Genetic Study of LCT- Enhancer, Y chromosome and Mitochondrial DNA Variation in Some Ethnic Groups in Ethiopia

Solomon Balemi Adugna

A Dissertation Submitted to the Department of Microbial, Cellular and Molecular Biology Presented in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy (Biology, Applied Genetics)

Addis Ababa University

School of Graduate Studies

Addis Ababa, Ethiopia

June, 2018

Genetic Study of LCT- Enhancer, Y chromosome and Mitochondrial DNA Variation in Some Ethnic Groups in Ethiopia

Solomon Balemi Adugna

A Dissertation Submitted to the Department of Microbial, Cellular and Molecular Biology Presented in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy (Biology, Applied Genetics)

Supervisor

Prof. Endashaw Bekele

Professor of Genetics, Addis Ababa University, Applied Genetics Stream, Ethiopia

June, 2018

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

Genetic Study of LCT- Enhancer, Y chromosome and Mitochondrial DNA Variation in Some Ethnic Groups in Ethiopia

By

Solomon Balemi Aduna

A Thesis Presented to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the PhD in Biology (Applied Genetics)

Approved By Examining Board:

Name	Signature
Prof. Endashaw Bekele	**(Advisor) _____**
Prof. Manjinder Sandhu	**(External Examiner)_____**
Dr. Kassahun Tesfaye	**(Internal Examiner)_____**
Gurja Belay (PhD)	**(Chairman)_____**

DECLARATION

I hereby declare that this thesis is my original work and has not previously been submitted by me or anybody else at another university. The references which I have made use of are well acknowledged at the appropriate place in the thesis.

Name	Signature	Date
Solomon Balemi	_____	_____

ABSTRACT

Genetic Study of LCT- Enhancer, Y chromosome and Mitochondrial DNA Variation in Some Ethnic Groups in Ethiopia

Solomon Balami Adugna (PhD Candidate)

Addis Ababa University, 2018

Ethiopia has been the centre of Human out of African migration and thus occupies a strategic place to study of human evolutionary population genetics. It is regarded as a centre where key evolutionary events have taken place; as a place of human origin, a corridor for human migration and centre for origin of earlier human culture. Despite of the importance of Ethiopian population genetics, the human population genetic studies across geographically, ethnically and linguistically diverse Ethiopia population is still not exhaustively covered and understood. To contribute to these, genomic DNA was collected from five ethnic groups of Ethiopia; Nuer, Berta, Gumuz, Shinasha and Mao to examine the correlation of LP phenotype with genotype and study the patterns of Y chromosome and Mitochondrial DNA Variations. From these groups, a total of 155 participants, representing two language families; Afro-Asiatic and Nilo-Saharan in Ethiopia, were recruited for genetic analysis.

The phenotypes of the lactase persistence (LP) were examined using breathe hydrogen test (BHT). The result showed that 83 (53.55%) and 68 (43.87%) of the total 155 study subject had a positive breath (≥20 ppm) and negative breath test (<20 ppm) respectively. Individuals with positive breath test are LP, while those with negative breath test are lactase non persistence (LNP). Moreover, 4 (2.58%) of NHP individuals were also observed. Milk drinking behavior and post gastrointestinal symptoms was also recorded. The higher frequency of LP was observed in the milk drinkers on the daily basis than the non milk drinkers. In all groups, majority of the LNP and very few LP individuals showed various post gastrointestinal symptoms.

Y chromosomes and mitochondrial variation were studied in Ethiopian populations. The Y chromosome haplogroup variations were studied by typing four major Y haplogroups (A, B E and F) in 120 unrelated adult males. The Y chromosome study showed that the Ethiopian populations studied contain haplogroup A-M13, B-M60, E-M78 and F-M89 in percentages of 44.04, 39.28, 86.11 and 21.42 % respectively, showing higher Y chromosome diversity in Ethiopia than elsewhere in Africa. The ancient and deepest haplogroups A-M13 and B-M60 and the most widespread haplogroup E-M78 was present at such high frequencies in Nilo-Saharan than Afro-Asiatic speaking groups. Intersetingly, haplogroup E-M78 which is defined by E1b1b1a1 terminal branch was found at remarkably high frequency in present study.

Based on sequencing of the mitochondrial Cytochrome C Oxidase subunit II (MT-COXII), **15 substitutions as SNPs were observed, of which the majority (80%) were Transitions. Of 15, 5 SNPs were new mutations.** ***From mitochondrial haplotype variations analysis, a total of 13 haplotypes with 138 polymorphic sites were observed. Also, the evidence for the sign of past demographic expansion and higher genetic diversity was noted. Moreover, phylogenetic analyses indicate that East Africans in general and Western Ethiopia in particular contains more ancestral lineages in comparison to various continental populations placing them at the root of the human evolutionary tree. The results also confirm East Africa as the most probable location from which dissemination of human out of Africa has taken place. The study shows the higher level of MT-COXII sequence variation within Ethiopia in comparison to other East Africans and the continental samples, and suggests further detailed studies on Ethiopian population to contribute in filling the existing gaps in the MT-COXII database. This data implicated in the need to carry out detailed studies of human genetic variation that includes more African populations; particularly Ethiopians.***

Key words: LP/LNP, BHT, Y chromosome, MT-COXII, Haplogroup, PCR-RFLP

Dedication

This work is dedicated to my late father, without the deep passion and devotion of whom it would not have been possible to attain this level of education. Beyond their almost unbelievable support, my mum and family are the most important people in my world and I also dedicate this thesis to them next to him.

ACKNOWLEDGMENTS

First of all I would like to thank my supervisor Professor Endashaw Bekele for his assistance, invaluable comments, constructive criticism, guidance and provision of suplies. Truly, he deserves profound thanks for his tremendous academic support and contribution starting from the beginning to the end during my stay. Specially, for the moral help he provided me to stay in the system even when I was about to give up many times when things became difficult, I am forever grateful to him. I am also grateful for his support throughout this work including struggling with correcting my poor English language. I heartily thank him for taking so much of his time throughout my PhD work.

Special thanks also go to Professor Muntaser Ibrahim for his continuous guidance in my lab work, and for his support hosting me in his lab at the University of Khartoum, Institute of Endemic Diseases, and Department of Molecular Biology. I am particularly indebted to him for his permit to do my lab work in his laboratory under very resource limited conditions. He has been enthusiastic and a truly dedicated mentor with amazing experiences. This and his continuous follow up and inspiration generously contributed to the work presented in this thesis.

In general, I am short of words to express the contribution of Professor Endashaw and Professor Muntaseer to bring the dissertation to this point. Thank you both for showing me such warm-hearted kindness, patience, hardworking help, endurance, etc. you both are my legends and I will always try to follow in your footsteps.

I am grateful to all the study participants of this research work for giving the time (especially for the amount of time you spent to participate in the BHT), for providing necessary milk drinking habit and health related information, and for providing mouth swab samples; without of which this dissertation would not have been completed.

I am grateful to people in the laboratory of Professor Muntaser El Ebrahim for their humble assistance starting from DNA Extraction to sending samples for sequencing. Special thanks must go to Mohammad Saad (Lab. Assistance) and Musab Ali (PhD Candidate at UOK) for their unreserved assistance in molecular lab work and in the process of sending samples for sequencing and in sequence data analysis respectively. I would also like to thank Dr.Tamiru Oljira, Dr. Ephrem Mekonnen, Dr. Dagmawit Chombe, Fikadu Gaddisa and Delelegn Woyessa for their contribution in sharing their experiences willingly and encouraging me continuously throughout the work of the thesis. The later also for his help in printing this dissertation.

I am particularly grateful to the Department of Microbial, Cellular and Molecular Biology for the provision of financial support and other multifaceted support from the start to the end of this thesis. I am also grateful to Jimma University, Biology Department. The University must be thanked for supporting by paying me salary. A number of people and colleagues from the University and from other field data collection areas deserve thanks for their contribution through organizing the field data collection. Some of them include Gosay Dasse, Gatluak Tut, Siyum Mengistu and many more. Among them Gosaye Dasse and Gatluak Tut deserves special thanks for providing me very good assistance in arranging and conducting field samples and data collection from Benishanguel Gumuz and Gambella regions respectively.

My parents and family deserve special thanks for their invaluable multifaceted support. They provided me with continuous encouragement, money and logistics supports at the right time when I completely lost a hope. In addition, I am grateful to my family for patiently waiting for my completion of this dissertation for this very long time. I understand the painful time you passed through to get this job done.

Thank you very much!

Solomon Balemi

TABLE OF CONTENTS **Page**

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ACCRONYMS

AMH	**Anatomically Modern Human**
AluI	***Arthrobacter luteus* First Derived Restriction Enzyme**
B.C. / YBP	**Before the Christ/ Years Before Present**
BHT	**Breath Hydrogen Test**
DYS	**DNA Y chromosome Segment**
KYA	**Kilo (thousands) Years Ago**
LCT	**Lactase Gene**
LP/LNP -	**Lactase Persistence /Lactase None Persistence**
LI	**Lactase intolerance**
MboI	***Moraxella bovis* initial Derived Restriction Enzyme**
MCM6	**Mini Chromosome Maintenance 6**
MJ	**Median Joining**
MT-COXII	**Mitochondrial encoded cytochrome C Oxidase Sub unit II**
mtDB	**Mitochondrial DNA Database**
mtDNA	**Mitochondrial DNA**
NJ	**Neighbor Joining**
NHP	**Non-Hydrogen Producing**
NlaIII -	***Neissereia lactamica* Third Isolated Restriction Enzyme**
NRY	**Non-Recombining portion of Y chromosome**
PCA	**Principal Component Analysis**
PCR-RFLP	**Polymerase Chain Reaction-Restriction Fragment Length Polymorphism**
PPM	**Parts Per Million**
SNPs	**Single Nucleotide polymorphisms**
STRs	**Short Terminal Repeats**
UEP	**Unique Event Polymoephisism**
YAP	**Y chromosome *Alu* Polymorphism**

1. INTRODUCTION

1.1. Background and Justification

Genetics of Lactase Persistence

The enzyme lactase is predominantly expressed in the microvilli of the small intestinal epithelial cells. Lactase is responsible to catalyze the hydrolysis of lactose to glucose and galactose, sugars that are easily absorbed into the bloodstream. Thus, lactase plays a crucial role in obtaining nutritional benefits from milk in humans. However, the ability to digest lactose declines rapidly after weaning and prevents the ability to digest lactose in the milk (Ingram, *et. al*., 2009). This in turn attributed to the progressively reduced expression of the lactase gene and thus the decreases in the production of the enzyme lactase in most human populations as age increases (Gerbault, *et. al*., 2010). Fermentation of undigested lactose produce H_2, CO_2 and CH_4 under the action of bacterial flora in the colon resulting in illness such as bloating, flatulence, abdominal pain and diarrhea (Robayo, *et. al.,* 2006). Individual with these characteristics are known as lactase non-persistence (LNP). However, some human populations continue to produce lactase throughout the adulthood and maintain their ability to digest lactose into adulthood (Gerbault, *et. al*., 2010). These individuals have the lactase persistence (LP) trait, due to continoues expression of the gene for lactase (LCT).

Phylogenetic analysis of the evolution of lactose digestion ability by Holden and Mace (1997) showed that the LP into adulthood is well correlated with the milk drinking traditions and confirmed that it is an evolutionary adaptation to pastoralist mode of life. Holden and Mace (2003) also suggested LP might have evolved and widespread in populations that had pastoralists' mode of life and the differences in human populations might be attributed to strong

recent and positive genetic selection of individuals with the ability to use much milk for its high nutritional advantages. It was hypothesized that, most likely, LP evolved from LNP common ancestors and dispersed in humans that had adopted cattle and drank milk (Mace, *et. al*., 2003; Holden & Mace, 1997; Ingram, *et. al*., 2009 and Arroyo, *et. al*., 2010) and it was supported by studies of ancient DNA from archeological skeletons (Burger, *et. al*., 2007).

Four decades ago, LP was wrongly considered to have evenly distributed worldwide (Romero, *et. al.*, 2011), but later, several studies indicated that the present distribution of LP is limited to certain geographical locations and is strongly correlated with dairy farming practices. In support of this, it was shown that LP is a population specific trait that varies widely in different populations across the world according to ethnicity (Ingram, *et. al*., 2009 and Romero, *et. al*., 2011). Moreover, three decades ago, little was known about the genetic basis of the LP into adulthood (Holden and Mace, 1997). Since then, numerous studies have been conducted in order to understand the genetic cause of LP or LNP. To this end scientists have investigated the regions within and surrounding the LCT gene. Investigations of within LCT gene have shown the existence of high diversity within the LCT gene, but none of them was shown to be associated with LP trait (Harvey, *et. al.*, 1998 and Hollox, *et. al.*, 2001). However, investigation of the LCT surrounding region have shown the existence of strong associations of LP trait with enhancer region upstream of the LCT gene located in the adjacent gene, MCM6 and have identified several mutations in the form of single nucleotide polymorphisms (SNPs) that appear to be highly correlated with LP (Ingram, *et. al.*, 2009). For instance, in Finnish families, Enattah, *et. al.* (2002) identified two mutations (SNPs); 13910C>T within intron 13 and 22018G>A within intron 9 of MCM6 gene which is located at 2q21 and control the production of lactase enzyme. The functional role of these SNPs was examined by Olds and Sibley (2003) and showed

their strong association with LP. Evidence showed that these mutations have rapidly evolved due to recent positive selections (Bersaglieri, *et. al.,* 2004). However, these variants were confirmed to be Europeans variants, and are absent in the major areas of East Africa including Ethiopia where high LP phenotype was reported. This have led to further work and several LP related mutations have been identified in different geographic regions and ethnic groups (Jones, *et. al.,* 2013; Ingram, *et. al*., 2009; Tishkoff, *et. al*., 2007 and Ingram, *et. al*., 2007).

Tishkoff, *et. al.* (2007) conducted a detailed genotype-phenotype association study in African pastoralists' populations and have identified (13907C>G (rs41525747), 13915T>G (rs41380347) and 14010G>C (rs145946881) mutations which are associated with LP trait, showing the action of convergent evolution of LP in different populations. Similar phenotype-genotype association study was conducted by Ingram, *et. al.* (2007) and confirmed the existence of multiple mutations and variations in different pastoralists' populations from sub-Saharan Africans and Middle East. Next to these, Enattah, *et .al.* (2008) identified two new mutations (13915 T>G and 3712 T>C) responsible for LP among Saudis population samples. Moreover, study by Ingram, *et. al.* (2009) and Jones, *et. al.* (2013) have reported a very high genetic diversity of the LCT regulatory region and have increased the number of identified mutations to five (14010 G>C, 13907 C>G, 13915 T>G, 13910 C>T and 14009 T>G) including the newly reported one by Jones, *et. al.* (2013). With reference to the case of European variants (13910C>T), functional studies were conducted for various mutant forms (13907*G, 13915*G, and 14010*C alleles). The results have shown that these variants have involved in maintenance of the LCT expression during adulthood by significantly enhancing transcription from the lactase gene promoter *in vitro* (Olds, *et. al.,* 2011).

All the expressed above studies have confirmed the absence of the 13910 C>T allele in the studied African samples. They have demonstrated that 13910 C>T allele has evolved from different allelic back ground and offered clear and direct evidence that it is not the causative of LP worldwide. Thus, the mutational basis of LP/LNP in Africa appears to be different from that in Europeans and the 13910 C>T variant did not account for LP in most people from African including Ethiopian where LP phenotype was reported to be high. So, either there may be several other variants to confer LP, yet unidentified in different populations of Africa, or that 13910C>T may not totally to be associated with LP and the true responsible mutations may be located elsewhere within the potential region close to this allele (where multiple independent mutations in around 100bp range were previously identified) or somewhat away from it.

To contribute on this, Phenotype-genotype correlation studies have been conducted using breath hydrogen test (BHT) and molecular methods. BHT is non-invasive, specific and sensitive diagnostic test based on the determination of exhaled hydrogen produced by bacterial flora in the colon after oral lactose load. Genetic analysis of the LCT enhancer regions were planned to conduct target DNA sequencing of 700 bp flanking the 13910 C>T allele or PCR-RFLP analysis of the PCR product of this region and exome sequencing. In comparison the PCR-RFLP is cost effective than target DNA sequencing and can be done at a conventional lab. In addition, the association studies using PCR-RFLP have been employed by several scientists (Tishkoff, *et. al.*, 2007; Ingram, *et. al*, 2009 and Jones, *et. al.*, 2013) and that these studies have confirmed the reliability of the association studies method to identify the mutations. In addition, exome sequencing looks useful to identify the genetic variants underlying LP phenotype. Hence, on these ground, breath hydrogen test and PCR-RFLP analysis of the PCR product were chosen to

examine the association of LNP/ LP phenotype with genotype, and also exome sequencing was chosen to generate complete information for the LCT gene (lactase coding) in Ethiopia samples.

The molecular study would provide better understanding of the frequency 1390C>T in ethnic groups who have had long tradition of pastoralism and milk drinking culture (Nuer, Bertha, Gumuz, Shinasha and Mao), where LP phenotype is reportedly higher. Representative study subjects were randomly selected based on their pure interest from these Ethiopian groups. The result of this study in combination with the similar studies previously conducted have provided the general picture of the existing frequency and distribution of *13910C>T* in Ethiopian populations and contributed to fulfill the knowledge gaps existed. The current study was designed to also evaluate whether the milk intakes by pastorals subjects are influenced by the LP genotypes and also to highlight the prevalence of mutations in LP alleles.

Genetic Relationships of Populations Using Y chromosome and mtDNA

Archeological and fossil evidence suggested that modern human have originated during the middle stone age around 160-200 KYA (Lopez, *et. al.,* 2015) and started settlement around 180 KYA in Africa, due to conducive environment provided by the Nile Valley (Phillipson, 1993). As described by Kaestle and Horsburgh (2002), anthropologists were immediately understood the abilities of new molecular techniques to investigate the question of human origin and population relationship, that could not have been addressed by traditional anthropological techniques and to provide additional lines of evidences. Biochemical and molecular markers on chromosomes (Y-chromosome, mitochondrial DNA (mtDNA), autosomal and X-chromosomal DNA) variations have revealed that African populations have ample and essential history from which we can learn to a great extent about our origin as human species and the way in which

genetic variability affects human phenotypes (Zietkiewicz, *et. al.,* 1997; Underhill, *et. al.,* 2000 and Chen, *et. al.,* 2000). The NRY chromosome and mtDNA do not recombine during meiosis. Because of their uniparental inheritance, the polymorphisims on these molecules are particularly useful for tracing the separate ancestry of maternal and paternal lineages in human populations including Ethiopian. This means that the NRY chromosome and mtDNA remains constant from generation to generation except for rare mutation events. Therefore, Y chromosome and mtDNA are known as lineage markers and hence are important markers for evolutionary studies and assessements of ethnic group variations. The timing out of Africa for anatomically modern human (AMH) estimated by archeological methods (160-2000 KYA) was refined by the study of Y chromosome and mtDNA and was estimated to be 40-60 KYA (Lopez, *et. al.*, 2015). Also, other studies have shown that analysis of human Y chromosomes and mtDNA play a pivotal role in the studying human origin, tracking human migration patterns, estimating the time of migration out of Africa and reconstructing past events in human history (White, *et. al*., 2003 and Trinkaus, 2005 and Lopez, *et. al., 2015*). Nuclear DNA markers together with the Y chromosome and mtDNA markers studies consistently indicate that Africa is the most diverse (genetically, ethnically, linguistically, culturelly, etc) region of the world (Ingman, *et. al*., 2000; Alonso and Amour, 2001 and Tishkoff and Williams, 2002). However, most of the studies have reported the extent and the patterns of genetic diversity only based on small number of populations and markers in diverse African populations, making very difficult to draw general conclusions about the extent and patterns of genetic diversity of these populations. Moreover, studies have shown the presence of extensive genetic variation among even geographically close African populations, which indicate that there is no single representative of African populations

that has been appropriately sampled, and studied (Tishkoff and Williams, 2002), except the recent studies made on several Ethiopian populations (Plaster, 2011).

The studies of human population genetics in Ethiopia and the whole of Africa is also crucial to understand the relationship of humans living outside Africa and the population genetic pattern within the African region. The deep population history, migration history, varying climate and ecological conditions, geological factors, variable culture and traditions of the the Ethiopian popualtuons, together with the pattern of migrations are likely to have contributed to the genetic variability. These factors jointly have created more genetic and/or phenotypic variation in Ethiopia than in any other geographically closer indigenous populations of similar size (Cruciani, *et. al*., 2002 and Pagani, *et. al.*, 2012).

Previous studies of classical genetic markers and Y chromosomal haplogroup distributions have shown that the Ethiopian gene pool of few studied groups posseess a considerable component indicative of admixture with sub-Saharan African populations of Arabian and/or Near Eastern origin (Cavalli-Sforza, 1994; Cruciani, *et. al.*, 2004; Luis, *et. al.*, 2004; Passarino, *et. al*., 1998 and Thomas, *et. al.*, 2000). A similar pattern of *mt*DNA variation was observed in other few studiedgroups from Ethiopian populations (Kivisilid, *et. al*., 2004 and Passarino, *et. al*., 1998) where both Sub-Saharan (L) and Eurasian mtDNA (M and N) haplogroups are observed. It was found that around one-half of Ethiopian (52.2%) mtDNA lineages belong to clades specific to sub-Saharan Africa, whereas the other half was divided between derived subclades of haplogroups M and N that are common outside Africa (Kivisilid, *et. al.*, 2004). These finding gave an intermediate position for the studied Ethiopian population, consistent with early admixture between sub-Saharan and Caucasian populations. This study has also indicated that

both Semitic and Cushitic speaking populations of Ethiopia were close to each other and did not reveal significant differences by comparing samples of Amhara and Oromo peoples. However, several of these previous studies are limited in a number of respects. They are mainly biased in terms of sampling appropriate and representative groups for genetic studies of Ethiopian populations. They included samples mainly from Semetic and to the lesser extenet from the Cushetic populations without giving attention to Omotic and Nilo-Saharan speaking populations.

In other study conducted by Hassan, *et. al.* (2008), the Y chromosome analysis have shown that the Sudanese populations are categorized into eight haplogroups namely A, B, E, F, I, J, K, and R. Where as haplogroupA and B occurs mainly in Nilosharan speaking groups including Nilotics, Fur, Borgu and Masalit, haplogroups F, I, J, K and R are more frequent among Afro-Asiatic speaking groups. It was indicated that Afro-Asiatic speaking groups appear to have sustained high gene flow from Nilo-Saharan speaking groups. More recently, Plaster (2011) genotyped 5756 Ethiopians for their NRY chromosome and identified 526 NRY haplotypes that were distributed over 8 haplogroups of which 232(44.1%) were singletones. The gene diversity in his studied ethnic gropoups at haplotypes level was comparable to the levels of diversity observed in the linguistically diverse cross river region of Nigeria which was 0.911 to 0.973. On the otherhand the 5756 sequenced for mtDNA HVSI distributed among 1328 of which 723(54.4%) were singletones. The gene diversity of the mtDNA HVSI haplotypes were also comparable to cross river Nigerians that ranged from 0.978 to 1.00. The NRY and mtDNA showed positive correlation. The majority of Ethiopian ethnic groups are loosely clustered.

In general, despite the great potential of studying human genetic variation in Ethiopia, which has crucial importance both regionally and globally, limited but encouraging progress has been made

only recently. Plaster (2011) and very recently Lopez and Hellenthal (2017) have conducted genetic studies with good ethnic coverage for populations in Ethiopia. However, studies involving Ychromosome and mtDNA with appropriate sample size are very limited even today. Hence, this study was designed in order to contribute to the better understaning of the genetic structure of Ethiopians through appropriate sample and so that better understanding of the relationships between Ethiopians and other African and non-African populations are worked out. To this end, sample of some Ethiopians populations for a set of mtDNA and Y chromosome polymorphisms were analyzed. The result obtained was compared with the data from other previous study on Ethiopian population, African and non-African populations.

1.2. Objectives of the Study

General Objective

- To Study the Phenotype and Genetic Diversity of LP, Patterns of mtDNA and Y Chromosome Variations in Selected Communities of Benishangul Gumuz and Gambella Regions of Ethiopia

Specific Objectives

The specific objectives of this study are:

- To determine the frequency of LP phenotypes in the selected populations.
- To examine the correlation of LP phenotype with LP Genotypes in the selected populations
- To examine the genetic affinities of the Ethiopian populations (Nuer, Bertha, Gumuz and Mao), using Y-chromosomes and mtDNA markers

2. LITERATURE REVIEW

2.1. Studies on Ethiopian Human Population Genetics

2.1.1. Peoples of the Current Ethiopia:

As described by Finneran (2007), the past historical records related to Ethiopia intentionally ignored the reality that Ethiopia is a multi ethnic society with diverse populations; that follow different religions, have various distinct traditions and cultures, with different languages and live in diverse climate with diverse livelihood. They were totally misidentified based on a monolithic identity that defines the country in terms of highland, Christian, agriculturalist and Semitic kingdom although what used to be called Semitic is partly at least Cushitic. So, both domestic and foreign written sources and Ethiopian traditions look very difficult to be accepted as genuine historical evidences and be considered as historical framework to describe the current Ethiopian peoples. As described by Jobling, *et.al*. (2004), numerous evidence that support the early existence of human in Ethiopia come from linguistic, Paleontological, archaeological, anthropological and genetic diversity studies and some of these are described as follow.

Paleontological evidence indicated that East Africa's Rift Valley were the main sites where the first humans have originated (Cela-Conde and Ayala, 2007). The Ethiopia's Omo and Hadar are the two sites in the early history of human origin. These are the major key sites where a number of fossil evidences have been found including the recent discovery of *Ardipithecus ramidus* (White, *et. al.*, 2009) and *Australopithecus afarensis* 'Lucy' by Johanson in 1974 to the oldest known recognizably AMH which appeared in the fossil records at Omo Kibish and Awash in Ethiopia around 200KYA (Haile-Selassie, 2001 and White, *et. al.*, 2003). Although, the other

sites in the Great Rift Valley are found in Kenya and Tanzania, evidences from paleontological studies show that Ethiopia is one of the main places where AMH might have originated, around 160-200 KYA and left Africa. Although, other ancient fossils may be found in the future, these represent our best understanding of approximately where and when humans have originated; the idea which is also strongly supported by genetic evidence (Cela-Conde and Ayala, 2007 and Campbel and Tishkoff, 2008). Also, studies from archaeology and linguistic have shown that a very complex interaction of early population led to the complex intermixes in the past starting from about 10 KYA to the very recent period which in turn resulted in the present diversity of ethnic composition of Ethiopian people's (Levine, 1974). These interactions of the ancient Ethiopians with very different populations gave origin to the present day diversity in Ethiopia.

Encyclopaedia Britannica (1964) have reported that around 8,000 B.C., the first early hunting peoples were living in the present southern and southwestern part of Ethiopia. In addition to this, Levine (1974) also reported that Afro-Asiatic languages speaking peoples (proto-Cushites, proto-Omotic, and proto-Semites) were found to live in Ethiopia around 4-5,000 B.C. They proposed that they might have been derived from the Sahara (roughly between the Nile River and Red Sea) or from Arabia, as opposed to the original hypothesis of a local Ethiopian origin of Afro-Asiatic languages speaker. In support to the later, some recent linguistic evidences by Finneran (2007) claim that the Afro-Asiatic language speaker locally emerged in Ethiopia around SouthWestern part of Ethiopia and were intermixed with early known inhabitants.

In either case, it was suggested that these early peoples of Ethiopians had undergone a strong diversification around 2,000 B.C and consequently resulted in different groups of peoples present today: the Omotic speaking groups; the Cushites (northern Cushites (Beja), the central Cushites

(Agaw), and the eastern Cushites (Oromo)), and the Semites (Ethio-Semitic)**.** The distributions of these Afro-Asiatic speaking groups were then subjected to continuous external pressure mainly from Sudanese, Arabian, and Mediterranean peoples. The first external influences come from the neigbour Sudanese peoples. It was suggested that Sudanese Nilotic speakers came to Ethiopian and settled in the South Western part of the country and partially intermixed with Omotic and Eastern Cushitic speakers at around 3,000 B.C. and 1,000 B.C. This was followed by migrations from the Arabian Peninsula into Ethiopia which brought about the introduction of Sabean cultures by Southern Arabians into Ethiopia. Further Semitic migrations continued and influenced the Ethiopian peoples from various perispectives. Finally, Mediterranean cultures introduced into Ethiopia, but they did not affect its ethnic composition (Levine, 1974; Passarino, *et.al.,* 1998 and Finneran, 2007).

2.1.2. Genetic Variation among Ethiopian Populations

The study of human genetic diversity has an evolutionary significance among geographically, linguistically, ethnically and culturally diverse Ethiopian populations. Studies of Ethiopian human genetics can help scientists understand about population history as well as how human groups are biologically related to one another. Earliest study of Ethiopian genetics using classical markers revealed that some of the Ethiopian population occupies an intermediate position between West African and North African in PCA may be due to population admixture between sub-Saharan Africans and Eurasia (Cavalli-Sforza, *et. al.,* 1994). In agreement with this, result from structural analysis of autosomal microsatellites indicated the intermediate position of some Ethiopian population between sub-Saharan African and Eurasia (Wilson, *et. al.,* 2001). More recently, because much of the key fossil evidence for human origins and evolution are

found in present day Ethiopia and also as it occupies key geographical position (between Africa and Eurasia), further extensive and in depth human genetic diversity studies was recommended (Pagani, *et. al.*, 2012).

However, until recently, despite the importance of Ethiopian to human genetic studies, Ethiopian populations have not been well represented; for instance, they are absent from widely used collections, such as the Human Genome Diversity Project (HGDP) and international HapMap consortium (Cann, *et. al*., 2002 and Altshuler, *et. al*., 2010). Also, for so long in Africa, practical studies of patterns of genetic diversity have been limited to populations from central and western Africa. This deficiency has led to an incomplete picture of African genetic diversity in general and in Ethiopia in particular, although it has implications for the study of our origins (where and when humans have originated) as a species, including the route followed during the dispersal(s) out of Africa and more recent demographic events involving East Africa. Few recent genetic diversity studies that have included or focused on Ethiopians are highlighted as follows.

Ethiopian and its two neighbors (south Sudanese and Somali) populations were genotyped on an Illumina Omni1M chip in the study conducted by Pagani, *et. al.* (2012). These genotypes were compared with published data from several African and non-African populations. Their analyses confirmed the existence of high within and between populations' genetic diversity in East Africa and observed strong association between genetic diversity and linguistic stratification. With in East Africa, this study confirmed Ethiopia as one of the most diverse African regions.

In addition, Browning (2008), studied human genetic Variation of the gene Cytochrome P450 1A2 (CYP1A2), that is known to be associated with differential efficacy of therapeutic drugs and adverse drug reactions, with implications for healthcares in Ethiopia. CYP1A2 gene and its

flanking region of 762 chromosomes of Ethiopians were re sequenced from members of five ethnic groups (Afar, Amhara, Anuak, Maale and Oromo). The result indicated that there was a high genetic diversity in Ethiopian, with more CYP1A2 diversity than other populations characterized to date and, in some respects, the rest of the world combined. It indicated that much of the variation found on a global scale was observed, supporting the proposition that AMHs migrated out of Africa from Ethiopia.

Also, in Ethiopia, other genetic studies on lactase persistence (LP) phenotypes showed high genetic diversity in LCT gene regulatory region. More recently, Oljira (2014) analyzed the LCT enhancer sequence in LP individuals from diverse ethnic group in Ethiopian and confirmed the existence of multiple mutations responsible for LP trait. His study reported a very high genetic diversity of the LCT enhancer region in LP groups than LNP and increased the number of major identified mutations correlated with LP into five.

Moreover, several numbers of population genetics studies utilizing samples collected from Ethiopian populations have analyzed sex specific inherited markers on the paternally inherited NRY and maternally inherited mtDNA. Their lack recombination, uniparental inheritance and much smaller effective population size as opposed the autosome markers are the major advantages of using NRY and mtDNA markers (Underhill and Kivisilid, 2007). The lack of recombination, as described elsewhere, permits to construct phylogenies easily, allowing for the easier identification of geographically structured haplogroups which could be indicative of past historical migration. Their uniparental inheritances showed distinct patterns of inheritance and as marker systems, they are unique indicator of gender in modifying the extant population structure. The much smaller effective population size of these markers (which is attributed to their haploid

nature of inheritance through one sex only) means that they are susceptible to genetic drift and therefore more sensitive to changes in the demography resulting in clearly visible genetic structure (Jobiling, *et. al.*, 2004).

For the first time, studies of both NRY and mtDNA variation in Ethiopian populations was conducted by Passarino, *et. al.* (1998). This study utilized 77 Ethiopians samples and analyzed for mtDNA and Y chromosome specific variations. The mtDNA was examined for the RFLPs using common enzymes (*Hpa*I, *BamHI, HaeII, MspI, AvaII, and HinCII*) and identified the African haplogroup L and the Caucasoid haplogroups I and T. For the same samples, the author employed other restriction enzymes to investigate mtDNA variation and show other Caucasoid haplogroups (H, U, V, W, X, J, and K) and detected simultaneous presence of the *Dde*I10394 and *Alu*I10397 sites, which defines the Asian haplogroup M, in Ethiopian samples. The Y chromosome was examined for four polymorphic systems namely: the *Taq*I/12f2 and the 49a, f RFLPs, the Y *Alu* polymorphic element (DYS287), and the sY81-A/G (DYS271) polymorphism. The results revealed that the Ethiopian populations; have experienced Caucasoid gene flow mainly through male lineages. It also indicated that the Ethiopian populations have contained an African genetic component attributed to Bantu migrations and to an *in situ* differentiation process from an ancestral African gene pool, showing some affinities with very ancient African group such as Tsumkwe San. All these evidences were suggested by Y chromosome analysis.

In the same study, mtDNA analysis which described above, showed high frequency of the "Asian" *Dde*I10394 and *Alu*I10397 mtDNA haplotype in Ethiopia, but did not reveal a particular relationship between Ethiopians and the Khoisan. Furthermore, Semino, *et. al.* (2002) undertook survey of NRY haplogroups in Ethiopian population. In addition to this servey, this author had

utilized data from an earlier study of NRY haplogroup present in global populations by (Underhill, *et. al.*, 2000) for comparison. This author was investigated the genetic structure 126 Ethiopian (78 Oromo and 48 Amhara) and 139 Senegalese for the deepest diagnostic markers of the major haplogroups of the YRY genealogy (Underhill, *et. al.*, 2001). The results were compared with Ethiopians and Khoisan samples from the Underhill, *et. al*. (2000). One of the results provided by this study was the evidence that the Ethiopians share ancestral paternity with the Khoisan, the deepest human Y chromosome clades (Group I and II); haplogroup A and B according to the Karafet, *et. al*. (2008). Ingeneral these earlier Y chromosome studies have confirmed the presence of the ancestral affinity between the Ethiopians and the Khoisan, which has previously been suggested by both archaeological and genetic findings (Seilested, *et. al.*, 1998). However, these clades were not observed in the Senegalese samples included for the sake of comparisions. Instead, very large number of Group III (haplogroup E according to Karafet, *et. al*. (2008) was observed in Senegalese samples than in Ethiopian samples of which over 80% of the Senegalese and none of the Ethiopian samples had the derived state for the sY81 marker that define the clade E1b1a according to the nomenclature of Karafet, *et. al.* (2008). Moreover, in the frequencies of derived state of M35 marker (defining E1b1b1 according to Karafet, *et. al.,* 2008) and p12f2 marker, significant differences was observed between Amhara and Oromo samples.

Next, Semino, *et. al*. (2004) have investigated the distribution of the NRY haplogroups sub clades E and J using the same Amhara and Oromo samples that was used in Semino, *et. al*. (2002) combined with the set of global population samples from previous work. The authors showed that where as the majority of the Oromo than Amhara belong to the sub clade E, of which the highest frequency subclade for both groups was E1b1b1a (that define the derived state of M78 according to Karafet, *et. al*. (2008)), the frequency of the haplogroup J (as defined by

derived state of p12f2 marker) was found by far at higher frequency in Amhara than in the Oromo, in agreement with previous observation (Semino, *et. al*., 2002). In other study, samples from Wolayta (12), Amhara (34) Oromo (25) and mixed Ethiopians (12) were investigated for terminal haplogroup present in E clade (Cruciani, *et. al*., 2004). From this study, they observed that all the typed haplogroup E sub clades were present in all genotyped groups. This same Ethiopian samples (but 7 Borana samples from Kenyan combined to the Ethiopian Oromo samples) were later analyzed for further marker in E1b1b1a (which is M78 derived terminal marker). It was observed that the E1b1b1a was present in all genotyped group but with variable frequencies. Detail for this is available in Cruciani, *et.al*. (2007).

In other study, Pagani, *et. al.* (2012) has genotyped Ethiopian and the two neighbors (South Sudan and Somalia) populations on an Illumina Omni1M chip and compared the resulted genotypes with published data from several African and non-African populations. Their analyses confirmed the existence of high within and between populations' genetic diversity in East Africa and revealed strong association between genetic diversity and the linguistic stratification. Interestingly, with in East Africa, these all expressed and other comprehensive study conducted recently by Plaster (2011) has confirmed Ethiopia as one of the most diverse African regions. With regard to the mtDNA, several comprehensive studies on the variation in mtDNA haplogroups in Ethiopian populations were conducted (Kivilisid, *et. al.,* 2004; Tishkoff, *et. al.,* 2007 and Poloni, *et. al*., 2009). All these studies were focused on the control region of mtDNA genome and showed higher diversity in Ethiopian populations and the details are described elsewhere in this thesis. In agreement with this, although the current study was focused on the coding region of the mtDNA (*MT-COXII*), the result has shown highe diversity in the studied population in Ethiopia.

2.1.3. Genetics and the Origins of Modern Human

Debate with respect to the modern human evolution largely centered on the most probable location of origin and the age of human species, and the amount of genetic contribution of the ancient humans to the modern human (Cavalli-Sforza, *et. al*., 1994; Relethford, 1995 and Ruvolo, 1993). Numerous paleontological evidences have now established East Africa in general and Ethiopia in particular as a strategic region where hominization gradually took place, with *H.erectus* representing the first hominid possibly migrating out of Africa continents (Lopez, *et. al*., 2015). With regard to the origin and dissemination of AMH, the widely accepted model is the 'Out of Africa' (OOA) model of modern human evolution and migration, according to which AMHs originated in Africa around 160-200 KYA and distributed to the rest of the world during the last 100KY through either Southern or Northern tips of Red Sea. Whereas, the Southern routes of migration possibly through the Bab-el-Mandeb strait at the Southern end of the Red Sea, the Northern route is through the Levantine corridor (Stringer, 1994; Cavalli-Sforza *et. al.*, 1993 and Pagani *et. al.*, 2012).

It is also generally believed that AMH, which are the main ancestors of all humans alive today, arose in one place in Africa, most likely in Ethiopia from where they are dispersed elsewhere out of Africa to Eurasia and the rest of the world at some point between 60 - 120 KYA (Lopez, *et. al.*, 2015). It has been suggested that these first migrant from Africa most likely left through present day Ethiopia (Figure 1), in favor of the exodus of Out of African model of human origin. This suggestion was supported by the discovery of earliest known AMH human fossils including *Ardipithecus ramidus* (White, *et. al.*, 2009), *Australopithecus afarensis* 'Lucy' (Johanson, 1974), Omo I from Kibish and Awash (Haile-Selassie, 2001and White, *et. al*., 2003) and *Hertho*

remains (Clark, *et. al.,* 2003). These coupled with its deep history make Ethiopia the key country to studying human genetics and learn about the origin of our species. The deep history of Ethiopia would be attributed to the greatest level of ethnic, linguistic, cultural and tradition diversity in the world. For instance, Ethiopia is the homeland for over eighty different ethnic groups and living languages.

For several decades, the archeological and linguistic evidences are strongly supported with Genetic data to address questions of human evolution and migration patterns. With this respect classical markers study (Cavalli-Sforza, *et. al.,* 1994, 1998), mtDNA analysis (Cann, *et. al.,* 1987; Ayala, 1995 and Relethford, 1995), Y-chromosome analysis (Thomson, *et. al.,* 2000), variation studies of nuclear DNA using RFLPs (Bowcock, *et. al.,* 1991) and analysis of microsatellites (Jorde, *et. al.,* 1997) have been employed. Of these, the first genetic evidence consistent with the Out of African model was provided by the study of mtDNA phylogenetic trees, which identified Africa as the source of human mtDNA gene pool. This was also supported by other studies of mtDNA (Relethford, 2001) and Y chromosome (Thomson, *et. al.*, 2000 and Underhill, *et. al*., 2000). The resulting data have been summarized and provided a broad picture of the human evolution across the world from genetics perspectives. Their results indicated small level of inter population genetic divergence between continental human populations, greatest genetic distance between African and non- African populations (indicating initial population split happened between them) and highest level of genetic diversity in African populations (showing they represent the oldest human populations) relative to populations from other continents. All these observations are in accord with the model of a recent African origin for modern humans and that suggested the existence of a common African ancestor.

