terms referring to hunting and gathering did not persist in any form. Therefore, the idea of complete language replacement by neighboring agriculturalists has been called into question (Blench and Spriggs 1999).

All pygmies without exception speak the languages of the agricultural group to which they are bound. The Bakola appear to have an exceptionally close relationship with the Bantu-speaking Kwassio, with local oral tradition tracing this relationship back at least a century (Ngima Mawoung 2001). They share with one another common dance, food, and clothing traditions. They are also often associated with the Bongom (of Gabon), the Fang, Bassa, and Boulou (of Cameroon, Equatorial Guinea, and Gabon), and the Mvae (of Cameroon and Gabon) (Franqueville 1971). It has been recorded that in Cameroon the Bakola prefer to live among the Kwassio, and receive poor treatment from the Fang and Bassa and Boulou groups (Ngima Mawoung 2001). The Bakola also share clan names with the Kwassio, suggesting intermarriage between the groups, though the clan system and their names may have merely been borrowed (Loung 1981). Likewise,

marriage between the Bakola and the Bassa groups is almost non-existent, and a Bassa woman who has a child with a Bakola man is marginalized, as is the child.

Much is now known about the people of the nearby Ituri Forest in Zaire (now the Democratic Republic of the Congo), from the Ituri Forest Project (Bailey and DeVore 1989; Bailey et al. 1992; Bentley et al. 1999), though very little literature is available concerning the Bakola, save for research concerning the deforestation of their environment. The Bakola and their neighboring Bantu-speaking peoples share the heavily forested region of southern Cameroon; however, this region has been subjected to extreme deforestation that has seriously threatened the livelihood and the welfare of the native peoples. Although they are considered hunter-gatherers, they are partially settled— they are not as mobile as the Baka from Cameroon or the Aka from the Democratic Republic of Congo. Currently, the Bakola sell their goods, and especially the game they capture in the forest, to the Bantu-speaking peoples in the local markets. Yet, in the past the Bakola were, for all practical purposes, slaves of the Bantu-speaking peoples, inherited like land from father to son, beaten publicly, and humiliated (Ngima Mawoung 2001). In some places in Cameroon, this "traditional" type of relationship between the Bakola and the Bantu-speakers is still found today.

As they are pygmies, they can be characterized by their common physical attribute of short stature. They are thought to differ from other non-pygmoid groups in that they do not show an adolescent growth spurt and have lower levels of a human growth hormone receptor. African pygmies are genetically distinct from other sub-Saharans, although a substantial level of gene flow from Bantus to pygmies is detectable. Using protein

polymorphism data, Cavalli-Sforza estimated that the differences between pygmies and their closest African neighbors are large enough to have required at least 10–20,000 years of isolation (Cavalli-Sforza 1986). Human Leukocyte Antigen (HLA) variation seems to be higher in the Bakola than in other African pygmies in terms of the number of alleles and suggests intermingling in this case with other sub-Saharan populations (Bruges Armas et al. 2003).

According to Murdock, marriage is monogamous but polygyny is not strictly forbidden and does occur. Their social structure is based on bands that can number from 20-100 (Murdock, 1959). Residence is patrilocal and descent is patrilineal; however, patrilineality and patrilocality appear to be borrowed from the peoples near whom the pygmies live, and it is suspected that originally their descent was bilateral (Murdock, 1959). Most groups are exogamous and tend to choose mates that are from relatively far distances (30-50 km away) (Cavalli-Sforza, Menozzi, and Piazza 1994).

Comparative Non-Human Primates

The "molecular clock" theory, first proposed by Zuckerkandl and Pauling (1965), was based on the supposition that all molecular changes accrue in a clock-like fashion. In general, the molecular clock is calibrated using a known time of divergence and then used to estimate other unknown divergence times. Geneticists estimate rates of evolution based on the number of mutations or substitutions present in a sample of individuals and the known divergence times to then estimate divergence dates and ages of lineages.

In recent years it has been relatively commonplace to use a human-chimpanzee divergence of 5 MYR based on the fossil record of early hominins. However, this date has been pushed back to 6 MYR based on more recent fossil finds such as *Sahelanthropus tchadensis* (Brunet et al. 2005), *Orrorin tugenensis* (Senut et al. 2001) , and *Ardipithecus ramidus* (Haile-Selassie 2001). Here I have chosen to sequence one representative of chimpanzee (*Pan troglodytes*), one gorilla (*Gorilla gorilla*), and one orangutan (*Pongo pymaeus*). At present, estimates of divergence times for each of these taxa are as follows: human-chimpanzee 6.6 MYR (range 6.0-7.0), human-gorilla 8.6 MYR (range 7.7-9.2), and human-orangutan 18.3 MYR (range 16.3-20.8) (Steiper and Young 2006).

IMPROVEMENTS

Patterns of genetic variation among loci can be influenced in different ways by stochastic and selective processes. The interpretation of patterns of nucleotide variation among loci can be confounded by different sampling strategies employed by different investigators. This presents a formidable challenge: How does one disentangle the effects of population history from those of selection, and how does one control for sampling variance across multiple genetic loci? I address these questions in three ways, by (1) examining multiple, independent loci from the mtDNA, Y chromosome, and X chromosome to separate the effects of population history from selection, (2) using direct re-sequencing data to avoid ascertainment bias, and (3) using a population-based

sampling scheme rather than a grid-based (global) scheme to avoid heterogeneity generated when comparing different samples across loci.

Multilocus Comparisons across Populations

Multilocus approaches are relatively new and have only been employed successfully by a few researchers studying human populations (Ingman et al. 2000; Nachman and Crowell 2000; Frisse et al. 2001; Harris and Hey 2001; Stephens et al. 2001; Yu et al. 2001; Pluzhnikov, Di Rienzo, and Hudson 2002; Hammer et al. 2004b; Wilder et al. 2004), and more recently by researchers studying *Drosophila* (Glinka et al. 2003). However, the approach had not been used to systematically analyze African populations which are so critical to our understanding of modern human diversity. Using multiple loci, I employ statistical methods to tease apart the relative contributions of recurrent gene flow processes and historical range expansion events, as well as selective *versus* neutral processes, as factors generating observed patterns of variation at each locus.

Direct Re-Sequencing Data

Past studies of the NRY have been based mainly on pre-ascertained SNPs discovered in small global panels and subsequently used to screen larger numbers of individuals. This is problematic in that it has been fundamentally biased towards the recovery of common SNPs. It has also been difficult to compare directly across studies because of the differing types of data-- SNP, microsatellite, and re-sequencing data. Due to the uncertainty associated with microsatellite mutation models and the bias associated with

pre-ascertained SNPs, these types of comparisons are troublesome (Goldstein et al. 1996; Pritchard et al. 1999). For these reasons, current NRY datasets are insufficient to test hypotheses of population growth and sex-specific migration. I improve upon previous work by generating nucleotide sequence data for mtDNA, the Y chromosome, and the X chromosome.

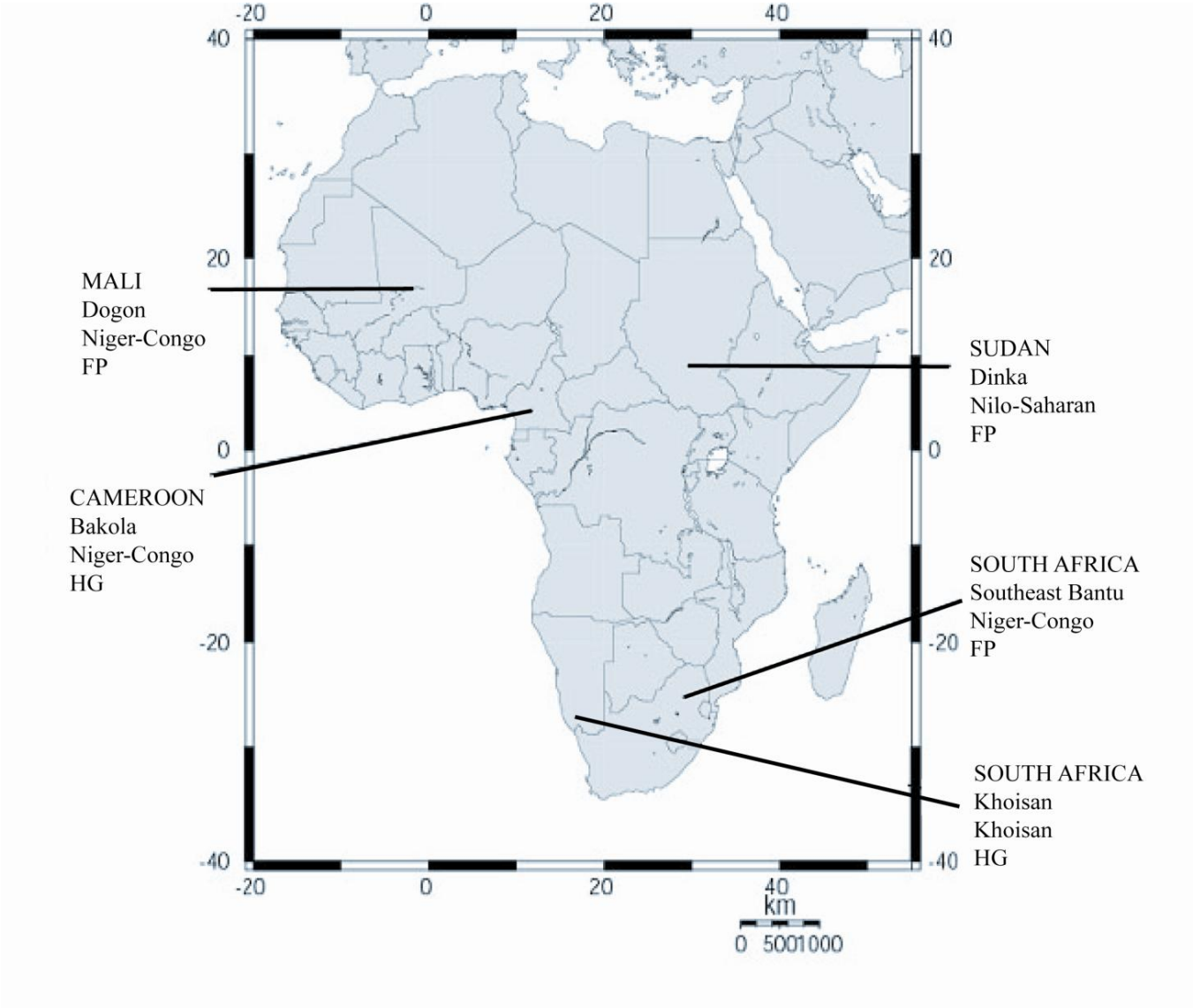
Population-Based Sampling

Finally, variation in the sampling schemes (global *versus* population-based) has made it extremely difficult to compare results across studies. Studies of mismatch distributions and frequency spectra are greatly affected by sampling strategy, and it is well-documented that surveys which sample only a few individuals from many localities tend to contain more rare alleles than surveys that sample many individuals from fewer localities (Ptak and Przeworski 2002; Hammer et al. 2003). A population-based sampling strategy avoids this bias by surveying many individuals from five distinct African populations. The use of population samples enables the estimation of African population coalescences, as well as their variances, which ultimately allows hypotheses about the structure of ancestral anatomically modern human populations to be tested.

TABLE 1.1 *Alu* names, families and gene regions used in this study.

ALU NAME	*ALU* FAMILY	GENE REGION
16e4	Yα5	SMCY
486	Yα5	DFFRY
DBYc2	Yc2	DBY
DBYd1	Yd2	DBY
AMELY	Y	AMELY
DFFRY30	Y	DFFRY
DFFRY40	Y	DFFRY
DFFRY50	Y	DFFRY
SMCY2	Y	SMCY
UTY62	Y	UTY
UTY87	Y	UTY
DFFRY04	Y	DFFRY
UTY83	Y	UTY

FIGURE 1.1 Map of populations examined in this study. The country of origin, group name, linguistic family, and major subsistence strategy (FP=food-producer, HG=hunter-gatherer) are shown for each.



CHAPTER 2: MATERIALS AND METHODS

OVERALL RESEARCH DESIGN

The overall research design was to amplify by polymerase chain reaction (PCR) and directly sequence 780 bp of the mtDNA *COIII* gene, 6.6 kb of non-coding DNA from the Y chromosome, ~2.5 kb of the X-linked *RRM2P4* pseudogene, and ~4.2 kb of the X-linked *PDHA1* locus in an African population panel which comprises five African populations: Dogon (DGN), Southeast Bantu (SEB), Khoisan (KHO), Bakola (BAK), and Dinka (DNK). Information concerning sample names, donors, and haplotypes for each locus are given in **TABLES 2.1A-E**.

Samples Used in this Study

In the past, different sampling strategies have complicated the interpretation of patterns of nucleotide variability among loci (e.g., grid versus population-based sampling) . For example, in order to distinguish the demographic signal of population structure from population growth one needs to sample thoroughly within populations (Hammer et al. 2003). The sampling strategy employed here minimizes sampling biases by surveying 25-50 individuals (depending on the locus) each from five distinct African populations. I compare patterns of variation among multiple loci, since any single locus represents only a single realization of a highly stochastic evolutionary process.

TABLE 2.1A Samples, population name, donor, and haplotype for the four loci under consideration.

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
GM3043	San	H. Soodyall	7	16	2	2
JR13	San	H. Soodyall	2	15	25	2
JR20	San	H. Soodyall	2	15	23	2
JR23	San	H. Soodyall	5	16	23	13
JR25	San	H. Soodyall	5	16	3	14
JR301	San	H. Soodyall	7	1	23	16
JR305	San	H. Soodyall	4	1	23	2
JR321	San	H. Soodyall	3	1	2	13
JR323	San	H. Soodyall	3	16	27	4
JR33	San	H. Soodyall	1	15	26	2
JR354	San	H. Soodyall	5	1	27	13
JR46	San	H. Soodyall	6	17	23	2
JR50	San	H. Soodyall	4	15	6	2
JR54	San	H. Soodyall	8	17	27	15
JR60	San	H. Soodyall	5	17	23	4
JR65	San	H. Soodyall	6	17	23	2
JR77	San	H. Soodyall	2	1	23	4
JR78	San	H. Soodyall	2	15	23	2
NAM16	San	H. Soodyall	7	18	1	4
NAM20	San	H. Soodyall	5	18	1	3
NAM24	San	H. Soodyall	3	18	23	4
NAM25	San	H. Soodyall	5	8	28	17
OM118	Khoi	H. Soodyall	5	18	27	4
OM127	Khoi	H. Soodyall	5	17	2	3
OM140	Khoi	H. Soodyall	5	4	2	4

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
ALB27	Southeast Bantu	H. Soodyall	9	1	6	2
ALB47	Southeast Bantu	H. Soodyall	5	4	2	4
ALB74	Southeast Bantu	H. Soodyall	10	4	5	2
ALB77	Southeast Bantu	H. Soodyall	11	11	23	4
NDE13	Southeast Bantu	H. Soodyall	11		33	2
PED53	Southeast Bantu	H. Soodyall	5	1	36	1
PED57	Southeast Bantu	H. Soodyall	12	4	10	4
RN21	Southeast Bantu	H. Soodyall	12	1	29	2
RN22	Southeast Bantu	H. Soodyall	13	6	13	4
RN23	Southeast Bantu	H. Soodyall	14	8	17	3
RN24	Southeast Bantu	H. Soodyall	5	1	30	3
RN25	Southeast Bantu	H. Soodyall	15	11	31	4
RN26	Southeast Bantu	H. Soodyall	16	8	29	2
RN27	Southeast Bantu	H. Soodyall	13	4	7	13
RN28	Southeast Bantu	H. Soodyall	12	1	31	2
RN29	Southeast Bantu	H. Soodyall	17	1	27	4
RN30	Southeast Bantu	H. Soodyall	18	1	5	4
RN31	Southeast Bantu	H. Soodyall	18	8	10	4
RN32	Southeast Bantu	H. Soodyall	10	4	6	4
RN33	Southeast Bantu	H. Soodyall	20	4	1	4
RN35	Southeast Bantu	H. Soodyall	21	4	7	4
RN37	Southeast Bantu	H. Soodyall	10	8	18	4
RN38	Southeast Bantu	H. Soodyall	10	6	32	2
RN39	Southeast Bantu	H. Soodyall	12	19	10	2
RN40	Southeast Bantu	H. Soodyall	12	1	23	4
RN41	Southeast Bantu	H. Soodyall	13	15	17	4
RN44	Southeast Bantu	H. Soodyall	12	1	3	4

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
RN45	Southeast Bantu	H. Soodyall	10	8	18	2
RN46	Southeast Bantu	H. Soodyall	22	8	23	12
RN47	Southeast Bantu	H. Soodyall	13	8	33	4
RN50	Southeast Bantu	H. Soodyall	1	4	34	4
RN51	Southeast Bantu	H. Soodyall	10	1	17	1
RN52	Southeast Bantu	H. Soodyall	10	4	8	4
RN53	Southeast Bantu	H. Soodyall	23	4	6	4
RN55	Southeast Bantu	H. Soodyall	12	8	17	4
RN56	Southeast Bantu	H. Soodyall	10	1	23	4
RN58	Southeast Bantu	H. Soodyall	10	6	1	2
RN60	Southeast Bantu	H. Soodyall	1	4	13	4
RN61	Southeast Bantu	H. Soodyall	17	8	13	4
RN62	Southeast Bantu	H. Soodyall	1	4	10	2
RN63	Southeast Bantu	H. Soodyall	5		35	4
RN67	Southeast Bantu	H. Soodyall	5	4	17	4
SOT58	Southeast Bantu	H. Soodyall	5	8	7	4
SWA61	Southeast Bantu	H. Soodyall	12	1	37	4
SWA64	Southeast Bantu	H. Soodyall	13	4	23	2
TSW26	Southeast Bantu	H. Soodyall	10		38	4
TSW30	Southeast Bantu	H. Soodyall	12	8	6	2
TSW48	Southeast Bantu	H. Soodyall	13	4	7	4
TSW54	Southeast Bantu	H. Soodyall	5		10	4
ZU44	Southeast Bantu	H. Soodyall	12	6	33	4
CMN001	Bakola pygmies	G. Destro-Bisol		1	1	1
CMN002	Bakola pygmies	G. Destro-Bisol	13	2	2	2
CMN007	Bakola pygmies	G. Destro-Bisol	24	3	3	3

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
CMN008	Bakola pygmies	G. Destro-Bisol	24	3	4	4
CMN009	Bakola pygmies	G. Destro-Bisol	13	3	2	3
CMN012	Bakol pygmies	G. Destro-Bisol	26	4	5	4
CMN013	Bakola pygmies	G. Destro-Bisol	13	4	4	2
CMN014	Bakola pygmies	G. Destro-Bisol	13	4	2	2
CMN016	Bakola pygmies	G. Destro-Bisol	13	4	6	4
CMN019	Bakola pygmies	G. Destro-Bisol	13	4	2	4
CMN030	Bakola pygmies	G. Destro-Bisol	13	5	6	2
CMN035	Bakola pygmies	G. Destro-Bisol	28	6	1	4
CMN036	Bakola pygmies	G. Destro-Bisol	27	7	7	4
CMN041	Bakola pygmies	G. Destro-Bisol	13	4	2	2
CMN042	Bakola pygmies	G. Destro-Bisol	13	8	3	5
CMN044	Bakola pygmies	G. Destro-Bisol	13	9	2	2
CMN046	Bakola pygmies	G. Destro-Bisol	28	2	1	3
CMN048	Bakola pygmies	G. Destro-Bisol	13	2		
CMN049	Bakola pygmies	G. Destro-Bisol	24	8	3	1
CMN065	Bakola pygmies	G. Destro-Bisol	13	1	5	1
CMN066	Bakola pygmies	G. Destro-Bisol	13	8		
CMN67	Bakola pygmies	G. Destro-Bisol	13	3	2	2
CMN68	Bakola pygmies	G. Destro-Bisol	25	3	1	6
CMN070	Bakola pygmies	G. Destro-Bisol	13	1	1	4
CMN073	Bakola pygmies	G. Destro-Bisol	13	7	8	6
DGN02	Dogon	B. Strassman		4		
DGN03	Dogon	B. Strassman	10	10	9	4
DGN04	Dogon	B. Strassman	10			
DGN06	Dogon	B. Strassman	13	4	2	7
DGN07	Dogon	B. Strassman	29	4		4

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
DGN09	Dogon	B. Strassman	10	11	10	4
DGN10	Dogon	B. Strassman	10	12		4
DGN11	Dogon	B. Strassman	13	4		
DGN12	Dogon	B. Strassman	10	12	3	4
DGN13	Dogon	B. Strassman	13	4		
DGN16	Dogon	B. Strassman	10	12	2	4
DGN17	Dogon	B. Strassman	30			
DGN19	Dogon	B. Strassman	13	12	2	4
DGN22	Dogon	B. Strassman	30	12	11	4
DGN26	Dogon	B. Strassman	13	12	12	2
DGN27	Dogon	B. Strassman	10	4	13	4
DGN29	Dogon	B. Strassman	10	12	2	4
DGN30	Dogon	B. Strassman	13	12	14	8
DGN31	Dogon	B. Strassman	13	6	1	9
DGN33	Dogon	B. Strassman	10	10	2	4
DGN35	Dogon	B. Strassman	13	4	2	4
DGN37	Dogon	B. Strassman		12	15	4
DGN39	Dogon	B. Strassman	13	4		4
DGN40	Dogon	B. Strassman	10	4	2	2
DGN41	Dogon	B. Strassman	13			
DGN42	Dogon	B. Strassman	10	12	2	4
DGN43	Dogon	B. Strassman	10	12	16	2
DGN44	Dogon	B. Strassman	10	12	2	4
DGN45	Dogon	B. Strassman	13	4	2	2
DGN48	Dogon	B. Strassman	13	13		
DGN50	Dogon	B. Strassman	31	4	7	4
DGN51	Dogon	B. Strassman	32	12	2	10

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
DGN52	Dogon	B. Strassman	10	4	10	4
DGN53	Dogon	B. Strassman	10	12		
DGN54	Dogon	B. Strassman	10	10	17	
DGN56	Dogon	B. Strassman	10	4	18	4
DGN57	Dogon	B. Strassman	10	12	1	2
DGN58	Dogon	B. Strassman	10	8	13	4
DGN59	Dogon	B. Strassman	10	4	19	4
DGN61	Dogon	B. Strassman	13	4	6	7
DGN62	Dogon	B. Strassman	10			
DGN63	Dogon	B. Strassman	13	4	2	
DGN65	Dogon	B. Strassman		4		
DGN66	Dogon	B. Strassman	10	14	3	4
DGN68	Dogon	B. Strassman	13			
DGN250	Dogon	B. Strassman	13			
DGN253	Dogon	B. Strassman	13			
DGN529	Dogon	B. Strassman	13			
DGN763	Dogon	B. Strassman	13			
DGN372	Dogon	B. Strassman	33			
DGN872	Dogon	B. Strassman	10			
DGN883	Dogon	B. Strassman	34			
DNK001	Dinka	T. Angoui	10	15		4
DNK002	Dinka	T. Angoui	10	6	7	4
DNK003	Dinka	T. Angoui	3	6	16	4
DNK004	Dinka	T. Angoui	35	15	7	2
DNK005	Dinka	T. Angoui	36	6	17	4
DNK006	Dinka	T. Angoui	10	6	20	2
DNK007	Dinka	T. Angoui	12	15	1	8

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
DNK008	Dinka	T. Angoui	10	15	13	4
DNK009	Dinka	T. Angoui	13	1	21	1
DNK010	Dinka	T. Angoui	13	6	2	11
DNK011	Dinka	T. Angoui	3	15	15	4
DNK012	Dinka	T. Angoui	37	15	22	2
DNK013	Dinka	T. Angoui	3	15	6	2
DNK014	Dinka	T. Angoui	13	6	23	4
DNK015	Dinka	T. Angoui	38	6	2	2
DNK016	Dinka	T. Angoui	39	4	23	2
DNK017	Dinka	T. Angoui	30	11	1	4
DNK018	Dinka	T. Angoui	13	1	23	12
DNK019	Dinka	T. Angoui	13	6	24	4
DNK020	Dinka	T. Angoui	13	6	2	
DNK021	Dinka	T. Angoui	22	1	7	
DNK022	Dinka	T. Angoui	10	6	2	4
DNK023	Dinka	T. Angoui	40	15	23	4
DNK024	Dinka	T. Angoui and M. Pilkington	41			
DNK025	Dinka	T. Angoui and M. Pilkington	42			
DNK030	Dinka	T. Angoui and M. Pilkington	43			
DNK031	Dinka	T. Angoui and M. Pilkington	10			
DNK032	Dinka	T. Angoui and M. Pilkington	44			
DNK034	Dinka	T. Angoui and M. Pilkington	10			
DNK035	Dinka	T. Angoui and M. Pilkington	3			
DNK036	Dinka	T. Angoui and M. Pilkington	37			
DNK037	Dinka	T. Angoui and M. Pilkington	36			
DNK038	Dinka	T. Angoui and M. Pilkington	3			
DNK039	Dinka	T. Angoui and M. Pilkington	44			

Sample	Population	Donor	Haplotypes			
			mtDNA	NRY	PDHA1	RRM2P4
DNK040	Dinka	T. Angoui and M. Pilkington	12			
DNK041	Dinka	T. Angoui and M. Pilkington	37			
DNK042	Dinka	T. Angoui and M. Pilkington	12			
DNK043	Dinka	T. Angoui and M. Pilkington	12			
DNK044	Dinka	T. Angoui and M. Pilkington	40			
DNK045	Dinka	T. Angoui and M. Pilkington	10			
DNK046	Dinka	T. Angoui and M. Pilkington	10			
DNK047	Dinka	T. Angoui and M. Pilkington	37			
DNK048	Dinka	T. Angoui and M. Pilkington	45			

TABLE 2.1B List of the haplotypes for mtDNA, related to **TABLE 2.1A.**

Haplotype Name	Haplotype
1	AACAGGTGAGAAAGCAACTGAGGTTATTACATTGCTGG
2	AACAGGTGAGAAAACAACTGAGGTTATTACATTGTTGG
3	AACAGGTGAAAAAGCAACTGAGGTTGTTACGTTGTTGG
4	AACAGGTGAGAAAGCAACTGAGGTTGTTATATTGTTGG
5	AACAGGTGAGAAAGCAACTGAGGTTATTACATTGTTGG
6	AGCAGGTGAGAAAGCAGCTGAGGTTATTACATTGTTGG
7	AACAGGTGAGAAAGCAACTGAGGTTATTACACTGTTGG
8	AACAGGTGAGAAAGCAACTAAGGTTGTTATATTGTTGG
9	AACAGGTGAGAAAGCAACTGAGGTTATTACATTGCTGA
10	GACAGGTGAAAAAGCAACTGAGGTTGTTACATTGTTGG
11	AACAGGTGAGAAAGCAACTGGAGTTATTATATTGTTGG
12	AACAGGTGAGAAAGCAACTGAGGTTATTATATTGTTGG
13	AACAGGTGAAAAAGCAACTGAGGTTGTTACATTGTTGG
14	AACAGGTGAGAAAGCAACTGAGGTTGTTACATTGTTGG
15	AACAGGTGAGAAAGCAACTGAGGTTATTACATTGTTGA
16	AACAGGTGAAAAAGCAACTGAGGTTGTTACATTGCTAG
17	AACAGGTGAAAAGGCAACTGAGGTTGTTACATTGTTGG
18	AACAGATAAAAAAGCAACTGAGGTTGTTACATTGTTGG
19	AACAGGTGAGAAAGCAATTGAGGTTGTTACATTGTTGG
20	AACAGGTGAGAAAGCAACTGAGGTTACTACATTGTTGG
21	AACAGGTGAGAGAGCAACTGAGGTTATTATATTGTTGG
22	AACAGGTGAAAAAGCAACTGAGGTTGTTACATTGTTAG
23	AACAGGTGAAAAAGCAACTGAAGTTGTTACATTGTTGG
24	AACAGGCGAAAAAGCAACTGAGGTTGTTACATTGTTGG
25	AACAGGTGAAAAAGCAACTGAGGTCGTTACATTGTTGG
26	AATAGGTGAAAAAGCAACTGAGGCTGTTACATTGTTGG
27	AATAGGTGAAAAAGCGACTGAGGCTGTTACATTGTTGG
28	AACAGGTGGAAAAGCAACTGAGGCTGTTACATTGTTGG
29	GACAGGTGAAAAAGCAACTGAGGTTGTTACATTATTGG
30	AACAGGTGAAAAAGTAACTGAGGTTGTTACATTGTTGG
31	GACAGGTGAAGAAGCAACTGAGGTTGTTACATTGTTGG
32	GACAGGTAAAAAAGCAACTGAGGTTGTTACATTGTTGG
33	GACAGGTGAAAAAGCAACTGAGGTTGCTACATTGTTGG
34	AACAAGTGAAAAAGCAACTGAGGTTGTTACGTTGTTGG
35	GACAGGTGAAAAAGCAACTGAAGTTGTTACATTGTTGG
36	GACAGGTGAAAAAACAACTGAGGTTGTTACATTATTGG
37	AACAGGTGAAAAAGCAACTGAGGTTGTTCCATTGTTGG
38	AACAGGTGAAAAAGCAACTGAGGTTGTTACATTATTGG
39	AACAGGCGAAAAAGCAACCGAGATTGTTACATTGTTGG
40	GACAGGTGAAAAAGCAACTGAGGTTGTCACATTGTTGG

41	GACAGGTGAAAAAGCAACTGAGGTTGTTACATTGTCGG
42	AACAGGTGAAAAAGCAACTGAGGTTGTTACATTGCTGG
43	AACAGGTGAAAAAGCAACCGAGATTGTTACATTGTTGG
44	AACAGATAAAAAAGCAACTGAGGTTGTTACATCGTTGG
45	AACGGGTGAGAAAGCAACTGAGGTTATTATATTGTTGG

TABLE 2.1C List of the haplotypes for NRY, related to **TABLE 2.1A**

Haplotype Name	Haplotype
1	CGGCTCTGCGTTTGATGAAGC
2	CGGCTCTGTATTCGAGGAGGC
3	TGGCTATGTGGCCGGTGAAGC
4	TGGCTATGTGTCCGATGAAGC
5	TGGCTAGGTGGCCGGTGAAGC
6	CGGCCATGTGTTCGATGAAGC
7	CAGCTCTGCGTTTGATGAAGC
8	TGGCTATGTGGCCGATGAAGC
9	TGGCTATGTGGCCTATGAAGC
10	CGGCTCTGCGTTTGATGAAGT
11	CGGCTATGTGTTCGATGAAGC
12	CGGCTATGTGTTCGATGAAAC
13	CGGCTCTTTGTTCGAGGAGGC
14	TGGCTCTGTGGCCGATGAAGC
15	CGGCTCTGTGTTCGAGAAAGC
16	CGGCTCTGTGTTCGAGGAAGC
17	CGGTTCTGTGTTCGAGGAAGC
18	CGGCTCTGTGTTCGAGACAGC
19	CGACTATGTGTTCGATGAAGC

TABLE 2.1D List of the haplotypes for *PDHA1*, related to **TABLE 2.1A**

Haplotype Name	Haplotype
1	CCCAGAGTGTTAGCACGCGTGAGAAGATCTATGCCGTGCGC
2	CCCCGACTGCCAAAATTGGCGAACAGACCCATGCCGTGCGT
3	CCCAGAGTGTTAAAACGCACGAGAAGATCTATGCCGTGCCC
4	CCCAGAGTGTTAAAACGCGCGAGAAGATCTATGTCGTACGC
5	CCCAGAGTGTTAAAACGCACGAGAAGGTCTATGCCGTGCCC
6	CCCAGAGTGTTAAAACGCGCGAGAAGATCTATGCCGTGCCC
7	CCCAGAGTGTTAAACCGCGCGAGAGGATCTATGCCGTGCCC
8	CCCAGAGTGTTAAAACGCGCGAGAAGATCTATGTCGTGCGC
9	CCCCGACTGCCAAAATTGGCGAGAAGATCTATATCGTGCGT
10	CCCAGAGTGTTAAAACGCGCGTGAAGATCTATGTCGTGCGC
11	CCCAGAGTGTTAGCACGCGTGAGAAGATGTATGCCGTGCGC
12	CCCAGAGTGTTAAAACGCACGAGAAGATCTATGCCATGTGC
13	CTCAGAGTGTTAAAACGCGCGAGAAGATCTATGTCGTGCGC
14	CCCAGAGTGTTAACACGCGTGAGAAGATCTATGCCGTGCGC
15	CCCAGAGTGTTAAAACGCACAAGAAGATCTATGCCGCGCCC
16	CTCAGAGTGTTAAAACGCGCGAGAAGATCTATATCGTGCGC
17	CCCAGAGTGTTAAAACGCGCGAGAAGATCTATGCCGTGCGC
18	CCCCAACTGCCAAAATTGGCGAACAGACCCATGCCGTGCGT
19	CTCAGAGTGTTAAAACGCGCGAGAAGATCTATGCCGTGCGC
20	CTCAGAGTGTTAAAACGCGCGAGAAGATCCATGTCGTGCGC
21	CCCAGAGTGTTAAAACGCACGAGAAGATCTATACTGCGCCC
22	CCCAGAGTCTTAAAACGCGCGAGAAGATCTATGCCGTGCCC
23	CCCAGAGTGTTAAAACGCGCGAGAGGATCTATGCCGTGCCC
24	CCCCGACTGCCAAAATTGGCGAACACACCCATGCCGTGCGT
25	CCCCGACTGCCAAAATTGGCGAGAAGACCCATGCCGTGCCC
26	CCCAGAGTGTTAAAACGCACGAGAAGATCTATGCCACGCCC
27	CCCAGAGAGTTAAAACGCACGAGAAGATCTATGCCGTGCCC
28	CCCAGAGTGTTTAAACGCGCGAGAAGATCTATGCCGTGCGC
29	CCCAGGGTGTTAAAACGCGCGAGAGGATCTATGCCGTGCCC
30	CCCAGAGTGTTAAAACGCACGAGAAGATCTATACCGTGCCC
31	CCCCGACTGCCAAAATTGGCGAGCAGACCCATGCCGCGCCC
32	CTCAGAGTGTTAAAACGCGCGAGAAGATCCATATCGTGCGC
33	CCCAGAGTGTTAAAACGCACGAGAAGATCTATGCCGCGCCC
34	CCCAGAGTGTTAAAACGCGCGAGAGGATCTACGCCGTGCCC
35	CCCAGACTGTTAAAACGCGCGAGAAGATCTATGCCGTGCCC
36	CCCAGACTGTTAGCACTCGTGAGAAGATCTATGCCGTGCGC
37	CCCAGAGTGTTAGCACGCGTGAGAAGATCCATGCCGTGCGC
38	GCTAGAGTGTTAAAACGCGCGAGAAGATCTCTGCCGTGCCC

TABLE 2.1E List of the haplotypes for *RRM2P4*, related to **TABLE 2.1A**

Haplotype Name	Haplotype
1	CGACTGCACAGCTGCC
2	CGATTGCACAGCTGCC
3	TGACTGCATGGCTGCC
4	CGACTGCATGGCTGCC
5	CGACTGCATGACTGCC
6	CCACTGCATGGCTGCC
7	TGACTGCATAGCTGCC
8	CGATTGCACGGCTGCC
9	TGACAGCATAGCGAAT
10	TGACAGCATAGCTGCC
11	CGATTGCATGGCTGCC
12	CGATTGCACAGGTGCC
13	CGACTGCATAGCTGCC
14	CGGCTGCATGGCTGCC
15	CGACTGCCCAGCTGCC
16	CGACTGTACAGCTGCC
17	CGATTACACAGCTGCC

Loci Used In This Study

Sequence data were generated from four independent loci, including the mtDNA *COIII* locus, non-coding portions of the NRY, the *RRM2P4* pseudogene, and a portion of the *PDHA1* gene (schematics presented below).

FIGURE 2.1 Schematic representations of loci amplified for this work. The mtDNA *COIII* locus is shown to the far left, regions from which the NRY *Alu* sequence data derive are shown in the middle, and the two X-linked loci are shown to the far right.

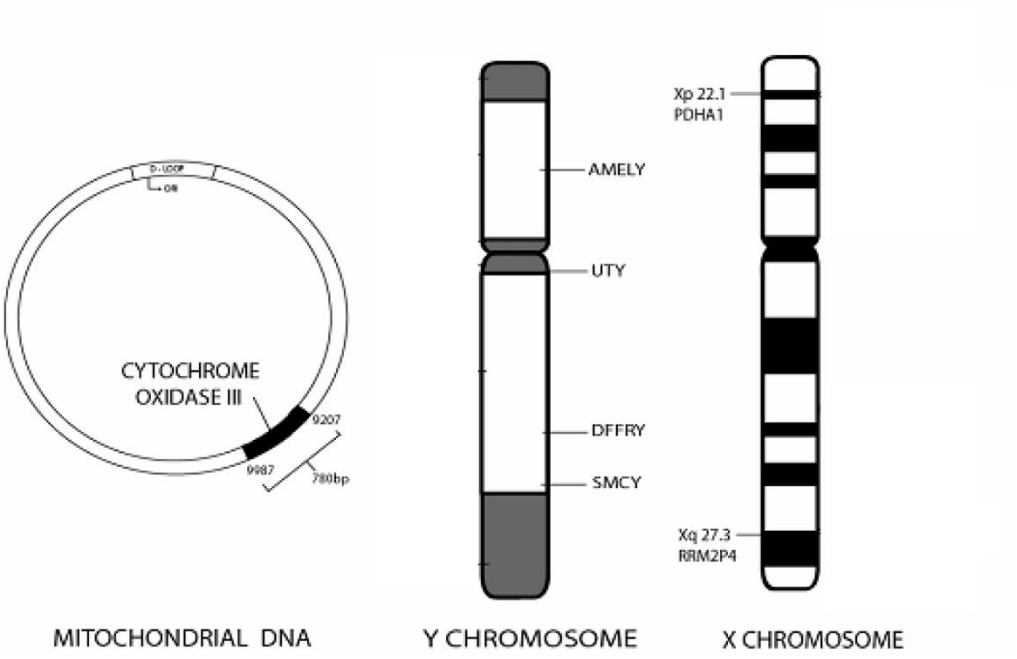


FIGURE 2.2 Schematic representation of *PDHA1*. Location of exons 7-10 marked in black boxes below the scale (in basepairs), amplification primers 1F/1R and 2F/2R and sequencing primers 1.1-2.4 below.

