Effects of female gene flow and effective population size on

Old and New World mitochondrial DNA patterns

By

KEN BATAI
B.S., Southern Illinois University, Carbondale, 2000
M.A., University of Illinois at Chicago, 2003

THESIS

Submitted as a partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Anthropology
in the Graduate College of the
The University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:

Sloan R. Williams, Chair and Adviser
Elizabeth T. Abrams
Crystal L. Patil
Mary Ashley, Department of Biology
M. Geoffrey Hayes, Northwestern University

This dissertation is dedicated to

My parents, who always supported my education,

My sister and her family, who are fulfilling my responsibilities at home,

My three crazy children, who are always my inspiration

And

My wife, who supported my dream.

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee members, Dr. Betsy Abrams, Dr. Mary Ashley, Dr. Geoffrey Hayes, and Dr. Crystal Patil for reviewing drafts of dissertation chapters and for their support and assistance, and Dr. Sloan Williams, my committee chair, for supporting my dissertation project from its infancy, for developing research design, guidance in laboratory methods, and reviewing drafts of the dissertation.

I also would like to thank people of the Aymra, Mijikenda, and Taita, who provided the samples. Special thanks to Dr. Virginia Vitzthum, for collecting the Aymara samples, Dr Kara Babrowski, for collecting the Taita and Mijikenda samples and producing initial mtDNA sequence data, Dr. Chapurukha Kusimba, the Field Museum, Chicago for facilitating the fieldwork and providing logistical support in Kenya for sample collection.

This work was supported by University of Illinois at Chicago Department of Anthropology Charles Reed Fund, Sigma Xi Grants In Aid of Research and UIC's Provost Research Award. Special thanks to the Pritzker Lab at the Field Museum, Chicago for providing sequencing facilities for this research.

KB

TABLE OF CONTENTS

CHAPTER PAGE

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF ABBREVIATIONS

AMOVA	Analysis of Molecular Variance
bp	base-pair
HPG	Haplogroup
HVRI	Hypervariable Region I
MDS	Multidimensional Scaling
mtDNA	mitochondrial DNA
np	nucleotide position
PC	Principal-Components
P-Value	Probability Value
RFLP	Restriction Fragment Length Polymorphism
S.D.	Standard Deviation
SSD	Sum of Square Deviation
STR	Short Tandem Repeat or microsatellite

SUMMARY

The goal of this dissertation project is to examine how much the impact female gene flow and female effective population size had on mtDNA variation. I examined 1) whether the Aymara and Bantu speakers expanded through spatial expansion by incorporating female migrants from other ethnic groups or demographic expansion due to increased female fertility rate, 2) whether female gene flow or effective population size had a bigger effect on mtDNA within-population genetic diversity than the other, and 3) the extent to which kinship structure affected the importance of each factor.

The results of the analyses suggest that the Aymara experienced population expansions, most likely because of rapid demographic expansion. Female gene flow was also an important factor influencing mtDNA variation among the Central Andeans as well as western South American populations, but female gene flow had a much greater effect on Bantu mtDNA variation. East African Bantu speakers interacted with non-Bantu east Africans. As a result, east African Bantu populations became genetically diverse and similar to non-Bantu east Africans. More close examination of the impact of female gene flow and effective population size reveals that female gene flow affect two within-population genetic diversity measurements and the genetic model that does not account for population subdivision and gene flow are poor fit. Finally, kinship structure is an important cultural practice that affects patterns and intensity of female gene flow and a good predictor of within-population genetic diversity and population subdivision.

1. INTRODUCTION

1.1 Roles of female gene flow in human evolutionary history

Mitochondrial DNA (mtDNA) genetic diversity data is often used in anthropological genetic projects to infer the demographic history of our species as well as that of individual populations. Demographic events in the past, such as migration (or gene flow) and changes in population size through demographic expansion with increased fertility, can be inferred from genetic analysis. MtDNA is inherited maternally, so it informs us of female evolutionary and demographic history. While anthropologists have used mtDNA analyses for a long time to understand demographic history, the effects of gene flow on mtDNA variation deserve further investigation. The role of cultural factors such as introduction of food producing technology, kinship or social structure, language, and religion of the society also affect genetic variation in important ways that merit careful consideration.

In this dissertation project, in order to examine how much effect female gene flow had on mtDNA variation, I analyzed the mtDNA patterns of the Aymara from Bolivia and the Taita and Mijikenda from Kenya and, compared them to the mtDNA variation in other Latin American and Bantu-speaking populations respectively. First, I asked whether the Aymara and Bantu expansions fit models of spatial expansions which incorporate female migrants from other ethnic groups or demographic expansion which are driven by high female fertility rates. Second, I used these two regions to examine and compare the effects of female gene flow and effective population size on mtDNA within-population genetic diversity. I also compared mtDNA patterns in patrilocal Andean highlanders with the mtDNA variation in Latin American matrilocal populations and compared patrilocal and matrilocal Bantu population genetic diversity.

1.2 Research questions

Among human populations, genetic diversity is often correlated with subsistence strategy. Many large agricultural populations are genetically diverse with evidence of demographic expansion (Excoffier and Schneider 1999; Rogers 1995; Watson et al. 1996). Increased female fertility rates lead to population growth and reduce the effects of genetic drift. Foragers, whose population size has stayed consistently small, are genetically homogeneous with no evidence of demographic expansion. Ray et al. (2003) and Excoffier (2004) note that smaller populations with increased gene flow can be genetically very diverse, however, and they argue that spatial expansions, population migrations which incorporate pre-existing local populations, can explain human demographic history better than the demographic expansion model. Their work is based on theoretical computer simulations, and the effect of spatial expansion on within-population genetic diversity is not fully examined with empirical data.

Among the Latin Americans, the highland Andeans exhibit much higher within-population genetic diversity than lowland populations. Researchers have proposed that they experienced a demographic expansion as a result of the introduction of intensive agriculture and/or a spatial expansion through increased female gene flow (Fuselli et al., 2003; Lewis et al., 2005, 2007b). Similarly, Bantu-speaking populations expanded from central Africa to east Africa through a massive demographic expansion replacing pre-existing populations (Cavalli-Sforza et al. 1994; Excoffier et al. 1987; Salas et al. 2002) and/or interacting with non-Bantu east African populations (Castrì et al. 2008; Castrì et al. 2009).

More recently, a number of studies have used sex-specific genetic markers (mtDNA and Y chromosomes) to analyze demographic history of males and females separately and have found asymmetrical genetic signatures of male and female demographic history (Chaix et al. 2007; Destro-Bisol et al. 2004; Marchani et al. 2008; Nasidze et al. 2004; Nasidze et al. 2005; Pérez-Lezaun et al. 1999; Salem et al. 1996). Comparative studies consistently show higher within-population genetic diversity and among-population homogeneity in mtDNA when compared with Y chromosome variation. In addition, the Y chromosome has a shorter coalescent time when compared to mtDNA (Wilder et al. 2004a). Reduced male effective population size (Hammer et al. 2008; Wilder et al. 2004a; Wilder et al. 2004b) or increased female gene flow (Seielstad et al. 1998) (or a combination these factors) has been suggested as explanations for this asymmetry. These researchers also argue that kinship structure, or mating pattern, was an important factor influencing these variables.

Effective population size is the long-term average of the number of individuals who contributed the genes to the next generation and is often indicative of demographic history. Effective population size is often correlated with census population size, but it is usually much smaller than census population size because not all individuals successfully contribute genes to the next generation. Effective population size can be inferred from the observed genetic diversity, but a number of factors affect this relationship.

One of the factors that affect this relationship is mating pattern. Differential reproductive success due to polygamous or monogamous marriage affects effective population size (Nunney 1993). Dupanloup et al. (2003), Wilder et al. (2004a; 2004b), and Pilkington et al. (2007) have argued that polygyny can dramatically reduce male effective population size by concentrating wives among a smaller pool of males and explains the observed smaller Y chromosome diversity and shorter coalescent time.

Alternatively, Seielstad et al. (1998) believe that higher rates of female gene flow contribute to the discrepancy between mtDNA and Y chromosome genetic diversity. Men are more likely to stay in their natal villages and women are more likely to move to their husbands' villages in populations with patrilineal descent systems and patrilocal post-marital residence patterns. These cultural practices lead to higher rates of female gene flow, so the mtDNA genetic diversity is higher than the Y chromosome diversity (Chaix et al. 2007; Hamilton et al. 2005; Oota et al. 2001). In societies with matrilineal descent systems and matrilocal post-marital residence patterns, women tend to stay in their natal communities, while the men move out, so these societies will have low rates of female gene flow and high rates of male gene flow that are reflected in the mtDNA and Y chromosome genetic diversities.

Three groups of researchers investigated the differences in mtDNA and Y chromosome genetic variation between patrilocal and matrilocal populations in northern Thailand (Besaggio et al. 2007; Hamilton et al. 2005; Oota et al. 2001). They used the same population sets to compare mtDNA and Y chromosome variation. The researchers found that mtDNA within-population genetic diversity was higher in patrilocal than in matrilocal populations and Y chromosome within-population diversity was higher in matrilocal than in patrilocal populations. Matrilocal populations were more differentiated when mtDNA variation was examined, while Y chromosome among-population differentiation was greater in patrilocal populations. They concluded that female gene flow was the most important factor influencing the observed pattern.

Kinship structure had no significant effect on two sex-specific markers in a similar study conducted in India, however, where ethnic or tribal endogamy is more strictly followed (Kumar et al. 2006). While these studies focus on the regional level to understand the effects of cultural and social processes, the number of populations analyzed was limited. Furthermore, the effect of female gene flow relative to effective populations on within-population genetic diversity has not been widely examined, yet. Finally, low mtDNA diversity is more difficult to explain in matrilocal populations, such as some sub-Saharan Bantu populations (Salas et al. 2002) and the Central American Chibchans (Batista et al. 1995; Kolman et al. 1995; Melton et al. 2007) who are known to have expanded in the recent past.

1.3 Objectives of the project

The main objective of this dissertation is to explore the roles of female gene flow on mtDNA variation in human populations. **Hypothesis**: Because many anthropological studies suggest that ethnic boundaries are open and inter-ethnic marriages are common (Barth 1969; Green and Perlman 1985; Moore 1994), I hypothesized that female gene flow was the major contributing factor affecting mtDNA variation. In order to test the hypothesis, I analyzed mtDNA variation in three populations that experienced population expansion 2,000-3,000 years ago, the Aymara of Bolivia and the Taita and Mijikenda Bantu-speaking ethnic groups of Kenya. Then, I compared them with other Latin American and Bantu populations respectively. I asked A) whether the Aymara and Bantu speakers expanded through range (or spatial) expansion by incorporating female migrants from other ethnic groups or through demographic expansion due to increased female fertility rates, B) whether female gene flow or effective population size had a greater effect on mtDNA within-population genetic diversity, and C) the extent to which kinship structure (patrilineal/patrilocal or matrilineal/matrilocal) affected the importance of each factor.

I focused on mtDNA variation for three reasons. First, comparisons of mtDNA and Y chromosome diversity can be confounded by many factors such as differences in mutation rates (Stoneking 1998). Second, using a matched population set for mtDNA and Y chromosome data severely reduces the numbers of populations available for comparison. By focusing on mtDNA, I could maximize the distribution and number of the population samples included in my analysis. Finally, researchers have already proposed that female gene flow was an important factor affecting the male-female asymmetrical genetic pattern in global and continental level (Seielstad et al. 1998), and noted that female gene flow is greater among patrilocal than matrilocal populations in regional studies (Besaggio et al. 2007; Hamilton et al. 2005; Oota et al. 2001). The main goal of this dissertation is to evaluate the roles of female gene flow during the population expansions and the difference in female gene flow pattern between matrilocal and patrilocal populations affecting within-population genetic diversity and pattern of population subdivision.

1.4 **Study Populations**

I used the Aymara of Bolivia and the Taita and Mijikenda of Kenya as case studies for several reasons. First, unlike other parts of the world, many Latin American and Bantu societies have matrilineal kinship systems (Burton et al. 1996). Matrilineal kinship systems are common among horticulturalists societies in tropical forests, and Latin American and Bantu societies tend to have more of these kinds of horticulturalists (Martin and Voorhies 1975). Patrilocal societies in the two areas differ, however; Bantu commonly practice polygyny, while polygyny is rare in Latin America. Second, the Latin American and Bantu populations make interesting data sets for comparison because their population histories are quite different. Africa has a long evolutionary history. The earliest anatomically modern *Homo Sapiens* remains have been found in east Africa (Clark et al. 2003; McDougall et al. 2005; White et al. 2003), and African populations have mtDNA diversity greater than that of populations in other parts of the world (Ingman et al. 2000; Jorde et al. 2000; Vigilant et al. 1991). The New World has a shorter history of human occupation. The Paleo-Indians entered the New World by 15,000 years ago and quickly travelled to southern tip of South America (Dillehay 2000). Consequently, New World populations exhibit small genetic diversity (Excoffier and Schneider 1999). While the magnitude of genetic diversity differs significantly between Latin American and Bantu populations, the differences in female gene flow and effective population size between patrilocal and matrilocal populations should have similar impacts on genetic variation in both areas. Using test cases from both areas provides an assessment of the general applicability of my conclusions.

Despite the differences in length of occupation, there are some interesting similarities in two areas. Climate change in the end of the Pleistocene and the beginning of the Holocene caused a cultural transition in Africa and the New World (Dillehay 2000; Moseley 2001; Phillipson 2005). Around 11,500 years ago, people in the world become less mobile. Archaeological data shows that material cultures become more diverse and heterogeneous, as people culturally adapted to diverse local microenvironments and begin to exploit more local resources. Also, people exploit various local resources, they become more knowledgeable about the plants and animals in the ecosystem, which lead to domestication of plants and animals in both areas of the world.

The Aymara and Bantu speakers also have similar recent demographic histories. Linguists and archaeologists (Browman 1994; Holden 2002; Phillipson 2005) believe that they were both small ethnic groups until approximately three thousand years ago when they experienced rapid population expansions, but the underlying mechanisms of expansion are not understood. The Aymara are traditionally pastoral-agriculturalists and live on the south-central Andean plateau, an important area in the development of prehistoric complex societies (Kolata 1993). The current Aymara population size is large (over 2 millions) and the Aymara are the second largest linguistic group in the Andes. The Aymara language is widely distributed geographically in Bolivia, northern Chile, and southern Peru, and majority of indigenous people who live in La Paz, Bolivia today are Aymara.

The ethnohistoric documents indicate that it was only one of several languages spoken in the area in prehistory. At the time of Inca conquest, numerous small ethnic groups lived across the Central Andes (Rostworowski de Diez Canseco 1999; Rowe 1946). The languages spoken are very diverse and each ethnic group had their own dialect or language. The Aymara coexisted with the Quechua, Uru, and Pukina in the Titicaca Basin (Browman 1994; Murra 1968). By the time of European contact, the Aymara were a large language group that occupied a wide area in the south-central Andes, but exactly when in the prehistory the Aymara expanded is debated (Browman 1994; Murra 1968).

One notable Aymara cultural practice is their vertical use of Andean ecosystem, vertical archipelago. The model of vertical archipelago was first presented by an ethnohistorican, John Murra (1968; 1985a), using historical documents on the Aymara. According to Murra, the Lupaqa, colonial period Aymara Kingdom, occupied from the Titicaca Basin to the colonies in the Pacific coast from Arica to Moquegua, controlling different resources from different ecological zones. This was not seasonal migrations. Instead, they established permanent colonies. They maintained their ethnic identity through reciprocal exchange, kinship, and ceremonies, and multi-ethnic groups coexisted without competition in the lower altitude areas of the Andean slope. The highland Tiwanaku state also used this cultural strategy and established a colony in the mid-valley region of Moquegua (Blom et al. 1998; Goldstein 1993).

The Aymara from La Paz were chosen because researchers (Batai and Williams 2007; Bert et al. 2001; Fuselli et al. 2003; Merriwether et al. 1995) have noted that the populations who live in the region encompassing southern Peru, lowland Bolivia, and northern Chile have very high frequencies of haplogroup (HPG) B and speculated about the genetic homogeneity. Other Latin American populations with intensive agriculture, such as Quechua from highland Peru and Quiche Mayan from Guatemala, are genetically very diverse with evidence of population expansion due to a combination of increased female gene flow and demographic expansion after the introduction of intensive agriculture (Fuselli et al. 2003; Lewis et al. 2005). The low HPG diversity in the area suggests that the Aymara could have experienced founder effects or bottlenecks in the past. However, mainly HPG variation in populations at the periphery of the Aymara territory has been analyzed, and no mtDNA sequence variation of Aymara living in the core area of Aymara expansion was not analyzed until recently (Barbieri et al. 2011; Gayà-Vidal et al. 2011).

The Bantu languages have an even wider distribution from central Africa to eastern and southern Africa and Bantu languages are spoken dominantly in many of these areas. Linguists, historians and anthropologists believe that the Bantu languages originated in northwestern central Africa and spread over such large areas because of massive migrations from central Africa (Ehret 2001; Holden 2002; Phillipson 2005).

Recent phylogenetic studies of Bantu languages support Guthrie’s view of the Bantu expansion (Holden 2002; Rexová et al. 2006). The Bantu languages in the root of the Bantu language trees are located in northwestern Central Africa in Cameroon. The East Bantu languages, distributed from East Africa and southeastern Africa, form a distinct sub-clade that are separated from Central African Bantu languages that spread in the initial expansion and West Bantu languages distributed in southwestern Africa.

Based on branching orders of Bantu languages and the similarities in the geographical distribution of the Bantu languages and archeological cultures, Holden (2002) believe that the Bantu languages spread with Neolithic culture in Central African and with Early Iron Age cultures from East Africa to southeastern Africa. For example, in East and South Africa, distribution of East Bantu languages overlaps with distribution of Chifumbaze Iron Age cultural complex (Phillipson 2005). Chifumbaze ceramic traditions are derived from Urewe ware that first appeared around Lake Victoria in East Africa. The people associated with the Chifumbaze complex culture were agropastoralists who used iron technologies that were absent in Late Stone Age forager cultures and replaced Late Stone Age cultures. However, because of poor preservation and lack of archaeological research in Africa, archaeologists have been heavily depending on linguistic evidence for reconstruction of the Bantu prehistory. Archaeologists and linguists tend to seek the supports of their work from each others' work without critical evaluation of their work (Eggert 2005).

The Taita and Mijikenda ethnic groups were chosen to provide a better representation of east African population samples for the analysis of Bantu mtDNA variation and demographic history. The Taita and Mijikenda are Bantu speaking agropastoralists who occupy in the Bantu expansion periphery in southeastern Kenya, where many non-Bantu speakers (Afro-Asiatic, Nilo-Saharan, and Khoisan) live as well. Unlike other ethnic groups in the east Africa, the Taita and Mijikenda are small pastoral-agricultural societies. The Taita and Mijikenda have estimated population sizes of 213,000 and 1,208,000 respectively, while major non-Bantu speaking ethnic groups, such as the Oromo, Amhara, and Somali, have estimated population sizes over 10 million.

The Taita and Mijikenda share a common origin oral history with other Bantu speaking ethnic groups in the area (Spear 1981; Spear 1974; Spear 1977). According their oral history, linguistically closely related Bantu speakers, including Taita, Mijikenda, Pokomo, and Swahili, left their mythological ancestral land, Singwaya, located in somewhere northeastern Kenya or Southern Somalia around the 16th century, and settled in the coastal region of southeastern Kenya and northeastern Tanzania. However, the Taita and Mijikenda are ethnically heterogeneous groups composed of formerly distinct groups and people of different ethnic origins, who reorganized to form new ethnic identities after European contact (Bravman 1998; Spear 1974; Spear 1977; Willis 1993).

Archaeologists (Phillipson 2005), linguists (Holden 2002), and human population geneticists (Salas et al. 2002) have noted the importance of east Africa in explaining Bantu history. They suggest that Bantu speakers migrated to east Africa from central Africa, and then expanded into southern Africa. Despite the importance of the area, mtDNA variation of only four other Bantu populations from east Africa has been reported until recently (Castrì et al. 2008; Knight et al. 2003; Tishkoff et al. 2007; Watson et al. 1996).

1.5 Overview of dissertation and contributions to the field of anthropological genetics

This dissertation consists of four research projects that address the roles of female gene flow and effective population size in human demographic history. In Chapter 2, I briefly explain the laboratory methods and mtDNA variation exists in sub-Saharan Africa and Latin America, and a discussion of the concept of population as used in human genetic studies. Then, I describe the statistical and analytical methods used in this dissertation focusing on the built-in assumptions, interpretation of data, and limitation of the methods.

In the next two chapters, mtDNA variation in Latin American and Bantu populations are examined to consider the effects of female gene flow and effective population size on regional mtDNA variation and past population expansions. In Chapter 3, I examine mtDNA genetic variation of the Aymara and compare it to that of other Central Andean and Latin American populations in order to understand the Aymara expansion. Based on genetic analysis of mainly Quechua populations, researchers have proposed that central Andeans exhibit genetic evidence of expansion because of either a demographic expansion as a result of introduction of intensive agriculture or a spatial expansion through increased female gene flow (Fuselli et al., 2003; Lewis et al., 2005, 2007b). Others propose that there were more interactions in the western South America compared to eastern South America (Cabana et al. 2006; Tarazona-Santos et al. 2001). Recent studies of Aymra mtDNA variation found distinctive mtDNA variation (Barbieri et al. 2011; Gayà-Vidal et al. 2011), but these studies did not consider whether the Aymara expanded differently from the Quechua. **Sub-hypotheses 1**: in chapter 3, I hypothesize that the spatial expansion had greater effects on mtDNA among the Aymara and other Andean populations. I describe the genetic characteristics of the Aymara expansion and evaluate the level of gene flow in western South America. This project contributes to overall understanding of evolutionary history of the Andes and South America by increasing our understanding of the demographic history of the second largest language group in the area.

In Chapter 4, I examine three models of Bantu expansion that describe possible interactions between the Bantu and non-Bantu groups living in east Africa. Many researchers initially proposed that the Bantu speakers experienced a massive demographic expansion

(Cavalli-Sforza et al. 1994; Excoffier et al. 1987; Salas et al. 2002), but more recently, other researchers have recognized the importance of female gene flow among east African Bantu-speaking populations interacting with non-Bantu east African populations (Castrì et al. 2008; Castrì et al. 2009). Unfortunately, only a few east African Bantu populations have been analyzed and the nature and the extent of interactions between the Bantu and non-Bantu populations in east Africa are poorly understood. **Sub-hypothesis 2**: Contrary to the traditional view of Bantu expansion, the Bantu-speakers experienced spatial expansion and had gene flow with non-Bantu speaking east African populations through various interactions. I address whether the Taita and Mijikenda maintained central African mtDNA characteristics through selective interaction with other Bantu populations or acquired new East African genetic characteristics through gene flow with non-Bantu East African populations. I also compare Taita and Mijikenda within-population genetic diversity to central African Bantu and non-Bantu east African populations. This project contributes to the understanding of the Bantu expansion by assessing the process of expansion as viewed from its northeastern periphery. East Africa was an important area of Bantus expansion. From east Africa, East Bantu language groups moved into southeastern Africa after acquiring new technologies (Phillipson 2005; Salas et al. 2002). East African Bantu populations, however, have only recently begun to attract the attention of human geneticists.

Chapter 5 evaluates the impact of female gene flow and effective population size on mtDNA within-population genetic diversity in both regions. When populations are interacting with large migration rates, they exhibit a genetic pattern that resembles populations that experienced pure demographic expansion (Excoffier, 2004; Ray et al. 2003). Many projects (Kivisild et al. 2004; Richards et al. 2000; Seielstad et al. 1998) focus on pattern of population differentiation or trace the evidence of gene flow and migration, but the impact of gene flow on within-population genetic diversity is not explored well. **Sub-hypothesis 3**: I hypothesize that female gene flow had a greater impact on mtDNA within-population genetic diversity than female effective population size. I investigate how female gene flow and effective population size affect three measurements of mtDNA within-population genetic diversity using a coalescent based computer simulation. The impact of female gene flow and effective population size on regional mtDNA genetic pattern is further analyzed with by comparing results from methods that assume that gene flow took place between subdivided populations to methods that assume that individuals are randomly mating in unsubdivided populations. This project contributes to the understanding of the human evolutionary history by evaluating the link between sex-biased demographic history and female gene flow and effective population size.

Finally, chapter 6 examines the effects of kinship structure on mtDNA variation in Latin American and Bantu populations. The role of kinship structure in influencing genetic variation has been discussed by some researchers (Besaggio et al. 2007; Hamilton et al. 2005; Oota et al. 2001; Seielstad et al. 1998), but the number of sampled populations used in these studies is very small or the cultural variability of sampled populations is not well considered, limiting our understanding of how kinship structures affect pattern of female gene flow and mtDNA variation in more general level. **Sub-hypothesis 4**: Kinship affects the pattern of female gene flow, so it greatly influences mtDNA variation of the populations. In this project, I increased the number of sampled populations by focusing on mtDNA to ensure statistically more robust analyses. The populations included in the study are sorted into matricentric, flexible, and patricentric categories using a simplified version of a system proposed by Burton et al. (1996) (details described in chapter 6). Then, I examine how kinship structure influences the pattern of female gene flow, within-population genetic diversity, and population subdivision. I also investigate whether kinship structure is a better predictor of mtDNA variation than current ethnic population size or subsistence strategy. This project contributes to anthropological knowledge by better elucidating the roles that cultural choices play in genetic variation. Kinship structure is an important cultural variable that affects sex-specific genetic variation, but exactly how kinship structure affects genetic patterning is still debated.

In the conclusion, I indirectly assess the cause of sex biased demographic history based on analyses of mtDNA variation in each chapter of the dissertation. Focusing on mtDNA variation, this dissertation investigates the role of female gene flow and cultural practices, such as post-marital residence pattern and polygyny, on mtDNA within-population genetic diversity and population subdivision. Female gene flow affects within-population genetic diversity and population subdivision, and kinship structure is an important factor affecting the pattern and intensity of female gene flow. Female effective population size was also an important factor affecting mtDNA variation, especially for highland Andeans with intensive agricultural technologies. Throughout the dissertation, I show that female gene flow was an important factor affecting mtDNA variation in Latin American and Bantu populations and that the role of gene flow influencing genetic variation deserves further investigation for the study of human evolutionary history.

2. METHODS FOR THE ANALYSES OF HUMAN MITOCHONDRIAL DNA

2.1 Introduction

In this dissertation, I address research questions that rely heavily on population genetic analytical methods and their models, so it is critical to address underlying assumptions and limitations of the methods. In this chapter, first, I will provide a brief description of human mtDNA variation and laboratory methods to analyze the mtDNA sequence variation. I will also define a study unit for human population genetic research using the biological concept of population. Finally, I will explain the statistical and analytical methods used in this dissertation, discuss the meaning of statistical values and describe the assumptions and limitations of these methods.

2.2 Human mtDNA variation

The hypervariable region I (HVRI) of the mitochondrial genome currently is the most often used genetic marker in anthropological genetic studies of demographic history and population relationships. The HVRI is one of two highly polymorphic short DNA segments in the control region. This area that does not encode for any gene, instead it serves as the site of replication initiation (Anderson et al. 1981). Based on restriction fragment length polymorphism (RFLP) analysis of human mtDNA, Cann and colleagues (1987), first, proposed that anatomically modern human originated in Africa. Later, HVRI sequences were used to support Cann and colleagues' assertion (Vigilant et al. 1991). More recently, these findings were confirmed with sequence analyses of whole mitochondrial genome sequences (Ingman et al. 2000).

Sub-Saharan African mtDNA sequences can be categorized into six macro-haplogroups (L0, L1, L2, L3, L4, and L5) defined by HVRI sequence motifs and/or RFLP patterns (Kivisild et al. 2004; Salas et al. 2002). Each mtDNA haplogroup's (HPG) place of origin has been inferred on the basis on HPG frequency. For example, HPGs L0a and L0f, which are found near the root of human mtDNA phylogenetic trees, are more common in East Africa, while Central African populations have the highest frequency of HPG L1c. All mtDNA HPGs found outside of sub-Saharan Africa are sub-groups of African HPG L3.

Similarly, Latin American mtDNA sequences be categorized into four HPGs (A, B, C, and D) based on HVRI sequence motifs, RFLPs, and/or a 9 base-pair deletion in the case of HPG B (Torroni et al. 1993a; Torroni et al. 1992; Torroni et al. 1993b). The HPG X is found in North America (Brown et al. 1998; Malhi and Smith 2002), but it has not yet been reported in Latin America.

2.3 **Laboratory methods**

In order to analyze the mtDNA HVRI sequence variation, first, DNA was extracted. Puregene® DNA purification kit was used for extraction from blood spot samples from the Aymara samples following the recommended protocol and MasterAmp™ DNA Extraction Solution (Epicentre Technologies, Madison, WI) was used for extraction from swab samples collected in Kenya. Second, HVRI was amplified using a touchdown PCR protocol and either of the primer sets (L15985 5’-GCACCCAAAGCTAAGATTCTAA-3’ and H404 5’-AAAGTGCATACCGCCAAAAG-3’or L15926 5’-

TCAAAGCTTACACCAGTCTTGTAAACC-3’ and H16498 5’-CCTGAAGTAGGAACCAGATG-3’). The amplified product was sequenced in both directions using the BigDye Terminator Cycle Sequencing Kit, version 3.1 (Perkin Elmer Biosystems) and analyzed on an ABI 3730 DNA sequencer housed at the Field Museum’s Pritzker Molecular Biology Laboratory.

Sequences were edited and aligned in Sequencher 4.1.4 (GeneCodes). The nucleotide positions (nps) between 16024 and 16383 relative to the Cambridge Reference Sequence (Anderson et al. 1981) were edited, but subsequent analyses for Latin American populations are confined to nps 16056-16383 to facilitate comparisons with previous work. When T→C transition at np 16189 was present, nps 16182 and 16183 were excluded from analysis because of heteroplasmy (Bendall and Sykes 1995; Pfeiffer et al. 1999). The T→C transition at np16189 creates a poly-C tract, a long stretch of C between np 16182-16193, and the length of the poly-C tract can vary within a single individual.

2.4 Population as a study unit (population ≈ deme ≈ ethnolinguistic group)

In this dissertation, I am following the human population genetics conventional method to define a study unit. The term ‘population’ is used as an operational term to define a study unit and population is defined using ethnolinguistic groups, because comparative population samples available were defined using ethnolinguistic grouping by human population. In a population genetics model, an idealized population is assumed to be panmictic (randomly mating within a population) (Cavalli-Sforza and Bodmer 1971), but anthropologists have realized that traditional human societies, or ethnolinguistic groups, are internally subdivided and are culturally and linguistically heterogeneous due to many cultural factors, such as inter-ethnic marriage, subsistence pattern, religion, and ethnic rebelling by colonial government, yet the subdivided segments of societies interact (Barth 1969; Eriksen 1993; Errington 2001; Fried 1968). Many human population geneticists and anthropological geneticists understand the heterogeneous and permeable nature of ethnic group, and the problem associated with applying the biological concept of ‘population’ into human societies (Cavalli-Sforza et al. 1994; Mielke and Fix 2007). During the development stage of the recent large international collaborative project, International HapMap Project, sampling strategy, participant inclusion criteria, and ethical issues were carefully considered. In the HapMap project, population is defined as “a group of people with a shared ancestry and therefore a shared history and pattern of geographic migration” (The International HapMap Consortium 2004). By recognizing that many individuals in a population may have multiple group identities or memberships, or that they will not share a recent common ancestor, they acknowledged the statistical and ethnical difficulties of working with human populations with great within-population variability.

Population geneticists use another concept, deme, a localized population or subdivided segments of a population, in human population genetic studies (Cavalli-Sforza et al. 1994; Ray et al. 2003). Within a deme, mating is random and individuals share a distinct gene pool. The demes are reproductively isolated to some extent, but mates are exchanged between different demes and can be incorporated into a broader population genetic model. Figure 1 shows gene flow taking place among neighboring demes. Incorporating mate exchanges between demes into the model allows us to measure the extent of gene flow and evaluate the impacts of gene flow on genetic diversity.

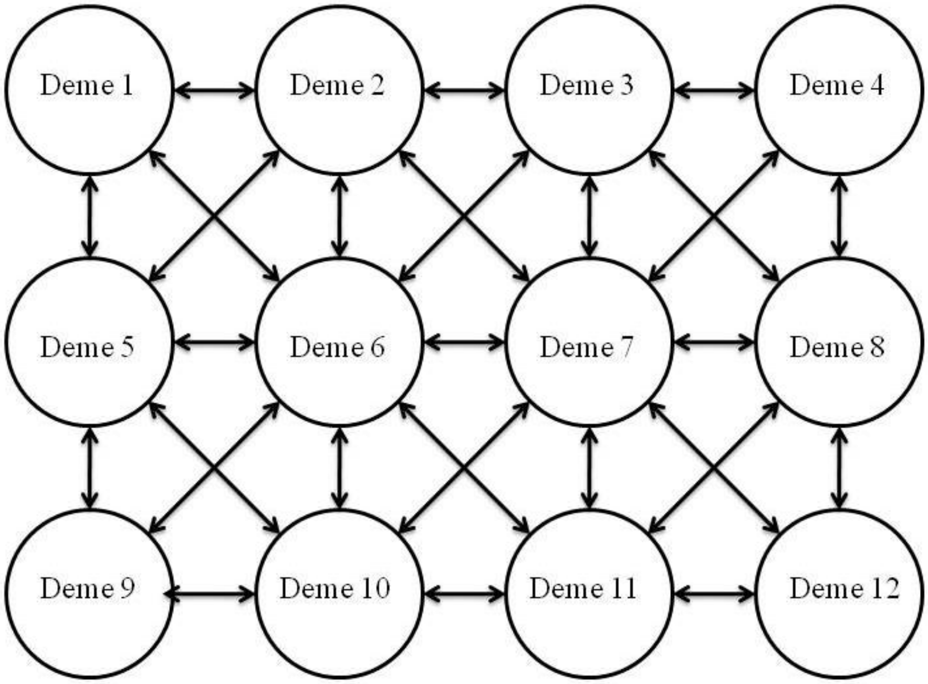

Figure 1. Demes and island model of migration. Each circle represents deme and arrow represents migration. Here equal deme size and migration rate are assumed

Although it is difficult to define a study unit using a concept of deme, many human population geneticists use ethnolinguistic groups, because the ethnolinguistic grouping has some similarities to the concept of deme (Cavalli-Sforza et al. 1994). Ethnolinguistic groups are to some extent endogamous, functioning as reproductive units. Mating between demes may be restricted or encouraged depending on the marriage practices (exogamy or endogamy) of the culture.

In this dissertation, the terms, deme and population, are used interchangeably. When the term ‘population’ is used, the concept of deme is employed. For example, the Mijikenda is an ethnolinguistic group. The Mijikenda can be further subdivided into nine local populations that share a common language, similar cultural traits, and an origin myth (Spear 1974). Because my research focuses on interactions between ethnolinguistic groups, I use the Mijikenda as a study unit, a population interacting with neighboring ethnolinguistic groups, while assuming that subdivision within ethnolinguistic groups, in this case, the Mijikenda, have only minor effects on their genetic variation for the purpose of this research project.

2.5 Assumptions and limitations of analytical methods

Many population genetic analytical methods assume that populations are panmictic, but recent analytic methods have begun to incorporate population subdivisions. Now, many methods have been developed to analyze patterns of population subdivision and gene flow between demes. These newly developed methods also allow estimate within-population genetic diversity adjusting for effects of gene flow. The development of computational technology has also played an important role in the field of population genetics allowing more computationally demanding analyses, such as maximum likelihood estimation and computer simulation.

2.5.1 Analysis of within-population genetic diversity

To reconstruct the demographic history of the Aymara, Taita, Mijikenda, and other populations, effective population size was inferred using within-population genetic diversity measurements. The genetic diversity parameter $\theta=2N_f\mu$, where N_f is the effective female population size (effective population size of mtDNA is a quarter of autosomal DNA, $2N_{f+m}$) and μ is the mutation rate, was estimated by three different values: θ_k, θ_S, and θ_π. The value of θ_k is based on the relationship between the number of sequences (k) and the sample size (Ewens 1972). The value of θ_S is based on the relationship between the number of polymorphic, or segregating sites (S), and the sample size (Watterson 1975). The θ_π is a measure of mean pairwise differences, the mean number of mutational differences between two sequences (π) (Tajima 1983).

Current population size affects the number of sequences (k) and mutations (S) more than it affects the average number of mutational differences (π), and the average number of mutational differences is more influenced by ancient population size or the original population size before any recent changes in population size (Fig. 2) (Aris-Brosou and Excoffier 1996; Rogers and Harpending 1992; Tajima 1989a). When population size increases, the genetic variation is less likely to be reduced, so mutations are accumulated quickly, creating new sequences that differ from other sequences only by a few mutational changes, and the values of θ_k and θ_S increase quickly without affecting $\theta_{\pi.}$ Longer time spans are needed to increase the value of θ_π because more mutations have to be accumulated on each sequence to increase the average number of mutational differences between two sequences. When population size declines quickly, rare sequences tend to be lost first which results in reduced θ_k and θ_S with less initial effect on θ_π.

Since mtDNA mutation rates should be stable across populations, differences in the θ values reflect the differences in female effective population size (N_f), the harmonic mean of the number of females involved in reproduction that transmit mtDNA to the next generation. Therefore, the θ_π value reflects population size in the distant past before recent demographic events, while the θ_k and θ_S values reflect population size in the recent past. For example, Helgason and colleagues (2003; 2000) calculated the genetic diversity measures for Icelanders and compared them with other European populations to infer that Icelanders experienced a population bottleneck recently. They observed large θ_π and small θ_k and θ_S values among

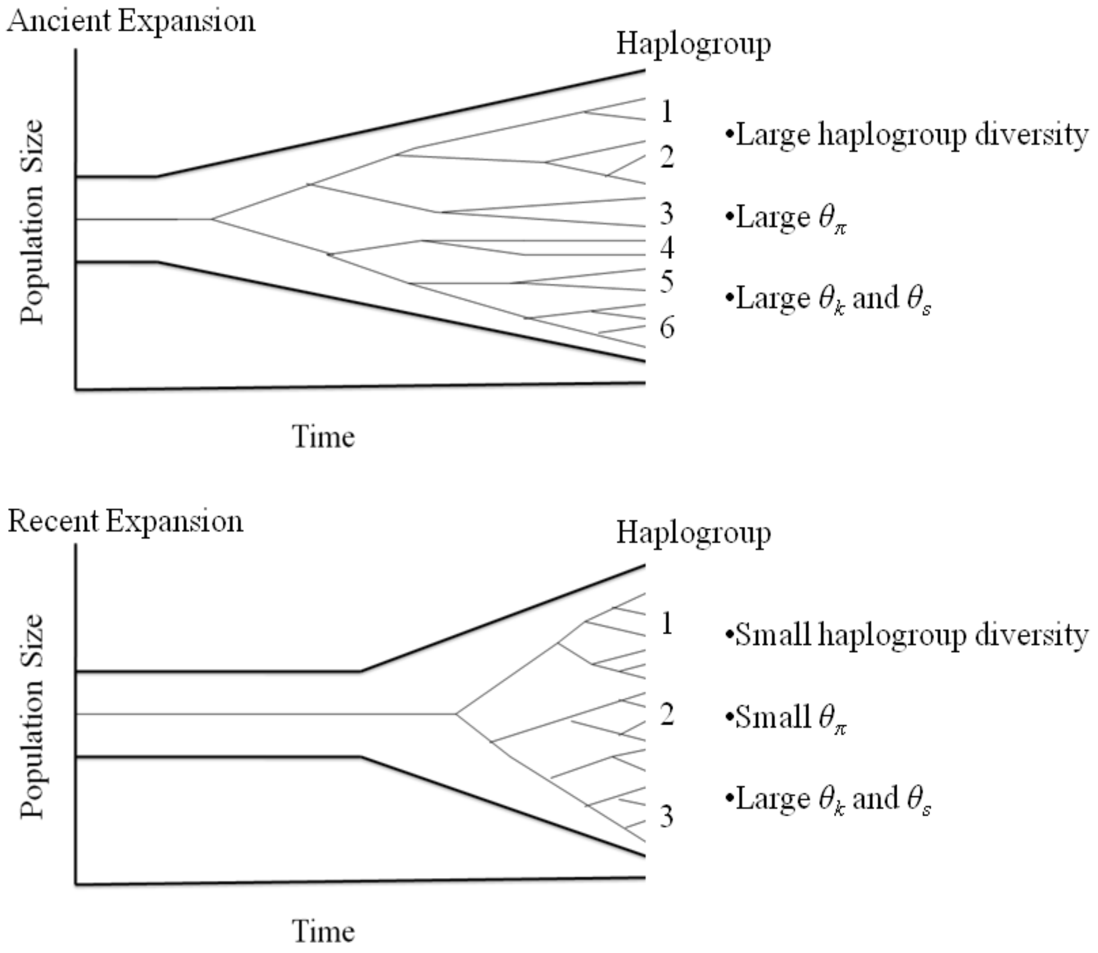

Figure 2. Gene trees of 15 sequences illustrating expected differences in genetic patterns between ancient and recent population expansion

Icelanders and argued that Icelanders experienced a reduction in genetic diversity after a series of historic population declines until the end of 18th century.

Two tests of neutrality, Tajima's *D* (Tajima 1989b) and Fu's *Fs* (Fu 1996), were used to detect genetic evidence of population expansions and bottlenecks on sequence diversity in the Latin American and Bantu populations. These statistics use the measurements of genetic diversity described above and test the assumptions of selective neutrality: that natural selection is not acting on the genes or genetic markers in the study and that population size is stable. Deviations from neutrality (large negative or positive values) show evidence of natural selection or changes in population size. The Tajima's *D* statistic is based on a comparison of θ_π and θ_S, expressed as

$D = \frac{\theta_\pi - \theta_S}{\sqrt{Var(\theta_\pi - \theta_S)}}$ (Tajima1989b).

Fu's *Fs* statistic evaluates the probability of observing the same or fewer numbers of alleles (*k*) given the observed θ_π, and is defined as

$Fs = \ln\left(\frac{S'}{1-S'}\right)$, where $S'=\Pr(K \geq k_{obs} \mid \theta=\theta_\pi)$ (Fu, 1996).

While the effects of natural selection on mtDNA variation need to be explored, if we assume that natural selection has not affected mtDNA HVRI sequence variation, then large negative values of Tajima's *D* and Fu's *Fs* can be used as an evidence of population expansions and large positive values of Tajima's *D* are indicative of genetic bottlenecks.

Numerous studies have detected evidence of population expansion using Tajima's *D* and Fu's *Fs* (Excoffier and Schneider 1999; Fuselli et al. 2003; Helgason et al. 2003). Detecting genetic bottlenecks using Tajima's *D* statistics, on the other hand, is more difficult because θ_S is sensitive to recent changes in population size (Helgason et al. 2003) and mtDNA variation tends to recover very quickly after a bottleneck, which will lower the Tajima's *D* (Fay and Wu 1999).

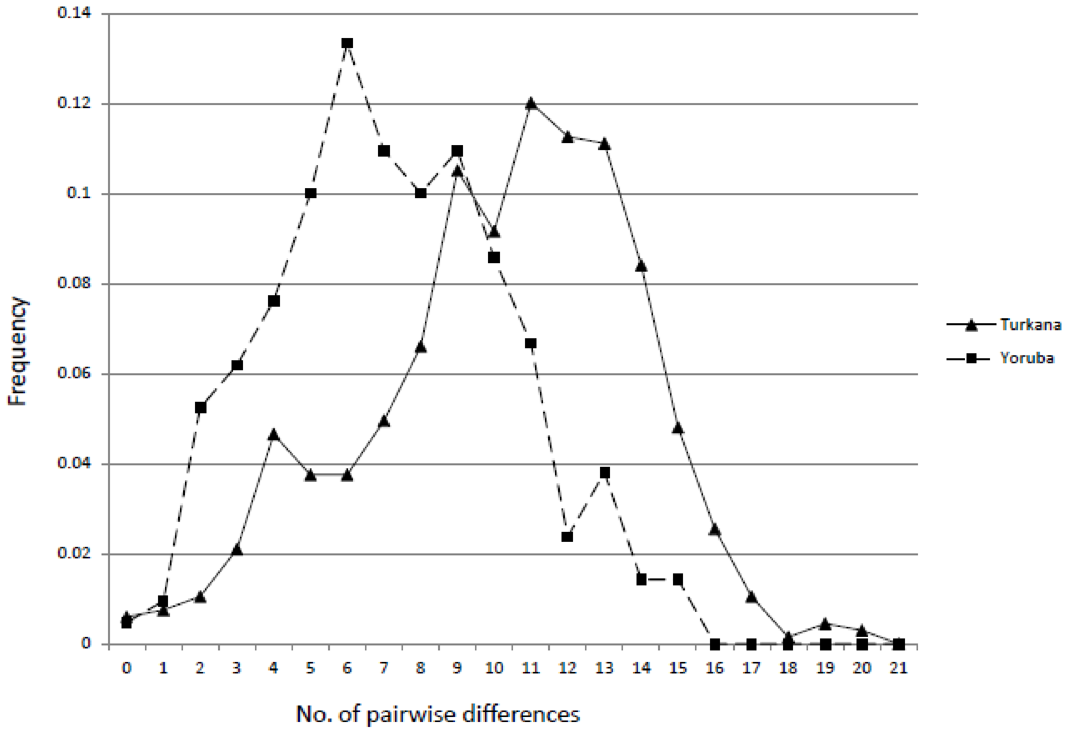

Figure 3. Mismatch distribution of Turkana (Kenya) and Yoruba (Nigeria) showing that the Turkana had population expansion further in the past than the Yoruba

Mismatch distributions were also used to analyze within-population genetic diversity. A mismatch distribution is an analysis of nucleotide differences between sequences from a single population, and the number of nucleotide differences can be graphically represented (Fig. 3). Under the pure demographic expansion model, unimodal mismatch distributions on a graph are interpreted as evidence of demographic expansion (Rogers and Harpending 1992; Slatkin and Hudson 1991). Populations that have experienced recent demographic expansions will have large θ_k and θ_S values, but small θ_π values and mismatch distributions that peak around several nucleotide differences. Populations that experienced ancient expansions will have large θ_π values and mismatch distributions that peak at a much higher number of pairwise nucleotide differences. For example, both the Turkana from Kenya and the Yoruba from Nigeria are genetically diverse populations with large θ_k and θ_S values, but the Turkana θ_π value is larger than the Yoruba value (Salas et al. 2002; Watson et al. 1996) and the Turkana mismatch peak is higher than the Yoruba's, suggesting that the Turkana experienced a population expansion in the more distant past than the Yoruba (Fig. 3).

2.5.2 Analysis of within-population genetic diversity when populations are subdivided

The statistical methods described above assume that populations are not subdivided and that mating is random. However, human populations usually violate these assumptions. Human populations tend to be subdivided into smaller social/reproductive units or demes that interact with each other in very complex ways. Ray et al (2003) and Excoffier (2004) demonstrated that when genetic exchange among demes is high, the demes exhibit genetic patterns that are similar to populations that experience pure demographic expansion.

When human populations are spatially expanding, migrants are incorporated into the demes in large numbers. Ray et al. (2003) demonstrated that, as migration rates increase, genetic diversity increases, so spatially expanding populations have high migration rates, large within-population genetic diversity with large negative values of the two neutrality tests, and unimodal mismatch distributions. Excoffier (2004) also showed that mtDNA variation in forager populations that do not show evidence of past demographic expansion and experienced a series of population size contractions will fit the mismatch distributions expected under a spatial expansion model better. These studies suggest that while the effective population size of the sampled populations can remain small, when mating networks extend beyond the deme through mate exchanges, the sampled populations may be more genetically diverse than expected from their population size and show evidence of population expansion.