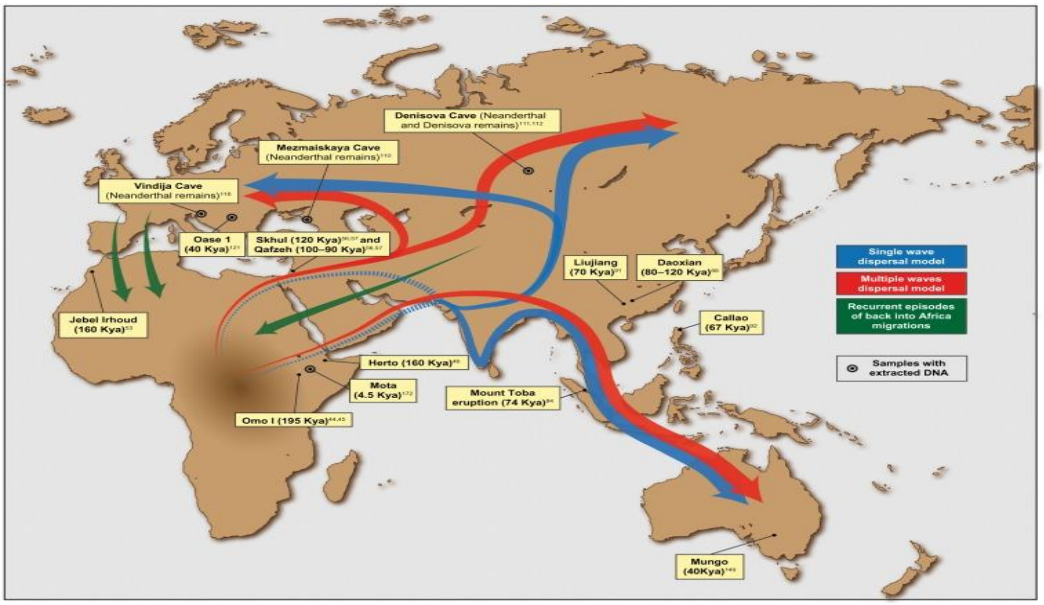

Figure1.1. Putative migration waves out of Africa and location of some of the most relevant ancient human remains and archeological sites. The placement of arrows is indicative (Taken from Lopez, *et. al.*, 2015).

2.1.4. The Short History of Presently Studied Ethiopian Ethnic Groups

Past historical account and culture of the various ethnic groups and their relationship with each other could influence their present day genetic structure, identity, distances and similarities within and between each group. It is thus important to sketch historical highlight of each as portrayed in the section below.

A. The Nuer

The non-cited information mentioned within here are obtained through personal communications from Nuer elders. The Nuer people are Nilotic ethnic group known with their pastoral traditions living in Ethiopia and Sudan and primarily inhabit the Nile Valley. They speak the Nuer

language, which belongs to the Nilo-Saharan family. In Ethiopia, they are one of the five tribes (Anyuak, Majanger, Oppo, Mao and Komo) living in Gambella region situated in the western part of Ethiopia. They call themselves 'Naath' meaning "human beings". The Nuer people are pastoralists who herd cattle for a living. Cattle occupy a central position and it is everything in Nuer lifestyle. They have historically considered their cattle to have highest religious, symbolic, political and economic value. Cattle also are important as they are used as bride wealth. That is they are given to the bride's family by the bridegroom's family. In addition to this, their role in sacrifice is something important also. The Nuer people give more focus to the quantity and quality of the cattle owned. They take the name of their favorite oxen or cows in ritual honor. Moreover, they are greeted by their cattle's names or colors.

Although large numbers of Nuer are follower of Christian religion, the majority of the Nuer has a traditional belief. They worship a creator which they called "kuoth''. They believe that all life comes from and returns back to the creator. They pray and offer sacrifice of cattle to "kuoth" at a designated time of the year such as the beginning of the raining season and the end of the year, hoping for well-being and health. They do this also as part of ceremonies like marriage. In the spiritual beliefs of Nuer culture, cattle play an important role. For instance, they make contact with their ancestor spirit by rubbing ashes along the backs of oxen or cows dedicated to them. They worship one God whom they call 'kuoth' and they say he is the high god and source of life. According to Evans (1956) "Nuer says that God is everywhere, that he is 'like wind' and 'like air". This is related to the traditional beliefs of Nuer people that God is in the air.

The traditional ruling system of Nuer is based on their traditional norms and societal system. The Nuer community is ruled by members who are selected based on their good habit in the society.

Their traditional culture allows them to share leadership among different clans or different sub-clan in a village. But the leaders are chosen after they show leadership qualities and when other people respect them. When conflict happens, elders are concerned bodies to resolve the conflict. The Nuer traditional administration system of villages is mainly in the hands of elders of the dominant clan. In fact, personal qualities including lineage, age, seniority in family, wealth in cattle, large number of wives and children, skill in debate and some spiritual powers are taken in to account to create a social personality for being considered as a leader. They also select the war general that can initiate and stop the war based on the above qualities.

B. The Bertha, Gumuz, Shinasha and Mao

As in for Nuer, the non-cited information mentioned within here are obtained through personal communications from elders of each ethnic group. Bertha, Gumuz, Shinasha, Mao and Komo are ethnic groups living in Benishangul Gumuz, the vast region (50,000 sq. km) in the Western border of Ethiopia. They regard themselves as 'indigenous' peoples, as owner nationalities of a region. In fact, in addition to these ethnic groups, the region contains a large 'non-indigenous' population which accounts for more than two third of the total population. The majority of these are the Semitic (Amhara and Tigrei) and Cushitic (Oromo and Agaw) ethnic groups. These show that Benishangul- Gumuz consists of all kinds of cultures such as Nilo-Saharan (mainly Bertha and Gumuz), Semitic, Cushitic and Omotic Shinashaa and Mao. The Nilotic culture in the region was once called "pre-Nilotes", by their so-called archaic cultural features as compared to other Nilotic groups (Grottanelli, 1948). According to 2007 Ethiopian census, the total population of Benishangul Gumuz region is 670,000 peoples. Like other border regions, Benishangul Gumuz is a fluid region, meaning that peoples living in the region experienced both peaceful and violent

relationships and interaction with different external groups and large cultural traditions. In addition, 'indigenous' peoples of Benishangul Gumuz was characterized by multi-dimensional inter ethnic conflict with their minor groups and with other neighbors Ethiopian ethnic groups before the establishment of present day Federal government of Ethiopia although conflict of various persuasions still exist.

The Berthas are living along the Ethiopian-Sudan border from Metemma in the North to the Dabus River in the South. Also, known as Bertha, Barta and Benishangul (Arabic designation), they are the predominant ethnic groups in the region. Although it is very difficult to obtain information about the exact origins of the Bertha from the literature, the Bertha elders claim to have originated from Eastern Sudan Arabs, the area of the former Funj sultanate of Sinnar (i.e., where they claim originally lived and participated in state formation) and may have migrated to Western Ethiopia, the modern Benishangul Gumuz Region. To convince that they have originated from Arab, they call themselves 'Mayu' or 'Bani Ummaya', from the alleged descent line claimed by the Funj (Triulzi, 1981) and use the term Bertha as a derogatory label for the rest to whom they also refer as "Blacks". It was suggested that migration of the Bertha from Sudan lowlands to the Ethiopian might have been triggered by difficult climate condition and other security issues. According to the 2007 Census, their total population in Ethiopian is about 183,000. They speak bertha language, as mother tongue, that belongs to the Nilo-Saharan language family that is not related to those of their Nilo-Saharan neighbors. The Bertha language consist a number of dialects (i.e., mutually understandable) spoken by the Bertha, the Jebelawin, the Watawit and others. In addition, due to of several centuries of Arab Sudanese influence, most of the Berthas are now follower of Islamic religion and many of them can speak fluent Arabic. Although Berthas are mostly Muslim, many of them have still maintained traditional religious

that are very similar to their Nilo-Sharan neighbors. They have ritual specialists named *Neri*, who have the powers of healing and divination. They believe that *Neri* know everything, for instance, how to deal with evil spirits (*shuman*). Also, like other Nilo-Saharan and Nilotic communities, Berthas have rain-making rituals.

The survival of Berthas was challenged by strong territorial conflict with among the diverse neighbor communities of the region. They were severely challenged by dominant powers from Sudanese Arabs and Ethiopian highlanders. The Bertha ethnic identity was largely shaped by their relationship they have maintained with these dominant powers and other minority ethnic groups living with them. This can be exemplified by their historic relationships with the Sudanese Arabs that defined their identity and internal hierarchy. With this regard, they have still maintained belief that ruling families of the Bertha should come from the half-cast called Watawit. Watawit are class of people of mixed ethnicity or race who were descendants of intermarriage between the ruling families of the 'pure-Bertha' and Arabic Islamic preachers who came to the region from the Sudan (Bahru, 1991 and Triulzi, 1981). As described by Abdussamad (1999), this class division between the half-caste Arabs and the pure- Bertha had created conducive conditions for Watawit to exploit the pure-Bertha and other minority ethnic group (like Gumuz and Mao) living together with them. They together formed the bulk of the 19th century slave trade in the region in which the Watawit participated as slave owners and transferor of slave to other areas. For instance, they used pure-Bertha as slave labour to extract alluvial gold. Even earlier, According to Triulzi (1981), the class division had shaped the broader ethnic identity of the Bertha which latter paved the way for the Ethiopian and Sudanese powers to use Watawit to control the area for region's resources exploitation (e.g., gold and slaves). Even today, some of the Watawit dislike the naming of their ethnic group as Bertha as that name

echoes abide and was used to call the pure-Bertha by both the Sudanese Arabs and the Watawit themselves (Triulzi 1981). In contrast, many educated descendants of the Watawit who today control important offices within the Benishangul Gumuz region on behalf of the Bertha support the designation of their ethnic group as Bertha and seek to capitalize on the victimization of the pure Bertha in the past as an important platform to advance their political careers. In this context, Bertha politicians promote a new Bertha identity that includes both the half-caste and the pure. Further and detail of useful information on Bertha can be found in Cerulli (1956) and Triulzi (1981).

Like Bertha, the Gumuz are the predominant ethnic group inhabiting the Benishangul Gumuz of western Ethiopia and as well they live in Sudan. They speak the Gumuz language, which belongs to the Nilo-Saharan family. The Gumuz language has many dialects than Bertha, but those who speak them can understand the others. According to the 2007 Census, their total population size was around 160,000 and was forcasted to be about 200,000 in 2017 (Lewis, 2009). Some Gumuz are Muslim, and few of them follow Christianity (either Orthodox or evangelical). The Muslim groups are living near the Sudanese borderland and Christians live in South and East, bordering Etiopian highlanders. Thus, the adoption of Islam and Christianity is the result of contact with neighboring Sudanese and Ethiopian highland neighbors. Interestingly, most Gumuz still maintained traditional religious (Wolde-Selassie Abbute, 2005) and even the beliefs of Gumuz Muslim and Christians are deeply rooted in their traditional practices. They are Naturalistic, which means they worship the "spirits of God" (the Spirits are called '*mus'a'*) who created certain natural things like rocks, trees, and animals, etc for their good health and everything. The supreme God of the Gumuz who knows all is known as 'Rebba'. They have ritual specialists

called *gafea*, which parallels priest of Orthodox Christians. Their livelihood depends on agro-pastoralist and shifting cultivation.

Similarly, as in the case of Berta, the ethnic history and identity of the Gumuz was mainly shaped and characterized by painful relationships with more dominant neighbors. Slave raids and displacement from their land to hot lowland areas characterize the history of interaction between the Gumuz and their dominant neighbours. James (1986) wrote that the Gumuz was subjected to continued pressure until the first few decades of the 20th century from both highland Ethiopia Christian and the Sudan Islam. Moreover, around the beginning of 20th century, the Oromo occupied the Metekel area by displacing the Gumuz to lowland areas, where the climate is very difficult to live in (Abdussamad, 1995). The continued pressure placed on the Gumuz from several sides (mainly as subjects of slaving raids) shaped their settlement pattern and left their indestructible mark on their identity and the way they relate to each other and their ethnic neighbours. The acculturation of the people to their country coupled with the inhospitality of the area because of its arid climate and malaria infestation, the Gumuz maintained their survival against the intense pressures of Sudanese and Ethiopia traditions. The violent relationships that maintained between the Gumuz and dominant powerful neighbors and their collective experience led to the development of a common ethnic identity, the Gumuz.

The Shinasha are the third largest 'indigenous' ethnic group living in Benishangul Gumuz region. According to oral saying of the society, the Shinasha are the only group that have originated in Ethiopia and is one of the ancient groups of the region. They also believe that there are few migrated Arabic speaking Shinasha groups living across the border in Sudan. Shinasha peoples have their own language, which they called Shinasha /Boro that is quite different from

their neighbors and belong to the Omotic language family. According to a 2007 Ethiopian census, 37,500 speakers of the Shinasha language were identified. Most of the Shinasha are Orthodox Christians and are rarely Muslims.

Like other groups, Shinasha also experienced a painful relationship with its more powerful neighbours. Like the Gumuz, they faced slave raiding by the highlanders of the Ethiopian empire. Because of raiding and displacement, encountered by both the Gumuz and Shinasha, they forced to escape together to the hot inhospitable lowlands of the Metekel region. Both groups shared the challenge and lowland areas and started cooperation to live together. Then, as Tsega (2002) mentioned, at the later, Gumuz and shinasha started inter-competition for natural resources in the area which in turn had initiated conflict between the two groups. Meanwhile Shinasha was facing fierce and continued pressures from the Agaw and Amhara neighbours. These pave the way for the Shinashs to have assimilated into the Oromo during the 18th century. It was reported that because of the insecurity encountered, unassimilated Shinasha had established their villages in unreachable area to protect themselve from violent invader in the late 19th century. More recently, the relationship between the Shinasha and their ethnic neighbors kept on changing depending on the conditions on the ground. According to Gebre (2004), the Shinasha, the Gumuz and the Agaw showed cooperation and had established strong relationships to oppose the settlement of several thousand people in their fertile lowlands of Metekel during the 1980s by the military regime, thinking that the settlement may affect their livelihood. Similarly, the ethnic regionalization of Ethiopia since 1991 has led to the emergence of a new alignment in the Metekel region between the Gumuz and the Shinasha excluding the non-indigenous groups.

Mao is small population group of Nilo-Saharan societies. In contrast to the other constituent ethnic groups of the Benishangul- Gumuz region, the Mao is a little studied group. Thus, less is known about the Mao people and their language. Mao is an external denomination historically used by the expanding Oromo to identify those groups they entered into contact with and with whom they maintained relations, mainly trade (James, 1980 and Negaso Gidada, 2001). This group is located along the Ethio-Sudanese border (Cerulli, 1956). They are also found within the Oromia and Benishangul Gumuz regions of Ethiopia. Generally, they speak the Mao language, a branch of Omotic language, which belongs to the Afro-Asiatic language family spoken in Ethiopia. However, the classification of the Mao into one of the language families of Ethiopia remains incomplete.

Like the others, the Mao had difficult relationships with their powerful neighbours. Indeed, the pressure on the Mao became more intense after Sheik Khojele of Assosa peacefully submitted to Menelik and became a loyal vassal at the end of the 19th century. His autonomy was contingent upon making regular payments of tribute in gold and slaves to the crown in Addis Ababa. This required him to apply harsh tactics on the Mao and other subordinate groups mainly Komo in order to raise revenue and slaves (Abdussamad, 1999). This forced the Mao and their neighbor groups further to the fringes of the periphery (Cerulli, 1956 and Johnson, 1986). Although powerless in the face of their dominant neighbors, the Komo played one predator against the other. They also exploited the frontier aspect of their location to escape the heavy exaction that their neighbors put on them by alternating their residence between the Sudan and Ethiopia (Johnson, 1986). The Mao has a long history of interaction with Komo because of their geographic proximity and shared difficulties. Thus, they have exchanged many cultural values between each other. For instance, Mao social organizations were influenced because of their

absorption of the Komo (Cerulli, 1956). Cultural exchange between the Mao and Komo gained new momentum after the establishment of the Mao-Komo special district and the recognition of both groups as owner ethnic groups of the Benishangul - Gumuz region. Hence, the formation of a local government organization, a political party and a development association under the newly invented identity of Mao-Komo could accelerate the amalgamation of the two ethnic groups. In fact, both groups may find it expedient to use this invented Mao-Komo identity in order to defend their interests within the Benishangul Gumuz region (González-Ruibal and Fernádez, 2005).

2.1.5. Correlation between Linguistics and Genetics

Human genetic and linguistic diversity have been proposed to be generally correlated among globally distributed human populations, either through a direct link, whereby linguistic and genetic affiliations reflect the same past population processes, or an indirect one, where the evolution of the two types of diversity is independent but conditioned by the same geographical factors (Chen, *et. al*., 1995 and Cavalli-Sforza, *et. al*., 1988).

Cavalli-Sforza, *et. al*. (1988) showed clear correlation between genetic distances and a tree of the world's language families in human populations. They proposed that this correlation showed considerable parallelism between genetic and linguistic evolution. Other studies have found variable degrees of correlation between linguistic and genetic variation at the sub continental scale (Excoffier, *et. al.*, 1991), depending on the region, the type of diversity measured, and the linguistic classification. Y-chromosome studies have shown the correlation between genetics and language in Siberia population (Karafet, *et. al.*, 2002), in Europe (Rosser, *et. al.*, 2000), Australia (Hurles, *et. al.*, 2002) and the Americas (Zegura, *et. al.*, 2004).

In addition, the pattern of populations' genetic affinities inferred from the Y chromosome specific markers studies suggested genetic structure are generally correlated to the linguistic relatedness of different populations, from Europe, Middle East, Africa and Asia (Quintana-Murci, *et. al.,* 1999). Moreover, inferences of human population structure made from the Y chromosome by Poloni, *et. al.* (1997) with the help of multivariate analysis showed that genetic distances between world populations correlate generally to language families in agreement with autosomal and mitochondrial data. For instance, according to the Fagundes, *et. al*.(2002), mtDNA studies showed that languages are better correlated with genetic affinities among South American populations, whereas, both geography and language are associated with mtDNA variation in Austronesia (Lurn, *et. al.,* 1998).

In Africa, the Y chromosome studies have revealed a strong partial correlation between genetic and linguistic distances but showed no correlation between genetic and geographic distances. In contrast, mtDNA variation is weakly correlated with both language and geography. In the study conducted by Hirbo (2011), using mtDNA and Y chromosome DNA, significant relationships between genetic and linguistic distances among East African populations was reported. Also, the author reported that the correlations between genetic and linguistic variation is stronger for autosomal and Y chromosome than for mtDNA lineages. Y chromosome and mtDNA lineage distributions seem to cluster linguistically. These data suggested that patterns of differentiation and gene flow in Africa have differed for males and females in the recent evolutionary history in agreement with the report of Wood, *et. al*. (2005).

2.2. Lactase Persistence in Human

2.2.1. The LP and LCT Enhancer (MCM6) gene

As was presented earlier in humans, LP individuals maintain lactase enzyme production through their whole life as a result of a mutation in the LCT enhancer gene (MCM6) enabling them to consume higher amounts of milk, and thereby higher amounts of lactose, without symptoms (Wang, *et. al.,* 1995). The LCT gene is located on chromosome 2q21 and roughly 50 Kb in size (Kruse, *et. al.,* 1988), containing 17 exons that can be translated into a 6kb transcript. Also, the MCM6 gene has been scanned by fluorescence *in situ* hybridization in chromosome two directly before LCT gene and have shown to be expressed in a variety of adult and fetal human tissues, having an important role in the regulation of cell cycle (i.e. DNA replication) as part of a complex associated with DNA helicase (Harvey, *et. al.,* 1996 and You, *et. al.,* 1999). Studies have revealed that the cause of LP does not appear to come from within the LCT gene itself, but rather within nearby MCM6 gene located upstream of the LCT gene transcription initiation site. Many of the key genetic mutations associated with LP are in this neighboring MCM6 gene which covers 36 kb in size and contain at least 43 SNP's and 9 insertion/deletions (Enattah, *et. al*, 2002). By investigating the intron 13 of MCM6, Enattah, *et. al.* (2002) found single SNP that allow LCT gene expression and maintain lactase activity in adults. Several further studies have shown that DNA fragments in intron 13 and 9 of the MCM6 gene have shown to have similar effects for LP traits (Ingram, *et. al.,* 2007; Enattah, *et. al.,* 2008 and Tishkoff, *et. al*., 2007). To date at least five major LP associated SNPs (T/C 13910, T/G 13915, C/G-13907, T/G 14009 and G/C-14010) have been detected. These SNPs have perfect to strong association with LP traits in different populations of the world. Functional studies have demonstrated that sequence

containing the mutant alleles in the introns of MCM6 acts as a stronger enhancer for LP gene expression *in vitro* than the normal variants (Troelsen, *et. al*., 2003; Jensen, *et. al.*, 2011; Olds, *et. al.*, 2011 and Jones, *et. al*., 2013). The allelic change initiates the binding of the transcription factor *OCT-1*, thereby enhancing LCT promoter activity (Lewinsky, *et. al*., 2005). The LP or LNP is, therefore, associated with a SNPs variation in MCM6 gene situated upstream of the LCT gene, in a region that appears to act as a *cis* element capable of enhancing differential transcriptional activation of the lactase promoter region (Olds and Sibley, 2003).

2.2.2. The Appearance LP and Culture of Dairy Farming

The enzyme, lactase, enable human to break down and utilize lactose. Post weaning onwards, its production declines progressively as the age increases in the majority of human populations. Whereas some humans can maintain the continuous production of lactase throughout adult life and are thus able to digest the lactose found in fresh milk without problem, the others cannot do so. To date, the co-appearance of LP (ability to consume lactose) and cultures of dairy farming is an established knowledge and is the text book example of gene culture co-evolution. Investigators have observed that the majorities of populations with high frequency of LP experienced similar history of dairying culture (Simoons, 1970 and Durham, 1991). From this, they have suggested that dairying induce the selection pressures that made alleles for LP to be high in frequencies. In agreement with this, strong correlation was observed between the frequency of LP and a history of dairy farming and milk drinking (Durham, 1991 and Myles, *et.al.,* 2005). Similar other studies now support this idea saying that dairying permitted the LP allele to have appeared and spreaded across the world indicating LP is a good example of niche construction (Gerbault, *et. al.,* 2011). Estimation of the age of LP allele using intra-allelic

microsatellite variations indicated that the LP is thought to have appeared in Eurasia before the Neolithic and after the emergency of modern human outside Africa due to mutations introduced in the MCM6 gene following a fully developed culture of dairying-based farming (Coelho, *et. al.,* 2005). Thus, the appearance of LP as a result of this mutation is thought to results from gene-culture co-evolution as milk is the only natural source of lactose on which the lactase gene product, lactase enzyme works.

Studies by Enattah, et. *al.* (2002) have shown that 13 910C>T mutation is the most prevalent in the MCM6 region and strongly associated with, and probably causative of LP in Europe. In support of this, other studies have reported that this mutation is one of the most prevalent and strongly selected alleles in the human genome in Europeans (Tishkoff, *et. al*., 2007; Bersaglieri, *et. al*., 2004 and Coelho, *et. al*., 2005). They have noticed that most likely natural selection and a supply of fresh milk would have driven the 13910 C>T allele to high frequencies and related the evolutionary appearance of LP to the culture of dairying through a gene-culture co-evolution process. In agreement with these, Swallow (2003) reported that the selection of this alleles and the distribution of the phenotype correlates well with that of populations with a history of cattle domestication and milk drinking. In the Hap Map collections (International HapMap Consortium, 2005), mutations occurred in the MCM6 gene in Europeans shows the strongest signal of positive selection, reflecting a powerful advantage that may have been more related to milk as a source of clean water than as a source of nutrition. Enattah, *et. al.* (2002) identified a causative regulatory variant in MCM6 gene in the populations of Europe with an estimated age of 2000-20 000 years (Bersagleiri, *et. al.,* 2004). However, lactase persistent populations elsewhere, including Africa, do not carry this variant. Studies on Kenyans, Tanzanians, Sudanese and Ethiopia (Tishkoff, *et. al*., 2007; Ingram, *et. al*., 2007 and Jones, *et. al.,* 2013) have revealed

five further nearby mutations causing LP. Examinations of the surrounding haplotypes also show that the five African variants arose independently of each other and of the European variant. Thus, different mutation results in similar phenotype, a further example of convergent evolution. Particularly in Africa the situations seem very complex and the known variants still do not account for all of LP; so further examples are likely to exist.

2.2.3. Origins and spreads of pastoralism with LP Trait

The genetic history of the origin of pastoralism is controversial and poorly understood. Pastoralism is found in many variations throughout the world and perhaps might have independently originated in a number of areas. Some of the literatures claim pastoralism was originated in the Levante region and from there spread to the African continent. According to Pagani, *et. al*. (2012) and Pickrell, *et. al*. (2013), the pastoralism was spread into Africa through dispersing farmers (demic diffusion) from a Levantine area and reached East Africa around 3000 YBP, from where it was spread to Southern Africa by around 2000-1200 YBP. These authors reported that the present day African populations have been subjected to back-admixture from Eurasian populations associated with the spread of pastoralism. Estimates of genetic Eurasian back-admixture are strongest in East African populations (Pagani, *et. al*., 2012 and Tishkoff, *et. al*., 2009), where admixture ratios in modern populations are up to 50% Eurasian ancestry proportion. The pastoralism mode of existence, i.e., dairy farming practice especially provided new food sources and has left a strong pressure on human genome. The LP trait, an ability to digest the lactose (main milk sugar) as an adult, has associated with several independently evolved alleles in Africa, at least five of the alleles are found mainly in Ethiopia (Ingram, *et. al*., 2009 and Oljira, *et. al*., 2014), where they have the highest frequencies in pastoral communities.

LP, which is genetically determined phenotypic trait, associated alleles are found to be under severe selective pressure (Bersaglieri, *et. al*., 2004) and can be found in the highest frequencies in dairy farming pastoralist communities (Holden and Mace, 1997 and Gerbault, *et. al.*, 2011). Positive selection is thought to have played a major role in maintaining LP in different human populations that practice pastoralism. Two major observations led to the idea that LP increased in frequency relatively recently as a result of positive natural (Cavalli-Sforza, 1993) as it is advantageous; i.e., increased nutritional benefit from dairy. The first is populations with long history of the culture of dairy farming have high frequency of LP and vice versa (Simoons 1969; Kretchmer, 1971 and MacCraken, 1971). The second is the polymorphism is genetically regulated. Evidence from population genetics analysis of the *LCT* and its surroundings including the enhancer supported this for European populations (Bersaglieri, *et. al*., 2004). Extended haplotype background different from the other derived alleles is supporting recent adaptation and multiple independent mutation causes. The LP variant frequencies differences between Eurasia and African populations would be possibly attributed to a phenomenon of admixture. The LP alleles have been used to track migration events, to identify admixture from East African pastoralists in southern African populations, and as an example of adaptive introgression (Myles, *et. al*., 2005 and Enattah, *et. al*., 2007). Hence, frequencies differences of LP alleles in Ethiopian and other population, for instance, Eurasia populations would also indicate the role of admixture. Hence, to identify variants associated with the LP trait admixture and to study its evolutionary history in Ethiopia, next-generation sequencing seems important.

2.2.4. Distribution of LP Phenotype

Earlier, it was wrongly considered that LP is uniformly distributed worldwide. However, several phenotype and genotype studies showed that the frequency distribution of LP in human populations is not uniform but varies widely depending on a history of dairying farming. The phenotype distribution of LP has been determined directly by intestinal biopsy or indirectly through measuring individual ability to digest lactose sugar. The later include breath hydrogen test (that measures the level of hydrogen produced from the metabolism of lactose by the action of colon bacteria in LNP individuals) and blood glucose level determination test (that measures the increase in blood glucose concentration), both after the oral load of 50g lactose (Metz, *et. al*., 1975 and Arola, 1994). Using the LP phenotype distribution and the genetic analysis involving SNPs genotyping in the LCT-enhancer region, the genotype distribution and prevalence of LP traits have been assessed in different human populations living in different geographic locations. In both cases, the modern frequencies of LP considerably vary throughout Europe and elsewhere in the world according to a geographic locations and population demographic history and are strongly correlated with dairy farming cultures. Moreover, although LP is the most common phenotype in humans and exclusively human trait, it is not universal to all human populations (Romero, *et. al*., 2011). Instead, it is a population specific trait that varies widely in different populations across the world according to ethnicity (Ingram, *et. al.,* 2009). Other study by Midgley (1992) and Swallow (2003) have showed that the frequency of LP varies widely in different populations, with high prevalence in North Western Europeans where milk dependent cattle pastoralism was developed very early. Similar high frequencies were reported in milk dependent African and Asian pastoral populations but low in other Asians and Africans populations. These authors have indicated that the patterns of the occurrence of LP are generally

decreases in prevalence to wards South and East. Similar study have reported that in North West of Indian subcontinent, high frequency of LP was reported than elsewhere, but almost approaching none toward South East (Romero, *et. al.,* 2011). They further showed the patchy distribution of LP in Africa with the existence of moderately high frequency for some ethnic groups that have traditions of using milk than neighboring tribes living in the same country. Another study by Bayoumi, *et. al.* (1982) has reported that there are higher occurrences of LP in both North and South African sub-Sahara pastoralists than non-pastoralists. Consistent with this, Hijazi, *et. al*. (1983) and Dissanyake, *et. al*. (1990) have indicated that LP have been reported to be more frequent in Bedouins populations who have a habit of drinking fresh milk than in neighboring non-Bedouin Arabs. Moreover, studies on East Africans and Middle Eastern groups revealed that the prevalence of LP was reported to be higher. They are greatest in adults of North Western Europe, gradually decreasing towards the South East and, in pastorals coming from African and Asian desert (Flatz, 1989 and Itan, *et. al.,* 2010) as these people have in common a long history of milk drinking. The global distribution of LP is depicted in figure 2.1 below.

Figure 2.1. Global map showing the frequency of the LP trait for populations (Ingram, *et.al*., 2009). LP is shaded in black. Each pie chart represents a unique population. The degree to which the pie chart is filled in illustrates the percentage of individuals in that population who are LP.

In order to better understand the co-appearance of LP and dairying in Europe, Itan, *et. al.* (2009) developed a demic computer simulation model and examines how demographic and evolutionary parameters could have shaped both the modern distribution of LP in Europe and the timing of the arrival of farming at different locations throughout Europe. The study modeled the spread of dairying and non-dairying farmers into Europe that was previously occupied by hunter-gatherers, under the plausible assumption that an LP associated allele would only be selected in dairying farmers. Moreover, Burger and Thomas (2011) were able to identify parameter values that best explained the modern distribution of LP in Europe and the timing of the arrival of farming at different locations throughout Europe. They concluded that LP allele is most frequent in Europe today and overlap with the arrival time of dairy farming at different locations.

2.2.5. Evolution of LP Associated Mutations

The most common proposed gene to have experienced strong recent positive selection is LCT gene, which codes for lactase enzyme. The ability to use this enzyme to digest lactose during adulthood varies widely across worldwide populations, with particularly high capacity among Europeans (Simoons, 1969 and Scrimshaw and Murray, 1988). Moreover, an ability of individuals to maintain the lactase activity into adulthood is determined by their genetics (Simoons, 1970; Scrimshaw and Murray, 1988 and Enattah, *et. al*., 2002). Because of unusual distribution of LP as well as its correlation with milk drinking habits, they have proposed that the actual LP distribution is explained by recent positive selection resulting from increased nutrition from milk consumption, the only dietary source of lactose, in populations that shifted their subsistence patterns to become truly milk dependent (Simoons, 1970; Flatz, 1987; Hollox, *et. al*., 2001; Poulter, *et. al*., 2003 and Holden and Mace, 1997).

The culture and population demographic history appear to be the primary leading force that led to the expansion of LP, and the frequencies of the LP we observe today in Europe and other parts of the world might be the result of recent and strong natural selection. Studies have proposed (Hollox, *et. al*., 2001; Enattah, *et. al.,* 2002 and Poulter, *et. al*., 2003) that a selective advantage based on additional nutrition benefits from dairy explains these genetically determined population differences for which population genetics-based evidence of selection has later been provided (Simoons, 1970; Scrimshaw and Murray, 1988; Enattah, *et. al.*, 2002). Other studies have shown that many non-European populations experience high rates of LP, raising questions about whether a single allele arose once and is shared by all LP individuals or whether different alleles have arisen independently in human history. Further investigation revealed the evidence

supporting that LP has evolved independently, in different part of the world over the last 10,000 years and is associated with different mutations in different regions (Enattah, *et. al.,* 2002; Ingram, *et. al.,* 2007; Tishkoff, *et. al*., 2007; Enattah, *et. al.,* 2007; Enattah, *et. al*., 2008 and Ingram, *et. al*., 2009) indicating an actual ongoing convergent evolution in human populations.

A variety of genetic signatures as an evidence of recent strong positive selection have been described (Bamshad and Wooding, 2003). The most interesting signatures are high frequency differences of LP alleles (that in linkage disequilibrium) among populations and unusually long haplotype of low diversity (i.e., haplotype that remains intact over unusually long distances). The former indicates that effects of selection that cause alleles to increase highly in frequency are different in some but not all of the populations. But, the later indicate an allele that increased rapidly to high frequency before recombination could disrupt the haplotype on which the allele lies. These signatures can be detected by genotyping common single base pair substitutions in one or more populations as these genetic variants have better power of showing the sign of strong recent positive selection (Sabeti, *et. al.,* 2002). Large differences in allele frequencies between populations have conventionally been detected using the population genetics measure F_{ST} (Akey, *et. al.,* 2002), whereas demonstration that a common haplotype is unusually long requires application of the recently described long-range haplotype test (Sabeti, *et. al*., 2002).

Other form of signature includes an excess of rare variants. This indicates a selective sweep followed by the accumulation of new and rare mutations. Two SNPs in the MCM6 gene near the LCT gene which were reported to be strongly correlated with LP (Enattah, *et. al*., 2002) in populations from Northern European origins was assessed to provide population genetics-based evidence for selection. Bersaglieri, *et. al*. (2004) has reported that these two alleles were rapidly

evolved due to recent positive selections in European populations. The evidence of selection comes from the fact that highly frequent 13910 C>T and 22018 G>C variants in Europe harbor a relatively conserved long region of haplotype together which suggest recent origin of LP due to strong selection. As indicated above, this can be considered as examples of gene culture coevolution and the strongest cases of recent selection on LCT gene and the LP alleles are an example of selection based evolutionary change in human that have traditionally practiced cattle domestication and milk drinking culture (Pereira, *et. al*., 2004). The selection in favor of LP alleles, produced by single gene mutations and widespread, in Europe enabled enhanced survival and growth and provides people the ability to produce lactase and drink milk throughout their lives. Coelho, *et. al*. (2005) have studied that the extents of haplotype variations in 13910C>T and 22018G>A associated with LP on chromosomes from ethnically different populations with different genetic backgrounds in Europe and Africa using highly evolving microsatellite loci. They have found that the haplotype, conferring LP, had a tightly clustered microsatellite allele distribution, irrespective of geographic location. They also indicated that LP most probably arose from different mutations in Europe and the majority of Africa, even if 13910 C>T is so rare to be the causative allele, again suggesting that selective pressure could have promoted the convergent evolution of the trait.

2.2.6. Molecular Basis of LP and Related Challenges

The roles of genetic differences in LCT gene expression were not established until the late 20th century. By the early 1970s, it was established that the LP variations in human populations, as a function of ethnic differences, had a genetic base, although the mechanisms was not clear. Only little was known about the genetic basis of the LP into adulthood life when the correlation of LP and milk drinking culture was examined (Holden and Mace, 1997). Since then, many studies

have been conducted in order to better understand the genetic basis of LP. One study showed that LP is inherited as an autosomal dominant manner as an individual only inherit on copy of the mutated variants to become a LP (Enattah, *et. al*., 2002). Other studies have showed that LP is genetically determined trait by a single gene, LCT gene, mapped to the long arm of chromosome 2 at a location of 2p21 (Figure 2.2), containing 17 exons and roughly 50 Kb in size (Kruse, *et. al*., 1988 and Harvey, *et. al*., 1993). There is one Kb long promoter preceding the LCT gene which transcribed to a mRNA transcript of 6, 274 bp that later translated to primary translation product 1,927 amino acids long (Boll, *et. al*., 1991 and Mantei, *et. al*., 1988).

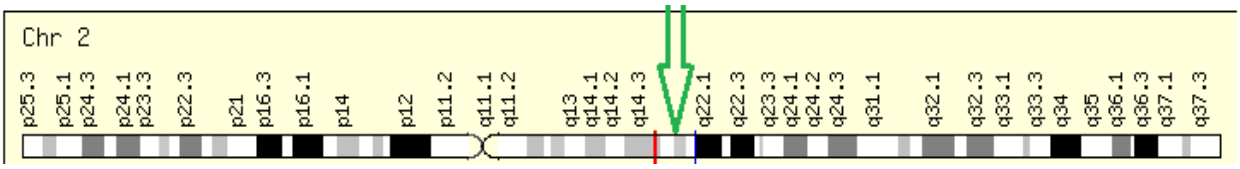

Figure 2.2. Chromosome 2 with marked location of the LCT. Genecards version 3, http://www.genecards.org/pics/loc/LCT-gene.png.

Hence, the LCT and MCM6 gene were both mapped on chromosome 2q21. Scientists have investigated the whole LCT gene and MCM6 region and have identified many polymorphisms in MCM6 as a molecular basis of LP (Figure 2.3). Within the LCT gene, several SNPs with high diversity have been identified, but none of them were shown to be associated with LP (Harvey, *et. al.,* 1998 and Hollox, *et. al*., 2001). Other study by Wang, *et. al*. (1995) for the first time found the existence of the sequence differences responsible for LP or LNP residing within MCM6 gene. Later, in this region, putative *cis*-regulatory regions that have affected the transcriptional activity of the LCT gene have been identified (Poulter, *et. al.,* 2003). Study by Enattah, *et. al.* (2002) was performed in order to better understand the actual molecular basis of

LP phenotype in the same region and showed the existence of perfect associations of LP trait with SNP (13910 C>T) in MCM6 gene. This polymorphism was in what appeared to be cis-acting regulatory element located in the intron 13 of a MCM6 gene. In populations from Europe and their descendants, 13910 C>T SNP was shown to be the common cause of LP trait, where as in the majority of populations of Africa multiple distinct SNPs have been discovered (Mulcare, *et. al.,* 2004; Ingram, *et. al.,* 2007; Tishkoff, *et. al.,* 2007 and Jones, *et. al*.,2013). The identification of several SNPs that appears to be correlated with LP in the same adjacent region indicate that LP have evolved several times independently in human evolution in different areas of the world (Ingram, *et. al*., 2009). The discovery of several independently derived LP associated mutations in different populations challenged the identification of the true molecular basis of LP in the world in general and in Africa in particular (where the case is very complex) including Ethiopia. Moreover, comparisons of LP phenotype frequencies with the distribution of currently known LP associated alleles have suggested that our knowledge of the genetic causes of LP is incomplete in some regions of Africa. It is thus likely that further genetic variants remain to be found (Itan, *et. al*., 2010; Tishkoff, *et. al*., 2007; Ingram, *et. al*., 2009 and Jones, *et. al*, 2013).

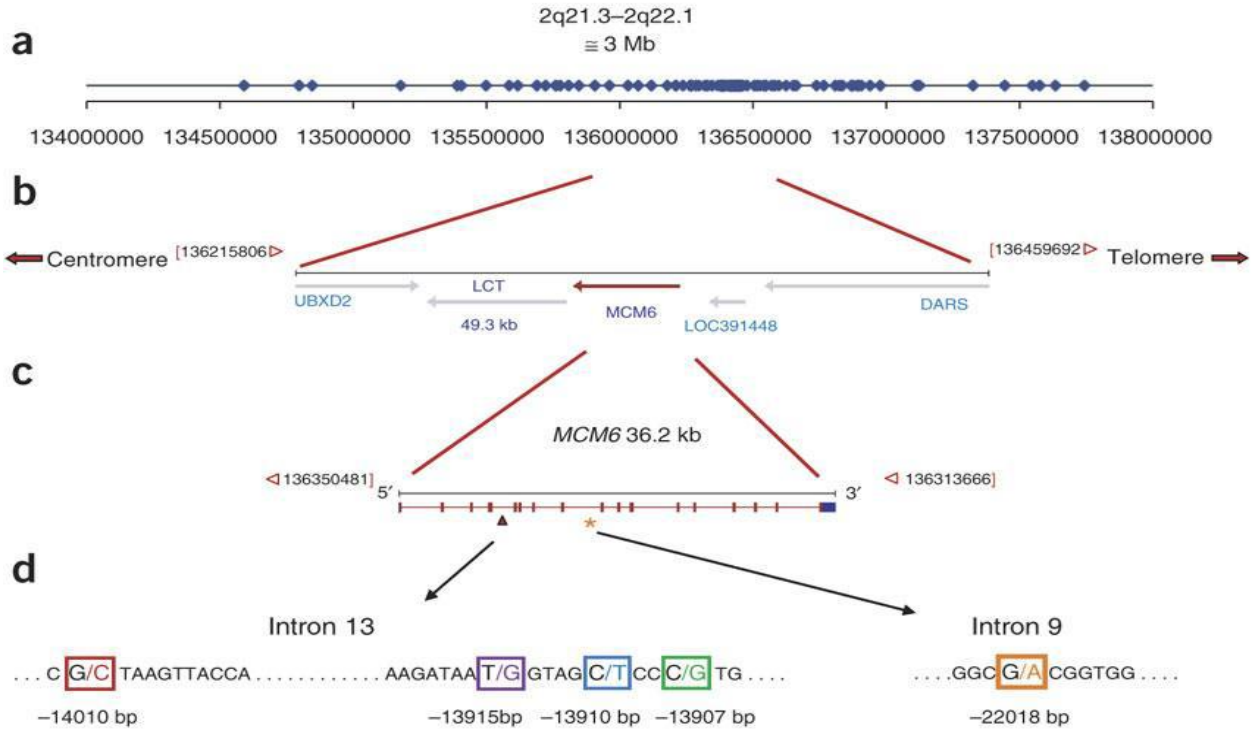

Figure 2.3. Map of the LCT and MCM6 gene region and location of genotyped SNPs.

(a) Distribution of 123 SNPs included in genotype analysis. (b) Map of the LCT and MCM6 gene region. (c) Map of the MCM6 gene. (d) Sequence of the lactase persistence-associated SNPs within Introns 9 and 13 of the MCM6 gene in African and European populations. The Intron 9 SNP and the T-13910 variant in Intron 13 (blue box) are associated with lactase persistence in Europeans, whereas the other three variants in Intron 13 (red, purple and green boxes) are the African variants (Taken from Tishkoff, *et.al.*, 2007).