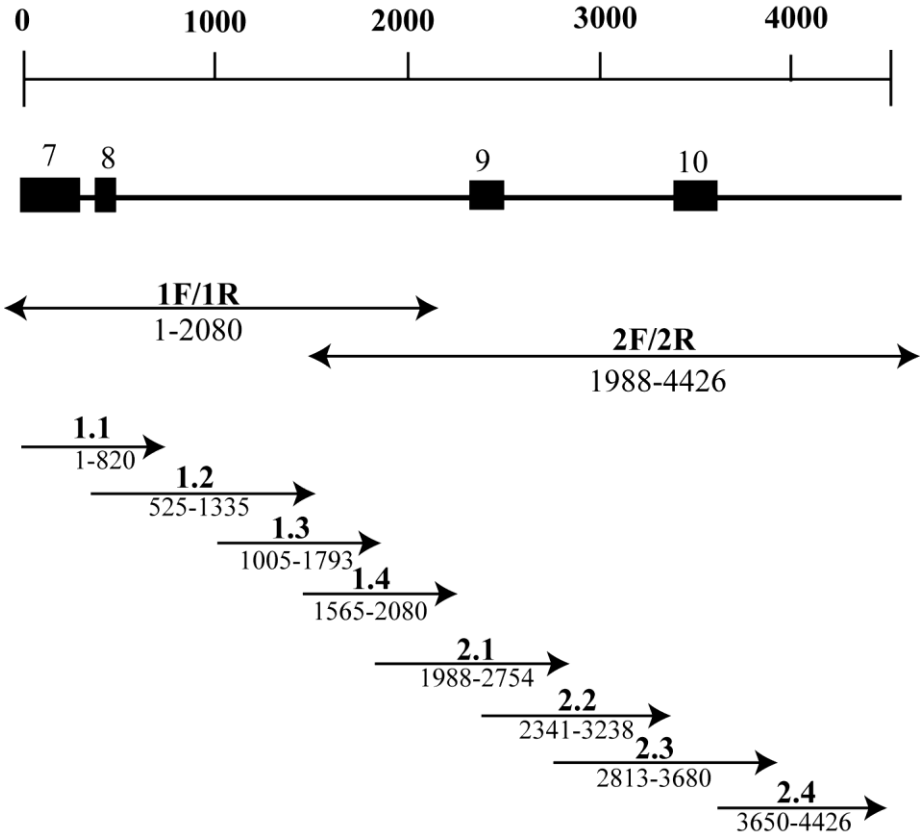


FIGURE 2.3 Schematic representation of *RRM2P4* pseudogene. Amplifying primers 1F/1R and sequencing primers 1F–4R shown below the scale (in basepairs).

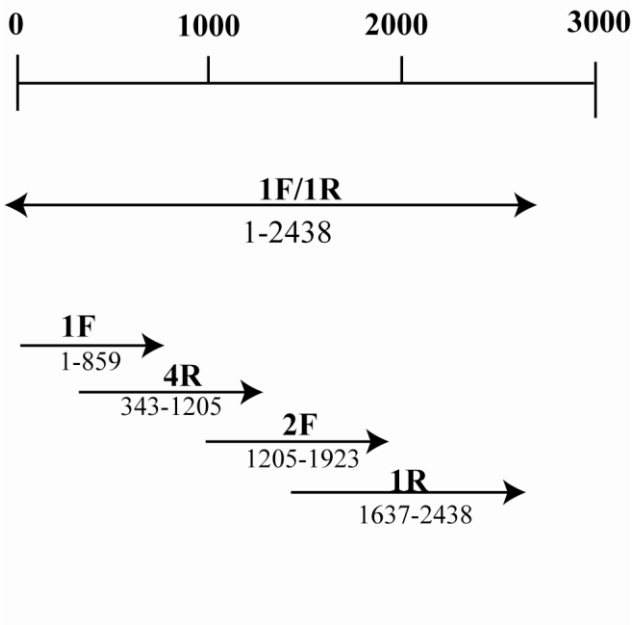


TABLE 2.2 Loci used in this study, alignment lengths and reference sequences for each population.

	COIII	**NRY**	**PDHA1**	**RRM2P4**
Location	mtDNA	Y chromosome	X chromosome	X chromosome
Alignment length	780 bp	6601 bp	1F-1R (2497 bp) 2F-2R (2439 bp) Total (4936 bp) Used 4559 bp	1F-1R (2414 bp)
Reference sequence (s) from Genbank	J01415 1…16569	Separate table	NM_000284 1…16052	AY694945 1…2387
Position in reference Sequence	9207…9987	**TABLE 2.03**	11281…15781	1…2387
KHO	25	25	25	25
SEB	50	46	50	50
DNK	37	23	22	21
DGN	49	40	32	32
BAK	24	25	23	23
Total number of individuals sampled	**185**	**159**	**152**	**151**

TABLE 2.3 NRY *Alu* nomenclature, family, Blat location (start and stop), and gene region.

Hammer Lab Name	Alu Name	Alu Family	Human Blat Start	Human Blat End	Region Information
16e4	Alu 16e4	Ya8	20,268,870	20,269,290	SMCY gene intron 14
486	Alu 486	Y	13,364,304	13,364,593	DFRY gene intron 31
DFFRY30	DRRFY I11	Y	13,288,401	13,289,063	DFFRY gene intron 11
AMELY	AMELY I2	AluSc	6,783,080	6,783,804	AMELY gene intron 2
DBYc2	DBYc2	Y	13,458,707	13,459,046	DBY gene intron 2
DBYd1	DBYc1	AluSg/x	13,457,672	13,458,259	DBY gene intron 2
DFFRY50	DFFRY I17	Y	13,324,140	13,324,551	DFFRY gene intron 17
DFFRY04	DFFRY I1	Y	13,257,814	13,258,333	DFFRY gene intron 1
DFFRY40	DFFRY i13	Y	13,307,361	13,307,906	DFFRY gene intron 13
SMCY2	SMCY I5	Y	20,290,644	20,291,049	SMCY gene intron 5
UTY83	UTY I19	Y	13,863,705	13,864,500	UTY gene intron 19
UTY62	UTY I10	Y	13,915,779	13,916,303	UTY gene inton 10
UTY87	UTY I19	Y	13,860,665	13,861,105	UTY gene intron 19

METHODS

Laboratory Procedures

DNA Extraction: Isolation of Genomic DNA from Buccal Swabs Using Phase Lock Gel (protocol from M. Kaplan)

Day 1

1. Use a sterile cytology brush to scrape cells from the inside of each cheek (approximately 1 minute per cheek). Place swab head in a 2.0 mL tube containing 650 uL of lysis buffer.
2. If the buffer volume, including the swab, is less than 750 uL, add up to 300 uL of H_20. Add 25 uL of 10mg/mL Proteinase K.
3. Incubate sample at 55˚ C (on an agitator if possible) overnight.

Day 2

4. Pour the sample into a 2mL phase lock "light" tube (Green) containing an equal (or excess) volume (650 uL) of saturated, pH adjusted (pH 8.0) phenol. Mix by rocking gently. Agitate 5 minutes.
5. Centrifuge 5 minutes at maximum speed to separate phases (keeping the swabs in tubes makes the separation of aqueous phase slightly more difficult, but has been shown to increase yield).
6. Pour the aqueous phase into a 2 mL tube of phase lock "heavy" gel (Yellow) containing an equal volume (650 uL) of $CHCl_3$:IAA (24:1). Agitate 5 minutes.
7. Centrifuge 5 minutes at maximum speed to separate the phases.
8. Pour the aqueous phase into a new tube.

9. Add 1/10 (75 uL) vol. of 3M NaOAc, and mix by rocking gently. Add 0.6 total volume of Isopropanol (500 uL), cold if possible. Invert the tube several times to precipitate the DNA.
10. Incubate at 0˚ C (on ice) overnight.

Day 3

11. Pellet DNA by centrifugation at maximum speed, 30 minutes.
12. Carefully decant the liquid and rinse the pellet with an excess (1000 uL) of 70% EtOH (cold if possible). Mix by rocking gently.
13. Pellet the DNA by centrifugation at maximum speed, 5 minutes.
14. Carefully decant the EtOH and rinse the pellet a second time with an excess (1000 uL) of 70% EtOH (cold if possible). Mix by rocking gently.
15. Pellet the DNA by centrifugation at maximum speed, 5 minutes.
16. Carefully decant the EtOH and rinse the pellet with an excess (1000 uL) of 95% EtOH (cold if possible). Mix by rocking gently.
17. Pellet the DNA by centrifugation at maximum speed, 5 minutes. Carefully decant the EtOH, then dry the DNA in a vacuum chamber with little or no heat overnight.
18. Resuspend the DNA in a 50 uL Low TE pH 8.0.

Reagents for DNA Extraction

Lysis buffer (with ½ x sucrose to reduce the density for better phase lock separation)	50 mM Tris pH 8.0 50 mM EDTA 25 mM Sucrose 100 mM NaCl 1% SDS
Proteinase K 10 mg/mL	
Low TE pH 8.0	10 mM Tris pH 8.0 0.1 mM EDTA pH 8.0
NaOAc 3M	Filter sterilize through 0.2 um cellulose acetate or nitrocellulose filter
EtOH (70%)	
EtOH (95%)	
Phenol	Equlibrated with Tris pH 8.0
$CHCl_3$:IAA (24:1)	
Omni Swabs	Whatman Scientific Cat # WB10 0004

Locus Amplification

All loci were amplified using the Polymerase Chain reaction. PCR primers used to amplify the loci of interest are listed in **TABLE 2.4**, and the conditions for amplification are listed in **TABLE 2.5**. All reactions were run in MJ 96 well thermal cyclers with heated lids. In general, each reaction contained 2-4 μl of 5ng/μl DNA, 0.1 U of Taq polymerase, 10 mM of dNTPs, and 20 mM of total primer in a 30 μl reaction. All PCR experiments contained a positive control (known DNA sample) as well as a negative control (water). PCR reactions were visualized on a 2.0% agarose gel stained with GelStar to check for the presence of a single amplicon. Schematic representations of the primers used for each locus are presented in **FIGURE 2.1** and **FIGURE 2.2**.

For each locus, amplification primers were designed to be in unique regions flanking the locus of interest, and both the flanking regions and the locus were amplified, visualized on a gel, purified, and then sequenced. For the NRY, only *Alu* elements located within introns of single copy genes were amplified. Each amplicon had to contain at least one element from the Y family (the most recent family) of *Alus*. The *Alu* family was determined using the program RepeatMasker (www.repeatmasker.org).

TABLE 2.4 Primers used for DNA amplification.

Locus	Upper Primer Sequence	Lower Primer Sequence
mtDNA COIII		
MT9203F-MT1R	AGCCTCTACCTGCACGA	TAATTGGAAGTTAACGGTACT
NRY ALUS		
AMELY ALU Y1	TTGTTTGCCTGCCTTGTG	TCTGAGAATAGTCAAGATGT
UTY ALUY62	ATTCAGTATCTCCAAAAGTC	GAAAAGCAAAATAAAATGTAG
DFFRY ALUY40	AGAAGATGATAAAGATGGTG	CTTATTTGTTTCAGAGCAGG
DBY ALUYD1	ACAGCGAGCAGTAAGTAA	ACTAACCTCACCCAATCT
DBY ALUYC2	TTTTCACTTTCCAACTTTTCAT	AGACTCTTATGTGCTATACTA
DFFRY ALUY30	ACATTACAGGACCTTGAT	TGCCTAACAATACTCCC
DFFRY ALUY50	TGGAGGGTAAGTGAGTAG	TTTTAATGGAACACCGTAG
SMCY ALUY2	TGAGGTTGATGTTTACTAAGATC	CCTGCTAAATCAGTTTCCACAC
UTY ALUY87	TTATTCCACCCAGCACTGTTA	AGGCACAAATGGTAAGGTCTT
486 YA5	TGTGGTAAGTGTAGTTTCAA	TCTGGACTGGAAACATAA
16E4 YA5	GAAGCAATACTCTGAAAAGT	TTTGGAGGGACATTATTCTC
UTY ALUY 83	ATTCTGTGTTCTCTTTTATT	ACTTCAGAAATAAATGCT
DFFRY ALUY 4	CTGATTATTCTTTTCTACCTTG	GTTATGCCAGGAAACATGCC
PDHA1		
1F-1R	TTTATATGGCGATGGTGCTG	GCTGGCAGCACTCCTACTTC
2F-2R	GGTACAATGGGCTGAGCAGT	CCTTCCTCACTTCCACATCAA
RRM2P4		
1F-1R	GTCTTTTATGTGATACCACCCG	

TABLE 2.5 Loci conditions for PCR amplification.

LOCUS	BUFFER	TEMP	TAQ	CYCLER
MTDNA				
COIII	G	65-55	CLONETECH WITH DIL BUFFER	TOUCHDOWN
PDHA1				
PDHA1 1F-1R	E	67-57	INVITROGEN PLAT TAQ	TOUCHDOWN
PDHA1 2F-2R	D	68-58	INVITROGEN PLAT TAQ	TOUCHDOWN
RRM2P4				
RRM2P4 1F-1R	D	58-48	INVITROGEN PLAT TAQ	TOUCHDOWN
NRY ALUS				
AMELY ALU Y1	E	65-55	INVITROGEN PLAT TAQ	TOUCHDOWN
UTY ALUY62	11	65-55	INVITROGEN PLAT TAQ	TOUCHDOWN
DFFRY ALUY40	11	65-55	INVITROGEN PLAT TAQ	TOUCHDOWN
DBY ALUYD1	C	60-50	CLONETECH WITH DIL BUFFER	TOUCHDOWN
DBY ALUYC2	11	65-55	CLONETECH WITH DIL BUFFER	TOUCHDOWN
DFFRY ALUY30	1	59-49	CLONETECH WITH DIL BUFFER	TOUCHDOWN
DFFRY ALUY50	7	62-52	CLONETECH WITH DIL BUFFER	TOUCHDOWN
SMCY ALUY2	3	68-58	CLONETECH WITH DIL BUFFER	TOUCHDOWN
UTY ALUY87	9	65-55	CLONETECH WITH DIL BUFFER	TOUCHDOWN
486 YA5	1	53	CLONETECH WITH DIL BUFFER	STANDARD
16E4 YA5	4	53	CLONETECH WITH DIL BUFFER	STANDARD
UTY ALUY 83	1	56-46	INVITROGEN PLAT TAQ	TOUCHDOWN
DFFRY ALUY 4	F	65-55	INVITROGEN PLAT TAQ	TOUCHDOWN

DNA Purification

Before DNA sequencing, all samples were purified using either the protocol listed below or automated DNA purification very similar to the manual protocol.

Manual Millipore Cleanup Protocol

1. Add 200 uL H_20 to each well and vacuum through until dry (~10-15 minutes).
2. Add 200 uL H_20 to each well and cover plate with aluminum sealing tape.
3. Shake plate for 5 minutes on vortex.
4. Discard H_20.
5. Add 100 uL of sample to each well and vacuum through till dry.
6. Add 100 uL of H_20 to each well and vacuum through till dry.
7. Add 100 uL of sample to each well and cover plate with aluminum sealing tape.
8. Shake plate for 5 minutes on vortex.
9. Pat down plate to remove sample from cover.
10. Transfer (elute) clean sample to a newly labeled catch plate.

DNA Sequencing

After DNA amplification, visualization, and purification, the DNA product was sequenced. All DNA sequencing was completed on a high volume, 96-capillary Applied Biosystems 3730xl DNA analyzer at The University of Arizona. The machine was capable of producing over 800 basepairs of high quality sequence data for each sample. In general, most loci were sequenced in both directions (except for a few of the NRY *Alu*s and gaps were avoided by the use of overlapping regions creating contigs.

Individuals with poor or ambiguous sequences were re-amplified and re-sequenced.

TABLE 2.6 presents the sequencing primers used for each locus.

TABLE 2.6 Primers used to sequence loci.

Locus	Upper Primer Sequence	Lower Primer Sequence
mtDNA COIII		
MT9203F-MT1R	AGCCTCTACCTGCACGA	TAATTGGAAGTTAACGGTACT
PDHA1		
PDH1.1	TTGCTCTACATCAGTGCTT	-
PDH1.2	GAAAGAAGCCAAATGAAACC	-
PDH1.3	GTTCAAAGACTGCCTCCCAT	-
PDH1.4	CTTAGGAGGTTTGGGTGTTC	-
PDH2.1N	CCTTCCTCACTTCCACAATCAA	-
PDH2.2NN	ATACTTGCTAGAAATGAGAACAG	-
PDH2.3NN	AAATACATCAATCAAAAAGC	-
PDH2.4NN	TTACTACTTTTCCCTCCCCAT	-
RRM2P4		
RRM2P4-1F	TTTCAATTAATTCCCGT	
RRM2P4-4R		AGAAGAAGGCAGGTTGGGTC
RRM2P4-2F	CAATAAATAGCAAGAAGG	
RRM2P4-1R		TGAGTGTATGAGCAGTGAAGCA
NRY ALUS		
AMELY ALU Y1	TTGTTTGCCTGCCTTGTG	TCTGAGAATAGTCAAGATGT
UTY ALUY62	ATTCAGTATCTCCAAAAGTC	GAAAAGCAAAATAAAATGTAG
DFFRY ALUY40	AGAAGATGATAAAGATGGTG	AATCTGGCTGGAAAACCC
DBY ALUYD1	ACAGCGAGCAGTAAGTAA	ACTAACCTCACCCAATCT
DBY ALUYC2	TTTTCACTTTCCAACTTTTCAT	AGACTCTTATGTGCTATACTA
DFFRY ALUY30	ATCTGGGCAGCACAGGTA	CAGGGTTTTTTATTATGGAA
DFFRY ALUY50	TTCTGTTTTGATGCTTGTC	-
SMCY ALUY2	TGAGGTTGATGTTTACTAAGATC	CCTGCTAAATCAGTTTCCACAC
UTY ALUY87	-	AGGCACAAATGGTAAGGTCTT

486 YA5	TGTGGTAAGTGTAGTTTCAA	-
16E4 YA5	GAAGCAATACTCTGAAAAGT	TTTGGAGGGACATTATTCTC
UTY ALUY 83	ATTCTGTGTTCTCTTTTATT	-
DFFRY ALUY 4	-	CTGGTTAGGGTTCACTGC

Analytical Procedures

Sequence Alignment

All direct sequencing results were aligned using Sequencher version 4.1. (Genecodes Corporation) and checked manually. Ambiguous bases that could not be resolved were labeled "N". For each individual, a consensus sequence was created from the aligned amplicons. These individual consensus sequences were grouped by population and exported to other programs (generally DnaSP version 3.51) for further analysis.

Population Genetic Analyses

Tests of Neutrality and Methods Used to Detect Population Size Change

There are several statistical tests widely used to evaluate deviations from the neutral equilibrium model. The strength of each of these tests to detect deviation from neutrality differs, often depending on the type of data being analyzed. Specifically, characteristics such as the number of polymorphic sites, the time since the onset of an expansion, the strength of the expansion, and the number of samples included in the analysis can greatly affect the strength of any given test (Ramos-Onsins and Rozas 2002). Therefore, it is recommended that several of tests should be used which will hopefully converge upon similar results. These tests can be based on the number of segregating sites, haplotype distribution, and pairwise sequence distribution or mismatch. Here I provide a description of some of the most commonly used tests in studies of population expansion.

Tajima's D

Effective population size (N_e) is defined as the number of breeding individuals in a population (Wright 1931). Polymorphism data can be used to calculate two measures of nucleotide diversity: the number of segregating sites, θ_s (Watterson 1975), and the average number of pairwise differences, θ_π (Nei and Li 1979). The effective population size can be estimated from within-population diversity measures such as θ_π or θ_s since both are estimators of the neutral parameter, θ_W, which equals $4Ne\mu$ for autosomal loci, $3Ne\mu$ for X-linked loci, and $Ne\mu$ for the Y-chromosome and mtDNA (where *Ne* is the effective population size and μ is the mutation rate). Assuming panmixia and no recurrent mutation, θ_π should equal θ_s at mutation-drift equilibrium. Tajima's *D* (Tajima 1989) is based on the difference between θ_π and θ_s, and may point to evolutionary forces causing deviations from neutrality. There are both selective and demographic explanations for Tajima's D results. Significantly positive values of Tajima's *D* can be consistent with balancing selection, a population bottleneck, or admixture, whereas significantly negative values of Tajima's *D* can be indicative of positive directional selection or population expansion. This test was employed using DnaSP version 4.01 (Ramos-Onsins and Rozas 2002).

R_2 Statistic

This is one of a number of tests developed by Ramos-Onsins and Rozas based on the frequency spectrum, or difference between the average number of nucleotide differences and the number of singleton mutations (Ramos-Onsins and Rozas 2002). The variables included in the calculation are: the sample size, the total number of segregating sites, the average number of nucleotide differences between two sequences, and the number of singleton mutations in each sequence. The number of singleton mutations on a branch of a genealogical tree after a growth

event is expected to be equal to the average number of pairwise differences divided by two. Therefore, low R_2 values are expected under such a scenario.

Fu's Fs

Fu's (1997) Fs statistic is based on the haplotype distribution and compares the number of alleles observed in a sample with the expected number in a population of constant size on the basis of the observed average mismatch. It is based on the probability of having a number of alleles greater or equal to the observed number in a sample drawn from a stationary population with parameter $\Theta = 2N_e u$ (where N_e is the effective population size and u is mutation rate for the whole sequence). Like Tajima's D, negative Fu's Fs values are generated when a population expands, and positive values are generated when it contracts. Selection can also affect the shape of the underlying gene tree and therefore mimic these same demographic changes. The significance of Fu's Fs is calculated using coalescent simulations (5000 iterations) to generate the null distribution from a stationary population.

Mismatch Distribution

A polymorphic nucleotide site generally is represented by only two states within a sample, one of which is considered ancestral and the other mutant. The site frequency spectrum depicts the fraction of the polymorphic sites at which the mutant form is found in one copy, in two copies, and so on. The mismatch distribution is the distribution of the number of pairwise differences between haplotypes and can be used to estimate parameters of a demographic expansion model. Population expansions generate a distinctive signature both in the site frequency spectrum and the mismatch distribution. If a population has expanded, the frequency

spectrum should have an excess of sites at which the minor allele is very rare, and a unimodal wave would portray the mismatch distribution (Harpending and Rogers 2000). The star-like phylogeny which is the product of population expansion tends to be represented by branches with mutations that are unique to single individuals. This can be reflected by the shape of the mismatch distribution, or the distribution of pairwise differences between sequences (Rogers and Harpending 1992). Smooth, unimodal mismatch distributions can generally be regarded as being representative of a population expansion, while ragged multimodal mismatch distributions are representative of a stationary or shrinking population (Rogers and Harpending 1992; Excoffier and Schneider 1999). Population subdivision (Marjoram and Donnelly 1994) and admixture (Bertorelle and Slatkin 1995) can act as confounding factors.

McDonald-Kreitman Test

The McDonald Kreitman (MK) test is a powerful test of neutrality that compares the within species to between species patterns of evolution (McDonald and Kreitman 1991). The MK test uses data from the proportion of synonymous sites (sites that do not change the resulting amino acid) to nonsynonymous sites (sites that do change the amino acid) across a pair of closely related species (in this study we draw comparisons between humans and chimpanzee). It relies on the assumption that the ratio of synonymous to non-synonymous substitutions should be similar to the ratio of synonymous to non-synonymous polymorphism under the neutral theory. A G-test of independence is used to test the significance of these differences—possibly indicating positive selection. Relevant to this work, Eyre-Walker (2002) found that a slight increase in effective population size can generate "artifactual evidence" of adaptive evolution when the amino acid substitutions are slightly deleterious.

Coalescent-Based Analysis Using GENETREE

To address the question of levels of past population structure, I estimated gene trees by maximum parsimony, assuming no recombination and an infinite sites model. If these criteria were met, I calculated the TMRCAs and the ages of exclusive mutations using a maximum-likelihood approach, as implemented in GENETREE (Griffiths and Tavare 1994). Gene flow events can be assigned parsimoniously to a gene tree simply on the basis of the gene tree topology and the population identification of the sampled individuals (Slatkin 1989; Slatkin and Maddison 1989). I utilized the estimates of the ages of the mutations to put bounds on the timing of the gene flow events. For each model, I calculated θ, the TMRCA, and the number of migrants per generation (*Nm*) for each deme. These data formed the basis of a likelihood ratio test to identify the model of population structure that best fit the data.

GENETREE (Bahlo and Griffiths 2000) uses a Markov chain Monte Carlo method (MCMC) to model DNA evolution using the coalescent and to generate maximum-likelihood estimates of population parameters from the data. Under the infinite sites model, unique gene trees describing the mutational history of the locus for a population are constructed. Estimates of the maximum likelihood θ (θ_{ml}) for the population, the effective size of the population (N_e for a stationary population, N_o for a growing population), and the Time to the Most Recent Common Ancestor (TMRCA), under models of constant population size and exponential growth (with an intrinsic growth factor of β) are obtained. A likelihood ratio test of the log likelihood of the θ_{ml} produced under a model of constant population size and the log likelihood of the θ_{ml} generated under a model of population growth is performed. The significance of θ_{ml} is then evaluated using a simple chi-square test.

Methods Used to Detect Population Structure

F_{ST}

To address the question of levels of current population structure in modern African populations, I calculated F_{ST} (a standardized measure of genetic variation among populations, or the variance among subpopulations relative to the total population), assuming an infinite sites model of mutation and an island model of population structure (Wright 1951). In the island model of migration a species is assumed to be subdivided into a number of subpopulations ("islands" or "demes") with the same effective size which exchange a proportion of migrants drawn at random from the whole population at a constant rate every generation (Hartl and Clark 1997; Hedrick 2000). Under this scenario, the population differentiation is described by a parameter (F_{ST}), which can be thought of as a kind of "fixation index" or coefficient of inbreeding . The island model assumes that migration from every subpopulation is equally likely to carry alleles to each other subpopulation and that alleles are neutral (Hartl and Clark 1997; Hedrick, 2000). At equilibrium between mutation and drift, $F_{ST} = 1/4N_e m+1$, where N_e is the effective population size of each population and m is the migration rate between populations.

Analysis of Molecular Variance

The Analyses of Molecular Variance (AMOVA) is available in the software package, ARLEQUIN (Excoffier, Laval, and Schneider 2005). The amount of variation between individuals within demes (Φ_{ST}), between demes within populations (Φ_{SC}), and between populations (Φ_{CT}) can be estimated and their significances tested by use of a nonparametric permutation procedure (Excoffier, Smouse, and Quattro 1992). Analysis of Molecular Variance (AMOVA) is used to measure levels of population structure and differentiation and to estimate

of levels of gene flow. The test may be used on a variety of molecular data including molecular marker data, direct sequence data, or phylogenetic trees based on molecular data, and both haplotype frequencies and molecular differences between haplotypes are taken into account with this approach (Excoffier, Smouse, and Quattro 1992).

Isolation with Migration Coalescent Simulations

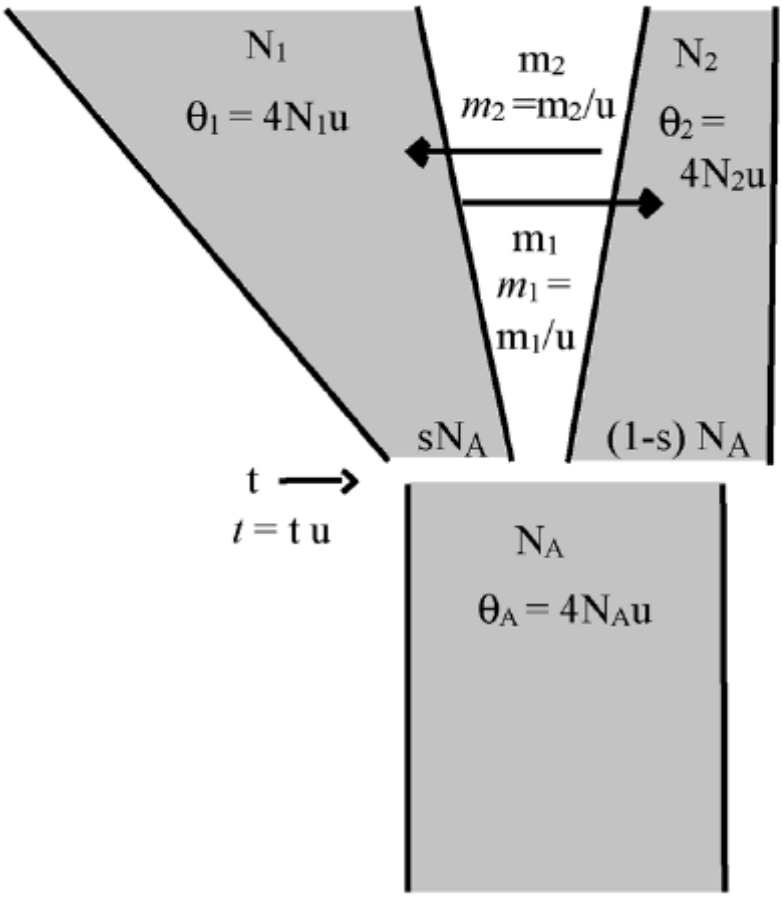
A general Isolation with Migration (IM) model of population structure is one of population splitting in which there is an ancestral population that gives rise to two descendent populations which may be connected by gene flow. Here we use the model modified by Hey and Nielsen (2004) to be used for multiple loci, which is an extension of the model developed by Nielson and Wakeley (2001) for data from a single, non-recombining locus. The model asks what range of possible gene trees are consistent with a given dataset.

The IM model is a Bayesian, Markov chain Monte Carlo (MCMC) method used to take into account the stochastic variance among loci. The model is used to capture many of the dynamics that occur during the early stages of population divergence. It has seven main parameters: constant effective population sizes for the ancestral population ($N_{e\mathrm{A}}$) and two descendant populations (N_{e1} and N_{e2}), migration rates between the descendant populations (m_1 and m_2), the time at which the ancestral population gave rise to the descendant populations (or the splitting time, t), and the proportion of the ancestral population that founded the descendent populations ($sN_{e\mathrm{A}}$) (**FIGURE 2.4**). Each of these parameters can be scaled by the rate of genetic drift or by the mutation rate.

For multiple loci, the model assumes that loci are independent, neutrally evolving, and non-recombining (or, at least do not show any evidence of recombination). It also assumes that there

are no unsampled populations that are more closely related to the sampled populations than they are to each other, an assumption that is difficult to fulfill using human populations.

FIGURE 2.4 Schematic of the Isolation with Migration under changing population size model, with associated parameters (from Hey 2005).



CHAPTER 3: CONTRASTING SIGNATURES OF POPULATION GROWTH FOR MITOCHONDRIAL DNA AND Y CHROMOSOMES AMONG HUMAN POPULATIONS IN AFRICA

Maya Metni Pilkington[1], Jason A. Wilder[2,3], Fernando L. Mendez[4], Murray P. Cox[2], August Woerner[2], Thiep Angui[2], Sarah Kingan[2], Zahra Mobasher[2], Chiara Batini[5], Giovanni Destro-Bisol[5], Himla Soodyall[6], Beverly I. Strassmann[7], Michael F. Hammer[1,2,4]

[1] Department of Anthropology, University of Arizona, Tucson AZ 85721

[2] ARL Division of Biotechnology, University of Arizona, Tucson AZ 85721

[3] Department of Biology, Williams College, Williamstown MA 01267

[4] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85721

[5] Department of Animal and Human Biology, University La Sapienza, Rome, Italy

[6] Human Genomic Diversity and Disease Research Unit, National Health Laboratory Service and University of the Witwatersrand, Johannesburg, South Africa

[7] Department of Anthropology, University of Michigan, Ann Arbor MI

Published in *Molecular Biology and Evolution* (2008), 25(3):517-25.

Reprinted with permission from publisher (License number 1912671206384, March 19, 2008). Figure, table names, and heading format are modified for this work.

Keywords: population growth, *Homo sapiens*, sub-Saharan Africa, mtDNA, NRY, hunter-gatherer

ABSTRACT

A history of Pleistocene population expansion has been inferred from the frequency spectrum of polymorphism in the mitochondrial DNA (mtDNA) of many human populations. Similar patterns are not typically observed for autosomal and X-linked loci. One explanation for this discrepancy is a recent population bottleneck, with different rates of recovery for haploid and autosomal loci as a result of their different effective population sizes. This hypothesis predicts that mitochondrial and Y chromosomal DNA will show a similar skew in the frequency spectrum in populations that have experienced a recent increase in effective population size. We test this hypothesis by re-sequencing 6.6 kb of non-coding Y chromosomal DNA and 780 basepairs of the mtDNA *cytochrome c oxidase subunit III* (*COIII*) gene in 172 males from five African populations. Four tests of population expansion are employed for each locus in each population: Fu's *Fs* statistic, the R_2 statistic, coalescent simulations and the mismatch distribution. Consistent with previous results, patterns of mtDNA polymorphism better fit a model of constant population size for food-gathering populations and a model of population expansion for food-producing populations. In contrast, none of the tests reveal evidence of Y chromosome growth for either food-gatherers or food-producers. The distinct mtDNA and Y chromosome polymorphism patterns most likely reflect sex-biased demographic processes in the recent history of African populations. We hypothesize that males experienced smaller effective population sizes and/or lower rates of migration during the Bantu expansion, which occurred over the last five thousand years.

INTRODUCTION

Several studies in the early 1990s concluded that the excess of low frequency polymorphisms in human mtDNA over the expected distribution under the standard neutral model was the signal of rapid population expansion (Di Rienzo and Wilson 1991; Slatkin and Hudson 1991; Lundstrom, Tavare, and Ward 1992; Aris-Brosou and Excoffier 1996). Many populations also exhibit unimodal peaks in the distribution of the number of pairwise differences (mismatch distributions) in mtDNA sequences, suggesting that populations expanded in size beginning ~30–130 kya (Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Sherry et al. 1994; Rogers and Jorde 1995). This pattern of rapid expansion from small ancestral effective population size (N_e) has not been generally observed for autosomal and X chromosomal DNA sequencing data sets (Hey 1997; Harpending and Rogers 2000; Wall and Przeworski 2000; Excoffier 2002). X-linked and autosomal sequences typically show no excess of low frequency polymorphisms, instead there is a tendency for non-African populations to have positive Tajima's D values, and African populations to have only slightly negative values (Przeworski, Hudson, and Di Rienzo 2000; Garrigan and Hammer 2006). There is disagreement on the causes of the discrepancy between mtDNA and nuclear data sets: some investigators favor models involving differential natural selection (Hey 1997; Harpending and Rogers 2000; Excoffier 2002), while others favor a demographic model featuring a recent bottleneck (Fay and Wu 1999; Hammer et al. 2004a). For example, Fay and Wu (1999) pointed out that for a period of time following a population bottleneck, there is expected to be an excess of low frequency polymorphisms in mtDNA and an excess of intermediate frequency autosomal polymorphisms simply because of the different effective population sizes of these loci.

Under a simple bottleneck model with an equal breeding sex-ratio, both the non-recombining portion of the Y chromosome (NRY) and mtDNA are expected to experience a similar reduction in N_e. As the population recovers from this bottleneck, the two haploid loci should respond similarly with respect to changes in the frequency distribution of polymorphisms over time (Fay and Wu 1999). To date, there are no published studies designed specifically to test the bottleneck hypothesis. While some studies of SNPs and STRs on the NRY have supported models of demographic expansion (Pritchard et al. 1999; Thomson et al. 2000), the grid-sampling schemes employed have reduced power to distinguish between the effects of population expansion and population structure (Ptak and Przeworski 2002). For instance, Hammer et al. (2003) showed that sampling a few individuals from many global populations may lead to an upward bias in the number of singleton variants. They concluded that it is necessary to thoroughly sample within demes to obtain a robust estimate of the frequency spectrum. Moreover, inferences of population growth based on Y-STR data are difficult to interpret as a result of uncertainty associated with microsatellite mutation models (Pritchard et al. 1999; Wall and Przeworski 2000).

In this study, we use a population-based sampling strategy and DNA re-sequencing data and compare patterns of DNA polymorphism in the mtDNA and on the NRY in the same sample of 172 males from five sub-Saharan African populations. We chose to sequence approximately 780 basepairs (bp) of the mtDNA *cytochrome c oxidase subunit III* (*COIII*) gene and 6.6 kb of non-coding Y chromosomal DNA encompassing 13 *Alu* elements because the former has lower levels of homoplasy than the mtDNA control region and the latter has a three-fold higher SNP density than other non-coding regions on the NRY (Wilder, Mobasher, and Hammer 2004). The extent of mutation rate heterogeneity in a given region is a concern as this phenomenon may

confound the signal of population growth (Aris-Brosou and Excoffier 1996). Previous DNA sequence surveys of the mtDNA control region indicated dramatically different signals of population expansion among African populations: many food-producing populations exhibited unimodal mismatch distributions consistent with past demographic expansions, while most food-gatherers showed ragged distributions. This was interpreted to be the consequence of a recent contraction in effective population size as Bantu-speaking farmers infringed upon hunter-gatherer territory (Excoffier and Schneider 1999). Our samples comprise both food-producing (the Dogon, Dinka, and a group of Southeast Bantu-speakers) and food-gathering groups (the Khoisan and Bakola). We ask whether mtDNA and Y chromosomal genealogies reflect similar population histories: that is, do the food-producing populations better fit a model of population growth and food-gathering populations better fit a model of constant size for both haploid loci.

MATERIALS AND METHODS

Populations and Loci Surveyed

We surveyed DNA sequence variation in five sub-Saharan African populations (**FIGURE 3.1**) including the Dogon of central Mali (n= 40-49), the Dinka of southern Sudan (n=23), the Bakola of southern Cameroon (n=24-25), the Khoisan from Namibia (n=25) and a group of Southeast Bantu speakers (SE Bantu) from southern Africa (n=46-50) (**TABLE 3.1**). The SE Bantu samples comprise a collection of Zulu, Ndebele, Khosa, Sotho, Swazi, Tswana, Pedi and Tsonga (Lane et al. 2002). Additionally, we analyzed orthologous DNA sequences from one common chimpanzee (*Pan troglodytes*) to determine ancestral states at each site. The Dinka DNA samples were obtained with written informed consent and the protocols were approved by the Human Subjects Committee at the University of Arizona. The SE Bantu and Khoisan

samples were obtained with either verbal or written consent with approval from the Committee for Research on Human Subjects, University of the Witwatersrand (protocol number M980553). The Dogon samples were collected with verbal consent with approval from the University of Michigan Health Sciences Institutional Review Board. The Bakola samples were collected by Gabriella Spedini and Giovanni Destro-Bisol with verbal informed consent and approval from the University of Rome "La Sapienza".

We generated DNA sequence data for 780 bp of the mitochondrial *cytochrome c oxidase subunit III* (*COIII*) (**SUPPLEMENTARY TABLE 2.1**) and for 6.6 kb of non-coding NRY encompassing 13 *Alu* elements in the Yα5 subfamily (see Wilder, Mobasher and Hammer 2004; **SUPPLEMENTARY TABLE 2.2,** primers and protocols available upon request). MtDNA and NRY re-sequence data were collected from the same individuals; however, sample numbers varied slightly between loci. Nucleotide sequences were obtained using an ABI Prism 3730 automated sequencer, and aligned using Sequencher version 4.1 (Genecodes Corporation). Insertion/deletion polymorphisms were excluded from all analyses.

Population Genetic Analyses and Tests of Population Growth

Population parameters such as nucleotide diversity (π and Θ) and Tajima's D were calculated using DnaSP version 4.0 (Rozas et al. 2003). Inferences of population expansion were made using four different methods. First, we calculated Fu's *Fs* statistic (Fu 1997), which is based on the probability of having a number of haplotypes greater or equal to the observed number of samples drawn from a constant-sized population. To complement this method, we calculated the R_2 statistic (Ramos-Onsins and Rozas 2002), which is based on the difference between the number of singleton mutations and the average number of nucleotide differences.

Ramos-Onsins and Rozas (2002) demonstrated that these statistics have the greatest power to detect population expansion for non-recombining regions of the genome under a variety of different circumstances, especially when population sample sizes are large (~50, Fu's *Fs*) or when sample sizes are small (~10, R_2). They also found that the power of the R_2 statistic is relatively high when the number of segregating sites is low (e.g., < 20). The significance of Fu's *Fs* and R_2 were obtained by examining the null distribution of 5,000 coalescent simulations of these statistics using DnaSP. Significantly large negative Fu's *Fs* values and significantly positive R_2 values were taken as evidence of a population expansion.