The Arlequin population genetics software program was used to test the fit of pure demographic and spatial expansion model (Excoffier et al. 2005). A Sum of Square Deviation (*SSD*) *P-value* is calculated based on the differences between observed and expected mismatch distributions from a coalescent simulation of the models specified.

A maximum-likelihood method was also used to estimate Θ ($\Theta=2N_f\mu$). The MIGRATE program estimates Θ accounting for gene flow (*M*) between demes (Beerli and Felsenstein 1999), unlike the previously discussed θ values where gene flow is not considered. MIGRATE uses a n-Island model (n is the number of subpopulations) and estimates the likelihood parameter $\Theta= [\Theta_1, \Theta_2,\ldots M_1, M_2\ldots]$, where Θ is $2N_f\mu$ for mtDNA and M is m/μ, by exploring genealogical trees, including the topology of branch lengths and various migration scenarios (Beerli and Felsenstein 2001). Rather than exploring all possible genealogical trees, parameters are calculated focusing on the trees with the highest likelihood using a Markov Chain Monte Carlo approach. After each chain, parameters are recalculated and the likelihood is re-evaluated. To obtain more accurate estimates, this process is repeated many times. The weakness of this method is the assumption that no unsampled populations are exchanging genes with the sampled populations. Beerli (2004) examined the effects of unsampled populations on this method. He found that the migration rate, $M=m/\mu$, is not seriously affected, but that the effective population size is upwardly biased. Beerli suggests that a more accurate estimation can be obtained by running the program with many populations (up to about seven populations for analysis of a single marker) at the same time.

2.5.3 Analysis of population subdivision and interactions at the haplotypic level

Considering ethnolinguistic groups as demes, several analytical methods were used to understand the pattern of population subdivision and interactions or gene flow. Genetic heterogeneity within linguistic families and geographic regions and genetic similarity between different linguistic and geographical groups were analyzed by examining the pattern of gene flow or estimating rate of gene flow at haplotypic or population level.

The Network phylogenetic program was used to analyze sequence sharing patterns among different groups of people at the haplotypic level (Bandelt et al. 1999). Network is a phylogenetic program that graphically shows evolutionary relationships among different sequences, or haplotypes. A network approach is more appropriate for intra-specific analysis than the traditional phylogenetic methods because ancestral haplotypes are not usually extinct (Posada and Crandall 2001). Multiple descendant haplotypes are derived from a single ancestral haplotype by accumulating mutations at different nucleotide positions. The ancestral haplotypes are often shared among many different populations, while rare derived haplotypes tend to be unique to a specific population. The sharing of rare derived haplotypes among different populations suggests the possibility of gene flow in the past.

If a linguistic family is genetically homogeneous and there was only limited gene flow between different linguistic families, many derived haplotypes should be shared only with individuals from different ethnic groups within a single linguistic family. On the other hand, if there is gene flow across language families, the members of ethnic groups from different language families will share rare derived haplotypes and movement of individuals, or direction of gene flow, can be traced. Under these circumstances, a language family will become genetically heterogeneous and different language families become genetically similar.

Unfortunately, Network does not give numerical estimates of how much gene flow occurred between demes and does not have the capability for statistical testing.

2.5.4 Analysis of population subdivision and interactions at the population level

There are more population-based methods for the analyses of intra-specific variation than phylogenetic approaches. In this dissertation, they were used to identify population subdivision and to estimate migration rates and the amount of genetic variation within linguistic or geographical groups. They have different built-in assumptions, so several methods were combined in my analyses.

First of all, the migration rate, *M*, was estimated from mismatch distributions under the spatial expansion model (Excoffier 2004). The scaled migration rate, *M*, is expressed as $M=2N_fm$, where N_f is the female effective population size of a deme and m_f is the rate of out-migrating female individuals in a population who are replaced by incoming immigrants in each generation. Assuming that population size was stable, Arlequin estimates the migration rates necessary to produce the observed mismatch distribution. Arlequin uses an infinite-island model, which is equivalent to the continent-island model (Fig. 4). A migration rate is estimated for each sampled population, but the migration model assumes that demes are exchanging mates with a deme with infinite population size and makes no assumptions about which populations are exchanging mates with a particular group.

The amount of genetic variation that exists within linguistic or geographic groups was analyzed using an Analysis of Molecular Variance (AMOVA) (Excoffier et al. 1992). One-group AMOVA estimates among-population variance, within-population variance, and fixation

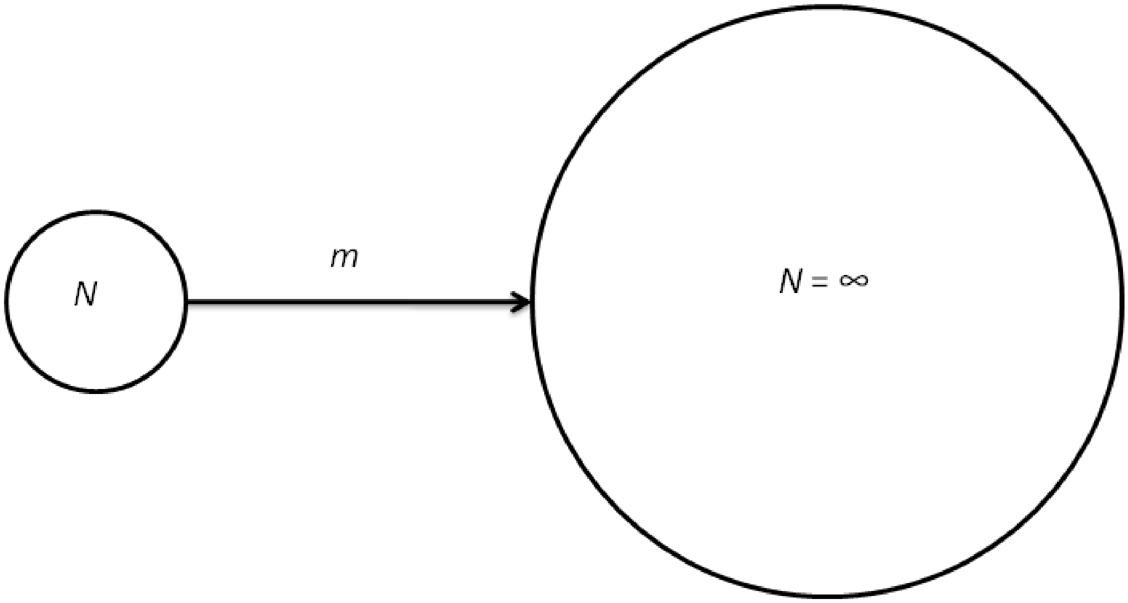

Figure 4. Infinite-island or continent-island model used in the spatial expansion model of mismatch distribution

indices. When populations are grouped together based on language and geography, small among-population variance and small Φ_{ST} values indicate genetic homogeneity within the linguistic or geographical groups. The Φ_{ST} is a fixation index and ranges between 1 and 0. When the Φ_{ST} value is small, it gives a statistically insignificant *P* value. Small Φ_{ST} values are obtained if all of the populations within a group are exchanging the mates or if the populations have diverged recently. In contrast, when gene flow is minimal, genetic drift causes populations to become differentiated and gives large among-population variances, large Φ_{ST} values, and statistically significant *P-values*.

The migration rates ($N_f m$) within linguistic or geographic groups were also estimated from among-population Φ_{ST} estimated using AMOVA. Based on an island migration model (Fig. 1), the relationship between Φ_{ST} and migration rate is expressed as $\Phi_{ST}{}^{mtDNA} = 1/(1+Nv)$, where *N* is the female effective population size and *v* is the sum of the migration (*m*), mutation rate (μ), and their product (v=m+μ +$m\mu$) (Seielstad et al. 1998). Since the mutation rate is very small, Nv is same as N_em, so the formula is Φ_{ST}^{mtDNA} = 1/(1+ N_em). This formula can be also written as N_em = 1/ Φ_{ST}^{mtDNA} – 1. When Φ_{ST} is small, N_em (or 2N_fm) is large. When Φ_{ST} is large, N_em is small. Unfortunately, Φ_{ST} only gives an average migration rate of populations within the group assuming that symmetrical gene flow occurred only among those populations.

Multidimensional Scaling (MDS) was used to visually examine the genetic differentiation pattern and genetic relationship among populations within a group and between different groups. The MDS uses distance measurements, such as population pairwise genetic distances (Φ_{ST}), the genetic distance between each pair of populations calculated using sequence data. The calculation of genetic distances was implemented using Arlequin and the genetic distances were visualized on the MDS plots. Because gene flow among human populations is common, the MDS plots represent genetic relationships better than traditional phylogenetic trees, which assume that two populations diverged from an ancestral population and that contacts between them were limited after the divergence (Sherry and Batzer 1997). Populations are plotted on a two or three-dimensional graph that shows the genetic distances among different populations without the bifurcation problem. When a linguistic or geographic group is genetically homogeneous with small AMOVA Φ_{ST} values, all of the populations in the groups have small population pairwise genetic distances and are plotted closely together. When the populations in a group are heterogeneous, they have large genetic distances and are scattered widely on the plots.

Populations within a group that are derived from a common ancestral population relatively recently may be homogenous, even without gene flow, because genetic drift has not yet had time to cause population differentiation. Also, in order to estimate migration rates using

AMOVA Φ_{ST}, the migration model assumes that 1) gene flow is symmetrical, 2) population sizes and genetic diversity values are equal, and 3) independent evolution of demes (Beerli and Felsenstein 2001; Long and Kittles 2003). In reality, human populations have different population sizes, unequal migration rates, and complex evolutionary histories.

To solve some of these issues, Beerli and Felsenstein (1999) developed a maximum-likelihood and coalescent theory based method to estimate migration rate and effective population size. MIGRATE uses an n-Island model to estimate migration rates (*M*) and Θ ($\Theta=2N_f\mu$, which should be reflection of effective population size) for three or more sampled populations at the same time (Fig. 5). Multiplying *M* and Θ gives $2N_fm$. Unlike the Φ_{ST} based method, the model assumes that demes have unequal effective population sizes, different gene diversities, and variable migration rates between each pair of demes.

Unfortunately, there are limitations to this method. First, it is difficult obtain reliable estimates when only single locus is used. Obtaining more accurate estimates from mtDNA sequence data requires large population samples sizes and long sequences, and no more than seven to eight populations can be analyzed in a single run. Second, this method is computationally intensive. More accurate estimates can be obtained, when more genealogies are sampled by running longer or when the Metropolis coupled Markov Chain Monte Carlo, or 'heating' is employed. This allows for running multiple chains and swapping between them to explore more genealogical spaces. The use of heating generally increases the running time. Finally, MIGRATE assumes that population size and migration rates are stable over time, but few if any, human populations meet this assumption.

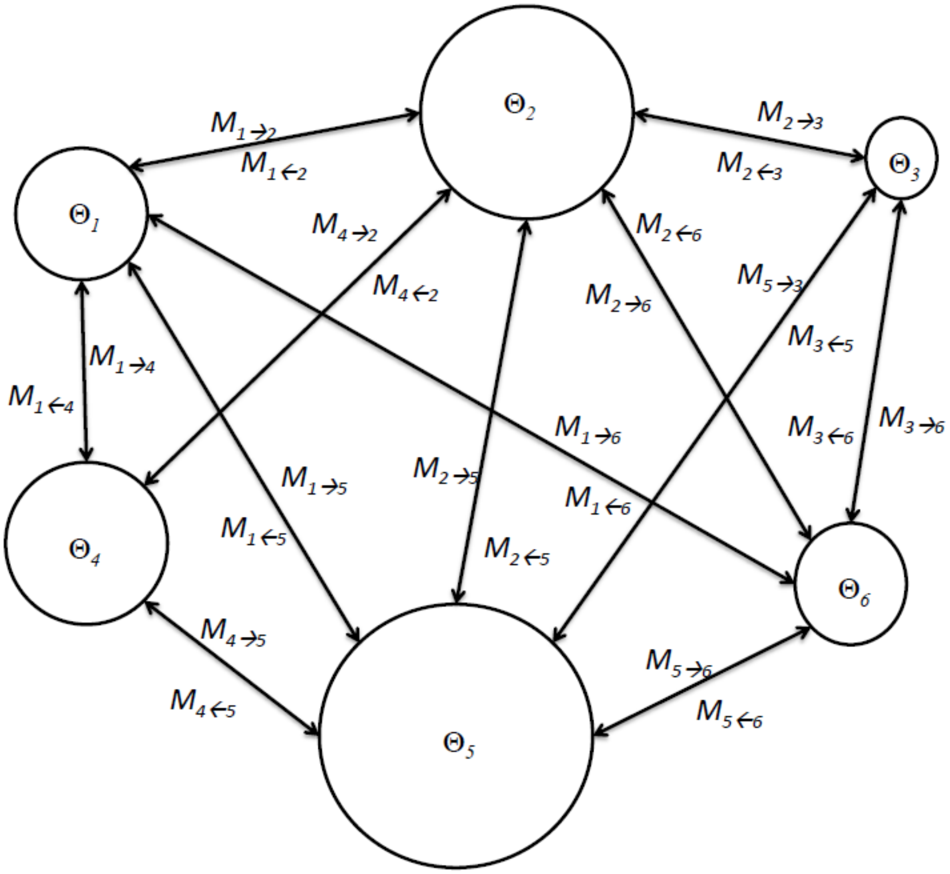

Figure 5 An n-island model with unequal population size and asymmetrical migration rates. This model is used in MIGRATE for estimation of rates (M=m/μ) and Θ=2$N_f\mu$

2.5.5 Computer Simulation

Computer simulations are often used to evaluate the effects of a specific demographic or genetic process on genetic variation. Unlike MIGRATE, coalescent based computer simulation programs such as SIMCOAL (Excoffier et al. 2000; Laval and Excoffier 2004) test more complex demographic scenarios where migration rates and population sizes shift at different points through evolutionary history. MtDNA sequence data were used as input data to estimate the demographic parameters in MIGRATE. SIMCOAL uses a very different approach. Demographic and migration models were developed based on archaeological, cultural, historical, and other evidence. Then, depending on the demographic model developed, the demographic parameters such as demes size and number, population growth and migration rates, and timing of historical events were entered into the simulation. For my analysis, six demes of three different sizes were analyzed and migration rates between the demes were specified in migration matrices. New migration rates after each demographic event were also specified.

For each demographic model, the simulations were replicated 1,000 times. After 1,000 replications, the simulation runs produced 1,000 output files. Each output file contains DNA sequences for each of six demes generated based on the demographic model. The simulation output files were imported into Arlequin to estimate the genetic diversity values for each sampled population in 1,000 replication runs. The results of Arlequin were imported to MS Excel to calculate the average and standard deviation of the estimated genetic diversity values. The fit of the demographic models to the observed mtDNA variation was evaluated by comparing the average genetic diversity values from the simulation to observed genetic diversity values.

2.6 Use of multiple methods to address the research question

Over time new more complex demographic models for population genetic analyses were developed. The earlier models assumed that populations are panmictic, while the newer models assumes that populations are subdivided and there is gene flow between subpopulations (or demes) (see TABLE I). Different analytical methods, however, need to be combined to effectively evaluate how female gene flow, effective population size, and kinship structure affect mtDNA within-population diversity and population subdivision. All of the methods have

TABLE I

NEW AND OLD DEMOGRAPHIC MODEL FOR POPULATION GENETIC ANALYSES

	Older Model	Newer Model
Time Frame	Up to 1990's	From the Late 1990's and 2000 to Now
Assumptions	Panmixia (random mating and no population subdivision)	Population subdivision and gene flow between demes
Examples		
Genetic Diversity Estimates	θ_k, θ_S and θ_π	Θ
Mismatch Distribution Models	Sudden Demographic Expansion	Spatial Expansion

unrealistic assumptions that human populations and demes violate. Also, each of the methods used to estimate migration rate have strengths and weaknesses and make different assumptions about the populations that exchange migrants. The use of multiple methods is necessary to examine the pattern of gene flow and its affect on within-population genetic diversity. The computationally simpler methods such as mismatch distribution and AMOVA have many simplifying assumptions and these methods should be complemented with more computationally intensive methods such as MIGRATE. While these methods still have unrealistic assumptions, they allow us to evaluate more complex patterns. The observed values of θ generated by different analytical methods can be evaluated using computer simulation. Therefore, if female gene flow had a great impact on mtDNA variation, various methods will consistently detect the evidence of population interactions and inflated within-population genetic diversity should be obtained from the populations that exhibit the evidence of population interactions.

3. GENETIC EVIDENCE OF THE AYMARA EXPANSION AND MITOCHONDRIAL VARIATIONIN THE CENTRAL ANDES

3.1 Introduction

Numerous studies have demonstrated that Andean populations show high within group variability with low population differentiation while Amazonian populations exhibit low within group variability with high population differentiation (Fuselli et al. 2003; Lewis and Long 2008; Lewis et al. 2007b; Lewis et al. 2005; Tarazona-Santos et al. 2001). The main focus of these studies has been on the differences between Andean and Amazonian populations; contrasting the different microevolutionary processes each region experienced and asking whether multiple migrations into South America could account for the differences. Fuselli et al (2003) note that several evolutionary processes could explain the Andean mitochondrial DNA (mtDNA) pattern of high within-group variability with an excess of rare alleles observed by researchers. Demographic expansion, range (also called a spatial) expansion with high gene flow among groups, or selection could each produce this pattern. Previous research projects (Bert et al. 2004; Bert et al. 2001; Corella et al. 2007; Fuselli et al. 2003; Lewis et al. 2007b; Lewis et al. 2005) have included large numbers of Quechua speakers living in the north and south central Andes and the eastern foothills, but Aymara-speakers, the other main language group found in the Andes, are relatively underrepresented, especially in their core area. This study attempts to examine the three possible scenarios (demographic expansion, range expansion, or selection) outlined by Fuselli et al (2003) to explain the area's mtDNA genetic variation by adding a large sample of Aymara speakers from La Paz, Bolivia, and uses a variety of analytical methods

3.2 Background

The south central Andes encompasses the area around Lake Titicaca that comprises southern Peru, western Bolivia, and the extreme northern part of Chile. Western and eastern mountain ranges and the high, wet plain or puna called the Altiplano between them characterize this area. The Pacific coast and western foothills of the Andes are extremely desertous. Rainfall is significantly higher on the eastern slopes that link the Andes to the Amazon basin.

The dominant indigenous languages spoken in this area are Quechua and Aymara. The Quechua is the largest linguistic group (approximately 7-11 million speakers) with a wide geographic distribution ranging from Columbia to Argentina and Chile (Stark 1985a; Stark 1985b). At the time of European contact, numerous ethnic groups lived in the Central Andes, and many spoke different dialects of Quechua language (Rowe 1946). The Quechua dialects are divided into groups designated Quechua I and II by Torero (1964) and Quechua A and B by Parker (1963). Linguists believe that the older group, Quechua I (B) spread through the northern and central Andes in 7th-10th centuries; the Inka carried the Quechua language with them as they conquered the Andes and re-introduced their dialect (Quechua II or A) into the north and central Andes (Stark 1985a; Stark 1985b). The Spanish expanded its distribution even further when they adopted Quechua II as a lingua franca during the Colonial era (Hardman 1985; Mannheim 1991).

The Aymara is the second largest language group (over 2 million speakers) and Aymara speakers are broadly distributed from highland Bolivia and Peru to mid-altitude valleys in Chile and lowland Bolivia (Briggs 1985b). The Aymara language has not been as thoroughly studied (Briggs 1985b). Researchers initially believed that Quechua and Aymara were closely related to each other, but have now determined that the languages belong to entirely different language families and any similarities are the result of borrowing (Mannheim 1985; Mannheim 1991). Aymara is a member of the Jaqi language family, which includes the nearly extinct Kawki and Jaqaru, which is still widely spoken in the northcentral Andes (Briggs 1985a).

The Aymara people are traditionally pastoral-agriculturalists who herd llamas and alpacas and grow a huge variety of potatoes. The majority of them live on the south central Andean plateau, Altipplano, an important area in the development of prehistoric complex societies (Kolata 1993). The Altiplano was a linguistically diverse area where Spanish identified several language speakers living around the Lake Titicaca at the time of contact, but historians disagree about the specifics. In addition to the Aymara and Quechua, Pukina and Uru were spoken on the Bolivian Altiplano around the Lake Titicaca (Browman 1994; Murra 1968), but are extinct now. The origins of the Aymara people are unclear, but their territory roughly overlaps the areas influenced by Tiwanaku prehistoric culture, so (Browman 1994) and Janusek (2004) have argued that they are descendants of the Tiwanaku Empire. Other researchers, such as a linguist, Torero (1987), have suggested that they were herders who arrived in the area after the Tiwanaku Empire collapsed around 900 years ago

Population movement has always been an important part of settlement history and economic practices in the Andes (D'Altroy 2002; Murra 1985a). The Quechua-speaking Inka sent mitmaq colonists into many parts of the Andes, including the Lake Titicaca Altiplano, to pacify the people they conquered and disrupt any brewing rebellions. Taxes were often collected in the form of labor that required men to travel long distances to fulfill their obligations to the state. Murra (1968; 1985a) has written about what he calls "vertical archipelago" systems that colonial Aymara kingdoms utilized to take advantage of the diverse and stacked Andean ecological zones. They established permanent colonies arranged vertically on both sides of the Andean slopes. Maize and other crops grown at lower elevations had nutritional as well as ritual importance for the people who lived on the Altiplano, where potato was a major agricultural crop. This vertical control of resources is ancient in the Andes, and the Tiwanaku Empire maintained colonies at lower elevations, for example in the Moquegua (Goldstein 1989; Goldstein 1993). Although ancient DNA analysis showed genetic differences between the Middle Horizon Chen Chen population and modern highland populations (Lewis et al. 2007a), bone chemistry and cranial data from the same site revealed the existence of highland migrants in the area (Blom et al. 1998; Knudson 2008).

The interactions among different ethnic groups in the Andes are likely to have influenced regional genetic patterns along the lines suggested by Fuselli and colleagues (2003). If the Aymara intermarried with other groups in the highlands and in the areas they colonized, or if the group defined as Aymara speakers includes biologically unrelated individuals who adopted the language and culture, this would be an example of a range expansion. Under these circumstances, linguistic, cultural and genetic patterns would not be expected to match well. However, the highland Aymara people may not have mingled with locals in n sufficient to affect genetic patterns. If they maintained a distinct identity with strong cultural and biological boundaries, then demographic expansion may explain the distribution of Aymara language and culture. Under this scenario, the Aymara would have expanded because of their population numbers increased, either through higher fertility, lower mortality or a combination of the two, and geographical distribution of Aymara language and genetic patterns will match closely.

Alternatively, the Aymara could have had a biological advantage that made them well adapted to living in the Altiplano. High altitude adaptation has a long history of study, both in the Andes and in Tibet (Beall 2007; Beall et al. 1999; Frisancho et al. 1999; Vitzthum et al. 2004; Vitzthum et al. 2009). A recent study through a genome scan identified candidate genomic regions that are positively selected in Andean populations (Bigham et al. 2010), but no study demonstrated the mitochondrial adaptation in the highland environment. However, many genetic studies proposed that natural selection affect human mtDNA variation, so the selective pressure in the Andean environment could have affected mtDNA variation. Numerous genetic studies, such as comparison of human to chimpanzee mitochondrial genome (Nachman et al. 1996) and mitochondrial genome comparison of human populations (Elson et al. 2004; Ingman and Gyllensten 2007; Mishmar et al. 2003; Ruiz-Pesini et al. 2004), demonstrated that natural selection affected mitochondrial genome. Mishmer et al. (2003) and Ruis-Pesini et al. (2004) proposed that Amerindian mitochondrial haplogroup (HGP) A, C, and D were naturally selected in the cold environment. Kivisild et al. (2006), on the other hand, found conflicting results. Researchers have also found that the 16189 T→C transition is associated with high body mass index and type 2 Diabetes Mellitus in Asian populations (Kim et al. 2002; Liou et al. 2007; Park et al. 2008). This mutation is part of the HPG B diagnostic motif, a HPG that occurs in high frequency in the Andean highlands, especially in the Aymara (Merriwether et al. 1995). While the south central Andean groups are characterized by high mtDNA sequence diversity, their mtDNA haplogroup diversity is extremely low (Lewis et al. 2005). Haplogroup B is the dominant haplogroup, reaching a frequency as high as 93.3% in one Aymara group sampled (Bert et al 2001).

3.3 Samples and methods

3.3.1 Samples

The genetic material used in the study was taken from blood spot samples collected from 61 healthy unrelated Aymara women from the Altiplano, highland plateau of Bolivia. The samples were collected in1995 in La Paz as a part of Project REPA (Reproduction and Ecology in Provincia Aroma) analysis of reproductive hormones (Vitzthum et al. 2002; Vitzthum et al. 2004; Vitzthum et al. 2009). The verbal consents and signatures (or marks) from the participants were obtained after the goals of project, procedures, and possible risks and benefits was explained to the participants in their native language.

3.3.2 Laboratory methods

The Puregene® DNA purification kit was used to extract DNA from blood spot samples following the protocol provided. The Hypervariable Region I (HVRI) of the mtDNA control region was amplified using a touchdown PCR protocol and either of the primer sets (L15985 5'-GCACCCAAAGCTAAGATTCTAA-3' and H404 5'-AAAGTGCATACCGCCAAAAG-3'or L15926 5'-TCAAAGCTTACACCAGTCTTGTAAACC-3' and H16498 5'-CCTGAAGTAGGAACCAGATG-3'). The amplified product was sequenced in both directions using the BigDye Terminator Cycle Sequencing Kit, version 3.1 (Perkin Elmer Biosystems) and analyzed on an ABI 3730 DNA sequencer housed at the Field Museum's Pritzker Molecular Biology Laboratory.

Sequences were edited and aligned in Sequencher 4.1.4 (GeneCodes). The nucleotide positions (nps) between 16024 and 16383 relative to the Cambridge Reference Sequence (Anderson et al. 1981) were edited, but subsequent analyses are confined to nps 16056-16383 to facilitate comparisons with previous work. When T→C transition at np 16189 was present, nps 16182 and 16183 were excluded from analysis because of heteroplasmy (Bendall and Sykes 1995; Pfeiffer et al. 1999). The HPGs were assigned based on the diagnostic mutations observed in the HVRI sequence, HPG A (16223, 16290, 16319, 16362), HPG B (16189, 16217), HPG C (16233, 16325, 16362), and HPG D (16223, 16325, 16362). Additionally, since 16189T→C mutation is present in non-B HPGs, the 9 bp deletion that characterizes HPG B was screened to confirm the HPG assignment.

3.3.3 Analytical methods

The mtDNA sequences were analyzed using several methods. First, the mtDNA HVRI sequence variation of the La Paz Aymara was compared to 40 published population samples from Latin America (Fig. 6 and TABLE II). Five additional population samples where only mtDNA HPG frequencies were reported were included in HPG frequency comparisons, but could not be included in other analyses (see TABLE V for the list). The Aymara sample from the lowland Bolivia analyzed by Corella et al. (2008) was excluded in my analyses due to very small sample size.

The Arlequin population genetics software program (Excoffier et al. 2005; Schneider et al. 2000) was used to estimate within-population genetic diversity, haplotype diversity (*h*) and parameter $\theta=2N_f\mu$ (θ_k, θ_S and θ_π), conduct two test of molecular neutrality (Tajima's *D* and Fu' *Fs*), and analyze mismatch distributions. While ancient demographic history affects mean pairwise nucleotide differences (π), number of alleles (*k*) and polymorphic sites (*S*), so the two

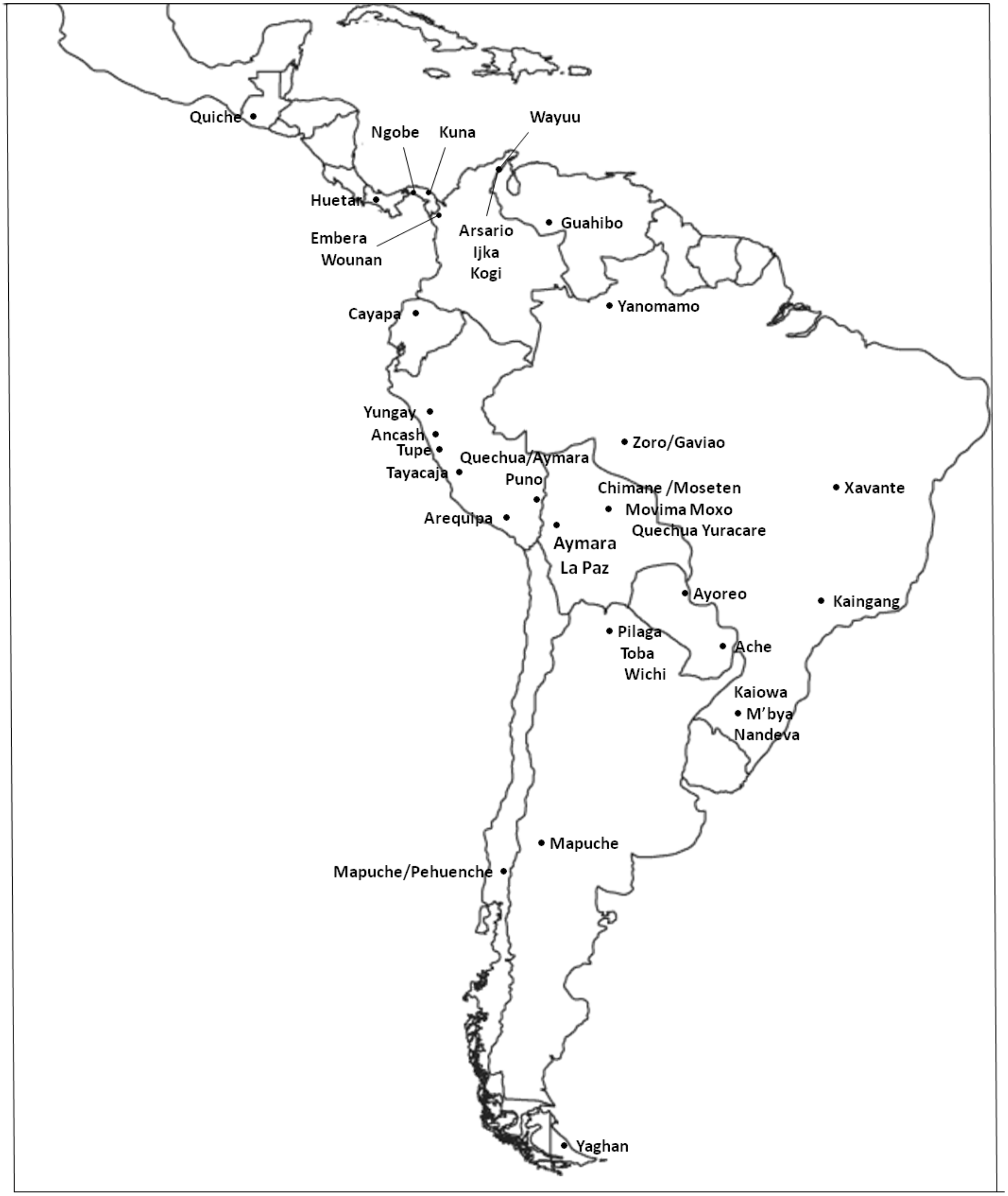

Figure 6. Map of Latin America showing the locations of sampled populations

TABLE II

NEW WORLD POPULATIONS ANALYZED[a]

Populations	Abbr.[b]	n	References
Mesoamerican			
Quiche	Qu	23	(Boles et al. 1995)
Central American/Chibchans			
Arsario	As	28	(Melton et al. 2007)
Huetar[c]	Hu	27	(Santos et al. 1994)
Ijka	Ij	31	(Melton et al. 2007)
Kogi	Ko	21	(Melton et al. 2007)
Kuna[c]	Ku	63	(Batista et al. 1995)
Ngobe[c]	Ng	46	(Kolman et al. 1995)
Western Lowland South Americans			
Cayapa[c]	Ca	30	(Rickards et al. 1999)
Embera[c]	Em	44	(Kolman and Bermingham 1997)
Wounan[c]	Wo	31	(Kolman and Bermingham 1997)
North-Central Andes			
Ancash (Quechua)	An	33	(Lewis et al. 2005)
Tayacaja (Quechua)[c]	Ta	61	(Fuselli et al. 2003)
Tupe (Jaqaru – Aymaran)	Tp	16	(Lewis et al. 2007b)
Yungay (Quechua)[c]	Yg	36	(Lewis et al. 2007b)
South-Central Andes			
Arequipa (Quechua)	Ar	22	(Fuselli et al. 2003)
Aymara La Paz[c]	ALP	61	This study
Aymara Puno	AP	14	(Lewis et al. 2007b)
Quechua Puno[c]	QP	30	(Lewis et al. 2007b)
Southern Andes			
Mapuche (Argentina)[c]	AM	39	(Ginther et al. 1993)
Mapuche/Pehuenche[c]	MP	58	(Moraga et al. 2000)
Yaghan	Yh	15	(Moraga et al. 2000)
Lowland Bolivians, Dept. of Beni			
Chimane/Moseten	CM	20	(Corella et al. 2007)
Movima	Mo	12	(Bert et al. 2004)
Moxo	Mx	27	(Bert et al. 2004)
Quechua Beni	QB	16	(Corella et al. 2007)
Yuracare	Yu	15	(Bert et al. 2004)

TABLE II (continued)

NEW WORLD POPULATIONS ANALYZED[a]

Populations	Abbr.[b]	n	References
Gran Chaco			
Pilaga[c]	Pi	38	(Cabana et al. 2006)
Toba[c]	Tb	67	(Cabana et al. 2006)
Wichi[c]	Wi	99	(Cabana et al. 2006)
Other Lowland South Americans			
Ache		63	(Schmitt et al. 2004)
Ayoreo		91	(Dornelles et al. 2004)
Guahibo	Gu	59	(Vona et al. 2006)
Kaingang	Kg	78	(Marrero et al. 2007)
Kaiowa	Ki	120	(Marrero et al. 2007)
M'bya	Mb	24	(Marrero et al. 2007)
Nandeva	Na	56	(Marrero et al. 2007)
Wayuu	Wy	29	(Melton et al. 2007)
Xavante		25	(Ward et al. 1996)
Yanomamo	Ya	129	(Merriwether et al. 2000)
Zoro/Gaviao	ZG	57	(Ward et al. 1996)

[a] Populations are arranged based on their linguistic/geographical grouping
[b] Abbreviations are used for multidimensional scaling analysis.
[c] Populations used for comparison of Θ to θ_k, θ_S, and θ_π in Chapter 5.

other θ estimators (θ_k, and θ_S), are sensitive to recent demographic events (Helgason et al. 2003; Helgason et al. 2000; Rogers 1995; Tajima 1989a). Mismatch distribution is an analysis of nucleotide differences between sequences from a single population, and number of nucleotide differences is graphical represented. Unimodal mismatch distribution on the graph is interpreted as the evidence of demographic expansion (Rogers and Harpending 1992; Slatikin and Hudson 1991), but Excoffier (2004) suggested that two populations from the same spatial expansion wave produce similar mismatch distribution.

To examine patterns of subdivision and interactions among the Aymara and other Latin American groups, first, pairwise population genetic distances, analysis of molecular variance (AMOVA) and Mantel tests were undertaken in Arlequin. The population pairwise genetic distances were, then, visualized using Multidimensional scaling (MDS) analysis with SPSS statistical software. Two different methods were used for estimation of migration rates. Arlequin was used to calculate the migration rate, M=2$N_f m$, under a mismatch distribution spatial expansion model (Excoffier, 2004), and migration rates, 2$N_f m$, between each pair of populations analyzed were estimated using MIGRATE, a coalescent based maximum likelihood method.

Based on cultural and linguistic similarities and ecology of the area, populations in western South America were grouped into six regional/linguistic groups: Aymara, Quechua, southern Andes, Lowland Bolivians, Gran Chaco, and northwestern lowland South Americans (See TABLE II). The parameters, 2$N_f m$ and $\Theta=2N_f\mu$, were estimated using averages of more than three independent runs in each regional set through 10 short chains (10,000 genealogy per chain) and three long chains (100,000 genealogy per chain) with increments of 20 and 200 steps respectively. The first 100,000 trees in each chain were discarded. Instead of sampling more genealogies, Metropolis coupled Markov Chain Monte Carlo, or 'heating' was used to explore a wider genealogical space by setting four temperatures (1, 1.5, 3, 6) and long chains were replicated.

3.4 Results

3.4.1 General patterns of South-Central Andean and Aymara mtDNA variation

As expected from previous reports (Merriwether et al. 1995), HPG B is the most common haplogroup in the La Paz Aymara sample (n=38; see TABLE III) and HPG D is rare (n=1). HPG A (n=7) and C (n=14) are relatively common. One sample could not be assigned a HPG from the HVRI sequence data. The most common haplotype, B05, has a 188-189-217 motif that is shared by another HPG B haplotype, B10. This motif is found among south central Andeans in

TABLE III

MITOCHONDRIAL DNA HVRI SEQUENCE OF AYMARA SAMPLES

Haplogroup		HVRI Sequence[a]	n
HPG A	A01	111 223 290 319 362	2
	A02	86 111 223 266 290 319 362	1
	A03	111 183C 189 223 290 319 362	1
	A04	111 188 223 290 319 357 362	1
	A05	111 217 223 290 319 343T 362	1
	A06	111 129 217 223 290 319 343T 362	1
HPG B	B01	183C 189 217	7
	B02	182C 183C 189 217	2
	B03	183D 217	1
	B04	182C 183C 189	1
	B05	183C 188 189 217	9
	B06	58T 111 183C 189 217	1
	B07	111 183C 189 217	1
	B08	92 183C 189 217	2
	B09	93 183C 189 217	2
	B10	93 183C 188 189 217	1
	B11	93 183C 184 189 217	1
	B12	129 182C 183C 189 217	1
	B13	168 183C 189 192 217	2
	B14	172 183C 189 217	2
	B15	172 189 217 256 288	1
	B16	182C 183C 189 217 295	1
	B17	183C 189 217 261 319	2
	B18	183C 189 217 362	1
HPG C	C01	223 298 325 327	5
	C02	223 298 325	2
	C03	140 223 298 325 327	1
	C04	93 183C 189 223 298 311 325 327	1
	C05	124 183C 189 223 298 325 327	3
	C06	223 263 298 325 327	1
	C07	223 298 325 327 345T	1
HPG D	D01	183C 189 223 310 325 362	1
Unknown	U01	248	1
Total			61

[a] Mutations at nucleotide position between 16024-16383 (without the prefix “16”) are reported. The number indicates the position of transition (C/T or A/G). Otherwise, transversion and deletion is indicated.

intermediate to high frequencies suggesting a south central Andean origin for this haplotype. This motif is common among the Quechua (Corella et al. 2007) and is also found among the Aymara from lowland Bolivia (Corella et al. 2007), the Arequipa Quechua (Fuselli et al., 2003), and two of the Gran Chaco groups (the Pilaga, and Wichi) (Cabana et al. 2006).

Although the 16,189T→C mutation is part of the haplogroup B motif, it has been observed on other haplogroups in other populations as well (TABLE IV, with reference). The transition observed in haplotypes from all four haplogroups in the La Paz Aymara sample - HPGs A (A03), C (C04 and C05), and D (D01). The presence of the mutation on a HPG A background has been observed among the lowland Bolivians (Chimane, Moseten, Moxo, and Yuracare), western lowland South Americans (Cayapa, Embera, Wounan), the Chibchans (Asario and Kogi), and Amazonians (Gaviao and Zoro), but not among the Quechua. Haplotypes with this mutation on a HPG C background are common among the Quechua in Arequipa, Puno, Tayacaja, and Yungay, but are also found among the Movima, Argentina Mapuche, and Guahibo. HPG D haplotypes with this mutation have been previously reported from the Puno Aymara and Quechua speaking groups in Ancash and Tayacaja. This same mutation is common among southern Andeans, but is paired with a 16187C→T substitution in those populations. The La Paz Aymara sample is the only population to date that includes haplotypes with this mutation in all four haplogroups.

The La Paz Aymara HPG frequencies are similar to other central Andean populations, especially other south central Andeans (TABLE V). However, the La Paz Aymara have the lowest HPG B frequency among the Aymara and the second lowest after the Puno Quechua among south central Andeans. The Aymara and Quechua of the lowland Bolivia, Department of

TABLE IV

16189T→C MUTATIONS IN HPG A, C, AND D BACKGROUNDS WITH NUMBER OF HAPLOTYPES, NUMBER OF INDIVIDUALS AND FREQUENCY[a]

	A	C	D	Poly-C tract frequency[b]	References:
Quiche	1 (n=1; 4.3%)			5 (n=6; 26.1%)	(Boles et al., 1995)
Arsario	1 (n=11; 39.3%)			1 (n=11; 39.3%)	(Melton et al., 2007)
Huetar			1 (n=2; 7.4%)	2 (n=3: 11.1%)	(Santos et al., 1994)
Kogi	1 (n=14; 66.7%)			1 (n=14; 66.7%)	(Melton et al., 2007)
Cayapa	1 (n=3; 10.0%)			3 (n=9; 30.0%)	(Rickards et al., 1999)
Embera	1 (n=2; 4.5%)			10 (n=25; 56.8%)	(Kolman and Bermingham, 1997)
Wounan	2 (n=5; 16.1%)			5 (n=9; 29.0%)	(Kolman and Bermingham, 1997)
Ancash			1 (n=1; 3.0%)	14 (n=18; 54.5%)	(Lewis et al., 2005)
Tayacaja		1 (n=1; 1.6%)	4 (n=5; 8.2%)	18 (n=24; 39.3%)	(Fuselli et al., 2003)
Arequipa		1 (n=1; 4.5%)		10 (n=11; 50%)	(Fuselli et al., 2003)
Aymara La Paz	1 (n=1; 1.6%)	2 (n=4; 6.6%)	1 (n=1; 1.6%)	17 (n=30; 49.2%)	This Study
Aymara Puno			1 (n=1; 7.1%)	4 (n=5; 37.5%)	(Lewis et al., 2007b)
Quechua Puno		3 (n=4; 13.3%)		11(n=14; 46.6%)	(Lewis et al., 2007b)
Mapuche (Argentina)		1 (n=1; 2.6%)	1 (n=1; 2.6%)	6 (n=16; 42.0%)	(Ginther et al., 1993)
Mapuche			2 (n=7; 20.6%)	2 (n=8; 23.5%)	(Moraga et al., 2000)
Pehuenche			3 (n=4; 16.7%)	6 (n=7: 29.2%)	(Moraga et al., 2000)
Yaghan			2 (n=5; 33.3%)	2 (n=5; 33.3%)	(Moraga et al., 2000)
Chimane	4 (n=4; 40%)			7 (n=9; 90.0%)	(Corella et al., 2007)
Moseten	3 (n=4; 40%)			7 (n=9; 90.0%)	(Corella et al., 2007)
Movima		2 (n=2; 16.7%)		3 (n=3; 25.0%)	(Bert et al., 2004)
Moxo	1 (n=1; 3.7%)			7 (n=8; 29.6%)	(Bert et al., 2004)
Yuracare	1 (n=1; 6.7%)			5 (n=7; 46.7%)	(Bert et al., 2004)

TABLE IV (continued)

16189T→C MUTATIONS IN HPG A, C, AND D BACKGROUNDS WITH NUMBER OF HAPLOTYPES, NUMBER OF INDIVIDUALS AND FREQUENCY[a]

	A	C	D	Poly-C tract frequency[b]	References:
Guahibo		1 (n=1; 1.8%)		3 (n=3; 5.1%)	(Vona et al., 2006)
Zoro/Gaviao	1 (n=1; 1.8%)		2 (n=3; 5.3%)	5 (n=9; 15.8%)	(Ward et al., 1996)

[a] number of individuals/total sample size x 100)

[b] Poly-C tract has CCCCCCCCCC between np 16,184-16,193, and its frequency includes HPG A, B, C, and D).

TABLE V

HPG FREQUENCIES OF THE ANDEAN POPULATIONS ARRANGED GEOGRAPHICALLY FROM NORTH TO SOUTH

Populations	n	A	B	C	D	Other
Yungay (Quechua)	36	2.8	47.2	36.1	13.9	0.0
Ancash (Quechua)	33	9.1	51.5	18.2	21.2	0.0
Tupe (Aymaran)	16	0.0	68.8	16.1	0.0	0.0
Tayacaja (Quechua)	61	21.3	32.8	13.3	29.5	3.3
Quechua (Pasco and Lima)[a]	52	3.9	53.8	17.3	19.2	5.8
Arequipa (Quechua)	22	9.1	68.2	13.6	9.1	0.0
Quechua Puno	30	6.7	60.0	23.3	10.0	0.0
Quechua (Beni, Bolivia)[b]	32	15.6	75.0	9.4	0.0	0.0
Aymara Puno (Peru)	14	0.0	71.4	14.3	14.3	0.0
Aymara La Paz	61	11.5	62.3	23.0	1.6	1.6
Aymara (Beni, Bolivia)[b]	33	0.0	93.3	3.0	3.0	0.0
Aymara (Chile)[c]	172	6.4	67.4	12.2	14.0	0.0
Atacamenos[c]	50	12.0	72.0	10.0	6.0	0.0
Quebrada de Humahuaca (Argentina)[d]	46	10.9	67.4	17.4	4.3	0.0
Mapuche (Argentina)[e, f]	136	10.3	35.3	21.3	28.7	5.1
Mapuche (Chile)[g]	111	0.0	7.2	44.1	48.7	0.0
Pehuenche[c]	100	2.0	9.0	37.0	52.0	0.0
Huilliche[c]	80	3.8	28.8	18.7	48.7	0.0
Tehuelche[f]	29	0.0	20.7	24.1	55.2	0.0
Yaghan[g]	21	1.3	8.0	43.0	47.7	0.0

References: [a] (Rodriguez-Delfin et al. 2001), [b] (Bert et al. 2001), [c] (Merriwether et al. 1995), [d] (Dipierri et al. 1998), [e] (Ginther et al. 1993), [f] (Goicoechea et al. 2001), and [g] (Moraga et al. 2000)

Beni, and the Quebrada de Humahuca and Atacameno populations living in the border area of northern Argentina, southern Bolivia, and Chile also have similar HPG frequencies. The north central Andeans have lower frequencies of HPG B and higher frequencies of various other HPGs when compared with the south central Andeans. Southern Andean populations have very different HPG frequencies; HPG C and D are common there (Moraga et al. 2000).

3.4.2 Aymara and Latin American mtDNA genetic diversity

TABLE VI shows the results of summary statistics for each of the 40 population samples, organized by regions and arranged in the order from high to low *h* within each group. All four measurements of within-population genetic diversity are significantly correlated ($P < 0.05$). The most genetically diverse populations are the central Andeans and Quiche Mayans who are the agriculturalists from areas where prehistoric state level societies once flourished, and they have many rare haplotypes (*k*) and variants (*S*), so large θ_k and θ_S values, parameters that reflect recent demographic history (Helgason et al. 2003; Tajima 1989a).

As Corella et al. (2007) and Cabana et al. (2006) observed, neighboring Andean populations from the southern Andes, lowland Bolivia, Grand Chaco area, and northwestern lowland South America have intermediate genetic diversity values, but the Pilaga, a forager population from Gran Chaco, and the Moxo, a horticulturalist population from lowland Bolivia, are also among the most genetically diverse populations and the Pilaga have the highest θ_π (Bert et al., 2004; Cabana et al., 2008; Corella et al., 2008). At the other end of the spectrum, low levels of genetic diversity are observed among the Central Americans and relatively isolated lowland South American groups from Amazonian areas.