2.2.7. Multiple Origin of LP Associated SNPs

SNPs associated with LP are shown below (Figure 2.4). Enattah, *et. al*. (2002) have identified two SNPs (3910C>T and 22018G>A) in Finnish families, that regulate the production of lactase enzyme. They showed that these variants are associated with LP in this populations, but by

comparisons, highest frequency and complete association of 13910 C>T polymorphism with LP was observed with 22018G>C. 13910 C>T and 22018G>A variants were shown to be located in intron 13 and 9 of MCM6 gene respectively. Investigators have showed that 13910C>T SNPs have been originated and tightly associated with LP in Euro-Asiatic populations but very low in the majority of African milk drinking pastorals groups where LP phenotype has been reported to be high in frequencies (Bersagleiri, *et. al*., 2004; Mulcare, *et. al*., 2004 and Coelho, *et. al.* 2005). Since the frequency of this allele is so low, it cannot explain the frequency of LP phenotype in the majority of African populations. This led scientists to the conclusion that there may be different cause of LP, or that the13910 C>T allele is not functional, and the actual casual mutation may be located elsewhere on the extended haplotype A indicating the needs for further investigation.

A detailed genotype-phenotype association study of Tishkoff, *et. al*. (2007) have identified 13907C>G (rs41525747), 13915T>G (rs41380347) and 14010G>C (rs145946881) variants within the same regions in pastoralist communities from African and Middle East populations and shown that the mutant form of these variants provided similar LP phenotype, like European variants, suggesting the roles of convergent evolution of LP in different populations. The former two were both identified in Ethiopia, Kenya, Saudi Arabia, Sudan, and Tanzania populations. Where as the 13915T>G was found in Ethiopian Somali, Morocco, and Jordan, the 14010G>C polymorphism was identified in Kenya, Tanzania. These SNPs originated on haplotype backgrounds different from the European 13910 C>T and from each other. Similar phenotype-genotype associations study detected the existing variations in different pastoralists' populations from sub-Saharan Africans and Middle East (Ingram, *et. al.*, 2007). In a sample from East Africa and Middle East participants, they have identified several variants; of these three SNPs: 13907

C>G, 13915T>G and 13913 T>C were new and found in close proximity to 13910 C>T. The former was most frequent in Sudan and Ethiopian (with highest frequency in Afar). While 13915T>G variants are moderately and prevalently distributed in East African and Middle East respectively, 13913 T>C variants were rarely observed. As the 13915 T>G allele had not been reported in Europeans, Ingram, *et. al*. (2007) suggested that it might have originated in the Middle East, where it is seen at highest frequency in Bedouin groups. They reported that it was fairly frequent in East Africa but rarest in West Africa. By this work, they provide clear and direct evidence that *13910 C*>T allele on the background of **A** haplotype is not the causative of LP worldwide. They reasoned that the frequency of *13910 C*>T is not enough to account for LP in these populations who have had milk drinking history, where LP phenotype was reported to be high in frequency, clearly contrasting what were obtained for Europeans except for few West African pastoralist Fulani and Hausa groups from Cameroon. These authors also indicated that the mutational basis of LP/LNP in Africa appears to be different from that in Europeans and the 13910 C>T variant did not account for LP in most people from African or Arabian.

All these studies have indicated that the 13910 C>T variant correlated perfectly with LP in European populations but is almost nonexistent among sub-Saharan African populations where high prevalence of LP phenotype exist. Moreover, Enattah, *et. al*. (2008) identified two new mutations responsible for LP among Saudis population samples and confirmed the absence of the 13910 C>T allele. These two new mutations present as a compound allele: 13915 T>G within the 13910 C>T enhancer region and a synonymous SNP, 3712 T>C, within the start codon of the LCT. They demonstrated that these two variants have evolved independently from different, highly divergent allelic back ground and driven to very high frequencies in different populations. Their results support the convergent evolution of the LP in different populations that presumably

had different histories of animal domestications and adaptation to dairy culture. More recently, Jones, *et. al.* (2013) analyzed the MCM6 gene in LP individuals from diverse ethnic group in Ethiopian and identified muntant variants namely; 13910 C>T, 14010 G>C, 13907 C>G, 13915 T>G and 14009 T>G and confirmed the strong association of 13907 C>G and 13915 T>G with LP. They also found the presence extremely rare (14010 G>C and 13910 C>T) variants and discovered other new variants (14009 T>G) which was shown to be strongly associated with LP. Moreover, they indicated that different mutations to the DNA exist to maintain the LCT gene to be switched on in some adults from East African population. These mutations would be attributed to the variability among different people for LP/LNP on which natural selection acts and provide adaptation to milk drinking, from which the phenomenon of soft selective sweep was inferred in the population of Ethiopia. In conclusion, the mutational basis of LP phenotype is not clearly identified yet. The suggestion is that still either there may be several other variants to confer LP, yet unidentified in different populations, or that 13910C>T may not totally associated to LP and the true responsible mutations may be located elsewhere within the potential region close to this allele where multiple independent mutations approximately within the 100 bp range were previously identified. This and other data further indicate the importance to investigate additional LP variants in global populations (Enattah, *et. al*., 2008).

In summary, the several independent derived alleles in the *LCT* enhancer region are spread through several ethnic groups across the world with different effects of mutation resulting in different status of the ability of milk consumption. In general week to perfect association of LP variants with the ability of milk consumption was reported (Oljira, *et. al*., 2014). For instance, 13910 C/T and 14009T/G variants are associated significantly with the ability of milk consumption in Europe and Ethiopia respectively.

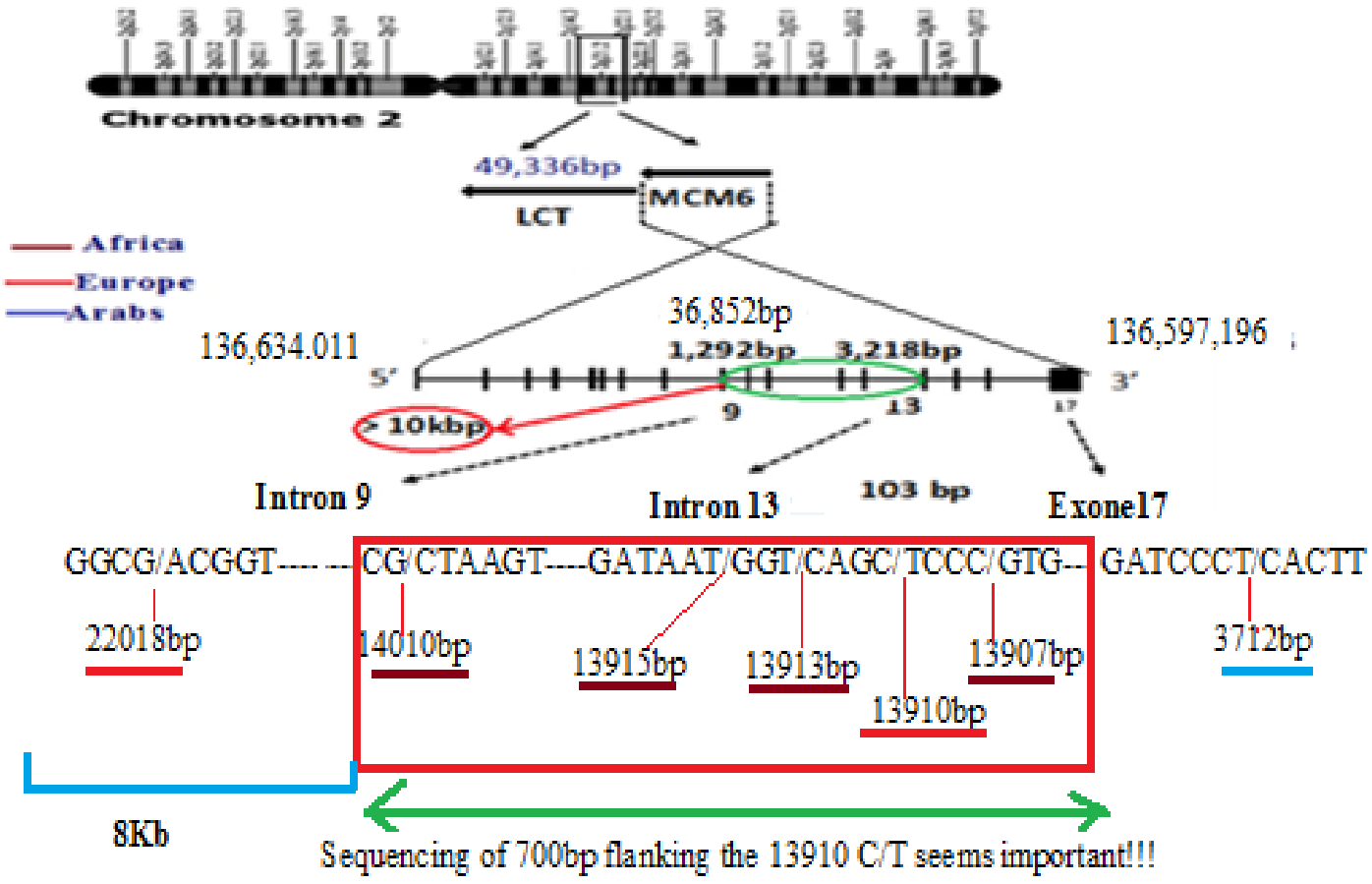

Figure 2.4. SNPS associated with LP. Map of the region of human chromosome 2q21 containing the human LCT and MCM6 genes. Intron-exon organization of the MCM6 gene is shown. Intron 9 and 13 and exon 17 are labeled. SNPs associated with LP in Europeans and certain African and Arabs pastorals populations. The location of each SNP relative to the initiation codon of the LCT is given in bps.

2.3. The Human Y Chromosome

2.3.1. Features of Y Chromosome

The Y chromosome is haploid and is inherited paternally. It is the second smallest chromosome with about 60 Mb lengths. About 95% of of Y chromosome length does not recombine and is known as non-recombining portion of Y chromosome (NRY), also known as the male-specific region. The NRY consist of euchromatic and heterochromatic regions flanked by the pseudoautosomal regions (PAR) at the terminal regions of both the short (Yp) and long (Yq) arms (Figure 2.5), were recombination process occurs during meiosis (Hammer and Zegura, 1996 and Skaletsky, *et. al,* 2003). In addition to short terminal repeats (STRs) and minisatellites, NRY chromosome has numerous bi-allelic markers (SNPs and YAP or other insertions/deletions 'indels'), as part of male specific region located on NRY. The bi-allelic markers have lower mutation rates and preserve a unique record of mutational events that occurred in paternal generations (along male lineages) through out evolution. Variation studies of NRY chromosome have been proposed for study of male mediated migration events and for reconstructing male ancestral history. Their features make the NRY chromosome variations are a suitable tool for the analysis of modern human evolutionary and genealogy studies. Uniparental mode of inheritance and haploid state of Y chromosome help to trace lineages found in present day male individuals to male ancestors who lived in the past. The absence of recombination would enable to construct the evolutionary histories of mutations found in contemporary divergent lineages to a common point in the past and provide the Y chromosome with the best haplotypic resolution in the human genome. Genetic Variation at many points along NRY chromosome is combined to form a Y haplogroup for a sample. However, the implication of lack of recombination is that the Y chromosome cannot be studied easily as part of a genome wide association study (GWAS)

pipeline. In contrast, it is analyzed using haplogroups which are stable lineages of the Y chromosomes that share a common ancestral allele.

Studies of the variation of the bi-allelic sites have been proposed to be very useful markers for evolutionary studies like detecting population migration and historical events (Malaspina, *et. al*., 1990 and Cruciani, *et. al*., 2002). It was shown that the haploid nature and the paternal inheritance of the Y chromosomes are key features in revealing highest level of population sub structuring than the Autosomes chromosome (Avise, *et. al*., 1987 and Underhill, *et. al*., 2001). They have reported that the absence of recombination helped the reconstruction of clear haplogroup phylogeny, which can be related to the geographic distribution of the haplogroups. NRY chromosome polymorphism was first discovered roughly three decades back (Casanova, *et. al*., 1985 and Ngo, *et. al*., 1986). Since then, numerous studies involving different aspects of human population genetics have been conducted. Recently, studies have revealed very large numbers of bi-allelic polymorphisms in Y chromosome and showed detailed phylogeographic picture of modern global population structure and past population movements and interactions (Rosser, *et. al.*, 2000; Underhill, *et. al*., 2000 and Hammer, *et. al*., 2001). The availability of these highly geographically structured sets of bi-allelic markers has stimulated the analysis of more restricted areas, giving clues regarding the peopling of Europe (Semino, *et. al.*, 2000 and Scozzari, *et. al.* 2001), Asia (Capelli, *et. al.*, 2001 and Karafet, *et. al.*, 2001), Oceania (Capelli, *et. al.*, 2001 and Underhill, *et. al.*, 2001), and the Americas (Underhill, *et. al.*, 1996).

Similarly, genetic studies of African populations, through the analysis of other DNA markers have been conducted. These have provided evidence that support an African origin for human species and have revealed that the African continent has had a central role in human evolutionary

history (Tishkoff, *et. al.*, 1996; Ingman, *et. al.*, 2000 and Underhill, *et. al.*, 2000). Small studies have been performed in African populations using the Y chromosome bilalleic markers. Even these studies were, either based on a small number of polymorphic markers (Seielstad, *et. al.*, 1994 and Scozzari, *et. al.*, 1999) and focused on specific geographic locations inside the continent (Passarino, *et. al.*, 1998 and Semino, *et. al.*, 2002). Indeed, Ethiopian population has not yet been well represented (Passirino, *et. al.*, 1998), although it occupies a central position with respect to human origin and "out of Africa" model of human migration. The present day complex ethnic and linguistic diversity of Ethiopia that thought to have originated from the complexity of the ancient Ethiopians' interactions with very different populations has not been fully addressed (Lewin, 1987; Levine, 1974 and Lopez, *et. al.*, 2015).

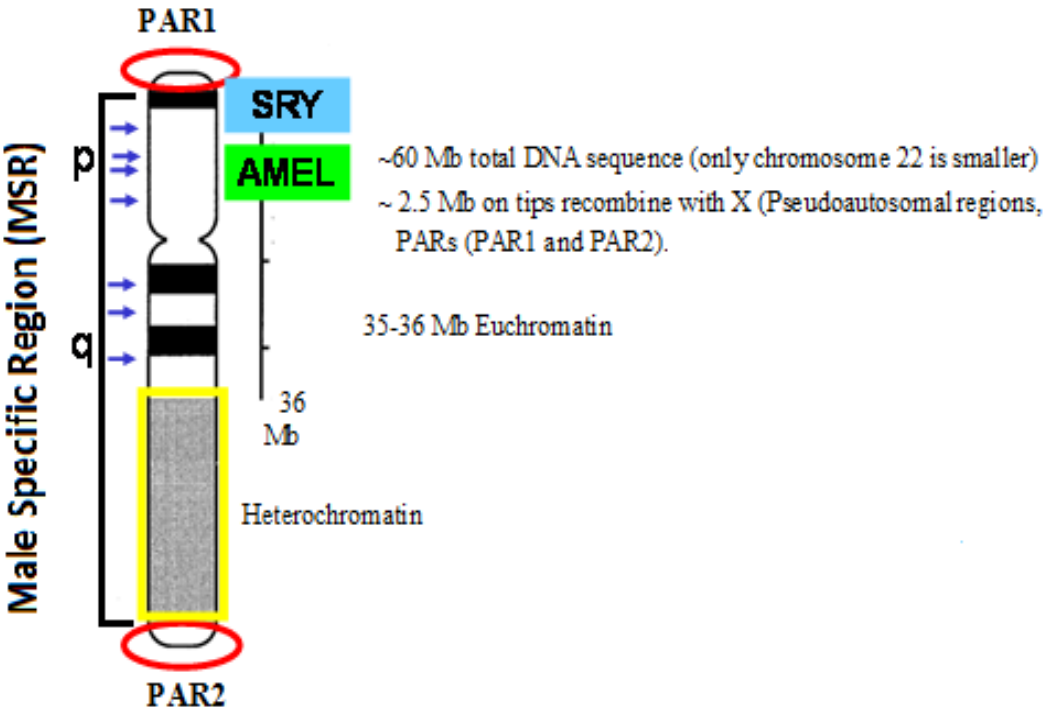

Figure 2.5. Schematic representation of the human Y chromosome structure including NRY and PARs. SRY and AMEL stand for Sex Determining Region Y and Amelogenin gene respectively. (Taken and modified from Nucleic Acids Res, 2000).

2.3.2. Y Chromosome Haplogroups

Due to the growing number of known bi-allelic polymorphisms there has been a corresponding increase in the number of different systems used to name these binary haplogroups. This led to inconsistency and increasingly inconvenient nomenclature among a scientific community. As a result, the Y chromosome Consortum (YCC) have constructed a detailed phylogenetic tree of NRY bi-allelic haplogroups and described a new nomenclature system that utilizes the capital letters A to R to define major Y chromosome DNA haplogroups (YCC, 2002). The haplogroup is defined by mutations that are shared by many people in the NRY. They represent major branches of the Y chromosome phylogenetic tree that share hundreds or even thousands of mutations unique to each haplogroup. The presence of a series of Ychromosome mutations serves as markers and defines the Y DNA haplogroups. Terminal mutations, the mutation furthest down in the Y chromosome defines the phylogenetic tree of subclades.

The YCC revealed the relationship among haplogroups of contemporary global population structures and past population movements and interactions based on bi-allelic markers of NRY in a globally representative set of samples (YCC, 2002). This system unified all past nomenclatures and was flexible enough to allow the inclusion of additional mutations that may result from the discovery of new polymorphisms and NRY lineage. One year later, a modified version of the YCC tree was subsequently devised by Jobling and Tyler-Smith (2003). Until 2008, although hundreds of polymorphisms have been examined and discovered in world human populations, efforts to revise and update the Y chromosome binary haplogroup tree were paused. Then, Karafet, *et. al.* (2008) reinitiated the effort to update the Y chromosome phylogenetic tree and extensively revised the tree containing many distinct haplogroups. They worked to bring the Y chromosome tree up to date and provide improved nomenclature system based on the rules put

forward by the YCC in 2002. They have described major changes in the topology of the tree and provided names for new and rearranged lineages within the tree following the rules presented in it. Also, they have mentioned that these several changes in the tree topology have important implications for studies of human ancestry (Karafet, *et. al.*, 2008). For the future, the current working Y Chromosome haplogroup nomenclature (since it has been revised in 2008) will be changed to accommodate the increasing number of NRY chromosome mutations being discovered and tested, and the resulting expansion of the Y chromosome phylogenetic tree.

Currently, the Y chromosome tree consists of 20 major Y chromosome haplogroups named with the capital letters A through T (Figure 2.6), with further subclades named using numbers and small letters. The tree incorporated more than 311 confirmed haplogroups defined by more than 600 Y chromosome mutations (Jobling, 2012 and van Oven, *et al.*, 2014). Subclades of Y DNA haplogroups are named by the first letter of the major Y DNA haplogroup followed by a dash and the name of the defining terminal mutation. Y chromosome short tandem repeats (Y STRs) can be genotyped to generate haplotypes, which are then used for finer resolution within the haplogroups (Underhill and Kivisild, 2007). The distribution of Y chromosome haplogroups are highly structured according to geographic origins of the studied populations. Of the major Y chromosome haplogroups, haplogroups A, B and E found most commonly in African populations although the later found at varying frequencies in the different groups tested in side and out side of Africa. The distribution of Haplogroup A and B are entirely limited to African populations implying an African origin of human Y chromosome. They represent the initial and the oldest split in the Y chromosome in the Y chromosome tree and are the most divergent haplogroups placed on the deepest branch of the global phylogeny. They are core haplotypes that are associated with the distribution of ancient hunter gatherer groups before the expansions of

pastoralists (Hammer, *et. al*., 2001; Underhill, *et. al*., 2001 and Wood, *et. al.*, 20050. The rest of the Y chromosome tree are represented by haplogroup E and other haplogroup and are present outside of Africa (Underhill and Kivisild, 2007; Tishkoff, *et. al.,* 2007; Karafet, *et. al.,* 2008; Batini, *et. al.,* 2011 and Pickrell, *et. al.*, 2013). Haplogroup E is the most common African lineages and all non-African clades. Some sub clades of haplogroup E spreaded out of Africa into the Middle East and Europe and has attained highest frequencies across the world. All of the African and non-African haplogroup distributions are briefly described as follows.

Haplogroup A was defined by the M91 and P97 mutations (Underhill, *et. al.,* 2000 and Karafet, *et. al.,* 2008). It was the first to split off the Y chromosome phylogeny and contained more several branches determined by numerous internal mutations (Cruciani, *et. al.,* 2011). A strict regional distribution was particularly observed for haplogroup A. The deepest split separated haplogroup A in to further sub caldes of haplogroup A, that exhibit strong geographical structuring. While haplogroup A exhibit a unique pattern of haplogroups lineages and is found most frequently in Mali, Morocco, Sudan, Ethiopia and Khoisan, it is found in lower frequencies in northern Cameroon (Underhill, *et. al.,* 2000; Semino, *et. al.,* 2002; Knight, *et. al.*, 2003 and Cruciani, *et. al.,* 2011).

Haplogroup B was inially defined by four mutations; M60, M181, P85, and P90 (Underhill, *et. al.,* 2000, Underhill, *et. al.,* 2001 and Karafet, *et. al.,* 2008) and was further increased by the discovery of seventeen branches with twenty-eight internal markers (Cruciani, *et. al.,* 2011). Haplogroup B occurred throughout Africa but have higher frequencies among hunter-gatherer populations such as Pygmies, Khoisan and Hadza, with some lineages segregating specifically to each of these groups. This haplogroup exhibit some geographic structuring in sub-Saharan

Africans, showing clear difference between the B sub haplogroups associated with the hunter-gatherer and all the other African populations. High frequencies of haplogroup B defined by the M112 mutation was found in hunter-gatherer populations, while M150 mutation was found in Bantu-speaking populations (Underhill, *et. al.,* 2001; Cruciani, *et. al.,* 2002; Semino, *et. al.,* 2002; Knight, *et. al.,* 2003; Tishkoff, *et. al.,* 2007 and Baitin, *et. al.,* 2011).

Haplogroup E of the human Y chromosome is massively the most common African haplogroup (Wood, *et. al.,* 2005). The very high frequency throughout Africa most likely reflects the Bantu-agricultural expansion in the last 3000 years (Underhill, *et. al.,* 2001). It is defined by the presence of a YAP, from which eighteen mutations and eighty-three polymorphic sites (that defines fifty six distinct lineages within the sub clades) are derived. Since, the YAP insertion is also found in Asian haplogroup D, this and the haplogroup E are the sister clade to one another. Since the DE*(ancestral paragroup) have been reported to have found in Africa (Weale, *et. al.,* 2003 and Rosa, *et. al.,* 2007) and Asia (Shi, *et. al.,* 2008), it is generally believed that haplogroup E might have originated in Africa; most likely in East African. Haplogroup E also found outside of Africa almost in all continents with varying frequencies. It is found with moderate frequencies in the Middle East and southern Europe but rare in Central and South Asia. Although haplogroup E groups are widespread all over Africa, the distributions of several unique haplogroups are not uniform across the continent (Hammer, *et. al.,* 1998; Cruciani, *et. al.,* 2004 and Karafet, *et. al.,* 2008). In agreement with this, the highest frequencies of E-M78 were reported in Oromo populations from Ethiopian (Semino, *et. al.,* 2002) and in Masalit and Fur populations from Sudan (Hassan, *et. al*., 2008) than others. These authors have also reported that whereas, E-M33 is largely distributed among Fulani and Hausa, E-M2 is limited to Hausa and E-M215 was found to have occur more in Nilo-Saharan rather than Afro-Asiatic speaking groups. Moreover, the bi-

allelic Y chromosome investigation in Eritrean populations by Gebremeskel and Ibrahim (2014) have revealed the highest distribution of haplogroup E and indicated the proto-Afro-Asiatic groups carrying E-M78 mutations are widely spread. This investigation showed major contribution of East African populations to the foundation of the E haplogroup, the origin and spread of Afro-Asiatic languages and the history of pastoralisms. In other studies, haplotypes carrying the mutations E-M35 found most commonly in North and East Africans, Mediterranean Islanders and Southern Europeans, but occurs at scarce frequencies within Bantu speakers. Among the different lineages carrying the E-M35 mutation, haplotypes defined by M78 (E3b1a, according to the Karafet, *et. al.*, 2008) occurs in East Africa, North Africa, the Middle East and Europe. In comparision, the E-M35 and E-M215 mutations have the highest frequency in Africa and the vast distribution outside of Africa. In other word, in addition of having a widespread African representation, these mutations have broad non-African distribution. The mutation E-M215 and E-M35 have East African origin and multiple exit routes out of Africa have been proposed (Cruciani, *et. al.,* 2007). While the E-M75 has been found mostly in the Bantu-speaking populations from East and South Africa, E-M33 was present at low frequencies across Africa and with different distributions. In contrast, haplogroup E-M2 and seven other mutations was commonly found in sub-Saharan populations and was associated with the expansion of Bantu-speaking populations (Underhill, *et. al.,* 2001; Cruciani, *et. al.,* 2002; Cruciani, *et. al.,* 2004 and Semino, *et. al.,* 2004). It should be noted that the subgroups of these haplogroups have different distributions and frequencies across the African region. For instance, the M191 mutation defines the most frequent E-M2 subgroup and was an evidence of a founder effect that resulted from the Bantu-expansions (Cruciani, *et. al.*, 2004 and Semino, *et. al.,* 2004).

The other haplogroups have non-African origin. Haplogroup C is found distributed widely across Eurasia, but at a low frequency in the Europe, Americans, and absent in Africa. This haplogroup also occur at high frequencies in Japanese and Tibetans. As a non-African lineage, haplogroup C is highly informative in tracing the migration route of the African exodus in prehistory. Haplogroups G and J have a wide spread in the Middle East and from there expanded to southern Europe probably with Neolithic farmers (Underhill, *et. al.,* 2001 and Cavalli-Sforza and Feldmann 2003). Hplogroup D, H and P are Asian specific, where as haplogroups F, L and N are shared primarily between Asia and the Europe and the Middle East region. In addition, haplogroups such as K, M, and O was found more frequently among the Asian populations. Further on these, haplogroups H and L together remained in India (Hammer and Zegura, 2002 and Cavalli-Sforza and Feldmann, 2003). While the distribution of haplogroup O is shared by Asia and Oceania, widest distribution of haplogroup N has been observed in Russia (Puzyrev, *et. al.,* 2003). Additionally, haplogroup H has been widely distributed in India, Pakistan and Sri Lanka (Qamar, *et. al.,* 2002 and Khurana, *et. al.,* 2014).

Haplogroup G and R are widely shared outside of Africa. Despite being predominantly found among Europeans, haplogroup R has been identified at higher frequency in Pakistan (Cavalli-Sforza and Feldmann, 2003). Haplogroup R has also been found in Central and West Africa in significant proportion (van Oven, *et. al.,* 2014). Whereas haplogroup Q is primarily distributed in the Americas and has low frequencies among the Asian populations (Hammer and Zegura, 2002), haplogroup J has been mostly found in Europe, Middle East, Asia and Africa. Haplogroups I and R were commonly found in Europe (Semino, *et. al.,* 2000). In summary, there are higher numbers and comparatively uniform distribution of haplogroups in Asia than in Africa, Americans and Europe. Predominant haplogroups such as E, R and Q are known to

characterize African, Europe and Middle East and Americans regions respectively. Both the high numbers of Asian haplogroups and their relatively uniform distribution undermine centrality of Asian for human dispersals (Hammer, *et. al.*, 2001).

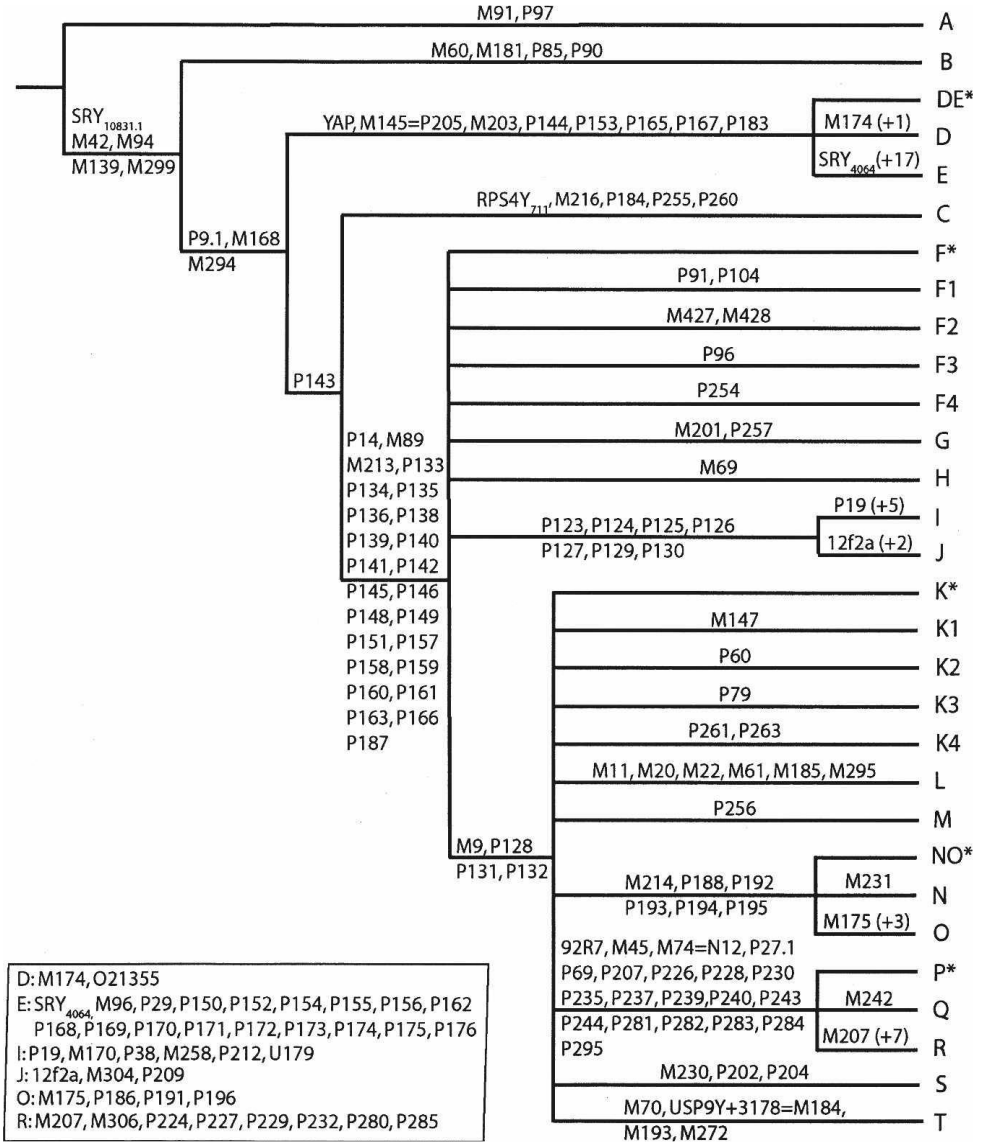

Figure 2.6. An abbreviated form of the Y chromosome tree. The maximum parsimony tree of 311 Y chromosome haplogroups is shown. Haplogroup names are given at the tips of the tree and major clades are labeled with large capital letters and shaded in color. Mutation names are given along the branches (The figure is taken from Karafet, *et. al.*, 2008).

2.3.3. Y *Alu* Polymorphisim (YAP)

There are two type of polymorphism on Y chromosome. The unique (that have arisen once) and recurrent (that have created independently many times) polymorphisms. Thus, the two polymorphisms can be distinguished by the nature of the mutation arisen on Y chromosome. Generally, the insertion of an *Alu* sequence and the origin of point mutations at a particular position are a rare event, so the YAP insertion (Hammer, 1994) and Point mutations (Whitefield, *et. al.,* 1995) can be safely assumed to have occurred only once and have a unique origin. The well described recurrent polymorphisms are changes in the number of repeats in minisatellites and microsatellites which are expected to be frequent and assumed to have happended several times. Although the frequencies expected for duplications and deletions are more difficult to predict, Y haplotypes can be examined for evidence of recurrent mutations. Both polymorphisms of Y chromosomes are useful for population studies. As expected, haplotype constructed using YAP and point mutations usually have simple relationships to one another (Hammer and Horai, 1995), while haplotypes constructed using microsatellites show very complex relationships (Roewer, *et. al.,* 1996).

As described by Hammer (1994), the YAP (DYS 287 locus) is unique event polymorphism (UEP) markers that define a deep-rooting clade of the Y chromosome genealogical tree. This marker is stable and originated some 65 KYA as an insertion element (~ 450bp) in human DNA, although, the geographic origin of YAP is still controversial. Studies have suggested that the YAP insertion present in two lineages world wide, corresponding to Haplogroup D and Haplogroup E polymorphisms. While haplogroup D is unique to Japan and South East Asia, haplogroup E is restricted to sub-Saharan Africa, Middle Eastern and Southern Europe populations. In contrast, haplogroup D is absent from areas where haplogroup E found and vice

versa. The Y chromosome haplogroup E carrying YAP insertion are known to have occur in high frequencies within and among African populations, followed by north African populations, and is rare among Europeans. Except the small presence in the Middle East and Central Asian populations, YAP insertion element in Y chromosome haplogroup E is absent in the majority of Asian populations investigated so far. YAP element in Y chromosome haplogroup D is present among Japanese populations in relatively higher frequency. Interestingly, YAP individuals in these two diverse geographical areas of Africa and Asia are found to have different lineages, haplogroup E and D respectivelly (Underhill, *et. al.*, 2000; Agrawal, *et. al.*, 2005; Rosa, *et. al.*, 2007 and Shi, *et. al*., 2008).

2.4. Mitochondrial DNA (mtDNA)

2.4.1. Feature of mtDNA and *MT- COXII*

Human mtDNA play a key role in cellular functions, apoptosis and aging. It is highly compact, double-stranded, circular genome of about 16,569 base pairs (bp) in length. mtDNA contains only thirty-seven genes; thirteen oxidative phosphorylation (OXPHOS) genes, two rRNAs (12S and 16S) and twenty-two tRNAs genes. There are no introns within mtDNA genes and almost no intergenic noncoding nucleotides exist except the 1.1 kb displacement loop (D-loop), which contains transcriptional promoters and at least one of the proposed replication origins (Figure 2.7). mtDNA has a high copy number in a single cell: every cell contains as many as several hundreds or thousands of copies of the mitochondrial genome. When single cell contains identical mtDNA molecules, it is known as homoplasmy, but when two or more different mtDNA molecules present in a cell, it is knowna as heteroplasmy.

Like Y chromosome, mtDNA lack the process recombination (their differences are only due to mutations). However, unlike Y chromosome (which is paternally inherited), mtDNA are strictly inherited through the maternal lineage. Also, mtDNA is characterized by very rapid evolution due to high mutation rate (2-4% per million years and 5-10 times more) than the nuclear genome. Thus, mtDNA is an extremely variable genome and it can show the high degree of variability between individuals (Wilson, *et. al.*, 1985). These features make mtDNA a remarkable molecule for evolutionary genetic studies as compared to the nuclear DNA. They have specific uni parental inheritance only from mothers to their child, which is useful for tracing matrilineal kinship in many generations.

There is only one mitochondrial ancestor in every generation. This direct inheritance of mtDNA led to the idea that all humans alive today had a single common mitochondrial ancestor at some point in the early past. It is the MCRA of all presently living humans. Since this ultimate common ancestor necessarily had to be female, and popularly named as "**Mitochondrial Eve**", and its male male analoge is the **Y chromosome Adam** (Y-MCRA). The variations in mtDNA in fact provide a reliable record of the maternal lineage of human species. The biggest non-coding, also called control regions (displacement loop 'D-loop') is the most variable part of the mtDNA. The most fast evolving segment of this (over ten times that of the protein coding) region are the hyper variable segment I (*HVS I*) and hyper variable segment II (*HVS II*) which are focuses of many studies and researches of the origins and evolution of human populations (Hammer and Horai, *et. al.*, 1995 and Tully, *et. al.*, 2000). These sequences are more useful and precise indicators of recent variation among and within human populations (Ward, *et. al.*, 1991). However, the coding region of the mtDNA (*MT-COXII* and others) can be used to provide a dependable record of past distant mutations which inturn can be used to trace matrilineal lines.

This is because, the mtDNA coding regions contain more slowly evolving protein coding genes and thus they are preferable than the fast-evolving region for the detection of slow differences among human population, showing that each region has its own merits and drawbacks.

Currently, to overcome the drawback of the fast-evolving non-coding and slow evolving coding region, complete sequencing the mtDNA have been extensively used and improved their importance in human evolutionary studies. So, the choice among which mtDNA region should be used is based on the available resources and objective of the study. The resulting mtDNA variation data has become an attractive source of information for studies of human evolution, migration and population history. It can be used to construct groups of stable haplotypes, called haplogroups that contain information about the order of evolutionary processes in time and space. DNA assessments in worldwide human population have demonstrated a continental structure in the distribution of mtDNA lineages. mtDNA haplogroups tend to be geographically restricted and they are used to genetically distinguish populations (Anderson, *et. al.*, 1981 and Pakendorf, *et. al.*, 2005).

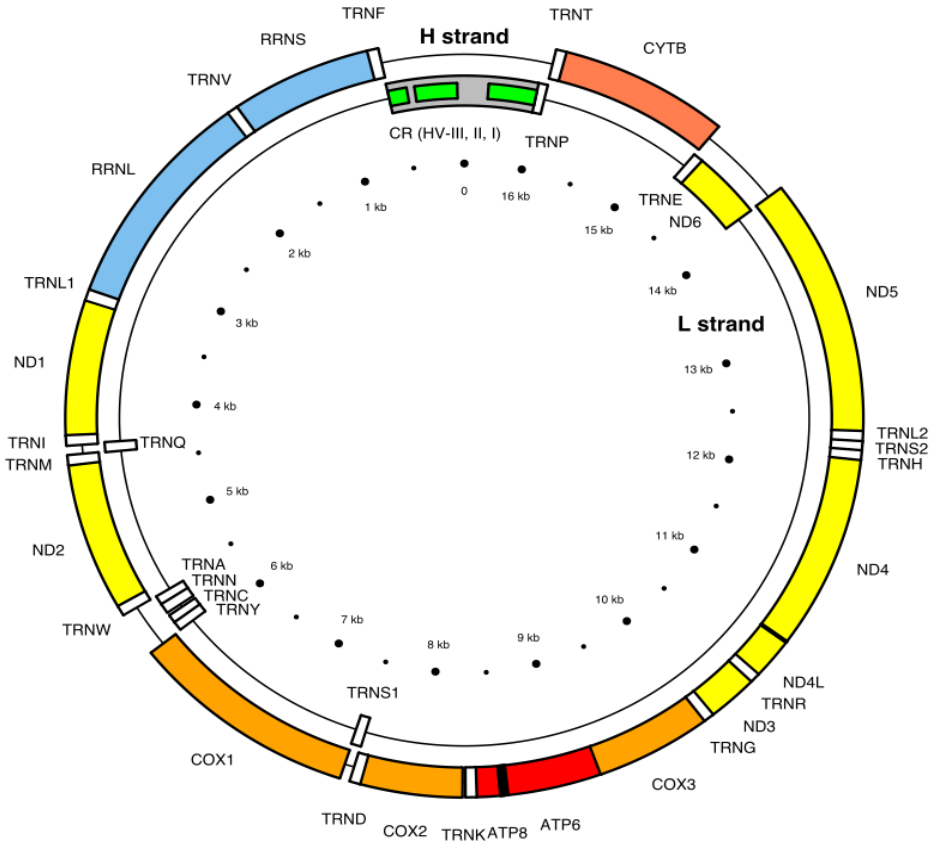

Figure 2.7. Map of the human mitochondrial DNA genome (16569 bp, NCBI sequence accession NC_012920 (Anderson, *et.al.*, 1981). The H (heavy, outer circle) and L (light, inner circle) strands are given with their corresponding genes. There are 22 transfer RNA (TRN), 2 ribosomal RNA (RRN) genes: S, small subunit, or 12S and L, large subunit, or 16S (blue boxes), 13 protein coding genes: 7 for NADH dehydrogenase subunits (ND, yellow boxes), 3 for cytochrome c oxidase subunits (COX, orange boxes), 2 for ATPase subunits (ATP, red boxes), and one for cytochrome b (CYTB, coral box). Two gene overlaps are indicated (ATP8-ATP6, and ND4L-ND4, black boxes). The control region (CR) is the longest non-coding sequence (grey box). Its three hyper-variable regions are indicated (HV, green boxes).

2.4.2. The Mitochondorial DNA (mtDNA) Heteroplasy

Hundreds to thousands of copies of mtDNA are present in each single human cell, in contrast to only two copies of nuclear DNAs. These mtDNAs can differ from each other as a result of mutations, which are variably distributed in mtDNA genes (Pereira, *et. al.,* 2009). The coexistence of multiple mtDNA variants in a single cell or among cells within an individual is called heteroplasmy. Earlier studies have shown that, for the vast majority of humans, all mtDNA molecules are identical at birth (Tayelor and Turnbull, 2005). However, recent work has shown that 25% of healthy individuals inherit a mixture of wild-type and mutant mtDNA (heteroplasmy), which have originated through mutations and almost exclusively involves the non-coding mtDNA D-loop (Li, *et. al.,* 2010). Even less frequently, small number people inherit a potentially pathogenic variant of mtDNA in the coding region (Elliott, *et. al.,* 2008). Thus, heteroplasmic mtDNA mutations are found in mtDNA coding genes and non-coding genes with variable degree and seem to increase as the age increase. Since mtDNA are exclusively maternally inherited and lack recombination, only variations introduced through mutations can be manifested at the population level. As a result, only mutation mediated variation of mtDNA has played a key role in determining population migrations on a global scale (Torroni*, et. al.,* 2006) and in the confident identification of biological samples in forensic medicine (Just, *et. al.,* 2009). That is, these variations of mtDAN can be used for ancestry and forensic identification. Moreover, mtDNA heteroplasy has been reported to be associated with large spectrum of human diseases (Wallace, *et. al.,* 2013).