To further test for evidence of population expansion, we used coalescent simulations to generate maximum-likelihood estimates of population parameters for mtDNA and the NRY under the infinite sites model using the program GENETREE version 9.0 (R.C. Griffiths, http://www.stats.ox.ac.uk/~griff/ software.html). GENETREE uses a coalescent Markov chain Monte Carlo approach to search the state space of DNA sequence evolution (Bahlo and Griffiths 2000). Unique gene trees describing both mtDNA and NRY mutational histories for each of the five African populations were constructed using Seq2tr (Wilder, Mobasher, and Hammer 2004). For each population, we generated maximum likelihood estimates of the population mutation rate (Θ_{ml}) under a model of constant population size and a model of exponential population growth, where we specified the range of growth parameters (β). The maximum likelihood estimate of θ_{ml} was then used to calculate N_e (or N_o) under each of the demographic models. Likelihood ratio tests of the log likelihood of the θ_{ml} produced under models of constant population size *versus* population growth were performed. The significance of θ_{ml} was evaluated using a simple chi-square test with one degree of freedom. To calculate effective population size we used previously published mutation rate estimates (1.58×10^{-8} mutations per site per year for

the mtDNA *COIII* region and 4.19 x 10^{-9} mutations per site per year for the NRY) with a generation time of 25 years (Wilder, Mobasher, and Hammer 2004; Wilder and Hammer 2007b).

The mismatch distributions were also examined (Rogers and Harpending 1992) using the program, ARLEQUIN version 2.000 (Excoffier, Laval, and Schneider 2005). The number of observed differences between pairs of mtDNA or NRY haplotypes was compared with the expected distribution of differences under a specified demographic model (i.e., constant population size or population growth). The mismatch distribution method uses estimated parameters of the expansion (τ, Θ_0, and Θ_1) to perform coalescent simulations of stepwise expansions and create new estimates of the same parameters (τ*, Θ_0*, and Θ_1*). The estimated demographic model is tested by obtaining the sum of the squared differences (SSD) between the observed and the estimated mismatch distributions (Schneider and Excoffier 1999). A significant SSD *P* value is interpreted here as departure from the estimated demographic model, which is population expansion when τ>o and $\Theta_1 > \Theta_0$, and population stationarity when $\tau = 0$ or $\Theta_1 = \Theta_0$ (these should read τ hat and Θ hat for "expected"). Population stationarity was also inferred if the 95% confidence intervals for Θ_1 and Θ_0 overlapped, even if the *P* value of the SSD was not significant (Excoffier and Schneider, 1999). To take into account known deviation from the infinite sites model for the mtDNA, the Kimura Two-Parameter model was used (Kimura 1980) and mutation rates followed a gamma distribution with a shape parameter α= 0.22 (Schneider and Excoffier1999).

Results

Patterns of Nucleotide Diversity

TABLE 3.1 summarizes nucleotide diversity in the sub-Saharan African populations studied here. The number of segregating sites (*s*) identified at the two loci is similar: for mtDNA *COIII s* ranges from 6 in the Bakola to 16 in the SE Bantu samples, while *s* ranges from 8 in the Dinka to 14 in the Bakola samples for the NRY. The average value of π per locus is 1.6 for mtDNA and 2.7 for the NRY, while the average value of Θ per locus is similar for both loci (2.6 and 2.5, respectively). In contrast, the average number of singleton substitutions is almost twice as high for mtDNA (5.0) compared with the NRY (2.6), although this difference is not statistically significant. The number of mtDNA singleton mutations ranges from 2 in the food-gatherer populations to 12 in the Dinka (mean of seven in the food-producers), while the number of NRY singletons only ranges from one to four (mean of three in the food-producers).

For mtDNA, Fu's *Fs* values for all populations are negative, with the three food-producing populations (SE Bantu and Dinka) having statistically significant negative *Fs* values ($p < 0.001$). For the NRY, two populations have positive *Fs* values (not statistically significant) and the others have values close to zero. Only a single population had a statistically significant R_2 value (the Dinka for mtDNA; $p < 0.001$). To test for the effect of differing sample sizes on Fu's *Fs* statistic, we randomly sub-sampled 20 individuals from each population 100 times for both mtDNA and the NRY (**3. 1**). The overall trend observed from reducing the sample sizes to 20 was for Fu's *Fs* to become less negative, except for situations where sample size was already small. For mtDNA the only noticeable effect of sub-sampling was for the Dogon, which no longer had a significant Fu's *Fs* value. For the NRY, sub-sampling did not change any of the inferences (i.e., no population has a statistically significant Fu's *Fs* value).

To test whether the mtDNA and NRY Fu's *Fs* values were significantly different from each other, we conducted coalescent simulations to create a sampling distribution of the difference between the mtDNA and NRY Fu's *Fs*. We simulated neutral genealogies, conditioned on the Θ values estimated from GENETREE under the null model of constant population size, and using the sample size of each population (which ranged from 23-50). We calculated Fu's *Fs* for each of 10,000 simulated datasets after assuring that the average number of segregating sites generated was comparable to those observed (**TABLE 3.2**). We arbitrarily subtracted the NRY Fu's *Fs* values from the mtDNA Fu's *Fs* values, and asked whether the observed mtDNA - NRY Fu's *Fs* was significantly small under the distribution ($\alpha = 0.05$). We find that the mtDNA and NRY Fu's *Fs* values are significantly different for the SE Bantu and the Dinka (both food-producers), while they were not significantly different for the Dogon (a food-producer), Khoisan and Bakola (both hunter-gatherers).

GENETREE Simulations

TABLE 3 lists the maximum likelihood estimates of the population mutation rate, growth rates, effective population sizes, and likelihood ratio tests for each population for each locus. For mtDNA, three food-producing populations better fit a model of exponential growth (LRS = 6.494, 6.720, 4.783 for the SE Bantu, Dinka and Dogon, respectively), while the two hunter-gatherer populations better fit a model of constant size (Khoisan and Bakola). In contrast, all NRY populations better fit a model of constant population size.

Mismatch Distributions

FIGURE 3.2 shows mismatch distributions for mtDNA *COIII* and the NRY. For mtDNA, the SE Bantu and the Dinka fit a model of population growth. The Khoisan mtDNA mismatch distribution also appears to fit a model of growth; however, the Θ_0 and Θ_1 95% confidence intervals are very close to overlapping (Θ_0 = 0.000 - 1.416, and Θ_1 = 2.269 - 602.722), and in fact do overlap for their 99% confidence intervals (Θ_0 = 0.000-2.455, and Θ_1 = 0.913 – 6818.972). Thus, similar to Excoffier and Schneider (1999), the Khoisan mtDNA mismatch distribution is interpreted as being consistent with population stationarity. SSD was significant for the Dogon and the Bakola mtDNA mismatch distributions, indicating a poor fit to the stepwise growth model. In contrast to mtDNA, all populations reject a model of population expansion for the NRY (**FIGURE 3.2**).

While there is a great deal of consistency among the methods used to infer changes in population size, some differences are noteworthy. For example, the SE Bantu mtDNA Fu's *Fs*, GENETREE and mismatch distributions are consistent in supporting a model of growth, while the R_2 statistic is not statistically significant (i.e., does not reject a model of constant population size). This apparent contradiction may be explained by the finding the R_2 statistic has less power than Fu's *Fs* to detect expansion when sample sizes are large, as is the case here (n=50). The mismatch distribution test provided the least consistent results as judged from the results of the other tests.

DISCUSSION

Here we present the first comparative analysis of DNA re-sequence data from both the NRY and mtDNA in the same set of sub-Saharan African population samples and ask whether

the genealogical histories of these two loci are concordant in their signals of population expansion. Similar to previous observations, we find that the mtDNA results indicate that, in general, the food-producers better fit a model of population expansion while the hunter-gatherers better fit a model of population stationarity. In contrast, the NRY re-sequence data indicate that all five populations better fit a model of constant population size. In the following sections we explore three possible explanations for why different patterns of population growth were inferred from these mtDNA and NRY re-sequencing datasets: (1) variation in mutation rate and mode, (2) natural selection, and (3) sex-specific demographic processes.

Variation in Mutation Rate and Mode

Two concerns with comparing patterns of mtDNA and NRY sequence variation are that mtDNA has much higher mutation and homoplasy rates than nuclear DNA (Horai et al. 1995; Takahata and Satta 1997). As pointed out by Fay and Wu (1999), after passing through a stepwise bottleneck, the more rapid and severe fluctuations in the frequency spectrum of mitochondrial *versus* autosomal DNA is due only to the mitochondrial genome's smaller population size, and has nothing to do with its higher mutation rate. However, the higher mutation rate for mtDNA compared with the NRY may contribute to a difference in power to detect population growth. For this reason, we sequenced much longer segments of the NRY to ensure that a comparable number of segregating sites were recovered from the two haploid systems. Indeed, π and Θ (calculated per sequence) are actually consistently higher for the NRY than for mtDNA (**TABLE 1**). As such, it does not appear that mutation rate differences between the NRY and mtDNA are contributing to the differing patterns of growth observed here. This is also illustrated by the fact that the Dogon mtDNA shows a signal of growth while the NRY does

not, despite a greater number of segregating sites on the NRY. We incorporated a shape parameter (α) for the gamma distribution of the mutation rate. This minimizes the influence of mutation rate heterogeneity, which can lead to erroneous inference of population growth. (Slatkin and Hudson 1991; Rogers and Jorde 1995; Aris-Brosou and Excoffier 1996). Importantly, different patterns of population growth for food-producers and food-gatherers were initially inferred from D-loop sequences, a region of the mitochondrial genome suffering the highest levels of rate heterogeneity (Meyer, Weiss, and Von Haeseler 1999; Ingman et al. 2000). Therefore, we do not believe that either mutation rate or rate heterogeneity are important factors contributing to the observed discrepancy between our mtDNA and NRY results.

Natural Selection

Positive directional selection is another process that may mimic the effects of population growth. Both haploid regions are particularly prone to the effects of periodic selective sweeps (genetic hitchhiking; Maynard-Smith and Haigh 1974), which are expected to lead to transient phases when there is an excess of rare variants over neutral expectations (Kaplan, Hudson, and Langley 1989; Braverman et al. 1995). While positive selection is thought to have reduced variation on the NRY (Malaspina et al. 1990; Dorit, Akashi, and Gilbert 1995; Jaruzelska, Zietkiewicz, and Labuda 1999; Pritchard et al. 1999) and to have influenced mtDNA variation in some geographic regions (Mishmar et al. 2003; Ruiz-Pesini et al. 2004), there is still no strong evidence that differential selection underlies contrasting patterns of mtDNA and NRY variation in human populations (Wilder, Mobasher, and Hammer 2004). None of the five populations examined here has a significant excess of rare NRY variants; and while it remains possible that directional selection is the underlying cause of an excess of rare mtDNA variants over neutral

expectations, this explanation requires a selective agent that affects food-producers but not food-gatherers.

Another form of selection that could mimic the signal of population growth is weak purifying selection, which has been implicated as the factor underlying an observed excess of low frequency non-synonymous polymorphism in mtDNA (Nachman, Boyer, and Aquadro 1994; Nachman 1998; Ballard and Dean 2001). However, the excess of rare mtDNA polymorphism in some African populations is not caused solely by replacement polymorphisms: when the mtDNA data are re-analyzed using only third position sites, results for Fu's *Fs* and the R_2 statistics are similar to results of analyses including all sites (data not shown). Interestingly, the mtDNA data for the food-producers do exhibit an excess number of polymorphic non-synonymous sites, while those of the hunter-gatherers do not (McDonald-Kreitman test Fisher's Exact Test p-value = 0.0001 and 0.070, respectively). This suggests weaker purifying selection on the food-producer *versus* the food-gatherer populations, which may in turn suggest larger effective population sizes for the food-gatherers in the past.

Sex-Specific Demographic Processes

In this section, we turn to demographic explanations for the observed discrepancy between mtDNA and NRY results. The different mtDNA patterns for food-producers and food-gatherers were initially explained by a model of Pleistocene demographic expansions followed by more recent population crashes for marginalized hunter-gatherers, effectively erasing any signature of population growth (Excoffier and Schneider 1999). Later work by Ray, Currat and Excoffier (2003) examined the effects of a spatial expansion in a subdivided population. Their computer simulations demonstrated that a population with low levels of gene flow among neighboring

demes (N_em) during a spatial expansion usually do not exhibit the signature of expansion, while demes experiencing high N_em values tend to show statistically significant negative values of Fu's Fs statistic (Ray, Currat, and Excoffier 2003). When range expansions are simulated with small deme sizes followed by recent demographic growth (i.e., deme sizes and levels of gene flow with surrounding demes increase well after the spatial expansion), the outcome is similar to simulations of spatial expansions with demes of constant size that always exchanged large numbers of individuals with their neighbors (Ray, Currat, and Excoffier 2003). In other words, spatial expansions followed by relatively recent population growth with differing rates of gene flow among neighboring demes (N_em) can also produce different signatures of population growth at the molecular level.

Drawing upon the simulation results of Ray, Currat and Excoffier (2003), we suggest that a similar difference between effective population sizes and/or rates of gene flow (N_em) for males *versus* females may explain the discrepancy observed between our mtDNA and NRY results. For example, food-producing males may have experienced either a smaller effective population size (N_e) or lower rates of migration (m) than females during a phase of expansion. What are the possible causes of lower N_e or m for males of food producing populations? Two common cultural practices may lead to lower values of N_em for the NRY: polygyny and patrilocality. Polygyny, a marriage practice that allows males (but not females) more than one spouse, is widespread in many parts of the world, especially in Africa (Dorjahn 1959; Konotey-Ahulu 1980; Spencer 1980; Strassmann 2003). When males father children with more females than females do with males, the result is an increase in the variance in reproductive success among males, which lowers their N_e relative to females (Low 1988). Interestingly, food-producers are generally described as substantially more polygynous than hunter-gatherers (Cavalli-Sforza

1986; Biesele and Royal 1999). It is also well established that the widespread practice of patrilocality (defined as the tendency of females to migrate and males to remain in their natal groups) can result in lower rates of male migration (Murdock 1981). While most agricultural societies are patrilocal (Murdock 1967), hunter-gatherer groups are typically referred to as bi-local, i.e., as spending time living with both the father's and the mother's families (Marlowe 2004). The importance of polygyny and patrilocality in shaping patterns of maternally- and paternally-inherited variation has been suggested previously for African populations (Destro-Bisol et al. 2004b; Wood et al. 2005). Assuming that these processes evolved as populations shifted from foraging to food-producing lifestyles, they may have played an important role in shaping the distinctive patterns of mtDNA and NRY polymorphism observed here.

CONCLUSIONS

This work constitutes the first direct comparison of mtDNA and NRY re-sequence data from the same African samples for the purpose of examining patterns of population expansion. The four tests that we employed (especially the Fu's Fs, the R_2 statistic and coalescent simulations methods) were very consistent in revealing different patterns of population size change for the two haploid compartments of the genome. While all tests may be sensitive to unknown structure within populations (Ptak and Przeworski 2002; Hammer et al. 2003), we emphasize that the mismatch distribution test actually has the weakest power to detect growth (Ramos-Onsins and Rozas 2002). We find that mtDNA *COIII* data from African food-producers better fit a model of population growth while those for food-gathering populations better fit a model of stationarity. In striking contrast, NRY data from both the food-producers and the food-gatherers sampled here better fit a model of constant population size. We hypothesize that these

patterns are the result of differences in sex-specific demographic processes—in particular, asymmetrical migration and/or reduced male effective population sizes. These sex-biased demographic processes are expected to significantly alter the frequency spectrum of mtDNA and NRY polymorphisms during large spatial/demographic expansions (Ray, Currat, and Excoffier 2003). Such an expansion is known to have occurred recently as farmers speaking Niger-Congo Bantu languages expanded from a southern Cameroonian homeland over most of subequatorial Africa beginning ~4000 years ago (Holden 2002b; Lane et al. 2002; Destro-Bisol et al. 2004a; Wood et al. 2005). It still remains to be resolved whether the Bantu expansions and the spread of farming in Africa, or changes in population size that took place much earlier in the Pleistocene (Excoffier and Schneider 1999; Harpending and Rogers 2000), are responsible for the differential patterns observed here. Along theses lines, further analysis of populations with known differences in marriage customs or migration patterns would be extremely valuable. For example, the finding of similar signals of strong growth for both mtDNA and the NRY in a non-polygynous (or bi-local) food-producing population (i.e., thereby mitigating a reduction in male effective population size) would support the hypothesis that recent changes in cultural practices made an impact on patterns of genetic diversity (Wilkins and Marlowe 2006). The collection of additional genetic datasets that minimize ascertainment biases, along with the development of more realistic models of human demography that incorporate non-equilibrium processes, will aid future analyses of the effects of sex-biased processes on patterns of genetic variation.

ACKNOWLEDGMENTS

We would like to thank Amy Russell, and Heather Norton, Stephen Zegura, and Michael Nachman for helpful discussions. Publication of this work was made possible by grant GM-

53566 from the National Institute of General Medical Sciences (to MFH). Its contents are solely the responsibility of the authors and do not necessarily reflect the official views of the National Institutes of Health. Funding was also provided by NSF Doctoral Dissertation Improvement Grant (to MMP) and an NSF IGERT grant (to MMP).

TABLE 3.1 MtDNA and NRY polymorphism data for each sub-Saharan African population.

	Population	*N* [a]	Sites [b]	Hp [c]	π [d]	Θ [d]	Tajima's D	Number of Singletons	Fu's Fs	R_2	*P* SSD[e]	Fu's Fs[f]
mtDNA	Khoisan	25	10	8	2.00	2.70	-0.800	2	-1.651	0.096	0.469	-1.737
	Bakola	24	6	6	1.00	1.60	-1.135	2	-1.956	0.084	0.892 [c]	-1.478
	SE Bantu	50	16	18	2.60	3.60	-0.857	6	-8.768**	0.077	0.328	-4.304*
	Dinka	23	15	12	1.90	4.10	-1.887*	12	-7.135**	0.065**	0.141	-6.735**
	Dogon	49	8	8	0.80	1.80	-1.472	6	-3.842**	0.060	0.011*	-1.038
NRY	Khoisan	25	10	7	2.50	2.40	0.089	1	0.183	0.133	0.514 [g]	-0.108
	Bakola	25	14	9	3.70	3.40	0.286	3	-0.127	0.137	0.428 [g]	-0.673
	SE Bantu	46	10	7	2.70	2.10	0.762	3	1.525	0.141	0.288 [g]	1.826
	Dinka	23	8	5	2.40	2.00	-0.696	2	1.771	0.162	0.000**	2.188
	Dogon	40	12	8	2.30	2.60	-0.402	4	-0.071	0.102	0.025*	1.300

* $p < 0.05$ ** $p < 0.001$

[a] Number of individuals.

[b] Number of polymorphic sites.

[c] Number of haplotypes.

[d] Calculated as % per gene per year.

[e] Probability of the SSD statistic (sum of the squared differences) from the mismatch distribution.

[f] Average Fu's Fs value calculated from 100 random subsample of 20 individuals.

[g] Populations for which 95% confidence intervals for Θ_0 and Θ_1 overlap, thereby rejecting population growth.

TABLE 3.2 Comparison of observed and simulated Fu's *Fs* values for mtDNA and the NRY

	mtDNA-NRY *Fs* [a]	5% Quantile [b]	Variance [c]	p [d]	mtDNA *s* [e]	NRY *s* [e]
Khoisan	-1.83	-5.15	9.85	0.263	11.3	9.60
Bakola	-1.83	-5.03	9.92	0.250	7.10	15.8
SE Bantu	-10.2	-6.49	15.7	0.009 *	22.5	10.1
Dinka	-8.91	-5.17	10.2	0.008 *	20.6	7.05
Dogon	-3.77	-5.82	12.7	0.131	9.86	12.6

[a] The observed difference between the mtDNA and NRY Fu's *Fs* values. Fu's *Fs* was calculated in DnaSP.

[b] The 5% lower bound of the distribution of simulated mtDNA - NRY Fu's *Fs* values.

[c] The variance of the distribution.

[d] The probability of the observed mtDNA-NRY Fu's *Fs* value.

[e] The average number of segregating sites generated from 10000 coalescent simulations (which can be compared to the observed values in **TABLE 1**).

* $p < 0.05$

TABLE 3.3 Population parameters estimated using GENETREE for constant size (upper row for each population) and exponential growth (lower row) demographic models.

Locus	Pop	Θ_{ml} [a]	β_{ml} [b]	N_e or N_o [c]	Likelihood Score (SE)	LRS [d]
mtDNA	Khoisan	2.97	--	4950	2.12E-10 (9.82E-12)	0.704
		3.90	1.12	6500	3.02E-10 (8.68E-13)	
	Bakola	1.90	--	3167	5.31E-07 (6.92E-09)	1.910
		4.50	6.00	7500	1.38E-06 (1.02E-08)	
	SE Bantu	5.00	--	8183	1.25E-16 (1.37E-18)	**6.494****
		10.50	9.05	17500	3.22E-15 (4.32E-17)	
	Dinka	5.54	--	9233	1.01E-07 (5.35E-09)	**6.720****
		17.50	14.4	29167	2.90E-06 (2.50E-07)	
	Dogon	2.21	--	3683	3.64E-07 (1.62E-09)	**4.783***
		7.70	15.15	12833	3.46E-06 (1.04E-08)	
NRY	Khoisan	2.54	--	2318	9.27E-11 (2.26E-12)	0.361
		2.50	0.10	2282	1.11E-10 (6.27E-13)	
	Bakola	4.16	--	3797	7.77E-13 (1.15E-14)	0.193
		5.00	0.80	4564	8.56E-13 (2.10E-14)	
	SE Bantu	2.30	--	2099	2.72E-11 (5.96E-13)	-0.231
		2.50	0.10	2282	2.42E-11 (6.28E-14)	
	Dinka	1.91	--	1743	1.55E-6 (2.34E-8)	0.057
		2.00	0.10	1826	1.59E-6 (2.19E-9)	
	Dogon	2.96	--	2702	1.02E-11 (1.98E-13)	-0.053
		3.00	0.10	2738	9.94E-12 (3.04E-13)	

* $p < 0.05$, ** $p < 0.01$

[a] Maximum likelihood estimate of Θ.

[b] Maximum likelihood estimate of the growth parameter, β.

[c] N_e for constant population size model, N_o for present day population size, exponential growth model.

[d] Likelihood ratio score.

FIGURE 3.1 Map showing the country of origin, the population name, the population name, abbreviation and the subsistence strategy (FP= food-producer, HG= hunter-gatherer) of the five sub-Saharan African populations.

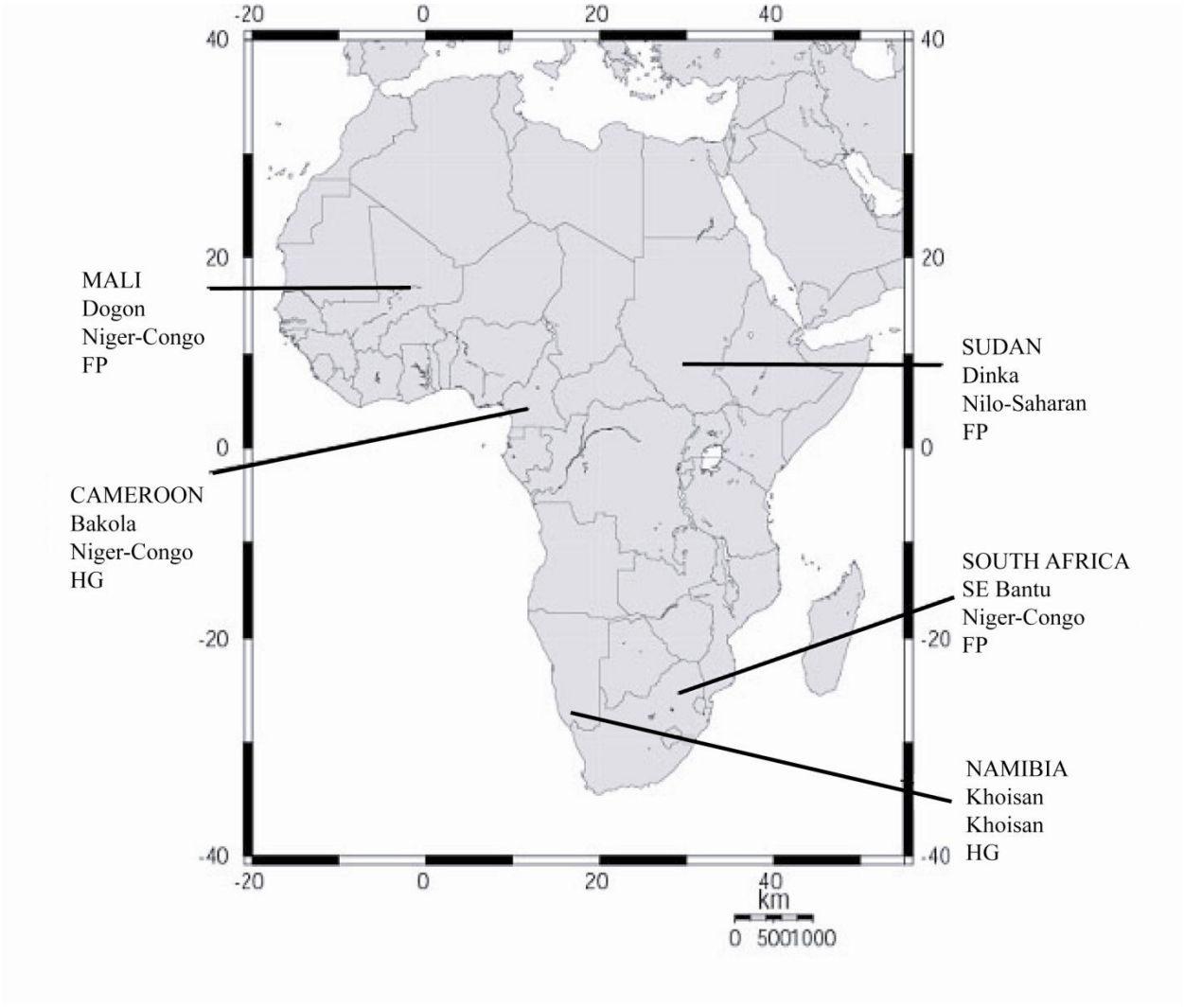
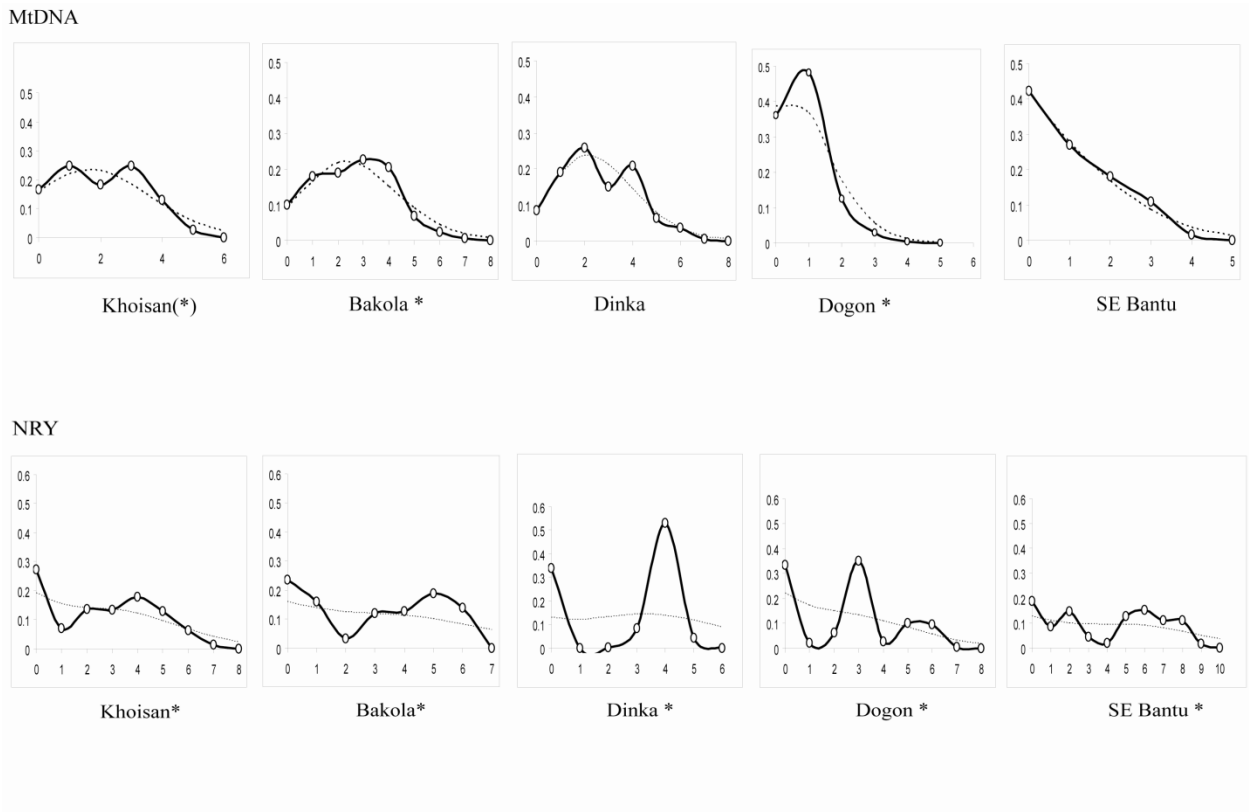


FIGURE 3.2 Mismatch distributions for mtDNA and NRY.



* Denotes rejection of growth model (by either significant SSD values or overlap of 95% confidence intervals for Θ_0 and Θ_1).

(*) Denotes the rejection of the growth model by the overlap of the 90% confidence intervals for Θ_0 and Θ_1.

SUPPLEMENTARY TABLE 3.1 MtDNA haplotype frequencies.

Haplotype	KHO	SEB	BAK	DGN	DNK	Total Individuals
1	0.04	0.06	0.00	0.00	0.00	4
2	0.16	0.00	0.00	0.00	0.00	4
3	0.12	0.00	0.00	0.00	0.13	6
4	0.08	0.00	0.00	0.00	0.00	2
5	0.36	0.14	0.00	0.00	0.00	16
6	0.08	0.00	0.00	0.00	0.00	2
7	0.12	0.00	0.00	0.00	0.00	3
8	0.04	0.00	0.00	0.00	0.00	1
9	0.00	0.02	0.00	0.00	0.00	1
10	0.00	0.20	0.00	0.47	0.21	38
11	0.00	0.04	0.00	0.00	0.00	2
12	0.00	0.20	0.00	0.00	0.04	11
13	0.00	0.12	0.67	0.39	0.26	47
14	0.00	0.02	0.00	0.00	0.00	1
15	0.00	0.02	0.00	0.00	0.00	1
16	0.00	0.02	0.00	0.00	0.00	1
17	0.00	0.04	0.00	0.00	0.00	2
18	0.00	0.02	0.00	0.00	0.00	1
19	0.00	0.02	0.00	0.00	0.00	1
20	0.00	0.02	0.00	0.00	0.00	1
21	0.00	0.02	0.00	0.00	0.00	1
22	0.00	0.02	0.00	0.00	0.04	2
23	0.00	0.02	0.00	0.00	0.00	1
24	0.00	0.00	0.13	0.00	0.00	3
25	0.00	0.00	0.04	0.00	0.00	1
26	0.00	0.00	0.04	0.00	0.00	1
27	0.00	0.00	0.04	0.00	0.00	1
28	0.00	0.00	0.08	0.00	0.00	2
29	0.00	0.00	0.00	0.02	0.00	1
30	0.00	0.00	0.00	0.04	0.04	3
31	0.00	0.00	0.00	0.02	0.00	1
32	0.00	0.00	0.00	0.02	0.00	1
33	0.00	0.00	0.00	0.02	0.00	1
34	0.00	0.00	0.00	0.02	0.00	1
35	0.00	0.00	0.00	0.00	0.04	1
36	0.00	0.00	0.00	0.00	0.04	1
37	0.00	0.00	0.00	0.00	0.04	1
38	0.00	0.00	0.00	0.00	0.04	1
39	0.00	0.00	0.00	0.00	0.04	1
40	0.00	0.00	0.00	0.00	0.04	1

Total	**25**	**50**	**24**	**49**	**23**	**171**

SUPPLEMENTARY TABLE 3.2 NRY haplotype frequencies.

Haplotype	KHO	SEB	BAK	DGN	DNK	Total Individuals	Lineage
1	0.20	0.26	0.12	0.00	0.13	23	B
2	0.00	0.00	0.12	0.00	0.00	3	A1B
3	0.00	0.00	0.20	0.00	0.00	5	E3A7A
4	0.04	0.33	0.24	0.43	0.04	40	EP1
5	0.00	0.00	0.04	0.00	0.00	1	E3A7A1
6	0.00	0.09	0.04	0.03	0.43	16	E2B
7	0.00	0.00	0.08	0.00	0.00	2	B2
8	0.04	0.24	0.12	0.03	0.00	16	E3A7A
9	0.00	0.00	0.04	0.00	0.00	1	E3A8A3
10	0.00	0.00	0.00	0.08	0.00	3	B2A2
11	0.00	0.04	0.00	0.03	0.04	4	E
12	0.00	0.00	0.00	0.38	0.00	15	E1B
13	0.00	0.00	0.00	0.03	0.00	1	A1A
14	0.00	0.00	0.00	0.03	0.00	1	EP1-excluded here
15	0.20	0.02	0.00	0.00	0.35	14	A3B1
16	0.16	0.00	0.00	0.00	0.00	4	A
17	0.20	0.00	0.00	0.00	0.00	5	A2B
18	0.16	0.00	0.00	0.00	0.00	4	A3B1A
19	0.00	0.02	0.00	0.00	0.00	1	E5
Total	**25**	**47**	**25**	**40**	**23**	**158**	

CHAPTER 4: SEX-SPECIFIC DEMOGRAPHIC PROCESSES IN SUB-SAHARAN AFRICA: EVIDENCE FOR WIDESPREAD MALE MIGRATION WITH REPLACEMENT

ABSTRACT

Previously, differing signals of population growth based on mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY) in sub-Saharan African populations were detected. Food-producing populations best fit a model of population growth for mtDNA but not the NRY, while hunter-gathering populations best fit a model of population stationarity for both loci. These results led me to hypothesize that food-producing males experienced either smaller effective population sizes and/or lower migration rates in Africa during a recent phase of demographic/range expansion which could have erased the signal of a past population expansion. To assess the relative effects of migration and effective population size on patterns of mtDNA and NRY nucleotide variability, I examined re-sequencing data from the mtDNA *COIII* gene and non-coding data from the NRY in 172 individuals from five distinct sub-Saharan African populations. Since past studies have been hampered by the co-estimation of effective population size and migration rates, I used an Isolation with Migration (IM) model to disentangle estimates of effective population size and migration. I find that in food-producing populations, male migration rate estimates are in fact higher, not lower, than those of females, while estimates of male effective population size are strikingly small. I infer that males have experienced a period of population size reduction due to replacement, and that this most likely occurred during the Bantu expansions, approximately 5,000 years ago.

INTRODUCTION

What is known about current population structure (i.e., gene flow and effective population size) in Africa comes mostly from mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY). Though both genomic compartments are uniparentally inherited, they are found to have distinct patterns of population structure. High levels of mtDNA differentiation have been indicated by previous studies of sub-Saharan African populations based on the large number of differences between mtDNA and nuclear data (Jorde et al. 1995; Mountain 1998). Support for distinct demographic histories for mtDNA and the NRY was also found in a recent study of 40 African populations, which showed a strong correlation between the patterns of Y chromosomal population structure and linguistics and mtDNA population structure and geography (Wood et al. 2005). But variation in mtDNA and NRY genetic structure is quite complicated and is deeply rooted in subsistence strategies; that is, the NRY reflects lower levels of population structure in food-producing groups than in hunter-gatherers (Destro-Bisol et al. 2004b).

Distinct patterns of population expansion were also inferred from the mtDNA and NRY data. A history of African population expansion is supported by mtDNA polymorphism data for food-producing populations and one of constant population size for hunter-gathering populations (Excoffier and Schneider 1999; Metni Pilkington et al. 2008). This is in sharp contrast to the NRY, where both hunter-gatherers and food-producers best fit a model of constant population size (Metni Pilkington et al. 2008). The distinct mtDNA and NRY polymorphism patterns most likely reflect sex-biased demographic processes in the recent history of African populations. It is possible that in Africa males experienced smaller effective population sizes (due to polygyny) and/or lower migration rates (due to patrilocality) in the past; however, the individual effects of

differential migration and effective population size on nucleotide variability have not been examined.

Previous studies estimated N_e (the effective population size) and m (the migration rate) together rather than as separate factors affecting patterns of population structure. Here we examine the relative effects of migration and effective population size on patterns of variability for mtDNA and the NRY. Generally, the number of migrants has been indirectly inferred using hierarchical *F*-statistics (Wright 1969), based on the underlying model of population structure (for instance, $F_{ST} = 1/4N_em + 1$ under Wright's island model). Yet, there are several implicit assumptions of Wright's island model that are often violated in the study of human populations, such as: equally-sized subpopulations, infinitely many subpopulations, and symmetrical migration (Whitlock and McCauley 1999). Estimating the number of migrants from *F*-statistics can be also be problematic because results are dependent on the particular model of population structure assumed, which may not necessarily be the appropriate model for the data under examination (Templeton, Routman, and Phillips 1995).

Newer methods incorporating the Isolation with Migration (or IM) model, were developed to allow for more realistic scenarios whereby all subpopulations are not the same size and migration is not symmetrical in two populations that split from a common ancestor (Beerli and Felsenstein 1999; Bahlo and Griffiths 2000; Wakeley 2000). These approaches involve calculating posterior probability estimates of migration rates and effective population sizes using the coalescent (Kingman 1982). The use of these methods also permits a focus on effective population size and migration rates separately, instead of estimating the number of migrants together (Nielsen and Wakeley 2001). This can be highly relevant for disentangling the effects of high variance in the reproductive success of one of the sexes and sex-specific migration, both of which greatly affect

our abilities to accurately estimate past demographic events such as ancient population expansions.

I use the IM coalescent-based method to assess the relative roles of N_e and m in shaping the patterns of genetic diversity of mtDNA and the NRY. To this end, I compare previously generated re-sequencing data from 6,601 basepairs (bp) of DNA sequence from the non-recombining portion of the Y chromosome (NRY) to 780 bp of the mitochondrial (mtDNA) *cytochrome c oxidase subunit III* (*COIII*) gene in five sub-Saharan populations. No study has directly estimated migration rates by examining sequence data from the same populations for both mtDNA and the NRY prior to this work. Ascertainment bias is corrected for by the use of direct mtDNA and NRY re-sequencing data, and sample heterogeneity is avoided by comparing the same populations across loci.