TABLE VI

SUMMARY STATISTICS OF LATIN AMERICAN POPULATIONS

	h	θ_k (95% CI)	θ_S (SD)	θ_π (SD)	Tajima's *D*	*Fs*	*M*[a]
Mesoamerica							
Quiche	0.945	17.606 (7.988-39.670)	6.774 (2.543)	6.177 (3.398)	-0.875	-4.949*	213.37
Central Americans/Chibchans							
Ngobe	0.763	2.057 (0.886-4.472)	2.730 (1.073)	5.198 (2.844)	1.684*	3.388	1.077
Arsario	0.725	1.032 (0.347-2.794)	2.570 (1.110)	4.878 (2.729)	1.926*	5.686	1.311
Huetar	0.709	2.728 (1.134-6.223)	3.113 (1.294)	4.018 (2.307)	0.413	1.179	2.107
Kuna	0.592	1.807 (0.778-3.860)	2.122 (0.853)	3.882 (2.190)	1.519	2.775	1.065
Kogi	0.524	0.703 (0.203-2.203)	2.780 (1.239)	3.851 (2.249)	0.581	5.398	1.153
Ijka	0.184	0.605 (0.177-1.849)	2.780 (1.150)	1.728 (1.151)	-1.488	2.811	0.249
Western Lowland South Americans							
Embera	0.940	12.155 (6.650-21.951)	5.057 (1.758)	6.673 (3.563)	0.312	-4.106	5.578
Wounan	0.912	9.256 (4.587-18.438)	7.009 (2.470)	7.569 (4.039)	-0.380	-1.303	11.303
Cayapa	0.837	3.226 (1.409-7.041)	4.544 (1.721)	7.253 (3.888)	1.155	2.873	6.291
North-Central Andes							
Ancash	0.981	67.034 (31.024-153.763)	9.609 (3.225)	6.869 (3.688)	-1.469	-20.108**	119.899
Tayacaja	0.967	53.609 (32.019-91.003)	10.043 (2.996)	7.087 (3.739)	-1.377	-25.286**	25.083
Yungay	0.954	17.738 (9.348-33.737)	6.511 (2.254)	6.203 (3.353)	-0.746	-7.380*	21.670
Tupe	0.867	7.691 (3.078-19.286)	5.425 (2.263)	6.426 (3.598)	-0.214	-0.766	1.206
South-Central Andes							
Arequipa	0.978	43.883 (17.646-117.364)	6.584 (2.503)	5.961 (3.298)	-0.933	-10.892**	NA[b]
Quechua Puno	0.972	35.666 (17.130-77.019)	8.077 (2.813)	6.150 (3.348)	-1.280	-12.724**	66.485
Aymara Puno	0.967	21.684 (7.586-66.408)	6.603 (2.768)	5.701 (3.263)	-1.117	-3.999*	NA[2]
Aymara La Paz	0.950	26.488 (16.003-43.803)	7.265 (2.258)	5.843 (3.135)	-1.075	-19.051**	NA[2]

TABLE VI (continued)

SUMMARY STATISTICS OF LATIN AMERICAN POPULATIONS

	h	θ_k (95% CI)	θ_S (SD)	θ_π (SD)	Tajima's *D*	*Fs*	M^a
Southern Andes							
Mapuche (Argentina)	0.908	6.417 (3.245-12.350)	4.730 (1.698)	6.427 (3.455)	0.460	-0.466	10.710
Yaghan	0.886	4.499 (1.717-11.565)	4.613 (1.999)	6.436 (3.635)	0.590	0.910	4.482
Mapuche/Pehuenche	0.875	10.380 (5.924-17.855)	5.833 (1.890)	6.823 (3.616)	-0.091	-3.242	7.245
Lowland Bolivians, Dept. of Beni							
Moxo	0.960	33.312 (15.399-75.066)	7.524 (2.700)	7.329 (3.942)	-0.753	-9.448**	12.726
Yuracare	0.943	11.941 (4.612-31.867)	6.151 (2.560)	7.326 (4.073)	-0.050	-1.479	20.990
Chimane/Moseten	0.926	9.252 (4.024-21.299)	6.201 (2.423)	6.599 (3.634)	-0.414	-1.350	9.815
Movima	0.894	9.317 (3.286-27.264)	3.974 (1.851)	3.319 (2.063)	-1.093	-2.632	15.969
Quechua Beni	0.758	5.696 (2.256-14.228)	5.424 (2.263)	5.163 (2.956)	-0.992	-0.556	0.545
Gran Chaco							
Pilaga	0.963	20.951 (11.213-39.355)	8.092 (2.695)	7.843 (4.147)	-0.692	-7.083*	23.443
Wichi	0.896	9.764 (5.990-15.582)	6.967 (2.018)	6.798 (3.573)	-0.604	-3.568	6.697
Toba	0.869	6.346 (3.511-11.129)	5.865 (1.855)	5.848 (3.138)	-0.487	-0.997	6.132

TABLE VI (continued)

SUMMARY STATISTICS OF LATIN AMERICAN POPULATIONS

	h	θ_k (95% CI)	θ_S (SD)	θ_π (SD)	Tajima's D	Fs	M^a
Other lowland South Americans							
Yanomamo	0.906	12.636 (8.207-19.126)	5.706 (1.645)	5.554 (2.974)	-0.482	-9.653	9.046
Guahibo	0.858	4.272 (2.191-7.997)	2.798 (1.053)	5.800 (3.121)	2.273*	0.935	9.384
Nandeva	0.844	1.892 (0.822-4.067)	2.395 (0.946)	3.876 (2.192)	1.277	2.776	1.649
Zoro/Gaviao	0.842	3.786 (1.894-7.239)	4.554 (1.546)	4.916 (2.695)	-0.248	0.691	3.516
Wayuu	0.773	1.016 (0.342-2.746)	3.310 (1.341)	6.739 (3.641)	2.252*	7.888	2.224
Kaingang	0.744	2.472 ((1.180-4.886)	3.898 (1.313)	6.883 (3.630)	1.323	4.692	2.551
Xavante	0.677	1.085 (0.302-2.963)	2.648 (1.157)	3.474 (2.043)	0.439	3.719	1.873
M'bya	0.652	0.666 (0.193-2.065)	1.339 (0.712)	2.707 (1.661)	2.466**	4.434	0.101
Kaiowa	0.593	1.172 (0.491-2.569)	1.865 (0.719)	1.791 (1.154)	-0.388	1.333	1.153
Ayoreo	0.473	1.927 (0.891-3.898)	1.968 (0.771)	2.761 (1.635)	0.386	0.886	0.439
Ache	0.204	0.485 (0.144-1.445)	1.485 (0.667)	1.241 (0.884)	-0.687	2.566	0.249

* Significant at P < 0.05, ** Significant at P < 0.001, [a] M=2$N_{fe}m$, [b] M could not be estimated.

Among the central Andeans, the Aymara are genetically less diverse than Quechua. The values of θ_k greatly vary from the highest observed in the Ancash to the lowest in the Tupe. The Ancash also has the highest *h*, and the Tayacaja has the highest θ_S. The central Andeans, especially the Aymara, have low HPG diversity due to very high frequency of HPG B, so they have low θ_π values, a parameter that generally reflects more ancient demographic events (Helgason et al. 2003; Helgason et al. 2000; Rogers 1995). Interestingly, very low within-population genetic diversity was observed among the Tupe.

Twelve genetically diverse Latin American populations, including central Andeans (Aymara and Quechua), Quiche Mayans, Pilaga, and Moxo show evidence of population expansion in significantly negative *Fs* values, but no statistically significant negative values of Tajima's *D* were found. Fu's *Fs* statistic calculates the probability of observing a number of haplotypes (*k*) the same or smaller than the observed *k* given an observed θ_π. Excoffier and Schneider (1999) have noted that populations with significantly negative *Fs* often exhibit other genetic evidence of population expansion, while Tajima's *D* is a more conservative test for detecting evidence of population expansion (Aris-Brosou and Excoffier 1996). The Aymara and Quechua have many rare haplotypes and large θ_k value relative to the θ_π value, so they have significantly negative *Fs,* suggesting that they have accumulated an abundance of rare haplotypes and variants through population expansion or genetic hitchhiking.

3.4.3 Identifying the pattern of interactions in Central Andes and western South America

The population pairwise genetic distances were calculated and the genetic relationships among the populations are illustrated on a MDS plot in Fig. 7. The plot illustrates how closely the Aymara and other central Andean groups are related to each other and their relationship to other western and eastern South American populations. The Aché, Ayoreo, and Xavante were included in the initial analysis, but are not included in the plot because their wide dispersion and extremely low within-population genetic diversity suggests that strong genetic drift has resulted in dramatic differentiation from their neighbors.

Although their HPG frequencies greatly vary, the western South American populations, including the central Andean, southern Andean, and lowland populations, all cluster together on the right side of the plot. The south central Andeans (Aymara, Arequipa, and Quechua Puno), who have high frequencies of HPG B, are found in the bottom right quarter of the plot and do not overlap with the north-central Andeans. The La Paz Aymara (ALP) clusters with other the south-central Andean groups and with the north-central Andean Tupe, who speak language related to Aymara. The Quechua from lowland Bolivia have extremely high frequencies of HPG B, but very low genetic diversity overall and are an outlier in this group. On the other hand, the Chibchans and Amazonian lowland populations scatter widely on the left and upper side of the plot.

AMOVA results support the pattern observed on the MDS plot and show similarly close relationships among the central Andeans (TABLE VII). Although the significant *P* values indicate population heterogeneity, the small Φ_{ST} values among the western South American groups suggest these populations are not very differentiated within each group. Lewis et al. (2007b) found smaller Φ_{ST} values among central Andean, southern Andean, and lowland Bolivian groups than among Amazonian populations. My result shows that the pattern remains the same when more populations are added. The south central Andeans are the most genetically homogeneous group and have non-significant Φ_{ST} *P* value. Among the western South

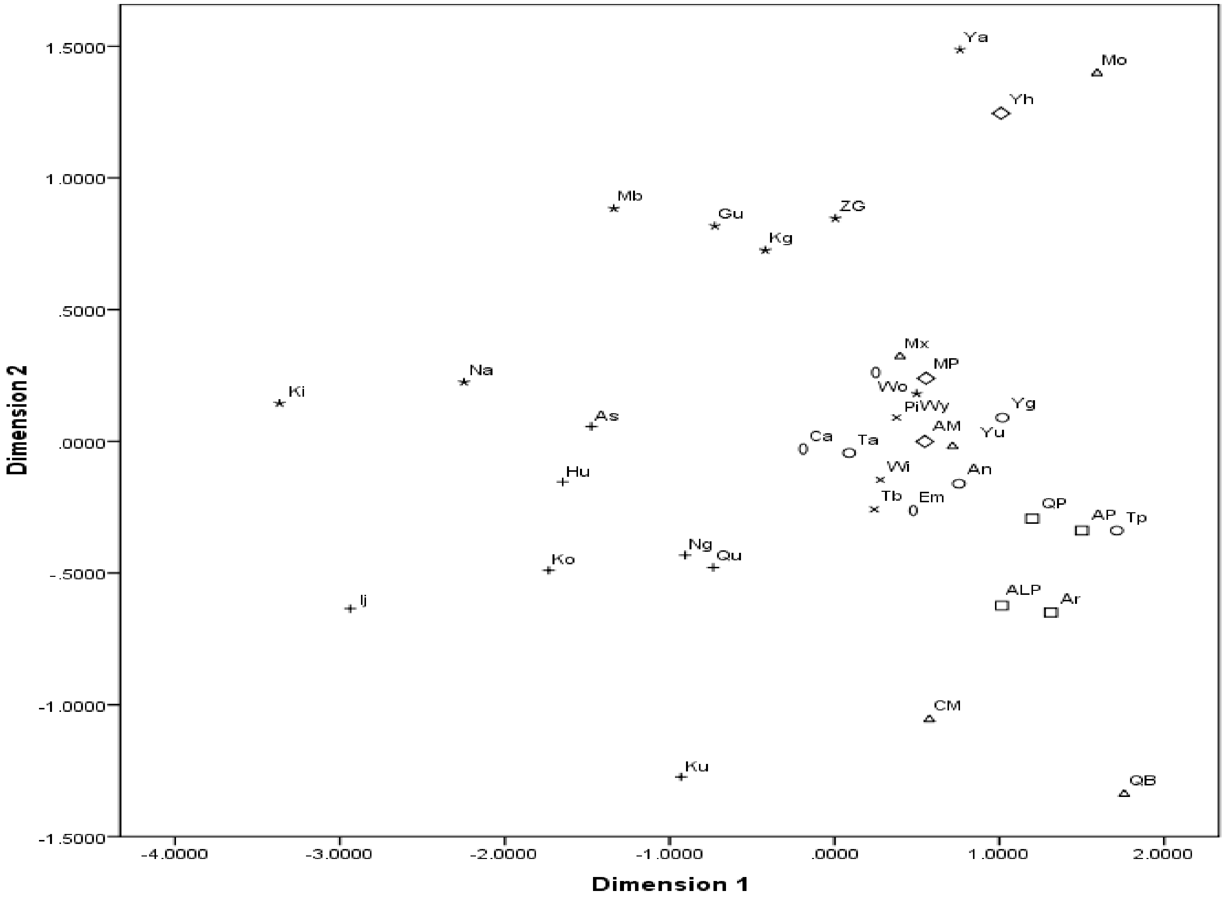

Figure 7. Multidimensional scaling (MDS) plot of Latin American populations. Populations are marked with shape indicating the regions: square (South-Central Andes), circle (North-Central Andes), diamond (Southern Andeans), x (Gran Chaco), triangle (lowland Bolivian, Dept. of Beni), oval (western lowland South Americans), star (other lowland South Americans), and + (Quiche and Chibchans).

TABLE VII

AMOVA RESULTS FOR LATIN AMERICAN POPULATIONS

	Number of Populations	Among Populations Variance (%)	Within Populations Variance (%)	Φ_{ST} (P)
All Andeans	11	6.35	93.65	0.064 (0.000)
Central Andeans	8	3.92	96.08	0.039 (0.000)
South-Central Andeans	4	0.49	99.51	0.005 (0.264)
North-Central Andeans	4	3.62	96.38	0.036 (0.008)
Southern Andeans	3	4.89	95.11	0.049 (0.003)
Gran Chaco	3	2.52	97.48	0.025 (0.011)
North-Western Lowland South Americans	3	8.79	91.21	0.088 (0.000)
Lowland Bolivia, Dept. of Beni	5	19.56	80.44	0.196 (0.000)
Western South Americans (All of above)	22	7.62	92.38	0.076 (0.000)
Chibchan	6	21.28	78.72	0.213 (0.000)

Andes include Central and Southern Andeans, and Central Andeans include South-Central and North-Central Andeans. Individual populations in the groups are followings; South-Central Andes (Aymara La Paz, Aymara Puno, Arequipa, and Quechua Puno), North-Central Andes (Ancash, Tayacaja, Yungay, and Tupe), Southern Andes (Argentina Mapuche, Mapuche/Pehuenche, and Yaghan), Lowland Bolivians (Chimane/Moseten, Movima, Moxo, Quechua Beni, and Yuracare), Gran Chaco (Pilaga, Whichi, and Toba), North-Western lowland South Americans (Embera, Wounan, and Cayapa), and Chibchan (Arsario, Huetar, Ijka, Kogi, Kuna, and Huetar)

Americans, the lowland Bolivians have the highest Φ_{ST} value, indicating greater genetic differentiation and explaining why they do not cluster together on the MDS plot**.**

The most genetically diverse populations tend to the large migration rates (M=2$N_f m$) (TABLE VI), and M was correlated with θ_k (P=0.001) and θ_S (P=0.003), but not with θ_π (P=0.165). The genetically diverse populations, such as the central Andeans, Quiche Mayans, Pilagá, and Moxo have large M, and these genetically diverse South American populations tend to cluster together on the MDS plot. The populations with low genetic diversity, including the Chibchan and Amazonian populations have low M, and they scatter on the plot. The migration rate could not be calculated for three South-central Andean populations (Arequipa, Aymara Puno, and Aymara La Paz), possibly because of their poor fit to the spatial expansion model. To further investigate effects of gene flow, isolation-by-distance model was evaluated with the Mantel test of 15 populations used for MIGRATE analysis, and the result shows significant correlation between population pairwise genetic distances (Φ_{ST}) and geographical distances (P = 0.025).

3.4.4 mtDNA variation of Aymara vs. Quechua

I examined the sequence diversity of each HPG found in the Central Andes to investigate whether all HPGs expanded equally or if only HPG B, the most common haplogroup, was affected (TABLE VIII). HPGs A and B are more diverse than HPG C and D, but large negative Tajima's D and Fu's F_S with significant P values indicate that all the HPGs show evidence of expansion. The mismatch distributions generally conform to the sequence diversity estimates (Fig. 8). All the HPGs have unimodal distributions showing evidence of expansion. HPG D

.

TABLE VIII

SUMMARY STATISTICS OF MITOCHONDRIAL DNA HAPLOGROUPS SEQUENCE DIVERSITY AMONG THE CENTRAL ANDEANS[a]

Haplogroups	n	h	θ_k (95% CI)	θ_S (SD)	θ_π (SD)	Tajima's *D*	*P-Value*	*Fs*[e]
Central Andes A	28	0.982	76.628 (31.732-200.984)	7.966 (2.818)	3.718 (2.154)	-2.116	0.003	-25.01
Central Andes B	146	0.969	67.350 (48.251-94.010)	12.236 (3.103)	2.585 (1.542)	-2.510	0.000	-26.949
Central Andes C	58	0.934	19.049 (11.289-31.936)	5.617 (1.831)	2.745 (1.638)	-1.770	0.021	-23.07
Central Andes D	38	0.859	12.761 (6.766-23.873)	4.522 (1.644)	2.162 (1.363)	-1.839	0.014	-13.551
Aymara[b] B	48	0.929	18.438 (10.472-32.336)	5.408 (1.831)	2.136 (1.356)	-2.056	0.004	-23.299
N.C. Quechua[c] B	54	0.940	35.092 (20.564-60.348)	7.900 (2.478)	2.644 (1.590)	-2.311	0.000	-26.789
S.C. Quechua[d] B	33	0.970	38.080 (18.899-79.299)	6.406 (2.258)	2.905 (1.740)	-2.026	0.007	-24.523

[a] Central Andeans includes Aymara and Quechua populations from highland Bolivia and Peru.

[b] The Aymara includes the Aymara from La Paz and Aymara Puno

[c] The Quechua from North-Central Andes includes Ancash, Tayacaja, and Yungay.

[d] The Quechua from South-Central Andes includes Arequipa and Quechua Puno.

[e] Fu's F_S is significant with *P*-value < 0.001 for all of the HPGs analyzed.

has the lowest θ estimates and has a peak at one nucleotide difference, while HPG A, which has the largest θ_k and θ_π values, has a peak at three nucleotide differences.

The HPG B sequences were analyzed separately for the Aymara and Quechua and then compared to assess similarities in genetic variation and demographic history. HPG B is slightly more diverse in the Quechua than in the Aymara, but the difference is not large and HPG B shows evidence of expansion in both groups. The Aymara and Quechua have similar HPG B with peaks at two nucleotide differences at similar frequency (~0.3), but the Aymara have more haplotypes with zero or one nucleotide difference and fewer haplotypes with four to six nucleotide differences than the Quechua.

The $\Theta = 2N_f\mu$ of regional/linguistic groups and migration rates ($2N_fm$) between each group were estimated using MIGRATE. The Quechua's effective population size is larger than other linguistic/regional groups analyzed, and the Aymara have the second largest estimate of Θ (TABLE IX). The migration rates estimated between these populations are large. The migration rate between Aymara and Quechua is the largest, but the Quechua have very large effective population size and high migration rate with all the groups. Compared to the Quechua, the Aymara have smaller migration rates between them and other groups.

3.5 Discussion

The mtDNA variation strongly supports the ethnohistorical evidence for an Aymara expansion. The comparison of sequence diversity among the Latin American populations, the Fu's *Fs* values, the sequence diversity estimated for each mtDNA HPG, and mismatch distributions clearly indicate that the Aymara and Quechua populations are highly diverse and

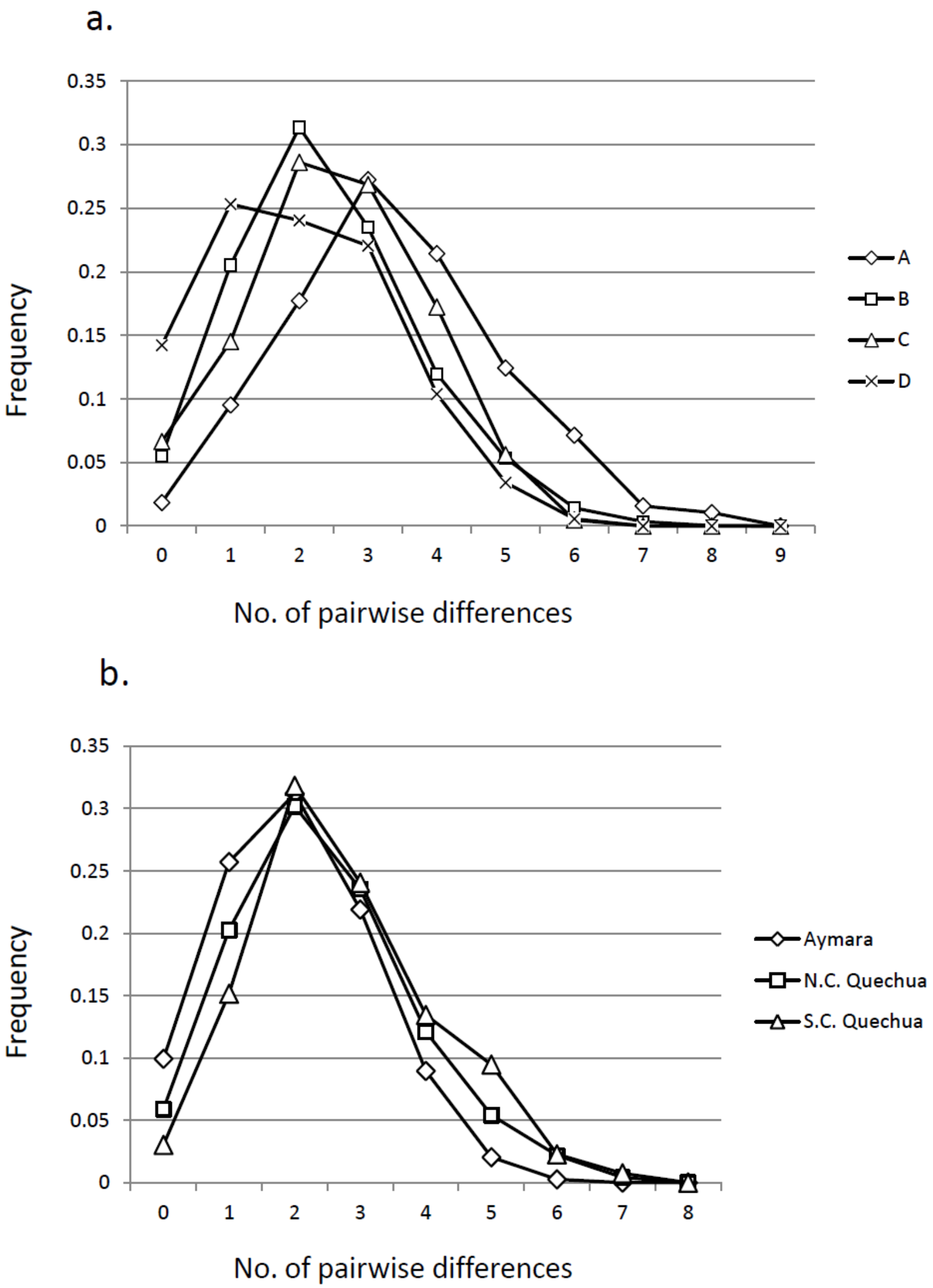

Figure 8. Mismatch distributions of mtDNA haplogroups. Fig. 8a compares the mismatch distribution of four major mtDNA haplogroups among the Central Andeans. Fig. 8b compares mismatch distribution of Aymara, North-Central Quechua, and South-Central Quechua mismatch distribution.

TABLE IX

Θ AND MIGRATION RATES ($2N_fm$) BETWEEN LINGUISTIC/REGIONAL GROUPS OBTAINED USING MIGRATE

	n	Freq. FH[1]	Θ	Aymara	Quechua	S. Andes	L. Bolivians	Gran Chaco
Aymara	75	0.240	0.0740					
Quechua	182	0.209	0.4798	195.9752				
S. Andes	97	0.340	0.0106	13.4695	42.4734			
L. Bolivians	74	0.338	0.0577	36.7992	63.9783	15.4234		
Gran Chaco	204	0.319	0.0086	25.3572	41.7774	7.7331	10.9738	
NWLSA	104	0.269	0.0137	20.4791	72.6017	7.8300	15.9892	6.3225

The estimates are average of five separate runs. [1] Total frequency of the founder haplotype of each of four major haplogroups Aymara (Aymara La Paz and Aymara Puno), Quechua (Ancash, Arequipa, Quechua Puno, Tayacaja, and Yungay), Southern Andes (Argentina Mapuche, and Mapuche/Pehuenche), Lowland Bolivians, Dept of Beni (Chimane, Moseten, Movima, Moxo, and Yuracare), Gran Chaco (Pilaga, Whichi, and Toba), North-Western lowland South Americans (Cayapa, Embera, and Wounan).

have recently experienced population expansion. Here, I examine the genetic signatures of expansion to evaluate whether a demographic, range expansion explains, or natural selection is a likely explanation for their expansion.

3.5.1 Evidence for demographic expansion

First, it should be noted that the Quechua and Aymara have large θ_k and θ_S that reflect recent demographic history and Θ estimated with MIGRATE. These populations accumulated rare variants relatively recently as the population size increased, possibly because of introduction of intensive agriculture. The program, MIGRATE, is designed to estimate $\Theta=2N_f\mu$, accounting for the gene flow took place with neighboring demes, or communities, so larger estimate of Θ in the Aymara and Quechua population compared to other Latin American populations suggests large female effective population size in these Andean populations. The chapter 5 further evaluates whether demographic or spatial expansion model explain the observed mtDNA variation in Latin American populations and differences in estimated Θ. The findings in chapter 5 support the data presented in this chapter.

The Titicaca Basin is highly productive area, where large variety of agricultural products were grown, the domesticated camelids are grown, and fish from the lake were caught (Stanish 2003). Agricultural products were grown in the area with rich soil, such as valley bottom and lake shore, as well as in the area with less productive soil, such as terraced field on the hillside. Along the shore of the Lake Titicaca and along the river, raised fields were constructed. The raised fields were labor-intensive construction, but the raised field, camels, and other agricultural technology were highly productive to support the large population size in the Altiplano (Erickson 1988).

3.5.2 Evidence for range expansion

Although the several of the analyses suggest that the expansion can be explained by a demographic expansion, the other analyses indicate that gene flow among the Aymara and Quechua was an important factor as well. The isolation-by-distance model was tested using the Mantel test, which examines the correlation between population pairwise genetic distance and geographical distance. The isolation-by-distance model predicts that human are more likely to find marriage partners from their neighboring communities, and the significant correlation between genetic and geographical distances indicates that there was constant gene flow with neighboring communities. The interaction with the neighboring populations seems to be more intense among the south central Andeans than other groups, and the south central Andeans have small AMOVA among population variance and Φ_{ST}. Central Andeans also have considerably larger values of migration rates (M) estimated using mismatch distribution and migration rates ($2N_fm$) estimated using MIGRATE. These findings suggest that they show genetic evidence of population expansion, partly due to high migration rates (Ray et al. 2003). These results agree with the result of simulation analysis examining effect of gene flow and female effective population size conducted by Fuselli and colleagues (2003). Their simulation analysis shows that mtDNA variation in the Andes can be explained with dense population living in the area and high migration rates before and after the European contact.

MIGRATE and MDS analyses further suggest that network of interaction or gene flow extends beyond Central Andes to the transition area between Andes and Amazon along the western side of South America (Corella et al. 2007). Although large effective population size, sharing founding haplotype of each HPG, unsampled populations, and small divergence time can inflate migration rates, migration rates between the Central Andeans and lowland populations are large.

3.5.3 Evidence for natural selection

Researchers suggested that mitochondrial genome has been naturally selected in human populations (Elson et al. 2004; Ingman and Gyllensten 2007; Mishmar et al. 2003; Ruiz-Pesini et al. 2004), but it has not been empirically demonstrated. Because T→C mutation at np 16189, which creates a 16184-16193 poly-C tract length polymorphism, is associated with high BMI and type 2 diabetes in Asian populations (Kim et al. 2002; Liou et al. 2007; Park et al. 2008), the distribution of this mutation was explored. Recurring T→C mutation at np 16189 in different HPG backgrounds was observed among the South American populations. While low altitude populations tend to have this mutation in only one HPG background beside HPG B, this mutation is observed in multiple HPG backgrounds among the highland Andeans. In the cold highland Andean environment, thrifty genotype/phenotype, cold climate adaptation, and adaptation to hypoxia could be some of selective mechanisms. For example, the thrifty genotype hypothesis predicts that hunter-gatherers who have genetic variations that allow them to store fat more efficiently would have a selective advantage, especially in the cold, high altitude Andean environment. The 16189 mutation is common among the New World populations, and obesity, diabetes, and other metabolic diseases are becoming a major health concern among Native Latin Americans, including the Aymara and Quechua (Barceló et al. 2001; Lindgärde et al. 2004; Mohanna et al. 2006). Therefore, it is possible that mtDNA among the highland Andeans exhibit a great genetic diversity with evidence of past population expansion, partly because of hitchhiking effect, accumulation (and increase in the frequency) of rare or low frequency variants that linked to positively selected loci. However, the association between this mutation and metabolic diseases has not been investigated in these populations.

However, the mutations that disrupt the poly-C tract between np 16184-16193 were also observed. The haplotype B05 and B10 have a mutation at np 16188, and 16189 mutation is paired with a 16187C→T substitution among the southern Andeans. These nucleotide positions are also mutational hot spots (Excoffier and Yang 1999; Hasegawa et al. 1993; Meyer et al. 1999; Wakeley 1993) and this mutation is observed in multiple HPG backgrounds, possibly because large enough samples have been collected from central Andean populations to capture their great genetic diversity during the analyses.

3.5.4 <u>Evolutionary history of the Aymara vs. Quechua</u>

No significant genetic differences were found between the south central Aymara and Quechua groups. These genetically diverse populations have high frequencies of HPG B and small AMOVA Φ_{ST}, and they cluster together on the MDS plot. They also have similar HPG B sequence diversity and mismatch distributions. Despite their different languages and histories, the genetic data suggests that the south central Aymara and Quechua groups belong to a single deme, in which members of different ethnic groups intermarried.

The Aymara and north central Quechua also show some similar genetic patterns. The north-central Quechua are slightly more diverse and their HPG B sequences show slightly more diversity than that of the Aymara, but overall they are both genetically more diverse than other Latin American populations and the HPG B mismatch distribution show similar pattern. If we assume that natural selection had a minor effect on mtDNA variation, similarity in genetic diversity suggests that they both experienced similar demographic expansion. Alternatively, extensive interaction and gene flow can result in a similar genetic diversity. Excoffier (2004) suggested that two populations from the same spatial expansion wave will have similar genetic diversity, so the similar HPG B sequence variation observed among the Aymara and Quechua may have been the result of long term interactions. In fact, all of the Aymara and Quechua population samples have large *M* values and the migration rates estimated between them are very high.

Geography could have played a role more than language influencing mtDNA variation, and the south central Aymara and Quechua populations differ from the north central Quechua populations. The north central Andean tends to be genetically more diverse than south central Andeans (including the Aymara and Quechua), and they form a separate cluster on the MDS plot. The Quechua language today includes many different dialects and has been spread over a wider area by both the Inka and the Spanish, which has likely increased the diversity of the people who have adopted the language. North central Andeans may have interacted more often with lowland populations than the south central Andeans. North central highland and lowland interactions are evident in the archaeological record by the Early Horizon, when the Chavín culture flourished between 900 BC and 200 BC (Burger 1995), and exotic items have been found in many archaeological sites from Peru, especially grave sites (Alva 2001). Migration rates estimated using MIGRATE between the Quechua and other linguistic/regional groups are very high and the north central Andeans have higher frequencies of the A, C and D HPGs than the south central populations.

3.6 Conclusions

The mtDNA data suggest that both Aymara and Quechua experienced population expansion, most likely because of rapid demographic expansion after introduction of intensive agriculture or a selective advantage. My results also suggest that female gene flow was an important factor homogenizing mtDNA variation among the Central Andeans as well. There were constant movements of people within Central Andes and into the transitional zones. In this region, language does not correlate with ancestry nearly as well as geography does and the south central Aymara and Quechua are virtually indistinguishable from each other. The evidence for natural selection on the mtDNA genome, and a mitochondrial adaption to a cold, high altitude environment is suggestive, but requires further investigation.

4. MITOCHONDRIAL DNA DIVERSITY IN TWO ETHNIC GROUPS IN SOUTHEASTERN KENYA: PERSPECTIVES FROM THE NORTHEASTERN PERIPHERY OF THE BANTU EXPANSION

4.1 Introduction

Until recently, theories about the Bantu expansion throughout sub-Saharan Africa relied most heavily on the fields of linguistics and archaeology. Linguists use various lines of evidence to argue that the Bantu languages originated in central Africa, in northern Cameroon, and spread relatively rapidly from central Africa to the eastern and southern sub-Saharan Africa (Ehret 2001; Holden 2002; Rexová et al. 2006). In the early study of Bantu language, Guthrie (1962) outlined the Bantu prehistory. The proto-Bantu language shared similar characteristics with many West African languages, so the proto-Bantu language originated in the western Central Africa. From there, the western Bantu languages were, first, separated from the proto-Bantu languages, and then the East Bantu languages diverged. The proto-Bantu speakers did not have knowledge of iron-working, but acquired probably in the Chad region.

Recent phylogenetic studies of Bantu languages support Guthrie's view of the Bantu expansion (Holden 2002; Rexová et al. 2006), and linguists classify Bantu languages geographically into Western and Eastern Bantu groups, which they believe corresponds to the migration routes that they took as they colonized southern and eastern Africa (see Fig. 9 routes A and B). Holden and Gray's (2006) phylogenetic analysis shows that the eastern Bantu languages are more homogeneous than the western Bantu languages, which they interpret as a indication

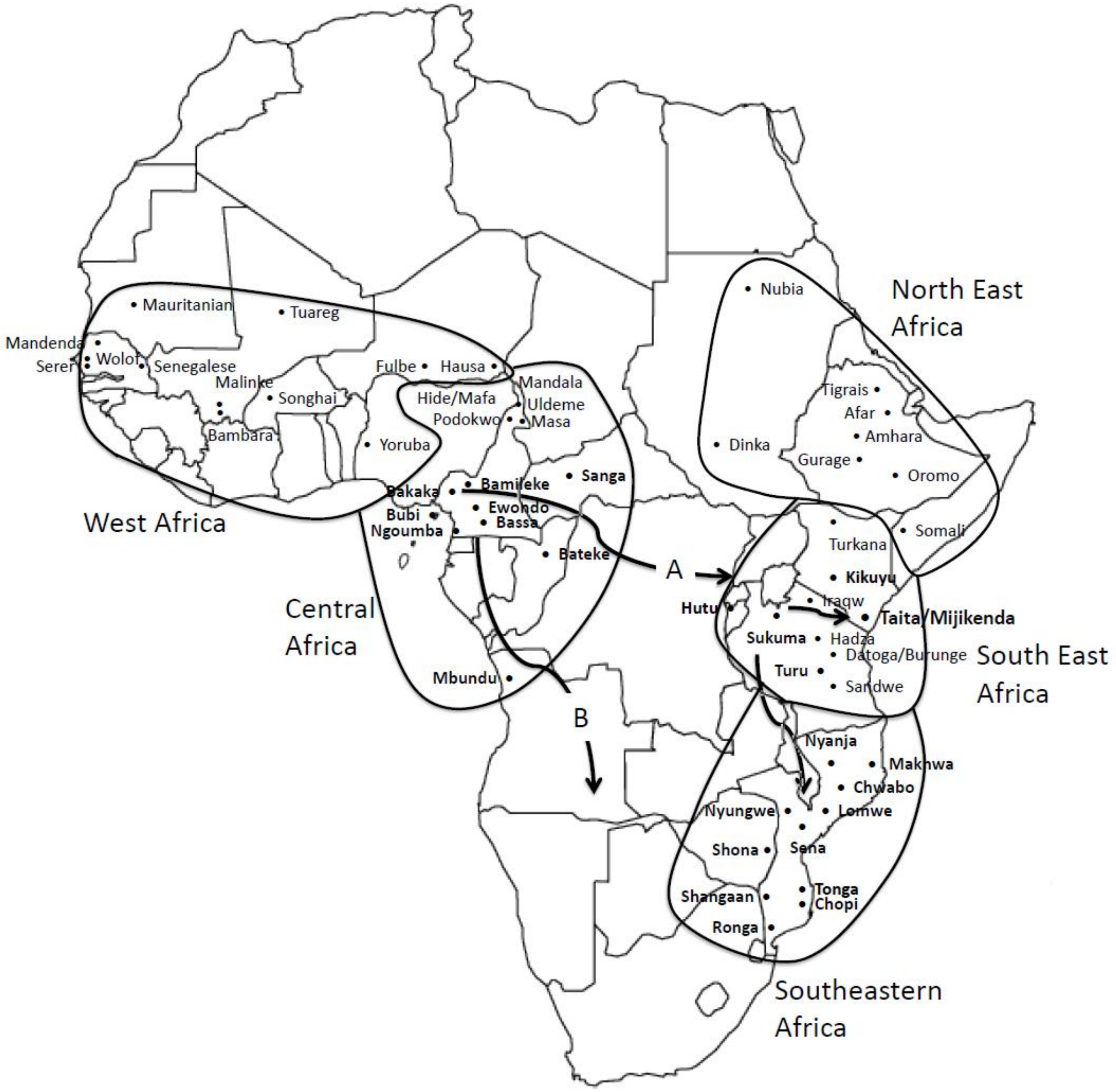

Figure 9. Map showing the location of the Bantus (bold) and non-Bantu populations used for analyses with geographical groupings. A and B in the map correspond to east and west route of Bantu expansion.

of either recent and rapid expansion or of extensive language borrowing among the eastern Bantu speaking groups.

Along the eastern half of Africa, the distribution of eastern Bantu languages overlaps with the distribution of the archaeological Chifumbaze Iron Age cultural complex (Phillipson 2005). Chifumbaze ceramic traditions are derived from the Urewe ware that first appeared around Lake Victoria in east Africa around 500 B.C. The people associated with the Chifumbaze complex culture were agropastoralists who used iron technologies that were absent in Late Stone Age forager cultures already present in the area. Consequently, many archaeologists believe that the Bantu languages spread from east Africa into southeastern Africa as a part of a cultural complex that includes distinctive ceramic, farming and iron technologies. Incorporating the archaeological dates with the linguistic data suggests that Bantu-speaking people began migrating into East Africa approximately 3,000 years ago, and then moved down into southeastern Africa around 2,000-3,000 years ago (Holden 2002; Phillipson 2005).

Although many researchers have assumed that the expansion involved the physical movement of people, the process of Bantu expansion (migration routes, number of expansion waves, and incorporation/replacement of pre-existing forager populations or continuity of local cultures) remains debated, and some have acknowledged that the Bantu languages could have spread because of language shift and borrowing (Holden 2002; Nurse 1997; Robertson and Bradley 2000; Vansina 1995). Vansina (1995) argues that there were at least nine expansion waves. In each expansion wave, languages were geographically dispersed and differentiated. Some of the expansion waves involved the migration of people, while other expansion waves were limited to the adoption of Bantu languages by non-Bantu speakers. Interactions between Bantu and non-Bantu speaking groups undoubtedly varied throughout the region. Schoenbrun

(1993) proposed that Bantu speakers in the Great Lakes region incorporated the food producing practices of Nilo-Saharan Sudanic language speakers and that non-Bantu speaking societies were culturally assimilated into Bantu societies, possibly through inter-ethnic marriage, and eventually adopted Bantu language and cultural practices. In other areas, archaeological evidence suggests that the Late Stone Age cultural traditions continued into the Early Iron Age, with foragers coexisting with Bantu farmers (Robertson and Bradley 2000).

Genetic studies generally support a model of a large-scale migration of people bringing their Bantu language and culture with them as they colonized new territories. Bantu-speaking populations today are genetically homogeneous (Cavalli-Sforza et al. 1994; Excoffier et al. 1987; Salas et al. 2002). They tend to have small genetic distances resulting in tight clustering on PC and MDS plots and small among populations variance. Researchers also have identified mtDNA and Y chromosome HPGs that are shared in high frequencies in Bantu speaking groups from the western central Africa to southeastern Africa (Castrì et al. 2009; Pereira et al. 2002; Salas et al. 2002). The places of origin for the mtDNA and Y chromosome HPGs were inferred from these frequencies and researchers (Scozzari et al. 1999; Underhill et al. 2001; Wood et al. 2005) propose that the wide distribution of Y chromosome HPG E3a (E-M2) is largely result of a Bantu population expansion. This HPG is common in many sub-Saharan populations, but central African Bantu populations have the highest frequencies (Berniell-Lee et al. 2009; Rosa et al. 2007).

The mtDNA HPGs, L1c and L3e, that are more common in West and Central African, show evidence of past population expansion with star-like network and unimodal mismatch distribution (Batini et al. 2007; Salas et al. 2002). Salas and colleagues (2002) analyzed mtDNA variation in Africa, and argue that, in addition to HPGs common in West and Central Africa (e.g., L1c and L3e), other HPGs (e.g., L0a from East Africa) were brought to southeastern Africa by Bantu migrants as well. They also conclude that Khoisan speakers contributed very little mtDNA (~5%) to the gene pool of modern Bantu-speaking populations in sparsely inhabited southeastern Africa and that Bantu-speaking migrants provided the majority of the mtDNA genetic variability there.

Although interest in the genetic diversity is intense and the number of samples is increasing quickly, sampling density is still relatively low in east Africa given how it population density and the demographic, ethnic, and linguistic diversity of the region. All four major African language families: Niger-Congo (Bantu), Nilo-Saharan, Afro-Asiatic, and Khoisan are spoken here; often by people who reside next to each other. Genetics studies of other east African populations indicate great genetic diversity as well (Cruciani et al. 2004; Hassan et al. 2008; Kivisild et al. 2004; Semino et al. 2002; Tishkoff et al. 2007; Watson et al. 1996). Kittles and Weiss (2003) have argued that our understanding Africa’s complex evolutionary history has been hampered by the fact that previous genetic studies tended to exclude east African populations because they are somewhat intermediate between sub-Saharan African and Eurasian populations.

In this project, I performed mtDNA analyses of the Taita and Mijikenda, two east African populations from southeastern Kenya, to identify potential female gene flow among Bantu-speaking populations and between Bantu and non-Bantu speaking populations in East Africa. I used these population samples to explore three different models or scenarios into Kenya on the northeastern periphery of the Bantu expansion. All three models assumed a rapid expansion of people who spoke Bantu languages and shared the agricultural and metalworking technologies identified by archaeologists. The models differ mainly in the amount of interactions the Bantu-speaking groups have with each other and with the non-Bantu speaking groups they encounter as they colonize east Africa. Although cultural diffusion undoubtedly played a role in the spread of Bantu languages and culture in some parts of sub-Saharan Africa, a model solely on cultural diffusion was not tested because it cannot account for the genetic similarities among geographically separated Bantu-speaking populations and the wide distribution of West and Central African HPGs in east and southeastern Africa observed in previous genetic studies (Salas et al. 2002; Tishkoff et al. 2007).

The following three models assume that west-central African people carried Bantu languages and cultural traits into new areas, but differ in the extent to which Bantu speakers interacted with each other and with non-Bantu-speaking people. Although the real expansion process was undoubtedly more complex, comparing simplified models allows us to compare the relative importance of the effects of isolation, within-group interaction, and interaction with non-Bantu populations on genetic variation in Bantu-speaking populations and demographic history. Each model predicts different within-population genetic diversities and patterns of population subdivision which can be observed in the Taita, Mijikenda, and other east African mtDNA variation examined in this project.

Model 1: Expansion without gene flow

This model resembles the more traditional view of Bantu expansion, where proponents conceptualize Bantu-speaking groups as rapidly expanding into new areas with little or no interaction with each other or with non-Bantu-speaking groups. In this model, the Bantu speakers outnumbered pre-existing forager populations and replaced them (Phillipson 2005), so they became genetically less diverse and more differentiated as they colonized new territories.

Model 2: Expansion with gene flow among Bantu-speaking populations

In this model, Bantu-speaking groups maintained contact with other Bantu groups in the core and nearby as they colonized new areas. As suggested by Cavalli-Sforza and others (Barbujani and Bertorelle 2001; Cavalli-Sforza et al. 1994; Cavalli-Sforza et al. 1988; Chikhi et al. 2002; Piazza et al. 1995; Sokal et al. 1991), linguistic barriers reduced the rate of gene flow among populations who spoke languages belonging to different linguistic families, so the expanding Bantu-speaking populations exchanged genes mostly with each other. This model predicts that central and east African Bantu-speaking groups will be relatively homogeneous with similar levels of diversity. It also predicts that Bantu and non-Bantu populations that live near each other will not resemble each other genetically.

Model 3: Expansion with gene flow with neighboring non-Bantu-speaking groups

In this model, Bantu-speaking groups interacted with non-Bantu-speaking groups as they colonized new areas (Schoenbrun 1993). If this model is correct, the Bantu-speaking populations in east Africa will resemble the non-Bantu speaking groups around them, and the gene flow will result in genetic diversity either equal to or greater than the genetic diversity observed in the west central Bantu-speaking populations living in the core territory.

4.2 Samples and methods

4.2.1 Samples

The mtDNA HVRI of 352 individuals from the Taita and Mijikenda, both Bantu-speaking agropastoralist groups (see Chapter 1), was sequenced using the methods described in the previous chapter (Chapter 3). The Taita live in the hills of the same name located about 150 km west of the port city of Mombasa. The Taita include three groups; the Davida, the Sagalla and the Kasigau, and their population size is small (213,000). The Mijikenda reside in the southeastern coastal region of Kenya, in and around Mombasa. The Mijikenda are composed of nine tribes (Digo, Duruma, Giriama, Jibana, Kambe Kauma, Rabai, and Ribe).has population size of 1,208,000.

Their mtDNA variation was compared to that of 58 published population samples from sub-Saharan Africa including 24 Bantu, 6 Nilo-Saharan, 17 Afro-Asiatic, 9 non-Bantu Niger-Congo, and 2 Khoisan groups (Fig. 9 and TABLE X). The 24 Bantu populations include 4 East African groups, 9 Central African groups, and 11 Southeastern African groups. I sequenced one Kikuyu sample and included it with the Kikuyu samples analyzed by Watson et al. (1996).

4.2.2 <u>Analytical methods</u>

The analyses used in this project are grouped into four categories; 1) HPG frequencies of Bantu and non-Bantu populations, 2) within-population genetic diversity and signature of expansion, 3) population differentiation, and 4) spatial patterning.

1) HPG frequencies of Bantus and non-Bantu populations

Based on mtDNA HVR sequence, HPG was assigned and the origin of mtDNA HPG proposed by Sales et al. (2002) and Kivisild et al. (2004) were used to evaluate the differences in frequencies between the Bantu and non-Bantu population. Haplotype sharing between East African Bantu and non-Bantu populations were examined using Network program (Bandelt et al. 1999).