2.4.3. MT-COXII in Phylogenetic studies

Although numerous studies have been conducted in the area of mtDNA genetics, little attention has been given to its coding regions. Most of the evolution and phylogenetic studies have employed the D-loop region that includes the *HVR1* and *HVRII* than the slow evolving coding genes (*MC-COXII* and others). The D-loop, which is the control region, is fast evolving region accompanied with complications of extreme variation in substitution rate and parallel mutations between sites. These caused the estimation of genetic distance difficult and phylogenetic inferences questionable (Maddison, *et. al.,* 1992 and Tamura, *et. al.,* 1993). With the advent of large scale sequencing technologies, it become fashionable to sequence the entire mitochondrial genome or step-by step sequencing of its genes. Employing genes with slow mutation for evolution and phylogenetic studies was suggested as an alternative for the whole mtDNA sequencing. However, very few people were carried out studies to evaluate the validity of using single conserved mtDNA gene as an evolutionary tool. *MT-COXII* was used to estimate divergence time between modern human and chimpanzee depending on class 1 mutations and maximuim likelihood analysis. This study depending on *MT- COXII* and *ND4* alone supported the out of African model of human origin and showed the validity of using slow evolving mtDNA genes for coalescence and evolutionary inference (Ruvolo, *et. al.,* 1993). Moreover, recent studies have shown the validity of mtDNA single genes in the regulation of cellular activities. Annotation of the 510-bp fragment as a part of *MT- COXII* and phylogenetic analysis of homologous sequences containing *COXII* showed DNA transfer events from mtDNA to nuclear genome. This observation provided a clue for further understanding of the roles of mtDNA in regulation of DNA methylation in the nucleus (He, *et. al.,* 2010).

2.4.4. Mitochondrial DNA Haplogroups

Large scale studies of global population have shown that there were marked differences in the genetic profile and haplogroup distribution of human mtDNAs in different geographic regions. Previously, these clear differences in mtDNA were attributed to founder effects. However, this model is difficult to accept because all of the African specific mtDNA lineages and the ancestors of the Eurasia radiation are found in North East Africa (Mishamar, *et. al.,* 2003). However, later it was suggested that natural selection may have played a role in shaping current distribution of mtDNA sequences variation in humans and that one of the selective influences was climate (Balloux, *et. al*., 2009).

Just as in the case of Y chromosome, the mtDNA genome has been the most widely used system for the investigation of human evolution. As mentioned earlier, it has become the system of choice because of its uniparental maternal inheritance and lack of recombination. These properties allow evolutionary histories to be reconstructed without the complexities imposed by recombination of paternal and maternal genomes. The lack of recombination would allow the data from coding and non-coding genes of mtDNA to be combined into the shape of a phylogenetic tree. To this end, while the data from the coding genes of mtDNA have not yet combined and converted to the phylogenetic tree, the data from the non-coding genes have combined and changed into the shape of phylogenetic trees. On the bases of the data generated from the non-coding genes, the most common branches of tree were assigned alphabetic labels that became to be known as mtDNA haplogroups. The oldest human mtDNA haplogroups originated in Africa about 130-200 years before present (YBP) and gave rise to a series of African specific haplogroups (L1, L2 and L3), which in aggregate form African

macrohaplogroup L. The African haplogroups gave rise to other macrohaplogroups and branches of the global phylogenetic tree during the migration waves from Africa all over the world. Of all the African mtDNAs, only two mtDNA lineages, macrohaplogroups M and N, successfully left Africa (about 65,000 -70,000 YBP) and colonized the rest of the world. The haplogroup L3 is ancestral to macrohaplogroups M and N. They arose in North East Africa and spread into Europe and Asia. In one migration, M and N left Africa and traveled along the tropical South East Asian coast, ultimately reaching Australia. Macrohaplogroup N mtDNAs also moved north into the Middle East and radiated to create submacrohaplogroup R. Both N and R lineages spread into Europe to generate European-specific haplogroups (H, I, J, N1b, T, U, V, W and X derived from haplogroup N) and at present they comprise the majority of mtDNAs in Europe.

In Asia, macrohaplogroup N radiated to form haplogroups A and Y, and the N derived R lineage generated haplogroups B and F. From South East Asia, macrohaplogroup M moved Northward to form an array of Asian-specific mtDNA haplogroups (C and D). The Asian haplogroups A, B, C, D, F and G derived from M and N. Ultimately, haplogroups A, B, C, D, and X migrated in to Americas and become frequent among the Native American populations (Figure 2.8).

In most studies on global samples, evidences that relate human origin, migration route and dispersal out of African using patterns of an established haplogroups was based on mitochondrial control region (*HVR I* and *HVR II*) variations (Cann, *et. al.*, 1987; Underhill, *et. al*., 2000; Ingman, *et. al*., 2000 and Tishkoff, *et. al.*, 2009). At the begining, since the control region has higher mutation rates, whole mitochondrial genome sequencing was regarded as the method of choice for evolutionary studies (Ingman, *et. al.*, 2002 and Fraumene, *et. al.*, 2003), gradually replacing the control region sequencing method.

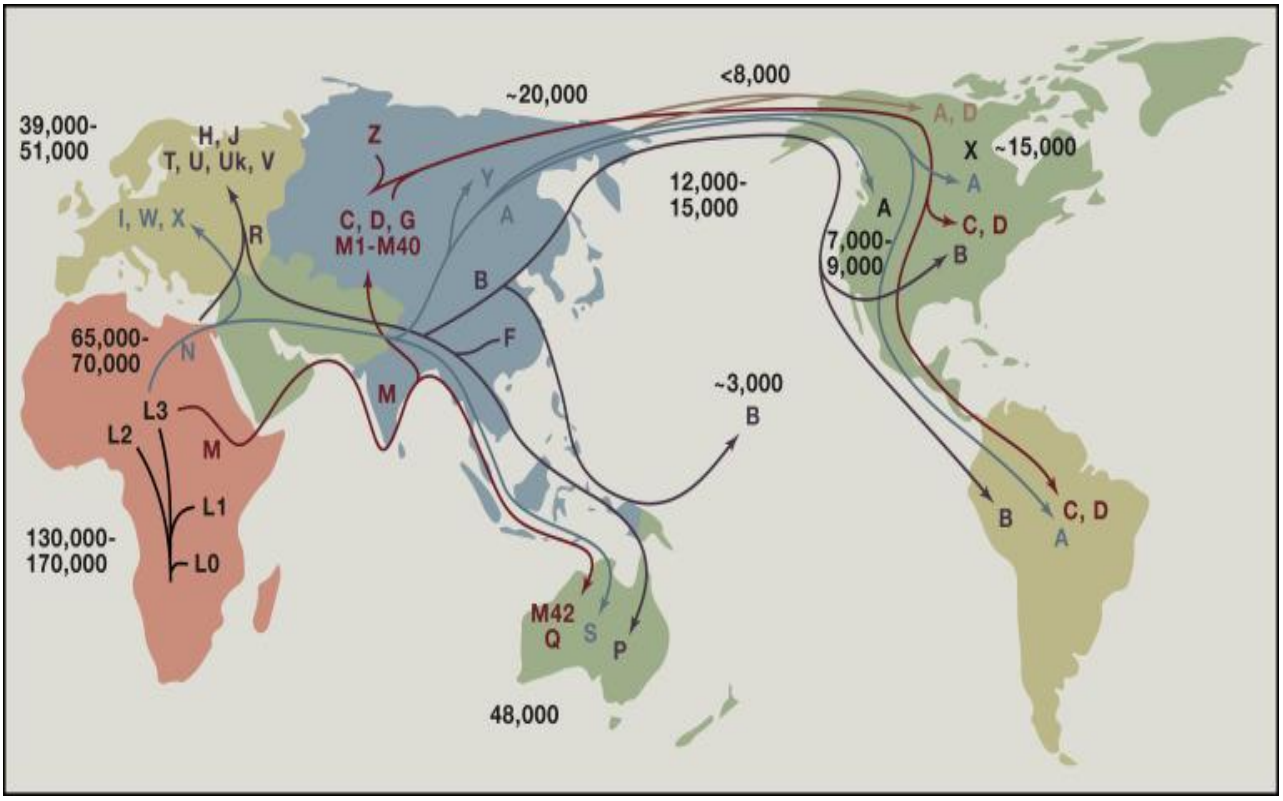

Figure 2.8. Regional Radiation of Human mtDNAs from their Origin in Africa and Colonization of Eurasia and Americas (Fig. modified from MITOMAP by Wallace, 2015).

Later, although, the whole mitochondrial genome sequencing is capable of providing reliable overall phylogeny based on averaged variation between sequences, it was noticed that estimation of dates, population size and genetic structure is greatly influenced by excess of mutations in the control region (Tamura, 1993). To overcome this, analysis of slow evolving mtDNA coding region was reported as a best alternative to the whole mtDNA sequencing. In support of this, the small non-recombining genome of the mitochondria have allowed a coding gene or sequence to answer specific evolutionary questions that would reflect patterns obtained from the whole sequencing (Elson, *et. al.*, 2004).

Although studies addressing sequence variation in mtDNA coding region have suggested that natural selection has significantly shaped the course of human mtDNA evolution (Mishmar, *et. al*., 2003), there is disagreement upon whether the distribution of specific human mtDNA haplogroups is due to an adaptation to different climates or if their distribution is a function of random drift assisted by purifying selection that eliminates non-synonymous changes (Kivisilid, *et. al*., 2006). The significance of a sequence displaying selective neutrality feature is its enhanced utility in estimating effective population size, one of the challenging and controversial activities in evolutionary genetics. Whole genome SNP data sets reflect features of selective neutrality in an averaged manner, as the intensities of selective signals vary widely between genomic regions. Despite a consensus on a large African effective population size as compared to populations from other continents, the extent of human genetic variation is still poorly understood in Africa particularly in Ethiopia. mtDNA coding gene is one of the most convenient tools in resolving questions of pattern of genetic diversity and population size given its extended haplotype structure, non-recombining nature and uniparental inheritance. These features combined are present no where else in the human genome, apart from the non-recombining portion of the Y chromosome, and hence make mtDNA coding gene variation ideal for human population genetics studies, for instance in tracing populations genetic history. Furthermore, inference made from the mtDNA hyper-variable control region or one of its coding genes are found to be highly correlated with whole mtDNA sequencing thus providing an added value for identifying population gaps in resource limited settings. In the current study, we explored the level of genetic diversity among Ethiopian populations utilizing *MT-COXII* sequences to gain insights into the extent of the contribution of Ethiopians to the genetic variation at regional and continental level.

3. MATERIALS AND METHODS

3.1. Ethical Aspects

The present study was approved by research ethics committees of Addis Ababa University, College of Natural Sciences, Department of Microbial, Cellular and Molecular Biology (MCMB). In addition, it was approved by National research ethics committee, Ministry of Sciences and Technology (MoST) of Ethiopia. Also written informed consent was obtained from each study subject. For this study, field breath hydrogen test was performed for LP/LNP phenotype screening and buccal cells were collected for human population genetic studies such as lactase genetics, population genetics affinities using Y chromosome and mtDNA.

3.2. The Studied Populations and Sampling Strategies

This study was conducted among two different regional states of Ethiopia; Gambella and Benishangul-Gumuz. Nuer, major ethnic group among the five tribes of Gambella region and Berta, Gumuz, Shinasha, Mao from Benishangul-Gumuz have been recruited based on the assumption that they are oldest from the perspective of human origin and also on the expectation that high occurrence of LP phenotype exists in this populations. They located South West and West North of Ethiopia respectively, along the Nile Valley. From evolutionary point of view, these areas look key strategic region and regarded as the corridors for the migration of human populations out of Africa. These groups speak Nilotic and Omotic languages and thus they represent groups that speak common languages belonging to the two major language families present and spoken in Ethiopia; Afro-Asiatic and Nilo-Saharan (Table 3.1).

55 individuals from Nuer, 30 individuals from each of Berta and Gumuz, and 20 individuals from each of Shinasha and Mao, all are above the age of 18 up to 60 years old were selected based on their full consent and interest. These make a total of 155 samples representing five different ethnic group and two different geographic regions of Ethiopia. Since this study involves human subjects, the number of participants were selected partially random (i.e., not purely random) and determined subjectively based on their availability and full interest, on the time, cost and other related factors. Numerous literatures have involved 20-30 numbers of human participants and reported that this numbers have a needed statistical power to detect sufficient genetic diversity. For instance, previous work on several ethnic groups of Ethiopia reported that the phenotype and genotype data generated from similar sample size would enables to capture the required diversity (Ingram, *et. al.,* 2007; Ingram, *et. al.,* 2009 and Jones, *et. al.,* 2013). This was considered as a framework to select 155 number of population size to see the expected phenotypic and genotypic differences (lactase genetics) and variation patterns (Y chromosome and mtDNA) among and within different selected ethnic groups. Each ethnic group in this study is considered as a population. The studies involve phenotype-genotype association of LP and study of genetic relationship of populations using Y chromosomes and mtDNA.

Figure 3.1 A map of Ethiopia showing the location of study populations and the two neighbors Sudan (www. Idpiuk.org): The stars indicate the approximate locations of study sites.

Table 3.1 Sample size, gender, socio-economic activities, linguistic Affilation and geographic location of the studied populations.

populations	Code	Sample size	Gender		Socio-economic activities	Lingustic Affilation	Region
			Male	Female			
Nuer	Nr	55	55	0	Pastorals	Nilo-Saharan	GM
Bertha	Bt	30	24	6	Mixed	Nilo-Saharan	BG
Gumuz	Gz	30	15	15	Mixed	Nilo-Saharan	BG
Shinasha	Sn	20	13	7	Mixed	Afro-Asiatic	BG
Mao	Mk	20	15	5	Mixed	Afro-Asiatic	BG
Total		155	124	31			

3.3. *MT-COXII* Global Sequences

In comparisons to the Ethiopian population, the global genetic diversity in representative world human populations was assessed. A total of 294 individuals representing most of the linguistic phyla were included in these groups. This sampling strategy attempts to avoid the bias inherent in selecting individuals based on the current world demography, such as current population size or geographic location. Selected populations were then grouped according to the continental location to simplify the analysis. Details of the geographical groups are illustrated (Table 3.2).

Table 3.2. mtDB Global Sequences

Groups	African	Asian	European	Americans	Australians	Total
No. of individuals	225	18	16	8	5	294

3.4. Breath Hydrogen Testing (BHT)

A total of 170 study subjects from five different ethnic groups were approached with the assistance of local elders and helper team and recruited. Then each volunteer study subject was asked to give their fully informed consent. The participants were provided with detail information about this study (appendix 1). The questions relating to environmental features (such as milk feeding habit, type of milk consumed, etc) and about personal health were assessed by questionnaire along with socio-demographic data. The consent form and health questionnaire were prepared in English and translated to Amharic. The Amharic versions were also further translated to respective local language in few cases with the help of local helpers as translator.

Then LP/LNP phenotype was tested using BHT for each individual sample. Each volunteer individual was requested to fast over night for 12 hours. To determine eligible participants fasting breath hydrogen measurements (base line reading) was taken, i.e., only subject with breath hydrogen $\geq 0 \leq 20$ ppm was considered. The base line reading excludes 15 individuals and they become ineligible to participate. This reduced the total population to 155. The test was performed after instructing each individual to orally ingest 50gm lactose dissolved in 250ml pure water. Breath hydrogen was measured using MicroH_2 meter as described by Ingram, *et. al.* (2007 or 2009) and the reading were recorded at every 30min. intervals after the ingestion of lactose

over a period of 3hrs. The individual was considered to be LNP when breath hydrogen or excretion is higher than 20 ppm as compared to the base line. But the individual was considered to be LP when breath hydrogen is less than 20ppm. Each subject was followed to see existence of post gastrointestinal illness over a period of 12 hrs after the test.

3.5. DNA Samples Collection

For DNA analysis, buccal swabs were collected from all 155 individual volunteer participants who had provided their informed consent and were eligible to participate. The buccal cells were used to obtain high quantities and quality genomic DNA for subsequent accurate analysis. Buccal swabs were collected from all study participants using a collection tube containing cap-attached cotton swabs. This was performed by brushing the cotton swabs firmly (pressing as possible without causing pain) ups and down against the inside of cheeks for approximately 30-40s. This was accompanied with rotating or twirling the swab while brushing it up and down, so that the entire surface of the swab tip comes in contact with the subject's cheek. This was done to obtain a sufficient amount of DNA. Upon collection, the buccal swabs sample was air dried for 15m and placed tip down directly in 15ml collecting tubes. Then the samples were kept in the ice box. These ice boxes with DNA samples were stored in the freezer until all the samples are collected from participants at the field. After completion of collection, the DNA samples were shipped within one day to the genetics research laboratory, Addis Ababa University, Ethiopia. From here the samples were transported to Molecular Biology Laboratory, Institute of Endemic Diseases, Medical campus, University of Khartoum, Sudan, where genomic DNA isolation and other molecular assay was performed. Since DNA is the most stable molecule, the transport process would not affect the swab analysis. The DNA samples were linked with health questionnaire and BHT data by code for lactase phenotype-genotype association study.

3.6. DNA Extraction

DNA extraction was performed using Guanidine Chloride method as described by Sambrook, *et. al*. (1990) with few modifications. The cotton tipped swab of each sample was cut and put into 1.5ml eppendroff tubes. To each sample, 500µl lysis buffer, 300µl guanidine chloride, 100µl ammonium acetate and 5µl proteinease K were added, mixed by shaking gently and incubated over night at 37^{o}C. Then the mixtures were allowed to cool at room temperature and transferred to another 15ml polypropylene falcon tubes. Then, 2ml of pre-chilled chloroform was added, vortexed and centrifuged at 3000rpm for 5min. The supernatant from each sample were carefully collected and transferred to another new falcon tubes to which 10ml absolute ethanol were added for concentrating, dehaydrating and desalting the DNA molecules. The tubes were inverted, shaken gently to precipitate the DNA and stored at -20 for overnight. The tubes were centrifuged at 3000rpm for 20min to obtain a DNA pellet. The supernatants were carefully discarded not to lose the pellet and the tubes were inverted on a tissue paper for 5min. To each tube, 4ml 70% ethanol was added to wash the DNA pellet and centrifuged for 20 min at 3000 rpm. The supernatant was discarded, and the pellet was allowed to air dry for 1.5 hours. The DNA pellet was then re-suspended in 100µl TE buffer and stored at 4^{o}c for overnight. The dissolved pellet was transferred to 1.5ml eppendroff & stored at -20 until use.

3.7. Agarose Gel Electrophoresis

After DNA extraction, agarose gel electrophoresis was performed to determine the presence or absence of genomic DNA in the involved samples. For this test, 1.5% agarose gel was prepared and used. 0.75gms of agarose powder was dissolved in 75ml of gel making mixtures (68dH2O + 7ml 10X TBE buffer). The mixtures were boiled in a microwave oven until a clear solution is observed and then was allowed to cool to 50^{o}c at room temperature. For later staining of the

DNA, 2µl EtBr was added to the mixture. The mixture was carefully poured in to a gel tray and the suitable size comb was placed correctly. The gel was left for 45min to set and solidify. The extracted DNA (3µl) was mixed with loading dye (3µl) and loaded into the running gel. The gel was run in a tap water for 20min and visualized by UV light, using SYNGENE^{R} Chemi Genius gel documentation system. Sample picture is shown below (Figure 3.2).

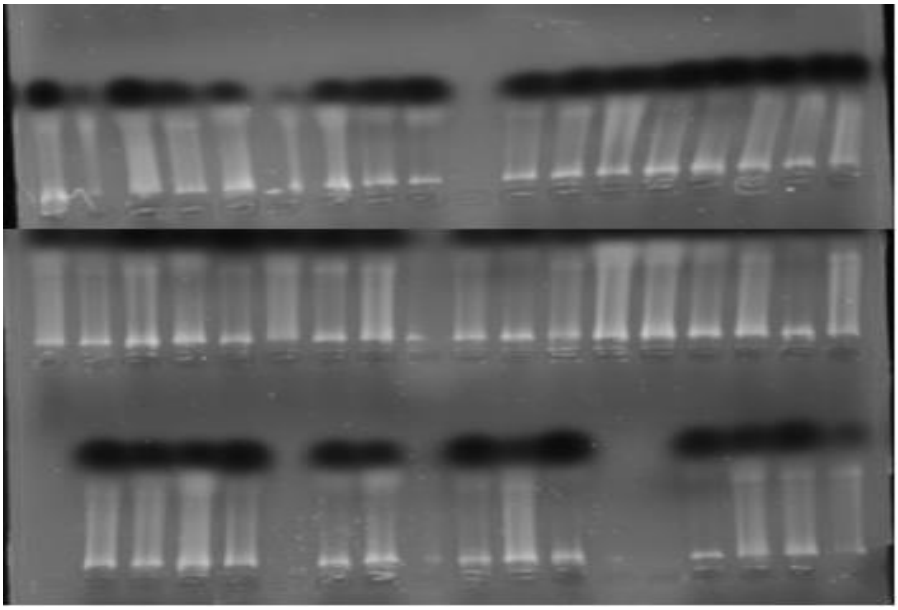

Figure 3.2. Sample gel pictures for Bertha, Sinasha, and Mao.

3.8. Quantification and Purity of Genomic DNA

The amount of DNA and quality was measured using a nanodrop device (ND-1000 Spectrophotometer). The device can measure 1µl sample with high accuracy and reproducibility. DNA concentrations were measured in ng/µl based on absorbance at 260 nm and the selected analysis constant (280 and 230). Firstly, the device was blanked using 2µl of sterile distilled water. Following brief vortexing the sample, 2µl of the DNA was loaded onto the lower device's

pedestal. After measuring, whereas the purity of the DNA samples was displayed on the computer connected to the device through wavelength ratios (260/280) and (260/230), the concentration was displayed as ng/µl (Table 3.1). Generally, approximately 1.8 - 2 was accepted as a pure sample for DNA at wavelength ratios for 260/280. From the present measurement, more or less purity values close to the standard 1.8 - 2 were obtained as shown below (Table 3.3) together with the samples concentration. Several different methods for detecting polymorphisms at marker regions were employed for both Y chromosomes and mtDNA. All these methods depend on PCR amplification of target DNA loci using primers specific for the loci under consideration.

Table 3.3. Sample quality and Concentration of the Genomic DNA for 120 samples

samples	ng/ µl	260/280	samples	ng/ µl	260/280	samples	ng/ µl	260/280	Samples	ng/ µl	260/280
Bt1	93.1	1.56	Gz1	57.8	1.65	Nr1	74.4	1.61	Sn1	102.4	1.7
Bt2	125.9	1.58	Gz2	283.5	1.73	Nr2	40.8	1.59	Sn2	110.2	1.75
Bt3	68.3	1.63	Gz3	123.1	1.6	Nr3	64	1.57	Sn3	132.1	1.67
Bt4	61.7	1.73	Gz4	456.7	1.7	Nr4	146.7	1.79	Sn4	104.8	1.67
Bt5	26.9	1.4	Gz5	171.1	1.54	Nr5	286	1.8	Sn5	78.9	1.67
Bt6	47.3	1.69	Gz6	344.3	1.83	Nr6	63	1.23	Sn6	47.7	1.74
Bt7	168.9	1.74	Gz7	294.8	1.55	Nr7	107.9	1.71	Sn7	135.1	1.63
Bt8	83.9	1.43	Gz8	228.9	1.59	Nr8	76.2	1.77	Sn8	74.9	1.85
Bt9	71.9	1.51	Gz9	97.5	1.53	Nr9	29.1	1.15	Sn9	93.3	1.57
Bt10	137	1.55	Gz10	165.2	1.66	Nr10	37.5	1.41	Sn10	3.6	0.72

Bt11	195.3	1.69	Gz11	161.2	1.43	Nr11	38.2	1.54	Sn11	144.6	1.66
Bt12	119.1	1.65	Gz12	135.5	1.33	Nr12	126.5	1.56	Sn12	35.7	1.66
Bt13	113.2	1.47	Gz13	179.1	1.51	Nr13	122.7	1.77	Sn13	-12.3	1.33
Bt14	109	1.68	Gz14	157.4	1.51	Nr14	91.9	1.68	Sn14	38.8	1.51
Bt15	215	1.75	Gz15	118.7	1.4	Nr15	3.7	1.18	Sn15	29.5	3.01
Bt16	143.5	1.79	Gz16	245.9	1.67	Nr16	269	1.48	Mk1	85.2	1.81
Bt17	128.7	1.64	Gz17	209.4	1.52	Nr17	25.4	1.67	Mk2	49.1	1.59
Bt18	143.6	1.79	Gz18	64.3	1.4	Nr18	26.4	1.71	Mk3	28.6	1.79
Bt19	55.7	1.72	Gz19	77.9	1.55	Nr19	114.7	1.65	Mk4	37.1	2.04
Bt20	198.5	1.66	Gz20	113.7	1.58	Nr20	66.8	1.65	Mk5	37.7	1.86
Bt21	107.3	1.64	Gz21	249.2	1.68	Nr21	78.9	1.76	Mk6	20.8	1.87
Bt22	76.6	1.59	Gz22	109.4	1.53	Nr22	39.3	1.53	Mk7	-6.4	3.13
Bt23	69.9	1.48	Gz23	145.6	1.68	Nr23	24.9	1.57	Mk8	96.8	1.83
Bt24	62.4	1.64	Gz24	142.5	1.64	Nr24	8.4	1.13	Mk9	191.6	1.66
Bt25	92.3	1.58	Gz25	162.6	1.75	Nr25	35.5	1.7	Mk10	44.2	1.75
Bt26	101.5	1.74	Gz26	281	1.78	Nr26	47.4	1.66	Mk11	67.2	1.79
Bt27	117.2	1.6	Gz27	181.6	1.57	Nr27	36.6	1.75	Mk12	69	8.45
Bt28	164.7	1.3	Gz28	111	1.6	Nr28	43.6	1.61	Mk13	38.3	2.05
Bt29	103.3	1.47	Gz29	157.9	1.83	Nr29	29.4	1.62	Mk14	80.4	1.55
Bt30	234.5	1.8	Gz30	40.4	1.78	Nr30	52.1	1.43	Mk16	35.6	1.48

3.9. Lactase Gene (LCT) - Enhancer Genotyping

As described above, LP phenotype was tested by HBT in 155 Ethiopian adults using MicroH_2 analyzer. Results are to be checked against genotyping results using PCR-RFLP and exome sequencing (work underway). So, the correlation of LP with genotype and the information from exome sequencing are examined in these populations after genotyping.

3.10. Y Chromosome Genotyping

The bi-allelic marker variation at Y chromosome specific polymorphism M1 (YAP), M13, M42, M60, M89 and M78 (Table 3.4) were assayed to produce male specific haplotypes. Firstly, the populations were categorized to YAP+ and YAP- using PCR amplification. Whereas YAP+ was further genotyped to investigate the distribution of M78 binary sub clades among the Ethiopian populations; using all the chromosomes carrying the E-M78 derived T-allele and C-allele, the YAP- further screened for the other binary markers (M13, M42, M60, M89). These male specific Y chromosome markers were genotyped using oligonucleotide primers (Table 3.4) and all genotypying work were done as follow.

3.10.1. YAP Genotyping

To genotype Y *Alu* polymorphism (YAP), touchdown PCR was used. Touchdown PCR comprises of two phases. The first phase consists of 5min initial activation at 95^{o}C. This was followed by 15 cycles of 20s denaturation at 94^{o}C, 45s annealing at 63^{o}C (temperature of annealing decreased by 0.5^{o}C per cycle), and 1min extension at 72^{o}C. The second phase consisted of 35 cycles, 20s denaturation at 94^{o}C, 45s annealing at 56^{o}C, 1.30min extension at 72^{o}C, and 5min final extension at 72^{o}C. Following optimization, the PCR condition for YAP genotyping was as following:

- 4 µl of Readymade master mix (consists of thermal Taq polymerase, 50mM MgCl2, in 10x PCR buffer and deoxynucleotide triphosphates, dNTPs)
- 1.5 µl of Forward primer (10um)
- 1.5 µl of Reverse Primer (10um)
- 3 µl of Genomic DNA
- 15 µl dH2O

These were mixed and made the final volume of 25 µl PCR mixtures. PCR was then carried out using SENSCQUEST thermal labcycler PCR machine.

Gel electrophoresis was performed for the determination of the size of the amplified DNA product and assign study participants as YAP + or YAP-. 2% Agarose gel was prepared by dissolving 1.5g of agarose in 75ml of 1X Tris Boric EDTA (TBE). The mixture was boiled in a microwave oven, and then was cooled to 50^{o}C at room temperature. About 2 µl of 10mg/ml EtBr was then added (for staining and later detection of the target DNA band). The mixture was poured into a gel tray with a suitable sized comb and left for 40min and allow solidified. The PCR products were loaded, and the gel was run in 1XTBE buffer for 45min and then visualized by UV light, using SYNGENE Chemi Genius gel documentation system. This enabled to score and categorize individual as YAP+ /YAP- according to the size of the PCR products of the same marker being tested and are depicted below (Figure 4.1).

3.10.2. Allele Specific- Polynerase Chain Reaction (AS- PCR)

Amplification refractory mutation system PCR (ARMS-PCR) primer used by Ye, *et. al*. (2001) was employed for allele specific PCR amplification for M78 marker. Similar primer designed by Hassan, *et. al.* (2008) using ARMS-PCR program was used to detect the ancestral (C) and derived (T) haplotypes. The program is found from the internet at (http://cedar genetics.soton.ac.uk/public_html/primer1html). PCR reactions were set up for two forward primers (F_1 or F_2) with common reverse (R) primer. Touch-down PCR was carried out to detect the presence or absence of the derived haplotypes. The PCR reactions were carried out using SENSCQUEST thermal lab cycler PCR machine. The PCR mix consisted of readymade master mix (containing thermal Taq polymerase, MgCl2, in 10x PCR buffer and dNTPs), genomic DNA, primers with sterile distilled water in the 25µl final PCR volume as follows:

Readymade master mix	4µl
Forward primer (10um)	1.5µl
Reverse Primer (10um)	1.5µl
Genomic DNA	3µl
dH_2O	15µl

Touch-down PCR program has been used for all Y chromosome markers as described above. The overall amplification success was determined by 2% agarose gel electrophoresis and presence or absences of C/T allele were scored for individual genotyped for M78.

Table 3.4. Oligonucleotide used to generate the different Y chromosomes haplogroups

PCR conditions			Polymorphism conditions		
Marker	Primer pairs (5' – 3')	Primer References	Enzymes	Ancestral alleles (bp)	Derived alleles (bp)
YAP	F:caggggaagataaagaaata R:actgctaaaaggggatggat	(Hammer and Horai, 1995)	-	YAP- (150)	YAP+ (450)
M13	F:tcctaacctggtggtctttc R:tgagccatgattttatccaac	(Underhill, *et. al*., 1997)	MboI (NEB)	G (156 + 77)	C (233)
M42	F:aaagcgagagattcaatccag R:ttttagcaagttaagtcaccagc	(Shen, *et. al*., 2000)	AluI (NEB)	A (340)	T (295+ 45)
M60	F:gcactggcgttcatcatctg R:atgttcattatggttcaggagg	(Shen, *et. al*., 2000)	MboI (NEB)	GATC (241 + 147)	1bp-insertion GTATC (389)
M78	F1:cacttaacaaagatacttctttcc F2:cacttaacaaagatacttctttct R:attactttcctaggttctccca	(Hisham, 2008)	-	C (319)	T (319)
M89	F:acagaaggatgctgctcagctt R:gcaactcaggcaaagtgagacat	(Akey, *et. al*., 2001)	NlaIII (NEB)	C (20 + 67)	T (87)

3.10.3. SNP Genotyping Using PCR-RFLPs

For Mutation M13, M42, M60 and M89, PCR-RFLP analysis was used for genotyping SNP genetic Markers. Touch-down PCR program was used to genotype these Y chromosome markers with the PCR conditions described above. In each case, the PCR was employed to flank the polymorphic site of interest and the PCR products were treated with appropriate restriction enzymes for digestion. MboI (NEB) was used to digest both M13 and M60 and as well NlaIII used for M89 digestion. AluI (NEB) was also used to digest M42 (Table 3.4 and Appendix 4). The digestion has been carried out as follows. The reactions were performed in PCR eppendrofs. 8.5µl of PCR product, 0.25µl of restriction enzymes and 1.25µl of 10xbuffer were mixed in a PCR eppendorf tubes. The tubes were caped and incubated at 37^{o}c for 2 hours in incubator. Gel electrophoresis was used to determine the presence or absence of the alleles (2% agarose gel was used to run the digestion product).

3.11. Mitochondrial DNA Genotyping

3.11.1. Mitochondrial Cytochrome C Oxidase II (*MT-COXII*) PCR

Mitochondrially encoded Cytochrome C oxidase II *(MT-COXII)* gene was amplified using conventional PCR using *COXII* forward and reverse primers (COXIIF and COXIIR) to obtain 623 bp fragment of the 684 bp region of the gene (Table 3.5 and Figure 3.3).

Table 3.5. Primer pairs of mitochondrial *COXII* genotyping

Primers	Primer sequence	Region	Primer Reference
COXII F	5'-TAGGTCTACAAGACGCTACTT-3'	623 bp	Elhassan, *et. al*., 2004
COXII R	5'-AATTAATTCTAGGACGATGG-3'		

Before actual amplification PCR conditions were optimized as following. PCR amplification was optimized using two samples from each of three populations (Berta, Gumuz and Nuer). They were checked for PCR amplification under two different PCR conditions. Firstly, 4µl of pre made mixtures (containing thermal Taq polymerase, 50mM MgCl2, 10x PCR buffer and dNTPs, gel loading buffer), 2µl of genomic DNA, 3µl (1.5µl each) of forward and reverse primers (both diluted 1:9 from 100µl stock). In the second case Maxime PCR PreMix Kit (i-Taq) for 25 µl reaction was used. Maxime PCR PreMix Kit (i-Taq) is the product what is mixed every component: 2.5U i-Taq^{TM} DNA Polymerase (5U/µl), dNTP mixture (2.5mM each), reaction buffer (1X) and gel loading buffer (1X), and so on - in one tube for 1 reaction PCR. To Maxime PCR PreMix, 2µl template DNA, 1µl of each of *COXIIF* primer and *COXIIR* primer were added.

Under both case sterile distilled H_2O was added to a total volume of 25µl (16µl and 20µl respectively). The amplification was performed in 35 cycles following a 95^{o}C preheating for 5min. in thermal cycler (PCR machine). Each cycle employed the following PCR temperature profiles: DNA denaturation at 94^{o}C for 30s, primer annealing at 60^{o}C for 1min, primer elongation at 72^{o}C for 1min. Final extension cycle at 72^{o}Cfor 5min. was used to help finish the elongation of many or most PCR products initiated during the last cycle of the PCR. To ensure the amplification of the target gene sequence, PCR products and 100bp ladder were electrophoresed in 1.5% agarose gel prepared as described earlier elsewhere. Best result was obtained with the second condition and this condition was used for PCR amplification of *MT-COXII* gene. The gel picture of PCR optimization is depicted in figure 3.4 below.

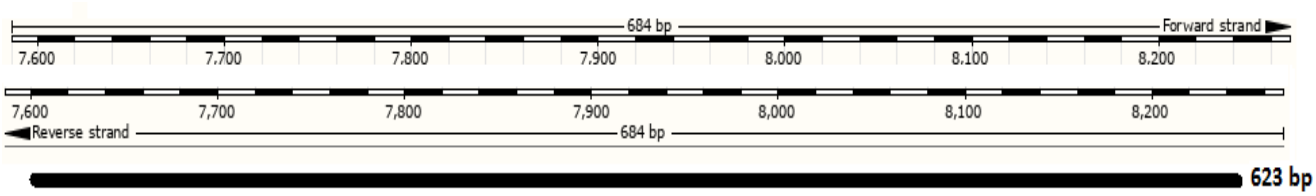

Figure 3.3. Cytochrome C Oxdase II location of mitochondrial Chromosome from 7586-8269 (Ensembl *Homo sapiens* version 91, 2017).

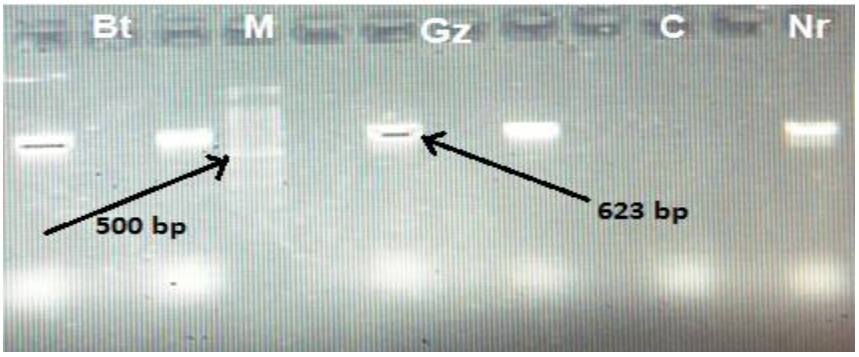

Figure 3.4. PCR Optimization for *MT-COXII* gene using Berta, Gumuz and Nuer, each two samples. The PCR product and 100 bp marker (M) were run on 1.5% agarose gel.

3.11.2. *MT-COXII* Sequencing

Following optimization, total genomic DNA was amplified by PCR using oligonucleotide primers (Table 3.5) specific for the COXII gene under the PCR condition of the second case, i.e., using Maxime PCR PreMix Kit (i-Taq) as described above to produce double stranded DNA. The amplification was successful for 30 PCR reactions out of fifty samples. Since some of these positive PCR products was found to contain multiple bands (may be due to high heteroplasy nature of mitochondrial gene), the products were sent for gel extraction sequencing at Macrogen, Inc. (Korea).

3.11.3 Sequence Quality Assessment

Sequences were aligned using Bioedit Software (Hall, 1999). Only sequences with clear electropherogram peaks and with back ground an almost imperceptible were used for analysis. Out of 30 original Ethiopian sequences, only 25 were included for *MT-COXII* sequence analysis. All electrophenogram peaks of the Mao sequences were not good and unclear and thus they were all excluded from the analysis. Sequences with gaps and double peaks were excluded and only clean genomic region was included in analysis after both end were trimmed. Both forward and reverse strands were used to check the correctness of the detected mutation.

3.12. Data Analysis

Lactase Persistence Phenotype diversity

Phenotype data were analyzed statistically using the IBM SPSS statistics software version 20 for screened individuals with and without LP trait.

Y chromosome Phylogenetic Trees

Y chromosome phylogenetic tree of the five Ethiopian populations and their haplogroups was designed by hand according to updated version (Karafet, *et. al.,* 2008) of YCC nomenclature (YCC, 2002). Neighbor-joining tree of Y chromosome was generated using Genetic Data Analysis (GDA) and tree view software based on Nei (1978) genetic distance. GDA is a population genetic program (Weir, 1990) and available online.

Mitochondorial DNA Sequence Alignment

A total of 269 *MT-COXII* DNA sequencs from Africans, Middle-East, Asia, Australians, Europeans and Americans were obtained from public mtDB database (http://www.mtdb.igp.uu.se/). These combined with the 25 *MT- COXII* DNA sequences from

present samples, making 294 in total, were aligned using the program CLUSTALX version 1.81 (ClustalX, 1997) on the sequence alignment editor Bioedit software (Hall, 1999).

BioEdit is biological sequence alignment editor software which can be accessed from the website: http://www.mbio.ncsu.edu/BioEdit/bioedit.html. Alignments were double-checked and refined manually. Yoruba reference sequence from African (Ingmann, *et. al*., 2000) and reconstructed Sapiens Reference Sequences/ RSRS (Behar, *et. al.*, 2012) were used separately as a reference sequence when aligning mtDNA sequences. *MT- COXII* sequence alignment with both references showed complete similarity and both could be used for sequence alignment. Since the comparision revealed complete similarty RSRS reference sequence was used as a reference sequence for the alignment. This reference sequence was used to detect the existing mutation among the presently sequenced samples

Genetic Distances

The aligned *MT-COXII* sequences were imported into MEGA version 5.0 for estimation of nucleotide sequence divergence between groups. *MT-COXII* genetic distances between populations were calculated based on Tajima's Nei model (Tajima and Nei, 1993). MEGA is an integrated tool for an automatic and manual sequence alignment, inferring phylogentic trees, mining web-based databases, estimating rates of molecular evolution (Kumar, *et. al.*, 2004). The program is freely available at: http://www.megasoftware.net/. Sites with alignment gaps and missing data were omitted from the analyses. A phylogenetic trees analysis was performed to examine the relationships among individuals. To evaluate significance levels and the consistency of nodes (tree topology) derived from the phylogenetic analysis, 1000 bootstrap replications of the original data set were performed.

Molecular Diversity and Neutrality Test

Molecular genetic diversity indices were estimated by computing the mean nucleotide composition, number of transitions, transversions, indels, polymorphic sites (Nps), number of segregating sites (S), haplotype (H) and nucleotide (π) diversities, and mean number of pairwise differences (M) and average number of nucleotide differences (K). These and haplotypes diversity were identified using DNASP version 5.10 (DnaSP, 2009). Test for deviations from expectations of neutrality, including Tajima's D, D* of Fu and Li, and F* of Fu and Li was performed using the same DnaSP software. Mismatch distribution between individuals was performed based on *MT- COXII* sequences to locate such an expansion in time among Ethiopian populations using ARLIQUIN 3.1 software. This is a program for population genetics analysis (Excoffier, *et. al.,* 2005). Mismatch distribution were not performed for Y chromosome data due to the small number of polymorphisms. Pair wise distances for Y chromosome and mtDNA were performed to describe the short term genetic distance between populations, using Markov chain within 10,000 steps.