MATERIALS AND METHODS

Subjects and Molecular Loci

All sequence data for this study were taken from previously published work (Metni Pilkington et al. 2008). The mtDNA *COIII* gene and the non-coding NRY *Alu* insertions were surveyed in a panel of 172 unrelated males, representing five African populations (**FIGURE 4.1**). These populations comprise the Dogon of central Mali (abbreviated DGN, n= 39-49), the Dinka of southern Sudan (DNK, n=23), the Bakola of southern Cameroon (BAK, n=24-25), the Khoisan of Namibia (KHO, n=25), and a group of Southeast Bantu speakers (SE Bantu) from southern Africa (SEB, n=47-50) (**TABLE 4.1**). For the purposes of these analyses, populations were separated into three linguistic groups: Niger-Congo (Bakola, Dogon, SE Bantu), Nilo-Saharan (Dinka), and Khoisan (Khoisan), and into three geographic regions: western (Bakola,

Dogon), eastern (Dinka) and southern (Khoisan, SE Bantu). Here the Khoisan and Bakola are classified as hunter-gatherers (HG) and the SE Bantu, Dinka, and Dogon as food-producers (FP). Additionally, orthologous DNA sequences from one common chimpanzee (*Pan troglodytes*) were analyzed.

Isolation with Migration (IM)

The Isolation and Migration (IM) model of population structure is one of population splitting in which there is an ancestral population that gives rise to two descendent populations which may be connected by gene flow. Here I employ the method modified by Hey and Nielsen (2004) for multiple loci, which is an extension of the method developed by Nielson and Wakeley (2001) for data from a single, non-recombining locus. By exploring the range of possible gene trees that are consistent with a given dataset, this method is able to capture many of the dynamics that occur during the early stages of population divergence. It has seven main parameters: effective population sizes for the ancestral population (N_{eA}), and two descendant populations (N_{e1} and N_{e2}), migration rates between the descendant populations (m_1 and m_2), the time at which the ancestral population gave rise to the descendant populations (the splitting time, t), and the fraction of the ancestral population that forms N_1 (sN_A) (**CHAPTER 2**, **FIGURE 2.4** in **MATERIALS AND METHODS**). Each of these parameters can be scaled by the rate of genetic drift or by the mutation rate. It has been noted by Nielsen and Wakeley (2001) that often the most difficult parameter to estimate properly is that of the splitting time.

IM uses a Markov chain Monte Carlo (MCMC) approach to take into account the stochastic variance among loci. For multiple loci, the model assumes that loci are independent, neutrally evolving, and non-recombining (or, at least do not show any evidence of recombination). When

there is evidence of recurrent mutation such that the four-gamete test is violated, individuals (rather than sites) are removed from the analysis (Hey and Nielsen 2004). Since IM works by comparing two populations at a time, it is not always the same individuals that violate the four-gamete test (i.e., sites change according to population pairwise comparison). The total number of sites examined in all comparisons was 21 (NRY) and 38 (mtDNA).

I ran the analyses with four loci: mtDNA, the NRY and two X-linked loci that are not included in results presented here. For the purposes of this study migration rates were estimated per locus. This allows for better resolution of the migration rates for mtDNA and the NRY since the method relies on estimating multiple parameters concurrently (i.e., it uses the multilocus posterior probability estimate of N_{e1}, N_{e2} and N_{eA}, t and sN_{eA} to estimate the per locus migration rates). The migration estimates obtained by running the loci jointly were similar to those obtained by running the loci separately. They differ slightly because the different parameters are not independent of each other (i.e., when all loci are constrained to share common population size and splitting time, their apparent migration histories might be affected). Estimates of population mutation rate parameters (Θ_1, Θ_2, and Θ_A) were converted into estimates of effective population size (N_{e1}, N_{e2} and N_{eA}), using a measure of the mutation rate on a scale of generations (a 25 year generation length was assumed for humans). The split time parameter, t, was converted into years by dividing the posterior probability estimate of t by the geometric mean number of mutations per year per gene. Population divergence times are reported in thousands of years (KYR). A six million year divergence from chimpanzee was assumed for all calculations.

All runs were set to have a "burn-in" time of 10 million steps, and use a two-step linear heating mode. Multiple Markov chains were run to help ensure proper mixing of the chains.

Datasets for all population pairs were run with the following uniform priors: Θ_1 (0-10), Θ_2 (0-10), Θ_A (0-10), *t* (0-4), m_1 (0-10), and m_2 (0-10); these priors were adjusted after the initial run to contain the peak of the marginal posterior probability distribution. Once reasonable ranges for all parameters were obtained, each population pair was run at least three times with up to six Metropolis-coupled chains and 10 million to 20 million steps to ensure proper convergence of results. All runs specified the changing population size model which includes the *s* parameter (per population, not locus), or the proportion of the ancestral population that founded the descendent population. All posterior parameter estimates presented here are the means of the replicate runs for each population comparison, and include no fewer than three runs. All confidence intervals presented are the 90% highest posterior density (HPD) intervals, or the boundaries of the shortest spans that include 90% of the probability density of the parameter estimate.

Population Genetic Analyses

Population parameters for mtDNA and the NRY such as haplotype frequency and haplotype diversity were calculated using the program DnaSP ver. 4.0 (Rozas et al. 2003). AMOVAs and F_{ST} for the haploid loci were calculated using Arlequin ver. 3.0 and include molecular distances (Excoffier, Laval, and Schneider 2005). The Kimura Two-Parameter model (Kimura 1980) of DNA substitution was used for the NRY data and the Tamura Nei model (Tamura and Nei 1993) for mtDNA data, with the gamma shape parameter α=0.22. Unweighted pair group method with arithmetic mean (UPGMA) trees of F_{ST} estimates for mtDNA and NRY datasets were estimated using MEGA version 3.1 (Kumar, Tamura, and Nei 2004)

RESULTS

IM Analyses—Migration and Effective Population Size

Results from the IM coalescent simulations estimating migration rates for mtDNA and the NRY are presented in **TABLE 4.1** and **FIGURE 4.2A**. There are only four significantly non-zero migration rate estimates for mtDNA, while there are 15 for the NRY. The general pattern of migration based on mtDNA is one of non-reciprocal migration (in an eastward and southward direction), while that of the NRY tends to be reciprocal across the entire range of the area examined. For mtDNA, high migration rate estimates are found from the Dogon into the Dinka (22.672) and from the Dogon into the SE Bantu (8.900), and in these two cases the 90% HPD confidence intervals for the migration rates between the two populations do not overlap, meaning that there is strong support for unidirectional migration. There are also high mtDNA migration rates from the Dinka into the SE Bantu (11.583) and from the Bakola into the Dinka (10.142). There are only two mtDNA comparisons where the migration appears to be bidirectional at low levels (Dinka-SE Bantu and Dogon-SE Bantu). For mtDNA, there is more migration out of hunter-gathering populations and more migration into food-producing populations, as expected (**TABLE 4.2**).

The migration patterns based on the NRY are strikingly different. The highest NRY migration rate estimate is from the SE Bantu into the Bakola (24.540), which is noteworthy given the large geographic distance between the two populations. The next highest migration rates are from the SE Bantu into the Dogon (6.505), from the SE Bantu into the Dinka (4.255), and from the Dinka into the Dogon (3.122). It is remarkable that in most cases, NRY migration is bi-directional, as opposed to that of mtDNA. Also in contrast to mtDNA, the overall NRY

pattern supports more migration into hunter-gathering populations and more migration out of food-producing populations.

To further address why food-producers do not show the signal of a population expansion based on the NRY (but do from mtDNA), NRY estimates of N_{e1}, N_{e2},and N_{eA} are presented (**TABLE 4.3**). Results indicate that in all comparisons involving only food-producing populations $N_{e1} + N_{e2} < N_{eA}$. Comparisons that involve a hunter-gatherer population are the only ones to result in $N_{e1} + N_{e2} > N_{eA}$. The pattern is the opposite for mtDNA, where the food-producing population comparisons always produced N_{eA} estimates that were lower than $N_{e1}+N_{e2}$ (data not shown). The implications of these findings are discussed further below.

Shared Haplotypes

A summary of the haplotype data for mtDNA and the NRY is presented in **TABLE 4.4**. The number of mtDNA haplotypes range from six for the Bakola to 18 for the SE Bantu and the NRY haplotypes range from five for the Dinka to nine for the Bakola. Consequently, mtDNA haplotype diversity is highest for the SE Bantu and NRY haplotype diversity is highest for the Bakola. The average haplotype diversity for the mtDNA and NRY data are similar (0.76 versus 0.78).

The haplotype frequencies and distributions (**FIGURE 4.2B** and **SUPPLEMENTARY TABLE 2.1** and **SUPPLEMENTARY TABLE 2.2**), reflect a great deal more NRY than mtDNA shared haplotypes in these populations sampled over a large geographic range. The Y chromosome Haplogroup E is shared among all populations, is likely associated with the Bantu expansions (Wood et al. 2005). The Y chromosome Haplogroup A is also shared among all five populations, though the Khoisan, SE Bantu, and Dinka alone share a subset of Haplogroup A

(A3b1). The Y chromosome Haplogroup B is shared among all populations except the Dogon. When populations are separated by subsistence strategy, the NRY of hunter-gatherers contain a large proportion of haplotypes that are unique to these populations (this is also true of the Dogon), whereas most of the food-producer's haplotypes are shared across very large distances (in fact, the Dinka have no unique haplotypes and the SE Bantu have only one at very low frequency).

Though the SE Bantu and Dinka have no (or low) unique haplotypes for the NRY, these same two food-producing populations have many unique mtDNA haplotypes (**FIGURE 4.2B)**. It is notable that there are no mtDNA haplotypes that are shared among all five populations. There is just one haplotype that is shared among four of the populations (i.e. excluding the Khoisan) at high frequencies (Haplotype 13), and one that is shared among three populations, the Dogon, Dinka, and SE Bantu (Haplotype 10). The geographically proximal Khoisan and SE Bantu share two mtDNA haplotypes at relatively high frequencies (Haplotypes 1 and 5). It is remarkable that the Khoisan and the Dinka share a mtDNA haplotype (Haplotype 3) to the exclusion of the other populations, and that it appears to be an ancestral haplotype.

Population Differentiation-- F_{ST} and AMOVA

Estimates of mtDNA and NRY F_{ST} values are presented in **FIGURE 4.3**, and reflect the information presented by the shared haplotypes. For mtDNA, the Dogon-Dinka pair has the lowest F_{ST} value (-0.041, effectively 0), despite great geographic distance. Two other populations showing close maternal genetic affinity are in geographical proximity: the Khoisan-SE Bantu (0.135) and the Dogon-Bakola (0.142). The populations that have the most differentiated mtDNA are the Dogon-Khoisan (0.701) and the Khoisan-Dinka (0.557), and are

also quite distant geographically. Comparisons of NRY F_{ST} values in these same populations revealed slightly different patterns (Wilcoxon rank test, $p = 0.11$). Based on the NRY, the populations that show the closest affinities are the Niger-Congo speakers: the Bakola-SE Bantu (0.003), Dogon-SE Bantu (0.070), and again the Dogon-Bakola (0.088). Those populations that appear to be the most differentiated based on the NRY are the Khoisan-Bakola (0.331), Khoisan-SE Bantu (0.358), and again the Dogon-Khoisan (0.404). In sum, the F_{ST} results support low levels of NRY population differentiation among the Niger-Congo speaking populations, whereas maternal ancestry reflects populations that are more similar based on geographical proximity, as illustrated by the UPGMA trees based on F_{ST} values (**FIGURE 4.4**)

The pattern observed from AMOVA for the NRY is again opposite of the pattern found for mtDNA; linguistics appears to influence NRY genetic structure more than geography. The overall mtDNA Φ_{ST} is 0.357, while the overall NRY Φ_{ST} is 0.218 (**TABLE 4.5**). For mtDNA, when populations are grouped according to linguistic affiliation, Φ_{SC} (differentiation among populations within groups) is much higher than when grouped according to geographic location (0.290 *versus* 0.134), whereas for the NRY, Φ_{SC} is much lower than when grouped by geographic location (0.057 *versus* 0.246). For mtDNA, when populations are grouped by geographic location, Φ_{CT} (or differentiation among groups) is nearly twice that when grouped by linguistic affiliation (0.305 *versus* 0.152), whereas for the NRY, Φ_{CT} is smaller than when grouped by linguistic affiliation (-0.048 *versus* 0.256). Taken together, these data suggest that linguistics and geography are playing different roles in shaping NRY and mtDNA variation in these populations.

DISCUSSION

The goal of this work is to assess the relative effects of migration and effective population size on patterns of population growth for mtDNA and the NRY in hunter-gatherer and food-producing populations. While previous studies have described differing patterns of mtDNA and NRY population structure in sub-Saharan Africa, this is the first to examine direct estimates of gene flow from mtDNA and the NRY separately from effective population size. The IM coalescent analyses indicate two main explanations for the signal of population stationarity from the NRY of food-producing populations: (1) there is more out-migration than in-migration for the NRY of food-producers, which is the opposite of the mtDNA pattern, and (2) estimates of the effective population size of food-producers' NRYs is quite low in comparison to the mtDNA effective population size of food-producers. Moreover, estimates of male migration rates are high (not low) in comparison to those of females, thus higher rates of patrilocality in food-producing populations do not appear to be lowering migration rates for the populations studied here.

Similar to previous analyses, mtDNA differentiation is more strongly influenced by geography and NRY differentiation is influenced more by language, with very low levels of population differentiation among the NRYs of Niger-Congo speakers even across very large distances. AMOVA results revealed an overall mtDNA Φ_{ST} of 0.357; much greater than those of Wood et al. (2005) and Destro-Bisol et al. (2004b) (Φ_{ST} = 0.15 and 0.172, respectively), but comparable to that Wilder et al. (2004) (Φ_{ST} = 0.320), which examined a similar dataset. The overall Φ_{ST} for the NRY (0.218) was slightly lower than that of Wood et al. (2005) and Hammer et al. (1997) (Φ_{ST} = 0.32 and 0.34, respectively), but greater than that of Destro-Bisol et al. (2004) (Φ_{ST} = 0.100), and comparable to that of Wilder et al. (2004) (Φ_{ST} = 0.208). My findings

are similar to many previous studies demonstrating lower levels of structure for males than females and more genetically homogenous male populations (Jorde et al. 1995; Hammer et al. 1998; Lum et al. 1998; Mountain 1998; Destro-Bisol et al. 2004b; Wilder et al. 2004), but differ from several key studies of sub-Saharan Africans (Lane et al. 2002; Wood et al. 2005). The differences most likely stem from four main sources: previous studies (1) did not compare the same populations for mtDNA and the NRY, (2) used autosomal data to estimate female migration, (3) used different detection assays for mtDNA and NRY (such as direct sequencing *versus* SNP typing or microsatellites), or (4) used grid-based sampling instead of population-based sampling.

The Bantu Expansions— Male Migration with "Replacement" Model

Studies of archaeology and linguistics indicate expansions of Bantu-speaking peoples from the area of present-day Cameroon eastward and southward across Africa occurring as recently as 3,000-5,000 years ago (Posnansky 1968; Greenberg 1972; Phillipson 1993; Cavalli-Sforza, Menozzi, and Piazza 1994; Poloni et al. 1997; Beleza et al. 2005; Wood et al. 2005; Rexova, Bastin, and Frynta 2006, **FIGURE 4.5**). Recent genetic work has suggested a strong correlation between Bantu languages and Y chromosomal variation, similar to results presented here (Wood et al. 2005). Yet, if males and females dispersed similarly during the Bantu expansions, then there should be similar associations between genetics, linguistics, and geography, similar levels of population structure, and similar migration rates based on mtDNA and the NRY. The sex-biased results (lower NRY N_em) presented here and elsewhere have caused researchers to posit that they resulted from the Bantu expansions (Destro-Bisol et al. 2004b; Wood et al. 2005; Metni Pilkington et al. 2008). Lower NRY N_em could result from

lower NRY migration rates (possibly due to patrilocality) or lower NRY effective population sizes (possibly due to polygyny), or a combination of the two factors.

In many patrilocal African societies marriage generally takes place between farmer males and hunter-gatherer females, with the children usually taking up residence in the male's village (Destro-Bisol et al. 2004b). It is unlikely that this factor alone has led to my results based on multiple lines of evidence suggesting that male migration has not been low in the past. First, and most importantly, we estimated that males are actually migrating more than females based on the greater number of significant NRY than mtDNA migration rates (15 for the NRY *versus* four for mtDNA, **FIGURE 4.2A**). Likewise, if NRY migration rates were low, then one would expect to see higher NRY than mtDNA F_{ST} values—the opposite of what is observed (**FIGURE 4.3**). And finally, there is similar pattern between NRY genetic variation and linguistics (but no such pattern for mtDNA, **TABLE 4.5** and **FIGURE 4.4**), suggesting that the expansion of males speaking Bantu languages acted to homogenize males along the path of this dispersion event either through replacement and/or polygyny, as noted by previous studies (Destro-Bisol et al. 2004b; Wood et al. 2005).

There are multiple lines of evidence that indicate a dramatic reduction in male effective population size in food-producing populations in comparison to those of females. Previous work estimated the effective population sizes in these same food-producing populations, and found the average NRY N_e to be far lower than the mtDNA N_e (~1,700 versus ~19,800) (Metni Pilkington et al. 2008). The newer estimates of N_e presented in this work can be compared to the previous study (which did not take into account migration). Theoretically, if estimates of N_e are the same, then migration is likely affecting estimates of population growth, since the IM model takes into account migration when estimating N_e. If the estimates for N_e differ, then migration is likely not

affecting the results. We find that the N_e estimates differ for the mtDNA data but are the same for the NRY, suggesting that migration has affected the estimates of population growth for the NRY. Since migration is high, and migration under the IM model is considered replacement by definition (since migration rates refer to the "proportion of migrants" in a population), then we can infer that the very low N_e estimates from IM and GENETREE are the result of a major replacement event in male food-producers (see the **SUPPLEMENTARY MATERIAL** for more on effective population size and TMRCA).

A "migration with replacement" model can be tested empirically. Under such a model, one would expect that $N_{e1} + N_{e2} < N_{eA}$ (**FIGURE 4.6A**), which is also a model of population contraction. This is precisely what is observed every time the effective sizes of the NRY of food-producing populations are compared to each other, as opposed to when the effective sizes of the NRY of the hunter-gathering populations are compared to each other (where $N_{e1} + N_{e2} > N_{eA,}$ **FIGURE 4.6B**). The replacement of males through high rates of migration can reduce the effective population sizes of the NRY as compared to mtDNA. Moreover, if polygyny caused there to be a higher variance in male reproductive success as males married local females along the path of the Bantu expansions, then the effective size of the NRYs would be further reduced compared to mtDNA. This effect is also likely to be more pronounced in food-producing than hunter-gathering populations, since (1) the Bantu expansions are thought to have been predominantly composed of food-producing populations, and (2) food-producers tend to have higher rates of polygyny than hunter-gatherers (Destro-Bisol et al. 2004b). My estimates of effective population size from the NRY are low for all populations, while those from mtDNA are high for food-producers and low for food-gatherers (**TABLE 4.2**). I find that for the NRY, food-producing males have higher rates of out-migration while hunter-gatherer males appear have

higher rates of in-migration. The opposite is true of the mtDNA, where food-producing females have higher rates of in-migration while hunter-gatherers have higher rates of out-migration, which follows a well known pattern of migration of females from hunter-gatherer populations (Destro-Bisol et al. 2004b). It is likely that as Bantu farmers migrated east and south, males took new wives along the way and consequently contributed to the local gene pools (Wood et al. 2005), replacing males along the path of the expansion and homogenizing the gene pool of males in sub-Saharan Africa.

CONCLUSIONS

While earlier studies have shown conflicting results concerning sex-specific migration in sub-Saharan African populations (i.e., some find higher rates of male migration while others find higher rates of female migration), this is the first to use data assayed in a similar way for both mtDNA and the NRY. Analyses of these haploid loci using newer MCMC methods correlate well with those based on shared haplotypes, F_{ST}, and AMOVA in indicating that Y chromosome population structure is less pronounced than that of mtDNA. Previous work showed no evidence of a population expansion from the NRY of these food-producing populations, while the same populations showed an expansion based on mtDNA. We postulated that this was because of a reduction in male effective population size (possibly due to a high variance in male reproductive success) and/or lower migration rates (due to patrilocality) in males. When migration rates are estimated separately from effective population size there are in fact higher (not lower) migration rates for males than females; however, estimates of effective population size are surprisingly low for the NRY. These new analyses demonstrate that data from the NRY of food-producing populations fit a model of population contraction (likely due to replacement), since the sums of

the effective population sizes of food-producers are always lower than the ancestral effective population size. I suggest that the Bantu expansions that occurred approximately 5,000 years ago likely served to homogenize the male gene pool across the continent through high rates of male migration and widespread male replacement.

TABLE 4.1 IM estimates of migration rates from mtDNA and the NRY (with 90% HPD confidence intervals in parentheses). Bold values indicate estimates where the 90 % HDP interval does not include zero. Populations were arbitrarily assigned to be population 1 (listed first) and population 2 (listed second).

	mtDNA		NRY	
Comparison	m_1 [a]	m_2 [b]	m_1 [a]	m_2 [b]
BAK x DGN	0.010 (0.002 - 4.080) [c]	0.012 (0.002 - 4.112) [c]	**2.985 (0.890 - 4.997)**	**1.742 (0.345 - 4.840)**
BAK x KHO	0.010 (0.002 - 1.757) [c]	0.397 (0.002 - 2.465)	0.880 (0.007 - 4.090)	**1.415 (0.605 - 4.582)**
BAK x DNK	0.045 (0.002 - 3.820) [c]	**10.142 (1.056-21.098)**	**1.627 (0.460 - 4.610)**	**1.617 (0.022 - 7.898)**
BAK x SEB	0.020 (0.010 - 4.250) [c]	1.325 (0.005 - 7.480)	**24.540 (10.50 - 39.980)**	0.071 (0.005 - 9.331)
KHO x DGN	0.436 (0.002 - 2.280)	0.034 (0.001 - 1.348) [c]	**0.902 (0.122 - 3.142)**	**0.576 (0.177 - 1.999)**
KHO x DNK	0.537 (0.003 - 3.171)	0.010 (0.005 - 2.250) [c]	0.804 (0.003 - 5.595)	**2.820 (0.520 - 8.565)**
KHO x SEB	0.547 (0.005 - 8.320)	0.092 (0.002 - 4.997) [c]	**2.795 (0.650 - 8.025)**	0.530 (0.002 - 3.422)
DNK x DGN	**22.672 (11.777-34.982)**	0.242 (0.003 - 5.518) [c]	**0.897 (0.032 - 4.112)**	**3.122 (1.029 - 6.186)**
DNK x SEB	0.721 (0.005 - 5.491)	**11.583 (4.323-19.900)**	**4.255 (1.427 - 5.997)**	**3.450 (0.490-13.330)**
DGN x SEB	0.035 (0.005 - 3.080)	**8.900 (3.600 - 9.995)**	**6.505 (2.900 - 9.995)**	0.365 (0.005 - 6.975)

[a] Migration is into population 1.

[b] Migration is into population 2.

[c] The estimated migration rate is effectively 0 because it was at the lowest range that IM could detect.

TABLE 4.2 MtDNA and NRY migration rate ratios and effective population size estimates from GENETREE and IM for hunter-gatherers (HG) and food-producers (FP). Migration rate ratios are the ratio of in-migration to out-migration, and were calculated from IM. Effective population sizes estimated in GENETREE used the best-fit model of population size change: constant population size or exponential growth.

Population	**mtDNA**	**NRY**	**Reference**
HG *m* ratio	0.166	3.726	This work
HG N_e (IM)	~18,000	~3,070	This work
HG N_e (GENETREE)	~4,060	~2,400	Pilkington et al., 2008
FP *m* ratio	1.219	0.445	This work
FP N_e (IM)	>50,000	~1,740	This work
FP N_e (GENETREE)	~19,800	~1,700	Pilkington et al., 2008

TABLE 4.3 Estimates of NRY N_{e1} (effective population size of the first population listed), N_{e2} (effective population size of the second population listed), and N_{eA} (ancestral effective population size) for the NRY. For bolded populations: $N_{e1}+N_{e2}< N_{eA}$.

Comparison	N_{e1}	N_{e2}	N_{eA}	$N_{e1}+N_{e2}$
BAK x DGN	5,040	1,310	6,970	**6,350**
BAK x KHO	5,770	1,800	1,130	7,580
BAK x DNK	3,230	390	5,620	**3,620**
BAK x SEB	2,290	3,390	5,370	5,680
KHO x DGN	2,560	2,990	2,860	5,550
KHO x DNK	2,400	1,200	4,640	**3,600**
KHO x SEB	1,490	2,270	2,400	3,760
DNK x DGN	1,530	1,300	10,190	**2,830**
DNK x SEB	1,370	1,200	4,410	**2,570**
DGN x SEB	760	3,240	4,851	**4,000**

TABLE 4.4 Number of haplotypes and haplotype diversity data for mtDNA and the NRY, with the mean below.

		COIII			*NRY*		
Population	**Linguistic Family**	***N*** [a]	**H** [b]	**HD** [c]	***N*** [a]	**H**	**HD**
Khoisan	Khoisan	25	8	0.83	25	7	0.86
Bakola	Niger-Congo	24	6	0.55	24	9	0.88
SE Bantu	Niger-Congo	50	18	0.90	47	7	0.78
Dogon	Niger-Congo	49	8	0.64	39	7	0.69
Dinka	Nilo-Saharan	23	12	0.89	21	5	0.70
Mean		34.2	10.4	0.76	31.2	7	0.78

[a] Number of samples.

[b] Number of haplotypes.

[c] Haplotype diversity.

TABLE 4.5 Analysis of molecular variance for mtDNA and the NRY (number of comparisons in parentheses).

Group	*No. of Groups*	Φ_{ST}	Φ_{SC}	Φ_{CT}
mtDNA total	**1**	**0.357 (166)***		
Linguistic groups	3	0.398 (166)*	0.290 (2)*	0.152 (2)
Geographic groups	3	0.399 (166)*	0.134 (2)*	0.305 (2)
NRY total	**1**	**0.218 (154)***		
Linguistic groups	3	0.299 (154)*	0.057 (2)*	0.256 (2)
Geographic groups	3	0.210 (154)*	0.246 (2)*	-0.048 (2)

* $P < 0.000$.

FIGURE 4.1 Map showing the country of origin, the population name, the language family and the subsistence strategy of the five sub-Saharan African populations. (Abbreviations: FP= food-producer, HG= hunter-gatherer, W= west, E= east, and S= south). Circles indicate geographic clustering.

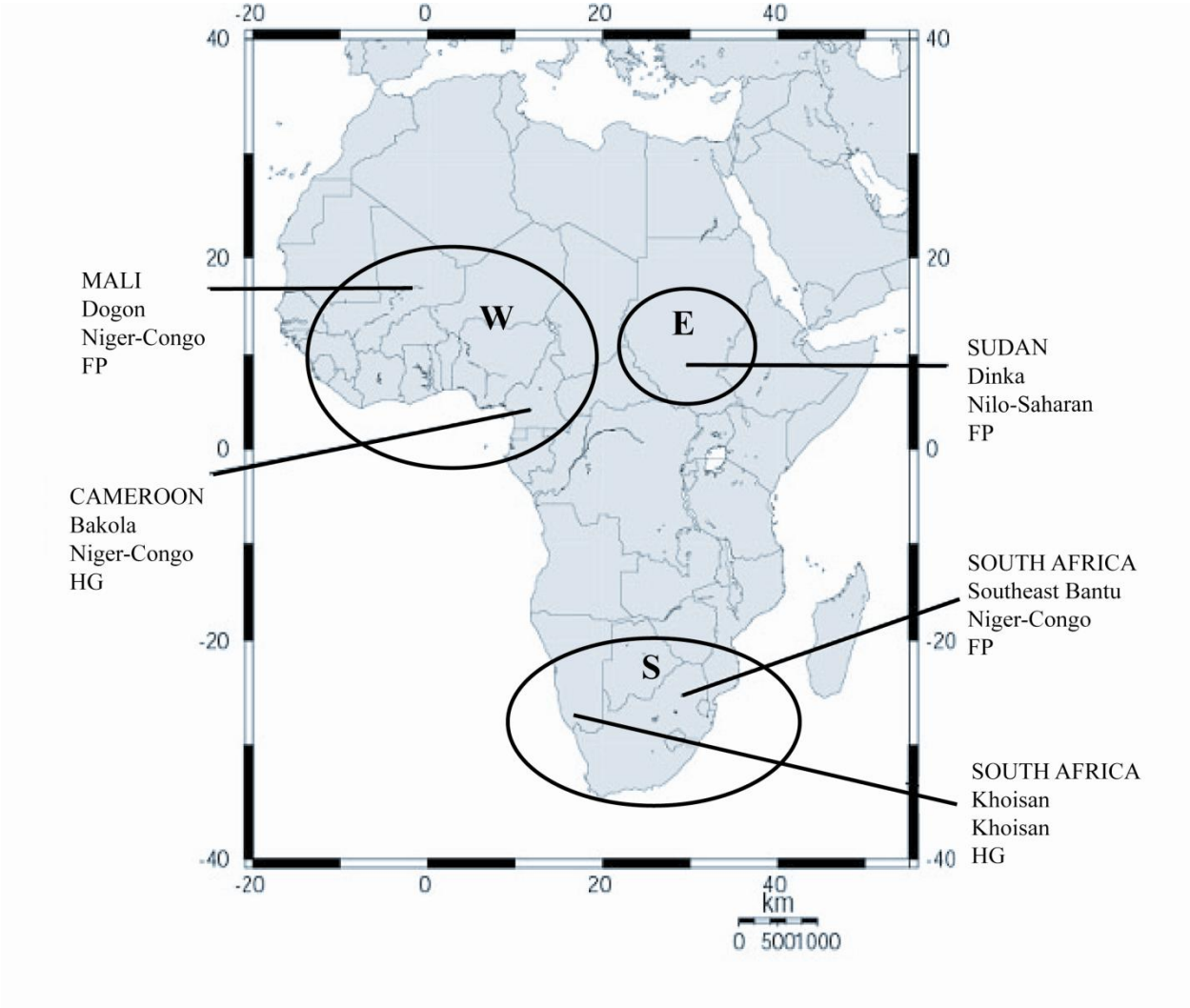


FIGURE 4.2 Distribution of mtDNA and NRY **(A)** Population migration rates (only migration rates whose 90% highest posterior density interval did not include zero are shown), and **(B)** shared haplotypes (unique haplotypes shown in grey), with NRY Haplogroups A, B, and E denoted.

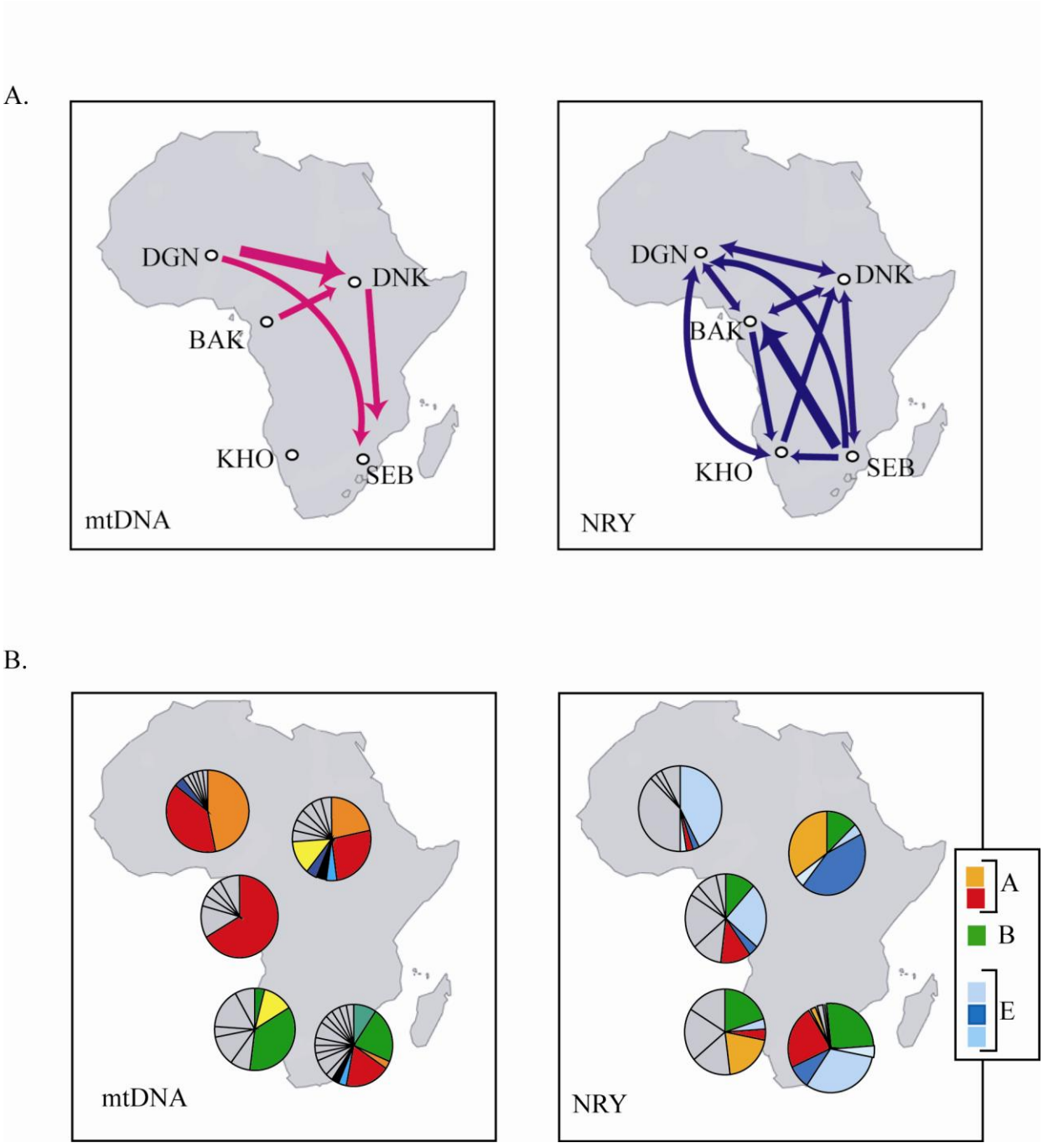


FIGURE 4.3 F_{ST} estimates for mtDNA and NRY, with focus on the Bantu-speaking populations denoted by black circles and unbroken lines (Bakola, Dogon, and SE Bantu). Pink represents mtDNA and blue represents NRY F_{ST} values.

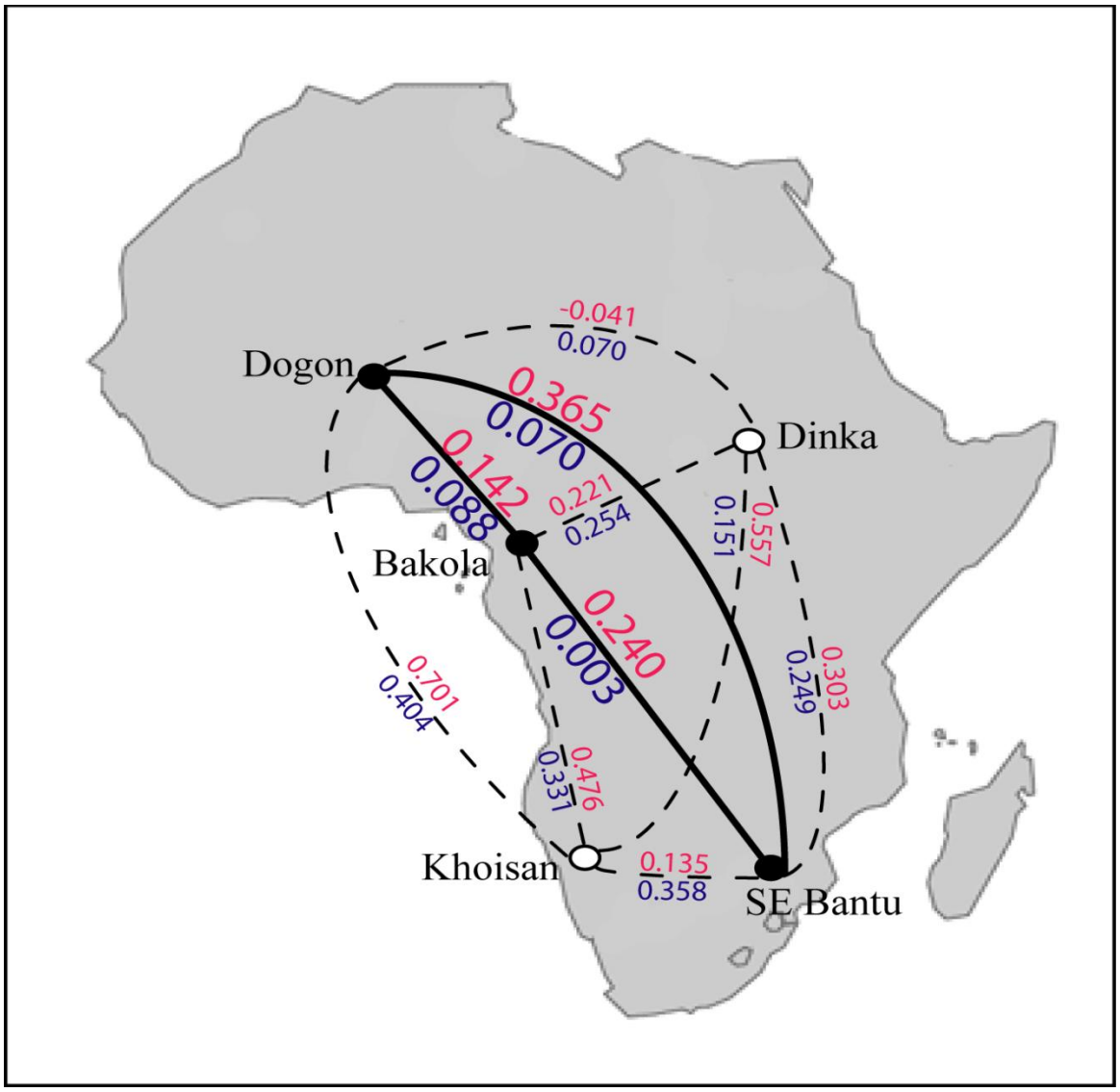
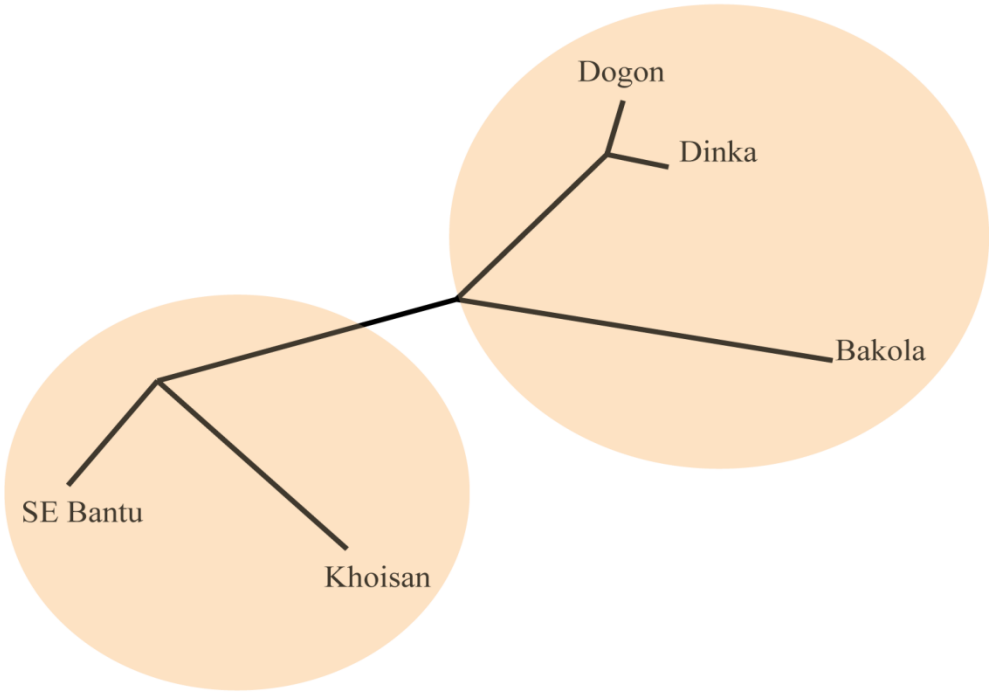


FIGURE 4.4 Un-rooted UPGMA trees for F_{ST} based on **(A)** mtDNA and **(B)** NRY for the five sub-Saharan African populations. The pink circles denote genetic similarity based on geographic proximity **(A)**; while the blue circles denote genetic similarity based on linguistics **(B)**.