TABLE XI

AFRICAN POPULATION INFORMATION CATEGORIZED BASED ON GEOGRAPHICAL LOCATION AND LANGUAGE

Populations	Abbr.[a]	n	References
Present Study			
Taita[b]	Taita	157	(Babrowski, 2007; Present Study)
Mijikenda[b]	Mijikenda	195	(Babrowski, 2007; Present Study)
East Africa - South	(Kenya, Tanzania, Burundi, and Rwanda)		
Bantus			
Kikuyu (Gikuyu)	Ki	25	(Watoson et al. 1997; Present Study n=1)
Sukuma	Su	32	(Knight et al. 2003; Tishkoff et al. 2007)
Hutu[b]	Hu	42	(Castrì et al. 2009)
Turu (Nyaturu)[b]	Tu	29	(Tishkoff et al. 2007)
Nilo-Saharan			
Turkana	Tk	37	(Watson et al. 1997)
Datoga	Da	57	(Knight et al. 2003; Tishkoff et al. 2007)
Afro-Asiatic			
Burunge	Bu	38	(Tishkoff et al. 2007)
Iraqw	Iq	12	(Knight et al. 2003; Tishkoff et al. 2007)
Khoisans			
Sandawe		82	(Tishkoff et al., 2007)
Hadza		96	(Knight et al. 2003; Tishkoff et al. 2007; Vigilant et al. 1991)
East Africa - North	(Ethiopia, Somalia, and Sudan)		
Nilo-Saharans			
Dinka	Di	46	(Krings et al. 1999)
Nubia	Nu	82	(Krings et al. 1999)
Afro-Asiatic			
Garages	Gu	21	(Kivisild et al. 2004)
Tigrais	Ti	46	(Kivisild et al. 2004)
Oromo	Om	30	(Kivisild et al. 2004)
Amhara	Am	88	(Kivisild et al. 2004)
Somali	So	24	(Watson et al. 1997)
Afar		13	(Kivisild et al. 2004)

TABLE XI (Continued)

AFRICAN POPULATION INFORMATION CATEGORIZED BASED ON GEOGRAPHICAL LOCATION AND LANGUAGE

Populations	Abbr.[a]	n	References
Southeastern Africa			
Bantu			
Ronga[b]	Ro	22	(Salas et al. 2002)
Shona[b]	Sh	59	(Castrì et al. 2009)
Nyungwe[b]	Nw	20	(Salas et al. 2002)
Shangaan (Tsonga)[b]	Sg	22	(Salas et al. 2002)
Chwabo[b]	Cb	20	(Salas et al. 2002)
Chopi	Cp	27	(Salas et al. 2002)
Tonga	Tn	20	(Salas et al. 2002)
Nyanja	Nj	20	(Salas et al. 2002)
Makhwa[b]	Mk	20	(Salas et al. 2002)
Lomwe[b]	Lo	20	(Salas et al. 2002)
Sena[b]	Sn	21	(Salas et al. 2002)
Central Africa			
Bantus			
Bassa[b]	Bs	46	(Destro-Bisol et al. 2004)
Ngoumba[b]	Ng	44	(Batini et al. 2007)
Mbundu	Mb	43	(Plaza et al. 2004)
Bamileke [b]	Bm	48	(Destro-Bisol et al. 2004)
Ewondo[b]	Ew	53	(Destro-Bisol et al. 2004)
Bakaka[b]	Bk	50	(Destro-Bisol et al. 2004)
Sanga	Sa	30	(Batini et al. 2007)
Bateke (Teke)[b]	Bt	50	(Batini et al. 2007)
Bubi[b]	Bb	45	(Mateu et al. 1997)
Non- Bantu Niger-Congo			
Fali		41	(Coia et al. 2005)

TABLE XI (Continued)

AFRICAN POPULATION INFORMATION CATEGORIZED BASED ON GEOGRAPHICAL LOCATION AND LANGUAGE

Populations	Abbr.[a]	n	References
Afrio-Asiatic			
Hide		23	(Černy et al. 2004)
Uldeme (Wuzlam)		28	(Coia et al. 2005)
Podokwo (Parkwa)		39	(Coia et al. 2005)
Mandara (Wandala)		37	(Coia et al. 2005)
Masa		31	(Černy et al. 2004)
Mafa		32	(Černy et al. 2004)
Western Africa			
Non-Bantu Niger-Congo			
Yoruba		33	(Vigilant et al. 1991; Watson et al. 1997)
Bambara		52	(González et al. 2006)
Senegalese		50	(Rando et al. 1998)
Serer		23	(Rando et al. 1998)
Wolof		48	(Rando et al. 1998)
Malinke		31	(González et al. 2006)
Fulbe		60	(Watson et al. 1997)
Mandenka		119	(Graven et al. 1995)
Nilo-Saharan			
Kanuri		14	(Watson et al. 1997)
Songhai		16	(González et al. 2006; Watson et al. 1997)
Afro-Asiatic			
Hausa		20	(Watson et al. 1997)
Tuareg (Tamahaq)		24	(González et al. 2006; Watson et al. 1997)
Mauritanian		64	(González et al. 2006)

[a] abbreviation used for the MDS plot

[b] Populations used for comparison of Theta using MIGRATE analysis in Chapter 5.

2) Within-population genetic diversity and signature of expansion

Estimates of within population genetic diversity, haplotype diversity (*h*) and parameter $\theta=2N_{fe}\mu$ (θ_k, θ_S and θ_π) and two tests of molecular neutrality (Tajima's *D* and Fu' *Fs*) were conducted using Arlequin population genetics software program (Excoffier et al. 2005; Schneider et al. 2000). While ancient demographic history affects mean pairwise nucleotide differences (π) most, the two other θ estimators (θ_k, and θ_S) based on the number of alleles (*k*) and polymorphic sites (*S*), are more sensitive to recent demographic events (Helgason et al. 2003; Helgason et al. 2000; Rogers 1995; Tajima 1989a). Arlequin was also used for mismatch distribution analyses. The spatial expansion model of mismatch distribution predicts that two populations from the same spatial expansion wave have similar mismatch distributions (Excoffier 2004).

3) Population differentiation

Population pairwise genetic distances (Φ_{ST}), Exact Tests of population differentiation, and Analysis of Molecular Variance (AMOVA) were performed in Arlequin. Population pairwise genetic distances were visualized with Multidimensional scaling (MDS) analysis using SPSS statistical software.

4) Spatial patterns

The correlations between genetic diversity (*h*, θ_k, θ_S, and θ_π) and distance from the center of Bantu expansion was examined. In the one-step expansion model, I used the direct distance from the center of expansion to the approximate central location of each Bantu population. Recent phylogenetic analyses of Bantu languages suggest western Central African origin of the Bantu (Holden 2002; Rexová et al. 2006), so I used Douala, a coastal city in northwestern Cameroon, as the center of expansion. However, the expanding Bantu populations expanded toward East Africa, then to southeastern Africa (Phillipson 2005), so for the two-step model, using the southern tip of the Lake Victoria as the center of the East Bantu expansion, the distance from Douala to the southern tip of the Lake Victoria, and then from there to approximate central location of each East Bantu population was calculated. Mantel tests were undertaken in Arlequin to test if the isolation-by-distance model explains the observed mtDNA variation. Arlequin was also used to calculate migration rate, $M=2N_{fe}m$, under spatial expansion model, and MIGRATE (Beerli and Felsenstein 2001) provides estimate of $2N_{fe}m$ between each pair of populations analyzed.

Different genetic variation among the Bantu-speaking populations in the northeastern periphery of the expansion is expected from the three models of Bantu expansions examined in this project. The TABLE XI summarizes the expected results of analyses under each Bantu expansion model.

Model 1: Expansion without gene flow

If this model holds, then the Bantu-speaking populations experienced a demic expansion, expanding rapidly with little gene flow and the populations at the periphery experienced series of founder effects. Consequently, these colonizing populations should show reduced within-population genetic diversity and increased between group heterogeneity (Austerlitz et al. 1997; Currat and Excoffier 2005).

We can expect to observe high frequencies of Central and West African HPGs and frequencies of each vary widely among eastern Bantu-speaking groups. Contrary, these Central African and West African HPGs should be rare in eastern non-Bantu-speaking populations. East African Bantu populations should be genetically less diverse than Central African Bantu populations, so the genetic diversity should negatively correlate with the distance from the center of Bantu expansion. They should have small migration rate, so there should be large distances among the East African Bantu-speaking populations and genetic distance should not correlate with geographical distances. Also, they should not cluster near Central African Bantu populations on the MDS plot.

Model 2: Expansion with gene flow among Bantu-speaking groups

When the spatially expanding populations exchange genes among themselves, even peripheral populations do not experience founder effects, but they show evidence of population expansion (Ray et al. 2003; Wegmann et al. 2006). High within-population genetic diversity, genetic homogeneity among the Bantu-speaking groups, and genetic similarity between East and Central African Bantu populations are expected.

High frequencies of Central and West African HPG should be found among eastern Bantu-speaking groups, but these HPGs are rare in non-Bantu East-speaking Africans. Both East African Bantu populations are expected to be genetically as diverse as Central African Bantu populations with genetic signature of population expansion, and the Bantu populations in peripheries should show similar unimodal mismatch distribution, so genetic distance should not correlate with distance from the center of Bantu expansion. Because of large migration rates among the Bantu-speaking groups, the Bantu populations should be homogeneous with small genetic distances, and East African Bantu populations should cluster closer to Central African Bantu populations. However, they should not cluster with non-Bantu-speaking populations, because they have small migration rates with non-Bantu populations. Genetic distances may correlate well with geographic distances.

TABLE XI

RESULTS OF ANALYSES EXPECTED AMONG THE TAITA, MIJIKENDA, AND OTHER EAST AFRICAN BANTU POPULATIONS UNDER DIFFERENT BANTU EXPANSION MODELS

	HPG frequencies of Bantu and non-Bantu populations	Within-population genetic diversity and population expansion	Population differentiation among EA Bantu populations	Spatial pattern
1. Pure demographic expansion	* High frequency of CA[a]/WA[b] HPGs, with a great variation * CA/WA HPG sequences not shared with non-Bantu populations	* Smaller than CA Bantu populations * Small CA/WA HPG sequence diversity * Founder Effect	* Heterogeneous * Large AMOVA Φ_{ST} * Scatter on the MDS plot away from CA Bantu populations	* Negative correlation between genetic diversity and distance from center of Bantu expansion * No correlation between genetic and geographic distances * Small migration rates
2. Expansion with gene flow among the Bantu-speaking populations	* High frequency of CA and WA HPGs * CA/WA HPG sequences not shared with non-Bantu populations	* Similar to CA Bantu populations * Population expansion * Unimodal mismatch distribution similar to other Bantu populations	* Homogeneous * Small AMOVA Φ_{ST} * Close to CA Bantu populations on the MDS plot	* No correlation between genetic diversity and distance from center of Bantu expansion * Significant correlation between two distances * Large migration rates within Bantu-speaking group, but small migration rates with non-Bantu populations
3. Expansion with gene flow with non-Bantu-speaking populations	* Higher frequency of EA[c] HPGs * CA/WA HPG sequences shared with non-Bantu populations	* Larger than CA Bantu populations * Population expansion * Unimodal mismatch distribution similar to EA non-Bantu populations	* Heterogeneous * Large AMOVA Φ_{ST} * Between CA Bantu populations and EA non-Bantu populations on the MDS plot	* No correlation between genetic diversity and distance from center of Bantu expansion * No correlation between two distance measurements * Large migration rates

[a] CA (Central African), [b] WA (West Africans), and [c] EA (East African)

Model 3: Expansion with gene flow with neighboring non-Bantu groups

Contrary to the arguments of Cavalli-Sforza and others (Barbujani and Bertorelle 2001; Cavalli-Sforza et al. 1994; Cavalli-Sforza et al. 1988; Chikhi et al. 2002; Piazza et al. 1995; Sokal et al. 1991), in this model, linguistic differences do not reduce gene flow rates. Instead, multilingualism and inter-ethnic marriage have been common as has been reported among many other ethnic groups (Barth 1969; Moore 1994). Bantu populations would show evidence of population expansion (Ray et al. 2003; Wegmann et al. 2006) and would be genetically similar to non-Bantu groups.

The east African Bantu-speaking populations should have high frequencies of East African HPGs and non-Bantu-speaking east Africans have central and west African HPGs. East African Bantu populations should be genetically more diverse than Central African Bantu populations because of genetic exchange with non-Bantu populations, with signatures of population expansion. They should have unimodal mismatch distributions similar to non-Bantu east Africans, so genetic diversity will not correlate with distance from the center of Bantu expansion. The east African Bantu populations are expected to be heterogeneous with large genetic distances, so genetic distance may not correlate with geographic distance. East Bantu populations should cluster between Central African Bantu and non-Bantu east African populations because of large migration rates among both Bantu and non-Bantu populations.

4.3 Results

4.3.1 Haplogroups and haplogroup frequencies

A total of 126 different haplotypes was found among the Taita and Mijkenda sampled (TABLE XII), and most (93.7%) were assigned to mtDNA HPGs as defined by phylogeographic

TABLE XII

TAITA AND MIJIKENDA MITOCHONDRAL DNA HVRI SEQUENCES

Haplotype	Haplogroup	16024-16383	Taita	Mijikenda	Kikuyu
H001	L0a	129 148 172 187 188G 189 223 230 311 320	3		
H002	L0a	129 139T 148 172 187 188G 189 223 230 311 320	1		
H003	L0a	129 148 172 187 188A 189 223 230 256D 301 311 320		1	
H004	L0a	129 145 148 172 187 188G 189 223 230 311 320	1		
H005	L0a	129 148 172 187 188G 189 223 230 260 311 320	7		
H006	L0a	129 148 172 187 188G 189 230 260 311 320	1		
H007	L0a	129 148 172 187 188G 223 230 260 311 320	1		
H008	L0a1	129 148 168 172 187 188G 189 223 230 311 320		2	
H009	L0a1	129 148 168 172 187 188G 189 223 230 234 311 319 320		1	
H010	L0a1	129 148 168 172 187 188G 189 192 223 230 234 311 319 320	1		
H011	L0a1	129 148 168 172 173 187 188G 189 223 230 234 311 319 320	1	4	
H012	L0a1	129 148 165 168 172 188A 189 223 230 311 320	1		
H013	L0a1	148 168 172 187 188G 189 223 230 264 287 293 311 320	1		
H014	L0a1	129 148 168 172 187 188G 189 223 230 278 293 311 320	2		
H015	L0a1	129 148 168 172 187 188G 189 223 230 278 293C 311 320		1	
H016	L0a1	148 168 172 187 188G 189 223 230 278 311 320		1	
H017	L0a2	148 172 187 188G 189 223 230 311 320	14	18	
H018	L0a2	93 148 172 187 188G 189 223 230 311 320	1		
H019	L0a2	111 148 172 187 188G 189 223 230 311 320		1	
H020	L0a2	148 172 173 187 188G 189 223 230 311 320		1	
H021	L0a2	148 172 187 188G 189 192 223 230 234 311 320	5		
H022	L0a2	148 172 187 188G 189 201 223 230 311 320	1		
H023	L0a2	148 172 187 188G 189 214 221 223 230 311		3	
H024	L0d	129 162 172 187 189 212 223 243 265 311		1	
H025	L0d	129 187 189 223 230 278 290 300 311	1		
H026	L0d	145 169 187 189 223 230 243 274 278 290 311 362		4	
H027	L0f	129 169 172 187 189 223 230 278 311 327 368	6		
H028	L0f	129 169 172 187 189 223 278 311 327 368	2	1	
H029	L0f	129 169 172 187 189 223 278 311 327 352 368	1		

TABLE XII (continued)

TAITA AND MIJIKENDA MITOCHONDRAL DNA HVRI SEQUENCES

Haplotype	Haplogroup	16024-16383	Taita	Mijikenda	Kikuyu
H030	L0f	129 169 172 187 189 274 278 311 327 368	1		
H031	L0f	129 169 172 187 189 223 230 278 311 325 327 354 368	3	1	
H032	L0f	129 169 172 187 189 223 230 278 290 311 325 327 354 368		1	
H033	L0f	52 129 169 172 187 189 223 230 278 290 311 325 327 354 360 368	1		
H034	L0f	52 129 169 183C 189 223 230 278 290 311 325 327 354 368	2	1	
H035	L0f	93 129 169 172 187 189 223 230 256 278 284 311 325 327 354 368	3		
H036	L0f	129 169 172 187 189 218 223 230 278 291 311 327 354 368	2		
H037	L1b	126 187 189 223 264 270 278 311		2	
H038	L1b	93 126 187 189 223 264 270 278 311		2	
H039	L1b	93 126 168 187 189 223 264 270 278 311		1	
H040	L1b	114 126 187 189 223 264 270 278 311		1	
H041	L1b1	126 166 187 189 193 223 264 270 278 293 311	3		
H042	L1c	129 172 173 188A 189 223 256 278 293 294 311 360 368		1	
H043	L1c	117 129 172 173 188A 189 223 256 278 291 293 294 311 360 368	1	3	
H044	L1c1	129 187 189 223 278 293 294 311 360	1		
H045	L1c1	86 129 187 189 223 241 278 293 294 311 360		1	
H046	L1c1	129 163 187 189 209 223 278 293 294 311 360	2	3	
H047	L1c2	129 163 187 189 259 265C 278 286G 294 311 320 360	1		
H048	L1c2	129 169 187 189 223 265C 278 286G 294 311 360	1		
H049	L1c2	129 187 189 214 223 265C 278 286A 291 294 311 360	1		
H050	L1c2	42 129 187 189 214 223 265C 274 278 286A 291 294 311 360		3	
H051	L1c2	71 129 145 187 189 213 223 234 265C 278 286G 294 311 360		1	
H052	L1c2	86 129 145 187 189 213 223 265C 278 286G 294 311 360		1	
H053	L2a	223 234 249 278 294 390		2	
H054	L2a	223 234 249 278 294 295 390	2	4	
H055	L2a	223 234 249 278 292 294 295 390		1	
H056	L2a1	223 278 294 309 390		2	
H057	L2a1	223 278 294 309 368 390	3	1	
H058	L2a1	93 183C 189 278 294 309 390		3	

TABLE XII (continued)

TAITA AND MIJIKENDA MITOCHONDRAL DNA HVRI SEQUENCES

Haplotype	Haplogroup	16024-16383	Taita	Mijikenda	Kikuyu
H059	L2a1	131 189 223 225 234 278 294 309 390	1		
H060	L2a1	182C 183C 189 223 278 294 309 390		3	
H061	L2a1	182C 183C 189 223 278 290 294 309 390		2	
H062	L2a1	93 182C 183C 188 189 223 278 290 294 309 390		1	
H063	L2a1	183C 189 223 224 255 278 309 390		1	
H064	L2a1	189 192 223 278 294 390		1	
H065	L2a1	183C 189 192 223 229 278 291 294 311 390	1	2	
H066	L2a1	172 189 223 278 294 390		3	
H067	L2a1a	223 278 286 294 309 390	2	1	
H068	L2a1a	92 223 278 286 294 309 390		1	
H069	L2b	114A 213 223 278 354 390		1	
H070	L3b	124 223 278 362	1		
H071	L3b	93, 124, 223, 278, 362	3	4	
H072	L3b	223 278 311 362	4		
H073	L3b2	124 223 278 311 362		1	
H074	L3d1	124 223 319	3	14	
H075	L3d1	93 124 223 319		1	
H076	L3d1	124 182 223 274 319		1	
H077	L3d1	124 223 254 319		1	
H078	L3d1	124 223 259 319		1	
H079	L3d1	124 223 287 319		2	
H080	L3e1	223 327	2	6	
H081	L3e1	176 223 327		6	
H082	L3e1b	223 325D 327	2		
H083	L3e2b	172 183C 189 223 320	1	2	
H084	L3e3	223 265T	2	32	1
H085	L3e3	169 223 265T	1		
H086	L3f	209 223 311	4	11	
H087	L3f	209 223 260 311		4	

TABLE XII (continued)

TAITA AND MIJIKENDA MITOCHONDRAL DNA HVRI SEQUENCES

Haplotype	Haplogroup	16024-16383	Taita	Mijikenda	Kikuyu
H088	L3f	209	1		
H089	L3f	111 209 223 311	1		
H090	L3f	126 209	1		
H091	L3f	126 209 224	1		
H092	L3f	171 178 209 223 325	1		
H093	L3f	171 178 209 223 309 325	2		
H094	L3f	171 178 209 223 293 309 325	1		
H095	L3i	153 223	3		
H096	L3i	153 174 179 223 319	1		
H097	L3i	153 174 189 223 301 311 319	2		
H098	L3x1	169 223 278		2	
H099	L3x1	169 223 278 317T	1		
H100	L3	223 254 316	1		
H101	L3	178 223 254 311 316	2		
H102	L3	148 192 223 311		1	
H103	L3	223 278 316	8		
H104	L3	209, 223, 278, 316	1		
H105	L3	92 126 223 299 320		1	
H106	L3	93 223 355	1		
H107	L3	192 218 223 303 360	1		
H108	L4	42 183C 189 223 234 243 311 362		1	
H109	L4a1	172 207T 223 260 274 295 311 362	1		
H110	L4g	223 293T 311 355 362		1	
H111	L4g	223 274 293T 311 355 362	2		
H112	L4g	172 223 293T 311 327 355 362	4		
H113	L4g	71 172 223 293T 311 355 362	1	7	
H114	L4g	145 172 223 293T 311 355 362		1	
H115	L4g	172 223 246 293T 311 355 362	1		
H116	L4g	75 154 223 274 293T 311 355 362	2		

TABLE XII (continued)

TAITA AND MIJIKENDA MITOCHONDRAL DNA HVRI SEQUENCES

Haplotype	Haplogroup	16024-16383	Taita	Mijikenda	Kikuyu
H117	L4g	93 223 293T 311 355 362	1		
H118	L4g	93 189 223 293T 311 344 355 362	1		
H119	L4g	93 172 223 293T 311 319 355 362	5		
H120	L4g	93 172 223 240 293T 311 319 355 362	1		
H121	L4g	192 223 293T 301 311 355 362		1	
H122	L4g	75 223 274 293T 311 355 362		1	
H123	M1a	129 189 223 249 311 359	1	1	
H124	M1a	129 183C 189 223 249 311 359	1	2	
H125	M1a	129 182C 183C 189 223 249 311 359	1		
H126	M1a	129 183 223 249 258 311 355	1		
Total			157	195	1

analyses (Kivisild et al. 2004; Salas et al. 2002; Watson et al. 1997), but eight L3 haplotypes (n=15, 6.7%) lacked the diagnostic mutations that define L3 sub-HPGs.

The HPG frequencies of the Taita and Mijikenda ethnic groups are different from Central African Bantu groups (TABLE XIII and TABLE XIV). Central African Bantu populations have high frequencies of L1c (32.5%), but the Taita and Mijikenda have lower frequencies of L1c (4.5 and 6.7%). L1c is also rare among other East African Bantus (3.9%). East African Bantu populations have higher frequencies of L0a, L0f, and L4, HPGs that are believed to have originated in East Africa. These HPGs are more common among the Taita and less common among the Mijikenda and Central African Bantu populations. Compared to the Taita, the Mijikenda have higher frequencies of HPGs that have West or Central African origins, such as L1b, L1c, L3b, L3d, and L3e. Overall, East African Bantu populations have higher frequencies of East African HPGs than Central African Bantu populations and lower frequencies of West and Central African HPGs. Particularly, the Taita have similar frequencies of East African HPGs to the Nilo-Saharan populations from Kenya and Tanzania.

Four M1a haplotypes (n=7) were found among the Taita and Mijikenda. M1 is a non-L HPG of East African or Middle Eastern origin (Kivisild et al. 2004; Quintana-Murci et al. 1999) and it is rare among the East African Bantus. The Sukuma are the only other East African Bantu populations where this HPG has been found, but the HPG is relatively common among east African Afro-Asiatic and Nilo-Saharan populations.

Haplotype sharing between East African Bantu and non-Bantu East Africans was examined and illustrated using Network (Fig. 10 and 11). L1c and L3e are the two most

TABLE XIII

MITOCHONDRIAL DNA HPG FREQUENCY (%) OF BANTU AND NON-BANTU EAST AFRICAN POPULATIONS

	Proposed			E.A.	C.A.	S.A.	S-E.	N-E.	S-E.	N-E.
Haplogrous	Origin[a]	Taita	Mijikenda	Bantu[b, c]	Bantu	Bantu	N-S	N-S	A-A	A-A
L0a	East Africa	26.1	16.9	14.8	9.9	24.8	37.2	11.7	26.0	7.1
L0d	Khoisan (SA)	0.6	2.6	0.0	0.0	5.0	1.1	0.0	2.0	0.0
L0f	East Africa	13.4	2.1	9.4	0.0	0.0	4.3	0.8	30.0	0.3
L0k	Khoisan (SA)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
L1b	West Africa	1.9	3.1	0.0	6.2	1.2	0.0	4.7	0.0	2.0
L1c	Central Africa	4.5	6.7	3.9	32.5	5.3	0.0	0.8	0.0	0.0
L2a	Wide Spread	5.7	14.4	7.0	14.5	32.9	8.5	17.2	4.0	13.8
L2 other	West/Central Africa	0.0	0.5	2.3	5.1	2.9	0.0	0.0	0.0	1.7
L4	East Africa	12.1	6.2	15.6	1.3	0.0	13.8	5.5	10.0	6.4
L5	East Africa	0.0	0.0	3.9	0.0	0.5	5.3	4.7	0.0	2.7
L3bd	West/Central Africa	7.0	12.8	14.8	7.6	8.4	2.1	2.3	0.0	3.0
L3e	Central Africa	5.1	23.6	7.8	15.4	15.4	0.0	0.8	2.0	0.0
L3f	East Africa	7.6	7.7	0.8	6.0	2.4	5.3	8.6	0.0	5.4
L3h	East or West Africa	0.0	0.0	0.0	0.7	0.0	6.4	2.3	2.0	0.7
L3i	East Africa	4.5	0.0	0.0	0.0	0.0	0.0	1.6	0.0	1.3
L3w	East Africa	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	3.4
L3x	East Africa	0.6	1.0	0.8	0.0	0.0	0.0	0.8	2.0	4.0
L3 other, L6, M, N, etc		10.8	2.6	17.9	0.8	1.2	16.0	38.3	22.0	48.1
n		157	195	128	1180	416	94	128	50	297

[a] The places of origin for mtDNA haplogroups were previously proposed (Kivisild et al. 2004; Salas et al. 2002; Watson et al. 1997).

[b] East African Bantu populations include the Hutu, Kikuyu, Sukuma, and Turu.

[c] The grouping of populations as shown on the TABLE X (East African Bantu, Central African Bantu, Southeastern African Bantu, south East African Nilo-Saharan, north East African Nilo-Saharan, south East African Afro-Asiatic, and north East African Afro-Asiatic populations).

TABLE XIV

MITOCHONDRIAL DNA HPG FREQUENCY (%) BASED ON PROPOSED ORIGIN

			E.A.	C.A.	S.E.	S-E.	N-E.	S-E.	N-E.
Proposed Origin	Taita	Mijikenda	Bantus	Bantus	Bantus	N-S	N-S	A-A	A-A
East Africa	66.2	33.8	53.1	17.2	27.6	67.0	33.6	80.0	30.6
Khoisan (SA)	0.6	2.6	0.0	0.0	5.0	1.1	0.0	2.0	0.0
West/Central Africa	18.5	46.7	28.9	66.8	33.2	2.1	8.6	2.0	6.7
Other	14.6	16.9	18.0	16.0	34.1	29.8	57.8	16.0	62.6

Individual mtDNA HPGs were grouped based on proposed place of origin; East African HPGs (L0a, L0f, L4, L5, L3f, L3h, L3i, L3w, and L3x), Khoisan HPGs (L0d and L0k), West/Central African HPGs (L1b, L1c, L2 other, L3bd, and L3e), and other HPGs (L2a, L3 other, L6, M, N, etc)

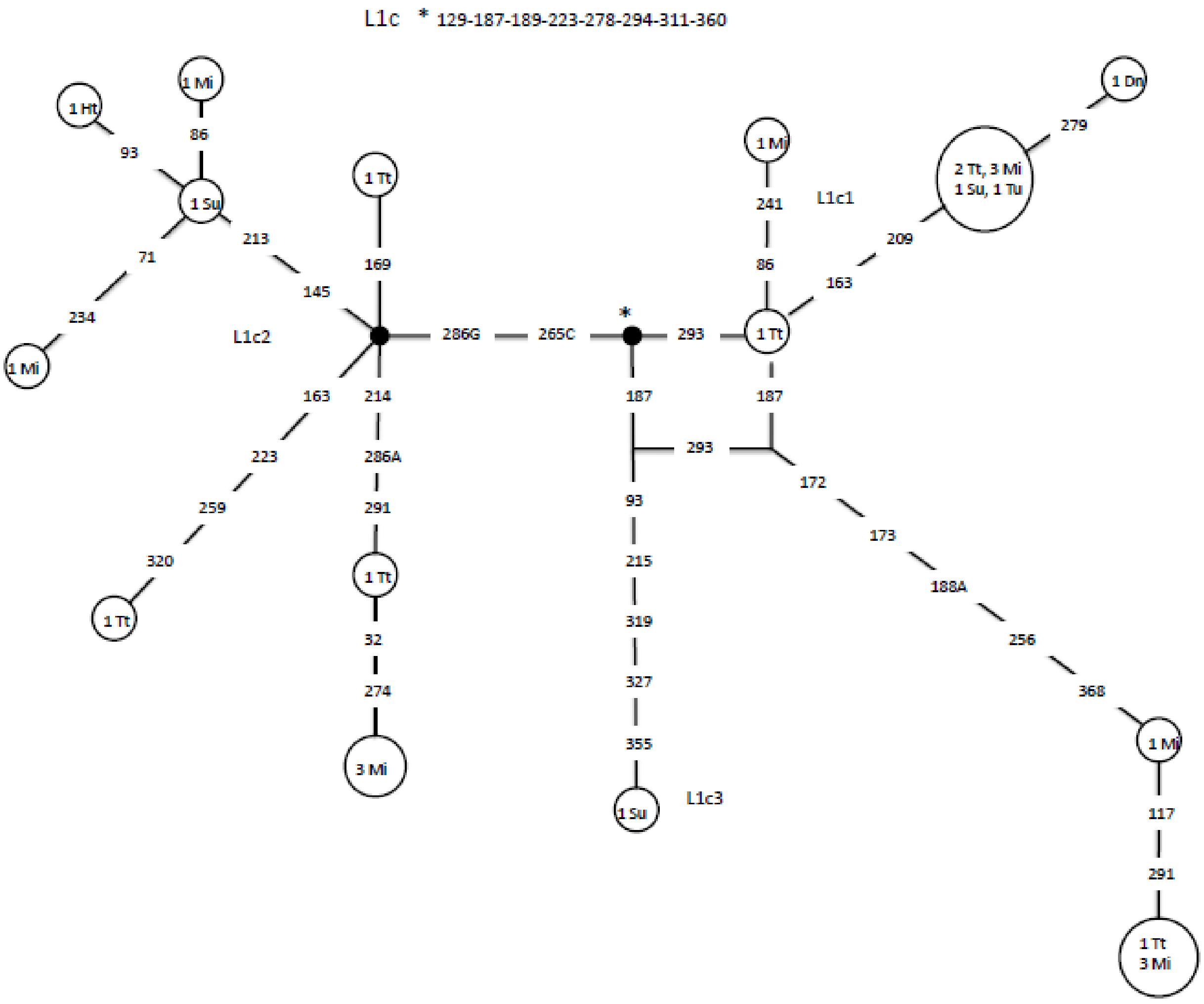

Figure 10. Network of East African L1c haplogroup. Abbreviation used in the Network tree: Tt (Taita), Mi (Mijikenda), Ki (Kikuyu), Ht (Hutu), Su (Sukuma), Tu (Turu), and Dn (Dinka).

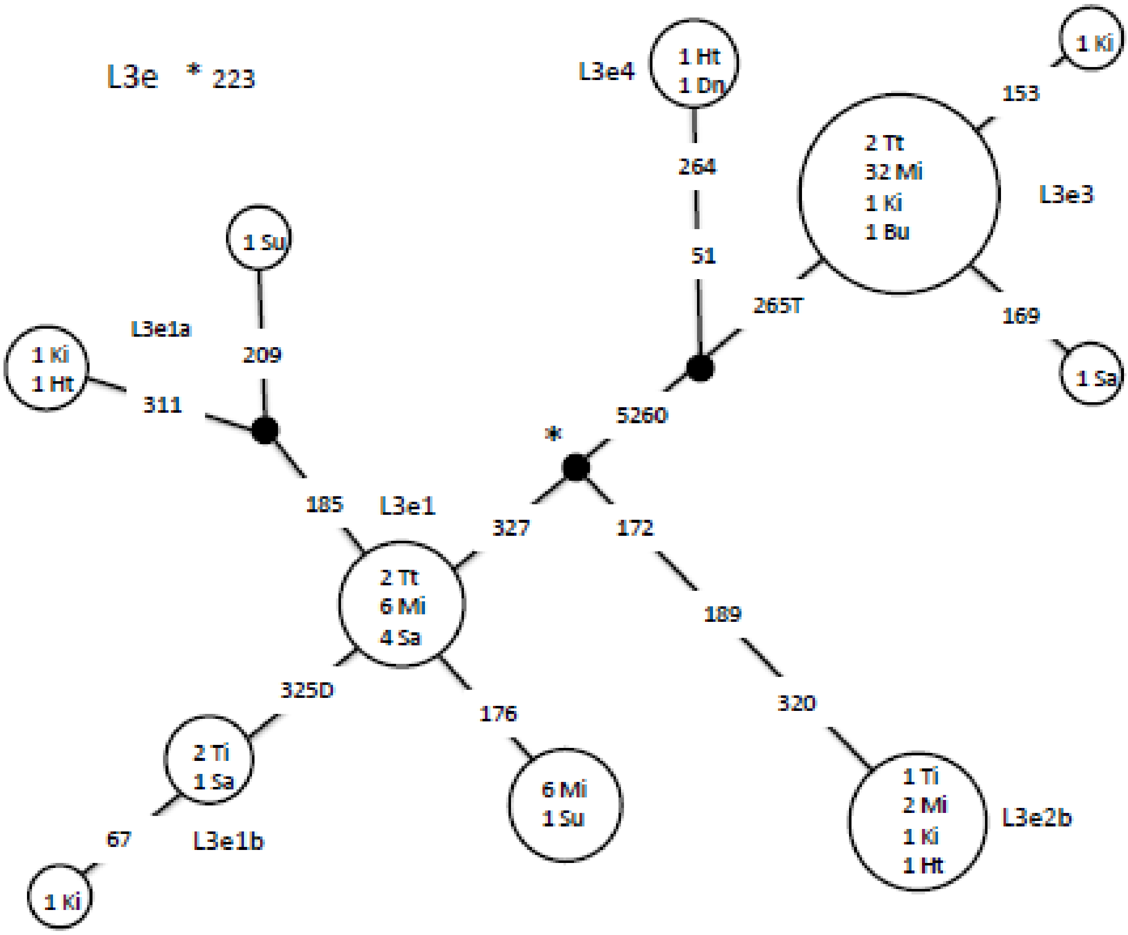

Figure 11. Network of East African L3e haplogroup. Abbreviation used in the Network tree: Tt (Taita), Mi (Mijikenda), Ki (Kikuyu), Ht (Hutu), Su (Sukuma), Bu (Burunge) Dn (Dinka), and Sa (Sandawe)

frequent HPGs in Central Africa (Batini et al. 2007; Quintana-Murci et al. 2008; Salas et al. 2002). L1c sub-HPGs, L1c1b, L1c1c, and L1c2 are proposed to be Bantu origin (Batini et al. 2007). Because the diagnostic mutations for sub-HPGs, L1c1b and L1c1c, are located in HVR II, none of the East African Bantu L1c1 haplotypes were assigned to them. However, both L1c and L3e HPGs are rare among East African Bantu and non-Bantu East African groups. Out of fifteen L1c haplotypes, only one L1c1 haplotype was identified among a non-Bantu population, the Dinka, a Nilo-Saharan from northern East Africa. L3e haplotypes are little more common among the non-Bantu populations and five haplotypes were found among them. Three haplotypes, one L3e1 (n = 4), one L3e1b (n = 1) and one L3e3 (n = 1), were found in the Sandawe, Khoisan foragers. The L3e1 haplotype is shared with two Taita and four Mijikenda individuals, while the L3e1b haplotype is shared with two Taita individuals. The Burunge has the L3e3 ancestral haplotype which is very common among the East African Bantu populations The only one L3e4 haplotype found is shared with the Dinka and Hutu.

4.3.2 <u>Within-population genetic diversity and population expansion</u>

The Taita, Mijikenda, and other east African Bantu-speaking populations are genetically very diverse in comparison to other Sub-Saharan populations, and they are as diverse as central African Bantu-speaking populations (TABLE XV). East African Bantu populations tend to have smaller θ_k values compared to East African Nilo-Saharan and Afro-Asiatic speaking populations, but they have θ_S and θ_π values as large as those groups. Among all the Bantu-speaking populations, the Taita are one of the most genetically diverse populations. The Taita have the largest θ_S value and the second largest θ_π value after the Hutu, another east African Bantu population. The Mijikenda are less genetically diverse, but they have larger θ_S values than many

TABLE XV

AFRICAN POPULATION SUMMARY STATISTICS

	Haplotype						Tajima's		
Populations	Diversity	*k*	θ_k (95% CI)	S	θ_S (SD)	$\theta\pi$ (SD)	*D*	*Fs*	*M*[b]
Taita[a]	0.981	78	61.419 (44.209-84.846)	84	14.936 (3.676)	11.946 (6.013)	-1.091	-24.330**	38.975
Mijikenda[a]	0.952	68	36.651 (26.798-49.814)	74	12.605 (3.065)	9.930 (5.048)	-1.068	-24.436**	16.781
East Africa - South									
Bantu									
Kikuyu	0.993	23	133.893 (44.790-450.024)	48	12.447 (4.316)	9.650 (5.100)	-1.325	-14.172**	NA
Sukuma	0.988	27	78.779 (34.916-190.912)	53	13.160 (4.317)	10.843 (5.634)	-1.145	14.350**	66.575
Hutu[a]	0.979	30	45.529 (24.384-87.209)	51	11.852 (3.720)	12.139 (6.216)	-0.595	-11.737*	94.190
Turu[a]	0.951	18	19.273 (9.487-39.686)	40	10.185 (3.486)	9.770 (5.128)	-0.712	-3.545	15.792
Nilo-Saharan									
Turkana	0.994	34	198.007 (77.407-565.238)	61	14.612 (4.619)	12.469 (6.398)	-1.098	-23.478**	
Datoga	0.984	40	58.278 (33.955-101.877)	68	14.529 (4.235)	12.453 (6.327)	-1.020	-19.557**	
Afro-Asiatic									
Burunge	0.937	22	20.951 (11.214-39.355)	44	10.472 (3.390)	11.353 (5.851)	-0.368	-4.036	
Iraqw	0.924	8	9.317 (3.286-27.264)	29	9.603 (4.021)	13.022 (7.108)	-0.370	0.838	
Khoisans									
Sandawe	0.831	30	16.604 (10.394-26.206)	49	9.634 (2.751)	9.103 (4.689)	-0.679	-6.058	
Hadza	0.796	32	16.401 (10.499-25.292)	46	8.956 (2.514)	6.292 (3.336)	-1.283	-11.600*	
East Africa - North									
Nilo-Saharan									
Dinka	0.995	42	228.871 (98.623-584.887)	62	14.107 (4.288)	10.332 (5.330)	-1.379	-24.807**	
Nubia	0.975	53	63.755 (40.859-100.331)	73	14.665 (4.007)	9.760 (5.013)	-1.481*	-24.785**	

TABLE XV (continued)

AFRICAN POPULATION SUMMARY STATISTICS

	Haplotype						Tajima's		
Populations	Diversity	*k*	θ_k (95% CI)	S	θ_S (SD)	$\theta\pi$ (SD)	*D*	*Fs*	*M*[b]
Afro-Asiatic									
Gurage	1.000	21	NA	43	11.952 (4.313)	9.989 (5.311)	-1.196	-15.273**	
Tigrais	0.994	46	162.663 (81.383-346.177)	67	14.764 (4.355)	9.868 (5.090)	-1.536*	-24.843**	
Oromo	0.992	30	154.674 (60.007-443.620)	57	14.045 (4.556)	10.466 (5.444)	-1.420*	-21.188**	
Amhara	0.994	88	147.293 (99.453-220.952)	95	17.722 (4.478)	10.558 (5.364)	-1.629*	-24.558**	
Somali	0.992	24	99.676 (38.058-288.718)	43	11.156 (3.842)	8.242 (4.391)	-1.387	-15.986**	
Afar	0.975	13	30.015 (10.753-90.518)	39	11.753 (4.521)	11.183 (6.006)	-0.899	-2.896	
Southeastern Africa									
Bantu									
Ronga[a]	0.987	19	63.012 (23.572-184.865)	38	10.424 (3.771)	10.416 (5.511)	-0.639	-8.080*	70.444
Shona[a]	0.988	39	49.001 (29.109-83.547)	60	12.914 (3.779)	10.558 (5.412)	-1.082	-19.950**	38.928
Nyungwe[a]	0.974	16	34.968 (13.864-94.418)	36	9.865 (3.611)	10.998 (5.829)	-0.296	-4.427*	35.391
Shangaan[a]	0.961	17	32.457 (13..673-81.462)	36	9.601 (3.500)	10.095 (5.351)	-0.437	-4.994*	24.703
Chwabo[a]	0.942	15	25.594 (10.596-65.022)	29	8.174 (3.091)	9.127 (4.896)	-0.217	-3.998**	10.446
Chopi	0.954	18	22.433 (10.697-48.149)	33	8.562 (3.027)	8.487 (4.511)	-0.578	-5.017*	11.121
Tonga	0.947	14	19.391 (8.250-47.238)	30	8.174 (3.091)	8.801 (4.733)	-0.348	-3.006	16.889
Nyanja	0.937	12	11.747 (5.113-27.309)	24	6.765 (2.615)	8.759 (4.712)	-0.370	-1.185	11.422
Makhwa[a]	0.905	12	11.747 (5.113-27.309)	33	9.302 (3.471)	11.297 (5.978)	-0.059	-0.421	7.808
Lomwe[a]	0.879	12	11.747 (5.113-27.309)	27	7.610 (2.901)	9.337 (5.001)	0.100	-0.967	5.241
Sena[a]	0.891	11	8.613 (3.801-19.444)	22	6.115 (2.370)	8.233 (4.438)	0.548	-0.328	5.389

TABLE XV (continued)

AFRICAN POPULATION SUMMARY STATISTICS

	Haplotype						Tajima's		
Populations	Diversity	k	θ_k (95% CI)	S	θ_S (SD)	$\theta\pi$ (SD)	*D*	*Fs*	*M*[b]
Central Africa									
Bantu									
Bassa[a]	0.991	38	99.974 (50.799-207.105)	61	13.652 (4.161)	11.165 (5.733)	-1.119	-24.283**	105.188
Ngoumba[a]	0.991	36	90.184 (45.619-187.442)	52	11.954 (3.717)	10.701 (5.514)	-0.900	-23.242**	107.463
Mbundu	0.989	34	73.060 (37.653-147.902)	57	13.174 (4.079)	11.062 (5.692)	-1.060	-19.271**	88.444
Bamileke[a]	0.988	36	63.764 (34.821-120.307)	53	11.942 (3.654)	9.379 (4.864)	-1.166	-22.539**	95.218
Ewondo[a]	0.983	38	58.673 (33.389-105.392)	54	11.679 (3.517)	12.331 (6.276)	-0.479	-18.723**	56.891
Bakaka[a]	0.983	36	56.184 (31.460-102.721)	59	12.725 (3.841)	11.807 (6.031)	-0.823	-17.540**	65.416
Sanga	0.970	21	29.755 (14.512-62.785)	37	9.340 (3.203)	10.783 (5.617)	-0.161	-5.913*	27.757
Bateke[a]	0.945	23	15.913 (9.074-27.687)	42	9.377 (2.919)	7.736 (4.067)	-1.003	-5.416*	18.569
Bubi[a]	0.908	15	7.469 (3.946-13.796)	31	7.089 (2.329)	7.629 (4.025)	0.148	0.208	9.978
Non-Bantu Niger-Congo									
Fali	0.976	25	26.243 (14.320-48.543)	41	9.583 (3.086)	9.601 (4.981)	-0.560	-7.691*	
Afro-Asiatic									
Hide	0.996	22	250.032 (61.477-1005.642)	46	12.433 (4.399)	10.427 (5.504)	-1.177	-13.931**	
Uldeme	0.992	25	108.009 (41.370-312.242)	38	9.508 (3.300)	9.209 (4.859)	-0.693	-15.899**	
Podokwo	0.991	33	98.528 (46.421-223.182)	54	12.299 (3.902)	9.562 (4.979)	-1.259	-22.883**	
Mandara	0.990	31	87.363 (40.936-198.621)	47	11.019 (3.568)	8.730 (4.581)	-1.176	-21.437**	
Masa	0.987	26	73.247 (32.352-177.923)	38	9.512 (3.235)	8.974 (4.726)	-0.784	-15.721**	
Mafa	0.980	23	35.193 (17.373-73.586)	44	10.926 (3.642)	9.021 (4.744)	-1.130	-9.261*	

TABLE XV (continued)

AFRICAN POPULATION SUMMARY STATISTICS

	Haplotype						Tajima's		
Populations	Diversity	k	θ_k (95% CI)	S	θ_S (SD)	$\theta\pi$ (SD)	D	Fs	M^b
West Africa									
Non-Bantu Niger-Congo									
Yoruba	0.996	31	242.559 (82.571-807.289)	46	11.334 (3.743)	8.576 (4.522)	-1.315	-24.938**	
Bambara	0.994	45	155.902 (77.888-332.131)	50	11.065 (3.361)	8.000 (4.191)	-1.306	-25.008**	
Senegalese	0.989	42	121.054 (61.991-249.347)	42	9.377 (2.919)	7.299 (3.856)	-1.127	-25.207**	
Serer	0.992	21	111.727 (37.139-376.894)	42	11.380 (4.046)	9.842 (5.214)	-1.036	-11.679**	
Wolof	0.992	40	110.264 (56.256-227.740)	45	9.914 (3.090)	8.878 (4.623)	-0.811	-24.959**	
Malinke	0.974	23	38.747 (18.714-83.327)	32	8.010 (2.776)	6.798 (3.661)	-0.952	-12.589**	
Fulbe	0.972	37	40.149 (24.116-67.360)	46	9.864 (2.957)	8.299 (4.324)	-0.970	-20.899**	
Mandenka	0.970	53	36.083 (24.794-52.281)	57	10.650 (2.831)	8.455 (4.362)	-1.150	-24.926**	
Nilo-Saharan									
Kanuri	0.989	13	82.112 (20.550-352.176)	41	12.893 (5.085)	10.208 (5.570)	-1.504	-4.975*	
Songhai	0.983	14	49.895 (16.012-171.787)	32	9.644 (3.770)	8.873 (3.770)	-0.941	-5.388*	
Afro-Asiatic									
Hausa	0.995	19	177.112 (45.396-750.663)	31	8.738 (3.281)	6.698 (3.683)	-1.290	-14.114**	
Tuareg	0.993	22	122.560 (40.875-412.660)	41	10.979 (3.881)	8.369 (4.476)	-1.367	-14.611**	
Mauritanian	0.979	43	56.232 (33.965-94.335)	45	9.517 (2.833)	7.469 (3.920)	-1.085	-25.164**	

** $P < 0.001$, * $P < 0.05$

[a] Bantu populations used for comparison of Theta using MIGRATE analysis.

[b] M was estimated only for Bantu populations

TABLE XVI

DIFFERENCES IN AVERAGE GENETIC DIVERSITY AMONG GEOGRAPHIC AND LINGUISTIC GROUPS IN SUB-SAHARAN AFRICA

	θ_k	θ_S	$\theta\pi$
Geography			
East Africa	73.474	12.479	10.706
Central Africa	73.349	11.013	9.882
Southeastern Africa	26.428	8.864	9.646
West Africa	103.415	10.259	8.290
ANOVA (*P*-values)	0.017	0.000	0.000

other Bantu populations. No east African Bantu population has a significantly negative Tajima's *D* value, but all of the east African Bantu-speaking populations analyzed, except for the Turu, show evidence of population expansion with significant *Fs* values.