Mitochondrial DNA Phylogenetic Tree

A neighbor-joining (NJ), Unweighted Pair Group Method Using Arithmetic Average (UPGMA), minimum Evolution (ME) trees were constructed for four Ethiopian populations in present study and other populations using MEGA5 (Tamura, *et. al.,* 2007). This was based on pair wise genetic distances of 25 *MT-COOXII* sequences. Similar trees were constructed by combining the sequences of the present populations with other Ethiopian, African and other countries. This was based on pair wise genetic distances of 166 *MT-COOXII* sequences. To convert the raw sequence data into a genetic distance matrix; individual sequences were grouped in to the different groups of populations, and pair wise genetic distances were calculated based on the

model of Tamura and Nei (1993) using bootstrap values (percentage of 100 total bootstrap replicates).

Analysis of Molecular Variance (AMOVA)

AMOVA test was performed to verify statistical differences between groups of populations. Using AMOVA, population pairwise F_{ST} values between populations were calculated. Fixation indices significantly different from zero were identified by comparison with the results of 10,000 data permutations. Both haplotypes frequencies and molecular differences of Y chromosome and mtDNA among haplogroups were taken into account. The populations were grouped in to two linguistic families. All calculations were performed using the ARLIQUIN version 3.1.

Principal Component Analysis (PCA)

Principal Coordinate Analysis (PCA) was used to plot the major patterns within the data set using population pair wise F_{ST} (Distance matrices). Two run of PCA were performed for F_{ST} genetic distance matrices of Y chromosome and mtDNA by employing PAST (paleontological statistics) algorithms, version 2.11 (Hammer, *et. al.,* 2001). The first run was performed for Ethiopian samples from the present study, whereas in the second run published data of other populations including Ethiopians, Sudanese, Senegalese (Semino, *et. al.,* 2002 and Hassan, *et. al.,* 2008) were used alongside population data from the present study to generate PC plots of Y chromosome. Also, two runs of PCA were performed for mtDNA data, the first for Ethiopian populations from the present study and the second to compare between Ethiopian populations with other African populations. Some of the populations used in the comparisons are single tribal groups, whereas others are pools of populations. The first PCA and the second PCA percentages were calculated and plotted on X and Y axis respectively.

4. RESULTS

4.1. The Phenotype of LP

One aspect of this study was to examine the diversity of LP phenotype in five ethnic groups of Ethiopia. The adult LP and LNP were determined by HBT using the Micro H_2 analyzer. In general, 83 (53.55%) and 68 (43.87%) of the total 155 study subject had a positive breath (≥20 ppm) and negative breath test (<20 ppm) respectively. Individuals with positive breath test are LP, while those with negative breath test are considered as LNP. Of 155 participants, 4 (2.58%) were NHP. In particular, the detail frequencies of LP, LNP and NHP in studied populations are indicated below (Table 4.1). The low frequencies of lactase unknown in the Nuer disagree with the hypothesis of natural selection in favor of the "lactase persistence gene" in milk-dependent pastoralists. Also, this could be due to some samples may be Anuak although they claim that they are Nuer.

Table 4.1. Frequency of LP, LNP and NHP in Studied Populations

Phenotyped Populations	No. of Samples	LP	%	LNP	%	NHP	%
Nuer	55	30	54.6	24	43.6	1	1.8
Berta	30	18	60	12	40	0	0
Gumuz	30	15	50	15	50	0	0
Shinasha	20	10	50	9	45	1	5
Mao	20	10	50	8	40	2	10
Total	155	83	53.5	68	43.87	4	2.5

Milk drinking habit and other health related information were assessed using questionnaire. Each individual was asked whether they drink milk daily or not and see the correlation. The lactase phenotype frequency was higher in the milk drinkers on the daily basis than the non-milk drinkers. In addition, post gasterointestinal symtoms were assessed. In all groups, the majority of the LNP and very few LP individuals showed various post gasterointstinal symptoms.

4.2. Y chromosome Genotyping

In this study, five Y chromosome binary markers were analyzed. Initially all individuals were genotyped for the YAP marker. At the beginning, they were categorized as YAP+ or YAP-. YAP+ individuals were further genotyped for haplogroupE sub clades. These polymorphisms were investigated by allele specific PCR amplification to obtain T/C SNPs. Similarly, YAP- individuals were analyzed for four major binary markers of M13, M60, M42 and M89 by PCR amplification followed by restriction enzymes mediated digestion (PCR-RFLP method). Thus, the mutations or SNPs in the fragments amplified by PCR were subsequently detected by RFLPs analysis and checked for the presence of haplogroups A, B and F. Y chromosome haplogroup affinity was determined according to the most recently updated phylogeny by such genotyping in hierarchical order (Karafet, *et. al.,* 2008). The nomenclature was according to the Y chromosome Consortium rules and its updated version (YCC, 2002 and Karafet, *et. al.,* 2008).

Sample picture of 2% gel electrophoresis containing the PCR product of YAP+ and YAP- (for Berta and Nuer) is illustrated below (Figure 4.1). The NRY region was successfully amplified as a 150 bp for YAP- (deletion) and 450bp forYAP^{+} (insertion). The absence of a negative control band indicated that the PCR is free from contamination. All samples carrying YAP + were further typed for E-M78 carrying derived allele C/ T and the PCR amplification products are indicated below (Figure 4.2). However, all YAP- samples were genotyped for binary markers

using PCR-RFLP analysis. The samples of gel electrophoresis picture of the results of RFLP genotyping of Y chromosome specific haplogroup (M42 and M60) are depicted in figure below (Figure 4.3 and 4.4). Of 124 male individuals where genotyping was performed, 4 individuals did not yield any amplified product for YAP+ and YAP- analysis.

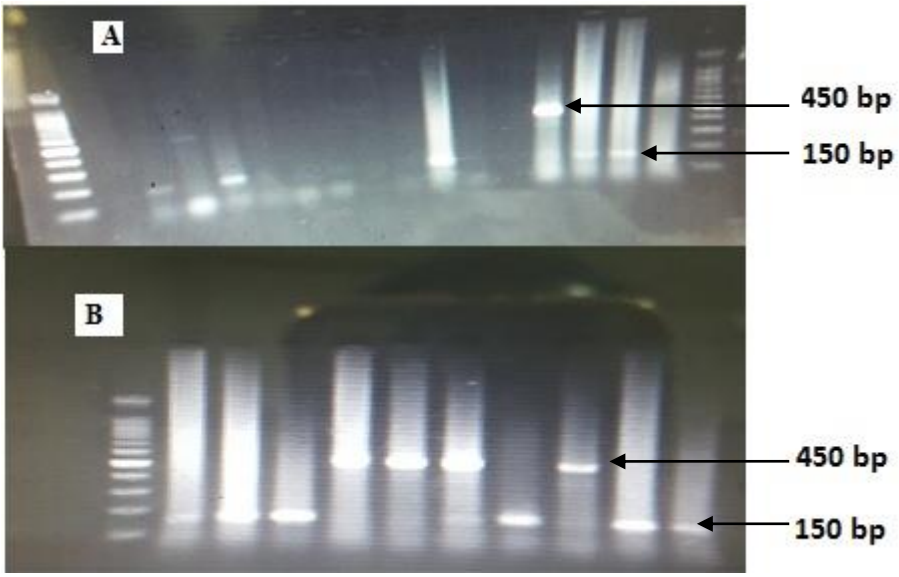

Figure 4.1. PCR Amplification of YAP locus to detect the presence or absence of *Alu* insertion. A) Bertha B) Nuer. 100bp DNA marker used. YAP+ is 450 bp and YaAP- is 150 bp. The second lane (A) and the last lane (B) are negative control. The products were run on 2% agarose gel.

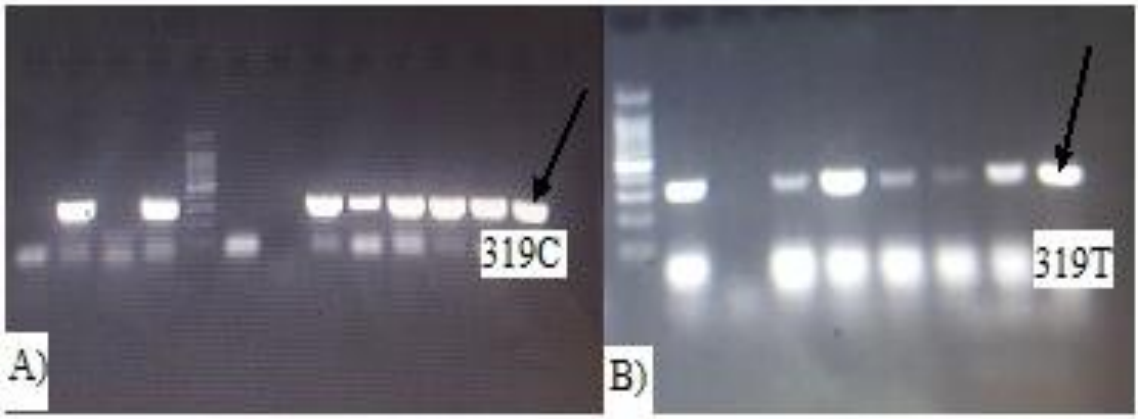

Figure 4.2. Sample PCR Amplification of YAP+ for M78. A) M78 carrying ancestral allele C B) M78 carrying derived allele T. 100bp marker used. The product were run on 2% gel.

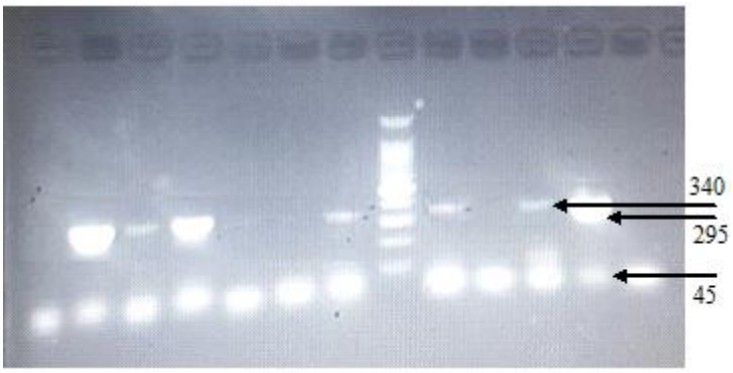

Figure 4.3. Sample RFLP Analysis of YAP- for M42. Lane 1, negative control, lane 8, 100bp marker.The product were run on 2% agarose gel.

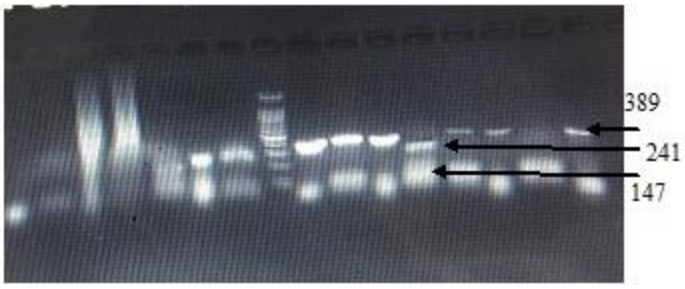

Figure 4.4. Sample RFLP Analysis of YAP- for M60. Lane one (from right to left) negative control, lane 8, 100bp marker. The product were run on 2% agarose gel.

Y chromosome Variation

Y chromosome haplogroup frequencies in five Ethiopian populations are shown below (Figure 4.5) according to YCC (2002) and the updated version (Karafet, *et. al*., 2008) nomenclature. In present study, all genotyped haplogroups (A, B, E and F) were observed among Western Ethiopian populations with percentages of 44.04, 36.9, 86.11and 21.42 % respectively. In general, although haplogroups E-M78 was by far the highest, both A-M13 and B-M60 were also present at such high frequencies in the present study. As compared to the Afro-Asiatic speaking groups, they were found at high frequencies in Nilotic speaking groups. More specifically, of the Nilo-Saharan speakers, haplogroup A-M13 was found at higher frequencies in Nuer with the frequency of 37.83% followed by Gumuz with 18.92%. In similar manner B-M60 was found in Nuer with frequency of 36.36% followed by Berta and Gumuz each with 21.21%. Moreover, E-M78, which is by far the most mutationally diverse of all major Y chromosme haplogroups, is present at the highest fequencies in Nuer (51.62%) and followed by Berta and Gumuz both with 19.35% and they all are Nilotics. Y chromosome data from which the mentioned frequenceise are calculated is illustrated in below (Table 4.2). Unexpectidely, haplogroup F-M89 were found in Nilo-Saharan speaking groups with high frequencies in Berta (55.55%) and Gumuz (33.33%). This haplogroup were present at low frequencies in Nuer, but completely absent from the Afro-Asiatic speaking groups. Y chromosome haplogroup variation and distribution among the five populations of the present study and few neighbour populations is illustrated below (Figure 4.5). Moreover, haplogroup B-M42 from which B-M60 desended was also genotyped and found to be present in all studied groups

E-M78 subclades

Haplogroup E-M78 which is defined by E1b1b1a was found in present study. It is the terminal branch of E3b1 that was found in Sudan population by Hassan, *et. al*.(2008). Haplogroup E1b1b1a (E-M78), previously called E3b1according to YCC (2002) was found at highest frequency in Nuer of this study with frequency of 51.62%, but was low in Nuer from Sudan (Hassan, *et. al.*, 2998). This result is in agreement with the original thought that E-M78 and its sub clades have originated in North East Africa (Cruciani, *et. al.,* 2007). It's distribution in studied Ethiopian group is highest than high frequency reported in Europe, middle east and East Africa. This may suggest the Ethiopian origin of this subclades although further detail anlaysis with more representative sample are required for confirmation. It has also higher frequency in other Nilo-saharan speaking groups, Bertha and Gumuz both having with similar 19.35%.

Table 4.2. Comparison of Frequency of Y chromosome Haplogroups (which is converted toY chromosome phylogenetic Tree)

Studied populations	Number of samples	YAP+	YAP-	M78 (E)	M13 (A)	M60 (B)	M89 (F)
Nuer	50	20	30	16	14	12	2
Bertha	22	7	15	6	5	7	10
Gumuz	23	6	17	6	7	7	6
Shinasha	11	2	9	2	4	3	0
Mao	14	1	13	1	7	4	0
Total	120	36	84	31	37	33	18

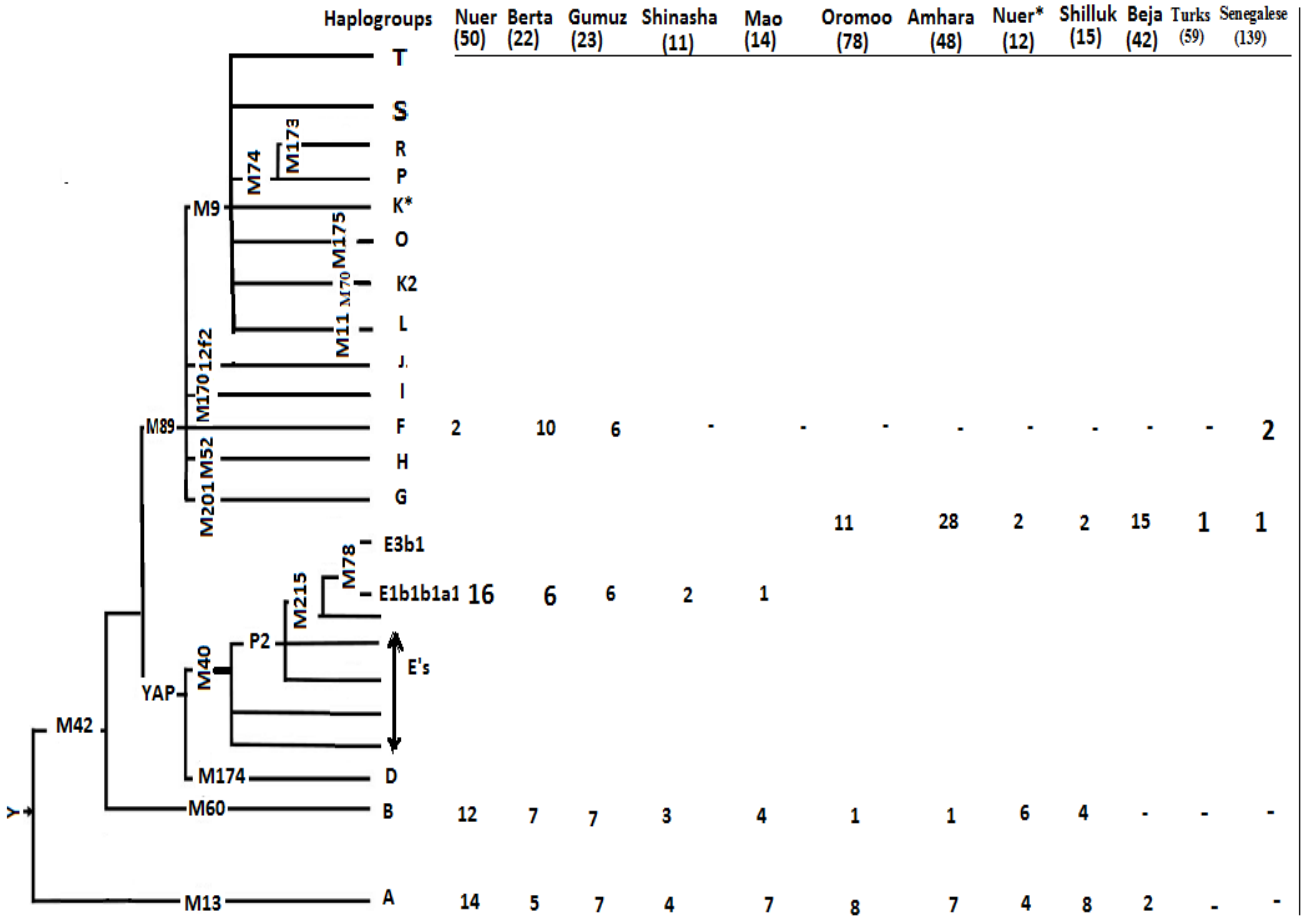

Figure 4.5. Phylogenetic distribution of the Y chromosome haplotypes and their frequencies in five Ethiopian populations in the present study, compared with three Sudanese (Hassan, *et. al.*, 2008), two Ethiopian ethnic groups (Amhara and Oromoo) and Senegalese (Semino, *et. al.*, 2002) Turks (Sanchese, *et. al.*, 2008). Numbering of Mutations and haplogroups nomenclature are according to YCC (2002) rules and updated by Karafet, *et. al.* (2008). Nuer* is those belong to Sudan. The arrow shows the root of the tree.

4.3. *MT-COXII* Genotyping

Affinities within mtDNA haplotypes was first inferred through the sequencing of a fragment of 623bps from the *MT-COXII*, one of mtDNA coding region. Sequence analysis of this coding region was performed by using primer COXIIF 5'-TAGGTCTACAAGACGCTACTT-3' and primer COXIIR 5'-AATTAATTCTACGATGG-3'. All the samples were sequenced both direction starting from primer *COXIIF* and primer *COXIIR*.

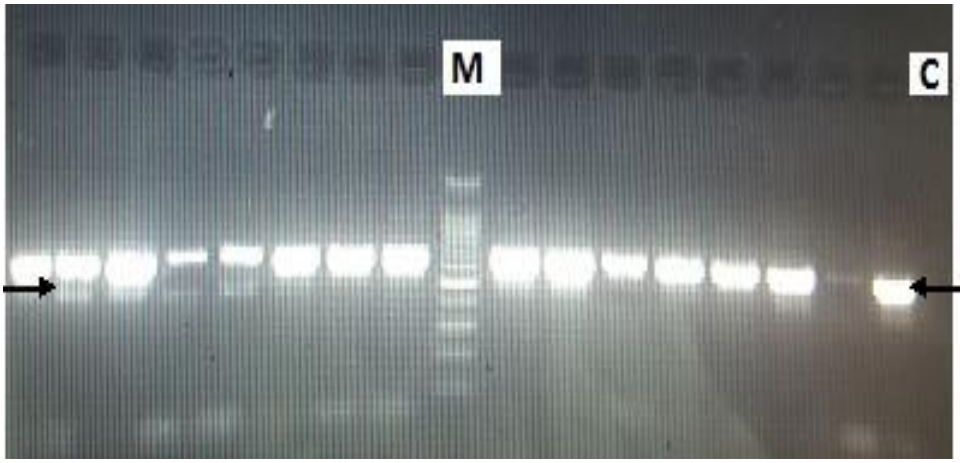

Figure 4.6. PCR Amplification of the *MT- COXII* gene (623bp). 100 bp marker (M) was used. Lane 1(C), is negative control. Multiple bands were observed for some *MT-COXII* positive samples. Arrows show additional and target bands. The product were run on 1.5% agarose gel eelctrophoresis.

SNPs Scoring and Haplotypes Frequencies of *MT-COXII* in Studied Populations

Sequences were aligned using BioEdit software (Hall, 1999) and Youruba refrence sequences (Ingman, *et. al*.,2000). From aligned sequences, mitochondrial sequence variants were visually counted in 25 Ethiopian sequences. Overall, fifteen different SNPs were identified of which, 12 (80%) were Transitions (synonymous substitutions) where as 3 (20%) were Transversions

(nonsynonymous substitutions). Thus, there were a total of 3 nonsynonymous (NS) and 12 synonymous (S) polymorphisms, yielding an NS/S ratio of 0.25. This indicates that transition is representing the majority of mutations observed in the present study. Out of the 15 mutations, 10 were previously reported in MitoMap data base and 5 were new mutations observed in this study (Table 4.3). Relatively the Nilo-Saharan groups exhibited highest number of synonymous substitutions and lower number of nonsynonymous mutations than the Afro-Asiatic groups.

The shared haplotypes and haplotypes frequencies among Ethiopian populations and other Africans and non-African populations were calculated. This was performed using Arlequin version 3.1 (Excoffeir, *et. al*., 2005) software. A total of 25 sequences from four ethnic groups representing two sampling locations were obtained in this study of Ethiopian group. Mao group was omited from sequence analysis because of low quality of the *MT-COXII* sequence output. Thrteen (13) haplotypes with 138 polymorphic sites were observed among Ethiopian populations with haplotype diversity, Hd: 0.9233. But, fewer shared haplotype were observed among groups of the populations. Where as haplotype 4 has shared bewteen Gumuz and Shinasha, Nuer has shared haplotype **1** with Berta and Gumuz. Haplotype **4** scored higher in Gumuz (66.7%) and in Shinasha (60%), followed by haplotype **7** in Nuer (44.44%) and haplotype **2** in Berta (37.50%). Haplotype frequencies distribution in populations from which these percentages are calculated is shown below (Table 4.4).

Table 4.3. The score and assignemnt of Mutation in *MT-COXII* sequences of the studied Ethiopian groups.

Mutation position	Substitutions:Transition/Transversion		MITOMAP database
7681**(C>T),** (C>A)	Transition, Transversion		Reported (both)
7702 (G>A)	Transition		Reported
7771 (A>G)	Transition		Reported
7849 (C>T)	Transition		Reported
7861 (T>C)	Transition		Reported
7910 (G>C)	Transversion		**New**
7972 (A>G)	Transition		Reported
7997 (G>A)	Transition		**New**
7999 (T>C)	Transition		Reported
8006 (C>T)	Transition		**New**
8010 (T>A),(T>C) *	Transversition (Transition)		**New**, Reported
8013 (T>A), (T>G)	Transversition		**New**, Reported
8047 (T>C)	Transition		Reported
8155 (G>A)	Transition		Reported
8188 (A>G)	Transition		Reported
15	12 Transitions = 80%	3Transversions = 20%	

*- was reported to be associed with diseases like develpmental delay, ataxia,seizure, hypotonia and lactic acidosis (Tang, *et. al*., 2013)

Table 4.4. mtDNA Haplotypes frequencies in four Ethiopian population

Haplotype	Nuer (9)	Berta (8)	Gumuz (3)	Shinasha (5)
Hap_1	1	1	1	0
Hap_2	0	3	0	0
Hap_3	2	0	0	0
Hap_4	0	0	2	3
Hap_5	1	0	0	0
Hap_6	0	0	0	1
Hap_7	4	0	0	0
Hap_8	1	0	0	0
Hap_9	0	1	0	0
Hap_10	0	0	0	1
Hap_11	0	1	0	0
Hap_12	0	1	0	0
Hap_13	0	1	0	0

Similarly, a total of 294 global *MT-COXII* sequences from mtDB (available on line at: www.mtdb.ipg.uu.se), including the present were analyzed using Arlequin to look at inter continental and inter populations diversity (Appendix 6). With regard to haplotype distribution, **66** haplotypes were detected with haplotype diversity, **Hd**: 0.7923. All haplotype was shared by Africans, except haplotype **62** and **63** that are only shared by Asians and Australians. Using these haplotype frequencies, relative friquencies for global popupations were calculated. Of all, haplotype **1** was shared at higher level by Africans (38.20%), Asians (66.70%), Americans (87.50%), Europeans (56.50%) and Australian (80%). In general high diversity of SNPS and haplotypes than the world populations was observed in present study, result tracing maternal ancestry back to Africa.

4.4. Genetic Distances

Table 4.5 and 4.6 show genetic distances of Y chromosome and *MT-COXII* variations respectively. For Y chromosome, genetic distances were greater than zero for all pair wise comparisons. Whereas three out of ten pair comparisons (30%) were significant, seven (70%) were not significant. Most of the non-significant genetic distances occur between populations of the same linguistic groups. For *MT-COXII* genetic distance, three pair wise comparisons (50%) were greater than zero and three other comparisons (50%) were equal to zero.

Table 4.5. Y chromosome Pair wise Genetic Distances of Different five Ethiopian Populations. List of labels for population samples used in this table: 1: Nuer, 2: Berta, 3: Gumuz, 4: Shinasha and 5: Mao.

	1	2	3	4	5
[1]	0.000				
[2]	0.036	0.000			
[3]	0.021	0.049	0.000		
[4]	0.003	0.034	0.019	0.000	
[5]	0.133*	0.159*	0.143*	0.130*	0.000

*Significant at multiple tests adjusted P-Value

Markov chain length: 10,000 steps

Table 4.6. *MT-COXII* Pair wise Genetic Distances of Different four Ethiopian Populations. List of labels for population samples used in this table: 1: Nuer, 2: Berta, 3: Gumuz, 4: Shinasha.

	1	2	3	4
[1]	0.000			
[2]	0.099	0.000		
[3]	0.072	-0.107	0.000	
[4]	0.133	-0.034	-0.126	0.000

Matrix of significant F_{ST} P values; Significance Level=0.0500

Similarly, the genetic distances of Y chromosomes and *MT-COXII* of Ethiopian populations were compared with some Ethiopian and some neighbor Africans and other populations as shown in Table 4.7 and 4.8, respectively.

Table 4.7. Y chromosome Pair wise Genetic Distances of Five Ethiopian populations compared with some other Ethiopian and Neighbor Africans. List of labels used in this table: 1: Nuer, 2: Berta, 3: Gumuz, 4: Shinasha, 5: Mao, 6: Oromo, 7: Amhara, 8: *Nuer, 9: Shilluk, 11: Beja**

[	1	2	3	4	5	6	7	8	9	10	11
[1]	0.0000										
[2]	0.0075	0.0000									
[3]	0.0037	0.0075	0.0000								
[4]	0.0019	0.0056	0.0019	0.0000							
[5]	0.0037	0.0075	0.0000	0.0019	0.0000						
[6]	0.0056	0.0093	0.0056	0.0037	0.0056	0.0000					
[7]	0.0019	0.0056	0.0019	0.0000	0.0019	0.0037	0.0000				
[8]	0.0037	0.0075	0.0037	0.0019	0.0037	0.0056	0.0019	0.0000			
[9]	0.0037	0.0075	0.0037	0.0019	0.0037	0.0019	0.0019	0.0037	0.0000		
[10]	0.0000	0.0075	0.0037	0.0019	0.0037	0.0056	0.0019	0.0037	0.0037	0.0000	
[11]	0.0037	0.0075	0.0037	0.0019	0.0037	0.0019	0.0019	0.0037	0.0000	0.0037	0.0000

*Nuer** belong to Sudanese group.

Table 4.8. *MT-COXII* Pairwise Genetic Distances of present Ethiopian populations compared with some other Ethiopian and Neigbour Africans and other populations. List of labels used in this table: 1: Hausa, 2: Nilotes, 3: Beja, 4: Nubian, 5: Nuba, 6: Ethiopian, 7: Ertrean, 8:Egypt, 9: Morocco, 10: Nigeria, 11:South Africa, 12:Tunisia, 13: San, 14; Pygymy, 15: Yoruba, 16: Saudi, 17: Yemen, 18: Israel, 19: Iraq, 20: Indian, 21: Chinese, 22: Japanese, 23: Ethiopian (present).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	0.00000																						
2	-0.85714	0.00000																					
3	-0.33333	-0.11702	0.00000																				
4	-0.27273	0.15625	0.22535	0.00000																			
5	-0.33333	0.05303	0.12281	0.18750	0.00000																		
6	0.00000	-0.21495	0.08046	0.11672	0.08046	0.00000																	
7	-1.00000	0.19631	0.42976	0.48454	0.45034	-0.32353	0.00000																
8	-0.82857	0.07527	0.14667	0.29139	0.26356	-0.24675	0.04245	0.00000															
9	1.00000	0.23529	0.50000	0.17647	0.36842	1.00000	0.90476	0.39048	0.00000														
10	0.00000	0.07143	0.33333	0.36364	0.33333	0.00000	-0.08434	-0.03337	1.00000	0.00000													
11	0.00000	-0.85714	-0.33333	-0.27273	-0.33333	0.00000	-1.00000	-0.82857	1.00000	0.00000	0.00000												
12	0.00000	-0.85714	-0.33333	-0.27273	-0.33333	0.00000	-1.00000	-0.82857	1.00000	0.00000	0.00000	0.00000											
13	0.00000	0.07143	0.33333	0.36364	0.33333	0.00000	-0.08434	-0.03337	1.00000	0.00000	0.00000	0.00000	0.00000										
14	0.00000	0.12641	0.33891	0.31653	0.30469	0.36842	0.47661	0.15556	0.75000	0.63415	0.00000	0.00000	0.63415	0.00000									
15	-0.80000	0.03150	0.13084	0.28532	0.24272	-0.20000	0.15523	-0.03226	0.55000	0.06250	-0.80000	-0.80000	0.06250	0.25000	0.00000								
16	1.00000	0.23304	0.46488	0.36652	0.37824	1.00000	0.80769	0.31915	1.00000	1.00000	1.00000	1.00000	1.00000	-0.20000	0.45455	0.00000							
17	-1.00000	0.00794	0.10714	0.21687	0.17808	-0.29032	0.11901	-0.00977	0.25000	0.00000	-1.00000	-1.00000	0.00000	0.09824	-0.02632	0.22830	0.00000						
18	0.16667	0.39474	0.50000	0.47826	0.47619	0.39394	0.66507	0.43637	0.54545	0.58333	0.16667	0.16667	0.58333	0.51163	0.48780	0.56044	0.25676	0.00000					
19	0.00000	-0.85714	-0.33333	-0.27273	-0.33333	0.00000	-1.00000	-0.82857	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.80000	1.00000	-1.00000	0.16667	0.00000				
20	0.00000	-0.21495	0.08046	0.11672	0.08046	0.00000	-0.32353	-0.24675	1.00000	0.00000	0.00000	0.00000	0.00000	0.36842	-0.20000	1.00000	-0.29032	0.39394	0.00000	0.00000			
21	0.00000	0.07143	0.33333	0.36364	0.33333	0.00000	-0.08434	-0.03337	1.00000	0.00000	0.00000	0.00000	0.00000	0.63415	0.06250	1.00000	0.00000	0.58333	0.00000	0.00000	0.00000		
22	-1.00000	-0.02644	0.20257	0.23286	0.20337	-0.20000	-0.02620	-0.04110	0.71429	0.18919	-1.00000	-1.00000	0.18919	-0.20000	0.01493	0.36842	-0.07989	0.45455	-1.00000	-0.20000	0.18919	0.00000	
23	0.00000	-0.21495	0.08046	0.11672	0.08046	0.00000	-0.32353	-0.24675	1.00000	0.00000	0.00000	0.00000	0.00000	0.36842	-0.20000	1.00000	-0.29032	0.39394	0.00000	0.00000	0.00000	-0.20000	0.00000

Matrix of significant Fst P values; Significance Level=0.0500.

4.5. Analysis of Molecular Variance (AMOVA)

The AMOVA results for the Y chromosome data in five Ethiopian populations is presented in table 4.9. The overall F_{ST} is 0.084. When the populations are grouped according to linguistic and geographic region, almost similar F_{ST} (0.83) with the overall was observed. The over all proportion of among group variance (F_{CT} = 0.07), which is different when they are grouped either according to languages (F_{CT} = 0.18) or geography (F_{CT} = 0.14). Similarly, the AMOVA results for the *MT-COXII* data in present Ethiopian populations is shown in table 4.9. The overall F_{ST} is 0.019, which is almost similar with when populations are either grouped in linguistic (F_{ST} = 0.018) or geographic regions (F_{ST} = 0.018). When the populations are grouped according to linguistic families, the proportion of among group variance (F_{CT} = 0.012), is also similar with when populations are grouped according to geographic regions (F_{CT} = 0.013), but different from the overall proportion of among group variance (F_{CT} = 0.034).

Table 4.9. Analysis of molecular variances (AMOVA) for Y chromosome.

Groups	No.of Groups	Within populations		Among popns within groups		Among groups	
		Variance (%)	F_{ST}	Variance (%)	F_{SC}	Variance (%)	F_{CT}
Overall	1	85.2	0.084	6.74	0.08	7.07	0.07
Linguistic	2	84.7	0.083	5.05	0.06	8.06	0.18
Geographic	2	85.5	0.083	6.99	0.07	4.65	0.14

Table 4.10. Analysis of Molecular Variances (AMOVA) for *MT-COXII*.

Groups	No.of Groups	Within populations		Among populations within groups		Among groups	
		Variance (%)	F_{ST}	Variance (%)	F_{SC}	Variance (%)	F_{CT}
Overall	1	98.60	0.019	2.0	0.020	2.0	0.034
Linguistic	2	98.13	0.018	1.9	0.019	0.10	0.012
Geographic	2	98.24	0.018	0.74	0.004	1.29	0.013

4.6. Pair wise Differences and Nucleotide Diversity

Computation of mismatch distribution was omitted for Y chromosome because of the smaller number of haplogroup genotyped in present Ethiopian population. But, based on the *MT-COXII* sequence data, the pair wise and nucleotide diversity was computed for four Ethiopian populations. Similar patterns of pair wise and nucleotide diversity observed, both were higher in Berta than Nuer, although the later is with larger sample size. Both the pair wise diversity (k: 33.714) and the nucleotide diversity (Pi: 0.06255) were higher in Berta, followed by Shinasha (k: 24 and Pi: 0.04436). Gumuz and Nuer occupy roughly intermediate pair wise and nucleotide diversity (Table 4.11).

Table 4.11. Pair wise and Nucleotide diversity of *MT-COXII* among four Ethiopian populations

Populations	Nuer	Berta	Gumuz	Shinasha	Total
No. of Variable sites	6	107	1	60	138
Number of Mutations	6	120	1	41	169
Pairwise differences (k)	1.944	33.714	0.667	24.000	16.790
Nucleotide diversity (Pi)	0.00361	0.06255	0.00123	0.04436	0.03121
Pi (SD)	0.00084	0.01957	0.00058	0.02791	0.01140
Number of haplotypes	5	6	2	3	13
Haplotype diversity (Hd)	0.80556	0.89286	0.66667	0.70000	0.923
Hd(SD)	0.120	0.111	0.314	0.218	0.030

Tajima's D: -2.48525 Statistical significance: ***, P < 0.001

4.7. F-Statistics and Migration Rate

The F_{ST} estimates and migration rate for the population analyzed is shown in table below (Table 4.12). The estimate was performed according to the island model of migration rate. For Y chromosome, Nilo-Saharan speaking groups were observed to have higher Nm as compared to the Afro-Asiatic speaking groups. Within Nilo-Saharan groups, higher migration rate (Nm =16.24) was observed in Nuer group, followed by Bertha (Nm = 11.19) group. For Afro-Asiatic speaking groups, the highest migration rate (Nm = 14.15) was observed in Shinasha. With

regard to *MT-COXII*, similar patterns of Y chromosome migration rate were observed for Nilo-Saharan groups. Also, within Afro-Asiatic group, the highest migration rate (Nm = 11.37) was observed in Shinasha but less than for the Y chromosome data.

Table 4.12. F_{ST} and Nm Estimates of Y chromosome and *MT-COXII* in two linguistic groups in Ethiopian populations.

Linguistic groups		Subgroups	F_{ST}[1]	Nm
Nilo-Saharan	Y chromosome	Nuer	0.058	16.24
		Berta	0.082	11.19
		Gumuz	0.184	4.43
Afro-Asiatic		Shinasha	0.066	14.15
		Mao	0.292	2.43
Nilo-Saharan	*MT-COXII*	Nuer	0.1654	7.04
		Berta	0.1641	5.09
		Gumuz	0.179	4.58
Afro-Asiatic		Shinasha	0.0421	12.37

Note: F_{ST}[1] and Nm is a measure of interpopulation variability and the effective number of migrants respectively.

4.8. Gene Diversity and Neutrality Test

The measures of genetic diversity estimated from Y chromosome data is depicted below (Table 4.13). Although the Nuer has the highest sample size, they were found to contain lowest number of haplotypes. The high numbers of haplotypes (3) was observed in Shinasha and Mao, which have small sample size. The intermediate and almost equal numbers of haplotypes were observed in Berta and Gumuz.

Neutrality tests were applied to search for demographic signs of population expansion. Tajima's D test based on the *MT- COXII* region sequences was significantly negative for three populations (Nuer, Berta and Shinasha) but zero and non-significant for Gumuz population. As in the case of Tajima's D statistics, the measure of Fu's Fs statistic also revealed values significantly different from zero in all analyzed individuals (Table 4.14). In addition, the overall non-significant values of Fu and Li's D* test statistic (-3.64183, $P < 0.02$), Fu and Li's F* test statistic (-3.84989, $P < 0.02$) and a significant Fu's Fs statistic (2.372) were obtained (Appendix 7).

Table 4.13 Measure of Molecular Diversity and Neutrality test Estimated from Y chromosome Data.

Populations	Nuer	Berta	Gumuz	Shinasha	Mao	Mean	s.d.
N	50	22	23	11	12	23.60	1.685
Hp	4	4	4	3	3	3.600	1.581
Pi	0.811	1.789	1.650	4.634	4.532	4.346	2.018
Tajima's D	0.910	3.672	0.200	6.036	5.865	2.249	2.763
Tajima's D p-value	0.212	0.927	0.387	0.991	0.837	0.629	0.336

Table 4.14. Molecular Diversity and Nuetrality Test Estimated from *MT-COXII* Data.

Populations	Nuer	Berta	Gumuz	Shinasha	Mean	s.d.
N	9	8	3	5	6.250	2.384
Hp	5	6	2	3	4.000	1.581
No. of Transitions	6	56	1	25	22.000	21.575
No. of Transversions	0	64	0	35	24.750	26.790
No. of Substitutions	6	120	1	60	46.750	48.205
No. of Transsion sites	6	52	1	25	21.000	20.012
No. of Transversion sites	6	52	1	25	24.750	6.790
No. of substitution sites	6	107	1	60	43.500	43.350
No. of indel sites	0	3	0	1	1.000	1.225
Pi	1.944	34.464	0.666	24.400	5.368	14.513
Tajima's D	-0.520	-0.996	0.000	-1.260	-0.694	0.480
Tajima's D p-value	0.332	0.150	0.979	0.003	0.366	0.372
Fu's FS test						
FS	-0.91082	3.67237	0.20067	6.03657	2.24970	2.76368
FS p-value	0.20100	0.94200	0.41000	0.98900	0.63550	0.33858

4.9. Principal Component Analysis (PCA)

Y chromosome principal component analysis of present Ethiopian and combined with few data from neighbour Sudan is depicted below (Figure 4.7 A and B) based on F_{ST} genetic distances. 92.42% of the total variation was contributed by the first PCA. This PCA showed the genereal genetic affinities between studied populations. Broadly three groups of genetically close populations were observed. This includes Mao, Shinasha and Gumuz. The Berta and Nuer are not only genetically distinct from each other, but also from the other groups. The second PCA run was used to analyze the combined data of this study with other available data from Ethiopia, Sudan, Senegalese and Turks. The PCA showed general genetic similarities between the populations. Three main groups of genetically closely related populations were observed. The first groups are Beja and Amhara, the second Berta and Gumuz and the third Shilluk, Shinasha, Mao and Nuer* (Sudan group).