A.



B.

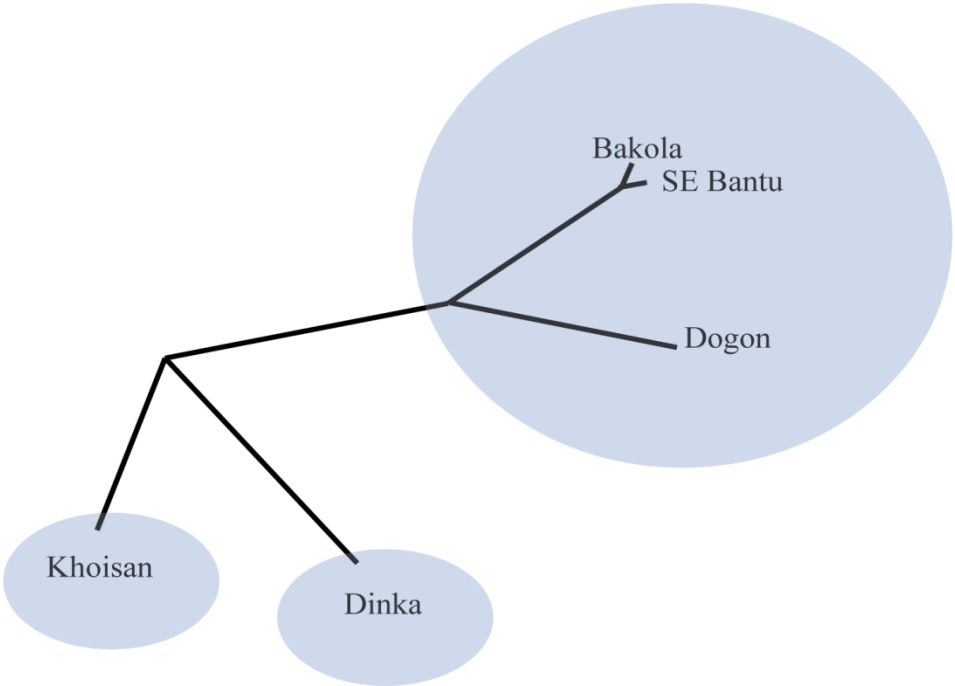


FIGURE 4.5 The estimated route of the Bantu expansions, based on linguistic and genetic data.

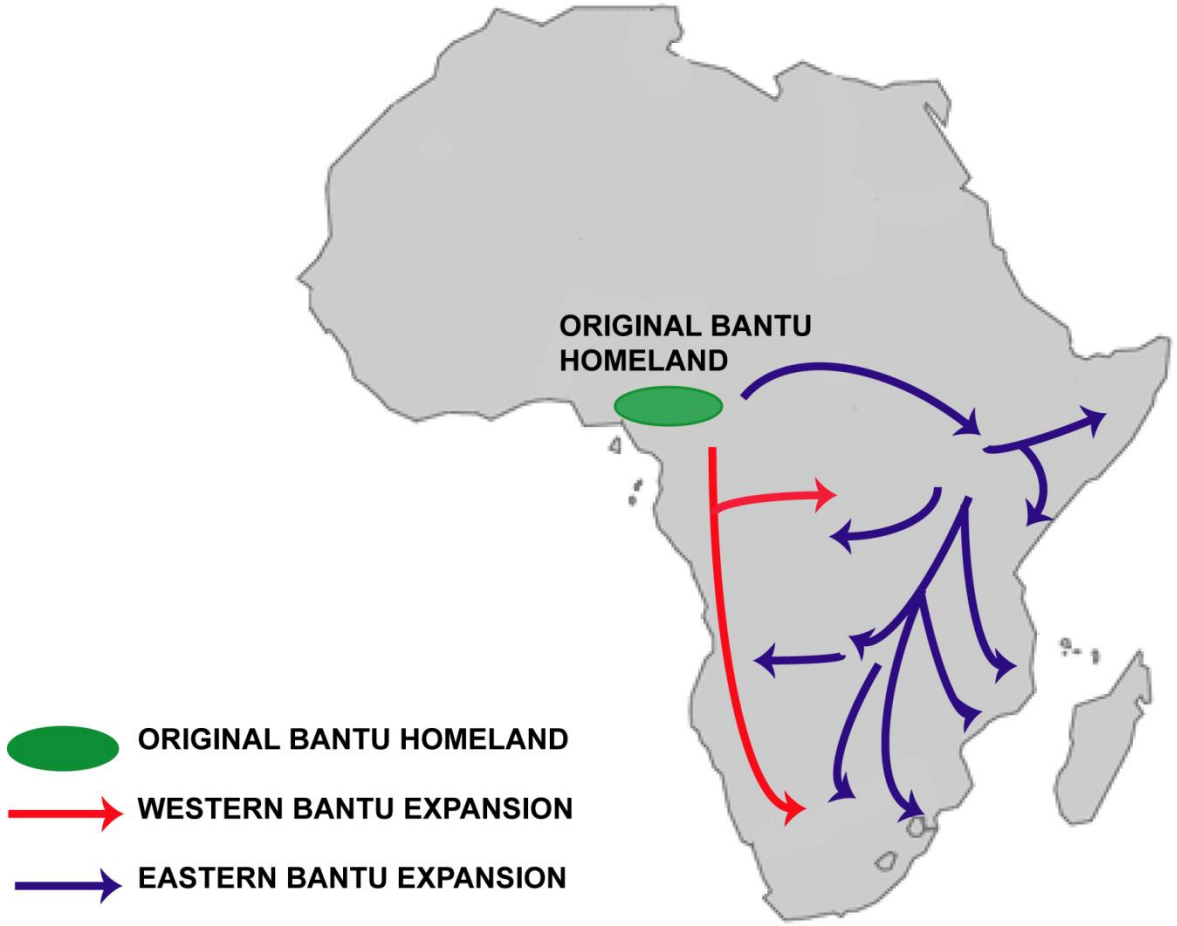
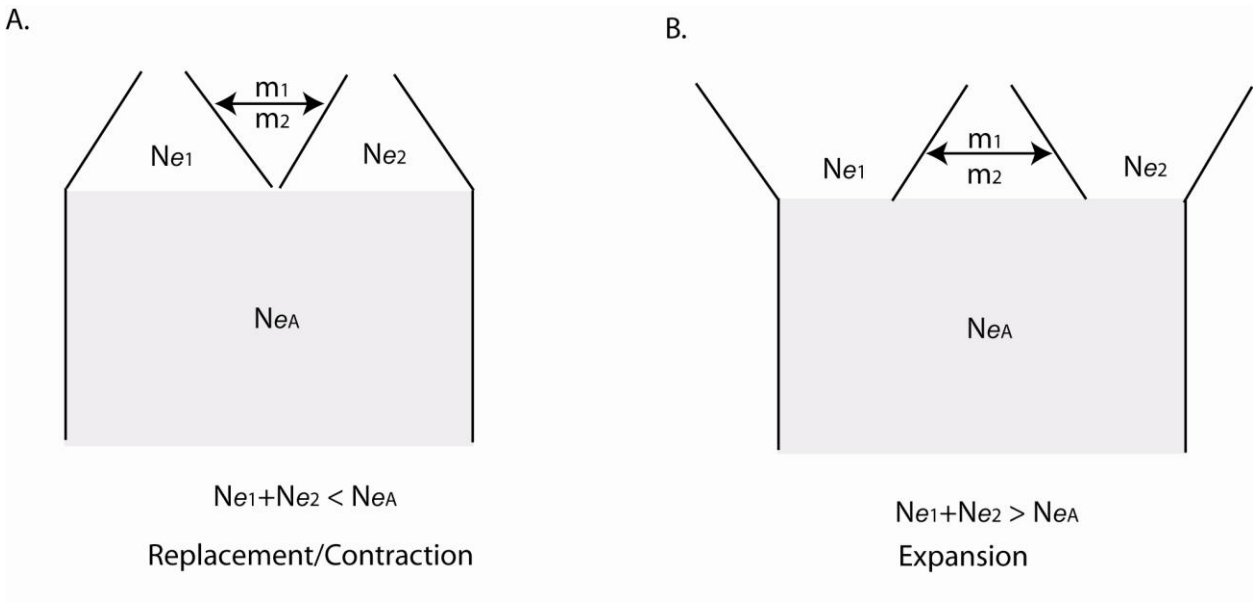


FIGURE 4.6 Models of (A) population contraction with replacement, and (B) population expansion. N_{e1} and N_{e2} are current population size, m_1 and m_2 are migration rates and N_{eA} is the ancestral population size.



SUPPLEMENTARY MATERIAL

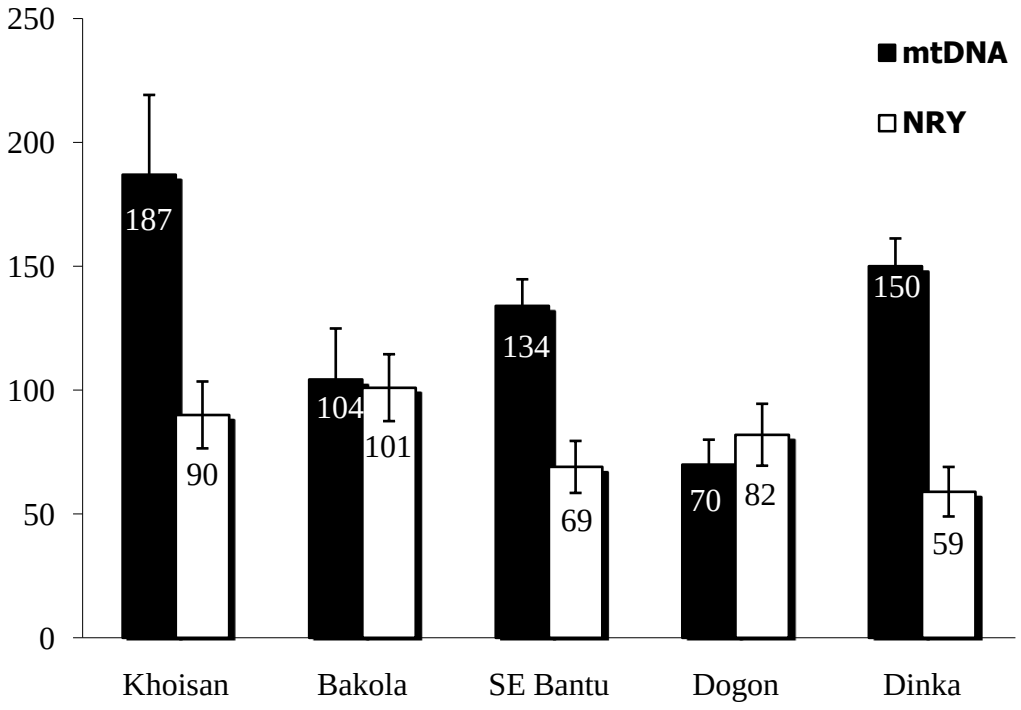
It has been demonstrated that the TMRCA estimates (which are highly dependent on estimates of effective population size) for mtDNA tend to be slightly more than double those of the NRY (Wilder, Mobasher, and Hammer 2004). There is a similar pattern in the TMRCA estimates of three of the five populations studied here (**SUPPLEMENTARY FIGURE 4.1**): the Khoisan (187 *versus* 90 thousand years; KYR), the SE Bantu (134 *versus* 69 KYR) and the Dinka (150 *versus* 59 KYR). Interestingly, the remaining two populations (Bakola and Dogon) have NRY TMRCA estimates that are close to or exceeding that of the mtDNA (104 *versus* 101 and 70 *versus* 82, respectively). These two populations appear to have low mtDNA haplotype diversity, i.e. lower than the mean, which has resulted in low mtDNA TMRCAs for these two populations (0.55 and 0.64 *versus* the mean of 0.76, **TABLE 4.1**). In the case of the Bakola, the NRY haplotype diversity is also highly elevated over those of other populations (0.88 *versus* the mean of 0.78).

MtDNA TMRCAs have been estimated to be approximately twice as old as those from the NRY, though whether these differences are due to differential male and female migration or rather to differences in male and female effective population size has not yet been resolved (Wilder, Mobasher, and Hammer 2004). To further investigate this issue, coalescent simulations were conducted assuming an island model of migration (Hudson, Slatkin, and Maddison 1992) using the mtDNA and NRY migration matrices for the five populations studied here. TMRCAs were calculated from 5,000 simulated datasets, holding Θ constant for the mtDNA and NRY. MtDNA results indicate that the observed TMRCAs can be generated by migration alone for the Khoisan, Bakola and SE Bantu

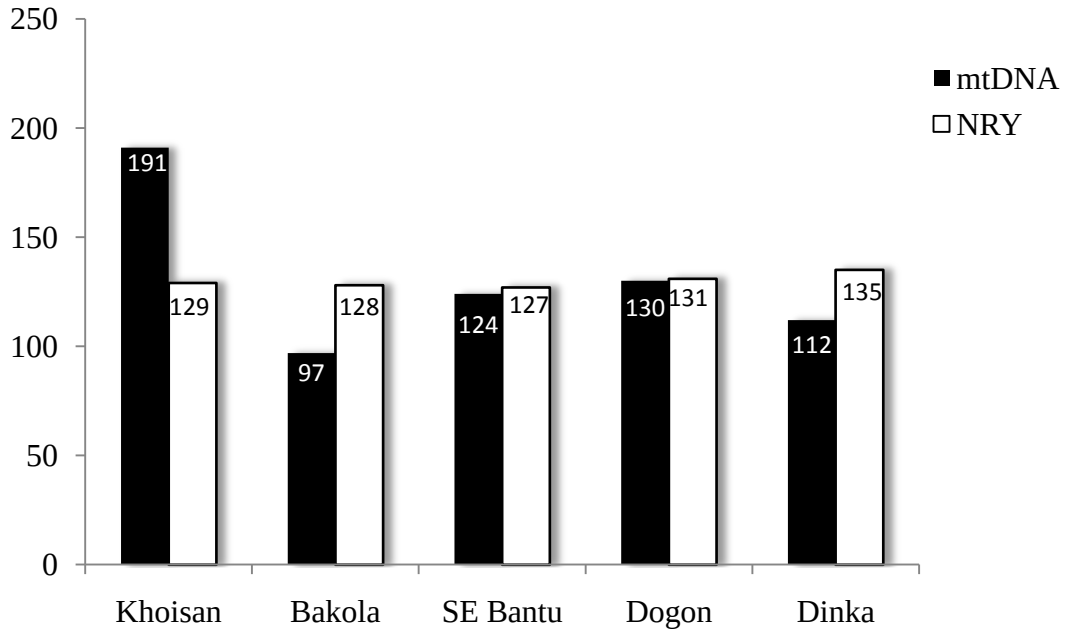
(**SUPPLEMENTARY FIGURE 4.2**). However, in the case of the Dogon, a decrease in the female effective population size must also have occurred in addition to migration for the simulated TMRCA to be comparable to that observed (**SUPPLEMENTARY FIGURE 4.1** and **SUPPLEMENTARY FIGURE 4.2**). In contrast, in the case of the Dinka, an increase in the female effective population size in addition to migration is necessary to obtain the observed TMRCA. The simulations also demonstrate that migration alone can account for the large amount of heterogeneity in TMRCAs based on mtDNA.

The NRY simulations are consistent in showing that migration alone cannot account for the observed TMRCAs (**SUPPLEMENTARY FIGURE 4.2**). It appears that a further reduction in effective population size for males is necessary to match the observed NRY TMRCAs, with a smaller reduction for the food-gatherers than the food producers. The simulated NRY TMRCAs are far more homogenous than those of the mtDNA in these same populations. Notably, the Dinka must have experienced a significant reduction in male effective population size as compared to the other populations. To summarize, the observed differences in mtDNA and NRY TMRCAs from these populations can generally be explained by differential male and female migration in combination with a reduction in male effective population size. The replacement of males coupled with polygyny could account for the increased variance in mtDNA over NRY TMRCAs, since a replacement of hunter-gatherer males during the dispersion event would likely have lead to lower variance in the NRY TMRCA estimates. On the other hand, if females were not as homogenized during the dispersion and marriage between farmer males and hunter-gatherer females occurred, then a higher variance in the mtDNA TMRCAs is expected, as observed here.

SUPPLEMENTARY FIGURE 4.1 Estimated mtDNA and NRY TMRCAs in thousands of years (KYR) for the five sub-Saharan African populations (with 95% confidence intervals) from GENETREE.



SUPPLEMENTARY FIGURE 4.2 MtDNA and NRY TMRCAs in thousands of years (KYR) for the five sub-Saharan African populations simulated from ms using the mtDNA and NRY migration matrices generated from IM.



CHAPTER FIVE: MULTILOCUS ESTIMATES OF ANCIENT POPULATION STRUCTURE AND THE ORIGIN OF MODERN HUMANS

.

ABSTRACT

Current models of anatomically modern human (AMH) origins make different predictions about population structure in humans that can be tested using genetic data. Here I present an assessment of sub-Saharan African population structure to address the questions: (1) How old is population structure in Africa? and (2) What are estimates of current and ancestral effective population sizes? I generated a multilocus dataset from the 172 individuals, and conducted Markov chain Monte Carlo coalescent-based simulations to generate estimates of current and ancestral effective population sizes, migration rates, and divergence times. I demonstrated that population divergence times in sub-Saharan Africa predate the timing of the emergence of modern humans outside of Africa, raising the possibility that modern humans dispersed from a structured African population. Population split times were similar to previous estimates in Africa, ranging from 17-142 thousand years ago (KYR). The Khoisan exhibited the oldest population split times (range, 102-142 KYR) and Niger-Congo speakers the most recent (range, 17-84 KYR). Current and ancestral effective population sizes for all populations were relatively similar, ranging from ~5,000-8,000 individuals. A comparison of estimates of ancestral population size to current effective population sizesdemonstrated that the two hunter-gatherer populations are not consistent in showing evidence of a population contraction, contrary to previous claims, and that nearly all populations appeared to have increased in size from the time of splitting.

INTRODUCTION

Traditionally, scientists have approached questions concerning anatomically modern humans' (AMH) place of origin and subsequent dispersal from paleontological and archaeological directions. Using such data, two extreme models for the origin of anatomically modern *Homo sapiens* were fashioned: the Recent African Origin (RAO) hypothesis and the Multiregional Evolution (ME) hypothesis. These models of AMH origins make different predictions about the history of population structure in humans. The Recent African Origin hypothesis posits that AMH developed in a small, isolated subpopulation that became ecologically successful and expanded and replaced other subpopulations across Africa and Eurasia (Harpending et al. 1993). In fact, early mtDNA data presented the signature of a rapid expansion event marked by short, star-shaped genealogies and led researchers to posit that a replacement event led to very little population structure in modern African populations. This model predicts that extant African populations should not exhibit population structure that predates the origin of AMH because the parent deme was relatively unstructured. In contrast to the Recent African Origin hypothesis, various Multiregional models contend that AMH evolved in the presence of long-standing population structure (and gene flow) without a population replacement event (Weidenreich 1943; Wolpoff et al. 1984). Consequently, one would predict that the signature of ancient population subdivision should still be evident in the gene pool of contemporary African populations. Population structure in Africa should, therefore, be older than the origin of AMH and neutral genealogies should reflect the larger effective population size of the archaic population system (i.e., at those loci that were not associated with selectively advantageous anatomically modern human traits). A

clear picture of ancient population structure is fundamental to our ability to test current models of modern human origins (e.g. the Recent African Origin model *versus* the Multiregional Evolution model).

The use of fossil data to address hypotheses of modern human origins is problematic because it is difficult to detect admixture between ancient populations from fossil remains. Yet the breadth of morphological variation associated with Middle Pleistocene hominin fossils leaves open the possibility that early AMH evolved in a subdivided population. Increasingly, population genetics studies have found evidence of highly divergent human lineages that could be the result of long-term population subdivision in humans (Harris and Hey 1999b; Yu et al. 2001; Reed et al. 2004; Garrigan et al. 2005a; Hayakawa et al. 2006; Plagnol and Wall 2006; Shimada et al. 2007; Cox et al. 2008). For example, based on a small sample of 16 African and 19 non-African individuals, data from the X-linked *PDHA1* locus were the first to present clear evidence for ancient population structure that preceded the emergence of early anatomically modern humans (AMH). This lead researchers to posit that the transition from early AMH occurred in a geographically subdivided ancestral population (Harris and Hey 1999b). Other evidence of ancient structure has come from studies of the ß-globin (Harding et al. 1997), Duffy blood group (Hamblin, Thompson, and Di Rienzo 2002), dystrophin (Labuda, Zietkiewicz, and Yotova 2000), and microcephalin (Evans et al. 2006) loci based on larger sample sizes. Support for deep African population structure was reported based on X-linked non-coding DNA in a small panel of African samples (Garrigan et al. 2005a). Thus, these several studies suggest the presence of ancient population subdivision based on modern Africans, but this has only recently been confirmed by comparisons of data

across multiple, neutrally evolving loci in the same populations (Garrigan et al. 2007). No studies to date have compared patterns of DNA sequence variation directly in modern sub-Saharan Africans to examine subdivision using data from multiple genetic loci.

Such data can be used to address several key questions concerning sub-Saharan Africans, such as: (1)What are the estimates of divergence times among populations, and how do these relate to the timing of the emergence of modern humans? (2) What are the estimates of effective population size in modern and ancestral African populations? and (3) What are the patterns of gene flow among current populations? Generally questions concerning population structure within and among populations have been addressed using F_{ST} (Wright 1951). In addition to traditional analysis, here I employ the Markov chain Monte Carlo (MCMC) method of Hey (2005) which fits the "Isolation with Migration" (IM) model under changing population size to these questions concerning both modern and ancient population structure. The IM model of population structure is one of population splitting in which there is an ancestral population that gives rise to two descendent populations that may be connected by gene flow, and carries with it no prior assumptions about population size or migration. I generated a large dataset of direct re-sequencing data from four loci in five sub-Saharan African populations for use in this analysis. The IM MCMC method was used to synthesize data from multiple loci and yield the posterior probability of each of seven demographic parameters, including current and ancestral effective population sizes, migration rates and divergence times, as well as the proportion of the descendent population founded by the ancestral population (Nielsen and Wakeley 2001).

MATERIALS AND METHODS

Samples and Loci

The loci used in this study are listed in **TABLE 5.1**. The loci include 780 bp of the *cytochrome c oxidase subunit III* (*COIII*) gene of the mtDNA and 13 concatenated *Alu* insertions comprising 6601 bp located on the non-recombining portion of the Y chromosome (NRY) (data presented in Metni Pilkington et al. 2008), 4559 bp of the X-linked pyruvate dehydrogenase alpha 1 gene (*PDHA1*), and 2414 bp of the X-linked ribonucleotide reductase m2 polypeptide pseudogene 4 (*RRM2P4*). I excluded insertion/deletion polymorphisms, and only single nucleotide polymorphisms were included in this study.

I generated re-sequencing data from males representing five sub-Saharan African populations: 25 Khoisan from Namibia (KHO), 25 Bakola from Cameroon (BAK), 42-49 Dogon from Mali (DGN), 23 Dinka from Sudan (DNK), and 50 Southeast Bantu speakers from South Africa (SEB). This resulted in 10 pairwise population comparisons. The use of male samples for this study removed the need to clone individuals in order to infer haplotypes for the X-linked loci. Two of the populations (Khoisan and Bakola) are hunter-gatherers and three of the populations are food-producers (Dinka, Dogon, and SE Bantu). Three of the populations are from the Niger-Congo language family (Bakola, Dogon, and SE Bantu), one is Nilo-Saharan (Dinka) and one is Khoisan. One common chimpanzee (*Pan troglodytes*) was sequenced as an outgroup for each locus. All sampling protocols were approved by the Human Subjects Committee at the University of Arizona and by the institutions of all collaborators who provided DNA samples.

Population Genetic Analyses

Population parameters such as the number of segregating sites, the number of haplotypes, nucleotide diversity (calculated as π and Θ per sequence), Tajima's *D*, Fu and Li's *D*, Fu's *Fs*, and F_{ST} were estimated using the program DnaSP ver. 4.0 (Rozas et al. 2003). Parameters were estimated for the entire population before individuals were excluded for comparisons in the IM analysis. Unweighted pair group method with arithmetic mean (UPGMA) trees of F_{ST} estimates for the four loci were estimated using MEGA version 3.1 (Kumar, Tamura, and Nei 2004)

IM Model and Computations

The Isolation and Migration (IM) model of population structure is one of population splitting in which there is an ancestral population that gives rise to two descendent populations that may be connected by gene flow. The method modified by Hey and Nielsen (2004) that accommodates multiple loci is used here (the method is discussed further in **CHAPTER TWO** and **CHAPTER FOUR** of this work).

The IM analysis requires that the samples used in the study do not show any evidence of recombination by the four-gamete test. When sequences do show evidence of recombination, this can be resolved in three ways: by breaking the sequence up into non-recombining blocks, excluding the site(s) that show evidence of recombination or excluding the individual(s) showing evidence of recombination. I chose to break up the sequence into non-recombining blocks, and to use the largest block (similar to Hey and Nielsen 2004). I did not use more than one block from a given locus as they are not truly independent of each other and should not be treated as separate loci with different

demographic histories. Therefore, I shortened *PDHA1* to 1623 bp and *RRM2P4* to 1643 bp of non-recombining sequence data. To adjust for differences in effective population size of the haploid and haplo-diploid loci, each locus is modified by an inheritance scalar relative to 1.0 for autosomal loci, such that the haploid loci (mtDNA *COIII* and the NRY *Alu*s) have a scalar of 0.25, while the X chromosomal loci (*RRM2P4* and *PDHA1*) have a scalar of 0.75. For the purposes of this analysis, I have arbitrarily assigned population 1 and population 2, as listed in **TABLE 5.2**. I used previously published mutation rates to calculate a geometric mean of the mutation rate per year per gene (Harris and Hey 1999b; Wilder, Mobasher, and Hammer 2004; Garrigan et al. 2005b; Wilder and Hammer 2007a).

RESULTS

A total of ~10.7 kb of re-sequencing data from two haploid and two haplo-diploid loci were obtained from 172 individuals sampled from five sub-Saharan African populations. Summary statistics describing the overall nucleotide diversity and the polymorphism frequency spectra of the loci examined are located in **TABLE 5.1.** The sequence divergence for all pooled populations is highest for *COIII* (0.289 %) and lowest for the NRY (0.054 %), with the two X-linked loci falling in between (0.073 % for *RRM2P4* and 0.168 % for *PDHA1*). The total number of segregating sites is greatest for *PDHA1* (22-32) and smaller for the other three loci, ranging from 5-16 for *COIII*, 8-14 for the NRY, and 4-9 for *RRM2P4*. *PDHA1* has the highest pooled haplotype diversity (0.93), *COIII* and the NRY have very similar pooled haplotype diversities (0.87 and 0.88, respectively), and *RRM2P4* has the lowest pooled haplotype diversity (0.66). The

estimates of π per sequence for *COIII*, the NRY and *RRM2P4* range from 0.827 to 3.721 and are much higher for *PDHA1*, ranging from 5.562 to 8.936. The same pattern is observed for Θ values as well. All loci, except for *COIII*, appear to be in mutation-drift equilibrium based on the frequency spectra presented here. In the case of *COIII*, significantly negative Tajima's D, Fu and Li's D*, and Fu's *Fs* values indicate an excess of rare alleles over that expected under equilibrium for the three food-producing populations (the Dinka, Dogon, and SE Bantu).

Pooled F_{ST} estimates indicate that population structure is far greater for the haploid loci (*COIII* F_{ST} = 0.264 and the NRY F_{ST} = 0.211) than it is for the haplo-diploid loci (*RRM2P4* F_{ST} = 0.016 and *PDHA1* F_{ST} = 0.049, **FIGURE 5.1**). The topologies of the UPGMA trees differ for each locus, but there are some overarching patterns. There is a strong association between the Khoisan and SE Bantu populations, based on *COIII* and *PDHA1*, and they exhibit many shared haplotypes to the exclusion of the other populations (**FIGURE 5.1** and **FIGURE 5.2,** and **TABLE 2.1** of the **MATERIALS AND METHODS**). Likewise, the Khoisan and Dinka are closely associated based on the NRY and *RRM2P4* (**FIGURE 5.1** and **FIGURE 5.2**). Finally, it is interesting to note that the SE Bantu sample captures all of the observed shared haplotypes, except for haplotypes shared exclusively by the Dinka and Dogon for *COIII*, *RRM2P4*, and *PDHA1*.

Population Divergence

For all populations the marginal posterior probability distributions of the population divergence parameter, *t*, had fairly sharp peaks, albeit with relatively wide ranges (**TABLE 5.2**, the 90% HDP intervals are included in **SUPPLEMENTARY TABLE 5.1,**

and **SUPPLEMENTARY FIGURES 5.1-5.10**). I estimated the minimum sub-Saharan African divergence time to be 17 KYR (Dinka-SE Bantu), while the maximum sub-Saharan African divergence time was estimated to be 142 KYR (Khoisan- SE Bantu). Three striking results emerge from the population comparisons of divergence times (**FIGURE 5.3** and **FIGURE 5.4**). First, the oldest divergence time of 142 KYR, between the Khoisan and the SE Bantu, predates the emergence of AMH outside of Africa based on current fossil evidence. Second, all estimates of the population divergence times that include the Khoisan are the oldest observed in this analysis (142, 123, 105, and 102 KYR). Third, the most recent divergence dates are from comparisons among the Niger-Congo speakers (specifically with the Dogon and other Niger-Congo speakers, the SE Bantu and Bakola). It should be noted that the Dinka-SE Bantu comparison resulted in the most recent divergence date (and highest migration rate), even though the Dinka are a Nilo-Saharan-speaking population, not Niger-Congo speakers.

Current and Ancestral Effective Population Size

Summaries of the parameter estimates generated from the IM analyses for each population are presented in **TABLE 5.2** (the 90% HDP intervals are included in **SUPPLEMENTARY TABLE 5.1**, **SUPPLEMENTARY TABLE 5.2,** and **SUPPLEMENTARY FIGURES 5.1-5.10**). The posterior point estimates of the descendent population sizes N_{e1} and N_{e2} range from 2,300 for the Bakola to 9,200 for the Dinka. These estimates are slightly lower than most estimates of effective population sizes of ~10,000 individuals based on autosomal loci. The range of ancestral effective population sizes is similar, from 1,600 for the Khoisan-SE Bantu to 7,800 for the Bakola-Dinka. Most population

comparisons reveal that the estimates of ancestral effective population size (N_{eA}) do not differ greatly from those of current effective population sizes (N_{e1} and N_{e2}). In the current analysis, the modern and ancestral effective sizes tend to range between 4,400 and 8,000 individuals. When I take into account the 90% HDP confidence intervals, the current and ancestral population sizes for all populations overlap, and are unable to reject a model of constant population size in these sub-Saharan African populations. However, when the size of the populations just after the time of splitting (sN_{eA} for population 1 and (1-s) N_{eA} for population 2) is compared to the current effective population sizes (N_{e1} and N_{e2}), all comparisons reveal that populations have increased in size (except Dinka-SE Bantu**; TABLE 5.2**). In fact, all comparisons with the SE Bantu demonstrate that nearly all of the ancestral population went on to found the SE Bantu and that only a small proportion of the ancestral population went on to found the Bakola, Dogon, and Dinka (the analysis yielded poor estimates for the SE Bantu-Khoisan comparison). Estimates suggest that since *t* (splitting time with the SE Bantu), the Dogon and Dinka have grown by a factor of about 10 (from 444 to 3,700 and 563 to 4,600, respectively) and the Bakola have grown by a factor of about 40 (from 56 to 2,300), despite the apparent similarity between current and ancestral effective population sizes (N_{e1}, N_{e2}, and N_{eA}). For the SE Bantu, the similarity between their estimated population size just after splitting from the ancestral population and current effective population sizes is consistent with constant population size. It is also remarkable that the two hunter-gatherer groups (Khoisan and the Bakola) both appear to have grown in size from the time of splitting, contrary to previous work that has demonstrated that most hunter-gatherers do not show the signs of a Pleistocene population expansion (Excoffier and Schneider 1999).

Migration

TABLE 5.2 and **FIGURE 5.3** present the patterns of migration among the five populations as estimated from multiple loci (the 90% HDP intervals are included in the **SUPPLEMENTARY TABLE 5.2** and **SUPPLEMENTARY FIGURES 5.1-5.10**). There are no population comparisons that reveal migration at the lower limit of the resolution of the IM analysis; therefore, data from all populations demonstrate that migration is reciprocal (and is never zero). Estimates of migration rate parameters range from 0.120 (Bakola-Dogon) to 7.230 (Dinka-SE Bantu). Levels of migration were divided into three intervals: low ($m \leq 1.0$), moderate ($1.0 < m < 5.0$), and high ($m \geq 5.0$). There is strong concordance between the results presented here using multiple loci, and the results obtained from the examination of mtDNA exclusively (**CHAPTER FOUR**). In particular, there are very low rates of emigration from the Khoisan to all other populations and moderate to high levels of migration into the Khoisan from all other populations. There is also strong support for low migration into the Bakola and much higher emigration from the Bakola into other populations. Finally, I observe very high migration rates into the SE Bantu from all other populations (with the exception of the Khoisan). I discuss the implications of these results in detail below.

DISCUSSION AND CONCLUSIONS

In recent years, the timing of human dispersals within Africa and beyond has been the source of considerable debate, particularly with regard to whether modern humans evolved from a single, panmictic population or many subdivided populations

(Harpending et al. 1993; Chen et al. 1995; Harpending et al. 1998; Zietkiewicz et al. 1998; Harris and Hey 1999b; Wakeley 1999; Labuda, Zietkiewicz, and Yotova 2000; Wall 2000; Relethford 2001b; Goldstein and Chikhi 2002; Garrigan et al. 2005a; Weaver and Roseman 2005; Garrigan et al. 2007). I conducted a detailed study of sub-Saharan Africans using both traditional population genetic analyses and the Isolation with Migration model, which has the advantage of estimating non-equilibrium demography and gene flow simultaneously. I used a multilocus sub-Saharan African dataset to estimate current and ancestral population sizes, migration rates, and population divergence times.

The haploid loci clearly reveal higher levels of population structure than the X-linked loci presented here. Patterns of gene flow estimated from the four loci are remarkably similar to those estimated from a single mtDNA locus (see **CHAPTER FOUR**), which might be expected considering that two of the four loci are X-linked and therefore represent a greater proportion of female population history. I increased the existing sample size of *PDHA1* in the literature from 33 to 151 (see **SUPPLEMENTARY FIGURE 5.11**), and refined the topology of the gene tree in sub-Saharan Africa. Based on the data from the four loci, I find strong support for low migration into the Bakola and much higher emigration from the Bakola into other populations. This proposed pattern of migration is compatible with data suggesting that the Bantu Expansions originated in the region where the Bakola reside presently. The *s* parameter estimates, which represent the proportion of the ancestral population that founded the descendent populations, reflect the recent divergence times of the Niger-Congo speakers and the Nilo-Saharan speakers. In comparisons of the SE Bantu with the Dinka, Bakola, and Dogon, the majority of the

ancestral population went into founding the SE Bantu while only a small fraction of the ancestral population went into founding the Dinka, Bakola, and Dogon. This is likely a reflection of the effects of the Bantu Expansions which are hypothesized to have swept over Africa approximately 3-5 KYR (Posnansky 1968; Greenberg 1972; Phillipson 1993; Cavalli-Sforza, Menozzi, and Piazza 1994; Poloni et al. 1997; Beleza et al. 2005; Wood et al. 2005; Rexova, Bastin, and Frynta 2006). Future work including more diverse sub-Saharan African populations could shed more light on the mode and tempo of movement during such major recent demographic events such as the Bantu Expansions as well as the proportion of migrants that populations exchanged *versus* those that were replaced entirely.

Similar to previous work using different methods, I estimated modern-day sub-Saharan African effective population sizes to be ~5,000-8,000 (Tenesa et al. 2007). These estimates are slightly lower than many estimates of effective population sizes of ~10,000 individuals based on autosomal loci (Harpending et al. 1998; Wilder, Mobasher, and Hammer 2004; Zhao et al. 2006), but similar to the estimates of Garrigan et al. (2007). Most population comparisons reveal that the estimates of ancestral effective population size (N_{eA}) do not differ greatly from those of current effective population sizes (N_{e1} and N_{e2}), supporting previous work showing that most African populations best fit a model of constant population size (Marth et al. 2003; Voight et al. 2005). However, when the size of the population just after the time of splitting is taken into account, nearly all populations show an increase in population size, even the hunter-gatherers. It has been proposed that the ancestral effective population size of hunter-gatherers was larger than their present day effective populations sizes, and that contractions in their

geographic ranges and population sizes were sustained due to interactions with neighboring food-producing groups (Excoffier and Schneider 1999). Such a population size contraction could possibly account for the fact that most hunter-gatherer populations do not show the signs of a Pleistocene population expansion that most food-producing populations exhibit based on mtDNA data. However, the hunter-gatherers presented here do not show evidence of a population size reduction from their ancestral population size. In fact, both the Bakola and the Khoisan show an *increase* in effective population size over that of the ancestral population in many of the comparisons. Thus, these patterns of increasing effective population size differ from those previously reported for sub-Saharan African populations.

The observed divergence times between sub-Saharan Africans (17-142 KYR) fall well within the range of expected African divergence times based on previous studies of protein polymorphism, mtDNA, and the Y chromosome (Cavalli-Sforza, Menozzi, and Piazza 1994; Knight et al. 2003; Gonder et al. 2007). Many of the Niger-Congo speakers' divergence times are quite recent (~20 KYR) while those that include the Khoisan are relatively ancient (102-142 KYR). Three fundamentally different questions rely on population divergence times: (1) Is there population structure in the source population for African AMH, (2) Is there population structure in early African AMH, and (3) Were non-African AMH derived from a structured African AMH population? Certainly, the breadth morphological variation associated with Middle Pleistocene hominin fossils leaves open the prospect that early AMH evolved in a subdivided African population. If the divergence dates within Africa had been much older, then there would have been good support for a subdivided African source population for AMH; however,

my results do not allow me to address this question fully. Therefore, I move to questions two. Again, one can not definitively say whether AMH evolved in an Africa that was genetically subdivided, since there is overlap between the fossil dates for the emergence of AMH (~200 KYR,McDougall, Brown, and Fleagle 2005) and the 90% confidence intervals for the earliest population split times. Other studies have suggested ancient African structure, with split times ranging from 145-214 KYR (Garrigan et al., 2007), though these must be examined with caution, as a more recent study using several X-linked and autosomal loci has estimated African split times to be less than 80 KYR (Murray Cox, personal communication).