TABLE XVI shows the differences in the average genetic diversity among geographic groups. The east African populations have significantly larger θ_S and θ_π values than populations from the other areas of Africa. The west African populations have the largest θ_k values, but they have the lowest θ_π values.

The effective population size (Θ) values estimated using MIGRATE resemble the estimated θ values (TABLE XVII). The Bantu-speaking populations have smaller Θ values than the east African Nilo-Saharan and northeast African Afro-Asiatic populations, but the Tanzanian Afro-Asiatic populations have smaller Θ values than the east African Bantu populations.

I compared Taita and Mijikenda mismatch distributions with those of the Turkana and Shona (Fig. 12). The Turkana, a Nilo-Saharan group from Kenya, has large θ_S and $\theta\pi$, similar to those of the Taita. The Shona is one of the most genetically diverse Bantu-speaking populations from southeastern Africa with large sample size. They have larger θ_S and θ_π values than the Mijikenda, but not the Taita. The spatial expansion model predicts that two populations from the

TABLE XVII

Θ AND MIGRATION RATES ($2N_fm$) ESTIMATED FOR EAST AFRICAN POPULATIONS USING MIGRATE

	Θ	Taita	Mijikenda	East Bantu	Other Bantu	Nilo-Saharan	N.E. Afro-Asiatic
Taita	0.059						
Mijikenda	0.037	28.071					
East African Bantu	0.105	25.925	18.696				
Other Bantu	0.121	25.453	36.270	37.884			
Nilo-Saharans	0.182	18.750	17.319	48.221	23.737		
N.E. Afro-Asiatic	0.146	6.123	9.704	18.074	11.588	42.136	
S.E. Afro-Asiatic	0.077	15.367	6.600	37.249	11.422	63.710	13.556

The population groupings are based on their language and geographical location and see the listing of population on the TABLE X.

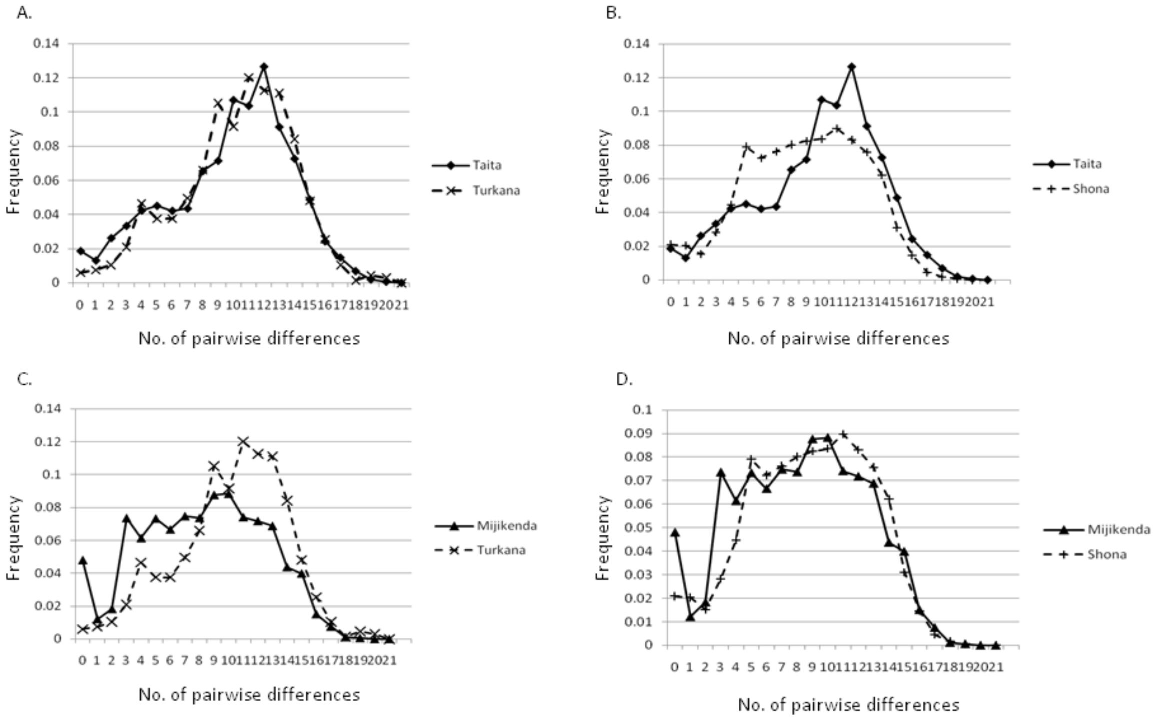

Figure 12. Comparison of mismatch distributions A) between the Taita and Turkana, B) the Taita and Shona, C) between the Mijikenda and Turkana, and D) the Mijikenda and Shona

same spatial expansion wave will have similar mismatch distributions (Excoffier 2004), so all the Bantu populations in the periphery should have similar mismatch distributions. The unimodal mismatch distribution of the Taita is similar to that of the Turkana, but not the Shona. Conversely, the Mijikenda mismatch distribution is similar to that of the Shona, but not the Turkana.

Sequence variation of two East African HPGs (L0a and L4) and two Central African HPGs (L1c and L3e) found among all East African populations were examined (TABLE XVIII). The East African HPGs are very diverse and show evidence of expansion. Khoisan speakers from Tanzania (Hadza and Sandawe) have high frequencies of HPG L4g and the Khoisan L4g sequences are over-represented in East African L4g data set. Nonetheless, whether Khoisan L4g

TABLE XVIII

SEQUENCE VARIATION OF TWO EAST AFRICAN HAPLOGROUPS (L0 AND L4) AND TWO CENTRAL AFRICAN HAPLOGROUPS (L1 AND L3E)

	n	*h*	k	θ_k (95% CI)	s	θ_S(SD)	$\theta\pi$ (SD)	Tajima's *D*	*Fs*
L0a	180	0.917	62	33.036 (23.832-45.481)	54	9.190 (2.342)	3.660 (2.060)	-1.905*	-25.976**
L0a1	70	0.920	25	13.475 (8.106-22.075)	23	4.773 (1.556)	3.296 (1.902)	-1.098	-14.067**
L0a2	78	0.633	19	7.687 (4.450-12.939)	22	4.465 (1.451)	1.593 (1.060)	-2.016**	-13.813**
L4	178	0.890	60	31.393 (22.560-43.366)	59	10.250 (2.579)	4.666 (2.543)	-1.804*	-25.601**
L4g	159	0.862	44	19.784 (13.648-28.351)	44	7.797 (2.071)	3.635 (2.050)	-1.737*	-26.143**
L1c	26	0.908	15	13.929 (6.640-29.404)	32	8.386 (2.994)	8.869 (4.707)	-0.334	-2.051
L1c1	10	0.533	4	1.956 (0.598-1.127)	5	1.767 (1.014)	1.377 (1.044)	-1.035	-0.312
L1c2	10	0.917	8	16.397 (4.890-59.016)	16	5.656 (2.615)	6.072 (3.572)	-0.209	-1.789
L3e	70	0.697	11	3.426 (1.732-6.465)	15	2.905 (1.057)	1.978(1.255)	-1.028	-2.131
L3e1	26	0.720	6	2.132 (0.840-5.083)	6	1.310 (0.693)	0.968 (0.758)	-0.837	-1.320
L3e3	38	0.104	3	0.564 (0.166-1.7090	2	0.476 (0.350)	0.109 (0.202)	-1.491*	-2.661*

** P < 0.001, * P < 0.05

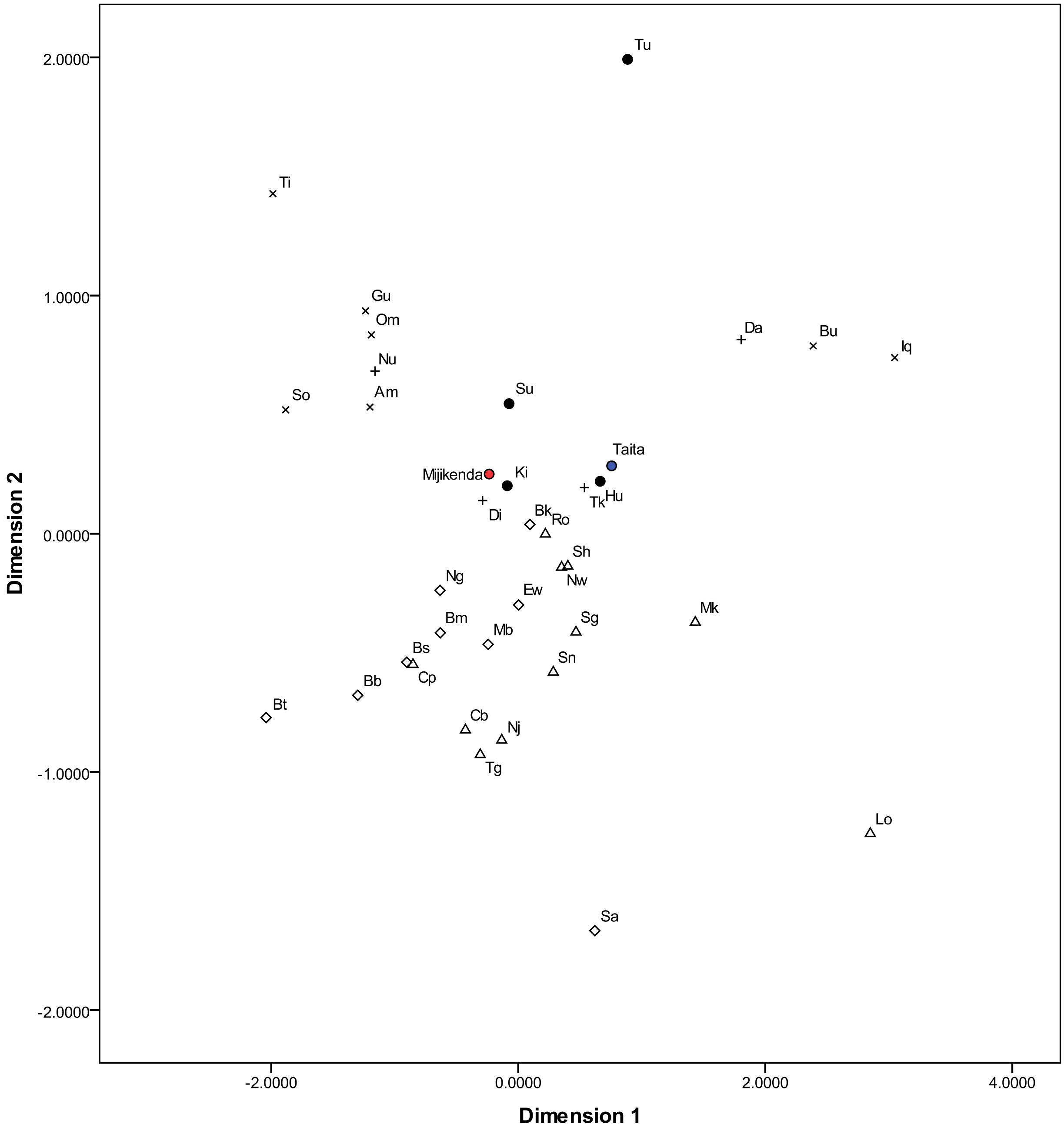

Figure 13. MDS plot of Bantu and East African populations. The symbols used in the plot: East African Bantu (O - Taita with blue fill, Mijikenda with red fill, and other East African Bantu populations with black fill), Central African Bantu (◊), southeastern African Bantu (Δ),Afro-Asiatic (×), and Nilo-Saharan (+) populations.

sequences are included or not, the analysis shows that L4g is diverse HPGs. On the other hand, Central African HPGs are rare in East Africa, especially among non-Bantu speakers. These two Central African HPGs are not diverse and do not show evidence of population expansion.

4.3.3 Population differentiation

Many east African Bantu-speaking populations are genetically distant from central African Bantu-speaking populations and do not cluster closely with central African Bantu or southeastern African Bantu populations on the MDS plot (Fig. 13). Instead, east African Bantu populations are dispersed on the MDS plot and many are located between non-Bantu east African populations and central African and southeastern African Bantu populations. Population pairwise genetic distance *P*-values between the east African Bantu and central African Bantu populations as well as between east African Bantu and non-Bantu east African populations tend to be significantly large. The Taita are clustered with the Hutu (Bantu) and Turkana (Nilo-Saharan) between the Tanzanian Nilo-Saharan and Afro-Asiatic populations on one side and Central African Bantu populations. Population pairwise genetic distance (Φ_{ST}) *P-values* between the Taita and these non-Bantu-speaking East Africans (Nilo-Saharan and Afro-Asiatic speakers) tend not to be significant (TABLE XIX). The Mijikenda are plotted between Central African Bantu and northern East African populations, but more closely with Central African Bantu populations.

Exact tests based on HPG frequencies show that the East African Bantu-speaking populations are genetically differentiated from each other (TABLE XX). The Taita and Mijikenda are genetically different from each other and from other East African Bantu populations. Genetic heterogeneity among the East African Bantu-speaking group is supported

TABLE XIX

PAIRWISE POPULATION GENETIC DISTANCE (Φ_{ST}) *P*-VALUES

	Taita	Mijikenda	Kikuyu	Sukuma	Turu	Hutu
Mijikenda	0.000					
Kikuyu	0.099	0.622				
Sukuma	0.009	0.108	0.315			
Turu	0.090	0.000	0.000	0.189		
Hutu	0.099	0.009	0.108	0.099	0.036	
Turkana	0.108	0.000	0.279	0.234	0.063	0.108
Datoga	0.000	0.000	0.009	0.009	0.000	0.000
Dinka	0.000	0.009	0.234	0.171	0.009	0.000
Nubians	0.000	0.000	0.117	0.000	0.000	0.000
Somali	0.000	0.000	0.117	0.018	0.000	0.000
Burunge	0.027	0.000	0.000	0.000	0.027	0.009
Iraqw	0.234	0.000	0.018	0.018	0.054	0.063
Amhara	0.000	0.000	0.117	0.009	0.000	0.000
Gurage	0.009	0.018	0.189	0.090	0.000	0.000
Oromo	0.000	0.009	0.108	0.018	0.000	0.000
Tigrais	0.000	0.000	0.000	0.000	0.000	0.000
Bassa	0.000	0.000	0.036	0.000	0.000	0.000
Ngoumba	0.000	0.009	0.117	0.009	0.000	0.000
Bamileke	0.000	0.018	0.171	0.054	0.000	0.000
Ewondo	0.000	0.000	0.000	0.009	0.000	0.000
Bakaka	0.000	0.018	0.523	0.108	0.009	0.018
Sanga	0.000	0.000	0.018	0.000	0.000	0.000
Bateke	0.000	0.000	0.000	0.000	0.000	0.000
Bubi	0.000	0.000	0.081	0.000	0.000	0.000
Mbundu	0.000	0.000	0.216	0.018	0.000	0.000
Sena	0.045	0.144	0.144	0.009	0.000	0.027
Ronga	0.063	0.270	0.586	0.117	0.018	0.189
Nyanja	0.009	0.072	0.171	0.018	0.000	0.018
Nyungwe	0.135	0.207	0.459	0.207	0.000	0.342
Makhuwa	0.144	0.000	0.099	0.018	0.000	0.036
Lomwe	0.000	0.000	0.000	0.000	0.000	0.000
Chwabo	0.009	0.036	0.108	0.000	0.000	0.045
Chopi	0.000	0.099	0.198	0.018	0.000	0.009
Shangaan	0.054	0.036	0.225	0.027	0.000	0.081
Tonga	0.000	0.009	0.225	0.027	0.000	0.009
Shona	0.000	0.009	0.189	0.045	0.000	0.054

TABLE XX

EXACT TEST OF EAST AFRICAN BANTU POPULATIONS BASED ON HPG FREQUENCIES

	Taita	Mijikenda	Kikuyu	Hutu	Sukuma
Mijikenda	0.000				
Kikuyu	0.000	0.000			
Hutu	0.009	0.000	0.035		
Sukuma	0.004	0.000	0.011	0.268	
Turu	0.171	0.000	0.002	0.240	0.295

by the AMOVA results as well. The East African Bantus group has greater among-population variance than the southeastern African Bantu and East African Afro-Asiatic speaking group and a significant Φ_{ST} *P*-value (TABLE XXI).

4.3.4 Spatial patterns

The correlation between genetic diversity and the distance from the center of the Bantu expansion was examined. If the Bantu expanded without gene flow, then the Bantu-speaking populations that occupy the periphery of the expansion should be less genetically diverse than the Bantu populations that occupy the core area of the expansion, meaning that the genetic diversity values would be negatively correlated with distance from expansion center. Although gene diversity was negatively correlated with distance here, the correlation was not statistically significant (TABLE XXII). Figure 14 illustrates the lack of this correlation largely due to great genetic diversity observed among East African Bantu-speaking populations. The figure shows that the East African Bantu populations are genetically as diverse as, or more diverse than Central African Bantu populations with large θ_S.

TABLE XXI

AMOVA RESULTS

	Number of Populations	Among Populations Variance (%)	Within Populations Variance (%)	Φ_{ST} (*P*)
All Bantus	26	3.88	96.12	0.039 (0.000)
East Bantu[a]	17	3.08	96.92	0.031 (0.000)
East African Bantu	6	2.59	97.41	0.026 (0.000)
Southeastern African Bantu	11	1.03	98.97	0.010 (0.122)
West Bantu[b]	9	2.25	97.75	0.022 (0.000)
Central African Bantu	8	2.56	97.44	0.026 (0.000)
S.E. Nilo-Saharan	2	2.64	97.36	0.026 (0.030)
S.E. Afro-Asiatic	2	0.37	99.63	0.004 (0.370)
N.E. Afro-Asiatic	5	0.05	99.95	0.001 (0.411)

[a] Speakers of East Bantu languages (East African and Southeastern Bantu)

[b] Speakers of West Bantu languages (Central African Bantu and Mbundu)

TABLE XXII

THE LACK OF CORRELATION BETWEEN GENETIC DIVERSITY AND THE DISTANCE FROM THE CENTER OF BANTU EXPANSION (*r* and *P*-value)

	h	θ_k	θ_S	θ_π
1-Step Expansion	-0.315 (0.117)	-0.300 (0.137)	-0.294 (0.145)	-0.170 (0.408)
2-Step Expansion	-0.330 (0.099)	-0.320 (0.111)	-0.354 (0.076)	-0.214 (0.293)

Next I tested an isolation-by-distance model using the Mantel test to examine gene flow among neighboring populations. A significant correlation between pairwise population genetic distances (Φ_{ST}) and geographical distances (*P value* = 0.001) was observed.

To quantify the intensity of gene flow, migration rates were estimated. Migration rates, M=2$N_f m$ estimated using the spatial expansion model of mismatch distribution, show the similar patterns as within-population genetic diversity estimated for each populations (TABLE XV). Genetically diverse populations, especially those with large θ_k, values, have large *M* values and populations with low genetic diversity have small *M* values. The Taita and Mijikenda have small θ_k and *M* vales. The genetically diverse Central African and southeastern African Bantu populations with high *M* values tend to cluster together on the MDS plot, while genetically less diverse populations with small *M* values tend to scatter on the plot. A migration rate could not be obtained from the Kikuyu, possibly because of poor fit of the data to the spatial expansion model.

The migration rates (2$N_f m$) obtained from MIGRATE analyses show that gene flow among different ethnic and language groups was common (TABLE XVII). As the geographical proximity and close social ties would predict, the migration rate estimated between the Taita and Mijikenda (2$N_f m$ = 28.071) tends to be higher than between the Taita and other groups or between the Mijikenda and other groups, ranging from 6.123 (between Taita and N.E. Afro-

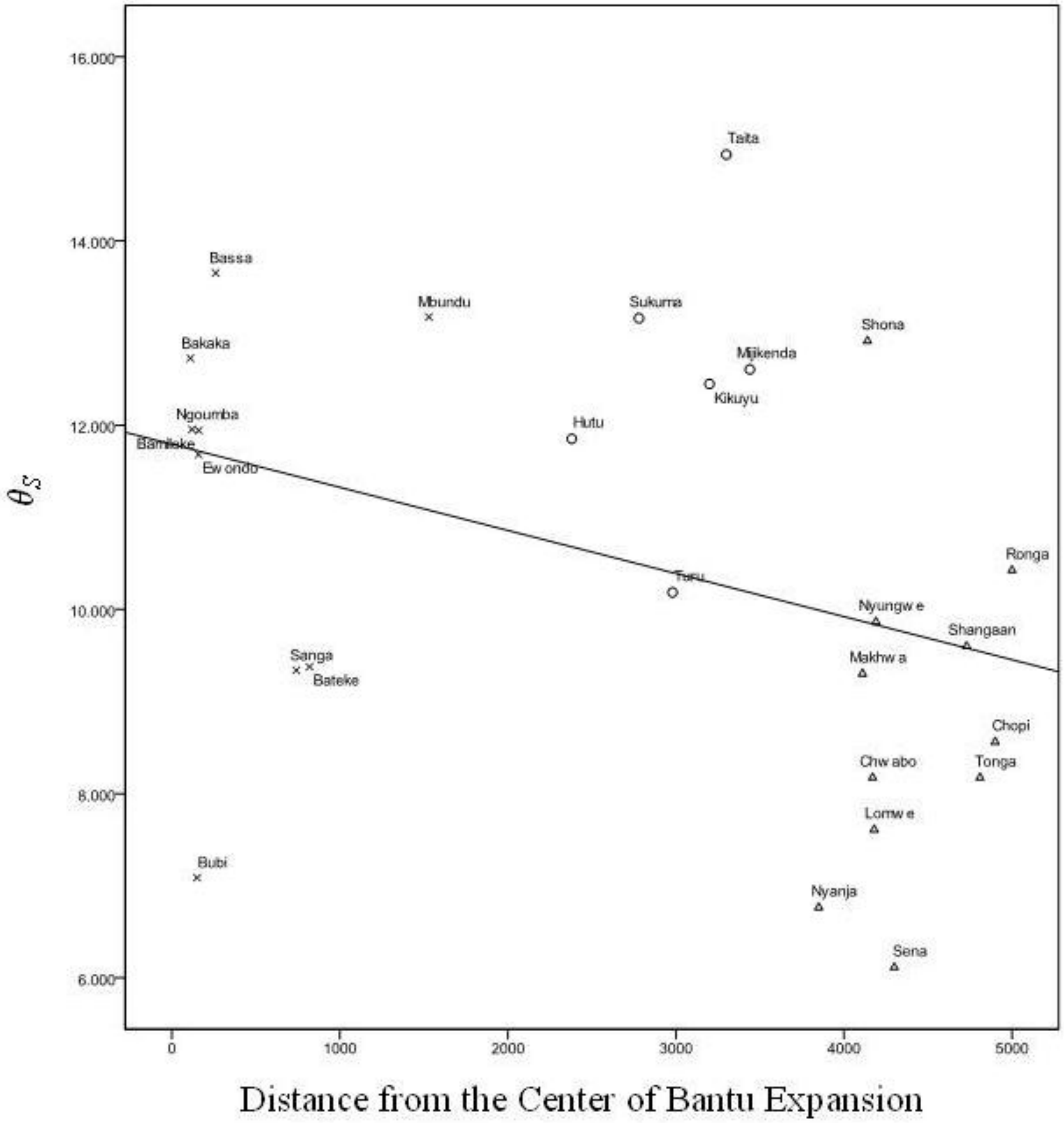

Figure 14 The correlation between θ_S and distance from the center of the Bantu expansion (2-step expansion model) with geographical region marked: Central Africa (x), East Africa (circle), and Southern Africa (triangle)

Asiatic speakers) to 36.270 (between Mijikenda and other Bantu speakers). The Taita and Mijikenda have higher estimated migration rates with other Bantu-speaking groups than with non-Bantu speakers. The other east African Bantu populations have high migration rate estimates with all other groups. The migration rates between Tanzanian Bantu, Nilo-Saharan and Afro-Asiatic populations are modest. The migration rate between Bantu speakers and northeastern African Afro-Asiatic speakers in Ethiopia and Kenya is very low.

4.4 Discussion

4.4.1 Evaluation of Bantu expansion models

I tested three models of Bantu expansion using mitochondrial DNA sequence data in order to better understand the northeastern periphery of the Bantu expansion. These models predict different within-population genetic diversity values and patterns of population subdivision under specific conditions.

Model 1: Expansion without gene flow

If the Bantu-speaking populations experienced a purely demic expansion, expanding rapidly without any interaction, small within population genetic diversity and heterogeneity among them due to founder effects was expected (Austerlitz et al. 1997; Currat and Excoffier 2005). Our data suggests that the expanding Bantu populations may have initially experienced founder effect in East Africa. Some east and southeastern African Bantu-speaking populations are not genetically diverse compared to Central African Bantu populations, and genetic diversity estimates show negative correlation with distance from the center of the Bantu expansion, though the correlation was not significant. The Central African HPGs found among the East Africans do not show evidence of expansion. These observations are consistent with Y chromosome data showing reduced Y chromosome STR diversity and male effective population size toward southeastern Africa (Coelho et al. 2009; Pereira et al. 2002; Thomas et al. 2000). Moreover, East African Bantu populations are rather heterogeneous, possibly because genetic drift may have affected them, including the Turu, who does not exhibit evidence of population expansion.

However, the evidence of genetic drift is not clear in observed mtDNA variation. The correlation between genetic diversity and distance from the center of expansion is not strong because many East African Bantu-speaking populations have large within population diversity and show evidence of population expansion with statistically significant *Fs* and unimodal mismatch distribution. Also, genetic drift did not reduce the genetic variation of the southeastern African Bantu populations significantly, and their $\theta\pi$ is only slightly smaller than that of the Central African Bantu populations.

Model 2: Expansion with gene flow among Bantu-speaking groups

If the expanding Bantu speakers preferably interacted wither other Bantu speakers, high within population genetic diversity, genetic homogeneity among the Bantu-speaking groups, and genetic similarity between East and Central African Bantu populations are predicted. As predicted from this model, East African Bantu and some southeastern African Bantu populations are genetically diverse. The Mijikenda have relatively high frequencies of Central African HPGs and were plotted closely to Central African Bantu speaking populations. They also have a peak of mismatch distribution at the similar position with the Shona, a southeastern African Bantu population. Although these two populations are geographically very distant from each other and they have very different cultural histories, the similar mismatch distribution peak suggests that they were in the same Bantu expansion wave or the Mijikenda maintained gene flow with other Bantu populations. Southeastern African Bantu group appears to be genetically homogeneous and some of these populations are genetically similar to Central African Bantu populations. The significant correlation between genetic distances and geographical distances among the Bantu populations suggest that they are interacting with other Bantu-speaking neighbors, so it is reasonable to believe that the Mijikenda had interaction preferably with other Bantu speakers, such as the Swahili (Willis 1993).

Contrary to the prediction, large AMOVA Φ_{ST} and significant Exact Test *P-values* indicate that the East African Bantu group is not genetically homogeneous and many of them, including the Taita, are not genetically similar to Central African Bantu populations with high frequencies of East African HPGs. Moreover, migration rates estimated between Bantu and non-Bantu east African populations using MIGRATE were large.

Model 3: Expansion with gene flow with neighboring non-Bantu-speaking groups

If the expanding Bantu speakers interacted with non-Bantu speakers in East African, high within population genetic diversity and genetic similarity to non-Bantu East African populations were expected. Many East African Bantu populations exhibit high within-population genetic diversity, especially θ_π, similar to that of non-Bantu East African populations, and mismatch distribution of the Taita and Turkana was similar. High frequencies of East African HPGs found among the East African Bantu-speaking populations can raise the value of θ_π. Previously, East African mtDNA HPG M1a had been found only in one Bantu population, the Sukuma (Knight et al. 2003), but this HPG was also found in the Taita and Mijikenda in this study. Central African and East African Bantu populations are genetically different. Central African Bantu and East

African clusters do not overlap on the MDS plot, and East African Bantu populations are plotted closer to non-Bantu East African populations. High migration rates between Bantu and non-Bantu populations estimated using MIGRATE confirm the pattern observed on the MDS plot as well as observation by Castrì et al. (2008; 2009) showing a great degree of interactions among East African populations.

Surprisingly, the migration rate estimated between the Taita and non-Bantu East African populations were not larger than between the Mijikenda and non-Bantu populations, considering larger East African mtDNA HPG contribution to the Taita than to the Mijikenda. Oral history of the Taita indicates that migrants of various ethnic origins were incorporated to the Taita (Bravman 1998). It is possible that migration rates were smaller, because non-Bantu populations, with whom the Taita had interactions, such as the Massai, are not sampled.

The third model seems to best explain the observed mtDNA variation among the East African Bantu-speaking populations. However, observed mtDNA variation among them exhibits mixed signatures of different demographic models, and none of three models were strongly supported or rejected, suggesting heterogeneous nature of their cultural and evolutionary histories. Central African Bantus initially experienced demographic expansion (Batini et al. 2007), but population size of Bantu ancestors that entered into East Africa may have been small and initially may have experienced founder effects reducing genetic variation of West and Central African HPGs. After the arrival, in East Africa, individual Bantu ethnic groups experienced various evolutionary histories. Different Bantu-speaking populations may have used unique strategies when they settled in East Africa and when they encountered non-Bantu groups. Some of the Bantu populations, such as the Kikuyu and Sukuma who has large θ_k, are genetically diverse because they subsequently re-expanded demographically as they acquired new technologies, such as metallurgy and more productive food producing methods (grain crops and pastoralism) (Phillipson 2005; Schoenbrun 1993). East African Bantu-speaking populations are genetically diverse and heterogeneous, also because the expanding Bantu populations, such as the Mijikenda, maintained close social relationship with other Bantu populations and exchanged marriage partners with them, while others, like the Taita and Hutu, interacted with non-Bantu populations and non-Bantu speakers were incorporated to the Bantu groups through varying degrees of gene flow that occurred both during and after the Bantus spatial expansion (Vansina 1995). Therefore, these three models should be used to explain the observed genetic variation of a population being studied rather than the three competing models.

It is important to note that gene and language have different evolutionary processes, and the rapid expansion of populations or languages has different consequences. The homogeneity among East Bantu languages can be interpreted as evidence for recent rapid expansion (Holden and Gray 2006). Population genetics theory suggests that after expansion without gene flow, populations tend to be genetically differentiated through genetic drift (Austerlitz et al. 1997; Currat and Excoffier 2005). Current mtDNA data do not support rapid expansion of Bantu populations very well. Some Bantu populations interacted with non-Bantu populations and the East African Bantu populations became genetically different from Central African Bantus, while others become genetically different from Central African Bantu populations through genetic drift. Genetic and linguistic homogeneity, on the other hand, can result from inter-ethnic interactions through gene flow and language borrowing.

4.4.2 mtDNA genetic contribution from the Bantu and non-Bantu speakers to modern East African gene pool

One way to evaluate the interaction between the Bantu and non-Bantu populations in East Africa is to examine the genetic contribution of each group to modern East African gene pool. Inferring from HPG frequencies, contribution of non-Bantu mtDNA to East Africa Bantu-speaking populations is possibly large. The Mijikenda have retained Central African Bantu genetic characteristics, and they have large Central African Bantu genetic contributions (~ 47%) and small East African contributions (~ 34%). Many other East African Bantu populations, including the Taita, have larger East African contributions (~ 66% for Taita and ~ 55% for other East African Bantus) and small Central African contributions (~ 18% for Taita and ~ 27% for other East African Bantus). Contrary to East African Bantu populations, contribution of non-Bantu East African mtDNA to southeastern African Bantu populations is smaller (~28%).

On the other hand, the contribution of Bantu mtDNA to non-Bantu-speaking East African population is small. Central African HPGs are rare among non-Bantu East Africans, and when the two most common HPGs among the Central African Bantus (L1c and L3e) were closely examined, these HPGs were found only in three non-Bantu populations. There are two possible explanations for this observed pattern.

First, female effective population size of the East African Bantu-speaking populations has been smaller than non-Bantu East African populations. Afro-Asiatic populations from northern East Africa and Nilo-Saharan from East Africa have larger female effective population size and genetic diversity estimates than East African Bantu populations. The small genetic diversity of two Central African HPGs analyzed suggests the possibility of founder effect during the initial settlement of the Bantu in East Africa. Larger population size of non-Bantu populations in East

Africa can also explain the larger contribution of mtDNA from pre-Bantu expansion local populations to East African Bantu populations than to southeastern African Bantu populations. The Bantu speakers expanded into the heavily populated area in East Africa, but they expanded from East Africa into less populated area in southeastern Africa carrying East African mtDNA, so they could maintain Central African Bantu genetic characteristics.

Second, the gene flows between the Bantu and non-Bantu East African populations were asymmetrical or unidirectional. It is possible that Central African Bantu male genetic contribution to modern East African Bantu populations is larger than female contribution. Y chromosome E3a (E-M2) HPG is relatively common among the East African Bantu populations (42-83%), even though they have high frequencies of East African Y chromosome HPGs (Luis et al. 2004; Tishkoff et al. 2007). If this HPG was brought to East and southeastern Africa by Bantu expansion from Central Africa (Scozzari et al. 1999; Underhill et al. 2001), the data supports the idea that the Bantu populations and languages spread with males (Wood et al. 2005) and local non-Bantu females were incorporated into the Bantu-speaking populations through inter-ethnic marriage.

Since East African Bantu populations have high female genetic contribution from the non-Bantu East Africans, if gene flows were asymmetrical, male Bantu genetic contribution to non-Bantu East Africans is expected to be larger than female Bantu contribution. Contrary to this expectation, the contribution of Bantu Y chromosome to non-Bantu East African populations is small. Two Nilo-Saharan populations, Masai from Kenya and Datoga from Tanzania, have low frequency of Central African Y chromosome HPG E3a (~11-16%) (Knight et al. 2003; Tishkoff et al. 2007; Wood et al. 2005), but this HPG is rare or non-existent among other East African populations (Hassan et al. 2008; Semino et al. 2002). Both mtDNA and Y chromosome data suggest that gene flow was unidirectional from non-Bantu to Bantu populations, but much larger non-Bantu population samples are necessary to examine how much the female and male Bantu speakers genetically contributed to non-Bantu modern populations.

4.5 Conclusion

The Bantu languages and associated culture spread over a large area of sub-Saharan Africa through migration(s) from Central Africa. In East Africa, the Bantu speakers encountered various populations with large population size. The Taita and Mijikenda mtDNA variation reveals that gene flow with other Bantu populations and non-Bantu East African populations was important factor influencing mtDNA variation of the Bantu-speaking populations from northeastern Bantu expansion periphery. Initially the expanding Bantu populations may have experienced founder effect, when they migrated to East Africa. Subsequently, they interacted with other Bantu and non-Bantu populations and exchanged genes. Through gene flow with non-Bantu populations, many East African Bantu populations became genetically similar to non-Bantu populations. Also, through gene flow with non-Bantu populations, they maintained high genetically diversity. The Bantu speakers expanded into southeastern Africa experienced different evolutionary history, because southern Africa was less populated.

5. IMPACT OF FEMALE GENE FLOW ON REGIONAL MITOCHONDRIAL DNA GENETIC PATTERN

5.1 Introduction

Numerous genetic research projects have demonstrated that gene flow and migration have been common in many parts of the world (Barbujani and Belle 2005; Reich et al. 2009; Serre and Pääbo 2004; Tishkoff et al. 2009). For example, the areas around the Mediterranean Ocean: north and northeastern Africa, the Middle East, and southern Europe have particularly complex demographic histories. Ethiopian populations have high frequencies of non-African mtDNA (Kivisild et al. 2004) and Y chromosome haplogroups (HPGs) (Luis et al. 2004; Semino et al. 2002). Sub-Saharan African mtDNA HPG frequencies in the Middle East range from 9 to 34%, with the highest frequencies to date found in Yemen (Cerný et al. 2008; Di Rienzo and Wilson 1991; Richards et al. 2000; Richards et al. 2003). Sub-Saharan African Y chromosome HPGs are rare, but not absent, in the Middle Eastern Arab populations (Cadenas et al. 2007; Richards et al. 2003). Neolithic farmers brought many Middle Eastern HPGs into Europe (Cavalli-Sforza 1993; Richards et al. 1996; Richards et al. 2000; Semino et al. 2004), but Mesolithic foragers contributed Middle Eastern HPGs as well (Battaglia et al. 2009). There is also Y chromosome evidence of more recent gene flow from northern Africa to southern Europe (Cruciani et al. 2004; Cruciani et al. 2007). The presence of European HPGs in the Middle East shows that the gene flow was not unidirectional; migrations from Europe to the Middle East occurred as well (Richards et al. 2000).

Many of these genetic studies focus on identifying patterns of gene flow or population subdivision, but the impacts of gene flow on within-population genetic diversity and the interpretation of past demographic events is addressed less often. Past demographic events, such as population expansions, founder effects, and population bottlenecks, are often inferred from effective population size estimates based on within-population genetic diversity measures. However, these methods generally assume that populations are not subdivided and that mating is random, which is seldom the case in human populations. Human populations are usually subdivided into smaller social/reproductive units, demes, that interact with other demes in complex ways (Cavalli-Sforza et al. 1994; Mielke and Fix 2007). Genetic exchange may occur between demes via sanctioned and unsanctioned means, at different rates, over small or large geographic areas and across cultural and linguistic boundaries (Barth 1969; Eriksen 1993; Fried 1968). Ethnic boundaries are permeable and membership of an individual to ethnic group can be easily shifted through interethnic marriage. Through interethnic marriage, many ethnic groups are heterogeneous and multilingual (Barth 1969; Green and Perlman 1985; MacEachern 2000; Wright 1999).

Ray et al (2003) and Excoffier (2004) demonstrated that when spatially expanding populations exchange genes with other populations at high migration rates (N_em), these populations exhibit the same signature of expansion as populations that experience pure demographic expansions through increases in population size due to high fertility rates. Ray et al (2003) and Excoffier (2004) argue that when human populations are spatially expanding, migrants move to new demes and are incorporated into those demes. Ray et al (2003) demonstrated that, like demographically expanding populations, spatially expanding populations have large migration rates, large within-group genetic diversity measures, large negative

Tajima's D and Fu's F_S, and unimodal mismatch distributions. Excoffier (2004) also demonstrated that forager mtDNA variation fits the expected mismatch distribution better under spatial expansion model.

In this project, I used two population sample data sets, Latin American populations and Bantu-speaking populations from sub-Saharan Africa. They have very different demographic histories and are interesting data sets with which to assess the relative importance of female gene flow and effective population size on within-population genetic diversity. Although central Andean highlanders experienced a demographic expansion, probably the result of population size increases after the introduction of intensive agriculture, gene flow had a homogenizing effect as well (Fuselli et al. 2003; Lewis et al. 2005)(see also Chapter 3). The traditional view of the Bantu expansion suggest that the Bantu speakers experienced massive demographic expansion that replaced pre-existing forager populations, but gene flow among the Bantu-speaking groups and between the Bantu and non-Bantu populations played an even more important role in east African, however, affecting within-population genetic diversity in the northeastern periphery of the Bantu expansion in Kenya (Cavalli-Sforza et al. 1994; Salas et al. 2004)(see Chapter 4).

I used three methods to assess the impact of female gene flow and effective population size on within-population genetic diversity in Latin American and Bantu populations. First, I examined the effects of female gene flow on mtDNA genetic diversity measures (θ_k, θ_s, and θ_π) in different demographic scenarios that varied the rates and timing of migration using a coalescent based computer simulation. Second, I investigated the influence of female gene flow on mtDNA genetic diversity measures (θ_k, θ_s, and θ_π) by comparing them to a genetic diversity measure (Θ) estimated using a maximum likelihood method that factors out the effects of gene flow on the genetic diversity estimates. Finally, I tested whether demographic or spatial expansion models fit the observed mismatch distributions better.

5.2 **Samples and methods**

The Latin American and Bantu populations that were the focus of the previous two chapters form the core of this chapter's analyses. For the MIGRATE analysis, subsets of Latin American and Bantu populations were used (the populations are marked on TABLE II and TABLE X).

The impact of female gene flow on mtDNA genetic diversity was first examined using computer simulation. The purpose of the simulation study was to examine 1) whether small populations can have large within-population sequence diversity when they are intensively interacting with larger populations and 2) which measurements of within-population genetic diversity more likely affected by female gene flow.

Then, the results of simulations were examined using empirical data. The Arlequin population genetics software program (Excoffier et al. 2005; Schneider et al. 2000) was used to estimate within-population genetic diversity, parameter $\theta=2N_f\mu$ (θ_k, θ_S and θ_π), of the Latin American and Bantu populations, and the estimates are listed in the previous two chapters (See Chapter 3; TABLE VI and Chapter 4; TABLE XV). While ancient demographic history affects mean pairwise nucleotide differences (π), number of alleles (k) and polymorphic sites (S), so the two other θ estimators (θ_k and θ_S), are sensitive to recent demographic events (Helgason et al. 2003; Helgason et al. 2000; Rogers 1995; Tajima 1989a). These diversity values were estimated and a mismatch distribution analysis under a demographic expansion model was conducted assuming that the populations are panmictic (un-subdivided randomly mating populations)

TABLE XXIII

TWO MODELS FOR ANALYSES OF WITHIN-POPULATION GENETIC DIVERSITY

	Model 1	Model 2
Assumptions	Panmixia (random mating and no population subdivision)	Population subdivision and gene flow between demes
Genetic diversity Estimates	θ_k, θ_S and θ_π	Θ
Mismatch Distribution Models	Sudden Demographic Expansion	Spatial Expansion

(TABLE XXIII; model 1), but in reality, human populations are often subdivided into smaller demes that are interacting with each other.

Therefore, three θ estimates (θ_k, θ_S and θ_π) and mismatch distributions under the sudden demographic expansion model were compared to the Θ estimates from MIGRATE (Beerli and Felsenstein 2001) and to mismatch distributions under the spatial expansion model (Excoffier 2004) (TABLE XXIII; model 2). These latter two methods assume that gene flow has been taking place, but that population size has been stable. If gene flow significantly impacted within-population genetic diversity, the three θ estimates (θ_k, θ_S and θ_π) and the Θ estimated from MIGRATE should not be correlated well and spatial expansion model should fit better to the observed mismatch distribution better than sudden demographic expansion model.

5.2.1 <u>Computer simulation</u>

A coalescent based simulation program, SIMCOAL (Excoffier et al. 2000; Laval and Excoffier 2004) was used for the computer simulation. The simulations were conducted under a stationary demographic model where population sizes did not change over time for two reasons. First, the main goal of this project is to understand the effect of female gene flow on mtDNA variation, not other factors, such as population expansion. Adding another variable would make the demographic scenarios more realistic, but would make interpretation of the results more difficult. Second, most of the statistical measurements and parameters estimated in this dissertation are based on stationary models, so its use in the simulations makes the results more directly comparable.

I considered six demographic scenarios that have six demes of three different effective population sizes interacting at different migration rates (Fig. 15). Based on previous simulation work done by Fuselli et al. (2003) and my preliminary runs, I chose to use three small demes with N_f=500, two medium sized demes with N_f =1,000, and one large deme with N_f =2,000). The ancestral population of these demes first split into three daughter populations around 11,500 years ago (425 generations ago assuming 27 years per generation). One of them is an ancestral population of two medium sized demes with N_f =1,000, which again split into two around 9,500 years ago (350 generations ago). Another daughter population is ancestral to three small demes with N_f =500, which further split into three small demes about 2,000 years ago (75 generations ago). The last daughter population is the deme with N_f =2,000.

The timing of the demographic events is intended to reflect the general cultural and demographic trends observed in Africa and the New World (Dillehay 2000; Moseley 2001; Phillipson 2005). Populations become more diverse in the end of the Pleistocene and the beginning of the Holocene as a result of cultural adaption to local microenvironments. Around 11,500 years ago, people in the world become less mobile and begin to exploit more local resources. By around that time, all the parts of the New World was occupied by the migrants from Asia. By 2,000 years ago, interactions between different groups of people began increasing. In East Africa, the Bantu speakers who left their homeland in Central Africa were present by this time. Also by 2,000 years ago, population size in the New World started to

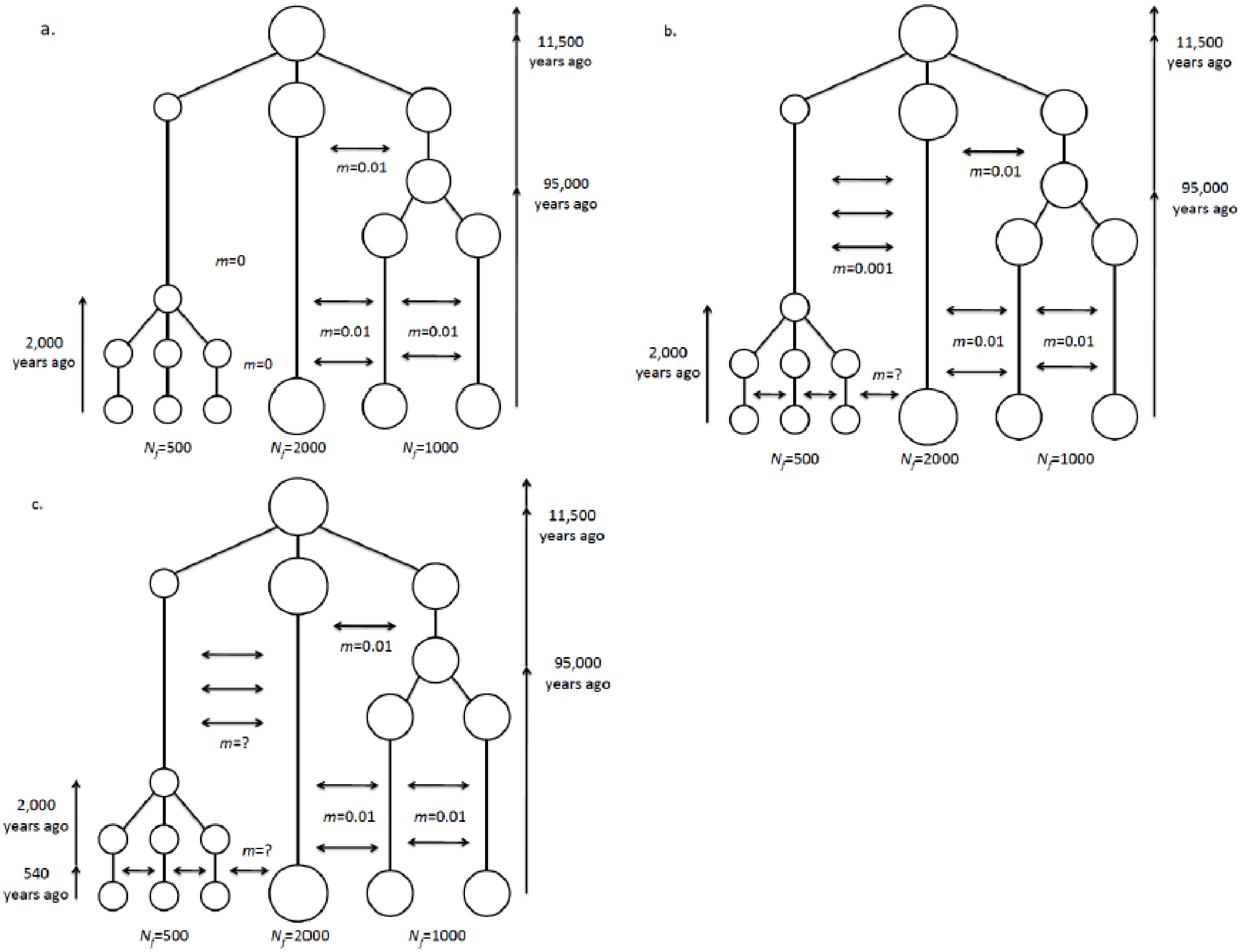

Figure 15 Demes of three different effective population sizes are interacting with different migration rates

increase, because many populations had already acquired food producing technology and food production had been intensified in the Andes and Mesoamerica.