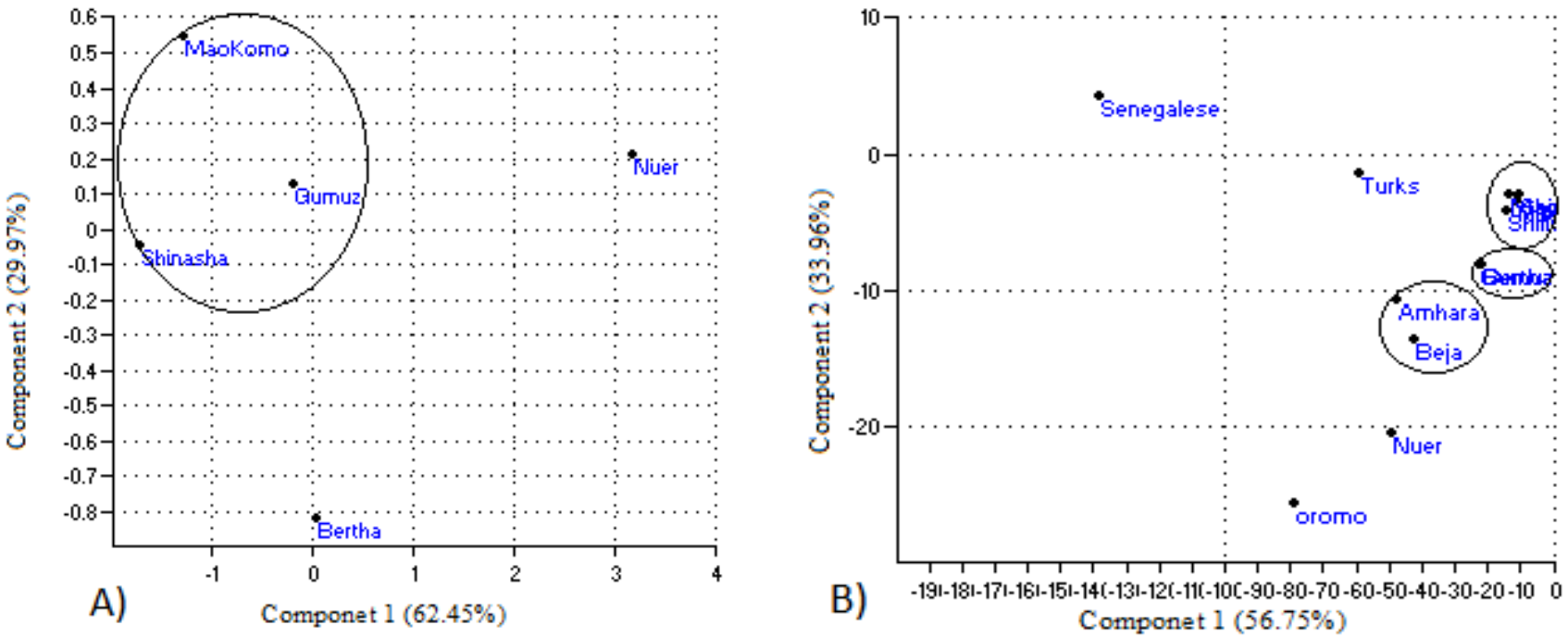

Figure 4.7. Principal Component Analysis based on Y chromosome haplogroups. A) Genetic affinities between populations in present study. B) Genetic affinities of Ethiopian populations of the present study compared with three Sudanese group (Hassan, *et. al*., 2008), two Ethiopian groups (Amhara and Oromo) and Senegalese (Semino, *et. al*., 2002) and Turks (Sanchese, *et. al*., 2005) (A) and present and Circle shows the genetic affinities between populations.

4.10. Y chromosome Phylogeny

Unrooted phenogram (Figure 4.8) based on pairwise F_{ST} shows two major clusters according to the population's geographic regions. The second major cluster further divided into two sub clusters, the first contains Berta and Gumuz and the second cluster contains the Shinasha and Mao following geographic distance. They are clustered together according to their region.

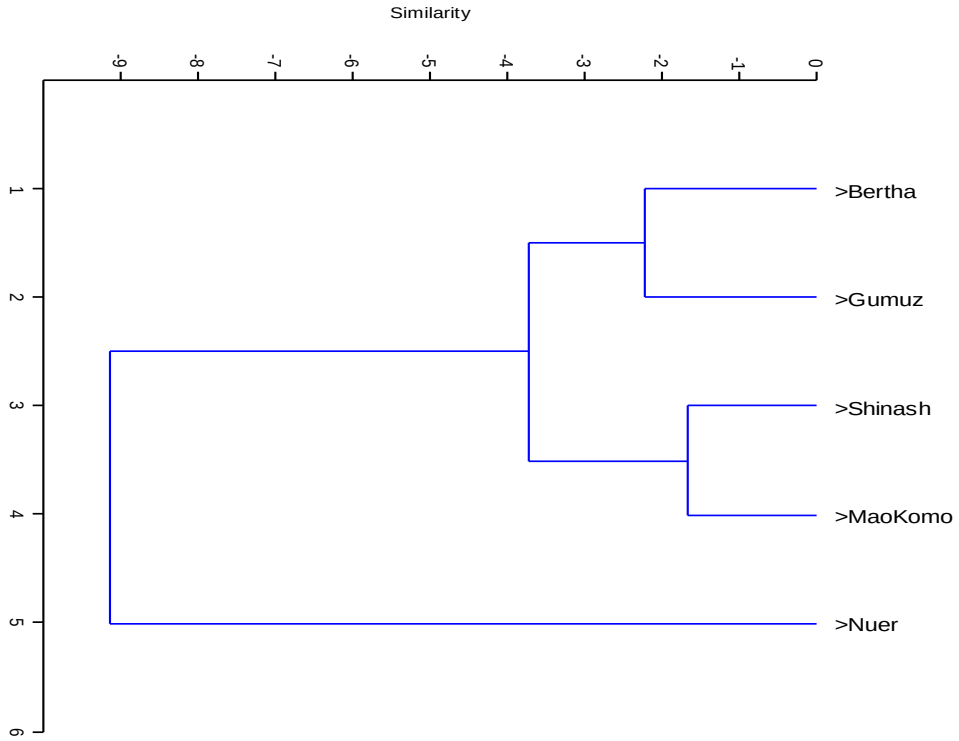

Figure 4.8. Similarity Tree of Five Ethiopian populations based on the genetic distances of Y chromosome binary markers.

4.11. *MT- COXII* Phylogeny

MT- COXII Phylogenetic Trees

In the case of *MT- COXII*, the individuals were not grouped to populations and they were used to construct the phylogenatic tree. The phylogentic tree were constructed from inter-individual distance and showed their genetic relatedness. In all topologies of the phylogenetic trees, each sub tree is made up of two branches. Both unrooted phylogenetic trees: Nelghbour-Joining (NJ) and Minimum Evolution (ME) based on F_{ST} of the *MT-COXII* sequences (Figure 4.9) and (Figure 4.10) respectively, showed similar topologies and branch length with few differences. Maximum Parsimony showed the evolutionary history for Ethiopian samples (Figure 4.11). In addition, when the data of the present study are combined with African and Asian populations, both NJ and ME phylogenetic tree of 166 individuals showed almost similar topology (Appendix 8), but a little bit different ftom the maximium Parsimony (Figure 4.12). In the later case, number of the 294 poold data is reduced to 166 individuals because of unmanageable size of the tree. Generally, the tree showed that individuals clustered according to to their geographic origin. The great majority of individuals' considerd formed separate clusters that coreespond with the continent of origin of the samples. Inaddition there were a tendency of forming subclusters within continents that correspond to the population of origin of samples.

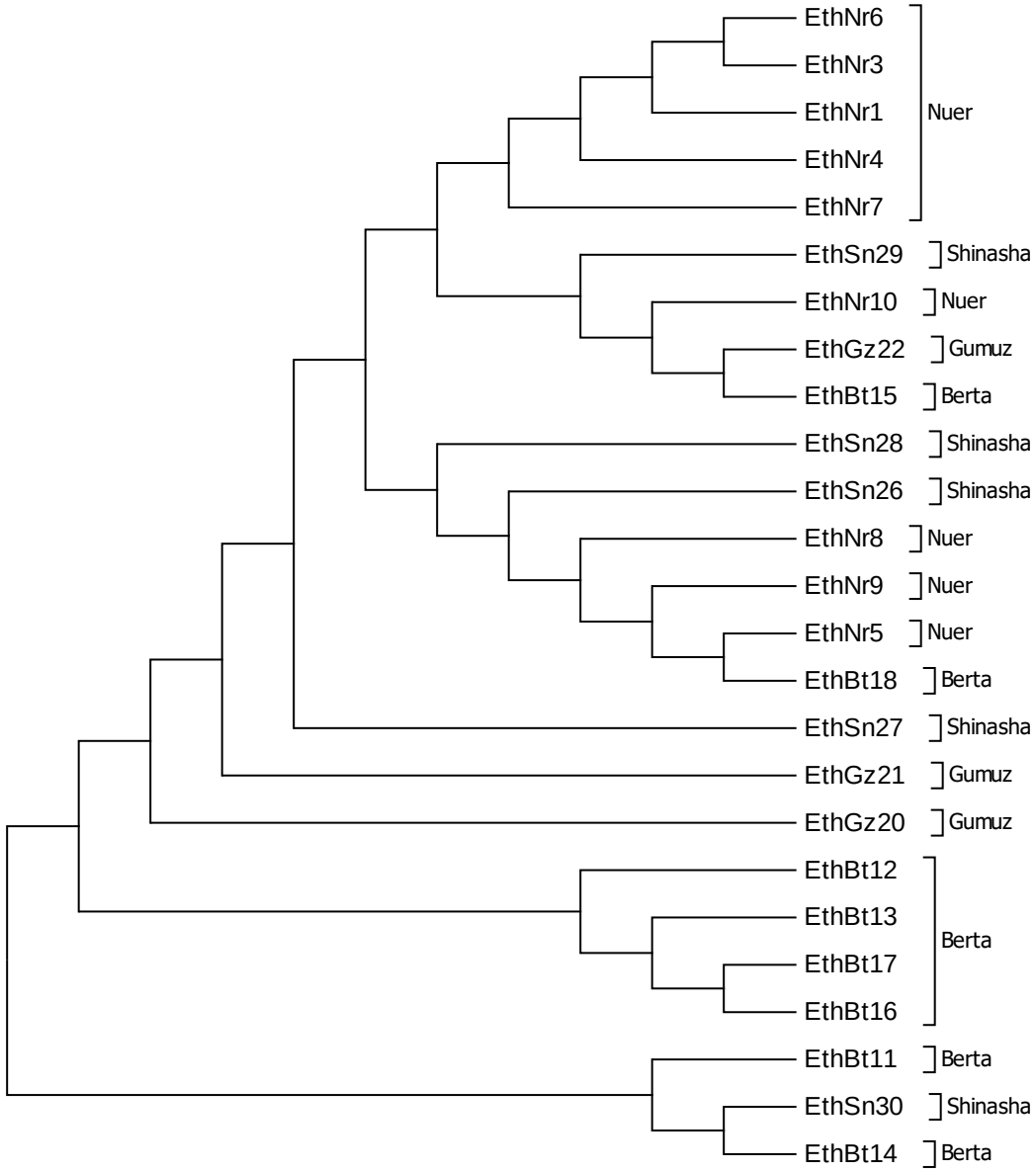

Figure 4.9. Evolutionary relationships using NJ Tree of 25 Ethiopian individuals based on the genetic distances of *MT-COXII* Sequences.

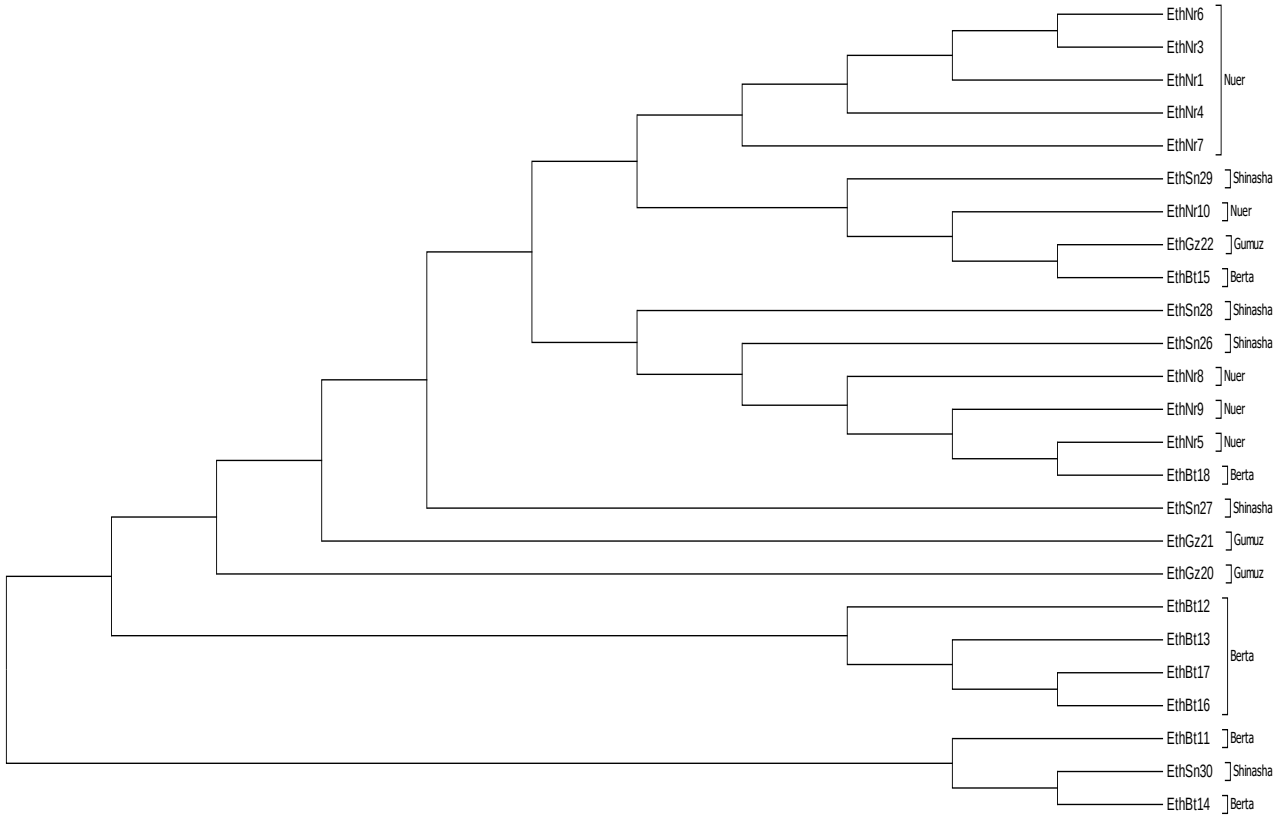

Figure 4.10. Evolutionary relationships using Minimuim Evolution Tree of 25 Ethiopian individuals based on the genetic distances of *MT-COXII* Sequences.

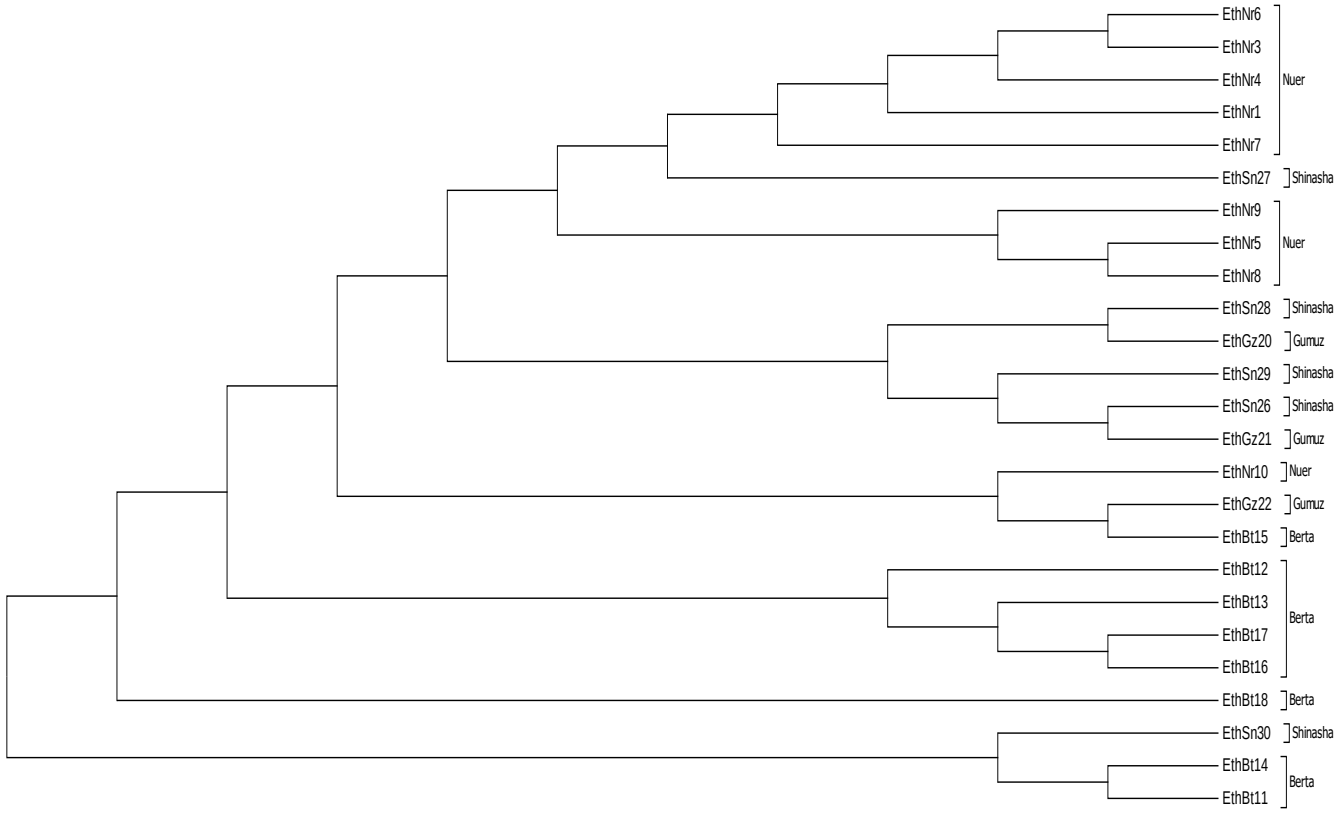

Figure 4.11. The evolutionary history using Maximum Parsimony Tree of 25 Ethiopian individuals based on the genetic distances of *MT-COXII* Sequences.

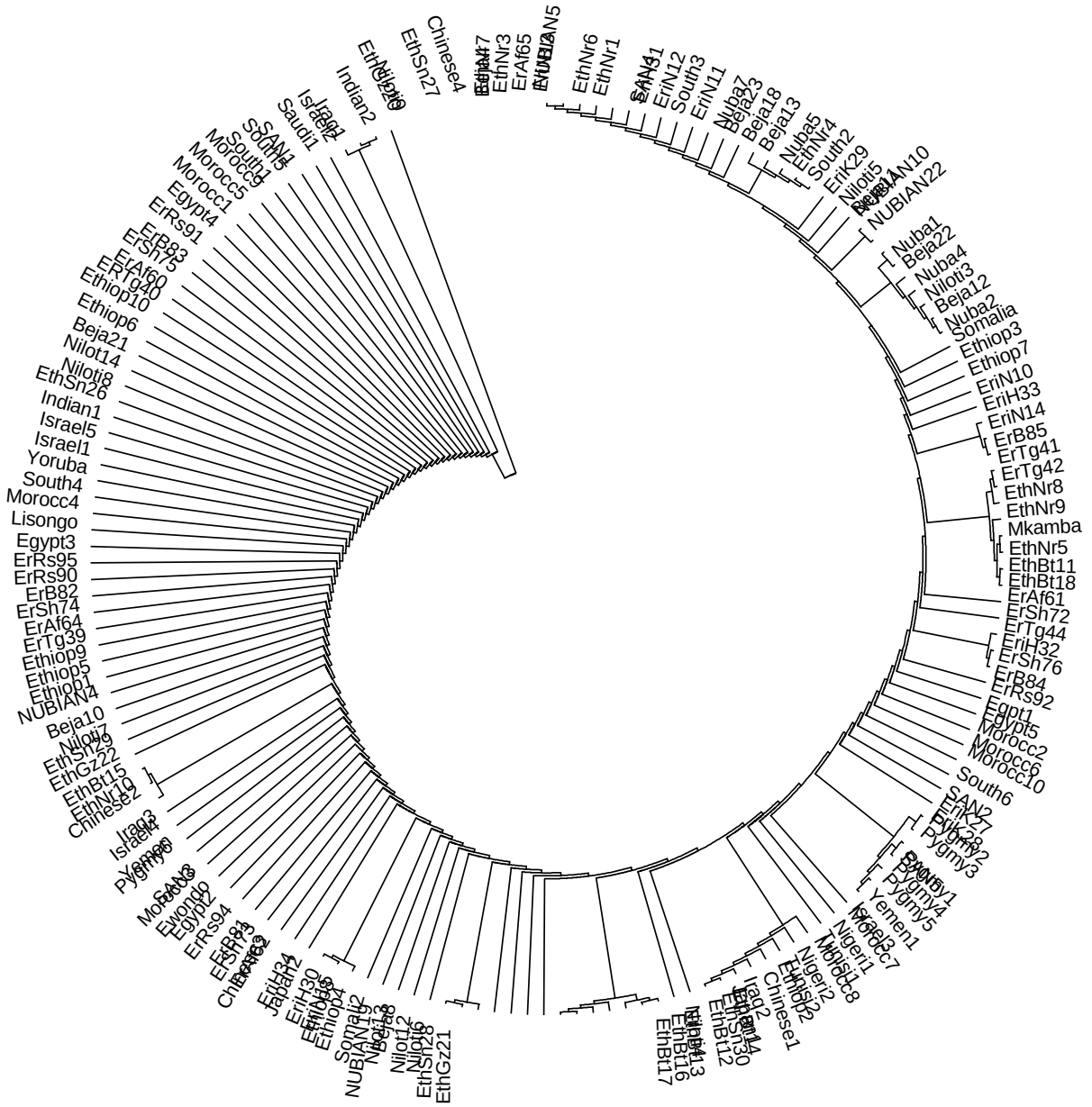

Figure 4.12. The evolutionary history using Maximum Parsimony Tree of 166 Africans and Asians populations based on the genetic distances of *MT-COXII* Sequences.

5. DISCUSSION

Ethiopia is a multi ethnic country containing highly diverse traditions, cultures, customs, etc. and is the key country in East Africa pertaining to the study of human population genetics. However, the human populations of Ethiopia have not extensively studied from various genetic perspectives. To contribute to this area, the association of LP phenotype with genotype and the genetic structure of some Ethiopian populations using Y chromosome and mitochondrial DNA (*MT-COXII*) markers were examined as an objective of this study. The mtDNA and Y chromosome polymorphisms were studied in samples from Ethiopia in order to better define the different components of the Ethiopian human population gene pool. Based on the polymorphisms generated from the Y chromosome and mtDNA study, haplotypes can be easily constructed, because of the lack of recombination in the genomic regions of these genetic materials, and permit inferences about the history of populations through female and male lineages separately. The results of this study are described in chapter four and here, all the results described above are discussed one by one.

5.1. The HBT in Ethiopian populations

Previous LP phenotype-genotype association studies were conducted on Ethiopian populations and revealed several causative markers for LP/LNP phenotype making the lactase genetics very complex (Ingram, *et al.,* 2009 and Oljira, *et. al.,* 2014) particularly in Ethiopia. These studies have suggested the need to undertake further phenotype-genotype association studies in diverse ethnic groups in Ethiopia to confirm the association of the already known markers with LP and other related information. Based on this ground phenotype-genotype association studies were planned in five ethnic groups (Nuer, Berta, Gumuz, Shinasha and Mao) from Ethiopia containing a total of 155 samples. The genotype study underway used PCR-RFLP and Exom sequencing.

The phenotype of LP was conducted using the validated 50g lactose dose in BHT as explained by Oljira, *et. al*. (2014). This author, after testing and confirming its validity, has reported that 50gm of the lactose sugar was the best amount to be used for Ethiopian adults in the BHT.

The BHT results showed that out of the total 155 study subject, 83 (53.55%) and 68 (43.87%) were LP and LNP respectively (Table 4.1). The frequency of LP in the studied Ethiopian reaches up to the percentages of 53.55% with noticeable differences between ethnic groups (between 50-60%). Similarly, the frequency of LNP was 43.87%, with high differences among ethnic groups (between 40-50%). This percentage is in between that of the Europeans (higher) and that of the Far Eastern (lower) populations (Swallow, 2003).

In comparison the highest frequency of LP in pastoralist Nuer group and in populations with mixed agriculturalist (Berta, Gumuz, Shinasha and Mao) was obsereved. This result was an unexpected on the basis of the notion that the frequency of LP has increased in response to the culture of milk drinking and that pastoralist populations are expected to have a higher frequency of LP (Swallow, 2003). The daily base milk drinking habit of each participant was assessed using questionairre. The lactase phenotype frequency was higher in the milk drinkers on the daily basis than the non-milk drinkers. In all groups, the majority of the LNP and very few LP individuals showed various post gastrointestinal symptoms after the BHT. This is in agreement with the expectation that LNPs cannot digest the large amount of the ingested lactose due to the lack lactase enzyme, and that lactose would move to the small intestine to be fermented by colon bacteria into number lactose products which initiates symptoms of lactose intolerance. In this study, also there were a total of four peoples with NHP (2.58%). Of these, the number of NHP phenotype was one both in in Nuer (1.8%) and Shinasha (5%) and two in Mao (10%). The low

frequencies of lactase unknown in the Nuer disagree with the hypothesis of natural selection in favor of the "lactase persistence gene" in milk-dependent pastoralist perhaps due to some individuals claim that they are Nuer although they are Anuak in real sense or it could be simply a chance event.

5.2. Y Chromosome Binary Markers in Ethiopia

Analyses of some Ethiopian Y chromosome haplogroup substructure as well as the region's phylogenetic relationships to few neighbour populations have provided information on a high haplogroup diversity shared among populations along the Eastern parts of Africa. In this study, all individuals were genotyped for the YAP marker, which is found in all other Y chromosome haplogroups outside of Africa. They were categorized as individuals with YAP+ and YAP- polymorphism with the percentages of 30 and 70 respectivelly. This study showed that the frequency of the YAP insertion was found to be moderately high in the studied populations. In agreement with this, the previous general studies of YAP + and YAP- polymorphism have reported higher frequencies of the YAP + within the African populations (Hammer, *et. al*., 1994 and Agrawal, *et.al*., 2005) suggesting that this marker would be an African ancestry. Moreover, very high frequency of the YAP+ was reported to have appeared in African, Japan and Tibetans population followed by the Western Eurasians populations (Hammer, *et.al*., 1994, 1995). Hence, the frequency of YAP, an *Alu* insertion polymorphism obtained in this study is in agreement with the report of high frequency of YAP in African. The YAP+ in Africa falls in to haplogroups E.

All samples carrying YAP+ insertion was further typed for E-M78 carrying derived allele C/T. These polymorphisms were investigated by allele specific PCR amplification to obtain T/C SNPs. Since, the YAP- individuals could fall into several haplogroups such as A, B, F, or other;

they were analyzed for major binary markers A-M13, B-M60 and F-M89 by PCR amplification. The mutations or SNPs in the fragments amplified by PCR were subsequently detected by RFLPs analysis and checked for the presence of these haplogroups. The set of Y chromosome binary alleles carried by a single individul is reffered to as Y haplogroups. Thus; in general, four Y chromosome binary markers A- M13, B-M60, E-M78 and F-M89 were analyzed. All these were observed among Ethiopian populations with higher frequencies of 44.04, 39.28, 86.11, 21.42 % respectively (Table 4.2). E-M78 haplogroup was found to be the most frequent followed by haplogroups A-M13 and B-M60, which were present at such higher frequencies and are exclusive to the African continent. In agreement with this, Gomes, *et. al*. (2010) have reported that the majority of Y chromosomes in sub-Saharan Africa carry haplogroup A and B with a frequency of approximately 33% each. According to this author, in comparisions, Nilotic speaking groups contain high frequencies of these haplogroups than the Afro-Asiatic speaking groups, which is consistent with the present study.

The frequency of haplogroup A-M13, also called A3b2-M13, obtained here support the previous studies. It was reported that it is the major haplogroupA subclade in East Africa; with a low level presence in central and West Africa. In terms of language, the largest proportion of A-M13 was found among Nilo-Saharan speaking populations than Afro-Asiatic speakers (for which intermediate frequency was observed here). Of Nilo-Saharan speaking groups, haplogroup A-M13 was found at higher frequencies in Nuer (37.84%) followed by Gumuz (18.92%). In support of this, A-M13 was reported to be prevalent in Nilo-Saharan populations (Wood, *et. al*., 2005 and Hassan, *et. al*., 2008). The high to moderate frequencies of M13 among both Nilo-Saharan and Afro-Asiatic populations in East Africa, as well as the presence of closely related and shared haplotypes observed in world mtDNA data are suggestive of extensive contact and

gene flow over a long period of time. Haplogroup A3b2-M13 was the most common haplogroup in the present sample with a frequency of 44.04%, greater than that already found in Nilotic groups from Sudan (30%) (Hassan, *et. al.*, 2008). Likewise, small frequencies were found in several Afro-Asiatic speaking populations from East Africa (Wood, *et. al*., 2005 and Hassan, *et. al*., 2008). This pattern of frequencies were also supported by autosomal data (Tishkoff, *et. al*., 2009). As described by Tishkoff, *et. al.* (2009), while East Africa contains most of the A3b2-M13 Y chromosomes, central and West Africans have high diversity of this haplogroup. These together with the potentially bi-directional nature of migration along the Sahel increase the potential for the haplogroup to have arisen anywhere along the belt. The lack of comparative data for the sub clades of haplogroup A-M13, haplogroups A-M171 and A-M118 make the analysis of these haplogroups difficult in Africa if not imposiible. Although not well characterized, both A-M171 and A-M118 were reported to be found be in Central African populations. Moreover, little other data exist on these sub clades apart from the findings of A-M118 in six Ethiopians and A-M171 in a Sudanese individual (Underhill, *et. al*., 2000).

Haplogroup B-M60, like clade A-M13, is a haplogroups that is deeply rooted within the human Y chromosome tree and is known to be common among populations in East Africa with varying degree of frequencies among different ethnic groups (Jobling and Tyler-Smith, 2003 and Wood, *et. al*., 2005). In this study haplograoup B-M60 was found at the highest frequencies in samples of Nilo-Saharan speaking groups than the Afro-Asiatic speakers. It was found in Nuer with frequency of 36.36% followed by Berta and Gumuz each with frequency of 21.21%. Similar study of the Y chromosome of populations in Sudan revealed higher frequencies of haplogroup B-M60 in Southern Sudanese (30%), Hausa (16%) and Nuba (14%) (Hassan, *et. al*., 2008), which are close to the result of the present study. According to another studies, haplogroup B-

M60B has been found in approximately 2.5% and 6% of Sudanese males and Malagasy males of Madagascar respectively (Underhill, *et. al*., 2000 and Hurles, *et. al*., 2005). Like A-M13, the higher frequency of haplogroup B-M60 in the present Ethiopian populations may suggest an African ancestory of this haplogroup. It also supports the report that haplogroup B-M60 is the oldest and one of the most diverse human Y haplogroups in African populations.

E-M78, which is by far the most mutationally diverse of all major Y chromosme haplogroups, is more wide spread in East Africa. In support of this, although E-M78 was present haplogroup is common to most of the present Ethiopian populations, it has exceptionally high frequencies in Nuer (51.62%) population, followed by Berta and Gumuz containing 19.35% and all are Nilotics. Interestingly, the frequency of the E-M78 observed in the Nuer, Bertha and Gumuz populations is higher than the percentages reoprted in the other African populations (Hassan, *et. al*., 2008). The high frequency of the E-M78 among Nilo-Saharan speakers may indicate that these populations are experiencing recent gene flow from other populations, have ancestors who were not involved in a back migration from Asia (Cruciani, *et. al*., 2002) or alternatively represent a genetically distinct Ethiopian populations, though these arguments all require further extensive and in depth investigation in order to confirm. Previously, it was reported that the E-M78 subclades identified M35-E3b1a that commonly found and widely distributed in Northern and East Africa, Western Asia, and Europeans (Cruciani, *et. al*., 2006, Hassan, *et. al*., 2008 and Karafet, *et. al*., 2008). Recently tremendous internal bi-allelic markers (UEPs) were observed within the E-M78 subclades. In this study, the analysis of E-M78 among Ethiopian samples confirmed the presences of deep terminal branch E1b1b1a (formerly called E3b1a sub lineage) and thus some what refined the phylogeny of haplogroup E-M78 in Ethiopia. The analysis was

done through genotyping of C/T SNP present on the E-M78 chromosomes which define the terminal branch of this haplogroup. Haplogroup E1b1b1a, the terminal branch of E-M78, was found at highest frequency in Nuer of this study, which is identical with the 51.62% mentioned above for parent E-M78. Moreover, its distribution in studied Ethiopian groups is highest than high frequency reported in Europe and Middle East (Cruciani, *et. al*., 2007). These results are in agreement with the previous findings of a high prevalence of M35 lineages in East African populations, but they contrast with those observed in Nilo-Saharan groups from Sudan (Hassan, *et. al*., 2008), where most lineages inside this haplogroup carry the M78 mutation. Although originally thought to have originated in North East Africa, this result may suggest the Ethiopian origin of this subclades, but further detail analysis with more representative samples are required for confirmation. It has also notable frequency in other Nilo-Saharan speaking groups, Bertha and Gumuz with identical frequency of 19.35%. In general, the present investigation indicated that the currently confirmed terminal branch with such high frequency within the E-M78 lineage may contain additional information in terms of new bi-allelic markers and that microsatellite marker should be applied to detect other UEPs. This marker was identified within a previously visualized terminal branch of the E-M78 tree. This may suggest by inference that, since numerous number of E-M78 UEPs has been reported in the past couples of years (Underhill, *et. al*., 2006 and Karafet, *et. al*., 2008), the apparent terminal branches of the E-M78 tree still contain a large amount of information, in terms of new bi-allelic markers, which can be important for human population studies, some of which for instance, for the detection of spatial patterns attributable to both ancient dispersals and more recent events of gene flow. The E-M78 haplogroup defined by E1b1b1a was also the most frequent haplogroup in Albanians and the Roma, Serb or Turk group followed by Macedonians. It is the most common haplogroup E

lineage in Europe with a frequency peak centered in the Balkans (cruciani, *et. al*., 2004 and Semino, *et. al*., 2004). It is also found in the Middle East and in eastern and northern Africa. Its high frequency in these mentioned populations and in Kosovar Albanians (46%) and Macedonian Roma (30%) is most likely a result of genetic drift by chance over time (Pericic, *et. al*., 2005).

Unexpectidely, haplogroup F-M89 were found in Nilo-Saharan speaking groups with high frequencies in Berta (55.55%) and Gumuz (33.33%), but relatively at low frequencies in Nuer (11.11%), and completely absent from the Afro-Asiatic speaking groups (Figure 4.5). From personal communication, elders from the two groups have claimed their Arab ancestry and tried to convience with their present numerious Arabs like traditions and Islam religion. Thus, this mutation might have disseminated to Out of Africa mainly to Asia, through the ancient Arabs, where it is currently found at highest frequencies. The F-M89 haplogroup individual was M89+ and thus considered to be in haplogroup F. Unfortunately, it was not tested for the presence of mutations that define the haplogoups descended from F (sub clades of F-M89) in order to confirm the present findings. Y chromosome haplogroup variation and distribution among the five populations of the present study and few neighbour populations are illustrated (Figure 4.5). Haplogroup M89 that defines haplogroup F is the very common Y chromosome haplogroup and is the direct ancestor of the majority of Y chromosomes present outside of Africa. Approximately 90% of the male populations carry the M89 and is primarily found throughout South, Southeast and parts of East Asia. The origin of F-M89 was attributed to North Africa and Eurasia (mainly Indian), in which the latter has numerous supporting evidences than the former. The high frequency of F-M89 observed in the present samples may suggest that Haplogroup F-M89 appear to have originated in Africa, from where it was migrated to outside of African

continent early past during prehistory. To shed light on the hypothetical "Back to Africa" migration, Y-chromosome haplogroups (J, R1b and T) associated with this hypothesis should be genotyped. Also, its high frequency may be due to small samples sizes and thus sampling effects of studied population.

Genetic Structure and Pattrens of Distributions

Population pair wise F_{ST} values were estimated to determine the level of genetic differentiation between studied populations, based on the four Y chromosome binary markers. A principal Component Analysis (PCA) was performed with PAST 3.exe using the typed haplogroup frequencies of populations living in two geographic regions (Gambella and benishangul). The PCA plot based on F_{ST} values of the present Ethiopian samples showed that populations were grouped according to their geographic regions. The Nuer and Berta are genetically differentiated from each other and also from other studied groups. Analysis of haplogroup diversity allowed understanding of these differences between the Nuer and Berta and other groups. The Nuer and Berta might have differentiated from one another and from the rest populations because of the presence of high frequency of haplogroups E-M78 sub clades in Nuer and high frequency of Hamplogroup F-M89 in Berta. PCA plot based on F_{ST} values of present Ethiopian data combined with few data from other Ethiopians, Sudanese, Turks and Senegalese populations, showed three main genetic clusters that was grouped according to linguistic and geographic variations except for Beja and Amhara (Figure 4.7). The first cluster includes populations who speak languages of the Nilo-Saharan family (Berta and Gumuz), one of linguistic family in Ethiopia. The second cluster includes populations (Beja and Amhara) who are essentially speakers of different languages belonging to the Afro-Asiatic family. The third cluster encompasses the Sudanese groups. These clusters are defined by the predominance of the ancient

haplogroups A-M13, B-M60, and E-M78. Both A-M13 and B-M60 are haplogroups that are deeply rooted within the human Y-chromosome tree, and they are known to be common among populations in East Africa (Underhill, *et. al*., 2000 and Semino, *et. al*., 2002). The placement of the Ethiopian Nuer, who speaks a language of the Nilo-Saharan family close to the Oromo group who speak language that belong to Afro-Asiatic family clearly revealed the importance of language relationship by revealing the presence of high frequencies of A-M13 in both groups and their geographic proximity. Although they are not close to each other, Amhara and Beja are clustered together may be because of the presence of shared ancient haplogroup A-M13 and B-M60. The distribution of E- M78 sub clades indicates that the Beja are perhaps related as well to the Oromo on the basis of the considerable frequencies of E3b1a among Oromo in comparison to Amhara (Cruciani, *et al*., 2007). These indicate the presence of the historical contact between Ethiopia and eastern Sudan and showed the strong correlation between linguistic and genetic diversity as these populations speak languages of the Afro-Asiatic family (Cavalli-Sforza, 1997; Passarino, *et al*., 1998 and Hassan, *et.al*., 2008).

The AMOVA results for the Y chromosome data is shown in table 4.8. The AMOVA analysis highlights the population differentiation and shows association of linguistic with genetics in the data set. The analysis revealed that 85.2% of the genetic variation occurred within populations, which might suggest a low and less stable regional genetic structure in the studied population. The overall F_{ST} estimated (F_{ST} = 0.084) values within populations showed strong population differentiation. When the populations are grouped according to linguistic and geographic families, almost similar F_{ST} with the overall was observed. The over all proportion of among group variance (F_{CT} = 0.07), is a bit different when they are grouped either according to languages (F_{CT} = 0.18) or geography (F_{CT} = 0.14). Thus, when populations are grouped

according to linguistic affiliation, the proportion of among group variance is similar with when they are grouped according to geography. This may indicate that Y chromosome variation is partitioned among both geographic and linguistic groups, showing the absence of gene flow supporting the PCA plots.

In contrast to this, Nm (migration rate) estimates on the basis of F_{ST} have indicated that gene flow is the recent phenomenon in the studied population. Most speakers of Nilo-Saharan languages, the major linguistic family spoken in the country, show high evidence of gene flow and demonstrate high migration rate as compared to the Afro-Asiatic speaking groups. Within Nilo-Saharan groups, higher migration rate (Nm =16.24) was observed in Nuer group, followed by Bertha (Nm = 11.19) group. For Afro-Asiatic speaking groups, the highest migration rate (Nm = 14.15) was observed in Shinasha. With regard to the *MT-COXII* similar patterns of migration rate were was observed for Nilo-Saharan groups. But, the highest migration rate (Nm = 11.37) was observed in Shinasha which is Afro-Asiatic (Table 4.12). In general, the high migration rate (Nm) obtained using F_{ST} showed that they all appear to have sustained considerable gene flow from neighbor by exchanging individuals. On the basis of this result, high gene flow would result in homogenization of the populations and have indicated that the populations under study are not differentiated. Other results from the Y chromosome and *MT-COXII* data have contrasted this result and they both showed that the population under study are differentiated, may be due to genetic drift which counteracts the homogenization effect of gene flow. The contradiction can be explained in such a way that, from evolutionary point of view, exchanging individuals may not necessarily have an impact on population structure as they may not contribute genes to the population that they migrate into. Another explation is overestimation of Nm from relationship of Nm and F_{ST} as the number of populations under study or exchanging

individuals is small. It also seems that the gene flow observed here might have resulted from sharing of genes due to common descent but may not due to actual movement of genes. However, the observation that the migration rate of the Nilo-Saharan speakers (which encompass a pastoral Nuer) is generally greater than Afro-Asiatic groups experiencing sedentary agriculturalists lifestyle is in accordance to the expectation that pastorals to be more able to admix, spread, and receive genes than their sedentary counterparts. For all studied populations, P values of Tajima's D (Table 4.13) were highly significant and, therefore, the hypothesis of demographic expansion could be accepted (at the P = 0.05 level). This may be due to the nature of past demographic expansion.

Accordingly, this study suggests that regional variation in Ychromosome haplogroups in present samples of Ethiopia is likely to have been shaped by combination of evolutionary forces which occurred in the far past. We found a contrast between demographic and genetic based Nm (migration rate) estimates. The high effective population size of Nilo-Saharan males could be explained by a recent higher migration rate with out actual movement of genes due to migration (together with the above expressed explanation); diversity among populations of Afro-Asiatic versus Nilo-Saharan speaking groups and the low male migration rate in Afro-asiatic may be due to the *in-situ* evolution of Afro-Asiatic in East Africa.

5.3. The *MT-COXII* Variation in Ethiopian Populations

This study is the first to sequence and analyze the *MT-COXII* coding region among Ethiopian populations. In this study it was attempted to analyze the *MT-COXII* gene sequences inorder to assess the level of genetic variation within some Ethiopian populations. For comparsions, the *MT-COXII* sequences of the present samples were combined with different global mtDB

sequences. Within the *MT-COXII* sequences of the studied populations, the polymorphic sites were allocated and counted visually along the alligned sequences against the Yoruba and rCRS and fifteen substitutins in the form of SNPs were detected. There was a total of 3 non synonymous and 12 synonymous polymorphisms, thus yielding an NS/S ratio of 0.25. This result is in agreement with the work of Elson, *et. al.* (2004) which indicates that negative selection operated on the coding region of the human mitochondrial genome during evolution. Of the total mutations, five SNPs were novel polymorphisims identified in this study. They were unique to the present samples and were not found previously reported in mitochondrial genome data base (mtDB) sequences. Of the total numbers of polymorphisms, six and four SNPs were observed in the Berta (40%) and Gumuz (26.67%) sequences, with three (20%) and two (13.33%) SNPs mutations respectivelly with small numbers of transverse mutations. This result is in agreement with ancestral sequence evolution concept that point out 'older' sequence changes in the human mtDNA coding sequences have relatively fewer transversion substitution than a set of 'younger' sequence changes. Transitions mutations are much more common in most species that have been studied. This is because, transition mutations are more easily generated because of the shape of the substituted molecules and are less likely to be removed by natural selection because they more often create synonymous substitutions, which encode the same amino acid sequence as the original DNA. With this regard, Elson, *et. al*.(2004), suggest that the relative lower number of transversion or nonsynonymous substitution in human mtDNA during course of evolution may be attributed to negative selection or to a relatively recent relaxation of selective constraints. For confirmation or investigation of the exact effect of such changes on the protein function, functional work is needed in future.