Finally, there is good reason to believe that population structure existed in Africa before modern humans migrated from the continent, since the earliest within-Africa divergence time (~142 KYR) predates the earliest evidence of modern humans outside of Africa based on Levantine fossils (100-135 KYR, Grun et al. 2005). Nevertheless, the question of whether all non-African populations are derived from a structured African AMH population can be addressed further by studies of multilocus datasets from multiple non-African populations. If African-non-African divergence times do not overlap with divergence times found within Africa but instead are much younger, then one might be able to say with some certainty that all non-African populations derive from a structured African population. In fact, this is exactly what Garrigan et al. (2007) demonstrated in a recent paper. The authors proposed that African populations were relatively structured prior to the founding of the non-African populations studied, based on the within-Africa and African-non-African divergence times. Population divergence times were greater than 140 KYR within Africa (with very large ranges), while the oldest African-non-

African divergence time was reported to be ~40 KYR (for the SE Bantu-Mongolian comparison).

In sum, my results are consistent with both the Recent African Origin model and the Multiregional Evolution model. Although population split times for ancestral African structure do not convincingly predate the emergence of AMH, there is strong evidence of long-standing population structure in Africa, and population structure that predates the emergence of AMH from Africa. Future analyses incorporating a greater number of neutrally evolving X-linked and autosomal loci should be able to estimate more accurately ancient population divergence times. However, there is a *caveat* to using multiple loci to characterize ancient population structure: the signal of ancient population structure may be lost in an analysis that pools together many loci and averages population split times because the signal of ancient population structure may only manifest itself in a single locus. Also, the haploid loci presented here have reduced power to test hypotheses of ancient structure since it is known *a priori* that their TMRCAs are never older than 200 KYR (Wilder, Mobasher, and Hammer 2004, and see **CHAPTER FOUR** of this work), and, consequently, their split times will be even more recent. Therefore, the haploid loci will never reveal any lineages that predate the origin of AMH, and will bias downward any multilocus analyses that attempt to characterize the origin of population structure. It is unclear at this time how many loci are necessary to accurately estimate population split times in Africa, but early work put this number at between 50-100 loci (Wall 2000).

TABLE 5.1 Summary of descriptive statistics describing nucleotide polymorphism for the mtDNA *COIII*, NRY-*Alu*, and X-linked *RRM2P4*, and *PDHA1* loci.

Locus	Scalar	Population	n[a]	S[b]	Hp[c]	H[d]	π[e] (%)	Θ_W[e] (%)	TD[f]	FLD*[g]	*Fs*[h]	F_{ST}
mtDNA	0.25	KHO	25	10	8	0.83	0.257	0.340	-0.800	0.387	-1.651	
(780 bp)		SEB	50	16	18	0.90	0.331	0.458	-0.857	-1.033	-8.768*	
		DNK	23	15	12	0.89	0.002	0.005	-1.882*	-2.784*	-7.135*	
		DGN	49	8	8	0.64	0.106	0.002	-1.472	-2.961*	-3.842*	
		BAK	24	6	6	0.55	0.129	0.206	-1.135	-0.238	-1.956	
		Pooled	**171**	**36**	**41**	**0.87**	**0.289**	**0.807**	**-2.529***	**-8.022***	**-37.424***	**0.264**
NRY	0.25	KHO	25	10	7	0.86	0.041	0.040	0.089	0.090	0.183	
(6601 bp)		SEB	46	10	7	0.78	0.044	0.034	0.762	-0.406	1.525	
		DNK	23	8	5	0.70	0.040	0.033	0.696	0.157	1.771	
		DGN	40	12	8	0.69	0.037	0.043	-0.402	-0.564	-0.071	
		BAK	25	14	9	0.88	0.061	0.056	0.286	0.337	-0.127	
		Pooled	**159**	**21**	**19**	**0.88**	**0.054**	**0.056**	**-0.136**	**-0.121**	**-2.386**	**0.211**
RRM2P4	0.75	KHO	25	8	8	0.80	0.083	0.088	-0.188	-1.082	-1.674	
(2414 bp)		SEB	50	5	6	0.56	0.060	0.046	0.685	0.134	0.048	
		DNK	21	4	6	0.67	0.070	0.048	1.300	0.158	-0.077	
		DGN	32	9	6	0.51	0.070	0.096	-0.811	-1.002	-1.000	
		BAK	23	6	6	0.81	0.081	0.067	0.630	0.509	-0.086	
		Pooled	**151**	**16**	**17**	**0.66**	**0.073**	**0.123**	**-1.081**	**-3.433***	**-6.789***	**0.016**
PDHA1	0.75	KHO	25	22	9	0.81	0.140	0.133	0.184	0.824	1.378	
(4559 bp)		SEB	50	32	23	0.96	0.122	0.164	-0.850	0.582	-7.522	
		DNK	22	27	13	0.93	0.176	0.172	0.091	0.801	-1.388	
		DGN	31	29	16	0.85	0.196	0.166	0.654	-0.115	-1.302	
		BAK	23	22	8	0.85	0.183	0.137	1.241	1.099	3.098	
		Pooled	**151**	**41**	**38**	**0.93**	**0.168**	**0.171**	**-0.046**	**-0.768**	**-8.403***	**0.049**

[a] Number of individuals.

[b] Number of segregating sites.

[c] Number of haplotypes.

[d] Haplotype diversity.

[e] Calculated per base pair.

[f] Tajima's D

[g] Fu and Li's D*

[h] Fu's *Fs*

*$p < 0.05$ of being observed under the standard neutral model.

TABLE 5.2 Multilocus estimates of scaled effective population sizes, migration rates, and divergence times using the IM model*. Populations were arbitrarily assigned to be population 1 (listed first) and population 2 (listed second). Bold indicates cases where 90% HDP confidence intervals for migration rates do not include zero.

Comparison	*t* (KYR)	N_{e1} [a]	sN_{eA} [b]	N_{e2} [c]	(1-s) N_{eA} [d]	N_{eA} [e]	m_1 [f]	m_2 [g]
BAK x DGN	20	6,300	3,500	5,500	2,900	6,500	0.120	**4.214**
BAK x KHO	105	5,800	--	6,400	--	5,200	0.403	**2.348**
BAK x DNK	36	5,700	3,000	5,700	4,700	7,800	**1.233**	**2.519**
BAK x SEB	84	2,300	56	6,000	5,600	6,000	1.050	**4.105**
KHO x DGN	123	7,000	2,500	6,300	1,800	4,400	**1.510**	**0.669**
KHO x DNK	102	7,500	5,900	7,700	513	6,400	**2.655**	0.815
KHO x SEB	142	3,800	--	6,800	--	1,600	**5.135**	**0.533**
DNK x DGN	71	9,200	5,700	3,600	971	5,000	**4.860**	**2.895**
DNK x SEB	17	4,600	563	4,500	5,200	5,300	**2.343**	**7.230**
DGN x SEB	18	3,700	444	7,000	6,500	7,000	2.970	**5.286**

* "--" indicates poor estimate of this parameter for this population comparison.

[a] Estimate of current effective population size of population 1.

[b] Estimate of the number of individuals from the ancestral population that founded population 1.

[c] Estimate of current effective population size of population 2.

[d] Estimate of the number of individuals from the ancestral population that founded population 2.

[e] Estimate of effective population size of the population ancestral to population 1 and population 2.

[f] Estimate of migration rate into population 1.

[g] Estimate of migration rate into population 2.

FIGURE 5.1 UPGMA trees based on F_{ST} for (**A**) mtDNA, (**B**) NRY, (**C**) *RRM2P4*, and (**D**) *PDHA1*.

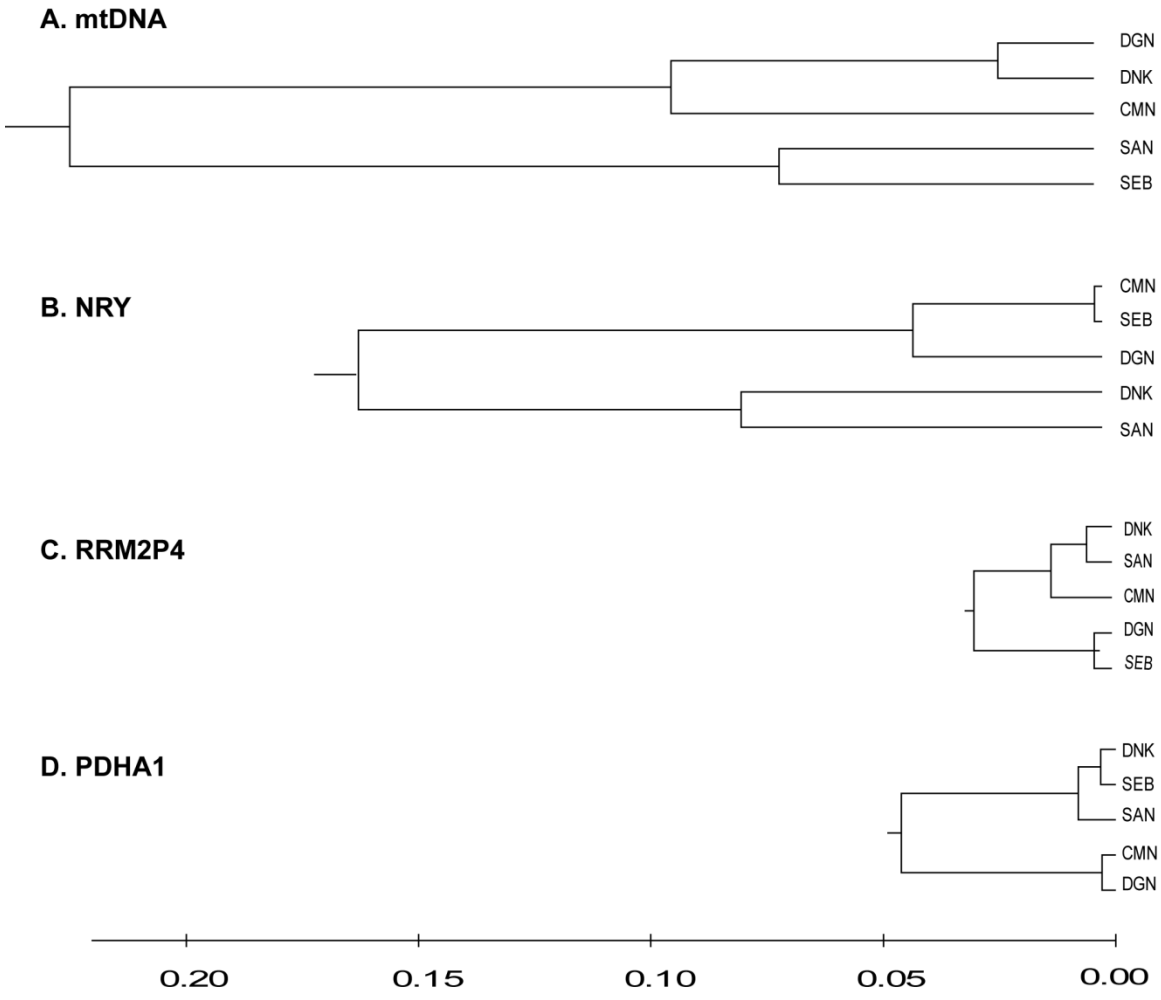


FIGURE 5.2 Shared haplotypes for (**A**) mtDNA, (**B**) the NRY, (**C**) *RRM2P4*, and (**D**) *PDHA1*. Grey indicates unique haplotypes.

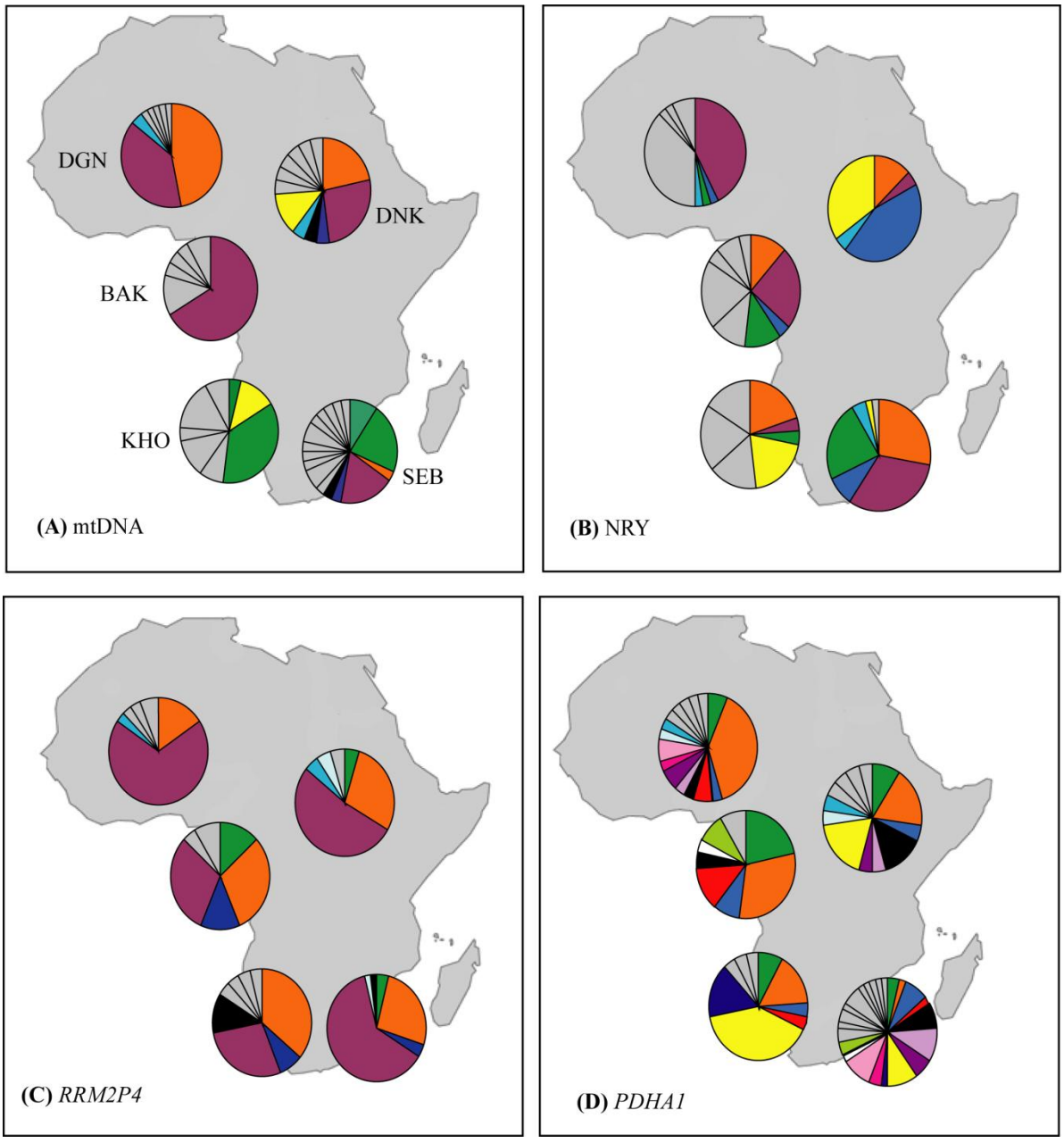


FIGURE 5.3 Estimates of migration rates and divergence times (KYR) for five sub-Saharan African populations (migration rates with 90% HDP intervals not including zero shown).

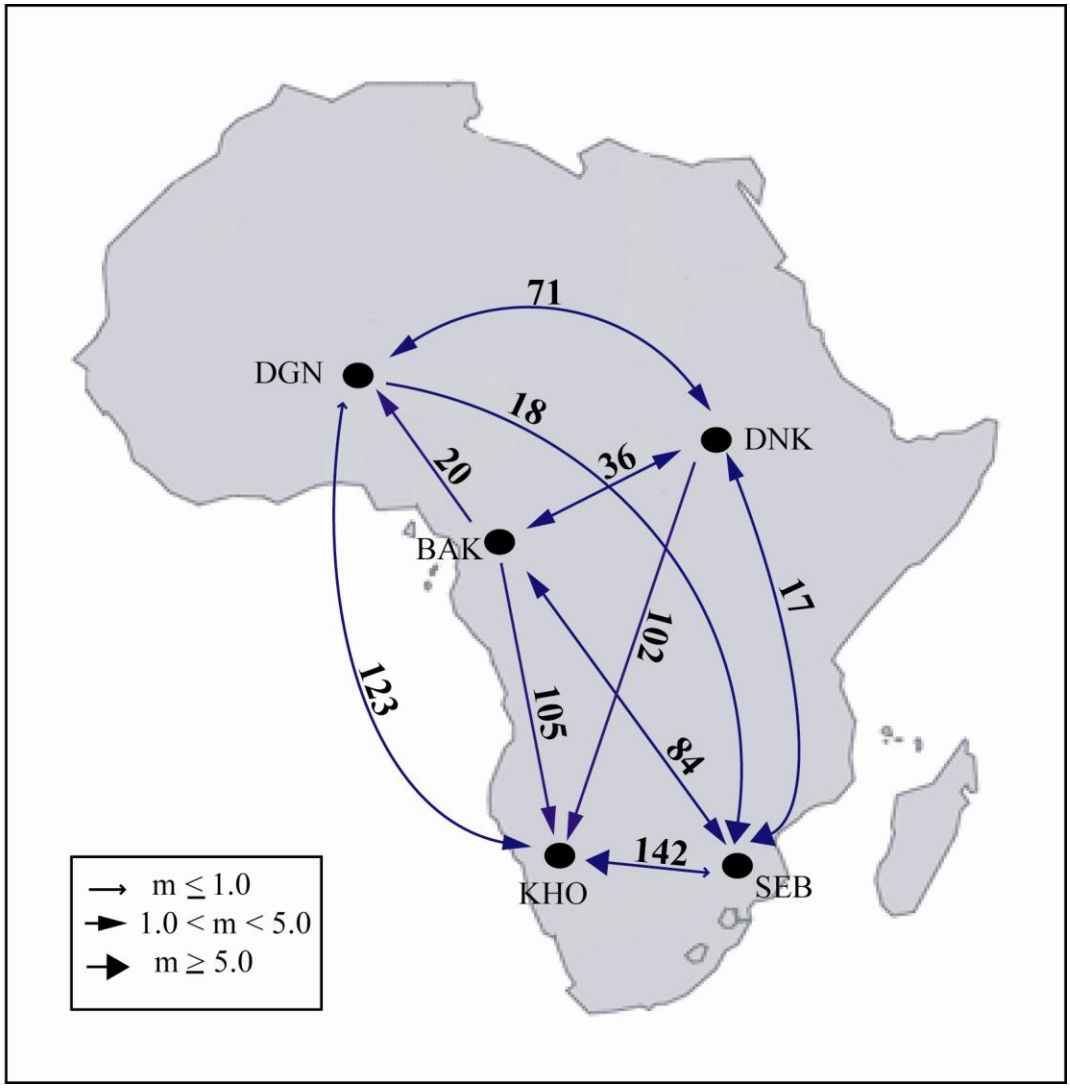
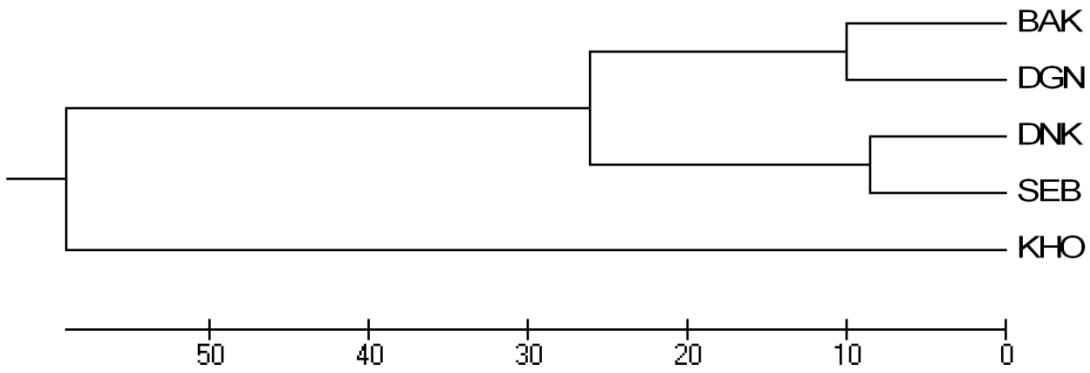


FIGURE 5.4 UPGMA tree of the split times (in thousands of years) for the five sub-Saharan populations.



SUPPLEMENTARY TABLE 5.1 Multilocus estimates of scaled effective population sizes and divergence times using the IM model (90% HDP confidence intervals in parentheses). Populations were arbitrarily assigned to be population 1 (listed first) and population 2 (listed second).

Comparison	N_{e1} [a]	sN_{eA} [b]	N_{e2} [c]	(1-s) N_{eA} [d]	N_{eA} [e]	*t* (KYR)
BAK x DGN	6,300 (3,600-11,100)	3,500	5,500 (3,200-9,700)	2,900	6,500 (300-13,400)	20 (9-897)
BAK x KHO	5,800 (3,100-11,300)	--	6,400 (3,500-12,400)	--	5,200 (120-15,400)	105 (34-1,817)
BAK x DNK	5,700 (2,900-10,900)	3,000	5,700 (2,600-14,600)	4,700	7,800 (2,800-14,900)	36 (15-341)
BAK x SEB	2,300 (1,500-5,500)	56	6,000 (3,500-10,900)	5,600	6,000 (500-16,500)	84 (23-2,665)
KHO x DGN	7,000 (4,000-14,300)	2,500	6,300 (4,000-10,500)	1,800	4,400 (140-15,900)	123 (56-908)
KHO x DNK	7,500 (4,300-13,800)	5,900	7,700 (4,600-12,100)	513	6,400 (1,800-11,700)	102 (34-352)
KHO x SEB	3,800 (2,000-7,700)	--	6,800 (4,000-11,800)	--	1,600 (21-31,200)	142 (85-1,817)
DNK x DGN	9,200 (4,300-19,400)	5,700	3,600 (1,900-6,100)	971	5,000 (600-13,200)	71 (38-1,470)
DNK x SEB	4,600 (2,600-9,200)	563	4,500 (2,200-9,400)	5,200	5,300 (800-10,300)	17 (8-909)
DGN x SEB	3,700 (1,900-7,600)	444	7,000 (4,500-9,900)	6,500	7,000 (1,000-14,400)	18 (13-363)

SUPPLEMENTARY TABLE 5.2 Multilocus estimates of effective population sizes and migration rates using the IM model (90% HDP confidence intervals in parentheses). Bold indicates cases where 90% HDP confidence intervals do not include zero. Populations were arbitrarily assigned to be population 1 (listed first) and population 2 (listed second).

Comparison	θ_1 [a]	θ_2 [b]	θ_A [c]	m_1 [d]	m_2 [e]
BAK x DGN	3.519 (2.006-6.217)	3.089 (1.768-5.434)	3.624 (0.148-7.502)	0.120 (0.003-4.563)	**4.214** (1.314-7.216)
BAK x KHO	3.262 (1.729-6.335)	3.575 (1.942-6.943)	2.928 (0.067-8.585)	0.403 (0.003-2.453)	**2.348** (0.828-4.178)
BAK x DNK	3.206 (1.636-6.087)	3.207 (1.453-8.163)	4.343 (1.552-8.321)	**1.233** (0.253-4.483)	**2.519** (0.033-8.129)
BAK x SEB	1.668 (0.825-3.085)	3.342 (1.958-6.065)	3.381 (0.283-9.187)	1.050 (0.001-11.890)	**4.105** (0.010-9.115)
KHO x DGN	3.908 (2.242-7.984)	3.523 (2.243-5.874)	2.440 (0.075-8.848)	**1.510** (0.374-3.118)	**0.669** (0.241-1.995)
KHO x DNK	4.177 (2.405-7.684)	4.315 (2.592-6.776)	3.590 (0.979-6.510)	**2.655** (0.525-5.259)	0.815 (0.005-4.725)
KHO x SEB	2.144 (1.118-4.320)	3.812 (2.236-6.611)	0.893 (0.012-17.443)	**5.135** (2.735-9.945)	**0.533** (0.008-3.493)
DNK x DGN	5.148 (2.383-10.815)	1.998 (1.072-3.406)	3.040 (0.340-7.378)	**4.860** (0.220-13.140)	**2.895** (1.565-9.995)
DNK x SEB	2.561 (1.432-5.137)	2.495 (1.242-5.232)	2.960 (0.474-5.767)	**2.343** (1.095-5.991)	**7.230** (2.790-16.690)
DGN x SEB	2.061 (1.037-4.240)	3.872 (2.522-5.506)	3.880 (0.595-8.063)	2.970 (0.005-9.186)	**5.286** (0.270-9.714)

[a] Estimate of θ per sequence for population 1.

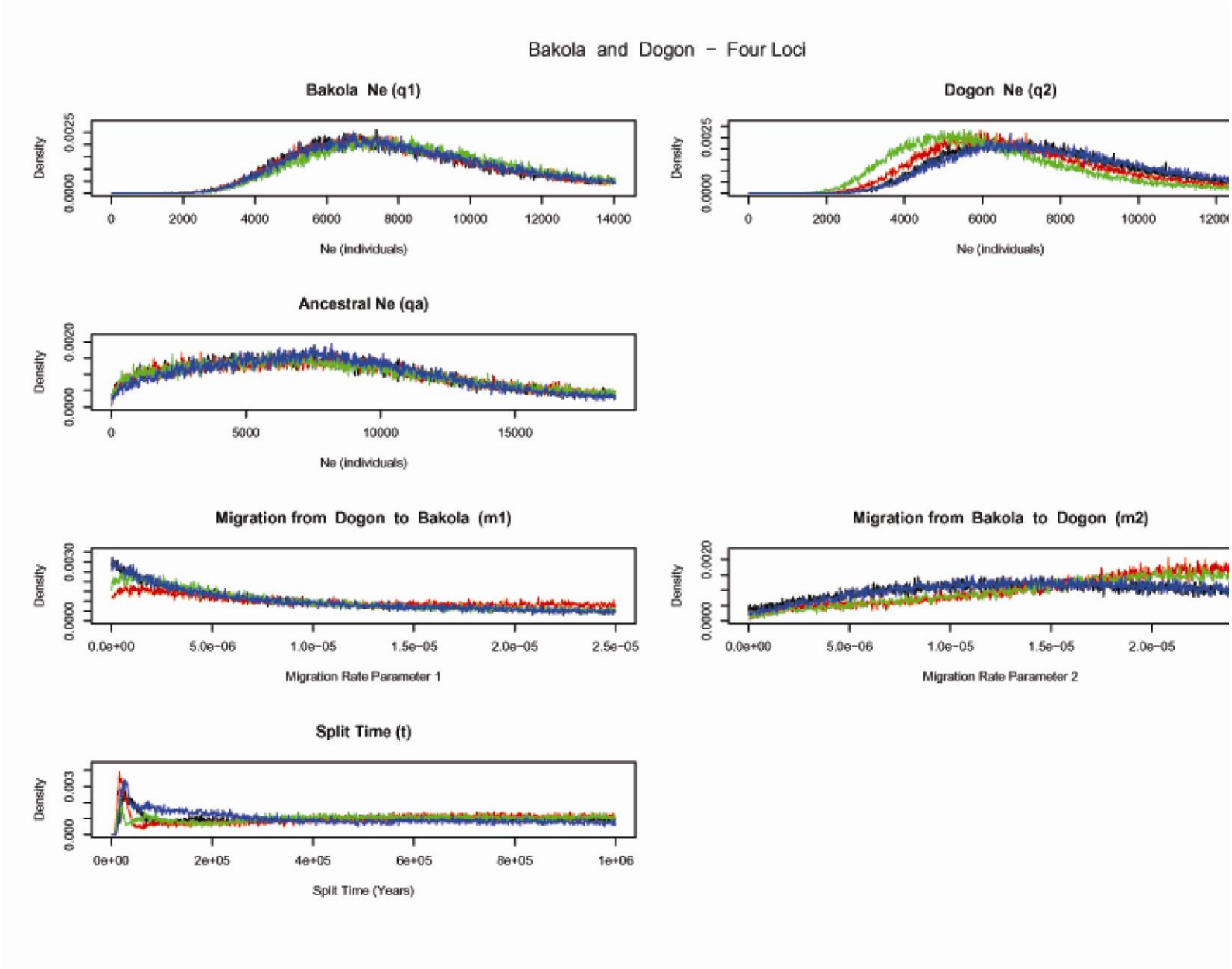
[b] Estimate of θ per sequence for population 2.

[c] Estimate of θ per sequence for population ancestral to population 1 and population 2.

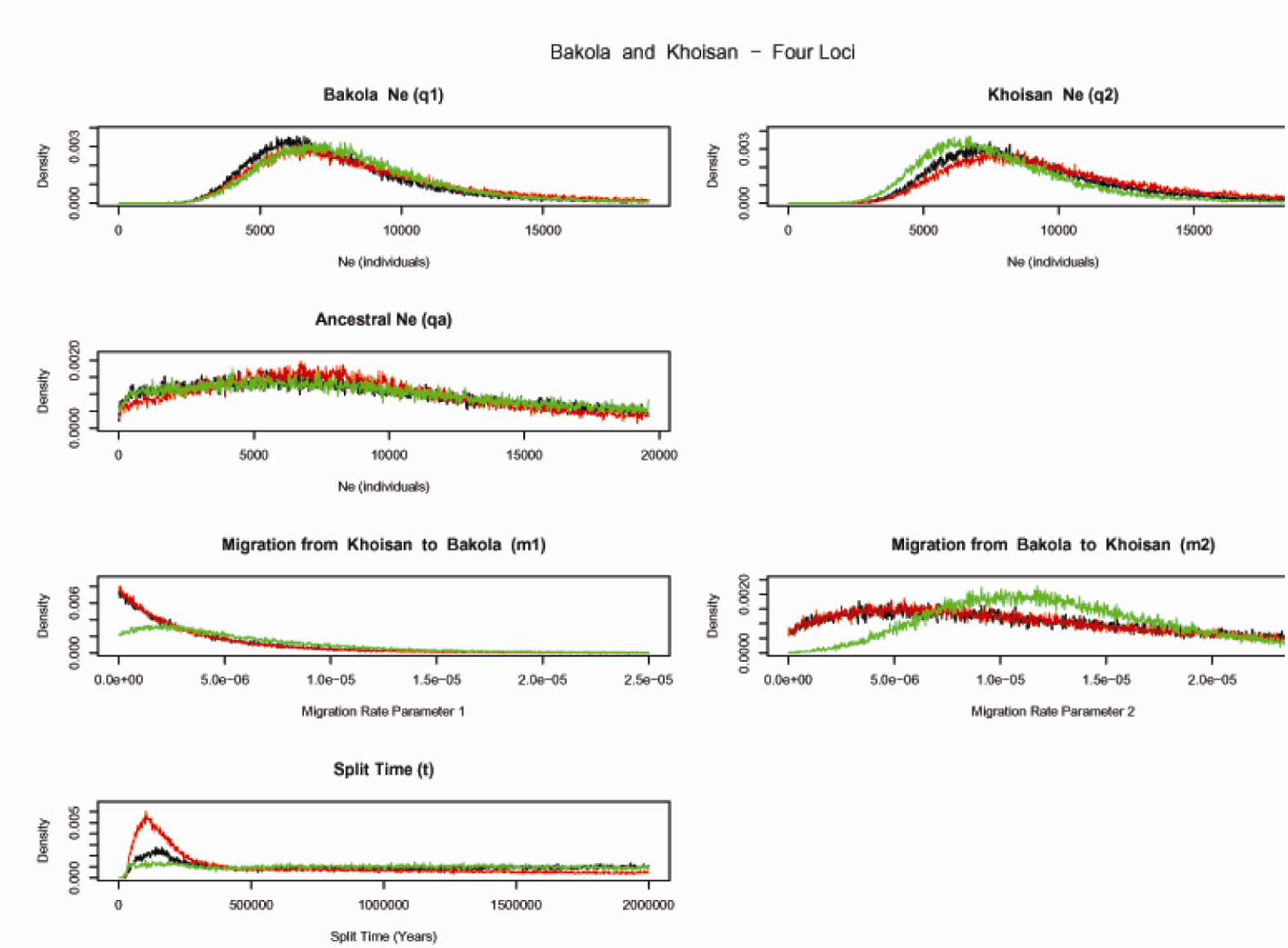
[d] Estimate of migration rate into population 1.

[e] Estimate of migration rate into population 2.

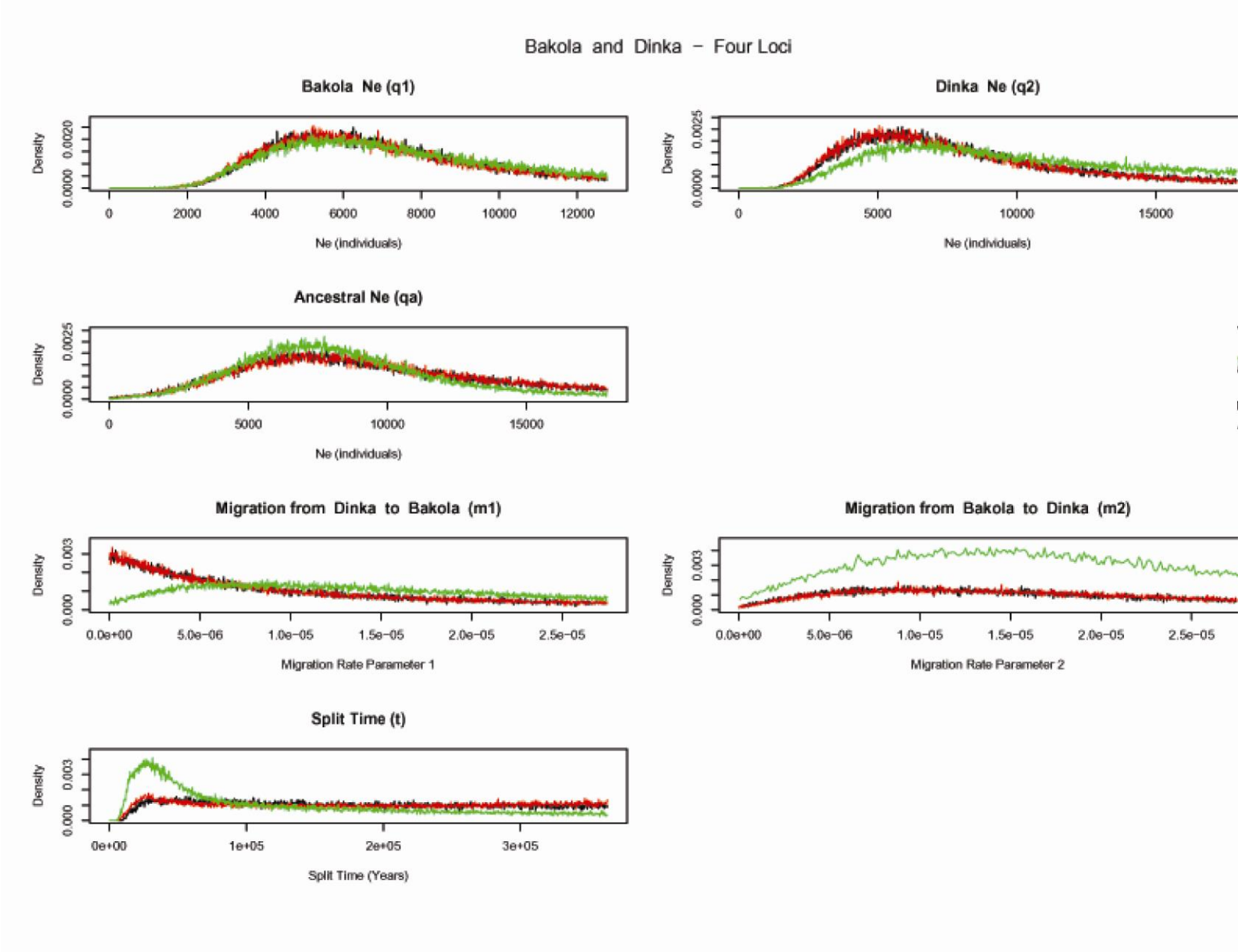
SUPPLEMENTARY FIGURE 5.1 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Bakola and Dogon comparison.



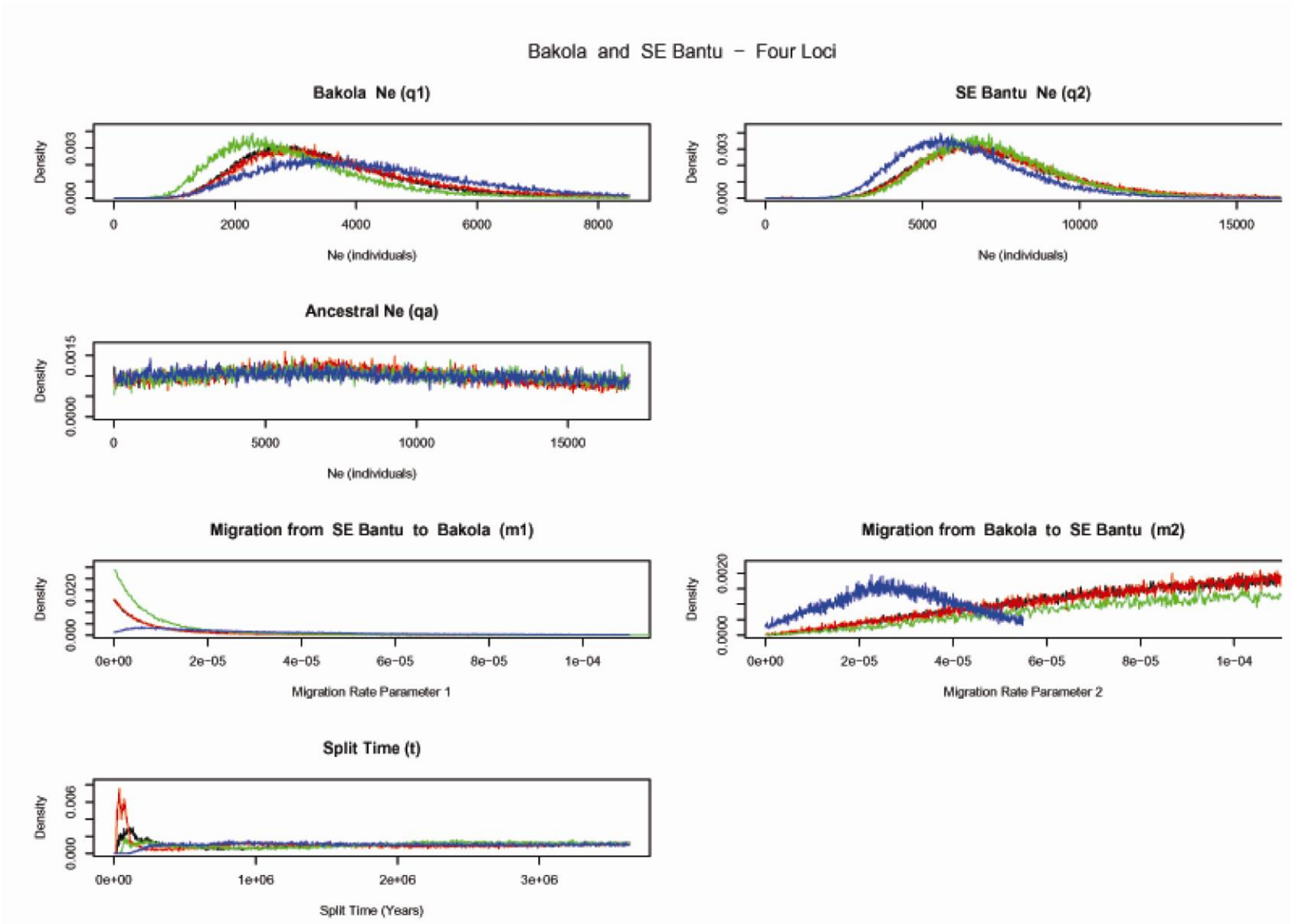
SUPPLEMENTARY FIGURE 5.2 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Bakola and Khoisan comparison.