All six demographic scenarios (a, b1, b2, b3, c1, and c2) have the same history of population divergence and same number of demes as described above, but migration rates (*m*) and the timing of its change vary among the seven scenarios (Figure 15).

Scenario a. Three small demes and their ancestral population was completely isolated from each other and from larger demes and never interacted with each other (*m*=0), while larger demes exchanged gene at constant rate of *m*=0.01 with each other (Fig. 15a). This scenario used as a base line for comparison to examine how altering migration rate influence genetic variation.

Scenario b (b1, b2, and b3) Three possible scenarios of interactions between three small demes with larger demes were examined by setting three different migration rate (*m*=0.001, 0.01, and 0.1) (Fig. 15b). The ancestral population of three small demes exchanged genes with other populations at very small rate (*m*=0.001) and larger demes exchanged gene at rate of *m*=0.01 with each other. I examined impacts of gene flow on genetic variation of small demes, when small demes exchanged genes with larger demes at very small (*m*=0.001; Fig. 15b1), small (*m*=0.01; Fig. 15b2), and large (*m*=0.1; Fig. 15b3) migration rate, after the small demes split from their ancestral population 2,000 years ago.

The migration rates among the small demes are proportion of population replaced by migrants. When *m*=0.1, 10% of population of a deme is replaced by migrants, so a deme is sending migrants to each of five other demes with *m*=0.02 and receiving migrants from each deme with *m*=0.02.

Senario c (c1 and c2). I considered the situation where small demes were completely (Fig. 15c1) or relatively (Fig. 15c2) isolated, but the nature of interactions between small isolated demes and larger demes changed about 540 years ago (20 generations ago) after the emergence of large state societies that swept through large part of continent and/or the European contact. Since then, small demes intensively interacted with other small demes and larger demes with large migration rate (*m*=0.1), so 10% of the population of a deme is replaced by migrants.

This scenario of complete isolation of small demes (c1) is same as the first scenario (Fig. 15a), where small demes did not interact with other demes, except for last 540 years when they interacted intensively. The scenario of relatively isolated small demes (c2) is similar to the second set of scenarios (Fig. 15b1), where the small demes interacted with larger demes with very small migration rate (*m*=0.001), but in this case, small demes interacted with other demes intensively in last 540 years.

Under each scenario, 1,000 genealogies were constructed. I set up the simulation runs for 40 sampled sequences from each deme and for 328 nucleotide long sequences. A mutation rate of 0.002 per generation over the whole sequence was assumed. The chosen mutation rate is similar to one of more conservative mutation rate estimates based on pedigree (Sigurðardottir et al. 2000). Following Meyer et al. (1999), mutation rate gamma distribution shape parameter α=0.26 and 10 classes of mutation rate were used. I used island model of migration similar to the n-island model used for MIGRATE. Output of the simulation were analyzed with Arlequin (Excoffier et al. 2005; Schneider et al. 2000). As described above, the program was used to estimate within-population genetic diversity, parameter $\theta=2N_f\mu$ (θ_k, θ_S and θ_π).

5.2.2 MIGRATE analysis

MIGRATE is a coalescent based maximum likelihood method. It uses an n-island model of migration with unequal population sizes and asymmetrical migration rates (Fig. 16). It estimates 1) the migration rates ($2N_fm$) and 2) the within-population genetic diversity ($\Theta = 2N_f\mu$) without the effects of gene flow (Beerli and Felsenstein 2001). The Θ should reflect real effective population size better than θ because the effects of gene flow are factored out. I reasoned that if Θ and θ are highly correlated, that would indicate that female effective population size is the most important determiner of genetic diversity, but if θ and Θ are not well correlated, then gene flow is an important source of genetic diversity.

Because only up to seven populations can be included in each MIGRATE run for the data set with a single marker, the MIGRATE runs were set by grouping the populations into six regions and I ran MIGRATE for each regional group. Each run usually consists of seven populations. For each single region, two to four populations from that region were run together with other regional groups where those populations were pooled together and treated as one large population, a potential source of migrants.

A subset of Latin American populations (N=15) with sample sizes greater than 25 individuals were selected to represent six regional/linguistic groups in western South America and Central America: south-central Andes, north-central Andes, southern Andes, Gran Chaco, northwestern lowland South Americans, and Central Americans. Populations with small sample sizes are not suitable for this analysis because reliable estimates cannot be obtained when using relatively short sequences from a single marker. Three population samples were included from

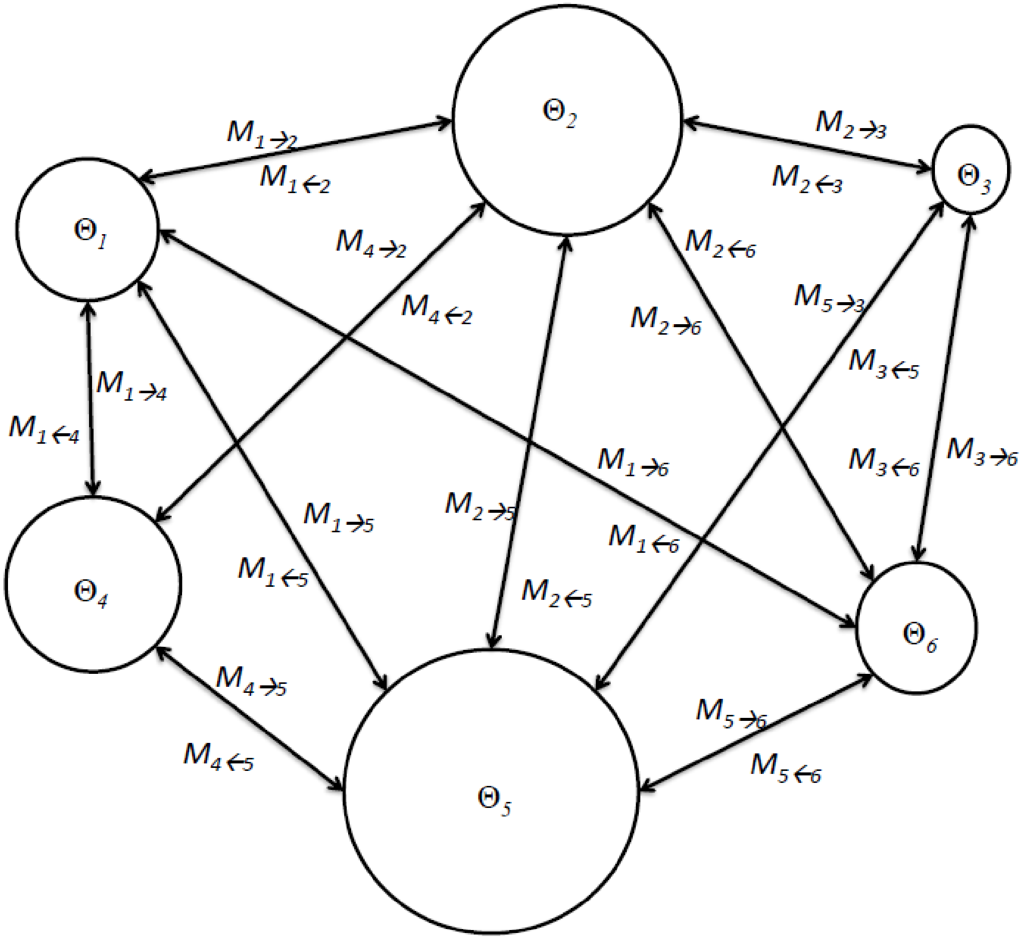

Figure16 n-island model used in MIGRATE. MIGRATE assumes unequal population size ($\Theta=2N_f\mu$) and asymmetrical migration rate ($M=m/\mu$). Multiplying M and Θ gives $2N_fm$.

each of the following regions: Central America, northwestern lowland South America, and the Gran Chaco area. Two population samples were included from each of the north-central, south-central, and southern Andean region. Regional/linguistic groups must have at least two populations for comparison, so the Moxo from the lowland Bolivia had to be excluded from the analysis, even though they live near the Aymara. I focused on western South America where the previous studies found evidence of increased female gene flow (see chapter 3).

Similarly, a subset of Bantu-speaking populations (N=19) were selected to represent three geographic groups, seven Central African, four East African, and eight Southeastern African Bantu populations. The nineteen Bantu populations were grouped into five (East African, patrilocal Central African, matrilocal Central African, patrilocal Southeastern African, and matrilocal Southeastern African Bantu populations) based on geographical origin and kinship structure. Five Bantu populations were excluded from the MIGRATE analyses. The Kikuyu were included during experimental MIGRATE runs, but were removed because of inconsistent Θ estimates, perhaps because their diversity values were high and their sample size was relatively small. The Sukumra, Chopi, and Tonga were excluded because their kinship structure could not be determined from written sources. The Mbundu was excluded because of the large geographic distance from other Central African Bantu populations.

The Θ estimates were obtained using averages of more than three independent runs in each regional set. Each run have 10 short chains (10,000 genealogies per chain) and three long chains (100,000 genealogies per chain) with increments of 20 and 200 steps respectively. The first 100,000 trees in each chain were discarded. Instead of sampling more genealogies, Metropolis coupled Markov Chain Monte Carlo, or ‘heating’ was used to explore a wider genealogical space by setting four temperatures (1, 1.5, 3, 6) and long chains were replicated.

5.2.3 Mismatch distribution

Mismatch distributions are another method to analyze within-population genetic diversity which allows us to evaluate whether observed within-population variation can better be explained by gene flow (spatial expansion) or effective population size (demographic expansion). Mismatch distributions are analyses of nucleotide differences between sequences in a single population. Number of nucleotide differences between two sequences and its frequency can be graphically represented. Under a sudden demographic expansion model, a unimodal mismatch distribution is interpreted as evidence of demographic expansion assuming that populations are panmictic (Rogers and Harpending 1992; Slatikin and Hudson 1991). Many large agricultural populations are genetically diverse and have unimodal mismatch distributions, while foragers whose population size has stayed consistently small are genetically not diverse and have multimodal mismatch distributions (Excoffier and Schneider 1999; Rogers 1995; Watson et al. 1996).

A mismatch distribution under a spatial expansion model, on the other hand, assumes that the population is exchanging genes with another population of infinite population size, without population growth (Excoffier 2004). In other words, spatially expanding populations are genetically diverse because of high levels of gene flow resulting in the same unimodal distribution as the demographic expansion. Ray and his colleagues (2003) analyzed effects of gene flow between subdivided populations on within-population genetic diversity and demonstrated that when the migration rate (*Nm*) is large, populations have unimodal mismatch distributions, a genetic signature of expansion similar to demographic expansions.

A Sum of Square Deviation (*SSD*) between observed and expected (simulated) mismatch distribution is used as a test statistic to determine whether a demographic expansion or spatial expansion model best explains the observed mismatch distribution of populations. The expected mismatch distribution for each demographic model is generated through coalescent simulation. The parameters estimated from the empirical data were used to simulate to test the hypothesis that estimated parameters under the model are real ones.

5.3 Results

5.3.1 Results of simulations

The results of the simulations show that female gene flow has considerable effects on within-population genetic diversity of small demes, especially when the small demes have prolonged interactions. As migration rates increase, the differences in genetic diversity between small and large demes decrease, but this effect is more notable on θ_S and θ_π estimates.

Although the standard deviation (S.D.) of the θ_k for the large deme and the θ_π for all of the demes are very large, the simulation results suggest that increased gene flow had greater impact on all of the θ estimates of the small demes (N_{fe} = 500) than the larger demes, and that the θ values of the small populations grew as their migration rates with larger demes increased (TABLE XXIVa). When the migration rate was less than 1%, the θ estimates of the small demes were much lower than those of the medium and large demes, so the ratios of the small to large deme θ values are large (TABLE XXV). Conversely, when 10% of the population in small demes was replaced, the genetic diversity differences between the small and large demes were reduced and the ratio of their θ became small.

Increasing the migration rates among the demes affected the estimates of θ_S and θ_π more than θ_k. The large and medium sized demes have much larger θ_k than small demes (TABLE XXIVa and XXIVb), so even when the migration rates are high, the θ_k ratios between the small and large demes remain large. Conversely, as the migration rates increased between the small and large demes, the differences in their θ_S and θ_π decreased. Most interestingly, the θ_π of the small demes began to increase quickly, even when migration rates were relatively small.

Increasing gene flow for 20 generations (540 years) after complete or relative isolation increased the genetic diversity of small demes, and θ_S and θ_π among the small demes became

TABLE XXIV

RESULTS OF COMPUTER SIMULATION SHOWING AVERAGE OF θ OVER 1000 SIMULATION RUNS AND STANDARD DEVIATION (S.D.) GROUPED BASED ON THE SIZE OF DEMES

a. Results for small demes (N_{fe} = 500)

Scenarios	Migration rate	θ_k	S.D.	θ_S	S.D.	$\theta\pi$	S.D.
A	0%	2.204	1.055	3.140	1.669	4.663	4.089
b1	0.1%	3.129	1.284	4.648	1.781	6.999	4.884
b2	1%	5.061	1.935	5.802	1.836	7.948	4.747
b3	10%	7.229	2.654	6.393	1.759	8.211	4.460
c1	0→10%	5.637	2.320	5.751	1.184	6.592	4.301
c2	0.1→10%	6.377	2.503	5.992	1.809	7.532	4.412

b. Results for medium demes (N_{fe} = 1,000)

Scenarios	Migration rate	θ_k	S.D.	θ_S	S.D.	$\theta\pi$	S.D.
A	0%	8.641	3.008	6.718	1.753	8.664	4.695
b1	0.1%	9.324	3.276	6.948	1.740	8.963	4.709
2b	1%	9.485	3.354	7.006	1.779	9.129	5.012
b3	10%	9.438	3.405	6.913	1.727	8.777	4.648
c1	0→10%	10.432	3.783	7.133	1.748	9.132	4.763
c2	0.1→10%	10.126	3.634	6.962	1.765	8.753	4.625

c. Results for large deme (N_{fe} = 2,000)

Scenarios	Migration rate	θ_k	S.D.	θ_S	S.D.	$\theta\pi$	S.D.
A	0%	13.252	4.710	7.574	1.705	9.163	4.719
b1	0.1%	13.685	4.971	7.683	1.696	9.404	4.816
b2	1%	14.142	5.261	7.755	1.796	9.568	4.974
b3	10%	12.744	4.726	7.498	1.709	8.985	4.711
c1	0→10%	15.118	5.612	7.887	1.726	9.573	4.933
c2	0.1→10%	14.567	5.392	7.657	1.741	9.166	4.586

TABLE XXV

RATIOS OF AVERAGE θ BETWEEN SMALL AND LARGE DEME

Scenarios	θ_k	θ_S	$\theta\pi$
A	0.166	0.415	0.509
b1	0.258	0.605	0.744
b2	0.358	0.748	0.831
b3	0.567	0.853	0.914
c1	0.373	0.729	0.689
c2	0.438	0.783	0.822

similar to θ_S and θ_π values for the larger demes. However, the θ of small demes were still smaller than values obtained for larger demes with large ratios of θ between small and large demes.

5.3.2 Comparison of Θ to θ_k, θ_S, and θ_π

The results of the computer simulation suggest that female gene flow has greater effect on the estimates of θ_π, and possibly θ_S, than on θ_k. Next, I continued my examination on the impact of female gene flow on θ estimates by comparing θ_k, θ_S, and θ_π to Θ estimates, which attempt to control for the effects of gene flow. Overall, Θ correlates well with θ_k, but not θ_S, and θ_π, and Θ correlates with θ_k and θ_S stronger in Latin American than in Bantu populations, with larger correlation coefficients.

The Θ estimates were strongly correlated with θ_k and θ_S, but not with θ_π in 15 Latin American population samples, which supports the simulation results indicating that θ_π estimates will be most affected by gene flow (TABLE XXVI). Close examination of the estimated θ and Θ values shows that many central Andeans have smaller θ_π values than expected, while some

TABLE XXVI

PEARSON'S CORRELATION BETWEEN Θ AND OTHER THETA ESTIMATES (*r* AND *P*-VALUES)

	θ_k	θ_S	$\theta\pi$
Latin Americans	0.914 (0.000)	0.736 (0.001)	0.269 (0.280)
Bantus	0.668 (0.002)	0.444 (0.057)	0.345 (0.147)

small lowland populations have higher θ_π than expected (See Supplemental Data Table S1). The twoSouth-Central Andean populations have high Θ, θ_k, and θ_s values, but smaller θ_π The two populations from North-Central Andes have similarly high Θ, θ_k, and θ_s values and low θ_π values, but their θ_π values are not as reduced as in the South-Central Andean populations. The Pilaga who are foragers, have large values for all four estimates. The two northwestern lowland populations, the Wounans and the Cayapa, have comparatively low Θ and θ_k values as expected, but unexpectedly high θ_π.

The Θ estimates were significantly correlated with θ_k values among the Bantu-speaking populations, but were not correlated with θ_S or θ_π values (TABLE XXVI). The east African Bantu populations have small Θ, but large θ_S and θ_π values, and these east African Bantu populations had an impact on the correlation of the estimates (See Supplemental Data Table S2). Despite their small current population size (~213,000), the Taita have relative large Θ values, but they have the largest θ_S and the third largest θ_π. Other east African Bantu-speaking populations had smaller Θ values. The Mijikenda had the third lowest Θ value, but their θ_k, θ_S, and θ_π were large, and they have the fifth largest θ_S. Another east African Bantu-speaking population, the Hutu, also showed a similar pattern. They have a relatively low Θ value, but they have moderately high θ_k and θ_S and the second largest θ_π.

5.3.3 Spatial expansion vs. demographic expansion models

The results of the MIGRATE analyses suggest that female gene flow did influence within-population genetic diversity measurements in both the Latin American and Bantu population samples I analyzed, so I next used mismatch distribution analysis and asked whether each individual population sample in both regions best fit the demographic expansion model which does not account for gene flow in subdivided populations or the spatial expansion model which assumes that gene flow was the significant contributor to influence within-population genetic diversity.

A few clear trends were observed (TABLE XXVII). First, the demographic expansion model was rejected more often than the spatial expansion model. Only one Bantu population rejected the demographic expansion model more than the spatial expansion model, but this pattern is the most prominent among the genetically less diverse Latin Americans, especially foragers (see Table S3 and S4 for more details). Six out of ten forager populations (60%) rejected the demographic expansion model, but none of them rejected the spatial expansion model. Second, a contrasting pattern was observed among the populations with intensive agricultures. Three out of nine (33.3%) intensive agriculturalists rejected the spatial expansion model, but none of them rejected the demographic expansion model.

Third, both models were rejected among the Latin Americans more often than among the Bantu-speaking populations. I included forager populations in Latin American data set, but the all the Bantus are food producers, so I calculated rejection rates for Latin American food producers. Only 11.5% of Bantu populations have significant SSD *P-value* showing the

TABLE XXVII

PERCENTAGE (AND NUMBER) OF POPULATIONS THAT REJECTED THE DEMOGRAPHIC AND SPATIAL EXPANSION MODEL

	Total (N)	Demographic Expansion Model	Spatial Expansion Model
All Latin Americans	40	40.0% (16)	20.0% (8)
Intensive Agriculturalists	9	0.0% (0)	33.3% (3)
Horticulturalists	21	47.6% (10)	23.8% (5)
Foragers	10	60.0% (6)	0.0 (0)
Food Producers			
Latin American	30	33.3% (10)	26.7% (8)
Bantu	26	11.5% (3)	7.7% (2)

The Latin American populations were further subdivided 1) intensive agriculturalists (highland Andeans and Quiche Mayans), 2) horticulturalists, and 3) foragers. Food producers include intensive agriculturalists, pastoralists, and horticulturalists.

deviation from demographic model and only 7.7% of Bantu populations have significant SSD *P*-values under the spatial expansion model. On the other hand, 33.3% of Latin American food producers rejected the demographic expansion model and 26.7% of them rejected the spatial expansion model.

5.4 Discussion

5.4.1 Evaluation of demographic models using Latin American and Bantu data sets

In this project, I compared the results from two types of methods, 1) methods based on a model that does not account for gene flow (assumes that populations are not subdivided and randomly mating) and 2) methods based on a model that does account for gene flow (assumes that populations are subdivided). The results suggest that panmixia for human populations should not be assumed when using within-population genetic diversity to infer the effective population size because gene flow will affect within-population genetic diversity measurements (θ_k, θ_S, θ_π, and mismatch distribution). I also observed a general tendency to reject demographic expansion model more than spatial expansion model. These results are consistent with the results in the two previous chapters that showed the importance of female gene flow affecting within-population genetic diversity estimates of Latin American and Bantu populations.

The results from this chapter show that female gene flow affects θ_π and possibly θ_S values at the regional level. Female effective population sizes (Θ) estimated removing the effects of female gene flow were not correlated with θ_π in Latin American samples or with θ_S and θ_π in Bantu populations. The simulation results also illustrate that when migration rates are large, the differences in $\theta\pi$, and also θ_S to a lesser extent, between small and large demes become small. The relative effects of genetic drift and gene flow on sequence variation can explain these simulation results. Helgason and colleagues (2003; 2000) suggested that genetic drift eliminates rare variants reducing number of alleles (*k*) and polymorphic sites (*S*), while admixture increases mean pairwise nucleotide differences (π). When the small demes are interacting with other demes, the effect of drift is strong, so it reduces the number of alleles and polymorphic sites, keeping θ_k and θ_S small, but the mean pairwise nucleotide differences ($\theta\pi$) is not reduced. Therefore, θ_π and possibly θ_S are no longer reliable to infer effective population size of populations.

I considered demographic scenarios with stationary population size, but if small demes are interacting with large demes that have experienced population expansion, the small demes may acquire the genetic signature of the population expansion that the large demes experienced. When I examined mtDNA variation of Bantu-speaking populations living at the northeastern periphery of the Bantu expansion in Kenya, female gene flow was an important factor in making small east African Bantu populations genetically diverse (Chapter 4). East African Bantu-speaking populations such as the Taita, Hutu, and Sukuma have large θ_S and $\theta\pi$ values but relatively small Θ and θ_k values. These East African Bantu populations have large non-Bantu genetic contributions through female gene flow with non-Bantu-speaking east Africans who are genetically diverse with very different HGP compositions and whose populations have large effective population sizes.

Similarly, some of lowland Latin Americans have large θ estimates. Two small Latin American populations, Pilaga (foragers) and Moxo (horticulturalists), have larger θ_S and θ_π than would be expected from their population size and subsistence. Especially, the θ_π estimated for these genetically diverse lowland populations was similar to or larger than the θ_π estimated for the Central Andeans. On the other hand, the Θ estimated using MIGRATE for many genetically diverse lowland populations was smaller than Central Andeans. These observations are consistent with ideas that female gene flow, possibly with the Central Andeans, increased within population genetic diversity among these lowland populations from the western South America (Chapter 3). The interactions of these small populations with surrounding populations may have intensified during the last several hundred years after European contact (Braunstein and Miller 1999), but their very large θ_S and θ_π estimates suggest that they were interacting with other populations before the European contact.

The results of the study are also consistent with previous simulation studies by Ray et al. (2003) and Excoffier (2004) that examined effect of spatial or range expansion on within-population diversity. They demonstrated that spatial and demographic expansions leave similar genetic signatures. As migration rates increase, demes become genetically more diverse and show signatures of population expansion. Their simulation results show that gene flow increases both the average number of pairwise differences (π) and the number of segregating sites (S), but that S is more drastically affected, causing Tajima's D and Fu's Fs to be significantly negative and for mismatch distributions to become unimodal. Central Andeans have large estimates of the migration rate, M, estimated using mismatch distributions (Chapter 3), suggesting that they have large θ_S with genetic evidence of a population expansion, partly due to increased female gene flow. These results agree with the simulation analyses conducted by Fuselli and colleagues (2003). Their simulation analysis shows that genetic variation in the Andes can be explained with dense populations living in the area and high migration rates before and after the European contact. However, Ray et al. (2003) did not consider the situations where demes of unequal sizes are interacting. In my simulation study, when small demes interact with large demes with small to large migration rates, gene flow has a greater impact on the small populations than on the large demes. Small populations, such as Moxo and Pilaga, can have large sequence diversity because an influx of new variants offsets the effects of drift. If small demes have more prolonged or intensive interactions with large demes, their θ estimates will resemble the large demes.

Ray et al. (2003) and Excoffier (2004) also argued that the spatial expansion model can explain the demographic history of human populations better than the demographic expansion model, even for forager populations that often reject the demographic expansion model (Excoffier and Schneider 1999). Our results support their arguments by demonstrating that no Latin American foragers rejected the spatial expansion model, but 60% of them rejected the demographic expansion model.

Among Bantu-speaking populations, neither sudden demographic expansion nor spatial expansion model was rejected, and there are two possible explanations. First, the sudden demographic expansion model was not rejected, because the Bantu populations experienced demographic expansion, but this model do not fit the data better than the spatial expansion model, since population growth in sub-Saharan African was gradual and African population did not experience massive demographic expansion that the populations with intensive agriculture did. Second, the spatial expansion model was not rejected, because as cultural anthropologists and historians have suggested, in Africa, gene flows through interethnic marriage had been common and many ethnic groups were social reorganized after the European colonization (Chimhundu 1992; Fried 1968; Niehaus 2002).

Both models were rejected in four Latin American populations. There could be unknown demographic factors that erased the evidence of past demographic expansion, but it is possible that the founder effect during the settlement of the New World and/or bottle neck after the

European contact more strongly affected these populations than the others (Excoffier and Schneider 1999).

5.4.2 Female gene flow vs. effective population size in the regional level

The objective of this project was to evaluate the impact of increased female gene flow and effective population size on mtDNA variation among the Bantu and Latin American populations. I predicted that female gene flow would have a greater role influencing mtDNA variation than female effective population size. However, the results of our analyses suggest that both female gene flow and effective population size were important factors affecting mtDNA variation, and the effects of these factors were detectable in different measurements.

First, because female gene flow was important factor, demographic models that account for gene flow, such as spatial expansion model, are more appropriate to explain human demographic history than models that do not. I have shown that female gene flow affects θ_π and possibly θ_S at the regional level. The results also suggest that female gene flow was an important factor affecting mtDNA variation among the Bantu more than the Latin American populations. Among Bantu populations, Θ was significantly correlated only with θ_k, but not with θ_S or θ_π, while Θ was significantly correlated with θ_k and θ_S among the Latin Americans.

Female effective population size was also important factor affecting within-population genetic diversity and Θ was significantly correlated with θ_k among the Latina American and Bantu populations. The results of simulation also show that even when migration is large, the difference in θ_k between small and large demes is still large, so θ_k should reflect effective population size. Female effective population size was particularly important factor among the Central Andeans, and Θ was larger among the Aymara and Quechua than other Latin American population, supporting high θ_k. Among these populations, the sudden demographic expansion model of mismatch distribution was not rejected, but three of them (33.3%) rejected the spatial expansion model. Also, some populations in the Bantu expansion periphery with large current population size such as the Kikuyu, Sukuma, Ronga, and Shona may have experienced more recent regional demographic expansion, increasing their θ_k. The Ronga and Shona also have large estimated Θ estimate.

Moreover, the model of sudden demographic expansion fits Central Andean demographic history better than the spatial expansion model, which suggests that although female gene flow was an important factor, increased fertility resulting from the introduction of intensive agricultural techniques was an important factor in increasing within-population genetic diversity in the Central Andes. If the Aymara and other Central Andean populations grew slowly and incorporated other genetically different populations, a spatial expansion model should not have been rejected. If their population size grew slowly and they interacted with genetically similar populations, either demographic model should have been rejected or there should have been no evidence of sudden population expansion. However, none of the Central Andeans departed from the expectations of the demographic expansion model, and the spatial expansion model, while compatible with some of the central Andean groups, did not fit the Ancash and Arequipa Quechua. The fact that the migration rate, *M*, could not be obtained from three south-central Andeans (Arequipa, Aymara Puno, Aymara La Paz) also supports demographic expansion model because reliable estimates of *M* cannot be obtained unless population size is stable. These results suggest that agricultural intensification greatly affected demography and mtDNA variation.

5.4.3 Limitations and applicability of this study

The limitations of the analytical methods used in this study have to be noted, first because human populations violate some important assumptions of the models. MIGRATE assumes that there are no unsampled populations exchanging genes with sampled populations. Beerli (2004) examined the effects of unsampled populations on these methods. He found that effective population size is upwardly biased under these conditions. Beerli suggests that more accurate estimation can be obtained by running the program with many populations at the same time. However, human populations tend to interact with many different populations and many of them are not sampled. The unexpectedly high estimates of Θ obtained from the Quechua (Tayacaja) and the Pilaga, a forager population, may be explained by this problem. MIGRATE also assumes that population size and migration rate do not change over time, but most human populations have experienced demographic expansions or bottlenecks at some point in their histories. If recent demographic expansions inflate Θ estimates, then the Aymara, Quechua, and some Bantu-speaking populations may have inflated Θ values. Rapidly expanding populations tend to accumulate rare haplotypes and polymorphisms without significantly affecting pairwise nucleotide differences, so they should have large estimates of Θ, θ_k, and θ_S relative to θ_π. The populations that have experienced recent bottlenecks may have smaller estimates of Θ. If these populations have lost rare haplotypes without losing the more common haplotypes of each HPG, as suggested by Helgason et al. (2003), they may have high θ_π values, but lower estimates of Θ, θ_k, and θ_S (e.g., Cayapa and Wounan).

Another limitation of this project is very simple general model of demographic history that could be applicable to many parts of the world. Actual θ parameters estimated should differ significantly depending on the areas of the world, because populations in different parts of the world experienced very different demographic events. For example, the Bantu-speaking populations have much larger θ than Latin Americans (Table S1 and S2). Also, in the real world, populations are interacting with more than five populations and some of them may have experienced pure demographic expansion.

5.5 Conclusion

In this chapter, I assessed whether female gene flow or effective population size affect mtDNA variation more than the other, and demonstrated that the demographic model that does not account for gene flow will provide only an inadequate understanding of human demographic history because of gene flow and effective population size are conflated. Although the greatest effects of effective population size are detectable among the Latin American populations with intensive agriculture, at the regional level, female gene flow greatly affected within-population genetic diversity of all Latin Americans. Among the Bantu-speaking populations, female gene flow was a more important factor influencing mtDNA diversity than effective population size, but neither model of mismatch distribution was rejected more often than the other. Unfortunately, because human populations violate many important assumptions that analytical methods based on, this issue should be addressed further using methods that incorporate more complex evolutionary processes with less built-in assumptions.

6. HOW DOES KINSHIP STRUCTURE EXPLAIN MITOCHONDRIAL DNA VARIATION?

6.1 Introduction

Genetic evidence of male and female asymmetrical demographic history is widely reported (Besaggio et al. 2007; Hammer et al. 2008; Keinan et al. 2009; Oota et al. 2001; Pérez-Lezaun et al. 1999; Ségurel et al. 2008; Wood et al. 2005) and geneticists have debated whether asymmetrical demographic history is the result of small male effective population size due to polygyny or increased female gene flow. Seielstad and colleagues (1998) compared Y chromosome variation to autosomal and mtDNA variation and found that Y chromosome data shows greater population differentiation. They believe that restricted male gene flow compared to female gene flow due to patrilocal residence is the cause of the population differentiation. Wilder et al. (2004b), on the other hand, show that Y chromosome variation is not more differentiated than mtDNA variation suggesting that there is no significant difference in male and female gene flow pattern. Instead, Wilder and colleagues (Hammer et al. 2008; Wilder et al. 2004a; Wilder et al. 2004b) have argued that reduced male effective population size due to polygyny and other factors is the more important factor than increased female gene flow.

In many regional level genetic studies, higher mtDNA diversity than Y chromosome diversity has been found in patrilocal populations, which the researchers have attributed to higher female gene flow through marriage exchange and postmarital residence patterns (Besaggio et al. 2007; Chaix et al. 2007; Hamilton et al. 2005; Oota et al. 2001). Women in societies with patrilineal descent systems and patrilocal post-marital residence are more likely to leave their home villages as adults, while men tend to stay in their natal homes near their paternal kin. These social practices lead to relatively high rates of female gene flow and low rates of male gene flow, which result in asymmetries in mtDNA and Y chromosome genetic diversity. The opposite patterns of diversity are seen in societies with matrilineal descent system and matrilocal post-marital residence pattern where the men are more likely to leave their natal villages.

In this project, I will examine how kinship structure affects mtDNA variation in a large number of population samples from two regions of the world, Latin America and Africa. Comparing mtDNA and Y chromosome variation from the same populations is the ideal approach (Jorde et al. 2000; Stoneking 1998), but it severely reduces the numbers of populations available for comparison. For example, out of six Bantu-speaking East Africa populations that mtDNA HVRI sequence data is available, Y chromosome variation has been analyzed in only two of them. Furthermore, direct comparisons of mtDNA and Y chromosome diversity can be confounded by a number of factors such as differences in mutation rates.

I sorted the population samples into three categories, patricentric, flexible, and matricentric kinship structures. The terms are a highly simplified version of Burton and his colleagues (1996) methods where they scored a number of variables to place societies along two axes: social organization and kinship terminology. In their scheme, patricentric societies are organized around males and are generally associated with patrilocal residence, patrilineal descent, and polygyny. Matricentric societies are characterized by matrilocal post-marital residence, matrilineal descent, and monogamy. Neither term describes many forager populations who practice bilateral descent or exhibit marital residence flexibility, however (Marlowe 2004; Martin and Voorhies 1975). In a flexible social system, couples may alternate between matrilocal and patrilocal residence during marriage or the post-marital residence pattern and other marriage practices may be open to choice. Although assigning societies into three simple categories ignores the variability that exists currently among different societies as well as existed in the past, I use it here to simplify the model and permit an initial assessment of the effects of kinship structure on mtDNA variation.

From this kinship structure model, I can infer the pattern of migration, or rate of gene flow between populations or demes, female effective population size, and then genetic diversity. Patricentric societies should have higher female migration rates because marriage exchange networks serve to cement relationships among kin groups and may extend beyond tribal and ethnic boundaries. Female effective population size may be larger than male effective population size if polygyny is practiced and some men have multiple wives while other men do not marry. In societies with flexible kinship systems, if both men and women migrate, female migration rates may be as high as patricentric rates. However, male and female effective population size may not be significantly different because of the prevalence of monogamy. In the matricentric societies, women stay in their natal societies and men move to live with their wife's family, so male migration rates will be higher, but the male and female effective populations sizes will be similar because matricentric societies generally practice monogramy (not polyandry).

Marriage practices (polygyny, monogamy or polyandry) may cause differences in male and female effective population size, but they are unlikely to cause large differences in female effective population size between patricentric and matricentric populations. Research in anthropological demography and reproductive ecology has shown that while first wives in polygamous marriages are reproductively more successful than wives in monogamous marriages, the reproductive advantages to women in general even out in both types of marriage, because the second and later wives in polygamous marriages are reproductively less successful than monogamous wives. Second and later wives in polygamous marriage have the smaller number of children born (Gibson and Mace 2007; Josephson 2002), children with the higher risks of mortality and malnourishment (Hadley 2005; Strassmann 1997), and the reduced maternal health (Bove and Valeggia 2009).

Moreover, population subdivision can be inferred from kinship structure. If matricentric populations have lower rates of female migration, they will likely lose mtDNA genetic variation through genetic drift and become genetically differentiated. Genetic drift should have less effect on mtDNA variation in patricentric and flexible populations because the women are likely to migrate at higher rates and over larger areas.

In order to examine the effect of kinship on mtDNA variation, first, I compared the differences in migration rates and genetic distances between patricentric and matricentric populations. Second, I examined the correlation between within-population genetic diversity and non-genetic factors (kinship structure, current ethnic population size, and subsistence strategies). I expected to observe higher female migration rates and higher within-population genetic diversity values (θ) in patricentric populations when compared to matricentric populations. On the other hand, if there is skewed male-female sex ratio among patricentric populations, patricentric populations should have large estimated female effective population size (Θ), even if some of them have small census population size, and the effect of polygyny should be observed among the Bantu-speaking populations who commonly practice polygyny.

6.2 Samples and methods

6.2.1 Samples

The Bantu and Latin American population samples from the previous two chapters were used in the analyses. There are 16 patricentric and 12 matricentric Latin American populations and eight Latin American populations have flexible kinship structure. The kinship structure of two Latin American populations could not be determined and they were dropped from this analysis. As for the Bantu-speaking groups, there are 16 populations with patricentric, 8 with matricentric, and two with flexible kinship structure. The Digo, one of nine Mijikenda tribes, traditionally have matricentric kinship structure, so they were removed from the Mijikenda population data set.

6.2.2 Non-genetic variables

The information on the non-genetic variables (kinship structure, current ethnic population size, and subsistence strategies) of population analyzed was collected from various sources (TABLE XXVIII and TABLE XXIX). Current ethnic population size estimates were obtained from Ethnologue when possible (web accessed between 2007-2008). Otherwise, the number of language speakers listed in Ethnologue was used. The information about kinship structure and subsistence strategies was obtained from Ethnographic Atlas, eHRAF, various anthropological literatures, or the genetic articles I used to obtain the population genetic data.

Subsistence strategy was used as a variable only in the Latin American analyses because the Bantu speakers are all food producers. The Bantu speakers traditionally practice slash-and-burn horticulture, but some of them adopted cereal agriculture and pastoralism in last 2,000-

TABLE XXVIII

NEW WORLD POPULATION INFORMATION

Populations	Population size	Kinship structure[a]	Subsistence[b]
Mesoamerica			
Quiche Mayans	2000000	3	3
Chibchans			
Arsario	3225	1	2
Huetar[e]	642[c]	1	2
Ijka	14301	1	2
Kogi	10000	1	2
Kuna[e]	57114	1	2
Ngobe[e]	133092	1	2
Western Lowland South Americans			
Cayapa	4250	3	2
Embera[e]	23480	2	2
Wounane	6000	2	2
North-Central Andes			
Ancash	856832	3	3
Tayacaja[e]	900000	3	3
Tupe	2000	3	3
Yungay[e]	300000	3	3
South-Central Andes			
Arequipa	532000	3	3
Aymara La Paz[e]	1790000	3	3
Aymara Puno	441743	3	3
Quechua Puno[e]	500000	3	3
Southern Andes			
Mapuche (Argentina)[e]	100000	3	1
Mapuche/Pehuenche[e]	928000	3	2
Yaghan	100	2	1
Lowland Bolivians, Dept. of Beni			
Movima	6528	NA[d]	2
Moxo	20805	3	2
Yuracare	3333	NA[d]	2
Gran Chaco			
Pilaga[e]	2000	2	1
Toba[e]	20656	2	1
Wichi[e]	25000	2	1

TABLE XXVIII (continued)

NEW WORLD POPULATION INFORMATION

Populations	Population size	Kinship structure[a]	Subsistence[b]
Other Lowland South Americans			
Ache	1500	2	1
Ayoreo	3771	3	1
Guahibo[e]	26425	1	1
Kaingang	18000	2	1
Kaiowa	15512	1	2
M'bya	16000	1	2
Nandeva	11900	1	2
Wayuu[e]	305000	1	2
Xavante	10000	1	1
Yanomamo	26653	3	2
Zoro/Gaviao	472	3	2

[a] Kinship Structures are categorized into 1 matricentric, 2 flexible, and 3 patricentric.
[b] Subsistence strategies are categorized into 1 forager, 2 horticulturalist, and 3 agriculturalist/pastoralist.
[5] Population size are taken from Santos et al. (
[d] Population size and social structure are unknown.
1994).
[e] Populations used for estimation of migration rate using MIGRATE.

3,000 years (Phillipson 2005). Unlike Latin Americans, however, the Bantu speakers do not have intensive agricultural technology. Subsistence strategy was used as a proxy for effective population size. While current population size may not be a good indicator of long-term population size in the past, subsistence practices can provide a general measure of expected long-term population size. Agriculture generally supports larger, sedentary populations because the amount of food produced per acre is higher, the food grown is more easily stored and larger families are more efficient and easily supported (Bentley et al. 1993). The sample populations were classified into one of three categories (agriculturalist/pastoralist, horticulturalist, and forager). The populations that were categorized in agriculturalists had intensive agricultural technologies (irrigation and terrace) before the European contact. Agriculturalists and

TABLE XXIX

BANTU POPULATION INFORMATION

	Population size	Kinship structure[a]
East Africa		
Taita[b]	213,389	3
Mijikenda[b]	991,000	3
Hutu[b]	11,000,000	3
Kikuyu	5,347,000	3
Sukuma	3,200,000	2
Turu[b]	556,000	3
Southern Africa		
Chopi	800,000	2
Chwabo[b]	786,715	1
Lomwe[b]	1,300,000	1
Makhwa[b]	2,500,000	1
Nyanja	497,671	1
Nyungwe[b]	262,455	3
Ronga[b]	727,565	3
Sena[b]	876,570	1
Shangaan[b]	3,275,105	3
Shona[b]	10,759,200	3
Tonga	223,971	3
Central Africa		
Bakaka[b]	30,000	3
Bamileke[b]	1,205,900	3
Bassa[b]	230,000	3
Bateke[b]	454,000	1
Bubi[b]	40,000	1
Ewondo[b]	577,700	3
Mbundu	3,000,000	3
Ngoumba[b]	17,500	3
Sanga	36,000	1

[a] Kinship Structures are categorized into 1 matricentric, 2 flexible, and 3 patricentric.

[b] Populations used for estimation of migration rates using MIGRATE.

pastoralists are grouped together because the only pastoralists included in this study were the highland Andeans who practice both agriculture and pastoralism. Foragers are defined as people who acquire more than 90% of their food from hunting, gathering, and fishing (Marlowe 2004). Horticulturalists are more intermediate between the two previous strategies. While they all cultivate some domesticated plants and animals, they also rely on hunting, gathering, and fishing in various degrees.

Some populations have shifted from matricentric to patricentric system or from foraging to horticultural economy since European contact (Martin and Voorhies 1975). For example, the Toba of Argentina were traditionally foragers, but they grow some of their own food today and would be considered horticulturalists according to this scheme (Metraux 1946). However, where modern shifts are known to have occurred, I classified those groups by their traditional occupation.

6.2.3 Analytical methods

Migration rates were estimated using three different methods (AMOVA Φ_{ST}, *M* under spatial expansion model using mismatch distribution, and MIGRATE) and pairwise population genetic distances (Φ_{ST}) were calculated. The AMOVA Φ_{ST}, the oldest of three methods and the most commonly used method for estimating migration rates, N_em (Seielstad et al. 1998) is used to estimate the average migration rate among populations within a group. The migration rate, $M=2N_fm$ is estimated from mismatch distributions under a spatial expansion model (Excoffier 2004; Ray et al. 2003). The migration rate is estimated separately for each population sample assuming that the population is exchanging genes with an infinite number of populations, so the correlation of *M* with non-genetic variables can be statistically examined. MIGRATE is a newer, but relatively untested method for estimating migration rate, $2N_fm$ (Beerli and Felsenstein 1999). The MIGRATE program provides migration rates between pairs of populations. The Arlequin population genetics software program was used to perform AMOVA, estimate M=$2N_fm$ under a spatial expansion model and calculate pairwise population genetic distances (Excoffier et al. 2005; Schneider et al. 2000). SPSS statistical software was used for MDS analyses.

Populations were organized into geographical/linguistic groups to compare migration rates. These groups were used for the AMOVA Φ_{ST} calculations. Then, the average M of populations and pairwise population genetic distances in the geographical groups and the average of the migration rates of pairs of populations in the geographical groups estimated using MIGRATE were calculated. The same grouping scheme was used for all three migration rate estimation methods and for the pairwise population genetic distance calculations. Human populations violate some of the underlying assumptions of the methods (see Chapter 2), so each method will likely yield slightly different values. Nonetheless, the estimates from one method should be supported by the estimates from other methods and if the patricentric and matricentric populations have different patterns of female gene flow or movement, the pattern should be reproduced in all three methods.

For the correlation analyses, the parameter $\theta=2N_f\mu$ (θ_k, θ_S and θ_π) was estimated using Arlequin and $\Theta=2N_f\mu$ was estimated using MIGRATE (Beerli and Felsenstein 2001), as described in previous chapters. The Pearson's correlation of the θ estimates, migration rate (M), and non-genetic variables (kinship structure, ethnic population size, and subsistence strategies) was performed using SPSS statistical software. When the non-genetic variables were correlated with each other, partial correlations between kinship structure and genetic diversity were calculated controlling for other variables.

6.3 Results

6.3.1 Effects of kinship structure on population subdivision

Most of the patricentric populations have higher migration rates by all three methods in comparison to matricentric populations in both the Latin American and Bantu data sets (TABLE XXX and TABLE XXXI). The exceptions are the migration rate estimates of the patricentric east African Bantu and matricentric southeastern African Bantu group obtained using AMOVA Φ_{ST}. The patricentric east African Bantu group has small migration rate, while the matricentric Southeastern African Bantu group has large migration rate. The migration rates among the patricentric groups, especially populations at the center of the population expansions, (the Central Andean and Central African Bantu-speaking populations) are high. The Latin American populations with flexible kinship structures also have larger estimated migration rates than the matricentric Latin American populations.

The patricentric populations have average pairwise population genetic distances (Φ_{ST}) that are consistently smaller than the matricentric populations. The Latin American population genetic distances are illustrated on an MDS plot (Fig. 17). The patricentric populations are shown in blue, the flexible populations in green and the matricentric populations in red. The same populations used to estimate migration rates are used in this analysis. Two populations, the Cayapa and Yaghan, are not included in the migration rate estimation because of their kinship structure differs from the two other populations in the same geographic group, but they are shown on the MDS plot. The populations with patricentric and flexible kinship structures from western Latin America tend to cluster together, while matricentric populations such as the Chibchans are widely scattered.

TABLE XXX

DIFFERENCES IN ESTIMATED MIGRATION RATES ($2N_fm$) AND GENETIC DISTANCE (Φ_{ST}) AMONG LATIN AMERICAN POPULATIONS WITH DIFFERENT KINSHIP STRUCTURE (number of populations analyzed)

	AMOVA	Mismatch distribution[a]	MIGRATE[a, b]	Genetic distance[a]
Patricentric				
Central Andeans	24.6 (8)	46.9 (5)[c]	43.9 (4)	0.031 (8)
Southern Andes[d]	63.0 (2)	9.0 (2)	4.9 (2)	0.016 (2)
Flexible				
Gran Chaco	39.0 (3)	12.1 (3)	5.2 (3)	0.029 (3)
NW Lowland S. Americans[e]	16.3 (2)	8.4 (2)	11.2 (2)	0.058 (2)
Matricentric				
Chibchan	3.7 (6)	1.2 (6)	0.7 (3)	0.207 (6)
NW Lowland S. Americans (Matricentric Arawakans)[f]	5.4 (2)	2.8 (2)	0.1 (2)	0.157 (2)
Guarani[g]	4.5 (3)	3.8 (3)		0.205 (3)

[a] Average pairwise population genetic distance and migration rates estimated from mismatch distribution under spatial expansion model and from MIGRATE are listed, except when the group contains only two population, pairwise genetic distance and migration rate obtained from MIGRATE are actual estimates.

[b] A subset of populations was used to estimate migration rates using MIGRATE. The populations used are listed on TABLE XXVIII.

[c] Migration rates could not be obtained from three Central Andean populations (Arequipa, Aymara Puno, and Aymara La Paz).

[d] The Yaghan is not included, because they have more flexible kinship structure, while two others have patricentric kinship structure.

[e] The Cayapa is not included, because they have patricentric kinship structure, while two others have flexible kinship structure.

[f] Matricentric Arawakans are Guahibo and Wayuu.

[g] Although the Gurarani (Kaiowa, M'bya, and Nandeva) were not included for MIGRATE analysis, they were included here for comparison, because matricentric populations are rare and patricentric populations are more often used for analyses in other studies (Chaix et al. 2007; Pilkington et al. 2007; Seielstad et al. 1998).