The haplotypes frequencies were calculated using Arlequin version 3.1 (Excoffeir, *et. al*., 2005) software. Haplotypes frequency is a basic measurement of genetic diversity and is equivalent to gene diversity in haploid genome. A total of 25 sequences from four ethnic groups representing two sampling locations and two linguistic families of Ethiopia were considered in this study. Mao group was omited from sequence analysis because of low quality of the *MT-COXII* sequence output. Thirteen haplotypes were identified among the studied populations with haplotype diversity, Hd: 0.9233. Haplotype 4 scored higher in Gumuz (66.7%) and in Shinasha (60%), followed by haplotype 7 in Nuer (44.44%) and haplotype 2 in Berta (37.50%). Haplotype frequencies distribution in populations is shown in table 4.2. This high value of haplotype diversity was observed and is almost similar to a previous report based on the same investigated species from populations in Sudan using the same *MT-COXII* region as marker (Nuha, *et. al.*, 2014). The very high values for haplotypee could be attributed to the complex and variable nature of studied population with deep evolutionary history in the Western Ethiopia, a high mutation rate in the study fragment and the large population size of the species. However, no shared haplotype were observed among all population groups in this study. Thus, the analysis of *MT-COXII* haplotype distribution showed the absence of single shared haplotypes (that are common) to all studied individuals. An absence of a common haplotype among groups is a pattern frequently attributed to populations that have undergone past demographic expansion as described by Slatkin and Hudson (1991) and Rogers and Harpending (1992). In addition, the absence of common haplotype between the studied groups indicate that there is no gene flow (Krings, *et. al.,* 1999). Where as haplotype 4 has shared bewteen Gumuz and Shinasha groups, Nuer has shared haplotype 1 with Berta and Gumuz groups showing a fewer extent of gene flow between these pairs of groups. Similarly, a total of 294 world *MT-COXII* sequences were

analyzed using Arlequin to look at inter continental and inter populations diversity (Appendix 6). In comparisions to the continental *MT-COXII* sequences, the diversity within Ethiopians by far is greater than among the Ethiopian and the global mtDB sequences. In the combined *MT-COXII* sequences of Ethiopian and mtDB global sequences, 66 haplotypes were detected and assigned numbers from 1 to 66 with haplotype diversity, Hd: 0.7923. All haplotype was shared by Africans, except haplotype 63 and 64 that are only shared by Asians and Australians showing Africans are the genetic ancestor of the other population. Of all, haplotype 1 was shared at higher level by Africans (38.20%), Asians (66.70%), Americans (87.50%), Europeans (56.50%) and Australian (80%). By comparing the present and combined data, it is clear that the present fewer sequences have contributed 19.69% of the total haplotyoes resolved. This result is consistent with the report of Elhassan, *et. al.* (2014). Haplotype relative and absolute frequencies in the studied populations were also calculated. Of the total 66 haplotypes 64 detected in African, 4 in European, 2 in Australians, 2 in Middle East, 2 in Asia and 2 in Americans. This show Africa exhibit high haplotype diversity than other continents. In general, this result clearly indicated that high diversity of SNPs and haplotypes are prsent in Ethiopian samples than the pooled continental samples trace the matrilineal ancestry back to Ethiopia. This pattern of diversity supports the idea that mtDNA in human populations of African origin are the most diverse in the world and that ancient human first existed and stayed in Africa for several thousands of years.

Genetic distances analysis for Y chromosome and *MT-COXII* showed that genetic distance was greater than zero for all pair wise comparison and most of the non-significant genetic distances occur between populations of the same linguistic groups. These indicate that there is a population divergence between Ethiopian populations and this is in agreement with results from analysis of population differentiation.

The AMOVA analysis of *MT-COXII* revealed that most genetic variation (98.6%) in the data was captured within populations with the overall F_{ST} of 0.019, which is almost similar with when populations are either grouped in linguistic (F_{ST} = 0.018) or geographic regions (F_{ST} = 0.018), in agreement with the Y chromosome data. The overall proportion of among group variance (F_{CT} = 0.034) was different from when the populations are grouped according to linguistic families or geographic regions which are almost similar. That is the proportion of among group variance (F_{CT} = 0.012) based on linguistic grouping was nearly similar with when populations are grouped according to geographic regions (F_{CT} = 0.013). The similar values of variation among geographic and linguistic groups indicate that both geography and language have limitted role in shaping the genetic structure of the Ethiopian populations. The absence of correlation between maternal genetic pattern and language may be because languages can be transmitted from a population to another without genetic change, and or genetic and linguistic evolution may proceed at heterogeneous rates (Diamond and Bellwood, 2003). However, Y chromosome data have shown a correlation between language and paternal genetics, as described above. This could suggest that matrilineal and patrilineal patterns of inheritance in Ethiopia could have different genetic history. But this argument should be taken with care as *MT-COXII* sequences detects high numbers of polymorphisms as compared to Y chromosome PCR-RFLP analysis, which allows to genotype only few numbers of polymorphisms.

Moreover, as in the case of Y chromosome, AMOVA result of *MT-COXII* indicates that there are restricted levels of gene flow and thus suggest the absence of gene movement between the populations. This might have helped the studied groups to maintain their uniqueness. Because Y chromosome and *MT-COXII* are not shuffled by recombination, they are transmitted intact from one generation to the next, revealing the paternal and maternal lineages of a population. Thus,

the studied populations shared Y chromosome and *MT-COXII* lineages as a result of common origins rather than gene flow. In this study, the *MT-COXII* diversity is relatively high than that of the Y chromosome. Relatively high genetic diversity for the *MT-COXII* can be attributed to a high substitution mutation rate and to the predominance of patrilocality in human populations as described by Elson, *et. al.* (2004) and Seielstad, *et. al*. (1998). That is, the highest *MT-COXII* variation indicate greater levels of mtDNA lineage sharing among studied populations, suggesting that females may have experienced more mobility during much of our history than have males. Moreover, since there is a culture of polygymy in some cases, some of males are expected to have multiple wives, further reducing the number of Y chromosomes variability in the population.

Computation of mismatch distribution was omitted for Y chromosome because the smaller number of haplogroup genotyped in present study. But, based on the *MT-COXII* sequence data the pair wise and nucleotide diversity was computed for four Ethiopian populations (Table 4.11). Both the pair wise diversity (k: 33.714) and the nucleotide diversity (Pi: 0.06255) were higher in Berta, followed by Shinasha (k: 24 and Pi: 0.04436). They were higher in Berta and in Shinasha than Nuer, although Nuer is with larger sample size. Gumuz and Nuer contain roughly intermediate pair wise and nucleotide diversity. Like haplotype diversity, the analysis of the *MT-COXII* sequences revealed high values of nucleotide diversity and considerable genetic variation in studied populations. The Berta showed the highest genetic diversity seems heterogeneous group as if they are subdivided in to groups and have different geographical origins. This result may be attributed to absence of recent gene flows that would allow the gradual creation of unique population, thereby facilitating the phenomena of genetic drift and selection. Moreover, when compared to other Nuer group, the Berta show little differentiation from the other studied

population. This result indicates to some extent that the genetic diversity of the Ethiopian poulation is characterized by complex and high genetic variability. However, the patterns observed in this study is incomparable to those obtained for Sudan population in a study carried out using the same *MT-COXII* fragment (Nuha, *et. al*., 2014) that reported that both populations do seem to experience gene flow in the region encompassed by their studies. Interestingly, it was noticed that on the basis of small data from few participants, no haplotypes were shared based on the *MT-COXII* sequences among groups of the present study although a local contact between them and other groups living around has been described. When the present data is combined with some Sudan data and analyzed together, few shared haplotypes were observed between in Berta, Gumuz and Sudan groups. The presence of Sudan haplotype in the Berta and Gumuz group may indicate the presence of gene flow and correspond to their oral claim that they have migrated in to the present-day area from Sudan during the early time and have Sudanese genetic background.

Tests of neutrality were done for Y chromosome and *MT-COXII* sequences to search for demographic signs of population expansion, to check if natural selection might have influenced patterns of genetic diversity or had an effect on the level of variation. The measures of genetic diversity and neutrality test estimated from Y chromosome data and *MT-COXII* data are shown in table 4.13 and 4.14 respectively. Tajima's D test based on the Y chromosome haplogroups was positive for all populations but varies among populations. The Y chromosome data showed statistically significant positive values of Tajima‟s D, indicating the less possibility that these positive values are due to recent population demographic expansions. Instead the positive valus could signify balancing selection. Likewise, based on the *MT-COXII* data, Tajima's D test, Fu's *Fs* analysis were performed for neutral evolution analysis to examine the historical demographic expansions of populations, which may play an important role in determining the patterns of

genetic variability. Tajima's D test based on the *MT- COXII* region sequences was significantly negative for three populations (Nuer, Berta and Shinasha) and zero for Gumuz population (non-significant). Non-significant values in Fu and Li's D* test statistics (-3.64183, P < 0.02) and Fu and Li's F* test statistic (-3.84989, P < 0.02) combined with a significant Fu's Fs statistic (2.372) have been interpreted by Fu (1997) as evidence of the possible presence of background selection, thus supporting indications of demographic expansions. Similarly, the measure of Fu's Fs statistics revealed values significantly different from zero in all analyzed individuals, suggesting a past expansion in the population studied (table not shown). It was reported that Fu's *FS* test is the most powerful tests for detecting population growth. The resulting negative values for Tajima' D and positive values for Fu's *Fs*-tests indicate a complex demographic history of Ethiopian populations, which might correspond to bottleneck and subsequent population growth. Moreover, study conducted by Elson, *et. al.*, (2004) assessed selection in the mitochondrial DNA (mtDNA) coding-region sequences and revealed results that support the present study. This author provided evidence that selection has been a significant force during the evolution of the human mitochondrial genome and indicated evidence for both gene-specific and lineage-specific selection. Unfortunately, there are only small data on present populations, and more extensive studies are needed. At the present, based on the present results, a simple model of isolation by languages can also explain the population genetic structure of populations within region since population genetic differentiation (estimated as F_{ST}) does appear to be related to linguistic distances separating populations.

To further see the Ethiopian population clustering patterns, genetic affinities and their relationship to neighbor Sudan, PCA was performed. The Y chromosome PCA of the present Ethiopian and combined with few data from neighbor Sudan were performed based on F_{ST}

genetic distances (Figure 4.7 A and B). 92.42% of the total variation was contributed by the first PCA. This PCA showed the general genetic affinities between populations. Brodely three groups of genetically close populations were observed. This includes Mao, Shinasha and Gumuz. It shows that the Berta and Nuer are not only genetically distinct from each other, but also from the other groups. The second PCA run was used to analyze the combined data of the present study and other available data from Ethiopians, Sudanese, Senegalese and Turks. The PCA showed general genetic similarities between some populations. Three main groups of genetically closely related populations were observed. The first groups are Beja and Amhara, the second Berta and Gumuz and the third Shilluk, Shinasha, Mao and Nuer* (Sudan group). This show that the Y chromosome lineage distributions seem to cluster geographically, and for some lineages, based on linguistic classification. The correlation between genetic distances with language is stronger than it is with geography classification. Additionally, correlation between genetic variation and geography/language is stronger for Y chromosome lineage data than for *mt*DNA lineages.

Y chromosome haplogroups from five ethnic groups in Ethiopia were analyzed for phylogenetic construction. Unrooted phenogram (Figure 4.8) based on pair wise F_{ST} showed two major clusters according to the population's geographic regions. The second major cluster further divided into two sub clusters, the first contains Berta and Gumuz and the second cluster contains the Shinasha and Mao according to their local proximity. They are clustered together according to their region. Similarly, *MT-COXII* sequences from four ethnic groups in Ethiopia were analysed for phylogenetic construction. As described by Ruvolo, *et. al.* (1993), *MT-COXII* was exploited in different phylogenetic analysis in human and vertebrates giving quite informative trees topologies. In all topologies of the phylogenetic trees, each sub tree is made up of two branches. Both unrooted phylogenetic trees: Neighbor-Joining (NJ) and Minimum Evolution

(ME) based on F_{ST} of the *MT-COXII* sequences (Figure 4.9) and (Figure 4.10) respectively, showed similar topologies and branch length with few differences. Maximum Parsimony showed the evolutionary history for Ethiopian samples (Figure 4.11). As for Y chromosome, the topology of the phylogenetic trees was an indication of a distinct pattern of phylogeographic structure among the 25 sequences, suggesting that the perceived population structure is likely due to restricted contemporary gene flow. The genetic distance between the two major groups also indicate the presence of one or more barriers to dispersal that may have prevented, or be preventing, migration between these two main geographic areas.

In addition, when the sequence data of the present study combined with the global populations sequences, both NJ and ME (Not shown) phylogenetic tree showed similar topology but a little bit different from the maximum Parsimony (Figure 4.12). This study represents the first survey of *MT-COXII* polymorphisms in Ethiopian population and showed the evolutionary relationships among Ethiopian and global *MT-COXII* sequences. Here most of the Ethiopian mtDNA types cluster together on the common lineage. In all cases, tree topologies were supported, basal branching patteren, with one deep branch lead generally to Ethiopians, followed by less deep branches of Africans, and non-Africans with intercalation of few Ethiopian sequences. In a variety of phylogenetic studies aimed at constructing human mtDNA trees, this deepest clade has always lead to sub-Saharan African individuals, and with the complete absence of Ethiopian sequences in the analysis, it was impossible to draw conclusion about the root of the tree encompassing the most ancient and diverse individuals. In the present study, mtDNA tree, for the first time, has been reconstructed, and found be consistent with the number of phylogenetic trees constructed by complete mtDNA sequences with the exception of the deepest clade constituent.

6. CONCLUSION AND RECOMMENDATIONS

In the present study, LP phenotype, Y chromosome genotyping and mtDNA *COXII* sequence variation was collected and analyzed from five ethnic groups; Nuer, Berta, Gumuz, shinasha and Mao living in two regions of Ethiopia; Gambella and Benishangul Gumuz. These populations speak languages belonging to the two major Ethiopian language (Nilo-Saharan and Afro-Asiatic) and practice pastoralists and agro-pastoralists mode of subsistence patterns. In comparisions, since they are located at a remote border, these populations have been under-represented in previous studies related to lactase, *mt*DNA and Y chromosome variation in Ethiopia.

The higher LP phenotype poplymorphism was observed in the studied populations. The distribution of LP was seen to be variable in the studied population with higest frequeny (60%) observed in Berta. This frequency does not explain the expected notion that LP phenotype is frequent in pastoral groups than other. So, further research into the molecular mechanisms is needed for shedding light on the genetic basis of this trait and additional causal variants or other mechanisms of the cause of lactase persistence non-persistence. The collection of additional phenotype data from additional geographic regions in Ethiopia will be informative for understanding the distribution of LP across the country. Thus, the combination of phenotypic data and the use of next-generation sequencing technology for sequencing large genomic regions sequences will be invaluable for discovering novel mutations associated with this trait across the genome, as well as identifying possible epistatic interactions among LP-associated mutations. This would help to have a general picture of the evolutionary history of LP and its association with the origin of pastoralism in Ethiopia.

Study of Y chromosome has revealed the presence of all genotyped haplogroups namely; haplogroup A, B, E and F. In this study, the Ethiopian populations have higher frequencies of ancient haplogroup A and B and older haplogroup E. Further analysis of E-M78 revealed that the haplogroup E-M78 sub clade E1b1b1a1 was found with elevated frequencies in Nuer than the other populations, suggesting its origin in Ethiopia in agreement with previous report that it is originated in North East Africa. Interestingly, the high frequency of F-M89 was unexpected and is new report, but this need to be confirmed with microsatellites markers along with the further refinements of other major haplogroups including the presentely typed ones.

Over all, these show that the studied populations are old enough to harbor high frequencies of ancient Y chromosome haplogroups from which other haplogroups have originated and widespreaded to the other regions and Ethiopians exhibit a complex pattern of genetic variation that is likely the result of a complex history of back migrations into the region. Although this study was focused on few populations, these results indicate the spread of may be 'pastoralism' through back migration of peoples. This need to be further investigated with appropriate samples size and population coverage to get evidence for this demic diffusion which is accumulating with modern Y chromosome (and also with mtDNA) that after analysis would reveal matrilineal and patrilineal relationships in space and time. Also, future Y chromosome and mtDNA study will need to be supplemented with the study of the genetics of LP for back migration admixture from Euraisa into Ethiopian populations in relation to the spread of pastoralisim. Although, both Y chromosome and *MT-COXII* gene have different mutation feature, their study has shown that the studied populations are diverse and genetically differentiated group indicating population divergence between Ethiopian populations exist. Interestingly, also the studied population showed a close genetic affinity with each other as well as with their neighbor populations for

instance with populations from Sudan than with other populations. Moreover, the current study of the slwoly evolving *MT-COXII* gene was able to show past population expansion and this need to be continued with more samples.

In addition, general recommendations can be mentioned from the current study of Ethiopian populations. It appears that extensive genetic studies with appropriate sample size and numbers of Ethiopian ethnic groups need to be made for both Y chromosome and mtDNA. Y chromosome haplogroup need to be genotyped in detail for various important haplogroups and this haplogroup analysis should be refined using microsatellite STR marker. This would enable to obtain the general picture of the pattern of Y chromosome marker variation. With regard to mtDNA study, since the costs of generating whole-genome SNP data and next-generation sequencing data has been decreasing, genome-wide variation across ethnically and geographically diverse Ethiopian populations would be informative. Detailed analysis of Y chromosome and mtDNA with representative using these approaches could help clarify the complex genetic history of Ethiopian populations in particular and East Africa in general.

7. REFERENCES

Abdussamad, H. A. (1999). "Trading in slaves in Bela-Shangul and Gumuz, Ethiopia: border enclaves in history, 1897-1938." *Journal of African History,* **40**: 433-46.

Abdussamad, H. A. (2001). "Bela-Shangul: The frontier in History 1897-1938." 3rd East Africa History Workshop, Addis Ababa, 29th-31st october.

Agrawal, S., Faisal, K., Pandey, A., Tripathi, M. and Herrera, R. (2005). YAP, signature of an African- Middle Eastern migration into northern India. *Curr Sci*, **88**: 1977.

Akey, J.M., Zhang, K., Xiong, M., Doris, P. and Jin, L. (2001). The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am J Hum Genet.* **68**:1447-1456.

Akey, J., Zhang, G., Zhang, K., Jin, L., Shriver, M. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: Pp 1805-1814.

Alonso, S. and Armour, J.A. (2001). A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. *Proc Natl Acad Sci U S A,* **98**:864-9.

Altshuler, D. M., Gibbs, R. A. and Peltonen, L. T. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature,* **467,** 52-58.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, **290**:457-465.

Andrews, R.M., Kubacka, I., Chinnery,P.F., Lightowlers,R.N.,Turnbull,D.M. and Howell,N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat.Genet.,* **23**: 147.

Arroyo, M.A.S., Lopes, A. C., Piatto, V. B., and Maniglia, J.V. (2010). Perspectives for Early Genetic Screening of Lactose Intolerance: -13910C/T Polymorphism Tracking in the *MCM6* Gene. *The Open Biology Journal,* **3***:* 66-71

Arola, H. (1994). Diagnosis of hypolactasia and lactose malabsorption. *Scand J Gastroenterol Suppl.,* **202**: Pp 26-35.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., et al. (1987). Intraspecific hylogeography: the mitochondrial DNA bridge between population genetics and systematics. Ann Rev Ecol Syst, **18**: 489-522.

Ayala, F (1995). "The myth of Eve: molecular biology and human origin". *Science*. **270** (5244): 1930-36

Bandelt, H.J., Forster, P. and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol,* **16**:37-48.

Bahru Zewde (2002). A history of modern Ethiopia, 1855-1991. Addis Ababa: Addis Ababa University Press; *Oxford: James Currey*; Athens: Ohio University Press.

Balloux, F., Handley, L.J., Jombart, T., Liu, H. and Manica, A. (2009). Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. *Proc R Soc B,* **276**: 3447-3455.

Bamshad, M. and Wooding, S.P. (2003). Signatures of natural selection in the human genome. *Nat Rev Genet.,* **4**: 99-111.

Batini, C., Ferri, G., Destro-bisol, G. et al (2011): Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Mol Biol Evol,* **28**: 2603-2613.

Bayoumi, R., Flatz, S.D., Kuhnau, W., Flatz, G. Beja and Nilotes (1982): Nomadic pastoralist groups in the Sudan with opposite distributions of the adult lactase phenotypes. *AmI Phys Anthropol,* **58**: 173-8.

Behar,D.M., van Oven,M., Rosset,S., Metspalu,M., Loogv¨ali,E.-L., Silva,N.M., Kivisild,T., Torroni, A. and Villems, R. (2012). A 'Copernican' reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.*, **90**: 675-684.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet*., **74**: 1111–1120.

Boll, W., Wagner, P., Mantei, N. (1991). Structure of the chromosomal gene and cDNAs coding For lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. *Am. J. Hum. Genet.* **48**:Pp 889–902.

Bowcock,A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd,K. and Cavalli-Sforza, L.L. (1991). Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc Natl Acad Sci U S A.* **88(3)**: 839-843.

Browning, S.L. (2008). Human Genetic Variation with Implications for Healthcare in Ethiopian Populations. *The Centre for Genetic Anthropology*, University College London

Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. (2007). Absence of the LP associated allele in Early Neolithic Europeans. *Proc. Natl Acad. Sci. USA*, **104**: 3736-3741. (doi:10.1073/pnas.0607187104)

Burger, J. and Thomas, M. G. (2011). The palaeopopulation genetics of humans, cattle and dairying in Neolithic Europe. In The bioarchaeology of the transition to agriculture (*eds R. Pinhasi & J. Stock*), Pp. 371–384. Chichester, UK: Wiley Blackwellnce.1124309)

Campbel, M. C., & Tishkoff, S. A. (2008). African Genetic Diversity:Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. of Genom. and Hum. Genet*., **9**: 403-433.

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

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B. and Cambon-Thomsen, A. (2002). Human genome Diversity cell line panel. *Science,* **296**: 261-262.

Capelli, C., Wilson, J.F., Richards, M., Stumpf, M.P., Gratrix, F., Oppenheimer, S., Underhill, P., Pascali, V.L., Ko, T.M. and Goldstein, D.B. (2001). A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am. J. Hum. Gen.*, **68:** 432-443.

Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., Bishop, C,. Fellous, M. and Purrello, M. (1985). A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science*, **230**:1403-1406.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1993. Demic expansions and human evolution. *Science*, **259**:639-646.

Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. (1994). The history and geography of human genes. *Princeton University Press*, Princeton.

Cavalli-Sforza, L.L. (1997). Genes, peoples, and languages. *Proceedings of the National Academy of Sciences of the United States of America*, **94**: 7719-7724.

Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J. (1998). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci US A*, **85**:6002-6.

Cavalli-Sforza, L.L. (1998). The Chinese human genome diversity project. *Proceedings of the National Academy of Sciences of the United States of America*, **95**: 11501-11503.

Cavalli-Sforza, L.L. and Feldman,M.W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet (Suppl)*, **33**: 266-275.

Cela-Conde, C. J., and Ayala, F. J. (2007). Human Evolution: Trails from the past. *Oxford. New York: Oxford University Press*.

Coelho, M., Luiselli, D., Bertorelle, G., Lopes, A.I., Seixas, S., Destro-Bisol, G., and Rocha, J. (2005). Microsatellite variation and evolution of human lactase persistence. *Hum. Genet.*, **117**: 329-339.

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

Chen, Y.S., et al. (2000). mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am J Hum Genet*, **66**(4):1362-83.

Clark, J.D., Beyene,Y. and WoldeGabriel, G. (2003). Stratigraphic, chronological and behavioural contexts of Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature*. **423:** 747-752.

ClustalX, version 1.81 (1997). Software for Multiple Alignment of Nucleic Acid and Protein Sequences. *The Conway Institute*: Dublin, Ireland.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P. and *et. al*. (2002). A back migration from Asia to sub-Saharan Africa is supported by highresolution analysis of human Ychromosome haplotypes. *Am. J. Hum. Genet.,* **70:** 1197-1214.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R. and *et.al*. (2004). Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am. J. Hum. Genet.,* **74:** 1014-1022.

Cruciani, F., La Fratta, R., Torroni, A., Underhill, P.A. and Scozzari, R. (2006). Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite- network-based approach through six new biallelic markers. *Hum Mutat,* **27**: 831-832.

Cruciani, F., La Fratta, R.,Trombetta, B., Santolamazza, P., Sellitto, D.,Colomb, E.B., Dugoujon, J.M., Crivellaro, F., Benincasa, T., Pascone, R. and *et. al*. (2007). Tracing past human male movements in northern/East Africa & western Eurasia: New clues from Y-chromosomal haplogroups E-78 and J-M12. *Mol Biol Evol*, **24**: 1300-1311.

Cruciani, F., Trombetta, B., Massaia, A., Destro-Bisol, G., Sellitto, D. and Scozzari, R. (2011). A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am J Hum Genet*., **88(6**):814-8.

Diamond, J. and Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science*, **300**: 597–603.

Dissanyake, A.S., El-Munshid, H.A. and Al-Qurain, A. (1990). Prevalence of primary adult lactose malabsorption in the eastern provine of Saudi Arabia. *Ann Saudi Med*, **10**:598-601.

DnaSP (2009) version 5.10. Software for Comprehensive Analysis of DNA Polymorphism Data. University of Barcelona, Barcelona, Spain.

Durham, W. H. (1991). Co-evolution: Genes, Culture and Human Diversity: A classic anthropological text that investigates the relationship between genes and culture. *Stanford Univ. Press*.

Elhassan, N., Gebremeskel, E.I., Elnour, M.A., Isabirye, D., Okello, J., Hussien, A., Kwiatksowski, D., Hirbo, J., Tishkoff, S., Ibrahim, M.E. (2014). The Episode of Genetic Drift Defining the Migration of Humans out of Africa Is Derived from a Large East African Population Size. PLoS ONE 9(5): e97674. doi:10.1371/journal.pone.0097674

Elliott, H.R., Samuels, D.C., Eden, J.A., Relton, C.L. and Chinnery, P.F. (2008). Pathogenic mitochondrial DNA mutations are common in the general population. *Am. J. Hum. Genet.*, **83**: 254-260.

Elson, J.L., Turnbull, D.M. and Howell, N. (2004). Comparative Genomics and the Evolution of Human mtDNA: Assessing the Effects of Selection. *Am J Phys Anthropol*, **74**: 229-238.

Enattah, N.S, Sahi, T., Savilahti, E., Terwilliger, J.D. and Peltonen, L. (2002). Identification of a variant associated with adult-type hypolactasia. *Nat Genet*., **30**:233-237.

Enattah, N.S., Trudeau, A., Pimenoff, V., Maiuri, L. and Auricchio, S. (2007). Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *Am J Hum Genet*., **81**: 615-625.

Enattah, N.S., Jensen, T.G., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J.K., Alifrangis, M., Khalil, I.F., *et. al.* (2008). Independent introduction of two Lactase-persistence allele into human populations reflects different history of adaptation to milk culture. *Am. J. Hum.Genet.*, **82**: 57-72.

Encyclopedia Britannica (1964). William Benton, London, Chicago, Geneva, Sydney, Toronto.

Ethiopian Central Statistical Authority Report (2007). Population and Housing Result, a first draft, Addis Ababa.

Evans (1956). Nuer Religion, New York: *Oxford University Press*: Clarendon Press.

Excoffier, L., Pellegrini, B., Sanchez-Mazas, A., Simon, C. and Langaney, A. (1991). Genetics and history of sub-Saharan Africa. *Yearbook Phys Anthropol*, **30**:151-194

Excoffier, L., Laval, L.G. and Schneide, L. (2005). Arlequin Version 3.0: an integrated software Package for population genetics data analysis. *Evol Bioinform Online,* **1**:47-50

Fagundes, N.J., Bonatto, S.L., Callegari-Jacques, S.M. and Salzano, F.M. (2002). Genetic, geographic, and linguistic variation among South American Indians: possible sex influence. *Am J Phys Anthropol,* **117**: 68-78.

Finneran, N. (2007). The Archeology of Ethiopia. *London and New YorkRoutledge,* Tylor and Francis Group.

Flatz, G. (1987). Genetics of lactose digestion in humans. In:Harris, H., Hirschhorn, K. (eds) Advances in human genetics. *Plenum Press, New York,* **16**. 1-77

Flatz, G. (1989). The genetic polymorphism of intestinal lactase activity in adult humans. In C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle (Eds.), The metabolic basis of inherited disease, (6th ed). New York7 McGraw-HillFraumene, C., Belle, E.M.S., Castrı, L., Sanna, S., Mancosu, G., *et. al*. (2003). High Resolution Analysis and Phylogenetic Network Construction Using Complete mtDNA Sequences in Sardinian Genetic Isolates. *Mol Biol Evol*., **23**: 2101-2111.

Gebremeskel,E.I. and Ibrahim, M.E. (2014). Y-chromosome E haplogroups: their distribution and implication to the origin of Afro-Asiatic languages and pastoralism. *Eur J Hum Genet*. **22**:1387-1392.

Gebre Muluneh (2004). Traditional practices among the Berta nationality. http://www.addistribune.com/Archives

Gomes, V., Sa´nchez-Diz, P., Amorim ,A., Carracedo, A. and Gusma˜o, L. (2010). Digging deeper into East African human Y chromosome lineages. *Hum Genet*., **127**:603-613.

Gerbault, P., Moret, C., Currat, M., Sanchez, M. A. (2011). Impact of selection and demography on the diffusion of lactase persistence. *PLoS One* 4:e6369

González-Ruibal, A. and Fernández, M. (2005). Exhibiting Cultures of Contact: A Museum for Benishangul-Gumuz, Ethiopia . *Stanford Journal of Archaeology.*

Grottanelli and Vinigi, L. (1948). "I Pre-Niloti: una arcaica provincia culturalle in Africa." *Annali Lateranensi,* **12**: 280-326.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95／98／NT. *Nucleic Acids SymposiumSeries*, **41**: 95-98.

Hammer, M.F. (1994). A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. *Mol Biol Evol.,* **11**:749-761.

Hammer, M.F. and Horai, S. (1995). Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet*, **56**: 951-962

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

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

Hammer, Ø., Harper, D.A.T. and Ryan, P.D. (2001). PAST: Paleontological statistics software package for education and data analysis. . *Palaeontologia Electronica* 4**,** 9.

Hammer, M. F., 1994 A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol*. **11**: 749-761

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

Hammer, M.F. and Zegura, S.L. (1996). The role of the Y chromosome in human evolutionary studies. *Evol. Anthropol.* **5:** 116-134.

Hammer, M.F. and Zegura, S.L. (2002). The human Y chromosome haplogroup tree: Nomenclature & phylogeography of its major divisions. *Annu. Rev. Anthropol.* **31:**303-321.

Harvey, C. B., Fox, M. F., Jeggo, P. A., Mantei, N., Povey, S. and Swallow, D. M. (1993). Regional localization of the lactase-phlorizin hydrolase gene, *LCT*,to chromosome 2q21. *Ann. Hum. Genet.*, **571**: 79-85.

Harvey, C.B., Pratt, W.S., Islam, I., Whitehouse, D.B., Swallow, D.M. (1996). DNA polymorphisms in the lactase gene. Linkage disequilibrium across the 70-kb region. *Eur J Hum Genet.*, **3**:27-41.

Harvey, C.B., Hollox, E.J., Poulter, M., Wang, Y., Rossi, M., Auricchio, S., Iqbal, T.H., Cooper, B.T., Barton, R., Sarner, M., et al. (1998). Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. *Ann. Hum. Genet.*, **62**: 215-223.

Hassan, H.Y., Underhill, P.A., Cavalli-Sforza, L.L. and Ibrahim, M.E. (2008). Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography and history. *Am J Phys Anthropol* , **137**: 316- 323.

He, Y., Wu, J., Dressman, D.C., Iacobuzio-Donahue, C., Markowitz, S.D., Velculescu, V.E., Diaz, L.A. Jr, Kinzler, K.W., Vogelstein, B. and Papadopoulos, N. (2010). Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature,* **464**: 610-614.

Hibro, J.B. (2011). Complex genetic Histpry of East African Human Populations. *PhD Thesis.*

Hijazi, S.S., Abulaban, A., Ammarin, Z. and Flatz, G. (1983). Distribution of adult lactase phenotypes in Bedouins and in urban and agricultural populations of Jordan. *Trop. Geogr. Med.,* **35**: 157-161.

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

Holden, C. and Mace, R. (1997). Phylogenetic analysis of the evolution of lactose digestion in adults. *Hum Biol,* **69**:605-628.

Holden, C., and Mace, R. (2003). Spread of cattle led to the loss of matrilineal descent in Africa: A coevolutionary analysis. *Proc. R. Soc. Lond. B.* **270**:2425-2433.

Hollox, E.J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A.I., and Swallow, D.M. (2001). Lactase haplotype diversity in the Old World. *Am. J. Hum. Genet.* **68**, 160-172.

Hurles, M.E., Nicholson, J., Bosch, E., Renfrew, C., Sykes, B.C., and Jobling, M.A. (2002). Y chromosomal evidence for the origins of Oceanic-speaking peoples. *Genetics,* **160:** 289-303.

Hurles, M.E., Sykes, B.C., Jobling, M.A. and Forster, P. (2005). The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages. *Am J Hum Genet,* **76**: 894-901. doi:10.1086/430051. PubMed: 15793703.

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

Ingram, C.J., Elamin, M.F., Mulcare, C.A., Weale, M.E., Tarekegn, A., Raga, T.O., Bekele, E., Elamin, F.M., Thomas, M.G., Bradman, N., and Swallow, D.M. (2007). A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence. *Hum. Genet*., **120**: 779-788.

Ingram, C.J., Raga, T.O., Tarekegn, A., Browning, S.L., Elamin, M.F., Bekele, E., Thomas, M.G.,Weale, M.E., Bradman, N., and Swallow, D.M. (2009). Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *J. Mol. Evol.*, **69**: 579-588.

Itan, Y., Powell, A., Beaumont, M.A., Burger, J., and Thomas, M.G. (2009). The origins of lactase persistence in Europe. *PLoS Comput*. Biol., **5**: e1000491.

Itan, Y., Jones, B. L., Ingram, C. J., Swallow, D. M. & Thomas, M. G. (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol*. Biol., **10**: 36.

James, W. (1980). "From aboriginal to frontier society in western Ethiopia." In *Working* papers on society and history in Imperial Ethiopia: The southern periphery from 1880 to *1974*, edited by Donham,D.L and James, W. Cambridge: *African Studies Center, Cambridge University Press.*

James, W. R. (1986). Lifelines: exchange marriage among the Gumuz. In Donham, Donald L. and James, Wendy R. (eds.), The southern marches of Imperial Ethiopia: essays in history and social anthropology, 119-147. *Cambridge Univ. Press*.

Jensen, T.G., Liebert, A., Lewinsky, R., Swallow, D.M., Olsen, J. (2011). The 214010*C variant associated with lactase persistence is located between an Oct-1 and HNF1a binding site and increases lactase promoter activity. *Hum. Genet.,* **130**: 483-493.

Jobling, M.A. and Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet,* **4:** 598-612.

Jobling, M. A., Hurles, M. E., and Tyler-Smith, C. (2004). Human Evolutionary Genetics: Origins, Peoples & Disease. New Delhi: *Garland Science; Tylor & Francis groups*.

Jobling, M.A. (2012). The impact of recent events on human genetic diversity. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**: 793-799

Johanson, D. (1986). On the Phylogenetic Analysis of Early Hominids. *Currenet Antropology*, **29 (4):** 361. DOI: 10.1086/203448.

Jones, B. L., . Raga, T. O, Liebert, A., Zmarz, P., Bekele, E. Danielsen, E.T., Olsen, A. K., Bradman, N., Troelsen, J.and Swallow. D. M. (2013). Diversity of Lactase Persistence Alleles in Ethiopia: Signature of a Soft Selective Sweep. *Am J Hum Genet*., **93**: 538-544.

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

Just, R.S., Leney, M.D., Barritt, S.M., Los, C.W., Smith, B.C., Holland, T.D. and Parsons, T.J. (2009). The use of mitochondrial DNA single nucleotide polymorphisms to assist in the resolution of three challenging forensic cases. *J. Forensic Sci.,* **54**: 887-891.

Karafet, T.M., Xu, L., Du, R., Wang, W., Feng, S., Wells, R.S., Redd, A.J., Zegura, S.L. and Hammer, M.F. (2001). Paternal population history of East Asia: Sources, patterns, and microevolutionary processes. *Am. J. Hum. Genet.* **69:** 615-628.

Karafet, T.M., Osipova, L.P., Gubina, M.A., Posukh, O.L., Zegura, S.L., et. al. (2002). High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol*, **74**: 761-789.

Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L. Hammer, M. F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*. **18**: 830–838.

Kaestle, F.A. and Horsburgh, K.A.(2002). Ancient DNA in anthropology: Methods, Applications, and Ethics. *Am J Phys Anthropol*, **45**:92-130

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 geneflow across and around the gate of tears. *Am J Hum Genet*, **75**:752-770.

Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., *et. al.* (2006). The role of selection in the evolution of human mitochondrial genomes. *Genetics*, **172**: 373-387.

Knight, A., Underhill, P. A., Mortesen, H. M., Zhivotovsky, A., Lin, A. A., Henn, B. M., Louis, D., Ruhlen, M., and Mountain J. L. (2003). African Y Chromosome and mtDNA Divergence Provides Insight into the History of Click Languages. *Current biology*, **13**: 464- 473

Kretchmer, N., Ransome-Kuti, O., Hurwitz, R. (1971). Intestinal Absorption of Lactose in Nigerian Ethnic Groups. *The Lancet*. 392 - 395

Krings, M., Salem, A.E.H., Bauer, K., et. al. (1999). mtDNA analysis of Nile River Valley populations: a genetic corridor or a barrier to migration*? Am J Hum Genet*, **64**:1166-1176.

Kruse, T.A., Bolund, L., Grzeschik, K.H., Ropers, H.H., Sjostrom, H., Noren, O., Mantei, N., Semenza, G. (1988). The human lactase-phlorizin hydrolase gene is located on chromosome 2. *FEBS Lett.* **240**: 123-126

Khurana, P., Aggarwal, A., Mitra, S., Italia, Y.M., Saraswathy, K.N., *et al.* (2014) Y Chromosome Haplogroup Distribution in Indo-European Speaking Tribes of Gujarat, Western India. *PLoS ONE,* **9(3):** e90414. doi:10.1371/journal.pone.0090414

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

Lewin, P.K. (1987). A unique ancient Egyptian mummified head, demonstrating removal of the brain from the foramen magnum. *Paleopathol. Newsl.*, **57**: 12-13.

Lewis, M. P. (2009). Ethnologue: Languages of The World (Sixteenth edition ed.). Dallas, Tex: *SIL International*.

Levine, D. N. (1974). Greater Ethiopia: The Evolution of a Multiethnic Society, 2nd ed., Chicago and London, *University of Chicago Press*.

Li, M., Schonberg, A., Schaefer, M., Schroeder, R., Nasidze, I. and Stoneking, M. (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet*., **87**: 237-249.

López, S., va Dorp, L. and Hellenthal, G. (2015). Human Dispersal Out of Africa: A Lasting Debate. *Evolutionary Bioinformatics.* **11**(s2) 57–68 doi: 10.4137/EBo.s33489.

Lopez, S. and Hellenthal, G. (2017). Exploring the geneteics and ancestry of peoples of Ethiopia. University College London (UCL), *WC1E 6BT, London (UK).*

Luis, J. R., Rowold, D. J., Regueirol, M., Caeiro, B., Cinnioglu, C., *et.al*. (2004). The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am. J. Hum. Genet.,* **74:** 532-544.

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

McCracken, R. D. (1971). Lactase deficiency: an example of dietary evolution**.** *Curr. Anthrop*. **12**: 479-517.

Maddison, D.R., Ruvolo, M. and Swofford, D.L.(1992). Geographic origins of human mtDNA: phylogenetic evidence from control region sequences. *Syst Biol*. **41**:111-124.

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

Mantei, N., Villa, M., Enzler, T., Wacker, H., Boll, W., James, P., Hunziker, W., Semenza, G. (1988). Complete primary structure of human and rabbit lactase-phlorizin hydrolase: implications for biosynthesis, membrane anchoring and evolution of the enzyme. *EMBO J*. **7:** Pp 2705-2713.

Metz, G., Jenkins, D.J, Peters, J.J., Newman, A. and Blendis, L.M. (1975). Breath hydrogen as a diagnostic method for hypolactasia. *Lancet*. **7917**: 1155-1157.

Midgley, M.S. (1992). TRB culture: the first farmers of the North European plain. *Edinburgh* University Press, Edinburgh.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., et. al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci U S A,* **100**: 171-176. doi:10.1073/pnas.0136972100.

MITOMAP (2007). A Human Mitochondrial Genome Database. http://wwwmitomaporg

Mulcare, C.A., Weale, M.E., Jones, A.L., Connell, B., Zeitlyn, D., Tarekegn, A., Swallow,D.M., Bradman, N., and Thomas, M.G. (2004). The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am. J. Hum. Genet*. **74**: 1102-1110.