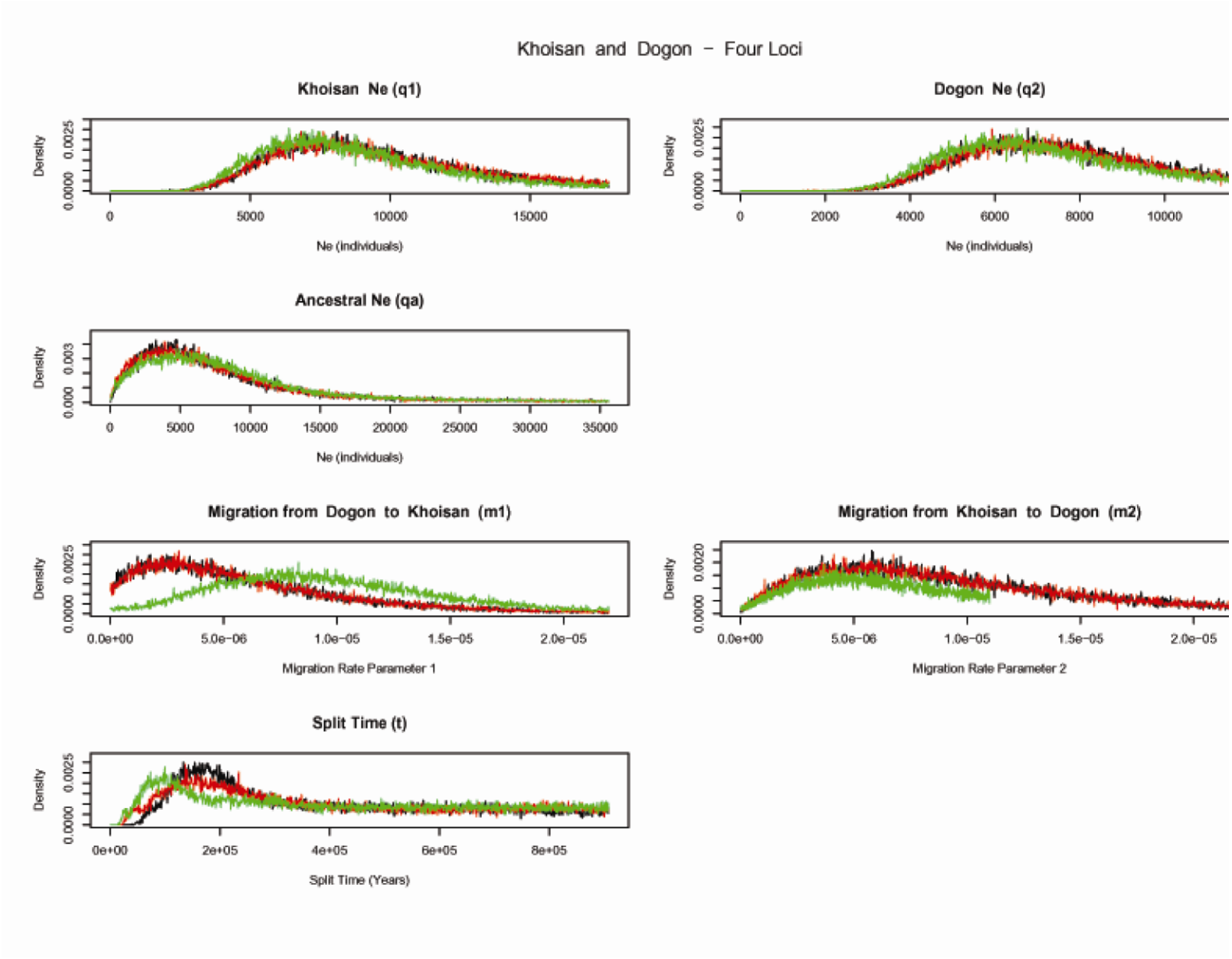
SUPPLEMENTARY FIGURE 5.3 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Bakola and Dinka comparison.



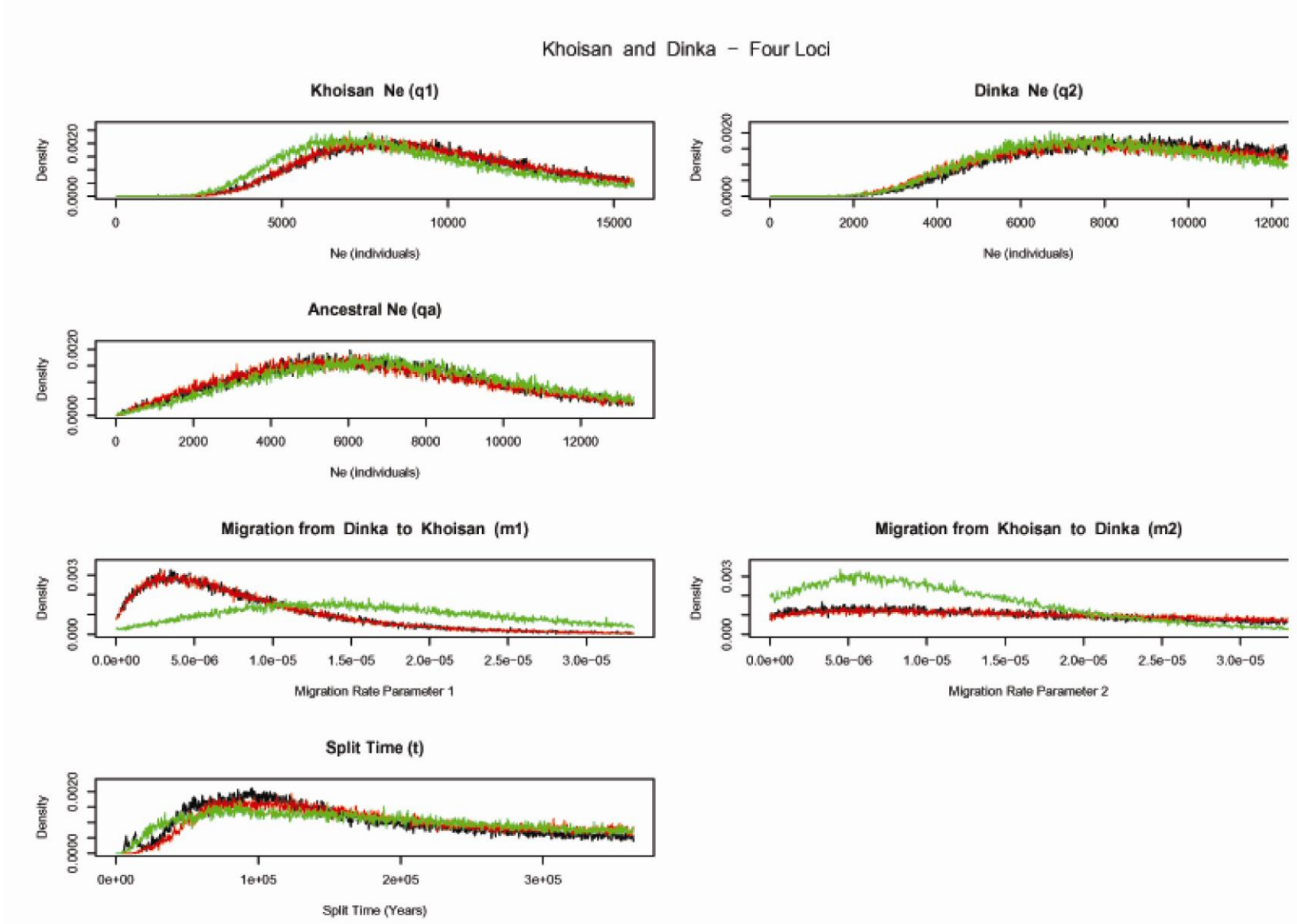
SUPPLEMENTARY FIGURE 5.4 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Bakola and SE Bantu comparison.



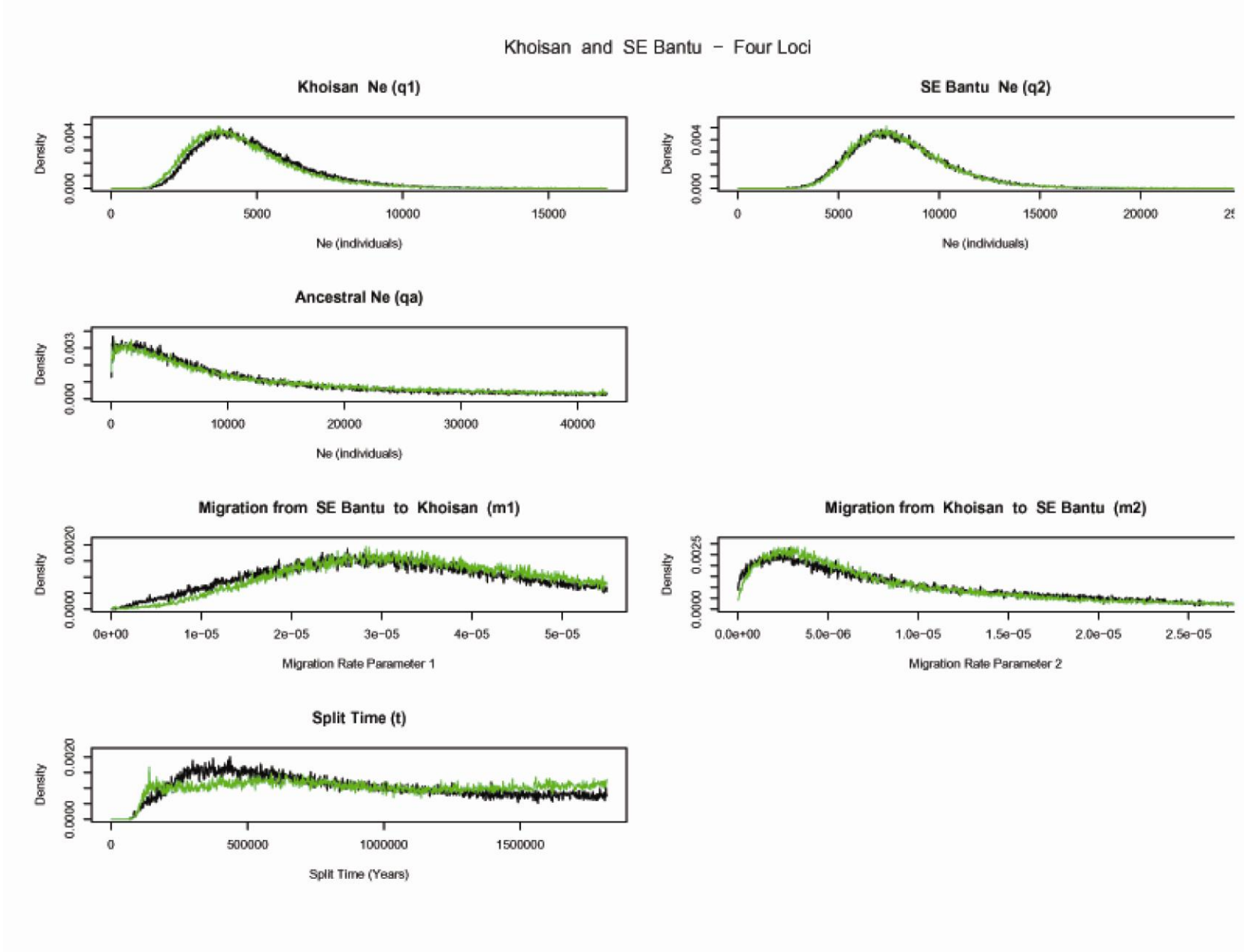
SUPPLEMENTARY FIGURE 5.5 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Khoisan and Dogon comparison.



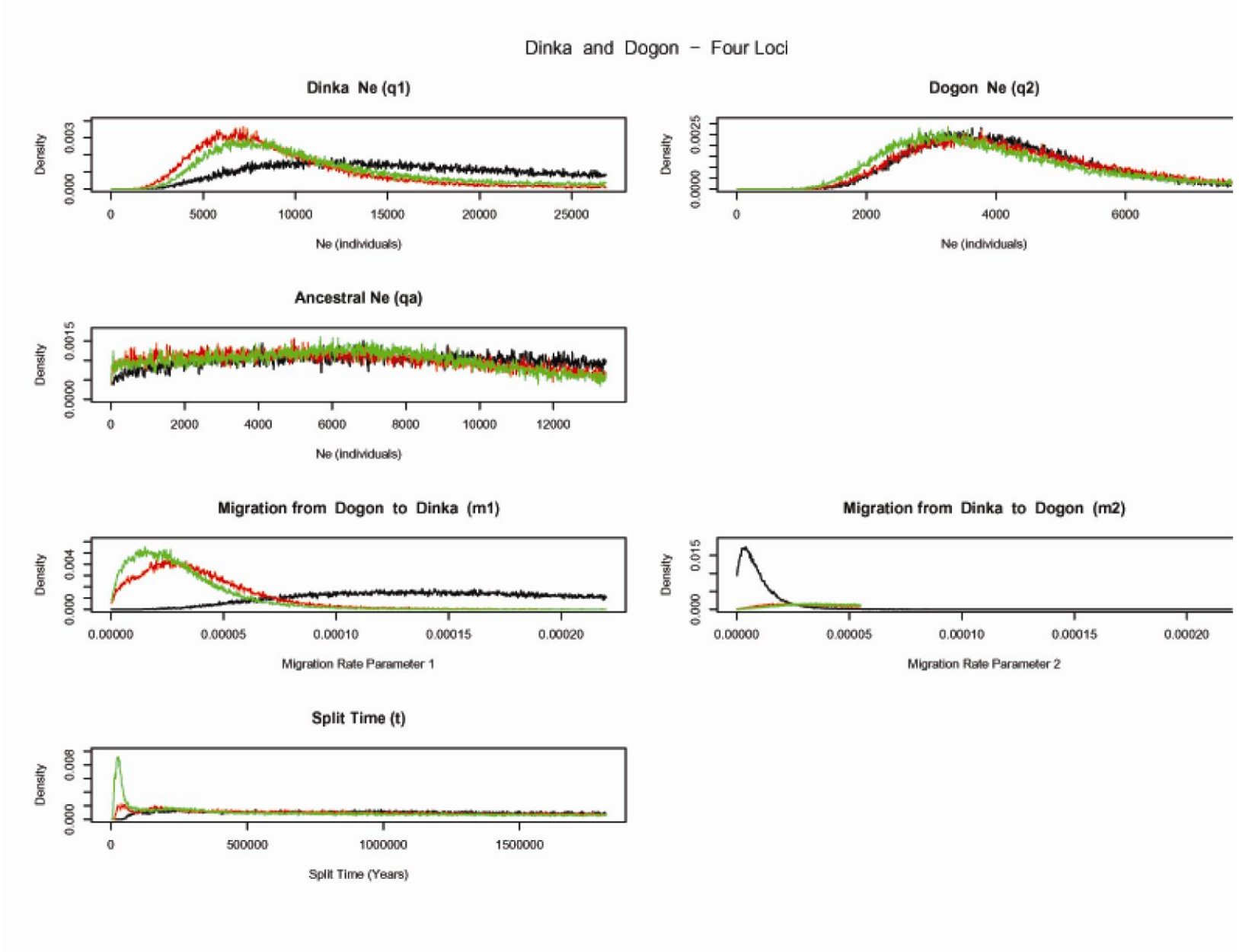
SUPPLEMENTARY FIGURE 5.6 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Khoisan and Dinka comparison.



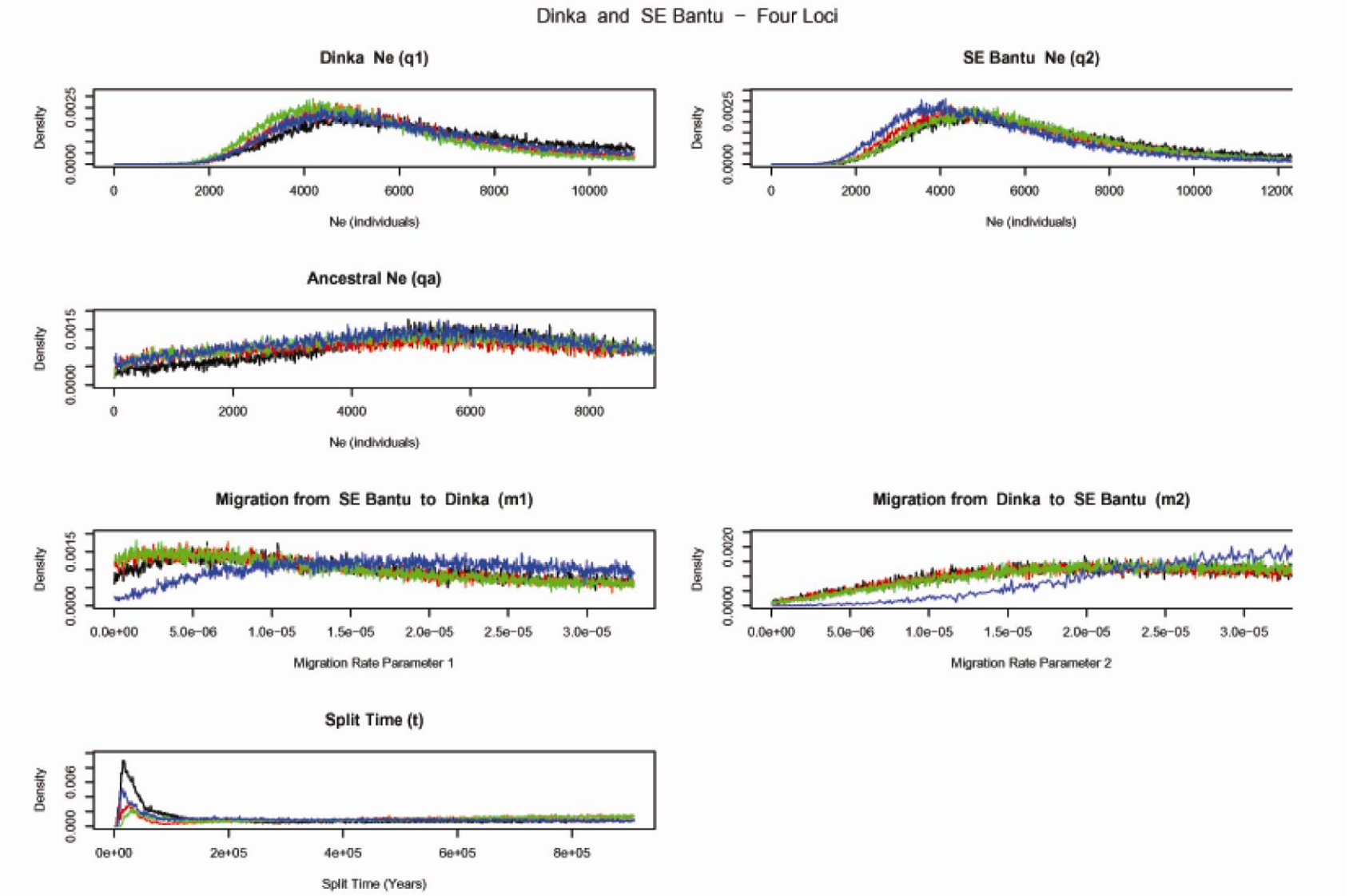
SUPPLEMENTARY FIGURE 5.7 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Khoisan and SE Bantu comparison.



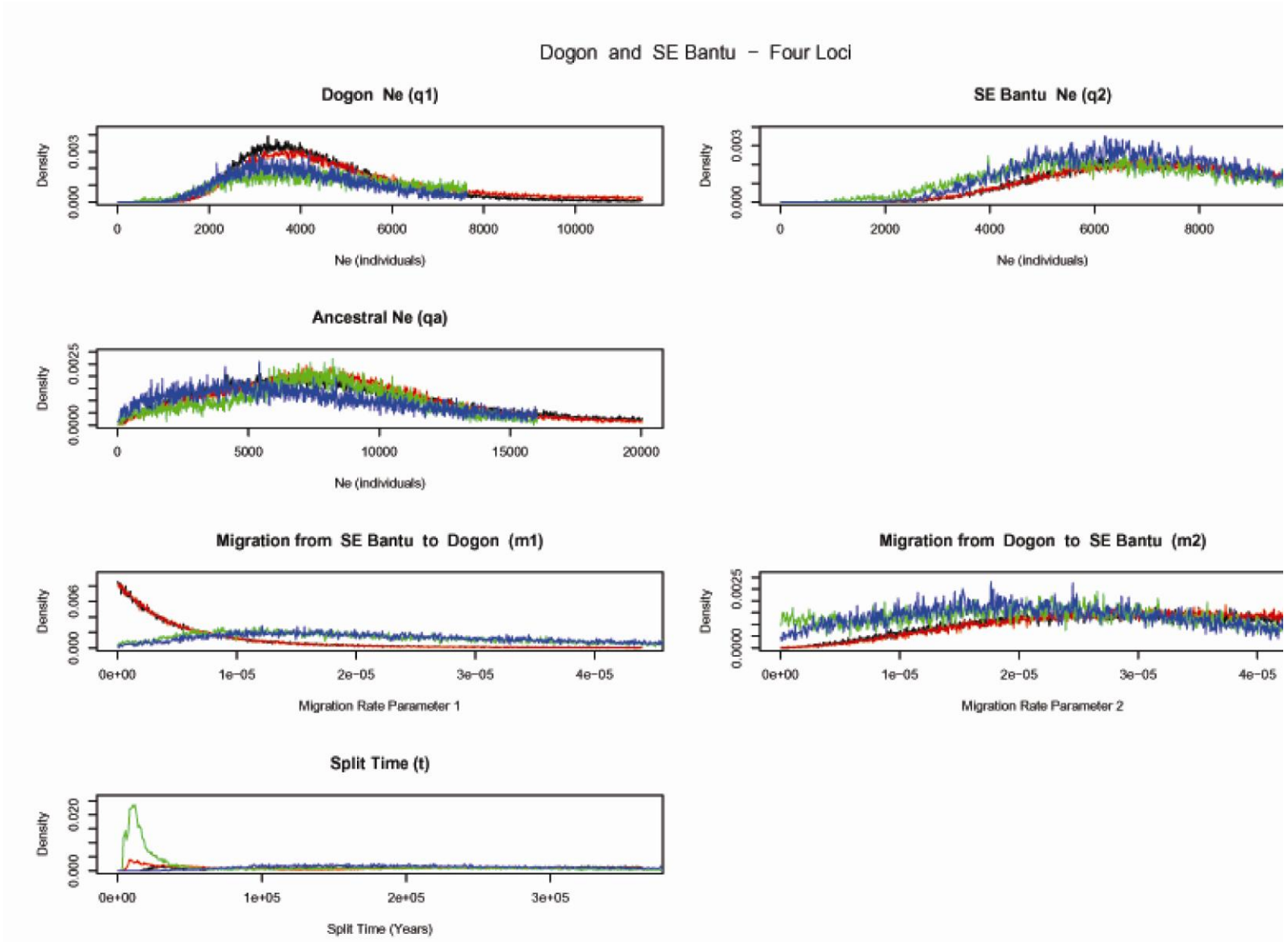
SUPPLEMENTARY FIGURE 5.8 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Dinka and Dogon comparison.



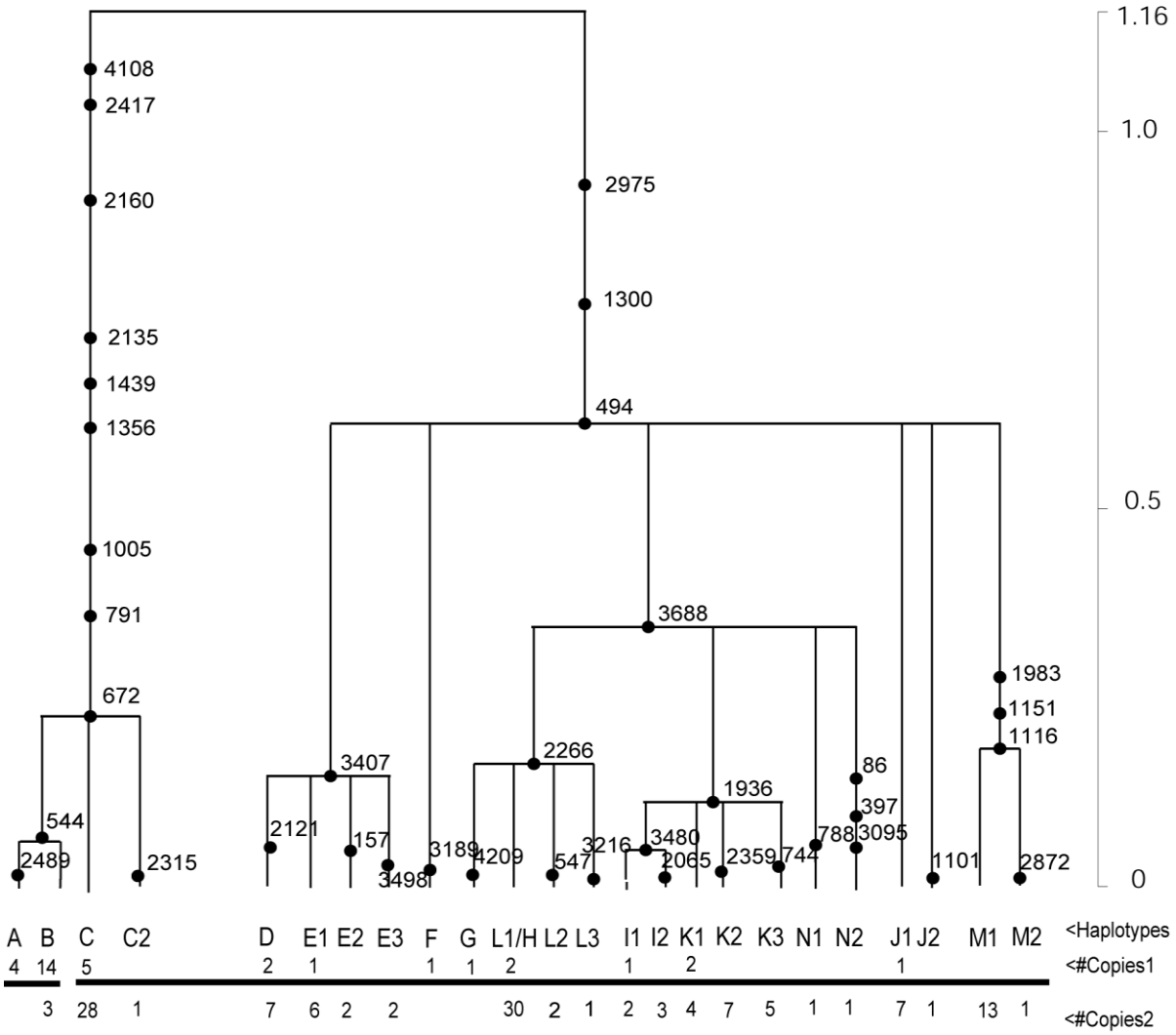
SUPPLEMENTARY FIGURE 5.9 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Dinka and SE Bantu comparison.



SUPPLEMENTARY FIGURE 5.10 The posterior probability distributions for the effective population sizes, migration rates, and split times for the Dogon and SE Bantu comparison.



SUPPLEMENTARY FIGURE 5.11: The gene tree estimate for *PDHA1*, with estimated ages of polymorphic mutations in millions of years (MYR). Mutations are identified by their position in the sequence and by their location on the branch as estimated by maximum likelihood (GENETREE). The position of mutations on branches having multiple mutations is arbitrary. Haplotypes are labeled below the tree, as are the number of individuals reported by Harris and Hey (1999) ("# Copies 1") and the number of individuals reported in this study ("# Copies 2"), with a line drawn below the non-Africans reported by Harris and Hey to the left and the Africans to the right. All of the current samples are sub-Saharan Africans.



CHAPTER 6: SUMMARY

SUMMARY

In an effort to better understand patterns of genetic variation in modern African populations, I surveyed nucleotide variability at four unlinked loci in five diverse sub-Saharan African populations. Detailed background, methods, results, and discussion for each of three studies are presented in **CHAPTER 3, CHAPTER 4** and **CHAPTER 5** of this dissertation. The following is a summary of some of the most important findings from these chapters.

CHAPTER 3 presents an analysis of mtDNA and NRY population history, asking specifically if similar models of population size change (i.e., constant size or exponential growth) can be fit to re-sequencing data from these two haploid loci when examined in the same populations. I used four tests of population growth: Fu's *Fs* statistic, the R_2 statistic, GENETREE coalescent simulations, and mismatch distributions. Results indicated that although mtDNA and the NRY are both haploid and are expected to fit similar patterns of population size change, the two loci in fact reveal different population histories for males and females. The mtDNA data indicate that food-producers best fit a model of exponential growth and hunter-gatherers best fit a model of constant population size (similar to previous analyses), while the NRY data are consistent in indicating that all five populations best fit a model of constant population size. I argue that these results are not the consequence of non-neutral evolutionary forces such as selection, but rather are more likely due to differences in sex-specific demographic processes in general, and asymmetrical migration or differences in the effective population sizes of males and females in particular.

CHAPTER 4 follows up on the results of **CHAPTER 3** by examining mtDNA and NRY population structure in these same populations. In this chapter, I combined traditional methods of examining population structure which assume an island model of migration (F_{ST} and

AMOVA) with a more realistic coalescent-based MCMC approach that employs a population splitting model with the possibility for subsequent gene flow (the "Isolation with Migration" model). Results indicated that the estimates of levels of female population structure are higher than those of males within this geographically broad sampling scheme. Interestingly, I also observed a pattern of unidirectional female migration and bidirectional male migration. I suggest that the Bantu expansion, which is estimated to have occurred approximately 3-5,000 years ago, may have differentially affected males and females in that it may have served to decrease the effective population sizes of males relative to females either through replacement or polygyny. This scenario is supported by estimates of the effective population sizes of males and females from **CHAPTER 3** which demonstrate the males in fact do have lower estimated effective sizes than females in the same populations. Through coalescent simulations, I demonstrate that bidirectional male migration and unidirectional female migration coupled with a reduction in the male effective population size could result in the differing mtDNA and NRY estimates of the TMRCA.

CHAPTER 5 further assesses population structure in these populations through an examination of four unlinked, neutrally evolving loci (mtDNA, NRY, and two X-linked loci). Here I used the IM MCMC method to simultaneously estimate the current and ancestral effective population sizes, migration rates, split times, and fraction of the ancestral population that contributed to the current populations in a total of ten population comparisons. The use of multiple loci allowed for a better estimate of each of these parameters. Current and ancestral effective population sizes ranged from ~5,000-8,000 individuals. Contrary to many previous studies, I found that most populations, including the hunter-gatherers, have increased in size relative to the ancestral population. Population split times ranged from 17-142 KYR, with the Khoisan split times being

the oldest and the Niger-Congo speaking populations' split times the most recent. Since the oldest population split times (142 KYR) precede the dates for the earliest modern humans outside of Africa, I posit that it is likely that modern humans evolved at a time when structured populations already existed in Africa.

REFERENCES

Aitken, R. J., and J. A. Marshall Graves. 2002. The future of sex. Nature **415**:963.

Alonso, S., and J. A. L. Armour. 2001. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. Proc.Natl.Acad.Sci.USA **98**:864-869.

Andrews, P. 1984. An alternative interpretation of the characters used to define *Homo erectus* Cour.Forsch.Inst.Senckenberg **69**:167-175.

Aris-Brosou, S., and L. Excoffier. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol.Biol.Evol. **13**:494-504.

Asfaw, B., W. H. Gilbert, Y. Beyene, W. K. Hart, P. Renne, G. WoldeGabriel, E. Vrba, and T. D. White. 2002. Remains of *Homo erectus* from Bouri, Middle Awash, Ethiopia. Nature **416**:317-320.

Bahlo, M., and R. C. Griffiths. 2000. Inference from gene trees in a subdivided population. Theor.Popul.Biol. **57**:79-95.

Bailey, R. C., and I. DeVore. 1989. Research on the Efe and Lese populations of the Ituri Forest, Zaire. American Journal of Primatology **78**:459-471.

Bailey, R. C., M. R. Jenike, P. T. Ellison, G. R. Bentley, A. M. Harrigan, and N. R. Peacock. 1992. The ecology of birth seasonality among agriculturalists in central Africa. J Biosoc Sci **24**:393-412.

Ballard, J. W., and M. D. Dean. 2001. The mitochondrial genome: mutation, selection and recombination. Curr.Opin.Genet.Dev. **11**:667-672.

Bamshad, M. J., S. Wooding, W. S. Watkins, C. T. Ostler, M. A. Batzer, and L. B. Jorde. 2003. Human population genetic structure and inference of group membership. Am.J.Hum.Genet. **72**:578-589.

Barnard, A. 1992. Hunters and Herders of Southern Africa. Cambridge University Press, Cambridge.

Barton, N. H. 2001. Speciation. Trends in Ecology and Evolution **16** 325.

Batzer, M. A., S. S. Arcot, J. W. Phinney, M. Alegria-Hartman, D. H. Kass, S. M. Milligan, C. Kimpton, P. Gill, M. Hochmeister, P. A. Ioannou, R. J. Herrera, D. A. Boudreau, W. D. Scheer, B. J. Keats, P. L. Deininger, and M. Stoneking. 1996. Genetic variation of recent *Alu* insertions in human populations. J.Mol.Evol. **42**:22-29.

Batzer, M. A., and P. L. Deininger. 2002. Alu repeats and human genomic diversity. Nat.Rev.Genet. **3**:370-379.

Beerli, P., and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152**:763-773.

Beleza, S., L. Gusmao, A. Amorim, A. Carracedo, and A. Salas. 2005. The genetic legacy of western Bantu migrations. Hum. Genet. **117**:366-375.

Bentley, G. R., R. Aunger, A. M. Harrigan, M. Jenike, R. C. Bailey, and P. T. Ellison. 1999. Women's strategies to alleviate nutritional stress in a rural African society. Soc Sci Med **48**:149-162.

Bertorelle, G., and M. Slatkin. 1995. The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. Mol.Biol.Evol. **12**:887-892.

Biesele, M., and K. Royal. 1999. Africa; Mbuti. Pp. 210–214 *in* B. Richard, and R. Daly, eds. The Cambridge encyclopedia of hunters and gatherers. Cambridge University Press, Cambridge.

Birdsell, J. B. 1951. Some implications of the genetical concept of race in terms of spatial analysis. Cold Spring Harb. Symp. Quant. Biol. **15**.

Blench, R., and M. Spriggs. 1999. Archaeology and Language III: Artefacts, Languages and Texts Routledge, New York.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140**:783-796.

Brown, W. M., M. George, Jr., and A. C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. Proc.Natl.Acad.Sci.USA **76**:1967-1971.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson. 1982. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. J.Mol.Evol. **18**:225-239.

Bruges Armas, J., G. Destro-Bisol, A. Lopez-Vazquez, A. R. Couto, G. Spedini, S. Gonzalez, C. Battaggia, M. J. Peixoto, J. Martinez-Borra, and C. Lopez-Larrea. 2003. HLA class I variation in the West African Pygmies and their genetic relationship with other African populations. Tissue Antigens **62**:233-242.

Brunet, M., F. Guy, D. Pilbeam, D. E. Lieberman, A. Likius, H. T. Mackaye, M. S. Ponce de Leon, C. P. Zollikofer, and P. Vignaud. 2005. New material of the earliest hominid from the Upper Miocene of Chad. Nature **434**:752-755.

Cameron, D. W. 2003. Early hominin speciation at the Plio/Pleistocene transition. Homo **54**:1-28.

Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature **325**:31-36.

Cavalli-Sforza, L. 1966. Population structure and human evolution. Proc.R.Soc.Lond.B Biol.Sci. **164**:362-379.

Cavalli-Sforza, L. 1986. African Pygmies. Academic Press, Orlando, Florida.

Cavalli-Sforza, L. L. 1959. Some data on the genetic structure of human populations Proceedings X International Congress Genetics **1**:389-407.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. The History and Geography of Human Genes. Princeton University Press, Princeton.

Cavelier, L., A. Johannisson, and U. Gyllensten. 2000. Analysis of mtDNA copy number and composition of single mitochondrial particles using flow cytometry and PCR. Exp. Cell. Res. **259**:79-85.

Cazes, M. H. 1990. Endogamy among the Dogon of Boni, Mali. J. Biosoc. Sci. **22**:85-99.

Cazes, M. H. 1986. Genetic origins of the Dogon population in the Arrondissement of Boni (Mali). Am.J.Hum.Genet. **39**:96-111.

Chen, Y. S., A. Olckers, T. G. Schurr, A. M. Kogelnik, K. Huoponen, and D. C. Wallace. 2000. mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. Am.J.Hum.Genet. **66**:1362-1383.

Chen, Y. S., A. Torroni, L. Excoffier, A. S. Santachiara-Benerecetti, and D. C. Wallace. 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am.J.Hum.Genet. **57**:133-149.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C. F. Sing. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am.J.Hum.Genet. **63**:595-612.

Clark, J. D., Y. Beyene, G. WoldeGabriel, W. K. Hart, P. R. Renne, H. Gilbert, A. Defleur, G. Suwa, S. Katoh, K. R. Ludwig, J. R. Boisserie, B. Asfaw, and T. D. White. 2003. Stratigraphic, chronological and behavioural contexts of Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. Nature **423**:747-752.

Clark, J. D., J. de Heinzelin, K. D. Schick, W. K. Hart, T. D. White, G. WoldeGabriel, R. C. Walter, G. Suwa, B. Asfaw, and E. Vrba. 1994. African *Homo erectus*: Old radiometric ages and young Oldowan assemblages in the Middle Awash Valley, Ethiopia. Science **264**:1907-1910.

Cox, M. P., F. L. Mendez, T. M. Karafet, M. M. Pilkington, S. B. Kingan, G. Destro-Bisol, B. I. Strassmann, and M. F. Hammer. 2008. Testing for archaic hominin admixture on the X chromosome: Model likelihoods for the modern human RRM2P4 region from summaries of genealogical topology under the structured coalescent. Genetics **178**:427-437.

Cracraft, J. 1983. Species concepts and speciation analysis. Curr.Ornithol. **1** 159-187.

Crow, J. F. 1958. Some possibilities for measuring selection intensities in man. Hum. Biol. **30**:1-13.

Cruciani, F., P. Santolamazza, P. Shen, V. Macaulay, P. Moral, A. Olckers, D. Modiano, S. Holmes, G. Destro-Bisol, V. Coia, D. C. Wallace, P. J. Oefner, A. Torroni, L. L. Cavalli-Sforza, R. Scozzari, and P. A. Underhill. 2002. A back migration from Asia to sub-

Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am.J.Hum.Genet. **70**:1197-1214.