TABLE XXXI

MIGRATION RATE ($2N_{fe}m$) AND GENETIC DISTANCE (Φ_{ST}) DIFFERENCE BETWEEN PATRICENTRIC AND MATRICENTRIC BANTU-SPEAKING POPULATIONS[a] (number of populations analyzed)

	AMOVA	Mismatch Distribution	MIGRATE[b]	Genetic Distance
Patricentric				
East African	32.1 (5)	42.0 (4)[c]	20.3 (4)	0.0285 (5)
Central African[d]	89.9 (5)	86.4 (5)	34.2 (5)	0.0106 (5)
Southeastern African	NA (5)	37.2 (5)	29.1 (4)	0.0030 (5)
Matricentric				
Central African	16.5 (3)	18.7 (3)	3.5 (2)	0.0599 (3)
Southeastern African	32.3 (5)	8.1 (5)	8.2 (3)	0.0360 (5)

[a] The Bantu populations with flexible and unknown kinship structure are not included.

[b] A subset of populations was used to estimate migration rates using MIGRATE. The populations used are listed on TABLE XXVIII.

[c] Migration rate from one East African population (Kikuyu) could not be obtained.

[d] The Mbundu is not included.

The MDS plot of the Bantu-speaking populations shows a similar pattern (Fig. 18). The Bantu-speaking populations with flexible and unknown kinship structure are also included in the MDS analysis. Patricentric Bantu populations tend to cluster together in the middle, while the matricentic populations are spread across the bottom of the plot. The only exception is the patricentric Turu, which is found at the top of the plot, far away from the other Bantu populations.

6.3.2 Correlation Analyses

In previous chapter, I demonstrated that female gene flow influences within-population genetic diversity. The results of analyses described above show that kinship structure influences the pattern of female gene flow, so I investigated the effects of kinship structure on mtDNA within-population genetic diversity by examining the correlation between kinship structure and within-population genetic diversity. The result of the previous analysis also shows that populations with patricentric and flexible kinship structures have larger migration rates than matricentric populations, so the correlations between migration rate, *M*, estimated from the mismatch distributions, and the non-genetic variables were examined.

Overall, kinship structure is most strongly correlated with mtDNA genetic diversity, and current ethnic population size and subsistence strategy do not predict genetic diversity well (TABLE XXXII). Kinship structure is strongly correlated with all three estimators of genetic diversity among the Latin American populations. Ethnic population size and subsistence strategy is significantly correlated with θ_k, and θ_S, that reflect recent demographic history. The θ_π, that measures ancient demographic history, is not significantly correlated with population size

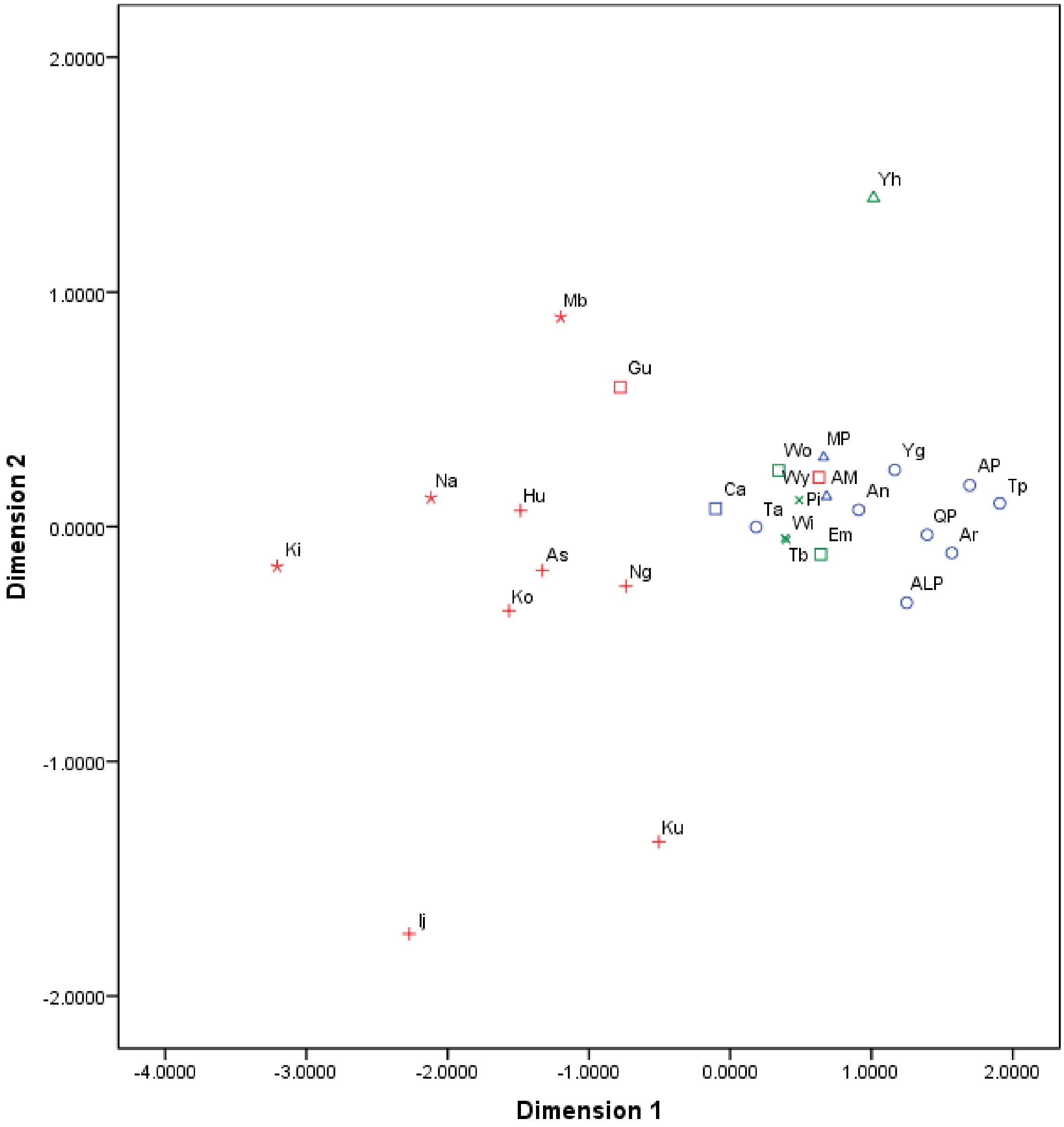

Figure 17. MDS plot of Lain American populations. The kinship structure colored with blue (patricentric), red (matricentric), and green (flexible or unknown). Populations are marked with shape indicating the regions: circle (Central Andeans), triangle (Southern Andeans), x (Gran Chaco), square (western lowland South Americans), star (Guarani), and + (Chibchans).

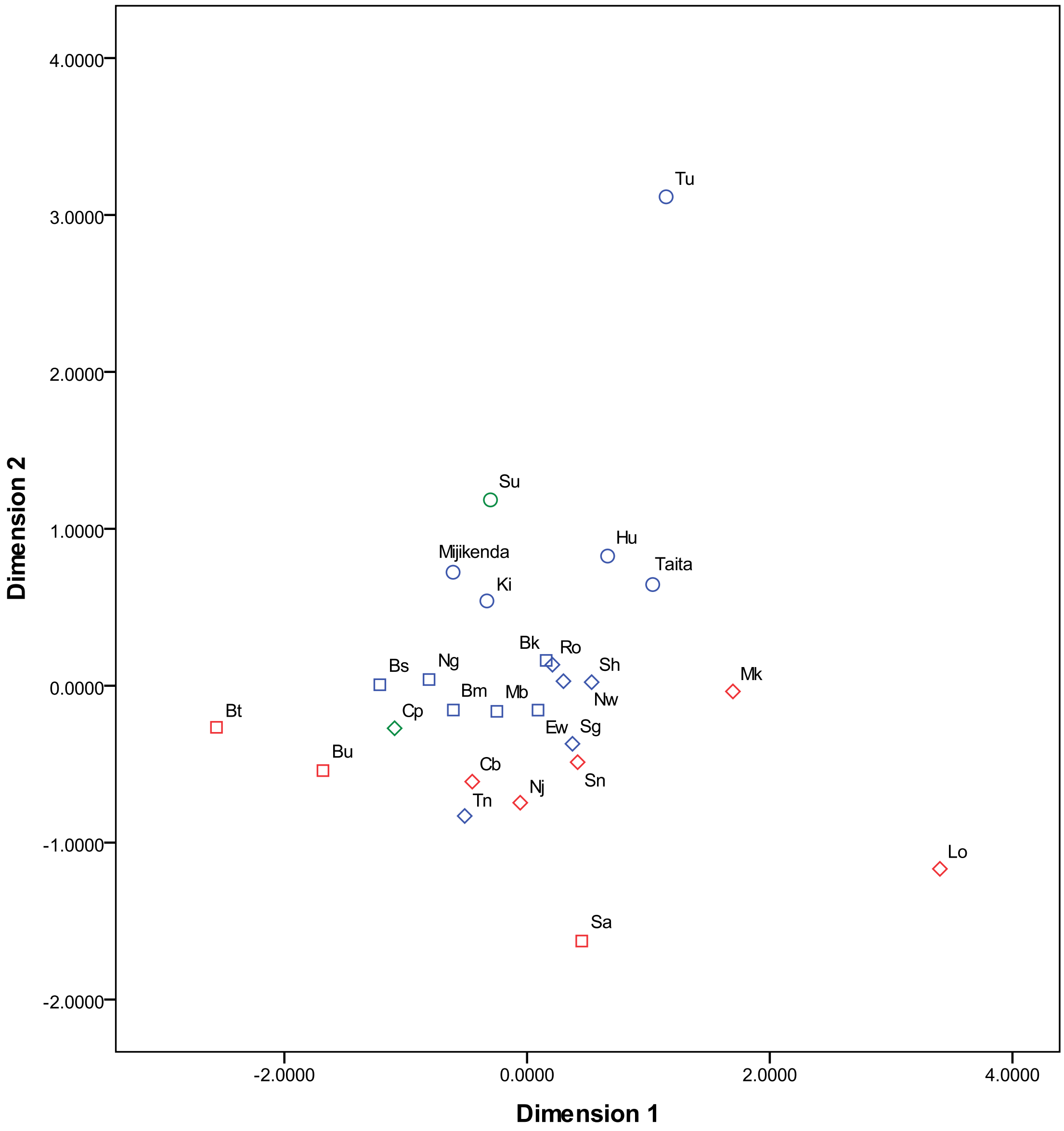

Figure 18. MDS plot of Bantu and East African populations. The kinship structure of the Bantus is colored with blue (patricentric), red (matricentric), and green (flexible or unknown). The symbols used in the plot: East African Bantus (circle), Central African Bantus (squre), Southern African Bantus (diamond).

TABLE XXXII

CORRELATION BETWEEN θ AND DIFFERENT VARIABLES (r and P-value)

	θ_k	θ_S	θ_π	M
Latin America				
Kinship structure	0.585 (0.000)	0.704 (0.000)	0.508 (0.002)	0.380 (0.029)
Population size	0.523 (0.001)	0.499 (0.001)	0.249 (0.132)	0.861 (0.000)
Subsistence pattern	0.588 (0.000)	0.467 (0.003)	0.176 (0.291)	0.490 (0.003)
Africa-Bantus				
Kinship structure	0.660 (0.000)	0.773 (0.000)	0.593 (0.001)	0.659 (0.000)
Population size	0.262 (0.196)	0.298 (0.139)	0.271 (0.180)	0.217 (0.297)

or subsistence strategy. Among the Bantu-speaking populations, all of the θ are strongly correlated with kinship structure, but population size is not significantly correlated with any of the θ estimates. The migration rate, M, is correlated with all of the non-genetic variables among the Latin Americans. The migration rate is strongly correlated with kinship structure, but not with population size among the Bantu populations.

The correlations among the non-genetic variables were examined. All the variables are significantly correlated in the Latin American groups (TABLE XXXIII), but the correlation between kinship structure and population size among the Bantus is not significant (r = 0.273 and P = 0.176). The partial correlations indicate that kinship structure and within-population genetic diversity are correlated well in the Latin American populations after controlling for population size and subsistence, but the correlation between kinship structure and M is no longer significant (TABLE XXXIV). Population size and subsistence pattern are not significantly correlated with the genetic diversity values, but population size is significantly correlated with M, suggesting that female population size (N_f) contribute to M=2$N_f m$ more than the migration rate (m). A partial correlation analysis of population size and subsistence pattern was not conducted because both are used here as proxies for effective population size.

TABLE XXXIII

CORRELATION OF THREE DIFFERENT NON-GENETIC VARIABLES AMONG LATIN AMERICANS (*r* and *P*-value)

	Kinship structure	Population size
Population size	0.450 (0.006)	
Subsistence pattern	0.398 (0.016)	0.584 (0.000)

Note: the correlation between kinship structure and population size among the Bantu is not significant (*r* = 0.273 and *P* = 0.176).

Finally, the correlation between Θ and non-genetic variables were examined using 15 selected Latin American and 19 Bantu populations (TABLE XXXV). None of non-genetic variables showed significant correlation with the Θ values among the Latin Americans. Kinship structure is significantly correlated with Θ among the Bantu populations, but current ethnic population size is not, so the patricentric Bantu populations tend have larger female effective population sizes than the matricentric Bantu populations.

6.4 Discussion: How does kinship structure influence mtDNA variation?

The regional studies of mtDNA and Y chromosome variation found increased female gene flow and higher within-population genetic diversity among patrilocal populations, when compared with matrilocal populations (Besaggio et al. 2007; Chaix et al. 2007; Hamilton et al. 2005; Oota et al. 2001). In this project, I focused on mtDNA variation and increased the number of population samples included for analyses, so the results of this study should be statistically more robust than previous studies. The results support the findings from previous studies and show that kinship structure is a good predictor of female gene flow. Furthermore, the patricentric populations have higher migration rates using all three different methods than the matricentric populations, and their genetic distances are smaller.

TABLE XXXIV

PARTIAL CORRELATION OF SEQUENCE DIVERSITY (θ) AND DIFFERENT VARIABLES CONTROLLING FOR OTHER FACTORS (r AND P-VALUES) AMONG LATIN AMERICANS

	Controlling factors	θ_k	θ_S	θ_π	M
Kinship structure	Population size	0.460 (0.005)	0.618 (0.000)	0.454 (0.006)	0.058 (0.751)
Population size	Kinship	0.359 (0.034)	0.301 (0.079)	0.039 (0.826)	0.843 (0.000)
Kinship structure	Subsistence	0.473 (0.004)	0.639 (0.000)	0.482 (0.004)	0.275 (0.173)
Subsistence pattern	Kinship	0.478 (0.004)	0.291 (0.090)	-0.025 (0.885)	0.423 (0.016)

TABLE XXXV

CORRELATION OF Θ ESTIMATED USING MIGRATE FOR 15 LATIN AMERICAN POPULATIONS AND 19 BANTU POPULATIONS WITH NON-GENETIC VARIABLES (*r* and *P*-value)

	Latin Americans	Bantus
Kinship Structure	0.378 (0.135)	0.529 (0.020)
Population Size	0.412 (0.101)	-0.063 (0.797)
Subsistence Pattern	0.475 (0.475)	

Using the new results from this study, the pattern of population subdivision observed previously could be reevaluated. In chapter 3, I demonstrated that there was considerable gene flow in western South America as suggested previously by some researchers (Fuselli et al. 2003; Lewis et al. 2005; Tarazona-Santos et al. 2001) and that the genetic distances among the western South American populations are small. Prehistoric state expansion, vertical use of Andean ecology, forced migration by the Inca and colonial government, and modernization were all important cultural factors for gene flow (D'Altroy 2002; Masuda et al. 1985; Murra 1968; Skeldon 1977), but in this chapter, I suggest that these groups tend to be genetically similar because of their patricentric kinship structure and the female migration associated with it. The Chibchans and some other lowland populations, on the other hand, exhibit genetic differentiation because of their matricentric kinship structure, not simply because these lowland populations are geographically isolated. Similarly, the Bantu-speaking populations were thought to be a genetically homogeneous group because of either recent common origin or considerable genetic exchange (Cavalli-Sforza et al. 1994; Excoffier et al. 1987; Salas et al. 2004), but in chapter 4, I demonstrated that east African Bantu-speaking populations are genetically heterogeneous because individual Bantu populations have different histories of isolation and interactions with non-Bantu Africans. These patricentric Bantu-speaking populations interacted with non-Bantu populations and became genetically differentiated from each other. In this chapter, I also showed that matricentric central and southeastern African Bantu populations are genetically differentiated because of reduced gene flow.

Moreover, the results suggest that kinship structure has a greater impact on within-population genetic diversity than ethnic population size or subsistence strategy, even though large agricultural populations tend to be genetically diverse and have patricentric kinship structure. The populations with patricentric and flexible kinship structure are genetically more diverse than matricentric populations, and kinship structure is strongly correlated with all three θ estimators. Kinship structure is also correlated with the Bantu Θ values estimated using MIGRATE.

Kinship structure is, however, strongly correlated both with population size and subsistence systems among the Latin Americans. The patricentric populations analyzed tend to have large population sizes and most of the agricultural populations in the New World, for example, the highland Andeans, have patricentric kinship structure. They also have large θ estimates based on sequence variation (Chapter 3) as well as large Θ estimates (Chapter 5). Therefore, kinship structure could be correlated with within-population genetic diversity because of these factors. To examine whether kinship structure was the real factor affecting within-population genetic diversity, the correlation between kinship structure and the θ estimates were examined controlling for population size and subsistence strategy. Kinship structure is well correlated with all three θ estimators, even after controlling for population size and subsistence pattern, but population size and subsistence pattern are only correlated with the θ_k when the effect of kinship structure is removed. Although foragers generally have lower migration rates (Excoffier 2004), foragers such as the Pilaga and Argentine Mapuche have relatively high M

values in this study. Despite their small population sizes, their marriage practices resulted in a higher number of migrants exchanged among demes (N_fm).

This observation is consistent with the findings from earlier chapters. The lack of correlation between Θ and θ_π suggests that gene flow can significantly inflate θ_π (Chapter 5), and θ_π correlates well with kinship structure, but not with population size or subsistence strategies. The significant correlation between Θ and θ_k, on the other hand, suggests that θ_k reflects effective population size and is correlated with ethnic population size and subsistence strategies.

Kinship structure may also be a better predictor of within-population genetic diversity than ethnic population size or subsistence strategy because of the way populations (study units) and ethnic groups are identified and defined. In human population genetic studies, ethnolinguistic groups are often used to define the population or sampling unit. I used ethnolinguistic groupings that many human population geneticists have used to collect samples and define the study unit. Then, the ethnic population sizes or the numbers of language speakers from various ethnic groups were obtained from Ethnologue, which use also ethnolinguistic groupings. However, many social scientists (Braun and Hammonds 2008; MacEachern 2000) have criticized the use of ethnolinguistic grouping in genetic studies and the interpretation of population genetic data requires consideration of the ethnohistory of study populations.

First of all, many indigenous ethnic groups in Latin America and Africa were reorganized after the European contact because of missionary activity, depopulation, and modern state expansion. During this process, substantial population movements in and out of ethnic homelands were not uncommon, and new ethnic groups formed (Braunstein and Miller 1999; Chimhundu 1992; Fried 1968; Whitehead 1994). Chimhundu (1992) argues that ethnic identity in Africa was artificially created after European contact. European colonial governments and missionaries categorized people based on cultural similarities and geographic location, often with little understanding of how African societies were organized. In some cases, multiple groups were merged into a single ethnic group (Chimhundu, 1992), while in other cases, groups shifted their ethnic affiliation in response to political and economic changes (Niehaus 2002). During this politically, socially, and economically unstable times, increased female gene flow might have occurred among different ethnic groups who had patricentric and flexible social structure. For example, Grand Chaco, the region in Argentine where the Ayoreo, Guarani (Kaiowa, Nandeva, and M'bya), Pilaga, Toba, and Wichi occupy, became an ethnic melting pot after the contact (Braunstein and Miller 1999).

Second, ethnic group membership is seldom determined only by the language that people speak (Barth 1969; Fried 1968) and ethnic groups in Africa and the New World often include people who speak multiple languages, especially where Europeans imposed a lingua franca (Errington 2001; Moore 1994; Whitehead 1994). In Africa, Europeans assumed linguistic and cultural homogeneity within ethnic boundaries and viewed linguistic diversity in a community as a sign of barbarism or savagery, so they chose a few languages as lingua franca (Errington, 2001). These languages were spread widely reducing the linguistic diversity in many areas. A similar situation occurred in Central Andes after the European contact; linguistic hegemony was achieved by the Europeans who used Quechua for administrative and religious use (Mannheim 1991).

Increased female gene flow is the most likely explanation for differences in mtDNA variation between patricentric and matricentric populations because patricentric populations have larger migration rates than matricentric populations and the kinship structure is well correlated with within-population genetic diversity (θ). However, differences in female effective population size due to polygyny could explain the higher mtDNA within-population genetic diversity levels, so the correlations between Θ and nongenetic variables were examined. If female gene flow is a more important factor, I predicted that the Θ should be significantly correlated with population size, since effect of gene flow is removed in Θ. If polygyny is the more important factor, then the Θ values should be significantly correlated with kinship structure rather than population size. There should be big discrepancy between female effective population size and actual population size in the patricentric populations. The patricentric populations can have larger female effective population size than male effective population size, because some males have multiple wives, while other males have smaller chance of reproduction. Matricentric populations have equal male and female effective population size.

Although the effects of gene flow may not have been completely removed, the fact that the Θ values are more strongly correlated with kinship structure than with population size among the Bantu-speaking populations suggests that some patricentric populations have larger female effective population sizes than would be expected from their population size. The effects of polygyny were more likely to be observed among the Bantu populations where it was traditionally practiced, but not in the New World populations, where monogamy is more predominant.

6.5 **Conclusion**

These results suggest that kinship structure is one of the most important cultural factors influencing female gene flow and within-population genetic diversity. Patricentric populations have larger migration rates than matricentric populations, so even patricentric populations with small population sizes can be genetically diverse and tend to be genetically similar to nearby populations. However, the influence of population size on mtDNA variation is difficult to assess because of the inter-correlation among many cultural factors and the difficulties associated with using ethno-linguistic classification to define populations. A potential effect of polygyny on male and female effective population size was also detected. Follow-up studies with Y chromosome data are necessary to investigate how kinship structure influences the pattern of male gene flow and Y chromosome variation.

7. CONCLUSION

7.1 Summary: Roles of female gene flow in human evolutionary history

The main objective of this dissertation was to understand the role of female gene flow affecting mtDNA variation, and I hypothesized that female gene flow was the major contributing factor affecting mtDNA variation. To test this hypothesis, I examined the impacts of female gene flow and effective population size on mtDNA variation. This research was motivated by two questions. How can the impact of gene flow through spatial expansion be observed in within-population genetic diversity? Was increased female gene flow a major contributing factor influencing human mtDNA variation? I asked three questions. A) Did the Aymara and Bantu speaking populations expanded through range (or spatial) expansions by incorporating female migrants from other ethnic groups or through demographic expansions with increased female fertility rates? B) Which of the two factors, female gene flow or effective population size, had greater effect on mtDNA within-population genetic diversity? C) to what extent did kinship structure (here defined as patricentric or matricentric) affect the importance of each factor?

7.1.1 Did the Aymara and the Bantu speaking populations expand through spatial or demographic expansion?

In Chapter 3, I examined how the Aymara expanded in the past, hypothesizing that a spatial expansion explained the mtDNA pattern among the Aymara and other Andean populations. Instead I found that both the Aymara and the Quechua most likely experienced rapid demographic expansions after the introduction of intensive agriculture around 3,000 years ago, though natural selection may have affected mtDNA variation. My data also show that female gene flow resulting from vertical archipelago systems, pre-historic state expansions, and European contact in the Central Andes also played an important role. The Aymara and Quechua have similar haplogroup B mismatch distributions, either because of similarities in their demographic expansions or because of long-term interactions between them. Population movements in Central Andes and into transitional zones between the Andean highlands and lowland South America increased within-population genetic diversity as well. These transitional populations generally exhibit intermediate within-population genetic diversity between highland Andeans and lowland Amazonians. The transitional populations were also plotted closely with Central Andeans on the MDS plot and migration rates between them were high. However, it is possible that the highland Andean mtDNA show evidence of past population expansion, partly because of a hitchhiking effect. The role of mitochondrial adaptation to high altitude environments needs to be investigated in the future.

Although gene flow was an important factor affecting Latin American mtDNA variation, it had a much greater effect on mtDNA variation among Bantu-speaking populations (Chapter 4). The traditional Bantu expansion model suggested that the Bantu experienced demographic expansion and replaced pre-existing forager populations. Contrary to this model of Bantu expansion, I hypothesized that the Bantu-speakers experienced a spatial expansion through gene flow with non-Bantu speaking east African populations. The results of the analyses support this hypothesis. The Bantu language and associated culture spread over a large area of sub-Saharan Africa through migration(s) from central Africa. In east Africa, Bantu-speaking immigrants encountered numerous large populations already living in the area. The Taita interacted with and incorporated non-Bantu speaking east Africans through marriage exchange and other sociocultural activities (Bravman 1998). As a result, they have many east African HPGs, and they are genetically as diverse as other east African populations. They are also genetically similar to east Africans, such as the Turkana (Nilo-Saharan speakers from Kenya). The Mijikenda, on the other hand, interacted more intensively with other Bantu-speaking populations, such as the Swahili (Willis 1993) and maintained largely Bantu mtDNA variation. They have more central and west African HPGs than east African HPGs, and they cluster near the central and southeastern African Bantu populations on MDS plots.

7.1.2 Did female gene flow or effective population size have a greater impact on mtDNA within-population genetic diversity?

Many previous anthropological genetic studies focused on population relationships and tracing evidence of gene flow and migration, but there was less attention on how female gene flow affected mtDNA within-population genetic diversity. I hypothesized that female gene flow was a more important factor than female effective population size influencing mtDNA within-population genetic diversity. In Chapter 5, I examined the roles of female gene flow and effective population size influencing mtDNA within-population genetic diversity by testing whether a demographic model that accounts for gene flow explains the observed mtDNA within-population genetic diversity better than a model that does not take the effects of gene flow into account. The results of analyses showed that neither model could explain the observed mtDNA variation very well. The demographic model that does not account for gene flow is poor fit for understanding human demographic history because female gene flow greatly influences within-population genetic diversity (θ_π, and possibly θ_S). Many populations, especially forager populations, rejected the demographic expansion model in mismatch distribution analysis, but not the spatial expansion model. However, female effective population size was also important factors affecting mtDNA variation. The Latin American populations practicing intensive agriculture, such as the Aymara and Quechua, have large Θ and θ_k, and did not reject the demographic expansion model in mismatch distribution analysis.

7.1.3 How did kinship structure affect mtDNA variation?

Previous anthropological genetic studies using small population samples showed that kinship structure affects mtDNA and Y chromosome variation. In Chapter 6, I increased the sample size to test whether the kinship structure affects female gene flow, and consequently mtDNA variation, in two areas of the world. I demonstrated that kinship structure is one of the important factors affecting mtDNA variation. As suggested by many researchers (Besaggio et al. 2007; Chaix et al. 2007; Hamilton et al. 2005; Oota et al. 2001), kinship structure influences the patterns and intensity of female gene flow and within-population genetic diversity. Patricentric populations exhibit larger female migration rates than matricentric populations do, so even small patricentric populations can be genetically diverse and genetically similar to other populations. On the other hand, matricentric populations have small female migration rates and large population pairwise genetic distances. The interpretations of population subdivision among Latin American and Bantu populations discussed in Chapter 3 and 4 were re-evaluated, focusing on the effects of kinship structure on population subdivision. In addition, significant correlation between Θ and kinship structure among Bantu-speaking populations, not between Θ and current ethnic population size, may reflect an unequal male and female population size, a potential consequence of polygyny on male and female effective population size.

My findings generally support the hypothesis that I proposed. Throughout the dissertation, the data shows that female gene flow was an important factor affecting Latin American and Bantu mtDNA variation. Female gene flow affects within-population genetic diversity and population subdivision, and kinship structure is the important factor affecting pattern and intensity of female gene flow. Female effective population size was also an important influence on mtDNA variation, especially for highland Andeans with intensive agricultural technology, which could support large population size.

7.2 How does gene flow affect other genetic markers among the Bantu and Latin American populations?

While acknowledging that female effective population size was an important factor affecting mtDNA, I focused more of my attention to the roles of female gene flow because the importance of gene flow in human evolution has long been recognized, yet not appreciated (Barbujani and Belle 2005; Livingstone 1962; Reich et al. 2011; Serre and Pääbo 2004). I demonstrated that gene flow has been common among the Bantu and Latin American populations, and it was an important factor that influenced their mtDNA variation, but how does gene flow and lack of gene flow influence other genetic markers among them? In Africa, other genetic markers generally support the mtDNA data. Based on comparison of mtDNA and Y-chromosome variation, Wood et al. (2005) argue that the Bantu languages were spread mainly by males and local non-Bantu females were frequently incorporated into Bantu populations through inter-ethnic marriage. The Bantu speaking populations in east and southeast Africa, however, do have some Y chromosome haplogroups of east African and Khoisan origin which suggests that non-Bantu males were also incorporated into Bantu populations (Luis et al. 2004; Pereira et al.

2002; Tishkoff et al. 2007). Classical markers show that the Bantu-speaking populations are genetically homogeneous with exceptions of the east African Bantu populations that tend to be genetically similar to non-Bantu East African populations as a result of interaction (Cavalli-Sforza et al. 1994; Excoffier et al. 1987; Salas et al. 2002). Autosomal STR data generally supports the genetic distinction between central and east African Bantu populations as well (Tishkoff et al. 2009). East African Bantu societies have heterogeneous ancestries resulting from gene flow with neighboring non-Bantu populations.

Other genetic markers support the mtDNA data in Latin America as well. Although some studies with classical markers did not find a clear recognizable geographic pattern, previous work generally indicated different population histories in western and eastern South America (Luiselli et al. 2000; O'Rourke and Suarez 1986). The Andean and neighboring populations have had large male effective population sizes and/or gene flow. Andean populations exhibit higher Y chromosome genetic diversity than lowland populations and they cluster together on the MDS plot (Tarazona-Santos et al. 2001). The Andean and Gran Chaco populations also show lower Y chromosome among populations variations than lowland populations (Demarchi and Michell 2004). In western South America, genetic drift had a great effect on Y chromosome variation, because of low male effective population size and/or male gene flow.

While many genetic studies show that gene flow heavily influenced genetic variation, no studies have examined the effects of male gene flow on Y chromosome variation. To further understand how kinship structure affects male gene flow, Y chromosome markers require further investigations. Understanding how male gene flow affected Y chromosome within-population genetic diversity will advance our understanding of sex-biased demographic history.

7.3 Assessment of sex-biased demographic history

In this project, indirect assessment of human sex-biased demographic history arguments using mtDNA data from the Latin American and Bantu population could not reject either argument (increased female gene flow vs. smaller male effective population size due to polygyny). Although the results in each chapter show that female gene flow had a great effect on mtDNA variation, the data suggest that both female gene flow and effective population size were important factors. Among Latin Americans who commonly practice monogamy, female gene flow had a great effect making western South American population genetically homogeneous and genetically diverse. The Andean populations with intensive agricultural technology experienced demographic expansion and increase in female effective population had considerable effect on mtDNA within-population genetic diversity. Among the east African Bantu-speaking populations, female gene flow with non-Bantu speakers increased their within-population genetic diversity estimates and made them genetically different from other Bantu-speaking populations. Larger than predicted female effective population sizes in patricentric populations suggests that polygyny could have influenced mtDNA within-population genetic diversity as well.

7.4 Limitations of this study and future prospects

This project analyzed only mtDNA variation. Focusing on mtDNA, I avoided the problems comparing the mtDNA and Y chromosome data, but I could examine only the female side of human demographic history. When the research design of this project was first developed, there were considerably fewer Y chromosome studies than mtDNA studies available for comparison. Focusing on mtDNA, I increased the number of sampled populations, so the results of this study would be statistically more robust than previous studies that included both mtDNA and Y chromosome data (Besaggio et al. 2007; Oota et al. 2001). European admixture among the Latin Americans also complicates the analysis of Y chromosome variation more than mtDNA variation. In the future, detailed analyses of Y chromosome variation are necessary to further evaluate impact of polygyny on male effective population size.

Second, human population geneticists and anthropological geneticists generally use ethnolinguistic grouping to collect individual samples and group them in order to conduct population genetic analysis (Cavalli-Sforza et al. 1994), and the use of this method of defining study units and sample collections continues to be used in large international collaborative projects (The 1000 Genomes Project Consortium 2010; The International HapMap Consortium 2003). This dissertation study followed this convention and used ethnolinguistic grouping to define the study units to include as many population samples as I could to make population genetic analyses possible and statistically more robust. However, ethnic group membership is not determined only by the languages that people speak (Barth 1969) and ethnic groups often include people who speak different languages (Moore 1994; Whitehead 1994). Human social systems tend to be fairly open and ethnic boundaries are often permeable (Barth 1969; Green and Perlman 1985). Therefore, I treated study populations as demes, where members of demes are replaced by members of other demes through gene flow. Unfortunately, there are many other issues, such as ethnic group reorganization during the state expansion (Braunstein and Miller 1999; Chimhundu 1992; Fried 1968; Whitehead 1994), which could not be resolved and incorporated in the demographic models for population genetic analyses. These recent historical events could have influenced mtDNA variation of the Taita, Mijikenda, and other populations in the area (Babrowski 2007).

Third, population sampling and the number of individuals sampled per population are additional issues in human population genetic studies. Considering the linguistic and cultural diversity that exists in Africa and Latin America, the number of populations sampled is small, and the population samples tend to geographically cluster. The number of individuals sampled per population should be increased because human populations, or ethnic groups, are internally heterogeneous and Africa is extremely diverse genetically. In this study, I included more individuals from the Aymara, Taita, and Mijikenda than many other populations analyzed previously. Unfortunately, the clustering of sampled populations may also affect results of some analyses, such as the Mantel test for geographical and genetic distance correlation, and the sampling of populations with equal geographical distribution will require a more carefully designed international collaborative effort.

Moreover, the population genetic models used in this dissertation projects may not be applicable for the evolutionary study of human demographic history. Many genetic models used for population genetic analyses assume that: 1) populations are panmictic, 2) population size is stable, and 3) populations have independent evolutionary histories (see detailed discussion in Chapter 2). Human populations usually violate these important built-in assumptions and the models are too simplistic. However, I believe that even methods that resort to these simplistic unrealistic models can be useful if the research plan is designed carefully. In this project, I examined the role of female gene flow affecting mtDNA variation in two areas of the world using multiple analytical methods. While each of these methods has limitations, consistent patterns observed in each of the analyses indicate the important role that female gene flow played affecting within-population genetic diversity, migration rate, and pattern of population subdivision.

There is a problem with computation for statistical inference, as well. Model-based methods of statistical inference, such as MIGRATE, are usually computationally intensive and demographic parameters may not be estimated accurately using a single marker. MIGRATE was designed to estimate only two basic demographic parameters (effective population size and migration rates) to make the use of the program easier for users, but analysis still takes long time. The methods with more complex models that allow users to estimate more demographic parameters (e.g., initial effective population size, growth rate, effective population size after expansion, and migration rates) and include multiple population samples per single run require more computational time. Considering that I used a short fragment of a single genetic marker, reliable estimations of many different demographic parameters would be very difficult, even after increasing number of sequences included for analysis.

To infer the female component of human demographic history more reliably using model-based statistical inference methods, we need to analyze much longer sequences (maybe the whole control region or entire mitochondrial genome) and we need to have more computational power. With the continued development of molecular genetics and increasing computer technology, the more accurate estimation of various demographic parameters and a greater understanding of female demographic history will be possible in the near future.

7.5 Conclusion

This dissertation project show that both female gene flow and effective population size were important factors affecting mtDNA variation (within-population genetic diversity and population subdivision). Cultural factors, such as subsistence pattern and kinship structure, are import forces that affect mtDNA variation in both of the areas that I examined. While the methods used in this study have many limitations which need to be further addressed, my findings make important contributions that can be used to direct future research.

CITED LITERATURE

Alva W. 2001. The royal tombs of Sipán: Art and power in Moche society. In: Pillsbury J, editor. Moche Art and Archaeology in Ancient Peru. New Haven: National Gallery of Art, Washington, distributed by Yale University Press. p 223-245.

Anderson A, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F and others. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457-470.

Aris-Brosou S, and Excoffier L. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Molecular Biology and Evolution 13:494-504.

Austerlitz F, Jung-Muller B, Godelle B, and Gouyon P-H. 1997. Evolution of Coalescence Times, Genetic Diversity and Structure during Colonization. Theoretical Population Biology 51(2):148-164.

Babrowski K. 2007. Ethnogenesis and genetic identity among the Taita and Mijikenda of Kenya: a mitochondrial DNA case study: University of Illinois at Chicago.

Bandelt H-J, Forster P, and Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution 16(1):37-48.

Barbieri C, Heggarty P, Castrì L, Luiselli D, and Pettener D. 2011. Mitochondrial DNA variability in the Titicaca basin: Matches and mismatches with linguistics and ethnohistory. American Journal of Human Biology 23(1):89-99.

Barbujani G, and Belle EMS. 2005. Genomic boundaries between human populations. Human Heredity 61:15-21.

Barbujani G, and Bertorelle G. 2001. Genetics and the population history of Europe. Proceedings of the National Academy of Sciences 98(1):22-25.

Barceló A, Daroca MdC, Ribera R, Duarte E, Zapata A, and Vohra M. 2001. Diabetes in Bolivia. Revista Panamericana de Salud Pública/Pan American Journal of Public Health 10:318-323.

Barth F. 1969. Ethnic group and boundaries London: Allen & Unwin.

Batai K, and Williams SR. 2007. Reconstructing the settlement history of the central Andes from mitochondrial DNA analysis. American Journal of Physical Anthropology 132(S44):69.

Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, and Calafell F. 2007. Phylogeography of the human mitochondrial L1c haplogroup: Genetic signatures of the prehistory of Central Africa. Molecular Phylogenetics and Evolution 43(2):635-644.

Batista O, Kolman CJ, and Bermingham E. 1995. Mitochondrial DNA diversity in the Kuna Amerinds of Panamá. Human Molecular Genetics 4(5):921-929.

Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, King RJ, Rootsi S, Marjanovic D, Primorac D and others. 2009. Y-chromosomal evidence of the cultural diffusion of agriculture in southeast Europe. European Journal of Human Genetics 17(6):820-830.

Beall CM. 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. Proceedings of the National Academy of Sciences 104(Suppl 1):8655-8660.

Beall CM, Almasy LA, Blangero J, Williams-Blangero S, Brittenham GM, Strohl KP, Decker MJ, Vargas E, Villena M, Soria R and others. 1999. Percent of oxygen saturation of arterial hemoglobin among Bolivian Aymara at 3,900–4,000 m. American Journal of Physical Anthropology 108(1):41-51.

Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Molecular Ecology 13(4):827-836.

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

Beerli P, and Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences of the United States of America 98(8):4563-4568.

Bendall KE, and Sykes BC. 1995. Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. American Journal of Human Genetics 57(2):248-256.

Bentley GR, Goldberg T, and Jasienska G. 1993. The fertility of agricultural and non-agricultural traditional societies. Population Studies 63:269-281.

Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, van der Veen L, Hombert J-M, Quintana-Murci L, and Comas D. 2009. Genetic and Demographic Implications of the Bantu Expansion: Insights from Human Paternal Lineages. Molecular Biology and Evolution 26(7):1581-1589.

Bert F, Corella A, Gene M, Perez-Perez A, and Turbon D. 2004. Mitochondrial DNA diversity in the Llanos de Moxos: Moxo, Movima and Yuracare Amerindian populations from Bolivia lowland. Annals of Human Biology 31:9-28.

Bert F, Corella A, Manel G, Pérez-Pérez A, and Turbón D. 2001. Major mitochondrial DNA haplotype heterogeneity in highland and lowland Amerindian populations from Bolivia. Human Biology 73(1):1-16.

Besaggio D, Fuselli S, Srikummool M, Kampuansai J, Castri L, Tyler-Smith C, Seielstad M, Kangwanpong D, and Bertorelle G. 2007. Genetic variation in Northern Thailand Hill Tribes: origins and relationships with social structure and linguistic differences. BMC Evolutionary Biology 7(Suppl 2):S12.

Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, Scherer SW, Julian CG, Wilson MJ, López Herráez D and others. 2010. Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. PLoS Genet 6(9):e1001116.

Blom DE, Hallgrimsson B, Keng L, Lozada Cerna MC, and Buikstra JE. 1998. Tiwanaku 'colonization': bioarchaeological implications for migration in the Moquegua Valley, Peru. World Archaeology 30(2):238-261.

Boles T, Snow C, and Stover E. 1995. Forensic DNA testing on skeletal remains from mass graves: a pilot study in Guatemala. Journal of Forensic Science 40:349-355.

Bove R, and Valeggia C. 2009. Polygyny and women's health in sub-Saharan Africa. Social Science & Medicine 68(1):21-29.

Braun L, and Hammonds E. 2008. Race, populations, and genomics: Africa as laboratory. Social Science & Medicine 67:1580-1588.

Braunstein J, and Miller E. 1999. Ethnohistorical Introduction. In: ES M, editor. Peoples of the Gran Chaco. Westport, Connecticut: Bergin & Garvey. p 2-22.

Bravman B. 1998. Making ethnic ways: Communities and their transformations in Taita, Kenya, 1800-1950. Isaacman A, and Allman J, editors. Portsmouth, NH: Heinemann.

Briggs LT. 1985a. A critical survey of the literature on the Aymara language. In: Manelis Klein HE, and Stark LR, editors. South American Indian languages: retrospect and prospect. Austin: University of Texas Press. p 546-594.

Briggs LT. 1985b. Dialectical variation in Aymara In: Manelis Klein HE, and Stark LR, editors. South American Indian languages: retrospect and prospect. Austin: University of Texas Press. p 595-616.

Browman DL. 1994. Titicaca Basin archaeolinguistics: Uru, Pukina and Ayrara AD 750-1450. World Archaeology 26:235-251.

Brown MD, Hosseini SH, Torroni A, and Bandelt H-J. 1998. mtDNA haplogroup X: An ancient link between Europe/Western Asia and North America? American Journal of Human Genetics 63:1852-1861.

Burger RL. 1995. Chavín and the Origins of Andean Civilization. London: Thames and Hudson.

Burton M, Moore C, Whiting J, and Romney A. 1996. Regions based on social structure. Current Anthropology 37:87-123.

Cabana GS, Merriwether DA, Hunley K, and Demarchi DA. 2006. Is the genetic structure of Gran Chaco populations unique? Interregional perspectives on native South American mitochondrial DNA variation. American Journal of Physical Anthropology 131(1):108-119.

Cadenas AM, Zhivotovsky LA, Cavalli-Sforza LL, Underhill PA, and Herrera RJ. 2007. Y-chromosome diversity characterizes the Gulf of Oman. European Journal of Human Genetics 16(3):374-386.

Cann RL, Stoneking M, and Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325:31-36.

Castrì L, Garagnani P, Useli A, Pettener D, and Luiselli D. 2008. Kenya crossroads: migration and gene flow in six ethnic groups from Eastern Africa. Journal of Anthropological Sciences 86:189-192.

Castrì L, Tofanelli S, Garagnani P, Bini C, Fosella X, Pelotti S, Paoli G, Pettener D, and Luiselli D. 2009. mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: Implications for peopling and migration patterns in sub-Saharan Africa. American Journal of Physical Anthropology 140(2):302-311.

Cavalli-Sforza LL. 1993. Demic Expansions and Human Evolution. Science 259:639-646.

Cavalli-Sforza LL, and Bodmer WF. 1971. The genetics of human population. San Francisco: W. H. Freeman and Company.

Cavalli-Sforza LL, Menozzi R, and Piazza A. 1994. The History and Geography of Human Genes. Princeton, NJ: Princeton University Press.

Cavalli-Sforza LL, Piazza A, Menozzi P, and Mountain JL. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proceedings of the National Academy of Sciences 85:6002-6006.

Černy V, Hájek M, Čmeila R, Bruzek J, and Brdička R. 2004. mtDNA sequences of Chadic-speaking populations from northern Cameroon suggest their affinities with eastern Africa. Annals of Human Biology 31(5):554-569.

Cerný V, Mulligan CJ, Rídl J, Zaloudková M, Edens CM, Hájek M, and Pereira L. 2008. Regional differences in the distribution of the sub-Saharan, West Eurasian, and South Asian mtDNA lineages in Yemen. American Journal of Physical Anthropology 136(2):128-137.

Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, and Heyer E. 2007. From social to genetic structures in Central Asia. Current Biology 17(43-48).

Chikhi L, Nichols R, Barbujani G, and Beaumont MA. 2002. Y genetic data support the Neolithic diffusion model. Proceedings of the National Academy of Sciences 99:11008-11013.

Chimhundu H. 1992. Early missionaries and the ethnolinguistic factor during the 'invention of tribalism' in Zimbabwe. Journal of African History 33:87-109.

Clark JD, Beyene Y, WoldeGabriel G, Hart WK, Renne PR, Gilbert H, Defleur A, Suwa G, Katoh S, Ludwig KR and others. 2003. Stratigraphic, chronological and behavioural contexts of Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature 423(6941):747-752.

Coelho M, Sequeira F, Luiselli D, Beleza S, and Rocha J. 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. BMC Evolutionary Biology 9(1):80.

Coia V, Destro-Bisol G, Verginelli F, Battaggia C, Boschi I, Cruciani F, Spedini G, Comas D, and Calafell F. 2005. Brief communication: mtDNA variation in North Cameroon: Lack of Asian lineages and implications for back migration from Asia to sub-Saharan Africa. American Journal of Physical Anthropology 128(3):678-681.

Corella A, Bert F, Pérez-Pérez A, Gen M, and Turbón D. 2007. Mitochondrial DNA diversity of the Amerindian populations living in the Andean Piedmont of Bolivia: Chimane, Moseten, Aymara and Quechua. Annals of Human Biology 34(1):34-55.

Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, Moral P, Watson E, Guida V, Colomb EB, Zaharova B and others. 2004. Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. American Journal of Human Genetics 74:1014-1022.

Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon J-M, Crivellaro F, Benincasa T, Pascone R and others. 2007. Tracing Past Human Male Movements in Northern/Eastern Africa and Western Eurasia: New Clues from Y-Chromosomal Haplogroups E-M78 and J-M12. Molecular Biology and Evolution 24(6):1300-1311.

Currat M, and Excoffier L. 2005. The effect of the Neolithic expansion on European molecular diversity. Proceedings of the Royal Society B: Biological Sciences 272(1564):679-688.

D'Altroy TN. 2002. The Incas. Cambridge, MA: Blackwell Publishing.

Demarchi DA, and Michell RJ. 2004. Genetic structure and gene flow in Gran Chaco populations of Argentina: evidence from Y-chromosome markers. Human Biology 76:413-429.

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Coglia A, Tofanelli S, Spedini G, and Capell C. 2004. Variation of female and male lineages in Sub-Saharan populations: the importance of sociocultural factors. Molecular Biology and Evolution 21:1673-1682.

Di Rienzo A, and Wilson AC. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. Proceedings of the National Academy of Sciences USA 88:1597-1601.

Dillehay TD. 2000. The settlement of the Americas: a new prehistory. New York: Basic Books.

Dipierri JE, Alfaro E, Martinez-Marignac V, Balliet G, Bravi CM, Cejas S, and Bianchi NO. 1998. Paternal directional mating in two Amerindian subpopulations located at different altitudes in northwestern Argentina. Human Biology 70(6):1001-1010.

Dornelles C, Battilana J, Fagundes N, Freitas L, Bonatto S, and Salzano F. 2004. Mitochondrial DNA and Alu insertions in a genetically peculiar population: the Ayoreo Indians of Bolivia and Paraguay. American Journal of Humam Biology 16:479-488.