Myles, S., Bouzekri, N., Haverfield, E., Cherkaoui, M., Dugoujon, J.M., andWard, R. (2005). Genetic evidence in support of a shared Eurasian-North African dairying origin. *Hum. Genet*.**, 117**: 34-42.

Negaso Gidada (2001). History of the Sayyoo Oromoo of Southwestern Wallaga, Ethiopia from about*1730 to 1886*. Addis Ababa: *Mega Printing Enterprise*.

Nei, M. (1978). The theory of genetic distance and evolution of human races. *Jinrui Idengaku Zasshi*. **23(4)**:341-369

Ngo, K.Y., Vergnaud, G., Johnsson, C., Lucotte, G. and Weissenbach, J. (1986). A DNA probe detecting multiple haplotypes of the human Y chromosome. *Am J Hum Genet*, **38**: 407-418.

Olds, L.C. and Sibley, E. (2003). Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Hum.Mol. Genet.,* **12**: 2333-2340.

Olds, L.C., Ahn, J.K., and Sibley, E. (2011). 13915*G DNA polymorphism associated with lactase persistence in Africa interacts with Oct-1. *Hum. Genet*., 129: 111-113.

Oljira, T (2014). Lactose Digestion and Variation in the Enhancer Region of Lactase Gene in the Context of Some Ethiopian Populations. *PhD Thesis, Addis Ababa University*.

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., Bekele, E., Bradman, N., Balding, D.J. & Tyler-Smith, C. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am J Hum Genet*, **91:** 83-96.

Passarino, G., Semino, O., Quninana-Murci, L., Excoffier, L., Hammer, M. and Santachiara-Benerecetti, A. S. (1998). Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphysim. *Am. J.Hum. Genet.,* **62**: 420-434.

Pakendorf, B. and Stoneking, M.(2005). Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet*, **6**:165-183.

Pereira, L., Freitas, F., Fernandes, V., Pereira, J.B., Costa, M.D., Costa, S., Maximo, V., Macaulay,V., Rocha, R. and Samuels, D.C. (2009). The diversity present in 5140 human mitochondrial genomes. *Am. J. Hum. Genet.,* **84**: 628-640.

Pericic,´ M., Lauc, L.B., Klaric,´ I.M. et. al (2005). High-resolution phylogenetic analysis of southeastern Europe traces major episodes of paternal gene flow among Slavic populations. *Mol Biol Evol*, **22**: 1964– 1975.

Phillipson, D. W. (1993). African Archeology (2nd ed.). Cambridge: *Cambridge University Press.*

Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Guldermann, B.K., Mpoloka, S.W.,Nagawa,H.,Naumann, C.,Lipson, M., Loh, P., Lachance, J., Mountain, J., Bustamante, C.D., Berger, B., Tishkoff, S.A., Henn, B.M., Stoneking, M., Reich, D., and Pakendorf, B. (2013). The genetic prehistory of southern Africa. *Nature communications*. **3:** 1143.

Plaster, A.C. (2011). Variations in Y chromosome, mitochondrial DNA and labels of Identity in Ethiopaia. *PhD Theiss, University College london*

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

Poloni, E.S., Naciri, Y., Bucho, R., Niba, R., Kervaire, B., Excoffier, L., Langaney, A. and Sanchez-Mazas, A. (2009). Genetic evidence for complexity in ethnic differentiation and history in East Africa. *Ann Hum Genet.,* **73:** 582-600.

Poulter, M., Hollox, E., Harvey, C.B., Mulcare, C., Peuhkuri, K., Kajander, K., Sarner, M., Korpela, R., Swallow, D.M. (2003). The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Ann. Hum. Genet.* **67**: 298-311.

Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., *et. al*.(2002). Y-chromosomal DNA variation in Pakistan. *Am. J. Hum. Genet.,* **70**: 1107-1124.

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K. and Santachiara-Benerecetti, A. S. (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through East Africa. *Nat Genet,* **23**: 437- 41.

Relethford, J.H. (1995). Genetics and modern human origins. *Evol Anthropol,* **4**:53-63.

Relethford, J.H. (2001). Absence of regional affinities of Neandertal DNA with living humans does not reject multiregional evolution. *Am J Phys Anthropol*, **115**:95-8.

Robayo-Torres, C.C., Quezada-Calvillo, R., Nichols, B.L. (2006). Dissaccharide digestion: clinical and molecular aspects. *Clin Gastroenterol Hepatol*, **4:**276-287.

Romero, I. G., Mallick, C.B., Liebert, A., Crivellaro, F. Chaubey, G., Itan, Y., Metspalu, M. Eaaswarkhanth, M. Pitchappan, R., Villems, R., Singh, L., Thangaraj, K.,Thomas, M. G., Swallow, D.M., Lahr, M. and Kivisild ,T.(2011). Herders of Indian and European Cattle Share Their Predominant Allele for Lactase Persistence. *Mol. Biol. Evol.,* **29(1):** 249-260.

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

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

Rosser, Z.H., et.al. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet*, 67**(6):** 1526-43.

Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., Bakker, E., Krawczak, M., de Knijff, P. (1996). Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. *Hum Mol Genet*, **5**:1029-1033

Ruvolo, M., Zehr, S., von Dornum, M., Pan, D., Chang, B. and J. Lin, J. (1993). Mitochondrial COII sequences and modern human origins. *Mol. Biol. Evol.,* **10:** 1115-1135.

Sabeti, P.C, Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel,S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D.,Ward, R. and Lander, E.S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature.,* **419**: 832-837.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1990). Molecular Cloning: A laboratory Manual (2nd Edition). Cold Spring Harbor Laboratory, New York.

Sanchez, J.J., Borsting, C. and Morling, N. (2005). Typing ofY chromosome SNPs with Multiplex PCR methods.Methods. *Mol Biol*, **297**: 209-228.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., *et.al*. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, **423**: 825-37. | www.nature.com/nature

Scrimshaw, N. and Murray, E. (1988). The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. *Am. J. Clin. Nutr*. **48**: 1079-1159

Scozzari, R., Cruciani, F., Santolamazza, P., Malaspina, P., Torroni, A., Sellitto, D., Arredi, B., Destro- Bisol, G., De Stefano, G., Rickards, O., Martinez-Labarga, C., Modiano, D., Biondi, G., Moral, P., Olckers, A., Wallace, D.C. and Novelletto, A. (1999). Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet*, **65**: 829-846.

Scozzari, R., Cruciani, F., Pangrazio, A., Santolamazza, P., Vona, G., *et. al*. (2001). Human Y-chromosome variation in the western Mediterranean area: implications For the peopling of the region. *Human immunology*, **62**: 871-884.

Seielstad, M,T., Hebert, J.M., Lin, A.A., Underhill, P.A., Ibrahim, M., Vollrath, D. and Cavalli-Sforza, L.L.(1994). Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum Mol Genet*, **3**:2159-2161

Seielstad, M.T., Minch, E. and Cavalli-Sforza, L.L. (1998). Genetic evidence for a higher female migration rate in humans. *Nat.Genet.* **20:** 278–280

Semino, O., et. al. (2002). Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet*, **70**(1): 265-8.

Semino, O., Magri, C., Benuzzi, G., Lin, A.A., Al-Zahery, N., et. al. (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on neolithization of Europe & later migratory events in the Mediterranean area. *Am J hum genet*. **74**:1023-1034.

Shi, H., Zhong, H., Peng, Y., Dong, Y.-L., Qi, X.-B., Zhang, F., Liu, L.-F., Tan, S.-J., Ma, R.Z., Xiao, C.-J., Wells, R.S., Jin, L. and Su, B. (2008). Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biology*, **6:** 45.

Simoons, F. (1969). Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. I. Review of the medical research. *Am. J. Dig. Dis.,* **14**: 819-836.

Simoons, F. (1970). Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. II. A culture historical hypothesis. *Dig. Dis. Sci.,* **15**: 695-710.

Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J. and Barbujani, G. (2000). Geographic patterns of mtDNA diversity in Europe. *Am J hum genet*, **66**: 262-278.

Shen, P. *et .al*. (2000). Population genetic implications from sequenc variation in four Y chromosome genes. *Proc. Natl Acad. Sci.USA*, **97:** 7354-7359.

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

Stringer, C. (1994). Out of Africa-a personal history. ***In***: *Origins of anatomically modern humans* (ed. M. Nitecki & D. Nitecki), pp. 149–172. New York: Plenum.

Swallow, D.M. (2003) Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet*, 37: 197-219.

Tamura, Y. Shimada, K. and Hibi, K. (1993). Wind response of a tower (typhoon observation at Nagasaki Huis Ten Bosch Domtoren). *J. Wind Eng. Ind. Aerodyn*., **50**: 309-318

Tamura, K. and Nei, M, (1993). Estimation of the Number of Nucleotide Substitutions in the Control Region of mtDNA in Humans and Chimpanzees. *Mol Biol Evol*, **10**: 512-526.

Tang, S., Wang, J. Zhang, V.W., Li, F.Y., Landsverk, M., Cui, H., Truong, C.K., Wang, G., Chen, L.C., Graham, B., Scaglia, F., Schmitt, E.S., Craigen, W.J. and Wong, L.J.C. (2013). Transition to Next Generation Analysis of the Whole Mitochondrial Genome: A Summary of Molecular Defects. **H***um Mutat*, Wiley Periodicals, Inc. **34**:882-893.

Taylor, R.W. and Turnbull, D.M. (2005). Mitochondrial DNA mutations in human disease. *Nat Rev Genet*, **6(5)**:389-402

Tishkoff , S.A. and Williams, S.M. (2002). Genetic Analysis of African Populaations: Human Evolution and Complex Disease. *Nat Rev Genet,* **3**: 611-661. doi:10.1038/nrg865.

Tishkoff, S.A., Dietzsch, E., Speed,W., Pakstis, A.J., Kidd, J.R., Bonne´-Tamir, B., Santachiara-Benerecetti, A.S., et. al. (1996). Global patterns of linkage disequilibrium at the CD4 locus And modern human origins. *Science*, **271**:1380-1387.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet*. **39**: 31-40.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., *et.al.* (2009), The Genetic Structure and History of Africans and African Americans. *Science,* **324**: 1035-1044.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature,* **437**:1299-1320. [PubMed: 16255080]

Thomas, M.G., Parfitt, T., Weiss, D.A.,et. al. (2000). Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lembathe 'Black Jews of Southern Africa'. *Am J Hum Genet,* **66**: 674- 686.

Torroni, A., Achilli, A., Macaulay, V., Richards, M. and Bandelt, H. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*. 22:339-45.

Trinkaus, E. (2005). Early modern humans. *Annu. Rev. Anthropol*., **34**: 207-230.

Triulzi, A. (1901). Salt, Gold and Legitimacy. Prelude to the history of a no-man's land. Bela Shangul, Wallaga, Ethiopia (ca. 1800-1898). Naples: *Istituto Universitario Orientale*.

Troelsen, J.T., Olsen, J., Møller, J. and Sjo¨stro¨m, H. (2003). An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology*, **125**: 1686-1694.

Tsega Engalew (2002). "Luba Basa and Harma Hodha: traditional mechanisms of conflict resolution in Metekel, Ethiopia." *Social Science Research Report Series* 24, Addis Ababa. (http://www.ossrea.net/publications)

Tully, L.A., Parsons, T.J., Steighner, R.J., Holland, M.M., Marino, M.A. and Prenger, V.L. A sensitive denaturing gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy inhypervariable region 1 of the human mtDNA control region. *Am J Hum Genet*., **67(2)**: 432-443.

Underhill, P.A., Jin, L., Zemans, R., Oefner, P.J. and Cavalli-Sforza, L.L. (1996). A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc Natl Acad Sci USA*, **93**:196-200.

Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., et. al. (2000). Y chromosome sequence variation and the history of human populations. *Nat Genet*, **26**: 358-361.

Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Mirazon Laher, M., Foley, R. A., et. al. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.*, **65**: 43-62.

Underhill, L. G., Crawford, R. J. M., Wolfaardt, A. C., Whittington, P. A., Dyer, B. M., Leshoro, T. M., Ruthenberg, M., Upfold, L., and Visagie, J. (2006). Regionally coherent trends in colonies of African Penguins Spheniscus demersus in the Western Cape, South Africa, 1987-2005. *African Journal of Marine Science*. **28**: 697-704.

Underhill, P.A. and Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Ann. Rev. Genet.*, **41**:539-564.

Y Chromosome Consortium (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res*, **12:** 339-348.

Ye, S., Dhillon, S., Ke, X., Collins, A.R. and Day, I.N. (2001). An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res*, **29**:88-98.

You, Z., Komamura, Y. and Ishimi, Y. (1999). Biochemical analysis of the intrinsic Mcm4-Mcm6- Mcm7 DNA helicase activity. *Mol Cell Biol*, **19(12)**:8003-8015

van Oven, M., Van Geystelen, A., Kayser, M., Decorte, R. and Larmuseau, M.H. (2014). Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Hum Mutat*., **35**:187-191.

Wallace, D.C. (2015). 1Mitochondrial DNA Variation in Human Radiation and Disease. *Elsevier Inc, Philadelphia, PA 19104, USA*. *Cell.* 163.

Wang, Y., Harvey, C. B., Pratt, W. S., Sams, V. R., Sarner, M., Rossi, M., Auricchio, S. & Swallow, D. M. (1995). The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. *Hum. Mol. Genet.*, **4**: 657-662.

Ward, R.H.,Frazier, B.L., Dew-Jager, K. and Paabo, ***S.*** (1991). Extensive mitochondrial diversity within **a** single Amerindian tribe. *Proc Natl Acad Sci USA,* **88**:8720-8724.

Weale, M.E., Shah, T., Jones, A.L., Greenhalgh, J., Wilson, J.F., Nymadawa, P., Zeitlin, D., Connell, B.A., Bradman, N. and Thomas, M.G. (2003). Rare deep-rooting Y chromosome lineages in humans: lessons for phylogeography. *Genetics*, **165(1)**:229-234.

Weir, B. S. (1990). Genetic data analysis: Methods for discrete population genetic data. *Sinauer Associates, Inc. Publishers, USA.*

Wilson, A. C., Cann, R.L., Carr, S.M., George, M., Gyllenstekn, U.B.,Helm-bychowskr.I .M., Higuchis, G., Palumrei, R., Pragerr, M., Sage, D. and Stoneking, M. (1985). Mitochondrial DNA and two perspectives on evolutionar genetics. *Biol. J. Linn. SOC*., **26:** 375-400.

Wilson, J.F., Weal, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bransman, N and Goldstein, D.B. (2001). "Population genetic structures of variable drug response. *Nat Genet*, **29 (3):** 265-9.

Whitfield, L.S., Sulston, J.E. and Goodfellow, P.N. (1995). Sequence variation of the human Y chromosome. *Nature*, **378**: 379-380.

White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G., *et. al*. (2003). Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature,* **423:** 742-747.

White, T. D., Asfaw, B., Beyene, Y., Haile-Selassie, Y., Lovejoy, C. O., Suwa, G., *et. al.* (2009). Ardipithecus ramidus and the paleobiology of early hominids. *Science,* **326**: 75-86.

Wood, E.T., Stover, D.A., Ehret, C., Destro-Bisol, G., Spedini, G., et, al. (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *European journal of human genetics* : *EJHG*, **13**: 867-876.

Wolde-Selassie Abbute (2005). Gumuz and Highland Resettlers: Differing Strategies of Livelihood and Ethnic Reaction in Metekal, Northwestern Ethiopia. Göttingen: *Gottinger Studien Zur Ethnologie*.

Zegura, S.L., Karafet, T.M., Zhivotovsky, L.A., and Hammer, M.F. (2004). High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol. Biol. Evol.*, **21:** 164-175.

Zietkiewicz, E., Yotova, V., Jarnik, M., Korab-Laskowska, M., Kidd, K. K., Modiano, D., Scozzari, R., Stoneking, M., Tishkoff, S., Batzer, M. & Labuda, D. (1997). Nuclear DNA diversity in worldwide distributed human populations. *Gene,* **205**: 161-171.

8. APPENDICES

Appendix 1: Lactase Phenotype related questionnaire to collect socio-demographic and some environmental features from each study participant.

I. Personal information: please tell the following?

Gender: a) male b) Female

Age _____ - Ethnicity _____

Language _____ -Region _____

II. Milk related information

1. How often do you have a milk drink? a) Never b) Daily c) Rarely
2. If you consume milk daily, how much milk do you take per day in liter

a) ½ b) 1 c) 1 and ½ d) 2

3. Have you ever aware of any self – perceived gastrointestinal symptoms and discomforts after drinking milk? a) Yes b) No

4. If your answer is 'yes' to question number '3' above, please indicate which symptoms?

a) Diarrhea b) Abdominal cramping C) Vomiting d) Flatulence e) Audible bowel sounds

5. Do you have a blood feeding habit? A) Yes_____ B) No_____

6. Have you engaged in the production of milk from dairy animals? a) yes b) No

7. If your answer is yes to question number '5' above, please indicate the types of milk produced and consumed frequently? a) Goat b) Cow c) Camel d) others (specify) _____

Appendix 2: Sample Table/forms used to record HBT results, collect post BHT symptoms associated with lactase intake, adverse effects (if happened): G**umuz Participants**

Participants code no	Times of breath hydrogen sampling (by ½ hr. intervals)							Post Lactose load symptoms (post gastrointestinal illness)						
	0	½	1	$1^{1/2}$	2	$2^{1/2}$	3	Di	Ap	Vm	fla	Gas	Tr	Remarks
Gmz1	2	2	6	8	12	6	3		X				LP	
Gmz2	0	0	0	13	8	5	0		X				LP	
Gmz3	5	25	59	56	45	41	35	X	X		X	X	LNP	Movement Dududu--
Gmz4	5	8	8	18	19	15	15		X		X	X	LP	
Gmz5	0	0	7	18	18	11	6	X	X				LP	Excessive crying
Gmz6	0	0	23	51	46	37	33		X	X	X	X	LNP	Prominent nausea
Gmz7	0	0	6	8	9	12	17	X	X		X	X	LP	
Gmz8	0	0	0	13	14	21	15		X		X		LP	Excessive bowl voice
Gmz10	0	3	12	20	25	24	21					X	LNP	Fullness due to gas
Gmz11	0	0	5	7	51	35	25		X				LNP	
Gmz12	0	32	69	51	25	24	20						LNP	No claim
Gmz13	0	7	8	9	7	6	0		X				LP	
Gmz14	5	15	19	18	17	13	11		X			X	LP	Nausea
Gmz15	7	14	17	18	18	16	13		X				LP	
Gmz16	11	12	18	23	34	67	46		X				LNP	Feeling of movement
Gmz17	0	0	0	7	13	18	15						LP	

Gmz18	0	0	4	11	24	15	13	X					LP	
Gmz19	0	0	16	24	50	26	26		X	X		X	LNP	Excessive voice, Naus
Gmz20	0	6	10	17	20	15	12						LP	
Gmz21	0	0	0	8	11	11	9	X	X		X	X	LP	
Gmz22	9	17	19	55	78	45	39		X			X	LNP	Sick after milk drinkin
Gmz23	0	0	41	55	39	36	35	X	X		X		LNP	Sick after milk drinkin
Gmz24	3	6	19	27	32	28	24		X			X	LNP	
Gmz25	0	0	8	13	21	19	18		X				LP	
Gmz26	5	11	16	20	15	14	10						LP	
Gmz27	9	13	34	52	47	41	34		X	X		X	LNP	Feeling burning, Naus
Gmz28	0	12	18	29	32	22	18		X				LNP	
Gmz29	0	0	0	6	13	16	19			X	X	X	LP	
Gmz30	0	5	13	26	38	30	28	X	X				LNP	

Appendix 3: Consent Form

My signature on this consent form means that:

The objectives, methodology, the contribution of the participants in this genetic research have been explained to me, I have been given the chance to discuss it and ask questions. All of my questions have been answered to my satisfaction,

I am aware of the minimal risks and benefits to me of participating in this study,

I agree to allow access to my personal health information as explained in this form,

I agree to drink lactose solution as explained and give information about associated symptoms. I also allow collection of my DNA samples using cotton swab and study data for the research purposes explained in this form, and

I voluntarily consent to take part in this genetic study.

I have been given enough time to think over before I signed this informed consent. It is therefore, with full understanding of the aims of the study that I gave my informed consent and cooperates at my will in the course of the work of the study

Name of Participant	Signature	Date

Investigator obtaining consent:

My signature below signifies that I have explained the objectives of the study and the risks involved to the study participant, and I have answered all questions to the best of my ability.

Name of investigator	Signature	Date

If the participant is illiterate and need assistance during the consent process:

The consent form will be read to the participant. The person signing below attests that the study as set out in this form will be accurately explained to, and appeared to be understood by the participant.

The person signing below acted as a translator for the participant during the consent process.

Language: _____

Name	Signature	Date

Appendix 4: Sequences of Y chromosome DNA genotyped in the study (taken from Hassan, 2008)

IUB code that defines polymorphic site

R = A or G

Y = C or T

K = G or T

M = A or C

S = G or C

W = A or T

H = A, C or T

M1 (YAP insertion) site located at position 253 • within the Alu element (italics). Large font bold alleles are specific to the DYS287 Alu element.

TCACATAATTTCATTTTCCCTATTGCAGATATGTTTTCCTGCATTTGTTAAGGA
ACAAGGGTCTTGAGAGGGAGAGCTTTTTGTCTTAAAGGGGAAGAGATACTTCT
GTGAGGCTAAGAGTTGCCTTTGACTTTGGAGATCTTCACAGGGTATAATAAGA
CAAGCATCAAAGGTAATAGTTTGGGGTCAACTTGACCTGGTTACGTTAATAAG
GAGAGGACTAGCAATAGcaggggaagataaagaaataTA•GGCCGGGCGCGGTGGCTCAC
GCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACGAGGTCAGGAG
ATCGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATACAAAAA
ATTAGCCGGGCGCGGTGGCGGGCGCCTGTATTCCCAGCTACTCGGGAGGCTGAGG
CAGGAGAATGGCGTGAACCC***A***GGAAGCGGAGCTTGCAGTGAGCCGAGATTGCGC
CATtgcagtccgcagtccggccGTCCGGCCTGGGC***A***ACAGAGCGAGACTCCGTCTCAAA…p
olyA tail

M13 = (233 bp) G to C at position 157

TCCTAACCTGGTGGTCTTTCATTGTTTTACAAAGGTGATTTAGTTTTGGGAAGG
ACTATTCTCCTTTAAACTATAGACTAAATTTTTCTCAAAGTTAGGTTAGTTTAT
GCCCAGGAATGAACAAGGGCAGTAGGTAGGTTAAGGGCAAGACGGTTA**S**ATC
AGTTCTCTGTTACTGTTATAATTTTCTCATTGTTATATTTTTTGCAAATGTGGTT
GGATAAAATCATGGCTCA

M42 = (340 bp) A to T substitution at position 297

AAAGCGAGAGATTCAATCCAGGATGACAGAATGCGTTCACCTTTAAAGGGAT
TAAAAGAAGTATAATACAGTCTGTATTATTAGATCACCCAGAGACACACAAA
ACAAGAACCGTGAATTGAATTAGTGGTATACTAATAGAGTGGTTTTACCTGAA
ATATTTACACATCAATCCTACTGAATTCTTACAACAAATGATTTAGATTAGCTA
TTGTATTCACCAGTTGAAAGAACAGAAAATATTGAGGGAGATAACTTGTGTCA
GTGCAACTTAATCAGATTTAGGACACAAAAGC**W**ACTACATAATGAAAAAGA
GAGCTGGTGACTTAACTTGCTAAAA

M60 = (388 bp ancestral); +1 bp insertion (389 bp = derived). Extra T inserted at position 242

GCACTGGCGTTCATCATCTGGGAGCAGCTCAAAAGCCTCTCGCTCAGCCTCCG
TGACGCCCTGGGGGTGTTCAACCCACATATACTGTAAAGACTAGGAGTAGGGT
TGTGGACACCCCACCTCAGCCAACACTGAGCCCTGATGTGGACTCAACCTTGT
AAGGAAAGCTGTAGAGAAATTGGAAGAAAAAATATAAACACATACAGACTCT
GTCTTTACATTTCAAAATGCATGACTTAAAG**T**ATCAGGCACACAGTGGTTACT
CAATGTTGGTCTGTGTCTCTGTAACGTAATATATGTGACTAAATCCCTAAGCTC
TGCTCTTGACCACCCACCTTCTCCAAAAGGGCCTTTCGTAGACGTCGCTCCTCC
TGAACCATAATGAACAT

M89 = (87 bp) C to T at position 64

ACAGAAGGATGCTGCTCAGCTTCCTGGATTCAGCTCTCTTCCTAAGGTTATGT
ACAAAAATCT**Y**ATGTCTCACTTTGCCTGAGTTGC

M78 = (319 bp) C to T at position 24

CACTTAACAAAGATACTTCTTTC**Y**GCCCTTCCAAATATTTCAAAATAAGCTGG
TCATAGTACTTGCTTTTCATAAAAAGATGGTAAGCTTCCAATATTTAGATTTAA
GGAAAGGTGAAGGAACACTATAGCTCTTCATTGATAATATCAAGATTTATACT
GTTCCTTTTATCTATCTCCATCAGAGTTTCAAGAGAAAAAAAATACGATGACT
GTCCATATTCCGGTTCACTGACAGATCATGAACCAGTATCATCAAACCTCATT
ATTATCCGAAATATGTGAAGACAACACTGTGTGGGAGAACCTAGGAAAGTAAT

Appendix 5: Y Chromodome Genotyping

Y Chromosom e Genotypying							
parts	YAP+/ YAP-	M78	M78	M13	M60	M89	M42
Nr1	-Ve					+Ve	
Nr2	-Ve			+Ve			
Nr4	-Ve				+Ve		+Ve
Nr5	-Ve				+Ve		+Ve
Nr6	-Ve				+Ve		
Nr8	-Ve					+Ve	
Nr9	+Ve		T				
Nr10	+Ve	C					
Nr11	-Ve					+Ve	
Nr12	-Ve			+Ve			
Nr13	-Ve					+Ve	
Nr14	-Ve			+Ve	+Ve		
Nr15	-Ve						
Nr16	+Ve	C					
Nr17	+Ve	C					
Nr18	-Ve			+Ve	+Ve		

Nr19	-Ve					+Ve	
Nr20	-Ve			+Ve	+Ve		+Ve
Nr21	+Ve		T				
Nr23	-Ve					+Ve	
Nr25	-Ve			+Ve	+Ve		
Nr26	+Ve	C					
Nr27	+Ve		T				
Nr28	+Ve	C					
Nr29	+Ve						
Nr30	+Ve						
Nr30	YAP+						
Nr31	YAP-					pos	
Nr32	YAP-			pos	pos		
Nr33	YAP+	C					
Nr34	YAP+		T	Pos			Pos
Nr35	YAP+						
Nr36	YAP-					pos	
Nr38	YAP-						Pos
Nr39	YAP+	C					
Nr40	YAP-			pos	pos		

Nr41	YAP-						
Nr42	YAP-						
Nr43	YAP+	C					
Nr44	YAP-			pos	pos		
Nr45	YAP-			pos			Pos
Nr46	YAP-			pos	pos		
Nr47	YAP-			Pos			Pos
Nr48	YAP-			pos			pos
Nr49	YAP+	C					
Nr50	YAP-				pos	Pos	
Nr51	YAP+		T				
Nr52	YAP+	C					
Nr53	YAP+		T				
Nr54	YAP-					pos	
Nr55	YAP+	C					
Gz1	YAP-						Pos
Gz2	YAP-			Pos			
Gz3	YAP-				pos		
Gz4	YAP+		T				
Gz5	YAP-				pos		

Gz6	YAP-			Pos	pos		Pos
Gz7	YAP+	C					
Gz8	YAP+		T				
Gz9	YAP+		T				
Gz10	YAP+		T				
Gz11	YAP-			Pos			Pos
Gz14	YAP+		T				
Gz18	YAP-				pos		
Gz20	YAP-			Pos			
Gz21	YAP-				pos	Pos	
Gz22	YAP-			Pos			
Gz23	YAP-			Pos	pos		
Gz24	YAP-						
Gz25	YAP-			Pos			
Gz26	YAP-						
Gz28	YAP-				pos		Pos
Gz29	YAP-					Pos	
Gz30	YAP-						
Bt1	YAP+	C					
Bt2	YAP-				pos		Pos

Bt3	YAP-					Pos	
Bt5	YAP-			Pos			Pos
Bt6	YAP-					Pos	
Bt7	YAP-				pos	Pos	
Bt8	YAP-					pos	
Bt9	YAP-				Pos		Pos
Bt10	YAP-						Pos
Bt11	YAP-			Pos	pos		
Bt12	YAP+	C					
Bt13	YAP-					pos	
Bt14	YAP+		T				
Bt17	YAP+	C					
Bt18	YAP-				pos	Pos	
Bt22	YAP-				pos		Pos
Bt23	YAP-			Pos			Pos
Bt24	YAP-			Pos			Pos
Bt27	YAP+	C					
Bt28	YAP+		T				
Bt29	YAP-			Pos	Pos		Pos
Bt30	YAP+		T				

Sn2	YAP-			Pos	pos		
Sn3	YAP-						
Sn5	YAP-			Pos	pos		
Sn7	YAP-			Pos			Pos
Sn8	YAP-						
Sn9	YAP-						
Sn12	YAP-			Pos			
Sn14	YAP+		T				
Sn18	YAP-				pos		Pos
Sn19	YAP-						Pos
Sn20	YAP+	C					
Mk1	YAP-			Pos			
Mk2	YAP-						
Mk3	YAP-			pos			
Mk4	YAP-			Pos	pos		
Mk5	YAP-						
Mk9	YAP-			Pos			
Mk10	YAP-						
Mk11	YAP-			Pos	pos		Pos
Mk13	YAP-						

Mk14	YAP-				Pos		
Mk15	YAP+		T				
Mk16	YAP-			pos	Pos		
Mk17	YAP-						
Mk20	YAP-			Pos			Pos

Appendix 6: Haplotype frequencies in populations

==========

Haplotype:	Africans (246)	Middle_East (13)	Asia (6)	Americans (8)	Europeans (16)	Australians (5)
Hap_1	94	12	4	7	9	4
Hap_2	2	0	0	0	0	0
Hap_3	1	0	0	0	0	0
Hap_4	3	0	0	0	0	0
Hap_5	1	0	0	0	0	0
Hap_6	2	0	0	0	0	0
Hap_7	1	0	0	0	0	0
Hap_8	13	0	0	0	0	0
Hap_9	1	0	0	0	0	0
Hap_10	1	0	0	0	0	0
Hap_11	2	0	0	0	0	0
Hap_12	23	0	0	0	0	0
Hap_13	3	0	0	0	0	0
Hap_14	1	0	0	0	0	0
Hap_15	1	0	0	0	0	0
Hap_16	1	0	0	0	0	0
Hap_17	1	0	0	0	0	0
Hap_18	1	0	0	0	0	0
Hap_19	2	0	0	0	0	0
Hap_20	1	0	0	0	0	0

Hap_21	1	0	0	0	0	0
Hap_22	1	0	0	0	0	0
Hap_23	1	0	0	0	0	0
Hap_24	1	0	0	0	0	0
Hap_25	1	0	0	0	0	0
Hap_26	1	0	0	0	0	0
Hap_27	1	0	0	0	0	0
Hap_28	1	0	0	0	0	0
Hap_29	1	0	0	0	0	0
Hap_30	1	0	0	0	0	0
Hap_31	4	0	0	0	0	0
Hap_32	4	0	0	0	0	0
Hap_33	3	0	0	0	0	0
Hap_34	1	0	0	0	0	0
Hap_35	1	0	0	0	0	0
Hap_36	2	0	0	0	0	0
Hap_37	7	0	0	0	0	0
Hap_38	2	0	0	0	0	0
Hap_39	5	0	0	0	0	0
Hap_40	1	0	0	0	0	0
Hap_41	14	1	0	0	1	0
Hap_42	1	0	0	0	0	0
Hap_43	1	0	0	0	0	0

Hap_44	4	0	0	0	0	0
Hap_45	1	0	0	0	0	0
Hap_46	5	0	0	0	0	0
Hap_47	1	0	2	0	0	0
Hap_48	2	0	0	0	0	0
Hap_49	1	0	0	0	0	0
Hap_50	1	0	0	0	0	0
Hap_51	1	0	0	0	0	0
Hap_52	1	0	0	0	0	0
Hap_53	2	0	0	0	0	0
Hap_54	1	0	0	0	0	0
Hap_55	1	0	0	0	0	0
Hap_56	3	0	0	0	5	0
Hap_57	1	0	0	1	0	0
Hap_58	1	0	0	0	0	0
Hap_59	1	0	0	0	0	0
Hap_60	3	0	0	0	0	0
Hap_61	2	0	0	0	0	0
Hap_62	0	0	0	0	1	0
Hap_63	0	0	0	0	0	1
Hap_64	1	0	0	0	0	0
Hap_65	1	0	0	0	0	0
Hap_66	1	0	0	0	0	0

Appendix 7: Neutrality Test Result

Input Data File: C:\...\For Arliqiun analysis*MT-COXII*.nex

Number of sequences: 25 Number of sequences used: 25

Selected region: 1-542 Number of sites: 542

Total number of sites (excluding sites with gaps / missing data): 538

Number of polymorphic (segregating) sites, S: 138

Total number of mutations, Eta: 169

Total number of singleton mutations, Eta(s): 137

Average number of pairwise nucleotide differences, k: 16.790

Nucleotide diversity, Pi: 0.03121

Theta estimated from k: 16.790

Theta estimated from Eta(s): 131.520

Theta estimated from Eta: 44.757

Fu and Li's D* test statistic: -3.64183

  Statistical significance: **, P < 0.02

Fu and Li's F* test statistic: -3.84989

  Statistical significance: **, P < 0.02

Calculated using the total number of mutations

Number of Haplotypes, h: 13

Haplotype (gene) diversity, Hd: 0.923

  Variance of Haplotype diversity: 0.00090

Fu's Fs statistic: 2.372

Strobeck's S statistic: 0.175

  (Probability that NHap <= 13)

Probability that [NHap = 13]: 0.090

Appendix 8: A) NJ Phylogenetic Tree and B) ME phylogenetic Tree

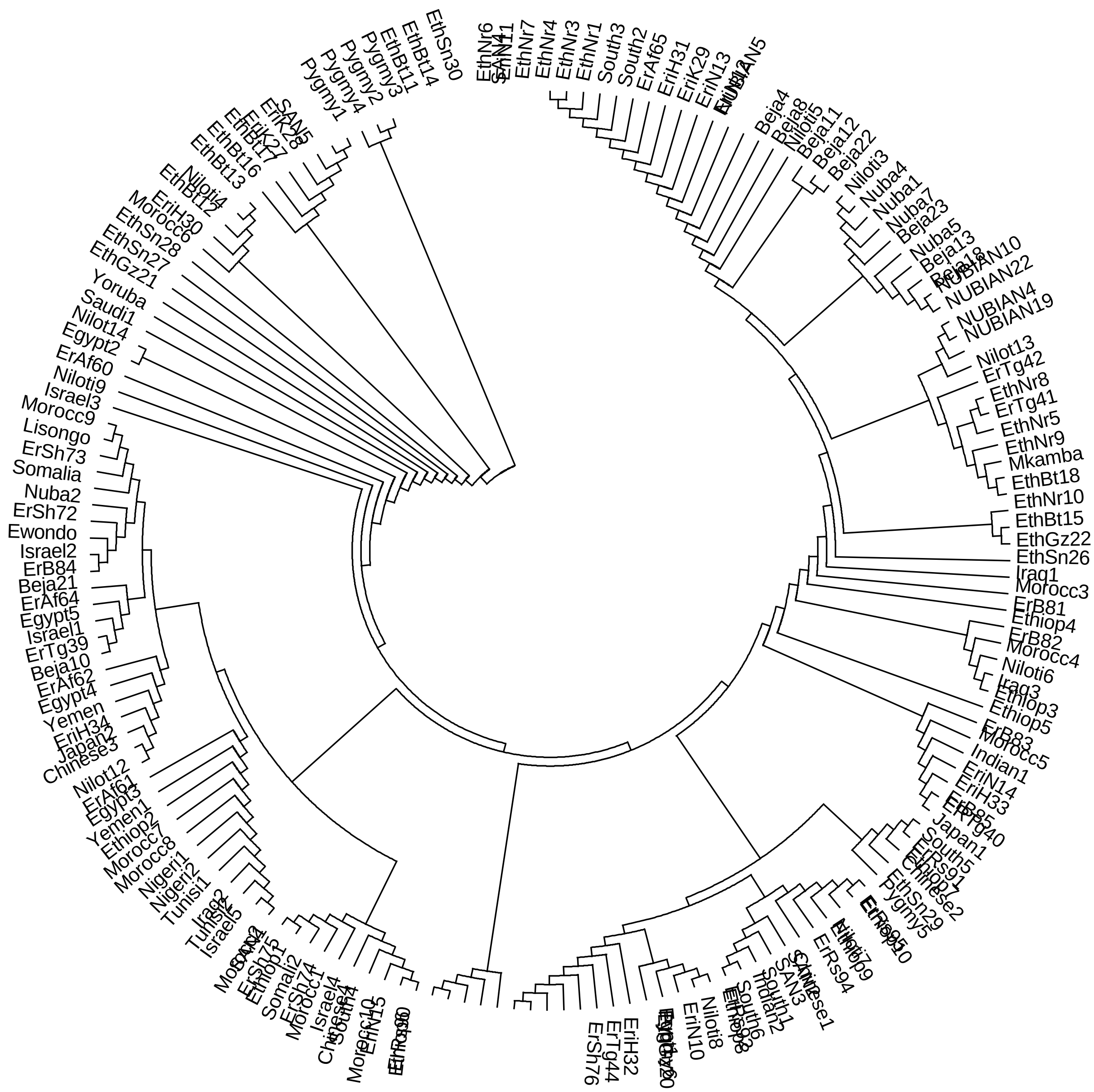

Figure A. Evolutionary relationships using Neighbour-Joining (NJ) method for 166 individuals based on the genetic distances of *MT-COXII* sequences.

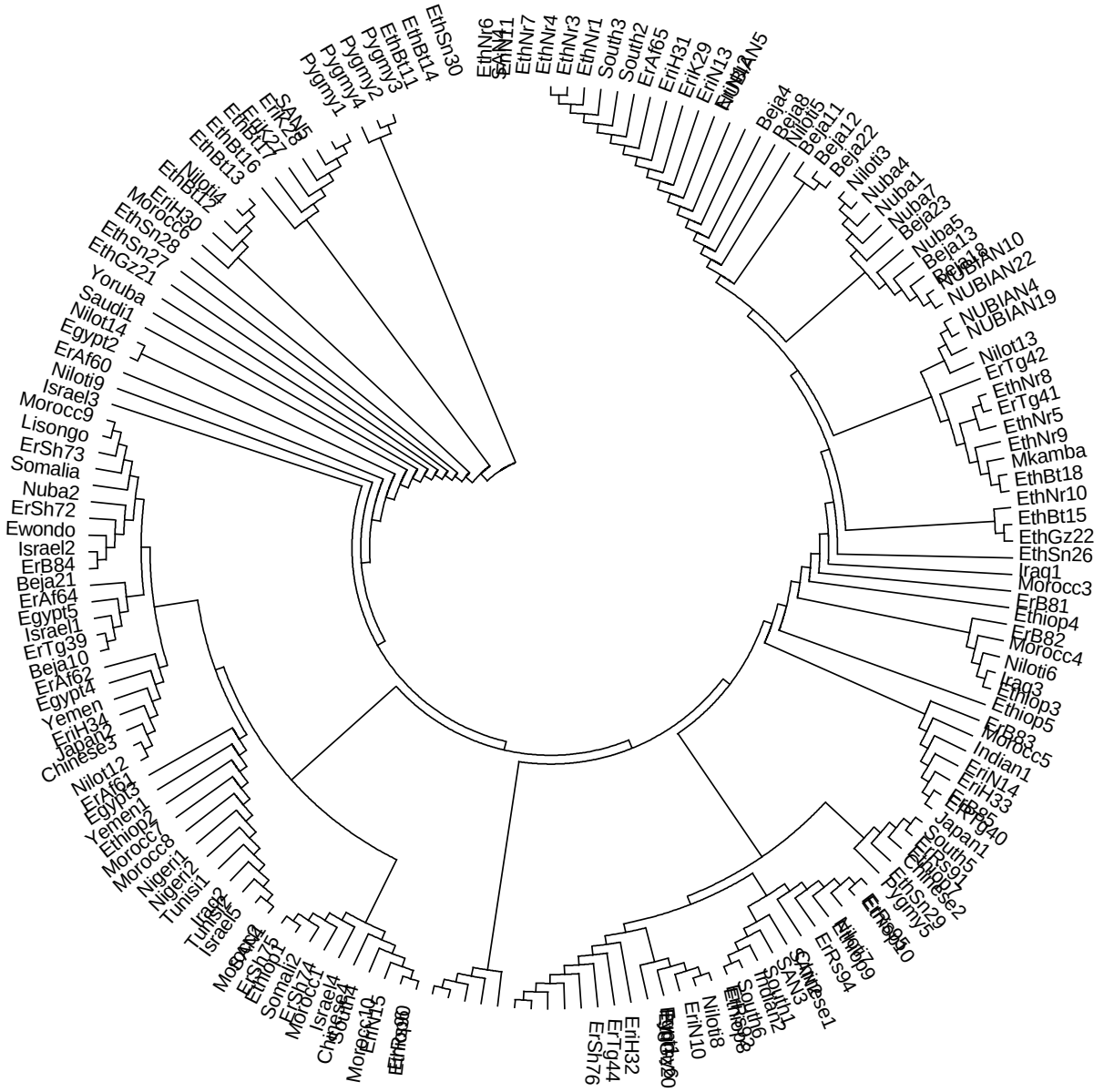

Figure B. Evolutionary relationships using the Minimum Evolution (ME) method for 166 individuals based on the genetic distances of *MT-COXII* sequences.