Deng, F. M. 1972. The Dinka of Sudan. Holt, Rinehart and Winston, New York.

Destro-Bisol, G., V. Coia, I. Boschi, F. Verginelli, A. Caglia, V. Pascali, G. Spedini, and F. Calafell. 2004a. The analysis of variation of mtDNA hypervariable region 1 suggests that Eastern and Western Pygmies diverged before the Bantu expansion. American Naturalist **163**:212-226.

Destro-Bisol, G., F. Donati, V. Coia, I. Boschi, F. Verginelli, A. Caglia, S. Tofanelli, G. Spedini, and C. Capelli. 2004b. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol.Biol.Evol. **21**:1673-1682.

Di Rienzo, A., and A. C. Wilson. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc.Natl.Acad.Sci.USA **88**:1597-1601.

Dobzhansky, T. 1937. Genetics and the Origins of Species. Columbia University Press, New York.

Dorit, R. L., H. Akashi, and W. Gilbert. 1995. Absence of polymorphism at the ZFY locus on the human Y chromosome. Science **268**:1183-1185.

Dorjahn, V. R. 1959. The Factor of Polygyny in African Demography. Pp. 87-112 *in* W. R. Bascom, and J. H. Melville, eds. Continuity and Change in African Cultures. University of Chicago Press, Chicago.

Dupanloup, I., L. Pereira, G. Bertorelle, F. Calafell, M. J. Prata, A. Amorim, and G. Barbujani. 2003. A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J.Mol.Evol. **57**:85-97.

Ehret, C. 2001. Bantu expansions: Re-envisioning a central problem of early African history. The International Journal of African Historical Studies **34**:5-41.

Ehret, C. 1972. Bantu origins and history: Critique and interpretation. Transafrican Journal of History **2**:1-9.

Eller, E. 2002. Population extinction and recolonization in human demographic history. Math Biosci **177-178**:1-10.

Elson, J. L., D. M. Turnbull, and N. Howell. 2004. Comparative genomics and the evolution of human mitochondrial DNA: Assessing the effects of selection. Am.J.Hum.Genet. **74**:229-238.

Evans, P. D., N. Mekel-Bobrov, E. J. Vallender, R. R. Hudson, and B. T. Lahn. 2006. Evidence that the adaptive allele of the brain size gene microcephalin introgressed into *Homo sapiens* from an archaic *Homo* lineage. Proc Natl Acad Sci U S A **103**:18178-18183.

Excoffier, L. 2002. Human demographic history: Refining the recent African origin model. Curr. Opin. Genet. Dev. **12**:675-682.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online **1**:47-50.

Excoffier, L., and S. Schneider. 1999. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc.Natl.Acad.Sci.USA **96**:10597-10602.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics **131**:479-491.

Eyre-Walker, A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics **162**:2017-2024.

Fay, J. C., and C. I. Wu. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial *versus* nuclear DNA variation. Mol.Biol.Evol. **16**:1003-1005.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Fix, A. G. 1978. The role of kin-structure migraton in genetic microdifferentiation. Annals of Human Genetics, London **41**:329-339.

Fix, A. G. 1999. Migration and Colonization in Human Microevolution. Cambridge University Press, Cambridge

Franqueville. 1971. Atlas régional du Cameroun : Sud-Ouest I. ORSTOM Paris.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack, and A. Di Rienzo. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am.J.Hum.Genet. **69**:831-843.

Fu, Y. X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**:915-925.

Gage, T. B. 2000. Demography *in* S. Stinson, B. bogin, R. Huss-Ashmore, and D. O'Rourke, eds. Human Biology: An Evolutionary and Biocultural Perspective. Wiley-Liss, New York.

Garrigan, D., and M. F. Hammer. 2006. Reconstructing human origins in the genomic era. Nat.Rev.Genet. **7**:669-680.

Garrigan, D., S. B. Kingan, M. M. Pilkington, J. A. Wilder, M. P. Cox, H. Soodyall, B. Strassmann, G. Destro-Bisol, P. de Knijff, A. Novelletto, J. Friedlaender, and M. F. Hammer. 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics **177**:2195-2207.

Garrigan, D., Z. Mobasher, S. B. Kingan, J. A. Wilder, and M. F. Hammer. 2005a. Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics **170**:1849-1856.

Garrigan, D., Z. Mobasher, T. Severson, J. A. Wilder, and M. F. Hammer. 2005b. Evidence for archaic Asian ancestry on the human X chromosome. Mol.Biol.Evol. **22**:189-192.

Gilad, Y., S. Rosenberg, M. Przeworski, D. Lancet, and K. Skorecki. 2002. Evidence for positive selection and population structure at the human *MAO-A* gene. Proc.Natl.Acad.Sci.USA **99**:862-867.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*. A multi-locus approach. Genetics **165**:1269-1278.

Goldstein, D. B., and L. Chikhi. 2002. Human migrations and population structure: What we know and why it matters. Annu.Rev.Genomics Hum.Genet. **3**:129-152.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and Feldman.M.W. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc.Natl.Acad.Sci.USA **92**:6723-6727.

Goldstein, D. B., L. A. Zhivotovsky, K. Nayar, A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1996. Statistical properties of the variation at linked microsatellite loci: Implications for the history of human Y chromosomes. Mol.Biol.Evol. **13**:1213-1218.

Gonder, M. K., H. M. Mortensen, F. A. Reed, A. de Sousa, and S. A. Tishkoff. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol.Biol.Evol. **24**:757-768.

Gordon, R. G., Jr. 2005. Ethnologue: Languages of the World, Fifteenth Edition, Online version: http://www.ethnologue.com/. SIL International, Dallas, Texas.

Graunt, J. 1662. Natural and Political Observations Made upon the Bills of Mortality, London.

Greenberg, J. H. 1963. The Languages of Africa. Indiana University Publication, Bloomington.

Greenberg, J. H. 1972. Linguistic evidence regarding Bantu origins. J. Afr. Hist. **23**:189-216.

Griffiths, R. C., and S. Tavare. 1994. Sampling theory for neutral alleles in a varying environment. Philos.Trans.R.Soc.Lond. B Biol.Sci. **344**:403-410.

Grun, R., C. Stringer, F. McDermott, R. Nathan, N. Porat, S. Robertson, L. Taylor, G. Mortimer, S. Eggins, and M. McCulloch. 2005. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. J.Hum.Evol. **49**:316-334.

Grun, R., C. B. Stringer, and H. P. Schwarcz. 1991. ESR dating of teeth from Garrod's Tabun cave collection. J.Hum.Evol. **20**:231-248.

Guthrie, M. 1963. Bantu Origins: A Tentative New Hypothesis. Journal of African Languages **1**:9-21.

Gyllensten, U., D. Wharton, and A. C. Wilson. 1985. Maternal inheritance of mitochondrial DNA during backcrossing of two species of mice. J. Hered. **76**:321-324.

Haile-Selassie, Y. 2001. Late Miocene hominids from the Middle Awash, Ethiopia. Nature **412**:178-181.

Haile-Selassie, Y., B. Asfaw, and T. D. White. 2004. Hominid cranial remains from upper Pleistocene deposits at Aduma, Middle Awash, Ethiopia. **123**:1-10.

Haldane, J. B. S. 1932. The Causes of Evolution. Longmans Green, London.

Hamblin, M. T., E. E. Thompson, and A. Di Rienzo. 2002. Complex signatures of natural selection at the Duffy blood group locus. Am.J.Hum.Genet. **70**:369-383.

Hammer, M. F. 1995. A recent Y common ancestry for human Y chromosomes. Nature **378**:376-378.

Hammer, M. F., A. Bigham, D. Garrigan, J. Krenz, Z. Mobasher, M. W. Nachman, J. A. Wilder, and E. Wood. 2004a. Heterogeneous patterns of variation among multiple X-linked loci in humans: The possible role of diversity reducing selection. Genetics.

Hammer, M. F., A. Bigham, D. Garrigan, J. Krenz, Z. Mobasher, M. W. Nachman, J. A. Wilder, and E. Wood. 2004b. Heterogeneous patterns of variation among multiple X-linked loci in humans: The possible role of diversity reducing selection. Genetics **167**:1841-1853.

Hammer, M. F., F. Blackmer, D. Garrigan, M. W. Nachman, and J. A. Wilder. 2003. Human population structure and its effects on sampling Y chromosome sequence variation. Genetics **164**:1495-1509.

Hammer, M. F., D. Garrigan, E. Wood, J. A. Wilder, Z. Mobasher, A. Bigham, J. G. Krenz, and M. W. Nachman. 2004c. Heterogeneous patterns of variation among multiple human x-linked Loci: the possible role of diversity-reducing selection in non-africans. Genetics **167**:1841-1853.

Hammer, M. F., T. Karafet, A. Rasanayagam, E. T. Wood, T. K. Altheide, T. Jenkins, R. C. Griffiths, A. R. Templeton, and S. L. Zegura. 1998. Out of Africa and back again: Nested cladistic analysis of human Y chromosome variation. Mol.Biol.Evol. **15**:427-441.

Hammer, M. F., T. M. Karafet, A. J. Redd, H. Jarjanazi, S. Santachiara-Benerecetti, H. Soodyall, and S. L. Zegura. 2001. Hierarchical patterns of global human Y-chromosome diversity. Mol.Biol.Evol. **18**:1189-1203.

Hammer, M. F., A. B. Spurdle, T. Karafet, M. R. Bonner, E. T. Wood, A. Novelletto, P. Malaspina, R. J. Mitchell, S. Horai, T. Jenkins, and S. L. Zegura. 1997. The geographic distribution of human Y chromosome variation. Genetics **145**:787-805.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox, J. A. Schneider, D. S. Moulin, and J. B. Clegg. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am.J.Hum.Genet. **60**:772-789.

Harpending, H., and A. Rogers. 2000. Genetic perspectives on human origins and differentiation. Annu. Rev. Genomics Hum. Genet. **1**:361-385.

Harpending, H., S. T. Sherry, A. R. Rogers, and M. Stoneking. 1993. The genetic structure of ancient human populations. Curr.Anthropol. **34** 483-496.

Harpending, H. C. 1994. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum.Biol. **66**:591-600.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers, and S. T. Sherry. 1998. Genetic traces of ancient demography. Proc.Natl.Acad.Sci.USA **95**:1961-1967.

Harris, E., and J. Hey. 1999a. Human demography in the Pleistocene: Do mitochondrial and nuclear genes tell the same story? Evol.Anthropol. **8** 81-86.

Harris, E. E., and J. Hey. 1999b. X chromosome evidence for ancient human histories. Proc.Natl.Acad.Sci.USA **96**:3320-3324.

Harris, E. E., and J. Hey. 2001. Human populations show reduced DNA sequence variation at the *factor IX* locus. Curr.Biol. **11**:774-778.

Hartl, D. L., and A. G. Clark. 1997. Principles of Population Genetics. Sinauer Associates, Inc., Sunderland.

Hauser, P., and O. D. Duncan. 1959. The Study of Population. Chicago University Press, Chicago.

Haussler, M. R., P. W. Jurutka, J. C. Hsieh, P. D. Thompson, S. H. Selznick, C. A. Haussler, and G. K. Whitfield. 1995. New understanding of the molecular mechanism of receptor-mediated genomic actions of the vitamin D hormone. Bone **17**:33S-38S.

Hayakawa, T., I. Aki, A. Varki, Y. Satta, and N. Takahata. 2006. Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics **172**:1139-1146.

Hedrick, P. W. 2000. Genetics of Populations. Jones and Bartlett Publishers, Sudbury

Henshilwood, C. S., F. d'Errico, R. Yates, Z. Jacobs, C. Tribolo, G. A. Duller, N. Mercier, J. C. Sealy, H. Valladas, I. Watts, and A. G. Wintle. 2002. Emergence of modern human behavior: Middle Stone Age engravings from South Africa. Science **295**:1278-1280.

Hey, J. 2005. On the number of New World founders: A population genetic portrait of the peopling of the Americas. PLoS Biol **3**:e193.

Hey, J. 1997. Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol.Biol.Evol. **14**:166-172.

Hey, J., and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics **167**:747-760.

Hiernaux, J. 1968. Bantu expansion: The evidence from physical anthopology confronted with linguistic and archaeological evidence. Journal of African History **IX**:505-515.

Hochstetler, J. L., and J. A. D.-B. Durieux, E.I.K. . 2004. Sociolinguistic Survey of the Dogon Language Area. SIL International.

Holden, C. J. 2002a. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. Proceedings: Biological Sciences **269**:793-799.

Holden, C. J. 2002b. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. Proc Biol Sci **269**:793-799.

Horai, S., K. Hayasaka, R. Kondo, K. Tsugane, and N. Takahata. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc.Natl.Acad.Sci.USA **92**:532-536.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics **132**:583-589.

Huxley, J. 1942. Evolution: The Modern Synthesis. Harper, New York.

Ingman, M., and U. Gyllensten. 2001. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J. Hered. **92**:454-461.

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

Jaruzelska, J., E. Zietkiewicz, and D. Labuda. 1999. Is selection responsible for the low level of variation in the last intron of the *ZFY* locus? Mol.Biol.Evol. **16**:1633-1640.

Jolly, C. J. 2001. A proper study for mankind: Analogies from the papionin monkeys and their implications for human evolution. Yearbook of Physical Anthropology:177-204.

Jorde, L. B., M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley, P. A. Krakowiak, K. D. Carpenter, H. Soodyall, T. Jenkins, and A. R. Rogers. 1995. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. Am.J.Hum.Genet. **57**:523-538.

Jorde, L. B., A. R. Rogers, M. Bamshad, W. S. Watkins, P. Krakowiak, S. Sung, J. Kere, and H. C. Harpending. 1997. Microsatellite diversity and the demographic history of modern humans. Proc.Natl.Acad.Sci.USA **94**:3100-3103.

Jorde, L. B., W. S. Watkins, and M. J. Bamshad. 2001. Population genomics: a bridge from evolutionary history to genetic medicine. Hum.Mol.Genet. **10**:2199-2207.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisited. Genetics **123**:887-899.

Ke, Y., B. Su, X. Song, D. Lu, L. Chen, H. Li, C. Qi, S. Marzuki, R. Deka, P. Underhill, C. Xiao, M. Shriver, J. Lell, D. Wallace, R. S. Wells, M. Seielstad, P. Oefner, D. Zhu, J. Jin, W.

Huang, R. Chakraborty, Z. Chen, and L. Jin. 2002. African origin of modern humans in East Asia: A tale of 12,000 chromosomes. **292**:1151-1153.

Keyfitz, N., and W. Flieger. 1968. World Population: An Analysis of Vital Data. University of Chicago Press, Chicago.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J.Mol.Evol. **16**:111-120.

Kingman, J. F. C. 1982. The coalescent. Stochastic Process. Appl. **13**.

Kivisild, T., P. Shen, D. P. Wall, B. Do, R. Sung, K. Davis, G. Passarino, P. A. Underhill, C. Scharfe, A. Torroni, R. Scozzari, D. Modiano, A. Coppa, P. de Knijff, M. Feldman, L. L. Cavalli-Sforza, and P. J. Oefner. 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics **172**:373-387.

Klein, R. G. 1986. The prehistory of Stone Age herders in the Cape Province of South Africa. Pp. 5-12 *in* M. Hall, and A. B. Smith, eds. Prehistoric Pastoralism in southern Africa. South African Archaeological Society: Goodwin Series, Vlaeberg.

Knight, A., P. A. Underhill, H. M. Mortensen, L. A. Zhivotovsky, A. A. Lin, B. M. Henn, D. Louis, M. Ruhlen, and J. L. Mountain. 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr.Biol. **13**:464-473.

Konotey-Ahulu, F. 1980. Procreative superiority index (MPSI): The missing coefficient in African anthropogenetics. Brit. Med. J. **281**:1700-1702.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. **5**:150-163.

Labuda, D., and G. Striker. 1989. Sequence conservation in *Alu* evolution. Nucleic Acids Res. **17**:2477-2491.

Labuda, D., E. Zietkiewicz, and V. Yotova. 2000. Archaic lineages in the history of modern humans. Genetics **156**:799-808.

Lahn, B. T., and D. C. Page. 1999. Four evolutionary strata on the human X chromosome. Science **286**:964-967.

Lane, A. B., H. Soodyall, S. Arndt, M. E. Ratshikhopha, E. Jonker, C. Freeman, L. Young, B. Morar, and L. Toffie. 2002. Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies. Am.J.Phys.Anthropol. **119**:175-185.

Leakey, M. G., F. Spoor, F. H. Brown, P. N. Gathogo, C. Kiarie, L. N. Leakey, and I. McDougall. 2001. New hominin genus from eastern Africa shows diverse middle Pliocene lineages. Nature **410**:433-440.

Lee, R. B. 1976. Kung Spatial Organization: An Ecological and Historical Perspective *in* R. B. Lee, and I. DeVore, eds. Kalahari Hunter-Gatherers. Harvard University Press, Cambridge, Massachusetts.

Lee, S. H., and M. H. Wolpoff. 2005. Habiline variation: a new approach using STET. Theory Biosci. **124**:25-40.

Li, W. H., and L. A. Sadler. 1991. Low nucleotide diversity in man. Genetics **129**:513-523.

Loung, J. F. 1981. La population pygmée de la région côtière Camerounaise. Yaoundé Institut des Sciences Humaines

Louw, J. A. 1979. A preliminary survey of the Khoi and San influence in Zulu. Pp. 8-21 *in* A. Traill, ed. Khoisan Linguistic Studies Department of Linguistics, Johannesburg.

Low, B. S. 1988. Measures of polygyny in humans. Current Anthropology **29**:189-194.

Lum, J. K., R. L. Cann, J. J. Martinson, and L. B. Jorde. 1998. Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. Am.J.Hum.Genet. **63**:613-624.

Lundstrom, R., S. Tavare, and R. H. Ward. 1992. Modeling the evolution of the human mitochondrial genome. Math Biosci. **112**:319-335.

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

Malaspina, P., F. Persichetti, A. Novelletto, C. Iodice, L. Terrenato, J. Wolfe, M. Ferraro, and G. Prantera. 1990. The human Y chromosome shows a low level of DNA polymorphism. Ann. Hum. Genet. **54 (Pt 4)**:297-305.

Manhire, A. 1987. Later Stone Age Settlement Patterns in the Sand-Veld of the South-Western Cape Province, South Africa. British Archaelogical Reports (International Series 351) Monographs in African Archaeology **21**.

Marjoram, P., and P. Donnelly. 1994. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136**:673-683.

Marks, S., and A. Atmore. 1970. The problem of the Nguni: An examination of the ethnic and linguistic situation in South Africa before the Mfecane. Pp. 120-132 *in* D. Dalby, ed. Language and History in Africa. Cass, London.

Marlowe, F. W. 2004. Is human ovulation concealed? Evidence from conception beliefs in a hunter-gatherer society. Arch. Sex. Behav. **33**:427-432.

Marth, G., G. Schuler, R. Yeh, R. Davenport, R. Agarwala, D. Church, S. Wheelan, J. Baker, M. Ward, M. Kholodov, L. Phan, E. Czabarka, J. Murvai, D. Cutler, S. Wooding, A. Rogers, A. Chakravarti, H. C. Harpending, P. Y. Kwok, and S. T. Sherry. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. Proc.Natl.Acad.Sci.USA **100**:376-381.

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. **23**:23-35.

Mayr, E. 1963. Animal Species and Evolution. Harvard University Press, Cambridge, Massachusetts.

Mayr, E. 1942. Systematics and the Origins of Species. Columbia University Press, New York

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**:652-654.

McDougall, I., F. H. Brown, and J. G. Fleagle. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature **433**:733-736.

Metni Pilkington, M., J. A. Wilder, F. L. Mendez, M. P. Cox, A. Woerner, T. Angui, S. Kingan, S. Mobasher, C. Batini, G. Destro-Bisol, H. Soodyall, B. I. Strassmann, and M. F. Hammer. 2008. Contrasting signatures of population growth for mitochondrial and Y chromosomes among human populations in Africa. Mol Biol Evol **25**:517-525.

Meyer, S., G. Weiss, and A. Von Haeseler. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics **152**:1103-1110.

Mielke, J. H., and A. G. Fix. 2007. The Confluence of Anthropological Genetics and Anthropological Demography. Pp. 112-140 *in* M. Crawford, ed. Anthropological Genetics. Cambridge University Press, Cambridge.

Mishmar, D., E. Ruiz-Pesini, P. Golik, V. Macaulay, A. G. Clark, S. Hosseini, M. Brandon, K. Easley, E. Chen, M. D. Brown, R. I. Sukernik, A. Olckers, and D. C. Wallace. 2003. Natural selection shaped regional mtDNA variation in humans. Proc.Natl.Acad.Sci.USA **100**:171-176.

Mishmar, D., E. Ruiz-Pesini, M. Mondragon-Palomino, V. Procaccio, B. Gaut, and D. C. Wallace. 2006. Adaptive selection of mitochondrial complex I subunits during primate radiation. Gene **378**:11-18.

Mountain, J. L. 1998. Molecular evolution and modern human origins. Evol.Anthropol. **7**:21-37.

Murdock, G. P. 1959. Africa: Its Peoples and Their Culture History. McGraw-Hill Book Company, Inc., New York.

Murdock, G. P. 1967. Ethnographic Atlas. University of Pittsburgh Press Pittsburgh.

Murdock, S. H., and D. R. Ellis. 1991. Applied Demography: An Introduction to Basic Concepts, Methods, and Data. Westview Press, Boulder.

Nabulsi, A., E. Emmerich, W. Cleve, L. G. Gurtler, and H. Cleve. 1993. Haptoglobin (HP), transferrin (TF) and group-specific component (GC) subtype distribution in Bantu-speaking people from Malawi. Hum.Hered. **43**:323-325.

Nachman, M. W. 1998. Deleterious mutations in animal mitochondrial DNA. Genetica **102-103**:61-69.

Nachman, M. W., S. N. Boyer, and C. F. Aquadro. 1994. Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice. Proc.Natl.Acad.Sci.USA **91**:6364-6368.

Nachman, M. W., W. M. Brown, M. Stoneking, and C. F. Aquadro. 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics **142**:953-963.

Nachman, M. W., and S. L. Crowell. 2000. Contrasting evolutionary histories of two introns of the Duchenne Muscular Dystrophy gene, *Dmd* , in humans. Genetics **155**:1855-1864.

Nei, M., and W. H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc.Natl.Acad.Sci.USA **76**:5269-5273.

Ngima Mawoung, G. 2001. The relationship between the Bakola and the Bantu peoples of the coastal regions of Cameroon and their perception of commercial forest exploitation. African Study Monographs **Suppl.26**:209-235.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson, J. Stengard, V. Salomaa, E. Vartiainen, E. Boerwinkle, and C. F. Sing. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat.Genet. **19**:233-240.

Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. Genetics **158**:885-896.

Nurse, G. T., J. S. Weiner, and T. Jenkins. 1985. The people of southern Africa and their affinities. Clarendon Press, Oxford.

Oliver, R. 1966. The Problem of the Bantu Expansion. Journal of African History **7**:361-376.

Olivio, P. D., M. J. Van de Walle, P. J. Laipis, and W. W. Hauswirth. 1983. Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. Nature **306**:400-402.

Parkington, J. E. 1987. Changing views of prehistoric settlements in the western Cape. Pp. 4-23 *in* J. E. Parkington, and M. Hall, eds. Papers in the Prehistory of the Western Cape, South Africa. British Archaeological Reports, International Series:332, Oxford.

Parr, R. L., J. Maki, B. Reguly, G. D. Dakubo, A. Aguirre, R. Wittock, K. Robinson, J. P. Jakupciak, and R. E. Thayer. 2006. The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. BMC Genomics **7**:185.

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature **441**:1103-1108.

Penny, D., M. Steel, P. J. Waddell, and M. D. Hendy. 1995. Improved analyses of human mtDNA sequences support a recent African origin for *Homo sapiens*. Mol.Biol.Evol. **12**:863-882.

Pereira, L., I. Dupanloup, Z. H. Rosser, M. A. Jobling, and G. Barbujani. 2001. Y-chromosome mismatch distributions in Europe. Mol.Biol.Evol. **18**:1259-1271.

Petit, V., and H. Vandewalle. 1991. The methodology and initial results of the census of the Sangha district (Dogon-Mali territory). Etudes Mali **44**:39-50.

Philipson, D. W. 1977. The spread of the Bantu language. Scientific American **236**:106-114.

Phillipson, D. W. 1993. African archaeology. Cambridge University Press, Cambridge.

Plagnol, V., and J. D. Wall. 2006. Possible ancestral structure in human populations. PLoS Genet **2**:e105.

Pluzhnikov, A., A. Di Rienzo, and R. R. Hudson. 2002. Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161**:1209-1218.

Poloni, E. S., O. Semino, G. Passarino, A. S. Santachiara-Benerecetti, I. Dupanloup, A. Langaney, and L. Excoffier. 1997. Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. Am.J.Hum.Genet. **61**:1015-1035.

Posnansky, M. 1968. Bantu genesis-- archaeological reflexions. J. Afr. Hist. **9**:1-11.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Mol.Biol.Evol. **16**:1791-1798.

Przeworski, M., R. R. Hudson, and A. Di Rienzo. 2000. Adjusting the focus on human variation. Trends Genet. **16**:296-302.

Ptak, S. E., and M. Przeworski. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. Trends Genet. **18**:559-563.

Ramos-Onsins, S. E., and J. Rozas. 2002. Statistical properties of new neutrality tests against population growth. Mol.Biol.Evol. **19**:2092-2100.

Rand, D. M., M. Dorfsman, and L. M. Kann. 1994. Neutral and non-neutral evolution of *Drosophila* mitochondrial DNA. Genetics **138**:741-756.

Ray, N., M. Currat, and L. Excoffier. 2003. Intra-deme molecular diversity in spatially expanding populations. Mol.Biol.Evol. **20**:76-86.

Reed, D. L., V. S. Smith, S. L. Hammond, A. R. Rogers, and D. H. Clayton. 2004. Genetic analysis of lice supports direct contact between modern and archaic humans. PLoS Biol **2**:e340.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. Nature **411**:199-204.

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a Paleolithic human population expansion in Africa. Proc.Natl.Acad.Sci.USA **95**:8119-8123.

Relethford, J. H. 1998. Mitochondrial DNA and ancient population growth. Am.J.Phys.Anthropol. **105**:1-7.

Relethford, J. H. 2001a. Global analysis of regional differences in craniometric diversity and population substructure. Hum.Biol. **73**:629-636.

Relethford, J. H. 2001b. Ancient DNA and the origin of modern humans. Proc.Natl.Acad.Sci.USA **98**:390-391.

Rexova, K., Y. Bastin, and D. Frynta. 2006. Cladistic analysis of Bantu languages: A new tree based on combined lexical and grammatical data. Naturwissenschaften **93**:189-194.

Robinson, B. H., N. MacKay, K. Chun, and M. Ling. 1996. Disorders of pyruvate carboxylase and the pyruvate dehydrogenase complex. J. Inherit. Metab. Dis. **19**:452-462.

Rogers, A. R., and H. Harpending. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol.Biol.Evol. **9**:552-569.

Rogers, A. R., and L. B. Jorde. 1995. Genetic evidence on modern human origins. Hum.Biol. **67**:1-36.

Roy-Engel, A. M., M. L. Carroll, E. Vogel, R. K. Garber, S. V. Nguyen, A. H. Salem, M. A. Batzer, and P. L. Deininger. 2001. *Alu* insertion polymorphisms for the study of human genomic diversity. Genetics **159**:279-290.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19**:2496-2497.

Ruiz-Pesini, E., D. Mishmar, M. Brandon, V. Procaccio, and D. C. Wallace. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science **303**:223-226.

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

Santa Luca, A. P. 1978. A re-examination of presumed Neandertal-like fossils. J.Hum.Evol. **7**:619-636.

Schneider, S., and L. Excoffier. 1999. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152**:1079-1089.

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

Senut, B., M. Pickford, D. Gommery, P. Mein, K. Cheboi, and Y. Coppens. 2001. First hominid from the Miocene (Lukeino Formation, Kenya). Comptes Rendus de l'Académie de Sciences **332**:137-144.

Shen, P., F. Wang, P. A. Underhill, C. Franco, W. H. Yang, A. Roxas, R. Sung, A. A. Lin, R. W. Hyman, D. Vollrath, R. W. Davis, L. L. Cavalli-Sforza, and P. J. Oefner. 2000.

Population genetic implications from sequence variation in four Y chromosome genes. Proc.Natl.Acad.Sci.USA **97**:7354-7359.

Sherry, S. T., A. R. Rogers, H. Harpending, H. Soodyall, T. Jenkins, and M. Stoneking. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. Hum. Biol. **66**:761-775.

Shimada, M. K., K. Panchapakesan, S. A. Tishkoff, A. Q. Nato, Jr., and J. Hey. 2007. Divergent haplotypes and human history as revealed in a worldwide survey of X-linked DNA sequence variation. Mol.Biol.Evol. **24**:687-698.

Simpson, G. G. 1944. Tempo and Mode in Evolution. Columbia University Press, New York.

Simpson, G. G. 1961. Some problems of vertebrate paleontology: The study of fossil vertebrates elucidates the general principles of evolutionary biology. Science **133**:1679-1689.

Slatkin, M. 1989. A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43** 1349-1368.

Slatkin, M., and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129** 555-562.

Slatkin, M., and W. P. Maddison. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123**:603-613.

Smith, A. K. 1973. The peoples of southern Mozambique: An historical survey. The Journal of African History **14**:565-580.

Spencer, P. 1980. Polygyny as a Measure of Social Differentiation in Africa *in* J. C. Mitchell, ed. Numerical Techniques in Social Anthropology. ISHI Press, Philidelphia.

Spoor, F., M. G. Leakey, P. N. Gathogo, F. H. Brown, S. C. Antón, I. McDougall, C. Kiarie, F. K. Manthi, and L. N. Leakey. 2007. Implications of new early Homo fossils from Ileret, east of Lake Turkana, Kenya. Nature **448**:688-691.

Spurdle, A. B., M. F. Hammer, and T. Jenkins. 1994. The Y *Alu* polymorphism in southern African populations and its relationship to other Y-specific polymorphisms. Am.J.Hum.Genet. **54**:319-330.

Steiper, M. E., and N. M. Young. 2006. Primate molecular divergence dates. Mol.Phylogenet.Evol. **41**:384-394.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya, S. E. Stanley, R. Jiang, C. J. Messer, A. Chew, J. H. Han, J. Duan, J. L. Carr, M. S. Lee, B. Koshy, A. M. Kumar, G. Zhang, W. R. Newell, A. Windemuth, C. Xu, T. S. Kalbfleisch, S. L. Shaner, K. Arnold, V. Schulz, C. M. Drysdale, K. Nandabalan, R. S. Judson, G. Ruano, and G. F. Vovis. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. Science **293**:489-493.

Stiner, M. C., N. D. Munro, and T. A. Surovell. 2000. The tortoise and the hare: Small game use, the broad-spectrum revolution, and Paleolithic demography. Curr.Anthropol. **21**:39-73.

Stoneking, M., J. J. Fontius, S. L. Clifford, H. Soodyall, S. S. Arcot, N. Saha, T. Jenkins, M. A. Tahir, P. L. Deininger, and M. A. Batzer. 1997. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res **7**:1061-1071.

Strassmann, B. 1992. The function of menstrual taboos among the Dogon: Defense against cuckoldry? Human Nature **3**:89-131.

Strassmann, B., and J. Warner. 1998. Predictors of fecundability and conception waits among the Dogon of Mali. Am.J.Phys.Anthropol. **105**:167-184.

Strassmann, B. I. 2003. Social Monogamy in Human Society. Pp. 177-189 *in* H. Reichard, and C. Boesch, eds. Mating Strategies and Partnerships in Birds, Humans and Other Mammals. Cambridge University Press, Cambridge.

Stringer, C. B. 1987. A numerical cladistic analysis for the genus *Homo*. J.Hum.Evol. **16**:135-146.

Sun, C., Q. P. Kong, and Y. P. Zhang. 2007. The role of climate in human mitochondrial DNA evolution: A reappraisal. Genomics **89**:338-342.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585-595.

Takahata, N. 1993. Allelic genealogy and human evolution. Mol.Biol.Evol. **10**:2-22.

Takahata, N., S. H. Lee, and Y. Satta. 2001. Testing multiregionality of modern human origins. Mol.Biol.Evol. **18**:172-183.

Takahata, N., and Y. Satta. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci U S A **94**:4811-4815.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol.Biol.Evol. **10**:512-526.

Templeton, A. R. 2002. Out of Africa again and again. Nature **416**:45-51.

Templeton, A. R., D. Otte, and J. A. Endler. 1989. The meaning of species and speciation: A genetic perspective. Pp. 3-37. Speciation and Its Consequences. Sinauer, Sunderland, Mass.

Templeton, A. R., E. Routman, and C. A. Phillips. 1995. Separating population structure from population history: A cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. Genetics **140**:767-782.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Research **17**:520-526.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman. 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc.Natl.Acad.Sci.USA **97**:7360-7365.

Tishkoff, S. A., and S. M. Williams. 2002. Genetic analysis of African populations: Human evolution and complex disease. Nat.Rev.Genet. **3**:611-621.

Tobias, P. V. 1978. The Bushmen: San Hunters and Herders from Southern Africa. Human and Rousseau, Cape Town.

Underhill, P. A., G. Passarino, A. A. Lin, P. Shen, L. M. Mirazon, R. A. Foley, P. J. Oefner, and L. L. Cavalli-Sforza. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann.Hum.Genet. **65**:43-62.

Underhill, P. A., P. Shen, A. A. Lin, L. Jin, G. Passarino, W. H. Yang, E. Kauffman, B. Bonne-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J. R. Kidd, S. Q. Mehdi, M. T. Seielstad, R. S. Wells, A. Piazza, R. W. Davis, M. W. Feldman, L. L. Cavalli-Sforza, and P. J. Oefner. 2000. Y chromosome sequence variation and the history of human populations. Nat.Genet. **26**:358-361.

Vansina, J. 2006. Linguistic evidence for the introduction of ironworking into Bantu-speaking Africa. History in Africa **33**:321-361.

Vansina, J. 1984. Western Bantu expansion J.Afr.Hist. **25**:129-145.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. Science **253**:1503-1507.

Voight, B. F., A. M. Adams, L. A. Frisse, Y. Qian, R. R. Hudson, and A. Di Rienzo. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc.Natl.Acad.Sci.USA **102**:18508-18513.

Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution Int.J.Org.Evolution **54**:1092-1101.

Wakeley, J. 1999. Nonequilibrium migration in human history. Genetics **153**:1863-1871.

Walker, A., and R. E. F. Leakey. 1993. The Nariokotome *Homo erectus* skeleton. Harvard University Press, Cambridge.

Wall, J. D. 2000. Detecting ancient admixture in humans using sequence polymorphism data. Genetics **154**:1271-1279.

Wall, J. D., and M. Przeworski. 2000. When did the human population size start increasing? Genetics **155**:1865-1874.

Watkins, W. S., A. R. Rogers, C. T. Ostler, S. Wooding, M. J. Bamshad, A. M. Brassington, M. L. Carroll, S. V. Nguyen, J. A. Walker, B. V. Prasad, P. G. Reddy, P. K. Das, M. A. Batzer, and L. B. Jorde. 2003. Genetic variation among world populations: Inferences from 100 *Alu i*nsertion polymorphisms. Genome Research **13**:1607-1618.

Watson, E., P. Forster, M. Richards, and H. J. Bandelt. 1997. Mitochondrial footprints of human expansions in Africa. Am.J.Hum.Genet. **61**:691-704.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor.Popul.Biol. **7**:256-276.

Weaver, T. D., and C. C. Roseman. 2005. Ancient DNA, late Neandertal survival and modern human-Neandertal genetic admixture. Curr.Anthropol. **46**:677-683.

Weidenreich, F. 1943. The skull of *Sinanthropus pekinensis*: A comparative study on a primitive hominid skull. Paleontol.Sinica **10**:1-484.

White, T. 2003. Paleoanthropology: Early hominids-- diversity or distortion. Science **301** 763-764.

Whitlock, M. C., and D. E. McCauley. 1999. Indirect measures of gene flow and migration: F_{ST} not equal to 1/(4Nm + 1). Heredity **82 (Pt 2)**:117-125.

Wilder, J. A., and M. F. Hammer. 2007a. Extraordinary population structure among the Baining of New Britain. Pp. 199-207 *in* J. S. Friedlaender, ed. Genes, Language, and Culture History in the Southwest Pacific. Oxford University Press, Oxford.

Wilder, J. A., and M. F. Hammer. 2007b. Extraordinary population structure among the Baining of New Britain *in* J. S. Friedlaender, ed. Genes, Language, and Culture History in the Southwest Pacific. Oxford University Press, Oxford.

Wilder, J. A., S. B. Kingan, Z. Mobasher, M. M. Pilkington, and M. F. Hammer. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females *versus* males. Nat.Genet. **36**:1122-1125.

Wilder, J. A., Z. Mobasher, and M. F. Hammer. 2004. Genetic evidence for unequal effective population sizes of human females and males. Mol.Biol.Evol. **21**:2047-2057.

Wilkins, J. F., and F. W. Marlowe. 2006. Sex-biased migration in humans: what should we expect from genetic data? Bioessays **28**:290-300.

Williamson, K., and R. Blench. 2000. Niger-Congo. Pp. 11-42 *in* B. Heine, and D. Nurse, eds. African Languages: An Introduction. Cambridge University press, Cambridge.

Wolpoff, M. H., X. Wu, A. G. Thorne, F. H. Smith, and F. Spencer. 1984. Modern *Homo sapiens* origins: A general theory of hominid evolution involving the fossil evidence from East Asia. Pp. 411-483. The Origins of Modern Humans: A World Survey of the Fossil Evidence. Alan R. Liss, New York.

Wood, B., and M. Collard. 1999. The Human genus. Science **284**:65-71.

Wood, E. T., D. A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini, H. McLeod, L. Louie, M. Bamshad, B. I. Strassmann, H. Soodyall, and M. F. Hammer. 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: Evidence for sex-biased demographic processes. Eur. J. Hum. Genet. **13**:867-876.

Wright, S. 1969. Evolution and the Genetics of Populations, vol 2, The Theory of Gene Frequencies. University of Chicago Press, Chicago

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics **15** 323-354.

Wright, S. 1931. Evolution in mendelian populations. Genetics **16** 97-159.

Xie, X. 2000. Demography: Past, present, and future. Journal of the American Statistical Association **95**:670-673.

Yu, N., Z. Zhao, Y. X. Fu, N. Sambuughin, M. Ramsay, T. Jenkins, E. Leskinen, L. Patthy, L. B. Jorde, T. Kuromori, and W. H. Li. 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol.Biol.Evol. **18**:214-222.

Zhao, Z., N. Yu, Y. X. Fu, and W. H. Li. 2006. Nucleotide variation and haplotype diversity in a 10-kb noncoding region in three continental human populations. Genetics **174**:399-409.

Zietkiewicz, E., V. Yotova, M. Jarnik, M. Korab-Laskowska, K. K. Kidd, D. Modiano, R. Scozzari, M. Stoneking, S. Tishkoff, M. Batzer, and D. Labuda. 1998. Genetic structure of the ancestral population of modern humans. J.Mol.Evol. **47**:146-155.

Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. J.Theor.Biol. **8**:357-366.