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

Eggert MK. 2005. The Bantu problem and African archaeology. In: Stahl AB, editor. African archaeology. Malden, MA: Blackwell Publishing. p 301-326.

Ehret C. 2001. Bantu expansions: re-envisioning a central problem of early African history. The International Journal of African Historical Studies 34(1):5-41.

Elson JL, Turnbull DM, and Howell N. 2004. Comparative genomics and the evolution of human mitochondrial DNA: Assessing the effects of selection. American Journal of Human Genetics 74:229-238.

Erickson CL. 1988. Raised field agriculture in the Lake Titicaca Basin: Putting ancient agriculture back to work. Expedition 30(1):8-16.

Eriksen TH. 1993. Ethnicity and nationalism: anthropological perspective. London: Pluto Press.

Errington J. 2001. Colonial Linguistics. Annual Review of Anthropology 30:19-39.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. Theoretical Population Biology 3:87-112.

Excoffier L. 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Molecular Ecology 13:853-864.

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

Excoffier L, Novembre J, and Schneider S. 2000. Computer note. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. Journal of Heredity 91(6):506-509.

Excoffier L, Pellegrini B, Sanchez-Mazas A, Simon C, and Langaney A. 1987. Genetics and history of Sub-Saharan Africa. Yearbook of Physical Anthropology 30:151-194.

Excoffier L, and Schneider S. 1999. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proceedings of the National Academy of Sciences 96:10597-10602.

Excoffier L, Smouse PE, and Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Excoffier L, and Yang Z. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. Molecular Biology and Evolution 16:1357-1368.

Fried MH. 1968. On the concepts of "tribe" and "tribal society". In: Helm J, editor. Essays on the problem of tribe. Seattle: University of Washington Press. p 3-20.

Frisancho AR, Juliao PC, Barcelona V, Kudyba CE, Amayo G, Davenport G, Knowles A, Sanchez D, Villena M, Vargas E and others. 1999. Developmental components of resting ventilation among high- and low-altitude Andean children and adults. American Journal of Physical Anthropology 109(3):295-301.

Fu Y-X. 1996. Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics 144:829-838.

Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, and Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Molecular Biology and Evolution 20(10):1682-1691.

Gayà-Vidal M, Moral P, Saenz-Ruales N, Gerbault P, Tonasso L, Villena M, Vasquez R, Bravi CM, and Dugoujon J-M. 2011. mtDNA and Y-chromosome diversity in Aymaras and Quechuas from Bolivia: Different stories and special genetic traits of the Andean Altiplano populations. American Journal of Physical Anthropology 145(2):215-230.

Gibson MA, and Mace R. 2007. Polygyny, reproductive success and child health in rural Ethiopia: why marry a married man? Journal of Biosocial Science 39(02):287-300.

Ginther C, Corach D, Penacino A, Rey JA, Carnese FR, Hutz MH, Anderson A, Just J, Salzano FM, and King M-C. 1993. Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: Mitochondrial DNA sequence variation and allele frequencies of several nuclear genes. In: Penn SDJ, Chakraborty R, Epplen JT, and Jeffreys A, editors. DNA Fingerprinting: State of the Science. Basel, Switzerland: Birkhäuser Verlag. p 211-219.

Goicoechea AS, Carnese FR, Dejean C, Avena SA, Weimer TA, Franco MHLP, Callegari-Jacques SM, Estalote AC, Simões MLMS, Palatnik M and others. 2001. Genetic relationships between Amerindian populations of Argentina. American Journal of Physical Anthropology 115:133-143.

Goldstein PS. 1989. The Tiwanaku occupation of Moquegua. In: Rice DS, Stanish C, and Scarr PR, editors. Ecol, Peruogy, Settlement, History in the Osmore Drainage. Oxford: British Archaeological Reports.

Goldstein PS. 1993. Tiwanaku temples and state expansion: a Tiwanaku sunken court temple in Moquegua, Peru. Latin America Antiquity 4:22-47.

González AM, Cabrera VM, Larruga JM, Tounkara A, Noumsi G, Thomas BN, and Moulds JM. 2006. Mitochondrial DNA Variation in Mauritania and Mali and their Genetic Relationship to Other Western Africa Populations. Annals of Human Genetics 70(5):631-657.

Graven L, Passarino G, Semino O, Boursot P, Santachiara-Benerecetti S, Langaney A, and Excoffier L. 1995. Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. Molecular Biology and Evolution 12(2):334-345.

Green SW, and Perlman SM. 1985. Frontiers, boundaries, and open social system. In: Green S, and Perlman S, editors. The archaeology of frontiers and boundaries. New York: Academic Press. p 3-13.

Guthrie M. 1962. Some Developments in the Prehistory of the Bantu Languages. Journal of African History 3(2):273-282.

Hadley C. 2005. Is polygyny a risk factor for poor growth performance among Tanzanian agropastoralists? American Journal of Physical Anthropology 126(4):471-480.

Hamilton G, Stoneking M, and Excoffier L. 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proceedings of the National Academy of Sciences USA 102:7476-7480.

Hammer MF, Mendez FL, Cox MP, Woerner AE, and Wall JD. 2008. Sex-Biased Evolutionary Forces Shape Genomic Patterns of Human Diversity. PLoS Genet 4(9):e1000202.

Hardman MJ. 1985. Aymara and Quechua: Languages in Contact. In: Manelis Klein HE, and Stark LR, editors. South American Indian languages: retrospect and prospect. Austin: University of Texas Press. p 617-643.

Hasegawa M, Di Rienzo A, Kocher T, and Wilson AC. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. Journal of Molecular Evolution 37:347-354.

Hassan HY, Underhill PA, Cavalli-Sforza LL, and Ibrahim ME. 2008. Y-chromosome variation among Sudanese: Restricted gene flow, concordance with language, geography, and history. American Journal of Physical Anthropology 137(3):316-323.

Helgason A, Nicholson GJ, Stefánsson K, and Donnelly P. 2003. A reassessment of genetic diversity in Icelanders: Strong evidence from multiple loci for relative homogeneity caused by genetic drift. Annals of Human Genetics 67:281-297.

Helgason A, Siguroardottir S, Gulcher JR, Ward R, and Stefansson K. 2000. mtDNA and the origin of the Icelanders: Deciphering signals of recent population history. American Journal of Human Genetics 66:999-1016.

Holden CJ. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-persimony analysis. Proceeding of the Royal Society of London Series B 269:793-799.

Holden CJ, and Gray RD. 2006. Rapid radiation, borrowing and dialect continua in the Bantu languages. In: Forster P, and Renfrew C, editors. Phylogenetic methods and the prehistory of languages. Cambridge: McDonald Institute for Archaeological Research.

Ingman M, and Gyllensten U. 2007. Rate variation between mitochondrial domains and adaptive evolution in humans. Human Molecular Genetics 16(19):2281-2287.

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

Janusek JW. 2004. Tiwanaku and its precursors: recent research and emerging perspectives. Journal of Archaeological Research 12(121-183).

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, and Batzer MA. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. American Journal of Human Genetics 66:979-988.

Josephson SC. 2002. Does polygyny reduce fertility? American Journal of Human Biology 14(2):222-232.

Keinan A, Mullikin JC, Patterson N, and Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41(1):66-70.

Kim JH, Park KS, Cho YM, Kang BS, Kim SK, Jeon HJ, Kim SY, and Lee HK. 2002. The prevalence of the mitochondrial DNA 16189 variant in non-diabetic Korean adults and its association with higher fasting glucose and body mass index. Diabetic Medicine 19(8):681-684.

Kittles RA, and Weiss KM. 2003. RACE, ANCESTRY, AND GENES: Implications for Defining Disease Risk. Annual Review of Genomics and Human Genetics 4(1):33-67.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, and Villems R. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the Gate of Tears. American Journal of Human Genetics 75:752-770.

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A and others. 2006. The Role of Selection in the Evolution of Human Mitochondrial Genomes. Genetics 172(1):373-387.

Knight A, Underhill PA, Mortensen HM, Zhivogtovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, and Mountain JL. 2003. African Y chromosome and mtDNA divergence provides insight into the history of Click languages. Current Biology 13:464-473.

Knudson KJ. 2008. Tiwanaku influence in the South Central Andes: Strontium isotope analysis and Middle Horizon migration. Latin American Antiquity 19(1):3-23.

Kolata AL. 1993. The Tiwanaku: Portrait of an Andean Civilization. Cambridge, MA: Blackwell.

Kolman CJ, and Bermingham E. 1997. Mitochondrial and nuclear DNA diversity in the Choco and Chibcha Amerinds of Panama. Genetics 147:1289-1302.

Kolman CJ, Bermingham E, Cooke R, ward RH, Arias TD, and Guinneau-Sinclair F. 1995. Reduced mtDNA diversity in the Ngöbé Amerinds of Panamá. Genetics 140:275-283.

Krings M, Halim Salem A-e, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, Welsby D, Di Rienzo A, Utermann G and others. 1999. mtDNA analysis of Nile River valley populations: a genetic corridor of a barrier to migration. American Journal of Human Genetics 64:1166-1176.

Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S, Thangaraj K, Singh L, and Reddy BM. 2006. Global Patterns in Human Mitochondrial DNA and Y-Chromosome Variation Caused by Spatial Instability of the Local Cultural Processes. PLoS Genet 2(4):e53.

Laval G, and Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20(15):2485-2487.

Lewis CM, Buikstra JE, and Stone AC. 2007a. Ancient DNA and genetic continuity in the south central Andes. Latin American Antiquity 18(2):145-160.

Lewis CM, Jr., and Long JC. 2008. Native South American genetic structure and prehistory inferred from hierarchical modeling of mtDNA. Molecular Biology and Evolution 25(3):478-486.

Lewis CMJ, Lizárraga B, Tito RY, López PW, Iannacone GC, Medina A, Martínez R, Polo SI, De La Cruz AF, Cáceres AM and others. 2007b. Mitochondrial DNA and the peopling of South America. Human Biology 79:159-178.

Lewis CMJ, Tito RY, Lizarraga B, and Stone AC. 2005. Land, language, and loci: mtDNA in Native Americans and the genetic history of Peru. American Journal of Physical Anthropology 127:351-360.

Lindgärde F, Ercilla MB, Correa LR, and Ahrén B. 2004. Body Adiposity, Insulin, and Leptin in Subgroups of Peruvian Amerindians. High Altitude Medicine & Biology 5(1):27-31.

Liou C-W, Lin T-K, Huei Weng H, Lee C-F, Chen T-L, Wei Y-H, Chen S-D, Chuang Y-C, Weng S-W, and Wang P-W. 2007. A Common Mitochondrial DNA Variant and Increased Body Mass Index as Associated Factors for Development of Type 2 Diabetes: Additive Effects of Genetic and Environmental Factors. Journal of Clinical Endocrinology and Metabolism 92(1):235-239.

Livingstone FB. 1962. On the non-existence of human races. Current Anthropology 3:279-281.

Long JC, and Kittles RA. 2003. Human genetic diversity and the nonexistence of biological races. Human Biology 75(4):449(423).

Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, and Herrera RJ. 2004. The Leveant versus the Horn of Africa: Evidence for bidirectional corridors of human migrations. American Journal of Human Genetics 74:532-544.

Luiselli D, Simoni L, Tarazona-Santos E, Pastor S, and Pettener D. 2000. Genetic structure of Quechua-speakers of the central Andes and geographic patterns of gene frequencies in South American populations. American Journal of Physical Anthropology 113(1):5-17.

MacEachern S. 2000. Genes, tribes, and African history. Current Anthropology 41(3):357-383.

Malhi RS, and Smith DG. 2002. Brief communication: Haplogroup X confirmed in prehistoric North America. American Journal of Physical Anthropology 119(1):84-86.

Mannheim B. 1985. Southern Peruvian Quechua In: Manelis Klein HE, and Stark LR, editors. South American Indian languages: retrospect and prospect. Austin: University of Texas Press. p 644-688.

Mannheim B. 1991. The language of the Inka since the European invasion. Austin: University of Texas Press.

Marchani E, Watkins WS, Bulayeva K, Harpending H, and Jorde L. 2008. Culture creates genetic structure in the Caucasus: Autosomal, mitochondrial, and Y-chromosomal variation in Daghestan. BMC Genetics 9(1):47.

Marlowe FW. 2004. Marital residence among foragers. Current Anthropology 45:277-284.

Marrero A, Silva-Junior W, Bravi C, Hutz M, Petzl-Erler M, Ruiz-Linares A, Salzano F, and Bortolini M. 2007. Demographic and evolutionary trajectories of the Guarani and Kaingang natives of Brazil. American Journal of Physical Anthropology 132:301-310.

Martin MK, and Voorhies B. 1975. Female of the species. New York: Columbia University Press.

Masuda S, Shimada I, and Morris C, editors. 1985. Andean Ecology and Civilization: An Interdisciplinary Perspective on Andean Ecological Complementarity. Tokyo: University of Tokyo Press.

Mateu E, Comas D, Calafell F, A.Pérez-Lezaun, Abade A, and Bertranpetit J. 1997. A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea. Annals of Human Genetics 61(6):507-518.

McDougall I, Brown FH, and Fleagle JG. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433(7027):733-736.

Melton PE, Briceño I, Gómez A, Devor EJ, Bernal JE, and Crawford MH. 2007. Biological relationship between Central and South American Chibchan speaking populations: evidence from mtDNA. American Journal of Physical Anthropology 133:753-770.

Merriwether DA, Kemp BM, Crews DE, and Neel JV. 2000. Gene flow and genetic variation in the Yanomama as revealed by mitochondrial DNA. In: Renfrew C, editor. America Past, America Present, Genes and Languages in the Americas and Beyond. Cambridge: The McDonald Institute for Archaeological Research. p 84-124.

Merriwether DA, Rothhammer F, and Ferrell RE. 1995. Distribution of the four founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World. American Journal of Physical Anthropology 98:411-430.

Metraux A. 1946. Ethnography of the Chaco. In: Steward JH, editor. Handbook of South American Indian. Washington DC: Smithsonian Institution, Bureau of American Ethnology. p 197-370.

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

Mielke JH, and Fix AG. 2007. The confluence of anthropological genetics and anthropological demography. In: Crawford MH, editor. Anthropological genetics: theory, methods, and applications. New York: Cambridge University Press. p 112-140.

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini SH, Brandon M, Easley K, Chen E, Brown MD and others. 2003. Natural selection shaped regional mtDNA variation in humans. Proceedings of the National Academy of Sciences USA 100:171-176.

Mohanna S, Baracco R, and Seclén S. 2006. Lipid Profile, Waist Circumference, and Body Mass Index in a High Altitude Population. High Altitude Medicine & Biology 7(3):245-255.

Moore J. 1994. Putting anthropology back together again: the ethnogenetic critique of cladistic theory. American Anthropologist 96:925-948.

Moraga ML, Rocco P, Miquel JF, Nervi F, Llop E, Chakraborty R, Rothhammer F, and Carvallo P. 2000. Mitochondrial DNA polymorphisms in Chilean aboriginal populations: implications for the peopling of the southern cone of the continent. American Journal of Physical Anthropology 113(1):19-29.

Moseley ME. 2001. The Incas and their Ancestors: The Archaeology of Peru. London: Thames and Hudson, Ltd.

Murra JV. 1968. An Aymara kingdom in 1567. Ethnohistory 15:115-151.

Murra JV. 1985a. "El Archipielago Vertical" Revisited. In: Masuda S, Izumi Shimada, and Craig Morris, editor. Andean Ecology and Civilization. Tokyo: University of Tokyo Press. p 3-13.

Nachman MW, Brown WM, and Stoneking M. 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics 142:953-963.

Nasidze I, Ling EYS, Quinque D, Dupanloup I, Cordaux R, Rychkov S, Naumova O, Zhukova O, Sarraf-Zadegan N, Naderi GA and others. 2004. Mitochondrial DNA and Y-Chromosome Variation in the Caucasus. Annals of Human Genetics 68(3):205-221.

Nasidze I, Quinque D, Ozturk M, Bendukidze N, and Stoneking M. 2005. MtDNA and Y-chromosome Variation in Kurdish Groups. Annals of Human Genetics 69(4):401-412.

Niehaus I. 2002. Ethnicity and the boundaries of belonging: reconfiguring Shangaan identity in the South African Lowveld. African Affairs 101:557-583.

Nunney L. 1993. The influence of mating system and overlapping generations on effective population size. Evolution 47(5):1329-1341.

Nurse D. 1997. The contributions of lingusitics to the study of history in Africa. Journal of African History 38:359-391.

O’Rourke DH, and Suarez B. 1986. Patterns and correlates of genetic variation in South Amerindians. . Annals of Human Biology 13:13-31.

Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, and Stoneking M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nature Genetics 29(20-21).

Park KS, Chan JC, Chuang L-M, Suzuki S, Araki E, Nanjo K, Ji L, Ng M, Nishi M, Furuta H and others. 2008. A mitochondrial DNA variant at position 16189 is associated with type 2 diabetes mellitus in Asians. Diabetologia 51:602-608.

Parker GJ. 1963. La Clasificatión genética de los dialectos Quechuas. Revista del Museo Nacional 32:241-252.

Pereira L, Gusmao L, Alves C, Amorim A, and Prata MJ. 2002. Bantu and European Y-lineages in Sub-Saharan Africa. Annals of Human Genetics 66(5-6):369-378.

Pérez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martínez-Arias R, Clarimón J, Fiori G, Luiselli D, Facchini F and others. 1999. Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. American Journal of Human Genetics 65:208-219.

Pfeiffer H, Brinkmann B, Hühne J, Rolf B, Morris AA, Steighner R, Holland MM, and Forster P. 1999. Expanding the forensic German mitochondrial DNA control region database: genetic diversity as a function of sample size and microgeography. International Journal of Legal Medicine 112(5):291-298.

Phillipson DW. 2005. African Archaeology. Cambridge: Cambridge University Press.

Piazza A, Rendine S, Minch E, Menozzi P, Mountain J, and Cavalli-Sforza LL. 1995. Genetics and the origin of European languages. Proceedings of the Davenport Academy of Sciences USA 92:5836-5840.

Pilkington M, Wilder J, Mendez F, Cox M, Woerner A, Angui T, Kingan S, Mobasher Z, Batini C, Destro-Bisol G and others. 2007. Contrasting signature of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa. Molecular Biology and Evolution 25(3):517-525.

Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, Carracedo Á, and Comas D. 2004. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Human Genetics 115:439-447.

Posada D, and Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. Trends in Ecology and Evolution 16(1):37-45.

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, and McElreavey K. 1999. Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. Nature Genetics 23:437-441.

Rando JC, Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM, and Bandelt H-J. 1998. Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with European, Near-Eastern, and sub-Saharan populations. Annals of Human Genetics 62(6):531-550.

Ray N, Currat M, and Excoffier L. 2003. Intra-deme molecular diversity in spatially expanding populations. Molecular Biology and Evolution 20:76-86.

Reich D, Patterson N, Kircher M, Delfin F, Nandineni Madhusudan R, Pugach I, Ko Albert M-S, Ko Y-C, Jinam Timothy A, Phipps Maude E and others. 2011. Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. American Journal of Human Genetics 89(4):516-528.

Reich D, Thangaraj K, Patterson N, Price AL, and Singh L. 2009. Reconstructing Indian population history. Nature 461(7263):489-494.

Rexová K, Bastin Y, and Frynta D. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. Naturwissenschaften 93(4):189-194.

Richards M, Côrte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt H-J, and Sykes B. 1996. Paleolithic and Neolithic lineages in the European mitochondrial gene pool. American Journal of Human Genetics 59:185-203.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T and others. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. American Journal of Human Genetics 67(5):1251-1276.

Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, Scozzari R, Macaulay V, and Torroni A. 2003. Extensive female-mediated gene flow from Sub-Saharan Africa into Near Eastern Arab Populaitons. American Journal of Human Genetics 72:1058-1064.

Rickards O, Martínez-Labarga C, Lum JK, De Stefano GF, and Cann RL. 1999. mtDNA history of the Cayapa Amerinds of Ecuador: detection of additional founding lineages for the Native American populations. American Journal of Human Genetics 65:519-530.

Robertson JH, and Bradley R. 2000. A new paradigm: the African Early Iron Age without Bantu migrations. History in Africa 27:287-323.

Rodriguez-Delfin LA, Rubin-de-Celis VE, and Zago MA. 2001. Genetic Diversity in an Andean Population from Peru and Regional Migration Patterns of Amerindians in South America: Data from Y Chromosome and Mitochondrial DNA. Human Heredity 51:97-106.

Rogers A, and Harpending H. 1992. Population growth makes waves in the distribution of pairwise differences. Molecular Biology and Evolution 9:552-569.

Rogers AR. 1995. Genetic evidence for a Pleistocene population expansion. Evolution 49(4):608-615.

Rosa A, Ornelas C, Jobling M, Brehm A, and Villems R. 2007. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. BMC Evolutionary Biology 7(1):124.

Rostworowski de Diez Canseco M. 1999. History of the Inca realm. Iceland HB, translator. Cambridge: Cambridge University Press.

Rowe JH. 1946. Inca Culture at the Time of the Spanish Conquest. In: Steward J, editor. Handbook of South American Indians. Washington, DC: Bureau of American Ethnology. p 183-330.

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, and Wallace DC. 2004. Effects of Purifying and Adaptive Selection on Regional Variation in Human mtDNA. Science 303(5655):223-226.

Salas A, Richards M, De le Fe T, Lareu M-V, Sobrino B, Sánchez-Diz P, Macaulay V, and Carracedo A. 2002. The making of the African mtDNA landscape. American Journal of Human Genetics 71:1082-1111.

Salas A, Richards M, Lareu M-V, Scozzari R, Coppa A, Torroni A, Macaulay V, and Carracedo A. 2004. The African diaspora: Mitochondrial DNA and the Atlantic slave trade. American Journal of Human Genetics 74:454-465.

Salem A-H, Badr FM, Gaballah MF, and Pääbo S. 1996. The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. American Journal of Human Genetics 59:741-743.

Santos M, Ward RH, and Barrantes R. 1994. mtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. Human Biology 66(6):963-977.

Schmitt R, Bonatto S, Freitas L, Muschner V, Kill K, Hurtado A, and Salzano F. 2004. Extremely limited mitochondrial DNA variation among the Aché Natives of Paraguay. Annals of Human Biology 31:87-94.

Schneider S, Roessli D, and Excoffier L. 2000. Arlequin 2.0: A software for population genetic data analysis. Geneva, Switzerland: University of Geneva Genetics and Biometry Laboratory.

Schoenbrun DL. 1993. We are what we eat: ancient agriculture between the Great Lakes. Journal of African History 34:1-31.

Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D, Arredi B, Destro-Bisol G, De Stefano G, Rickards O and others. 1999. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to inter affinities among African populations. American Journal of Human Genetics 65:829-846.

Ségurel L, Martínez-Cruz B, Quintana-Murci L, Balaresque P, Georges M, Hegay T, Aldashev A, Nasyrova F, Jobling MA, Heyer E and others. 2008. Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study. PLoS Genetics 4(9):e1000200.

Seielstad M, Minch E, and Cavalli-Sforza LL. 1998. Genetic evidence for a higher female migration rate in humans. Nature Genetics 20:278-280.

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ and others. 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: Inferences on the Neolithization of Europe and later migratory events in the Mediterranean area. American Journal of Human Genetics 74:1023-1034.

Semino O, Sanchiara Benerecetti AS, Falaschi F, Cavalli-Sforza LL, and Underhill PA. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. American Journal of Human Genetics 70:265-268.

Serre D, and Pääbo S. 2004. Evidence for Gradients of Human Genetic Diversity Within and Among Continents. Genome Research 14(9):1679-1685.

Sherry ST, and Batzer MA. 1997. Modeling human evolution - to tree or not to tree? Genome Research 7:947-949.

Sigurðardottir S, Helgason A, Gulcher JR, Stefansson K, and Donnelly P. 2000. The mutation rate in the human mtDNA control region. American Journal of Human Genetics 66(5):1599-1609.

Skeldon R. 1977. The evolution of migration patterns during urbanization in Peru. Geographical Review 67(4):394-411.

Slatikin M, and Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555-562.

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

Sokal RR, Oden NL, and Wilson C. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351:143-145.

Spear T. 1981. Traditions of origin and their interpretation: the Mijikenda of Kenya. Athens, OH: Ohio University Center for International Studies.

Spear TT. 1974. Traditional myths and historian's myths: variations on the Singwaya theme of Mijikenda origins. History in Africa 1:67-84.

Spear TT. 1977. Traditional myths and linguistic analysis: Singwaya revisited. History in Africa 4:229-264.

Stanish C. 2003. Ancient Titicaca: the evolution of complex society in southern Peru and northern Bolivia. Berkeley: University of California Press.

Stark LR. 1985a. Ecuadorian highland Quechua: history and current status. In: Manelis Klein HE, and Stark LR, editors. South American Indian languages: Retrospect and prospect. Austin: University of Texas Press. p 443-479.

Stark LR. 1985b. The Quechua language in Bolivia. In: Manelis Klein HE, and Stark LR, editors. South American Indian languages: retrospect and prospect. Austin: University of Texas Press. p 516-545.

Stoneking M. 1998. Women on the move. Nature Genetics 20:219-220.

Strassmann BI. 1997. Polygyny as a Risk Factor for Child Mortality among the Dogon. Current Anthropology 38(4):688-695.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437- 460.

Tajima F. 1989a. The effect of change in population size on DNA polymorphism. Genetics 123:597-601.

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

Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Martinez Labarga C, Rickards O, Tyler-Smith C, Pena SDJ, and Santos FR. 2001. Genetic Differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. American Journal of Human Genetics 68(6):1485-1496.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-1073.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426(6968):789-796.

The International HapMap Consortium. 2004. Integrating ethics and science in the International HapMap Project. Nat Rev Genet 5(6):467-475.

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, and Goldstein DB. 2000. Y Chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba-the "Black Jews of southern Africa. American Journal of Human Genetics 66:674-686.

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U and others. 2007. History of Click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Molecular Biology and Evolution 24(10):2180-2195.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O and others. 2009. The Genetic Structure and History of Africans and African Americans. Science 324(5930):1035-1044.

Torero A. 1964. Los dialectos quechuas. Anales Cienifico de la Universidad Agraria 2:446-478.

Torero A. 1987. Lenuas y pueblos altiplanicos en Torno al Siglo XIV. Revista Andina 5.

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, and Wallace DC. 1993a. Asian affinities and continental radiation of the four founding Native American mtDNAs. American Journal of Human Genetics 53:563-590.

Torroni A, Schurr TG, and Yang CC. 1992. Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. Genetics 130:153-162.

Torroni A, Sukernik RI, and Schurr TG. 1993b. mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. American Journal of Human Genetics 53:591-608.

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

Vansina J. 1995. New linguistic evidence and 'the Bantu Expansion'. Journal of African History 36:173-195.

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

Vitzthum VJ, Bentley GR, Spielvogel H, Caceres E, Thornburg J, Jones L, Shore S, Hodges KR, and Chatterton RT. 2002. Salivary progesterone levels and rate of ovulation are significantly lower in poorer than in better-off urban-dwelling Bolivian women. Human Reproduction 17(7):1906-1913.

Vitzthum VJ, Spielvogel H, and Thornburg J. 2004. Interpopulational differences in progesterone levels during conception and implantation in humans. Proceedings of the National Academy of Sciences of the United States of America 101(6):1443-1448.

Vitzthum VJ, Thornburg J, and Spielvogel H. 2009. Seasonal modulation of reproductive effort during early pregnancy in humans. American Journal of Human Biology 21(4):548-558.

Vona G, Falchi A, Moral P, Caló C, and Varesi L. 2006. Mitochondrial sequence variation in the Guahibo Amerindian population from Venezuela. American Journal of Physical Anthropology 127:361-369.

Wakeley J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. Journal of Molecular Evolution 37:613-623.

Ward RH, Salzno FM, Bonatto SL, Hutz MH, Coimbra CEA, and Santos RV. 1996. Mitochondrial DNA polymorphism is three Brazilian Indian Tribes. American Journal of Human Biology 8:317-323.

Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, and Pääbo S. 1996. mtDNA sequence diversity in Africa. American Journal of Human Genetics 59:437-444.

Watson E, Forster P, Richards M, and Bandelt H-J. 1997. Mitochondrial footprints of human expansions in Africa. American Journal of Human Genetics 61:691-704.

Watterson G-A. 1975. On the number of segregation sites in genetical models without recombination. Theoretical Population Biology 7:256-276.

Wegmann D, Currat M, and Excoffier L. 2006. Molecular Diversity After a Range Expansion in Heterogeneous Environments. Genetics 174(4):2009-2020.

White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, and Clark Howell F. 2003. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature 423(6941):742-747.

Whitehead NL. 1994. The ancient Amerindian polities of Amazon, the Orinoco, and the Atlantic coast: a preliminary analysis of their passage from antiquity to extinction. In: Roosevelt A, editor. Amazonian Indians from prehistory to the present. Tucson: University of Arizona Press. p 33-53.

Wilder J, Mobasher Z, and Hammer M. 2004a. Genetic evidence for unequal effective population sizes of human females and males. Molecular Biology and Evolution 21:2047-2057.

Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, and Hammer MF. 2004b. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nature Genetics 36:1122-1125.

Willis J. 1993. Mombasa, the Swahili, and the making of the Mijikenda. Hargreaves JD, Twaddle M, and Ranger T, editors. Oxford: Clarendon Press.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyal H and others. 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. European Journal of Human Genetics 13:867-876.

Wright DR. 1999. "What do you mean there were no tribes in Africa?": thoughts on boundaries- and related matters-in precolonial Africa. History in Africa 26:409-426.

APPENDICES

APPENDIX A

SUMMARY OF ANALYTICAL METHODS USED IN THIS DISSERTATION

Analysis of Molecular Variance (AMOVA) is a method to detect a evidence of population subdivision (Excoffier et al. 1992).

Fu's *Fs* statistic is a test of selective neutrality, and it evaluates the probability of observing the same or fewer numbers of alleles (k) given the observed θ_π, and is defined as

$Fs = \ln\left(\frac{S'}{1-S'}\right)$, where $S'=\Pr(K \geq k_{obs} \mid \theta=\theta_\pi)$ (Fu, 1996).

Populations that experienced expansion have significantly large negative *Fs*.

Migration rate, $M=2N_f m$ is estimated from the spatial expansion model of mismatch distribution assuming infinite-island model of migration (Excoffier 2004). $2N_f m$ is also estimated using AMOVA Φ_{ST} (Seielstad et al. 1998) and MIGRATE (Beerli and Felsenstein 2001). Both methods assume island model migration.

Mismatch distribution is an analysis of nucleotide differences between sequences from a single population. The test statistics available based on two demographic model (sudden demographic or spatial expansion) (Excoffier 2004)(Ray et al. 2003; Rogers and Harpending 1992; Slatikin and Hudson 1991).

Network is a phylogenetic method for intra-specific comparison (Bandelt et al. 1999).

APPENDIX A (continued)

SIMCOAL is a coalescent-based computer simulation program to test demographic model (Excoffier et al. 2000; Laval and Excoffier 2004)

Tajima's *D* is a test of selective neutrality, and it is based on a comparison of θ_π and θ_S, expressed as

$$D = \frac{\theta_\pi - \theta_\text{S}}{\sqrt{Var(\theta_\pi - \theta_\text{S})}}$$ (Tajima1989b).

Populations that experienced expansion have significantly large negative *D*.

$\boldsymbol{\theta_k}$**=2*N_fμ*** is estimated based on the relationship between the number of sequences (*k*) and the sample size assuming that the populations are panmictic (Ewens 1972), and it reflects recent demographic history.

$\boldsymbol{\theta_S}$**=2*N_fμ*** is estimated based on the relationship between the number of polymorphic, or segregating sites (*S*), and the sample size assuming that the populations are panmictic (Watterson 1975), and it reflects recent demographic history.

$\boldsymbol{\theta_\pi}$**=2*N_fμ*** is a measure of mean pairwise nucleotide ifferences, the mean number of mutational differences between two sequences (π) assuming that the populations are panmictic (Tajima 1983), and it reflects ancient demographic history.

Θ=2*N_fμ* is estimated using MIGRATE program accounting for the gene flow that took place between demes (Beerli and Felsenstein 2001).

APPENDIX B

CHAPTER 5 SUPPORTING MATERIALS

MIGRATE Analysis

The effective population size (Θ) was estimated for the 15 selected Latin American populations (TABLE XXVI). The lowland Bolivians were excluded from the analysis because of small sample size. Ancash were initially included in the analysis, but were dropped in the final analysis because the Θ values were inconsistent. The Θ estimates of three runs ranged from 0.0467 to 3.3583 suggesting that the sample size (n=33), although larger than some of the other population sampled, was too small for this group's higher genetic diversity. The Quechua and Aymara samples included in the analysis have high estimates of Θ. The Tayacaja have the highest Θ, which is significantly higher than the other groups. The Aymara La Paz have the third largest Θ (0.0374) following two Quechua populations. Like the other estimates of effective population size, the Chibchan groups have the lowest Θ values.

The effective population size (Θ) was estimated for the 19 selected Bantu populations (TABLE XXXVII). The Sanga and Nyanja were initially included for Θ estimates, but were removed from the final analysis because estimated Θ was inconsistent suggesting confidence interval is too large. East African Bantus tend to have small estimate of Θ compared to Central African and Southeastern African Bantus.

APPENDIX B (continued)

TABLE XXXVI

COMPARISON OF Θ TO θ_k, θ_s, AND θ_π ESTIMATED FOR SELECTED LATIN AMERICAN POPULATIONS

Θ		θ_k		θ_S		θ_π	
Tayacaja	0.3359	Tayacaja	53.609	Tayacaja	10.043	Pilaga	7.843
Quechua Puno	0.0925	Quechua Puno	35.666	Pilaga	8.092	Wounan	7.569
Aymara La Paz	0.0374	Aymara La Paz	26.488	Quechua Puno	8.077	Cayapa	7.253
Pilaga	0.0288	Pilaga	20.951	Aymara La Paz	7.265	Tayacaja	7.087
Yungay	0.0123	Yungay	17.738	Wounan	7.009	Mapuche/Pehuenche	6.823
Mapuche/Pehuenche	0.0110	Embera	12.155	Wichi	6.967	Wichi	6.798
Embera	0.0095	Mapuche/Pehuenche	10.38	Yungay	6.511	Embera	6.673
Wichi	0.0081	Wichi	9.764	Toba	5.865	Argentina Mapuche	6.427
Toba	0.0071	Wounan	9.256	Mapuche/Pehuenche	5.833	Yungay	6.203
Wounan	0.0058	Argentina Mapuche	6.417	Embera	5.057	Quechua Puno	6.15
Cayapa	0.0048	Toba	6.346	Argentina Mapuche	4.73	Toba	5.848
Argentina Mapuche	0.0044	Cayapa	3.226	Cayapa	4.544	Aymara La Paz	5.843
Huetar	0.0030	Huetar	2.728	Huetar	3.113	Ngobe	5.198
Ngobe	0.0030	Ngobe	2.057	Ngobe	2.73	Huetar	4.018
Kuna	0.0010	Kuna	1.807	Kuna	2.122	Kuna	3.882

Red (South-Central Andeans), Blue (North-Central Andeans), and Black (other populations)

APPENDIX B (continued)

TABLE XXXVII

COMPARISON OF Θ ESTIMATES AMONG THE BANTUS

	Θ		θ_k		θ_S		$\theta\pi$
Ronga	0.1804	Bassa	99.974	Taita	14.936	Ewondo	12.331
Ngoumba	0.1305	Ngoumba	90.184	Bassa	13.652	Hutu	12.139
Bakaka	0.1030	Bamileke	63.764	Shona	12.914	Taita	11.946
Shona	0.1017	Ronga	63.012	Bakaka	12.725	Bakaka	11.807
Bamileke	0.0905	Taita	61.419	Mijikenda	12.605	Makhwa	11.297
Taita	0.0877	Ewondo	58.673	Ngoumba	11.954	Bassa	11.165
Bassa	0.0783	Bakaka	56.184	Bamileke	11.942	Nyungwe	10.998
Ewondo	0.0681	Shona	49.001	Hutu	11.852	Ngoumba	10.701
Nyungwe	0.0674	Hutu	45.529	Ewondo	11.679	Shona	10.558
Chwabo	0.0672	Mijikenda	36.651	Ronga	10.424	Ronga	10.416
Shangaan	0.0651	Nyungwe	34.968	Turu	10.185	Shangaan	10.095
Turu	0.0633	Shangaan	32.457	Nyungwe	9.865	Mijikenda	9.930
Lomwe	0.0596	Chwabo	25.594	Shangaan	9.601	Turu	9.770
Bateke	0.0469	Turu	19.273	Bateke	9.377	Bamileke	9.379
Hutu	0.0425	Bateke	15.913	Makhwa	9.302	Lomwe	9.337
Makhwa	0.0405	Makhwa	11.747	Chwabo	8.174	Chwabo	9.127
Mijikenda	0.0381	Lomwe	11.747	Lomwe	7.610	Sena	8.233
Bubi	0.0248	Sena	8.613	Bubi	7.089	Bateke	7.736
Sena	0.0218	Bubi	7.469	Sena	6.115	Bubi	7.629

Blue (East African Bantus), Red (Southeastern African Bantus), Black (Central African Bantus)

APPENDIX B (continued)

TABLE XXXVII

LATIN AMERICAN POPULATIONS MISMATCH DISTRIBUTION SUM OF SQUARE DEVIATION (*SSD*) AND *P*-VALUES

Populations	Subsistence Pattern[a]	Demographic expansion model *SSD*[b] (*P-Values*)[c]	Spatial expansion model *SSD* (*P-Values*)
Central Andes			
Ancash	3	0.012 (0.37)	**0.019 (0.03)**
Arequipa	3	0.038 (0.17)	**0.035 (0.04)**
Aymara La Paz	3	0.021 (0.25)	0.026 (0.07)
Aymara Puno	3	0.015 (0.69)	0.018(0.39)
Quechua Puno	3	0.003 (0.69)	0.003 (0.67)
Tayacaja	3	0.002 (0.81)	0.004 (0.36)
Tupe	3	0.046 (0.22)	0.051 (0.20)
Yungay	3	0.005 (0.70)	0.007 (0.56)
Mesoamerica			
Quiche	3	0.043 (0.21)	**0.043 (0.00)**
Western Lowland South Americans			
Cayapa	2	**0.097 (0.00)**	**0.078 (0.04)**
Embera	2	**0.043 (0.01)**	**0.052 (0.04)**
Wounan	2	**0.048 (0.01)**	**0.048 (0.02)**
Lowland Bolivians, Dept. of Beni			
Chimane/Moseten	2	0.011 (0.63)	0.015 (0.33)
Movima	2	0.006 (0.81)	0.005 (0.87)
Moxo	2	0.013 (0.44)	0.023 (0.19)
Quechua Beni	3	**0.600 (0.00)**	0.033 (0.58)
Yuracare	2	**0.052 (0.01)**	**0.054 (0.04)**

APPENDIX B (continued)

TABLE XXXVII (continued)

LATIN AMERICAN POPULATIONS MISMATCH DISTRIBUTION SUM OF SQUARE DEVIATION (*SSD*) AND *P*-VALUES

Populations	Subsistence Pattern[a]	Demographic expansion model *SSD*[b] (*P-Values*)[c]	Spatial expansion model *SSD* (*P-Values*)
Gran Chaco			
Pilaga	1	0.002(0.87)	0.003 (0.67)
Toba	1	0.012(0.26)	0.006 (0.88)
Wichi	1	**0.025 (0.01)**	0.025 (0.19)
Southern Andes			
Mapuche (Argentina)	1	**0.021 (0.02)**	0.016 (0.37)
Yaghan	2	**0.032 (0.03)**	0.034 (0.34)
Mapuche/Pehuenche	1	**0.032 (0.00)**	0.023 (0.30)
Chibchans			
Arsario	2	0.131 (0.07)	**0.107 (0.00)**
Huetar	2	0.071 (0.15)	0.035 (0.61)
Ijka	2	**0.048 (0.02)**	0.025 (0.27)
Kogi	2	**0.389 (0.00)**	0.083 (0.25)
Kuna	2	0.128 (0.20)	0.066 (0.15)
Ngobe	2	0.097 (0.81)	0.067 (0.08)

APPENDIX B (continued)

TABLE XXXVII (continued)

LATIN AMERICAN POPULATIONS MISMATCH DISTRIBUTION SUM OF SQUARE DEVIATION (*SSD*) AND *P*-VALUES

Populations	Subsistence Pattern[a]	Demographic expansion model *SSD*[b] (*P-Values*)[c]	Spatial expansion model *SSD* (*P-Values*)
Other Lowland South Americans			
Ache	1	**0.042 (0.00)**	0.022 (0.25)
Ayoreo	1	**0.292 (0.00)**	0.025 (0.41)
Guahibo	1	0.026 (0.22)	0.026(0.18)
Kaingang	1	**0.081 (0.02)**	0.047 (0.32)
Kaiowa	2	**0.031 (0.00)**	0.057 (0.13)
M'bya	2	0.111 (0.07)	0.062 (0.15)
Nandeva	2	0.025 (0.44)	0.024 (0.22)
Wayuu	2	**0.109 (0.01)**	0.076 (0.15)
Xavante	1	0.104 (0.09)	0.094 (0.11)
Yanomamo	2	0.029 (0.30)	0.031 (0.09)
Zoro/Gaviao	2	0.020 (0.40)	0.019 (0.68)

[a] Subsistence strategies are categorized into 1 forager, 2 horticulturalist, and 3 agriculturalist/pastoralist.

[b] Sum of Standard Deviation (*SSD*) between observed and expected mismatch distribution

[c] *P-values* of the SSD statistics

APPENDIX B (continued)

TABLE XXXIX

BANTU POPULATIONS MISMATCH DISTRIBUTION SUM OF SQUARE DEVIATION (*SSD*) AND *P*-VALUES

	Demographic expansion model *SSD* (*P-values*)	Spatial expansion model *SSD* (*P-Values*)
East Africa		
Taita	0.0029 (0.39)	0.0060 (0.10)
Mijikenda	0.0036 (0.63)	0.0051 (0.60)
Kikuyu	**0.0218 (0.02)**	0.0094 (0.15)
Sukuma	0.0087 (0.19)	0.0107 (0.06)
Hutu	0.0163 (0.07)	**0.0216 (0.01)**
Turu	0.0135 (0.31)	0.0166 (0.12)
Southeastern Africa		
Ronga	0.0044 (0.69)	0.0051 (0.45)
Shona	0.0029 (0.59)	0.0042 (0.39)
Nyungwe	0.0170 (0.16)	0.0219 (0.07)
Shangaan	0.0099 (0.29)	0.0112 (0.39)
Chwabo	0.0067 (0.89)	0.0106 (0.60)
Chopi	0.0049 (0.74)	0.0077 (0.72)
Tonga	0.0249 (0.10)	0.0248 (0.23)
Nyanja	0.0242 (0.10)	0.0233 (0.22)
Makhwa	0.0207 (0.19)	0.0208 (0.47)
Lomwe	0.0282 (0.21)	0.0259 (0.56)
Sena	0.0204 (0.36)	0.0216 (0.32)
Central Africa		
Bassa	0.0013 (0.87)	0.0022 (0.54)
Ngoumba	0.0023 (0.77)	0.0042 (0.21)
Mbundu	**0.0124 (0.04)**	0.0055 (0.08)
Bamileke	0.0052 (0.19)	0.0056 (0.11)
Ewondo	0.0026 (0.45)	0.0032 (0.27)
Bakaka	0.0075 (0.37)	**0.0112 (0.04)**
Sanga	0.0022 (0.88)	0.0033 (0.84)
Bateke	0.0064 (0.13)	0.0045 (0.67)
Bubi	**0.0208 (0.02)**	0.0144 (0.46)

VITA

NAME: Ken Batai

EDUCATION: B.A., Anthropology, Southern Illinois University, Carbondale, 2000

M.A., Anthroplogy, University of Illinois at Chicago, 2003

Ph.D., Anthropology, University of Illinois at Chicago, 2012

PROFESSIONAL PREPERATIONS: Graduate Researcher, Institute of Human Genetics, University of Illinois at Chicago, 2011

Research Assistant, Department of Anthropology, University of Illinois at Chicago, 2009

Adjunct Instructor for Biological Science 102: Human Genetics at Triton College, 2006

Teaching Assistant, University of Illinois at Chicago, 2002

GRANT AWARD: University of Illinois at Chicago, Graduate Student Council Travel Award, 2011

University of Illinois at Chicago Department of Anthropology Charles Reed Fund, 2008

Sigma Xi Grant-in-Aid of Research, 2007

Provost's Award, University of Illinois, Chicago, 2004

MEMBERSHIPS: American Association of Cancer Research

American Association of Physical Anthropologists

American Society of Human Genetics

Lamba Alpha National Collegiate Honors Society for Anthropology

PRESENTATIONS: Batai K, Shah E, Ruden M, Newsome J, Murphy A, Kittles RA. Fine-mapping of IL-16 variants associated with prostate cancer risk in African Americans. Paper presented at UIC Cancer Center Research Forum, March 6, 2012

Kittles RA, Batai K, Newsome J, Ruden M, Shah E, Murphy A. Predictors of Serum Vitamin D Levels in African American and European American Men in Chicago. Paper presented at UIC Cancer Center Research Forum, March 6, 2012

Williams SR, Batai K, Vitzthum. A mtDNA population genetic study of the Aymara. Paper presented at the 40th Annual Midwest Conference on Andean and Amazonian Archaeology and Ethnohistory in Chicago, February 25-26, 2012

Batai K, Shah E, Murphy A, Kittles R. IL-16 variants are associated with prostate cancer risk in African Americans. UIC College of Medicine Research Forum 2011, November 11, 2011.

Batai K, Kusimba CM, Leenheer E, Williams SR. Mitochondrial DNA (mtDNA) heterogeneity within and among East African Bantu ethnic groups and their complex evolutionary histories. The 12th International Congress of Human Genetics and the 61st American Society of Human Genetics annual meeting in Montreal, Canada, October 11-15, 2011.

Batai K, Beisner E, Shah E, Castaneda L, Smith D, Murphy A, Kittles RA. IL-16 variants associated with prostate cancer risk in African Americans. The 4th American Association for Cancer Research Conference on the Science of Cancer Health Disparities in Racial/Ethnic Minorities and the Medically Underserved in Washington, DC, September 18-21, 2011

Batai K, Vitzthum V, Williams SR. Aymara mtDNA variation and demographic history in the Central Andes. The 80th annual meeting of the American Association of Physical Anthropologists in Minneapolis, MN, April 12-16, 2011.

Arroyo JP, Batai K, Williams SR. Mitochondrial DNA variation at position 16189 and diabetes: frequency amongst South Eastern Kenyan populations. The 79th annual meeting of the American Association of Physical Anthropologists in Albuquerque, NM, April 14 to April 17, 2010.

Williams SR, Kusimba CM, Batai K. Ancient Swahili origins: a mitochondrial study of ancient inhabitants of the Kenyan coast. The 79th annual meeting of the American Association of Physical Anthropologists in Albuquerque, NM, April 14 to April 17, 2010.

Batai K, Babrowski KB, Williams SR. The origin of the Taita and Mijikenda ethnic groups of Kenya: A mitochondrial case study. The 78th annual meeting of the American Association of Physical Anthropologists in Chicago, IL, April 2 to April 4, 2009.

Batai K, Williams SR. Mitochondrial DNA genetic diversity and New World demographic history. The 77th annual meeting of the American Association of Physical Anthropologists in Columbus, OH, April 9 to April 12, 2008.

Batai K, Williams SR. Reconstructing the settlement history of the central Andes from mitochondrial DNA analyses. The 76th annual meeting of the American Association of Physical Anthropologists in Philadelphia, PA, March 27 to April 1, 2007.